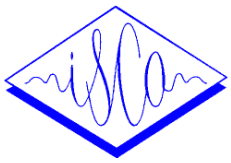


3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ACCURATE MEASUREMENT OF VOCAL TRACT SHAPES FROM MAGNETIC RESONANCE IMAGES OF CHILD, FEMALE AND MALE SUBJECTS

Chang-Sheng YANG and Hideki KASUYA

Faculty of Engineering, Utsunomiya University
2753 Ishii-machi, Utsunomiya 321, Japan

ABSTRACT

We have developed an accurate method to measure vocal tract (VT) shape and length from magnetic resonance (MR) images acquired during sustained phonation of Japanese vowels. The 3-dimensional (3D) VT shape was reconstructed by using coronal MR images for the oral cavity and axial MR images for the pharyngeal and glottal regions. A mid-sagittal image was used as a reference for the reconstruction. MR images of dental impressions of a subject were also incorporated into the reconstruction. All the MR images were directly transferred to a workstation where an intercative program was used to measure the VT shapes. Three Japanese subjects, a child, a female and a male, participated in the experiment. Formant frequencies were computed from the 3D VT shapes which were measured for the three subjects assuming one-dimensional sound wave propagation through the vocal tract. The first three formant frequencies were compared with the ones measured directly from real voice. Except for a few cases, differences between the two measurements were all less than the difference limen (DL) of the formant frequencies.

I. INTRODUCTION

Accurate measurement of vocal tract (VT) shape during phonation is an important part in the study of speech production. Chiba and Kajiyama[1] and Fant[2] used X-ray photography for the measurement of the vocal cavity. Magnetic resonance (MR) techniques have no known harmful side effects, and are capable of measuring 3-dimensional (3D) shapes. In recent years, MR techniques were used to measure 3D VT shapes[3–6]. Baer *et al.*[4] used a gridplane system as a reference to reconstruct 3D VT shapes from coronal and axial MR images. Matsumura and Sugiura[5,6] used only axial MR images to reconstruct 3D VT shapes. In these studies[4,5], however, differences between the formant frequencies measured from the 3D VT shapes and the ones measured directly from original voice samples are far bigger than the difference limen (DL)[7] of formant frequency. This implys that problems still remain with the accuracy of the measurement.

We have devised an accurate method[8] to measure 3D VT shape from multiple directional MR images by referring to a mid-sagittal image. Coronal MR images were used for the oral region and axial MR images were used for the pharyngeal and glottal regions. MR images of dental impressions of every subject were also incorporated into the measurement.

In this paper, we first describe briefly our measurement method, which is used to measure 3D VT shapes and VT lengths of a boy, a female and a male during phonation of Japanese vowels /a, i, u, e, o/. 3D VT shapes of a child have not been measured in previous studies. Formant frequencies were computed from the 3D VT shapes measured for the three subjects to compare with the ones of the real voice. Results show that except for a few cases, the differences between the two measurements are all less than DL.

II. METHOD

2.1 MR image acquisition

A General Electric SIGNA machine (1.5T) was used for the MR image acquisition. A fast SPGR (Spoiled Gradient Recalled Acquisition in the Steady State) volume acquisition pulse sequence was used in this experiment. This sequence was called T1 emphasis gradient echo fast scan method where short TR (repetition time) was used to achieve a higher imaging speed. The parameters for this sequence were as follows: TR = 11.2 ms, TE(echo time) = 4.2 ms, slice thick = 5 mm, gap = 0 mm, number of excitations NEX = 1(NEX = 2 for sagittal images), with image size = 240 × 240 mm^2 and image matrix = 256 × 256 pixels. About 1.43 seconds were required for one slice image (TR × N_y × NEX).

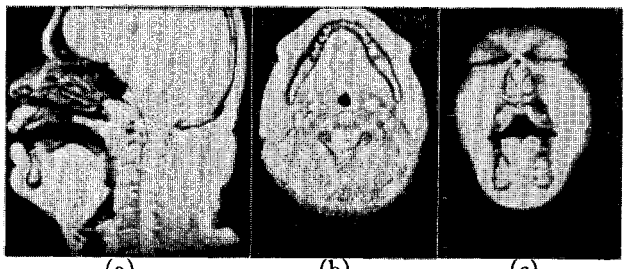
Three Japanese subjects, a male, a female and a boy of 11 years old, participated in this experiment. MR images of the subjects during sustained phonation of Japanese vowels /a, i, u, e, o/ were obtained. For each vowel, image acquisition process was divided into three sessions: (1) A mid-sagittal image was acquired as a reference for the reconstruction of 3D VT shape, and for the calculation of VT length; (2) axial images were obtained for the pharyngeal and laryngeal regions to get accurate shape for these regions; (3) coronal images were gained for the oral cavity. A head coil was used for all these three sessions. If a subject could not sustain a vowel during imaging, a quiet respiration was permitted without changing the vocal tract configuration. No obvious effect of the respiration was observed on the measurement.

