

MEASUREMENT AND SYNTHESIS OF THE VOCAL TRACT OF KOREAN MONOPHTHONGS BY MRI

Byunggon Yang
Donguei University

ABSTRACT

The author collected and compared midsagittal, coronal, coronal oblique, and transversal images of Korean monophthongs /a, i, e, o, u, i, ʌ/ produced by a healthy male speaker by MRI. Then, area was measured by computer software after manually tracing the cross-section at different points along the tract. Results showed that the width of the oral and pharyngeal cavities varied compensatorily with each other on the midsagittal dimension. Formant frequency values estimated from the area functions showed a strong correlation with those analyzed from the spoken vowels. Moreover, almost all of 35 subjects who listened to the synthesized vowels using the formant values estimated from area data perceived the synthesized vowels as the equivalent spoken ones. The results may be helpful to verify an exact measurement of the vocal tract area through the vowel synthesis and to make a simulation study before any surgery on the abnormal vocal tract area.

1. INTRODUCTION

Vowels are produced by varying vocal tract shapes moving such articulators as tongue, jaw, and lips. Jaw movement leads to the expansion or contraction of oral cavity while the tongue controls the size of the front, mid, or back part of the cavity. The tongue with its constant volume moves forward or backward to vary the oral and pharyngeal cavities. Lips play a role in producing round vowels. Those articulators modify a speaker's vocal tract shape to give various auditory impressions to a listener. Quantitative study on the vocal tract shape may be helpful in the understanding of complex vowel qualities.

Recently many studies on the various vocal tract shapes using Magnetic Resonance Imaging techniques have been made because of their advantages: They are not harmful to the body and multiplane images may be taken [1, 2, 3]. Disadvantages are low resolution images after a fast shot, and lack of real time representation of continuous speech. Baers et al. compared formant values estimated from the MRI vocal tract area function of two speakers and those from the actual speech production [1]. There were a few discrepancies between the formant values but they found that the synthesized vowels sounded similar in their perceptual study. Demolin et al. measured MRI images of French oral vowels produced by two male and two female speakers [4]. They did not find any systematic relation among the four speakers because of the different vocal tract lengths but observed some constriction points at the velo-pharyngeal region for back vowels. The groove size of the tongue was found to be important for front vowels but not back vowels. They reported difficulty finding the border of air and tissue. Yang and Kasuya studied three males and three female Japanese speakers' production of five vowels [3]. They compared normalized vocal tract length dividing the oral, pharyngeal, and glottal region. Their results

showed that vocal tract dimensions for male and female speakers are continuously distributed and the vocal tract length and the ratio of the oral cavity to the pharyngeal cavity lead to the individual differences. They compensated for the area formed by the teeth after modelling the speakers' teeth. They pointed out that formant values did not significantly vary with or without tooth information. Gracco et al. analyzed three dimensional views of normal and abnormal people [5]. They could predict fairly well the formant values of patients with or without tongue root, epiglottis, and false vocal folds. This suggests a possibility of simulating speech output before any physical removal of the vocal tract section.

The purpose of this study is (1) to gather midsagittal images of vowels; (2) to obtain cross-sectional area of the vocal tract from the glottis to the lips; (3) to find correlational coefficient between the formant values estimated from the area function and those from the actual pronunciations; (4) to test auditory impressions of the synthesized vowels from the formant values estimated from the area; (5) to investigate how each vowel varies from the average values of the corner vowels /a, i, u/.

2. METHODS

The author collected and compared midsagittal, coronal, coronal oblique, and transversal images of Korean monophthongs /a, i, e, o, u, i, ʌ/ produced by a normal male speaker using 1.5T MR, Vision. The area was obtained while the speaker was producing the Korean vowels for 19 seconds in the supine position. Pulse sequences for image acquisition are Flash 2D, TR 100 ms, TE 5 ms, Flip angle 30°, Matrix 98x128, FOV 130 mm, NEX 2, Slice thickness 10 mm. Before obtaining the cross-sectional area, a midsagittal image was taken to locate proper points of imaging around which area varied greatly along the vocal tract. Four coronal cuts cover the oral cavity including a lip cut. Two or three coronal oblique cuts cover the oro-pharynx and velum regions. Three transversal cuts cover the pharynx and larynx. The 3rd to 5th cervical vertebrae were traced to correctly overlap midsagittal images for comparison. The top-most trace of the 5th cervical vertebra was connected to the front most part of the glottis and designated as the vocal fold. Then, area was measured after manually tracing the cross-section at different points along the tract. Cross-sectional area was drawn by a pencil on a transparent paper on a window. The author consulted a doctor in a hospital when any suspicious border occurred. Then, midsagittal images and cross-sectional areas were scanned into a computer to measure area by computer software, Area Properties(v. 3.2). The software counted the screen pixels of the image and scaled them to the actual area. A rectangular area was tested to confirm that the program measures the area correctly. The formant values arising from the vocal tract area were determined by Formfrek [6]. The program calculates formant values by finding zeroes from the system radiation

phase and angle. It corrects radiation effects from the lips. Five formant values from the lowest were collected.

