euronoise

Extraction of vocal tract area function from three-dimensional magnetic resonance images using digital waveguide mesh

Kenji Inoue[a], Hironori Takemoto[b], Tatsuya Kitamura[c], Shinobu Masaki[d] and Hirotake Nakashima[e]

[a]Osaka Institute of Technology, 4-18-15, Tanabe, Higashisumiyosi-ku, 546-0031 Osaka, Japan
[b]ATR Cognitive Information Science Laboratories, 2-2-2 Hikaridai, Seika-cho Soraku-gun, 619-0288 Kyoto, Japan
[c]Konan University, Okamoto 8-9-1, Higashinada, 658-8501 Kobe, Japan
[d]ATR-Promotions, 2-2-2 Hikaridai, Seika-cho Soraku-gun, 619-0288 Kyoto, Japan
[e]Osaka Institute of Technology, 1-79-1, Kitayama, Hirakata, 573-0196 Osaka-fu, Japan
chikuwa.bushi@gmail.com

A method is proposed in this paper to extract the vocal tract area function from three-dimensional magnetic resonance images. The proposed method uses the digital waveguide mesh, an implementation of the finite-difference time-domain (FDTD) method, to simulate wave propagation in the vocal tract from the glottis to the lips. The dimensions of the vocal tract areas are then calculated along the traveling wavefront that emerges from the simulation. Formant analysis has been conducted for Japanese vowels to show the validity of the proposed method. The calculated formant frequencies of the area functions obtained by the proposed method and other existing methods have been compared to those measured from solid models of the imaged vocal tract shapes.

1 Introduction

In recent years, improvements are made on the techniques to produce fine three-dimensional (3D) data of the human vocal tract shapes during speech production using magnetic resonance imaging (MRI) techniques.

Some applications, such as articulatory speech synthesis based on the 3D vocal tract shapes, are, however, difficult in modern computers to be carried out in real-time due to the computational complexity introduced by the 3D model. A solution to the problem is to use an approximated 1D tubular model, described by an area function, estimated from the 3D vocal tract shape.

There have been many methods proposed to estimate the 1D vocal tract area function from the 3D vocal tract shape. [9, 3, 7, 11] While most of these methods try to find the geometrical centerline of the 3D vocal tract shape, the actual path of the acoustic wave propagated in the vocal tract was found not to pass through the geometrical center points above the supraglottis [6, 10]. Nakai *et al.* [7] proposed a technique which exploits the finite element method (FEM) to estimate the area function from the sound intensity at near the first formant frequency.

The present study is intended to address the issue by proposing a method which computes the propagation path of the acoustic wave solving the wave equation in time domain and estimates the vocal tract area function from the centerline along the propagation path. To solve the partial differential equation (PDE) of wave propagation, digital waveguide mesh (DWM) [8, 5] is exploited to discretize the equation.

The proposed and other existing methods have been applied to the MRI dataset of Japanese vowels contained in the "ATR MRI database of Japanese vowel production" [1] for comparison. Lower four formants have been obtained from the estimated vocal tract area functions and the formants have been compared to those measured from the solid models which were created from the same dataset [2].

2 Extraction of vocal tract area funcion

In this paper, as the algorithms that calculate the centerline of the region of the vocal tract shape, the proposed method using DWM and other three methods, namely a method using the Manhattan distance proposed by Takemoto *et al.* [11], the iterative bisection method adopted by Story *et al.* [9], and a method using the nearest points on the opposite edge [4], were implemented and investigated.

The whole process of the estimation of the vocal tract area function from the 3D MRI data was conducted in the following manner. The same procedure is taken for all the methods except the calculation of the centerline.

1. Segmentation of the airway from the surrounding tissue by thresholding.
2. Manual specification of the glottis and lip positions.
3. Segmentation of the main vocal tract region from the branches (e.g. nasal tract).
4. Calculation of the centerline of the vocal tract.
5. Smoothing of the centerline applying the spline interpolation.
6. Determination of the grid lines and their resliced oblique sections locally perpendicular to the centerline.
7. Derivation of the vocal tract area function by counting the number of airway voxels in the oblique sections, starting just above the glottis.