For each vowel, 3 slices of sagittal (9sec.), 23 slices of coronal (33sec.) and 27 slices of axial (39sec.) images were acquired. A mid-sagittal image was selected from the 3 slices of the sagittal images.

Examples of MR images of the male subject during sustained phonation of Japanese vowel /a/ are shown in Figure 1.

Since the image data were collected in the three sessions as mentioned above, subject's reproducibility of an identical vocal tract configuration must be examined. According to the study of Kasuya *et al.*[9], once a subject

is accustomed to the phonation of a vowel, variations of the pitch frequency was at most 3% and those of the formant frequencies were within several percentages. Before image acquisition, therefore, the subjects were all trained to reproduce the same articulatory configuration. Errors resulting from the different sessions can be considered negligible in the measurement.



(a) (b) (c)
Fig.1 (a)Mid-sagittal, (b)axial and (c)coronal MR images of /a/ for the male subject.

Because the teeth contain little mobile hydrogen, the measurement of an airway of the oral region may incorrectly include space occupied by the teeth. To solve this problem, dental impressions of the subjects were made at a dental clinic. Dental impressions were placed into a plastic vessel filled with water. Then MR images of the dental impressions in coronal plane with a slice thickness of 1 mm were obtained. These data were used afterwards to compensate for space occupied by the teeth.

All the data were directly transferred from the SIGNA to an HP750 workstation in binary form for further processing.

2.2 Extraction of vocal tract boundary

An interactive program was developed to extract VT boundaries from MR images by utilizing the fact that VT airway is darker than muscles or other tissues. The program was based on binary valued imaging and boundary tracing which were usually used in image processing areas. When a boundary interception occurs, it is modified manually by a mouse pointer in our program.

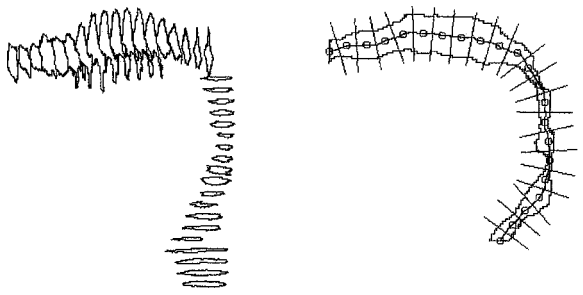
2.3 Reconstruction of 3D vocal tract shape

3D VT shapes are reconstructed from the VT boundaries which were extracted from coronal (the oral region) and axial (the pharyngeal and glottal regions) images. A mid-sagittal image was used as a reference to determine their proper positions. Fig.2(a) shows the reconstructed 3D VT shape of a Japanese vowel /a/ for the male subject.

The positions of the coronal boundaries extracted from the images of dental impressions were determined by referring to the palate contour of the mid-sagittal image for the upper teeth, and to the teeth root contour of the mid-sagittal image for the lower teeth. The coronal boundaries of the vocal tract were then modified by deleting the space occupied by the teeth.

2.4 Measurement of area functions

The mid-sagittal VT boundary was used not only for the 3D VT reconstruction, but also for the measurement of VT length. A center line of the tract was calculated by finding all the mid-point between the upper boundary (boundary that includes palate contour) and the lower boundary (boundary that includes tongue contour) from the glottis to the lips. Fig.2(b) illustrates the center line and the planes perpendicular to it for a Japanese vowel /a/ of the male subject. The length of the tract center line was regarded as VT length.



(a) (b)
Fig.2 (a) Reconstructed 3D VT shape of /a/.
(b) Boundary of mid-sagittal vocal tract of /a/ and its center line.

It is well known that a soundwave front below 4kHz propagates perpendicularly to the center line of the vocal tract. We estimate the cross-sectional area corresponding to the soundwave front.

The center line was divided into several sections at an equi-interval of about 0.85cm (see Fig.2(b)). At the mid-point of each section, a plane that is perpendicular to the center line was made. The VT area was estimated from the tract boundary that lies on each of the plane. By accumulating the area, the area function of the 3D VT is obtained.

An estimated area function of /a/ for the male subject is shown in Fig.3. A short line is drawn near the lips to indicate the area of the lip opening.