The formant values from the actual speech were obtained by SFS(Speech Filing System) at 1/3 point of the total vowel duration because that point shows sustained formant patterns [7]. The speaker produced the vowel in the supine position after the MRI session. Speech syntheses were made by formant synthesis software, SenSyn1.0(Sensimetrics). Amplitude, duration, pitch variation for the vowel synthesis were measured for each vowel and formant values estimated from the MRI area function were input to the synthesis software. Perceptual tests were done in a quiet classroom at a comfortable level. Listeners circled one of the seven vowels on the answer sheet while listening to each stimulus.

3. ANALYSIS AND DISCUSSION

3.1 Comparison of midsagittal images. The vocal tract length of the speaker was measured from the midsagittal images. A Hypertalk program was made to automatically add up the total length of the vocal tract by tracing its central line. The total length was measured to the edge of lips in the midsagittal images. However, the cross-sectional area was measured just a little behind the lip ends because it was hard to measure the area at the lip ends, the section of which showed only partial area. In the procedure, the central line of the tract around epiglottis, velum, between the lower teeth and tip of the tongue abruptly changed its direction but a natural curve was used to trace the length.

The rounded vowel /u/ was 18 cm long. That of the vowels /e/ and /i/ is 16.2 cm long. An average value for the seven vowels was 16.9 cm. Thus there is not much length difference except the rounded vowel /u/. Then, midsagittal images of the seven vowels were compared. Figure 1 shows the images of the three corner vowels /a, u, i/. The third and fifth cervical vertebrae were employed to overlap them. The subject's head was fixed but still a little vertical movement was possible. From Figure 2, one can notice that for the vowel /i/, the tip of tongue approaches to the hard palate closely. The pharyngeal section forms a wider area. On the other hand, the tongue and jaw for the vowel /a/ move down greatly to form a wider oral cavity and narrower pharyngeal cavity.

The distance between the tongue and pharyngeal wall of the vowels /a, i/ shows a difference of 0.8 cm. Lips for the back vowel /u/ were protruded 1 cm more than those of the vowel /i/. The subject produced the vowel /a/ by both raising the palatal area and lowering the jaw. Conventionally the palatal region was always fixed and only jaw movements were observed to discuss a midsagittal image of each vowel production. Also, the subject made most of the frontal part of the vocal tract to secure proper cross-sectional area for the front and back vowels. The vowel /i/ has a similar vocal tract shape like the vowel /i/ but the frontal section of the tongue is wider by 1.5 cm thus the tongue volume moves backward with a narrower pharyngeal cavity. The vowel /i/ with the tip of the tongue spread to the lower teeth is different from the vowel /u/ with the tip of the tongue drawn backward to form a pocket between the lower teeth and the tip of the tongue. Difference between the rounded vowels /u/ and /o/ exists in the protruded lips and the size of the pocket formed. The vowels /Λ/ and /a/ have similar pharyngeal width but the jaw opens more for the



Figure 1 The images of the three corner vowels /a, u, i/ collapsed. The thickest line denotes the vowel /i/; the thinnest line indicates the vowel /a/; the medium line describes the vowel /u/.

vowel /a/. The vowels /o/ and /e/ have similar pharyngeal width but the lower lip goes down more for the vowel /e/.

So far we have compared the width of labial, oral, pharyngeal cavities from the midsagittal images. For the production of vowels, tongue movements seem very important. Especially, since the tongue has a constant volume, the tongue movement leads to a negative correlation between the width of oral and pharyngeal cavities. Because the jaw moves diagonally to the glottal area, the amount of area change increases with the distances from the glottis. The greatest width of lips was observed for the vowel /a/ followed by /e, Λ, i, i, o, u/. There is not much variation in the width of oral and pharyngeal cavities from the midsagittal images. Also, the length variation seems small enough to be perceptually negligible. When one considers that 1 cm difference in the vocal tract length led to about 200 Hz for F₃ [7] which causes little perceptual difference [8], the length seems to contribute little to the vowel variation.