2.1 Airway identification

Regions whose intensity value is below a given threshold are identified as airway regions. Since this thresholding may leave some voxels in the midst of airway region as misidentified as tissue region, such isolated noise voxels within the airway are removed using the connectivity of region.

2.2 Segmentation of the main vocal tract from the branches

The algorithms used in this study require that the branch regions such as the nasal tract, piriform fossa, and epiglottic vallecula are also segmented from the airway region beforehand. Since it is usually obtained manually and no such segmentation algorithm is found in the literature, an automatic branch region detection algorithm is developed in this study as follows.

First, a distance map from the glottis is calculated using the method of Manhattan distance, whose detail is explained in the later section. A distance map $dm(p)$ denotes a correspondent step distance at position p from the glottis. Next, since the main vocal tract region is considered as the region such that the sound wave originated from the glottis to the lips can reach there without backward propagation, the main vocal tract region is ex-

tracted by finding the backward propagation path from the lips to the glottis with the following algorithm.

1. Assign $t = 0$. Define $R(t) = R(0)$ as a set of voxel positions consisting the lip region.
2. For all the positions $p_i \in R(t)$ $(0 \leq i < |R(t)|)$, calculate the set $Q_i = \{q \mid q \in FN(p), q \notin R(t), dm(q) \leq dm(p_i), dm(q) \geq 0\}$, and find the set $R(t+1) = R(t) \cup Q_1 \cup \cdots \cup Q_n$, where $FN(p)$ is the four neighbors in 2D (or six neighbors in 3D) of position p.
3. Iterate procedure 2. until $R(t+1) = R(t)$ satisfies. If it satisfies, the algorithm ends and the $R(t)$ is the voxel positions consisting the main vocal tract region.

2.3 Centerline calculation

In this study, two methods are used as the algorithms which compute the centerline along the propagation path of the acoustic wave; one is a method using the DWM to solve the wave equation of the wave propagation in time domain, and the other is a method using the Manhattan distance [11] as an approximation of the wave propagation solution with lower computational complexity.

Besides them, another two methods are used as the algorithms which compute the centerline from the geometrical center points of the vocal tract region; one is the iterative bisection method [9], and the other is a method using the nearest points on the opposite edge [4].

The method using Manhattan distance and nearest points only use the mid-sagittal plane to compute the centerline (2D), and iterative bisection method uses whole 3D shape. DWM has been implemented in both 2D and 3D versions.

2.3.1 Method using digital waveguide mesh

In the 1D acoustic tubular model of the vocal tract, sound wave propagated in the 3D space is considered as planar. Thus, it is the principle way to project the 3D shape to 1D tube using the actual wave propagation path.

In this study, DWM is adopted to compute the wave equation of acoustic wave propagation. DWM is an implementation of the finite-difference time-domain (FDTD) method which can efficiently compute the numerical solution of the wave equation in time domain.

For sound pressure $p(t, x, y, z)$ in the Cartesian coordinate, 2D or 3D wave equation

$$\frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = \frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 p}{\partial z^2} \tag{1}$$

is digitized and solved with rectilinear mesh (where c is the speed of sound).

In this study, a junction is set to each voxel of the image (Fig. 1). Each junction has a waveguide for each voxel of its four- or six-neighbors.

For a junction J with M waveguide connections, an outgoing pressure to the k-th connection p_k^- is formulized as

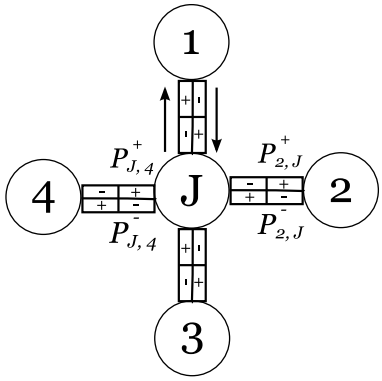

Figure 1: A waveguide junction with four connections.