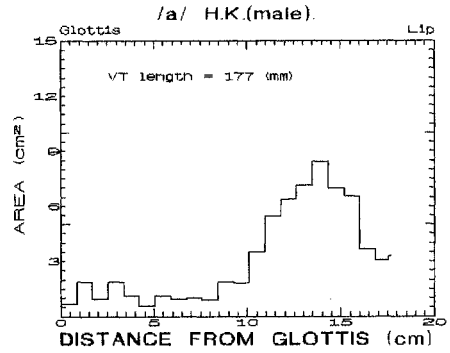


Fig.3 Area function of /a/.

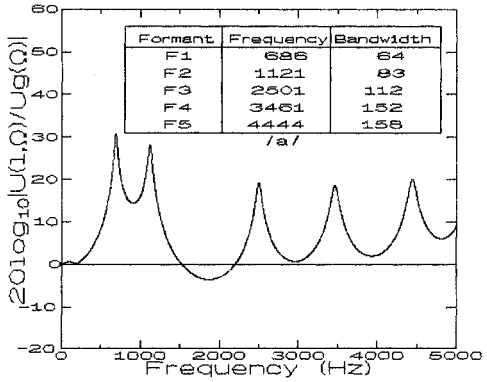


Fig.4 Transfer function of /a/.

Transfer functions of the measured area functions were computed by an algorithm proposed by Sondhi and Schroeter[10], in which losses of the tract wall were considered. Formant frequencies were computed from these

transfer functions. The transfer characteristic computed from the area function of Fig.3 is shown together with the formant frequencies and bandwidths in Fig.4.

2.5 Acoustic analysis

To evaluate the measurement accuracy of VT area functions, formant frequencies of vowel signal were also measured. Sustained phonation of each vowel was recorded in an MRI room just before starting image acquisition. The recorded vowels were intelligible but too much degraded to measure accurately the formant frequencies because a significant amount of noise was generated by the MRI machine during imaging. Therefore vowel utterances were differently recorded in a soundproof room. In the recording, the subjects in supine were required to imitate their own voices given through a headset that were recorded in the MRI room. Formant frequencies were computed from these utterances by means of a 10th order (12th for male subject) LPC analysis of successive 25 ms (30 ms for male subject) long frames with a sampling rate of 10 kHz. The average formant values over 4 seconds long were used as the original formant frequencies.

Table1 Vocal tract area functions of a male.

Section number	Cross-sectional areas(cm^2)				
	/a/	/i/	/u/	/e/	/o/
1	0.668	1.099	1.503	0.483	0.220
2	1.846	1.529	1.749	0.958	0.352
3	0.940	3.771	2.268	0.932	2.479
4	1.863	3.085	3.076	2.232	3.832
5	1.099	2.347	2.944	1.626	1.767
6	0.563	4.913	2.531	1.178	0.949
7	1.090	4.860	2.760	1.336	0.809
8	0.923	3.507	1.723	2.470	0.659
9	1.028	2.795	1.767	1.564	0.747
10	0.914	1.784	1.090	1.125	0.721
11	1.881	1.415	1.310	1.468	1.116
12	1.793	0.510	0.853	0.712	1.002
13	3.498	0.369	0.879	1.002	1.723
14	5.432	0.527	0.642	0.773	3.182
15	6.381	0.457	1.011	0.905	4.888
16	7.146	0.378	1.635	1.397	7.233
17	8.402	0.519	2.312	1.916	8.446
18	6.935	0.817	1.986	2.136	6.785
19	6.521	2.171	3.639	4.219	6.231
20	3.665	2.892	3.313	3.700	3.252
21	3.041	2.681	0.554	1.811	0.703
22	3.252	2.944	0.984	1.951	1.389
L	17.7	18.2	17.7	17.5	18.2
Δl	0.843	0.868	0.843	0.834	0.865

III. AREA FUNCTIONS ACROSS MALE, FEMALE AND CHILD SUBJECTS

Tables 1-3 illustrate the VT areas of the Japanese vowels /a, i, u, e, o/ measured for the three subjects. The VT areas are represented as a concatenation of uniform acoustic tube sections of an equi-length. The sections are numbered from the glottis to the lips, indicating the last one to be the area of the lip opening. In Tables 1-3, L (in cm) represents a vocal tract length and Δl (in cm) the length of an unit section. Note that Δl is different for each area function to avoid a round off error for VT length calculation.

Formant frequencies computed from the area functions shown in Tables 1-3 and the ones estimated from the original utterances are listed in Table 4.