3.2 Comparison of vocal tract area. Table 1 lists the distance from the glottis and its area at each crossing point in the vocal tract. If one looks at the area difference, the vowels /i/ and /e/ at 7.5 cm from the glottis differ from each other around 6 cm²; around 3.3 cm², at 14 cm. The pharyngeal area difference is almost twice that of the oral cavity. For vowel /i/ and /i/, the difference at 7 cm amounts to 8 cm². Compared to the width difference of 1.5 cm in the midsagittal images, the area results in greater difference. Thus, midsagittal comparison alone may not be sufficient to reflect the anatomical difference. The back vowels /Λ, o, u/ form constriction points around 8~9 cm from the glottis. The vowel /a/ has a lower

e		i		Λ		a		i		o		u	
dist	area	dist	area	dist	area	dist	area	dist	area	dist	area	dist	area
0.00	2.88	0.00	2.16	0.00	2.78	0.00	2.10	0.00	2.60	0.00	2.53	0.00	2.99
0.87	1.59	0.96	2.48	2.34	6.59	1.75	1.68	3.01	2.71	1.86	3.91	1.65	8.03
3.19	5.85	2.96	5.29	4.87	2.67	4.12	2.75	4.56	7.19	3.87	3.69	3.82	8.51
4.80	6.09	4.71	5.87	7.25	2.95	6.66	1.87	7.27	10.04	5.60	1.80	5.45	5.22
7.33	3.89	7.62	1.40	8.48	2.24	8.23	2.19	8.77	4.70	7.72	1.51	7.42	1.71
8.86	4.04	9.18	1.45	10.90	6.01	10.90	7.67	10.80	2.56	9.37	1.35	9.24	1.49
10.60	2.80	11.10	2.39	12.20	8.28	12.40	10.98	12.30	0.84	11.52	6.91	11.70	3.96
12.10	2.52	12.60	2.97	13.80	6.82	14.20	12.29	13.90	0.71	13.28	10.04	13.60	13.04
13.80	4.05	13.90	4.44	16.50	7.66	16.70	8.99	15.10	0.75	15.01	7.06	15.20	6.10
16.20	8.30	16.20	3.27					17.10	4.30	17.38	2.94	18.00	0.74

Table 1. Vocal tract areas of the seven Korean vowels at the distance(dist) from the glottis.

constriction point (6.7 cm) because the subject lowered the jaw. The front vowel /i/ has a constriction point at 14 cm. For the vowel /e/ it is at 12 cm in the middle of the oral cavity. The vowel /i/ has a constriction point at 8 cm, thus it might be closer to the back vowel /u/. The vowel /Λ/ has similar area difference for the vowel /a/ at both oral and pharyngeal cavity. Area difference for the vowel /i/ and /u/ is about 1 cm². Lowering the back portion of the tongue widens both the oral and pharyngeal cavity by 3 cm².

How does the volume of the vocal tract vary with each vowel? To check the volume, data for each area were summed up by using the Area Properties. The volume varies from the vowel /o/ (78.8 cm³) to the vowel /i/ (103.4 cm³). The general trend shows that the high vowels have smaller volume than the low vowels. This might come from the combination of tongue and jaw movements.

We obtained 17 cm³ at the pharyngeal section at 8 cm from the glottis while 74 cm³ at the oral section for the vowel /a/. For the vowel /i/, the pharyngeal section amounts to 58 cm³, and the oral section, 28 cm³. There is around 5 cm³ difference in volume. Thus, the tongue and jaw movements seem to change the ratio of oral and pharyngeal cavities but the total volume does not vary much. For the vowel /u/, the two cavities are divided into about the same volume (43 cm³) at 9 cm from the glottis. Thus, tongue movement seems to play a greater role in the volume control of the vocal tract.

3.3 Formant value estimation by area function and synthesis. Table 2 lists formant values analyzed from actual speech by SFS as well as those from the area functions. The values were input to the formant synthesizer. The synthesized vowels were presented to fifty students with normal hearing ability in a quiet classroom at a comfortable level. On an answer sheet, three sets of seven vowels were randomly presented. They listened to the stimuli and circled the vowel they hear. To avoid the initial confusion, only the answers from the 9th row were counted. Also fifteen students who could not tell the vowel /Λ/ from the vowel /i/ were rejected because their decision would be done by chance for the

vowels. Five cases were reported in which the vowel /o/ was heard as the vowel /Λ/. One case showed a confusion on the vowels /i/ and /o/. Almost all of the synthesized vowels were heard as the corresponding vowels.

Vowel	Formant from Speech			Formant from Area		
	F ₁	F ₂	F ₃	F ₁	F ₂	F ₃
a	744	1167	2816	762	1243	2691
e	544	2027	2672	492	1726	2487
Λ	526	1120	2818	545	1229	2482
o	412	738	2620	473	928	2727
i	311	1387	2445	430	1387	2803
u	296	706	2344	244	729	2802
i	284	2377	3093	264	2141	2679

Table 2. Formant values obtained from actual speech and estimated from area function.