$$p_k^- = r_k p_k^+ + \sum_{i=1, i\neq k}^{M} (1 + r_i) p_{J,i}^+ \tag{2}$$

where p_k^+ is an incoming pressure from k-th connection, $p_{J,i}^+$ is an incoming pressure from i-th connection a unit time step before, and r_k is a reflection coefficient for k-th connection defined as

$$r_k = \frac{2Y_k - \sum_{i=1}^{M} Y_i}{\sum_{i=1}^{M} Y_i} \tag{3}$$

where Y_k is an admittance of k-th waveguide connection. Impedance Z_k is defined as the inverse number of an admittance Y_k as

$$Z_k = \frac{1}{Y_k} \tag{4}$$

The input pressure of the simulation is given in the initial time step $t = 0$ such that excited pressure 1.0 is set to every voxel consisting the glottis and it is released just after the pressure is given, provided that the total excited pressure 1.0 is equally divided by the number of its waveguide connections to be set to each waveguide (i.e. 0.25 is set for each waveguide at a four-neighbor junction).

In the simulation steps, for each voxel of the main vocal tract region value of distance map $dm(p)$ at voxel position p is set to the step time when the sum of the energy passed through the voxel by the step time first exceeds a certain threshold T_{dm}, regarding this step time as a distance from the glottis. If the threshold T_{dm} is properly configured, contour planes of sound pressure distribution at each step distance from the glottis can be obtained by this method.

In this study, Euclidean norm is used for the summation energy. The value of the distance map $dm(p)$ at voxel position p is thus the step time $t_s = 0, 1, 2, ...$ that first satisfies the condition

$$\sum_{t=0}^{t_s} |p_t(p)| > T_{dm} \tag{5}$$

where $|p_t(p)|$ is an absolute value of the real-valued pressure at point p at step time t.

Centerline is obtained by calculating the centroids for every step distance in the computed distance map. The

centerline, however, contains small but sharp fluctuations, so that the smoothing process explained in a later section is applied.

2.3.2 Method using Manhattan distance

The method that utilizes the Manhattan distance, proposed by Takemoto *et al.* [11], calculates the distance map from the glottis using the Manhattan, or L_1, metric. The distance map computed in this way can be considered as a result of roughly approximated solution of the wave propagation (see Fig. 2).

The distance map for this method using the Manhattan distance, $dm_M(p)$, is given by

$$dm_M(p) = \min_{q \in FN(p)} \{dm_M(q)\} + 1 \tag{6}$$

where p is a voxel position, $FN(p)$ is the four neighbors in 2D (or six neighbors in 3D) of position p, and the value of $dm_M(p)$ is set to 0 for voxels consisting the glottis, and ∞ for voxels not in the main vocal tract region.

2.4 Smoothing

In the two methods using the DWM and Manhattan distance which are based on the computed distance map, centerline is computed by calculating the centroid for each step distance. The centerline computed in this way, however, contains a little variations which results to produce an unstably discontiguous vocal tract area function.

The centerline is, therefore, smoothed by first recalculating the centroids by each predefined step distance D_{smooth} to cut off the number of centerline points to be its $1/D_{smooth}$ points and then applying the spline interpolation.

3 Experiments

The 3D MRI data of the Japanese five vowels /aeiou/, spoken by an adult male Japanese speaker, contained in the "ATR MRI database of Japanese vowel production" [1] is used for the experiments and evaluations of the centerline calculation algorithms.

There are solid models formed by a stereo-lithographic technique [2] for those five vowels and their acoustical characteristics are well investigated. By comparing to the formant frequencies of the solid models, as opposed to those of the recorded utterance which many disturbance factors may involve, the acoustical difference between the 3D MRI data and its 1D derivation are thought to be precisely evaluable.