Table2 Vocal tract area functions of a female.

Section number	Cross-sectional areas(cm^2)				
	/a/	/i/	/u/	/e/	/o/
1	1.090	0.747	1.301	0.721	1.881
2	2.074	2.109	3.076	0.773	2.830
3	0.668	2.619	2.575	1.301	1.969
4	0.800	4.966	2.962	1.749	0.703
5	0.844	4.869	3.234	2.355	0.721
6	0.949	6.073	2.918	5.247	0.360
7	1.063	4.052	2.417	5.001	0.308
8	1.230	2.531	1.547	3.226	0.624
9	2.909	1.072	0.334	2.197	1.758
10	6.513	0.563	0.598	1.863	3.437
11	8.815	0.281	1.345	1.468	7.031
12	10.362	0.290	1.881	1.644	9.088
13	9.097	0.519	2.522	1.696	8.112
14	8.622	1.907	3.067	1.758	8.728
15	8.332	2.918	2.856	3.419	6.231
16	6.759	2.769	1.521	4.271	3.806
17	5.546	1.626	0.264	3.349	1.424
18			0.729	3.243	1.802
L	13.9	13.9	14.3	14.2	14.8
Δl	0.871	0.869	0.844	0.833	0.873

Table3 Vocal tract area functions of boy.

Section number	Cross-sectional areas(cm^2)				
	/a/	/i/	/u/	/e/	/o/
1	1.274	0.598	0.571	1.266	1.125
2	1.898	1.354	3.006	0.879	1.688
3	1.547	2.479	3.981	1.002	2.470
4	1.178	3.199	2.997	2.109	1.617
5	0.976	3.727	2.953	1.978	1.116
6	0.668	3.665	2.514	1.441	0.905
7	0.694	2.329	1.415	1.819	0.729
8	0.615	1.336	1.512	1.898	0.334
9	4.043	0.870	1.705	1.213	1.758
10	6.337	0.765	0.738	0.958	3.041
11	7.646	0.712	0.563	1.257	3.990
12	7.875	0.290	0.694	0.747	3.938
13	6.539	0.299	0.606	0.870	5.142
14	6.750	0.747	1.239	1.925	4.843
15	6.170	0.589	2.848	1.679	5.309
16	4.368	1.160	2.417	1.362	4.632
17	3.164	2.259	0.457	1.846	1.213
18			1.248		0.431
19					0.896
L	13.4	13.3	14.4	13.7	14.9
Δl	0.840	0.831	0.848	0.854	0.829

Perceptual difference limen (DL) of formant frequencies of vowels is about $3 \sim 5\%$[7]. Except for F1 of /i/ (22Hz, 7.2%) of the male subject, F1 and F2 of /a/ (79Hz, 7.8%; 70Hz, 5.2%) of the female subject, F2 of /a/ (75Hz, 5.8%) and F3 of /i/ (240Hz, 6.8%) of the boy subject, differences between the measured values and the original ones of the first three formant frequencies are all less than 5%. The measurement error is somewhat large for /a/ of the female subject.

IV. DISCUSSION

A SIGNA machine was used to acquire MR images of 3D VT shapes of the five Japanese vowels across male, female and child subjects. This machine provided high air-tissue contrast images with fast imaging (1.43second/slice). Data were directly transferred to an HP workstation for measurement of the 3D VT shapes. An interactive program was developed to extract VT boundaries.

Table4 Comparison of original and Measured formant frequencies in Hz.

	For the male subject							
	Original				Measured			
	F1	F2	F3	F4	F1	F2	F3	F4
/a/	703	1105	2420	3475	686	1121	2501	3461
/i/	306	2007	2482	3257	328	2098	2584	3431
/u/	449	1268	2257	3337	430	1313	2296	3357
/e/	557	1702	2286	3574	543	1724	2314	3450
/o/	534	839	2470	3354	522	874	2558	3692
	For the female subject							
	Original				Measured			
	F1	F2	F3	F4	F1	F2	F3	F4
/a/	1012	1343	3251	4148	933	1413	3215	4215
/i/	353	2965	3243	4359	368	3004	3390	4589
/u/	412	1138	3072	4552	424	1148	3049	4163
/e/	633	2460	3118	4295	633	2409	3059	4182
/o/	593	905	3292	4117	601	936	3389	4168
	For the boy subject							
	Original				Measured			
	F1	F2	F3	F4	F1	F2	F3	F4
/a/	862	1292	3641		834	1367	3728	4288
/i/	423	2737	3541		412	2637	3781	4529
/u/	451	1701	2768	4246	441	1664	2634	4325
/e/	590	2210	2984	4091	613	2152	3055	4044
/o/	493	967	2842	3916	517	990	2980	4045

(Underlines indicate the error be less than 5%)

To obtain accurate 3D VT shapes, image data acquisition process was divided into three sessions: for midsagittal, coronal and axial images respectively in this experiment. The variations of the images in the different sessions were considerably small once the subjects were trained to stably phonate the same vowel[9]. The longest session took 39 seconds for image acquisition. No obvious effect of the respiration was observed in this experiment.