Firstly, an investigation was conducted to show which values can be used for the parameters of the proposed method using DWM to stably or precisely estimate the vocal tract area function. There are three major parameters for the proposed method: a threshold value T_{dm} to calculate the distance map, waveguide impedances of the air inside vocal tract Z_{air} and the vocal tract wall Z_{wall}. In this study, threshold value T_{dm} to calculate the distance map was configured in all cases as

$$T_{dm}(t) = 0.01/N_g(t^2 + 1) \tag{7}$$

for 2D simulation and

$$T_{dm}(t) = 0.1/N_g(t^3 + 1) \tag{8}$$

for 3D simulation, where t was the step time in the simulation. The impedance of the air Z_{air} was always set to a constant 1 and the impedance of the vocal tract wall Z_{wall} was changed to 1 (same as air), 100, or 10000 (close to real ratio) to see that the calculated contour distance points give proper contour sound pressure planes. Note that all the pressure that goes to the wall junction was set to be lost (no energy back from the wall) in the simulation. Setting the impedances of the air and wall same is unrealistic, but, in this study, only the first wave propagation from the glottis to the lips is considered and steady state is not considered, so that it can be exploited.

Secondly, comparison of DWM 2D, DWM 3D, and other three existing methods were conducted. From the vocal tract area functions obtained by applying those methods, transfer functions were calculated in frequency domain using a transmission line model which included energy losses due to viscosity, heat conduction, and radiation. The yielding wall effect was not adopted (i.e. the wall is rigid) to make the simulation condition same as the reference solid models by Kitamura *et al.* The lower four formants were determined by finding the peaks in the transfer function and compared to those measured from the solid models.

4 Results and discussion

4.1 Investigation of parameter values

Contour distance maps were obtained by the methods using the DWM and Manhattan distance. While most of them show adequate contour sound pressure maps of continuous wave propagation, some conditions for vowel /e/ and /i/ failed to compute them (see Fig. 2 for vowel /i/). In the conditions $Z_{wall} = 100, 10000$ for 2D (left in figure) and $Z_{wall} = 10000$ for 3D (right), scattered dots are depicted just after the places where narrow constrictions occur in the vocal tract.

Considering that the same conditions tend to work for their 3D cases, it can be said that the shortage of the number of waveguide junctions in the constrictions, which is directly derived from the mesh size, introduced a numerical instability in the simulation that resulted to the false contour maps. The square or cubic mesh size was 0.5 mm in the simulation, and the narrowest constrictions were about 4 pixels wide for /e/ and 3 pixels wide for /i/ in the mid-sagittal planes on which the 2D simulations were performed.

As the result of this experiment, the most numerically stable condition over all the vowels in this experiment was the condition $Z_{air} = 1$ and $Z_{wall} = 1$. The condition that says the air and wall impedances are same

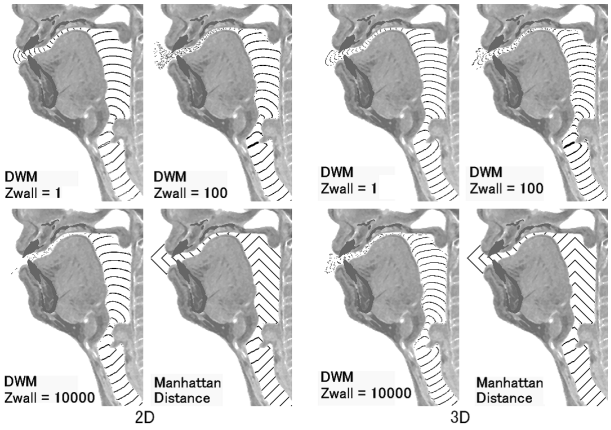

Figure 2: Contour distance maps for vowel /i/ produced by DWM and Manhattan distance methods. Z_{wall} is the wall impedance to the air impedance 1.0.

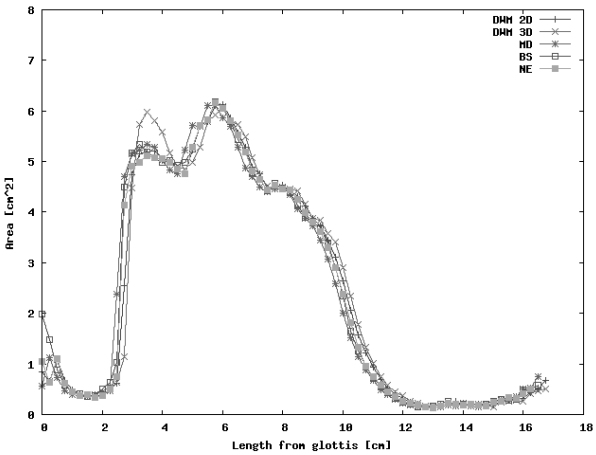

Figure 4: Comparison of vocal tract area functions for vowel /i/.

is unrealistic, but as, in this study, only the first wave propagation is used for centerline calculation, and another experiment not included in this paper showed that the results of varying Z_{wall} were not much different, hence the condition can be considered better for nonsteady-state analysis.

4.2 Vocal tract area functions and their transfer functions

The frequencies of the lower four formants, F1-F4, of the calculated transfer functions were found by peak picking and their four relative percent errors (RPEs), Δ1-Δ4, to those of the reference solid models were computed. The mean sum of the four RPEs, $\sum \Delta/N$, was also computed for each area function (where N = 4 for most vowels, but N = 3 for the vowel /i/ whose first formant of reference was not available).

Vocal tract area functions obtained by all the methods are depicted in Fig. 3 for vowel /a/ and Fig. 4 for vowel /i/. The general form of the area functions are quite similar for all the methods. DWM 3D shows greater area around the junction of piriform fossa due to the fluctuation of the centerline in the sagittal direction.

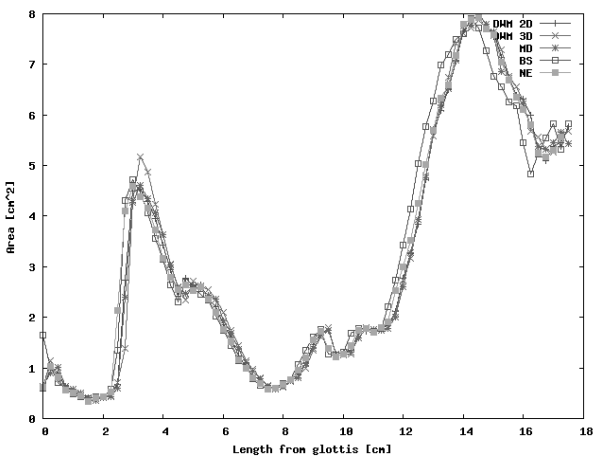

Figure 3: Comparison of vocal tract area functions for vowel /a/.

The comparison of the formant frequencies are shown in Table 1 for vowel /a/ and Table 2 for vowel /i/. The formant frequencies of solid models look like lower than typical values, but it is because of the rigid wall. The first and second formants showed around 20% of relative errors for vowel /a/. The reason is under investigation, but one of the causes can be the existence of piriform fossa.

In general, there were not so big different by method, and the winner method depend on the vowel applied. For example, Manhattan distance was best for vowel /a/ and DWM 3D was best for vowel /i/. Considering that the results are sensitive to various factors, including the smoothing method, it cannot say in this stage that which method is better one.

5 Conclusion

It was shown in this study that the proposed method using the DWM can be used to estimate the vocal tract area function from 3D MRI data. The parameters of the DWM simulation need to be set adequately to conduct the estimation. In this paper, setting both air and wall impedances to 1.0 worked stably with a fixed threshold value for the distance map.

The comparison to the reference solid model, with other existing methods, showed that the proposed method can produce the vocal tract area functions as the same level as other existing methods.

Acknowledgments

The MRI data used in this work is part of the “ATR MRI database of Japanese vowel production”, which was acquired and published by the ATR Human Information Science Laboratories based on the research “Enhancing human-machine communication technologies” commissioned by the National Institute of Information and Communications Technology. The use of the database and the publication of the results are under the license agreement with ATR-Promotions.