Mid-sagittal images of each vowel of the subjects were not only useful for reconstruction of 3D VT shapes from boundaries extracted from coronal and axial images, but were also effective for VT length measurement.

The area functions corresponding to the soundwave front were estimated from the reconstructed 3D VT shapes along the center line at an equi-interval. By using the equi-interval, a round off or truncated error of VT length was removed in the measurement process.

We calculated formant frequencies based on a planar sound wave propagation model[10]. According to the FEM simulation by Lu *et al.*[11], it is a good approximation in the frequencies below 4kHz.

Using the estimated formant frequencies and the original pitch patterns, we synthesized the vowels /a, i, u, e, o/ to compare their qualities with the original ones. It turned out that they were very similar to the original vowels. This gives us a perceptual evidence for the reliability of our measurement methods of vocal tract shapes.

V. SUMMARY

We have developed an accurate method to measure 3D VT shapes from MR images. VT shapes of a boy, a female and a male subjects during sustained phonation of Japanese vowels /a, i, u, e, o/ were measured by using this method. The first three formant frequencies measured from the original utterances were compared with those computed from the 3D VT shapes. Except for a few cases, the differences between them were less than DL. Vowels synthesized by using the formant frequencies computed from the 3D VT shapes and the original pitch patterns were very close in phonetic quality to the original vowels.

Data obtained by this work are indeed useful for the study of speech production and speech synthesis.

ACKNOWLEDGMENTS

The authors are very grateful to Dr. Sigeru Kano of National Hospital of Tochigi and Dr. Toshihiko Sato of Washiya Hospital for MRI data acquisition, Dr. Hisao Takahashi for making dental impressions. This work was partly supported by Grant-in-Aid for Scientific Research "Speech and Dialogue," from the Ministry of Education, Science and Culture of Japan.

References

[1] T. Chiba and M. Kajiyama, "The Vowel, Its Nature and Structure," Tokyo-Kaiseikan, Tokyo, 1941.

[2] G. Fant, "Acoustic Theory of Speech Production," Mouton, The Hague, The Netherlands, 1960.

[3] M. Rokkaku, S. Imaizumi, S. Niimi and S. Kiritani, "Measurements of the three dimensional shape of the vocal tract on the magnetic resonance imaging technique," Ann.Bull. RILP 20:47-54, 1986.

[4] T. Baer, j. C. Gore, L. C. Gracco and P. W. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging : Vowels," J. Acoust. Soc. Am., **90**, 2, 799-828(1991).

[5] M. Matsumura and A. Sugiura,"Modeling of 3-Dimensional Vocal Tract Shapes Obtained by Magnetic Resonance Imaging for Speech Synthesis," Proc. ICSLP, 425-428(1990).

[6] M. Matsumura, "Measurement of three dimensional shapes of Vocal Tract and nasal cavity using magnetic resonance imaging technique," Proc. ICSLP92, 779-782(1992).

[7] J. L. Flanagan, "A difference limen for vowel formant frequency," J. Acoust. Soc. Am., **27**, pp.613-617(1955).

[8] C.S. Yang, H. Kasuya, S. Kanou and S. Satou, "Considerations on Measurement Method of Vocal Tract Shapes Using Magnetic Resonance Imaging," Trans. IEICE (in print, 1994). (in Japanese)

[9] H. Kasuya, H. Suzuki and K. Kido, "Changes in Pitch and First Three Formant Frequencies of Five Japanese Vowels with Age and Sex of Speakers," J. Acoust. Soc. Jpn. (J) **24**, 6, 355-364 (1968). (in Japanese)

[10] M. M. Sondhi and J. Schroeter, "A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer," IEEE Trans. Acoust. Speech Signal Process. **35**, 7, 955-967(1987).

[11] C.X. Lu, T. Nakai and H. Suzuki, "A Three-Dimensional FEM Simulation of the Effects of the vocal Tract Shape on the Transfer Function," Proc. ICSLP93, 771-774(1992).