	VTL	F1	F2	F3	F4	Δ1	Δ2	Δ3	Δ4	$\sum \Delta/N$
Solid model		450	1070	2407	2696					
DWM 2D	17.75	539	1318	2531	2865	19.8	23.2	5.2	6.3	13.6
DWM 3D	17.75	533	1318	2514	2889	18.4	23.2	4.4	7.2	13.3
MD	17.75	533	1318	2496	2877	18.4	23.2	3.7	6.7	13.0
BS	17.5	539	1307	2467	2906	19.8	22.1	2.5	7.8	13.1
NE	17.75	527	1301	2625	2941	17.1	21.6	9.1	9.1	14.0

Table 1: Comparison of lower four formants on each method against the solid model for vowel /a/. VTL is vocal tract length in cm. Δ values are relative percent error to the solid model.

	VTL	F1	F2	F3	F4	Δ1	Δ2	Δ3	Δ4	$\sum \Delta/N$
Solid model		-	2038	2853	3134					
DWM 2D	17.0	170	2104	2690	3492	-	3.2	5.7	11.4	6.8
DWM 3D	17.0	164	2109	2707	3481	-	3.5	5.1	11.1	6.6
MD	16.5	170	2191	2906	3609	-	7.5	1.9	15.2	8.2
BS	16.75	170	1986	2572	3481	-	2.6	9.8	11.1	7.8
NE	16.75	170	2109	2695	3539	-	3.5	5.5	12.9	7.3

Table 2: Comparison of lower four formants on each method against the solid model for vowel /i/. VTL is vocal tract length in cm. Δ values are relative percent error to the solid model.

References

[1] ATR-Promotions. *Manual of the ATR MRI database of Japanese vowel production*, fourth edition, Apr. 2007.

[2] T. Kitamura, H. Takemoto, and K. Honda. Acoustic characteristics of solid models based on vowel production MRI data. *Technical Report of IEICE. EA*, EA2007-89:19–24, 2007.

[3] B. J. Kröger, R. Winkler, C. Mooshammer, and B. Pompino-Marschall. Estimation of vocal tract area function from magnetic resonance imaging: Preliminary results. In *Proceedings of 5th Seminar on Speech Production: Models and Data*, pages 333–336, Bavaria, Germany, May 2000.

[4] K. Mochizuki and T. Nakai. Estimation of area function from 3-D magnetic resonance images of vocal tract using finite element method. *Acoustical Science and Technology*, 28(5):346–348, 2007.

[5] J. Mullen, D. M. Howard, and D. T. Murphy. Acoustical simulations of the human vocal tract using the 1D and 2D digital waveguide software model. In *Proceedings of the 7th International Conference on Digital Audio Effects (DAFX-04)*, pages 311–314, Naples, Italy, Oct. 2004.

[6] T. Nakai, K. Satoh, and Y. Suzuki. Sound pressure distribution and propagation path in the vocal tract with the supraglottis and the pyriform fossa. *Technical Report of IEICE. HIP*, 98(504):31–38, 1999.

[7] T. Nakai, M. Shota, and S. Tetsuya. Estimation of cross sectional area of 2-d vocal tract by analysis of finite element method. *Technical Report of IEICE. SP*, 102(248):1–4, 2002.

[8] J. O. Smith III. Physical modeling using digital waveguides. *Computer Music Journal*, 16(4):74–91, 1992.

[9] 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*, 100(1):537–554, 1996.

[10] Y. Suzuki, T. Nakagawa, T. Nakai, J. Dang, and K. Honda. Propagation paths in the vocal tract of vowels by finite element method. In *Proceedings of the 2001 Spring Meeting of the Acoustical Society of Japan*, pages 247–248, Tsukuba, Japan, Mar. 2001.

[11] H. Takemoto, K. Honda, S. Masaki, Y. Shimada, and I. Fujimoto. Measurement of temporal changes in vocal tract area function from 3D cine-MRI data. *The Journal of the Acoustical Society of America*, 119(2):1037–1049, 2006.