This result was supported by a strong correlation coefficient (0.978) between the formant values estimated from the area function and actual speech. Here F₃ values show some difference but considering that 800 Hz variation in F₃ did not affect the auditory impression of vowels, the results seem to be expected.

If the constriction points of the vocal tract moved, what would be the resulting formant frequency and auditory impression of the synthesized vowel? This study may be helpful to simulate patients with cancerous cells or edema on the vocal tract before their removal [7]. In fact, vocal tract area can vary greatly but there is a physical limitation to the movement of articulators. Especially, there are some regions in which small articulatory variation leads to large acoustic change in formant values and vice versa [9, 10]. For example, there was not much variation in synthesized vowel qualities even though the constriction point moved around the velum region. Thus one may need a standard point to start with

because otherwise the combination of the size of oral and pharyngeal cavities would be unlimited. Figure 2 shows the chart of the corner vowels /a, u, i/ and their average.

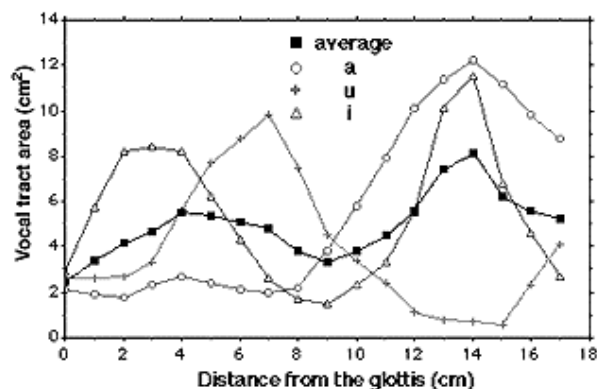


Figure 2 The area of the three corner vowels /a, u, i/ and their average.

In Figure 2 a constriction point of the average was made at 9 cm from the glottis. Subjects answered to the perceptual stimulus of the synthesized vowel estimated from that area function as /Λ/ (549 out of 560 responses). Seventeen subjects heard it as the vowel /i/ while two as the vowel /o/. If one looks at the three vowels from the average area values, the area variation should be made by compensating the one cavity for the other. In other words, since the total vocal tract is somewhat constant, if the oral cavity is reduced by the tongue raising, then the pharyngeal cavity must be expanded accordingly. This way one could account for the physical limitation and naturalness in the articulatory movement.

4. CONCLUSION

In this study, midsagittal images and area values of the seven vowels produced by a healthy Korean male were examined. Then, perceptual experiment on the synthesized vowels estimated from the area function was made. Also, the average values of the three corner vowels were determined to reflect the physical limitation to the area variation. Results showed that the width of the oral and pharyngeal cavities varied compensatorily with each other on the midsagittal dimension. Formant frequency values were estimated from the area functions of the seven vowels using Formfreq. Statistically, the values showed a strong correlation with those analyzed from the spoken vowels. Moreover, almost all of the subjects who listened to the synthesized vowels using the formant values estimated from the area data perceived the synthesized vowels as the equivalent spoken ones. Movements of constriction points of the vowel /u/ with a wider lip opening sounded /i/ and led to slight changes in vowel quality. The formant values estimated from the average area function of the three corner vowels /a, i, u/ were employed to synthesize a vowel which was identified as the vowel /Λ/ by almost all of the subjects. The volume of each vowel was estimated from the integral sum of the area along the tract. Jaw and tongue movements led to the major volume variation within the anatomical limitation. Each corner vowel varied

systematically from a rather constant volume of the average area of the three corner vowels. Thus, the author proposes that any simulation studies related to the vocal tract area variation reflect its constant volume. The results may be helpful to verify an exact measurement of the vocal tract area through the vowel synthesis and a simulation study before any surgery on the abnormal part in the vocal tract.

REFERENCES

- [1] Baers, T., Core J.C., Gracco L.C., and Nye P.W. 1991. Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels. *Journal of Acoustical Society of America*, 90, 2, 799-828.
- [2] Yang, C.S and Kasuya H. 1994. Accurate Measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects. *Proceedings International Congress of Speech and Language Processing 94*, 623-626.
- [3] Yang, C.S. and Kasuya H. 1996. Speaker individualities of vocal tract shapes of Japanese vowels measured by magnetic resonance images. *Proceedings International Congress of Speech and Language Processing 96*, 949-952
- [4] Demolin, D., Metens T., and Soquet A. 1994. Three-dimensional measurement of the vocal tract by MRI. *Proceedings International Congress of Speech and Language Processing 96*, 272-275.
- [5] Gracco, C., Sasaki C.T., McGowan R., Tierney E., and Gore J. 1994. Magnetic resonance imaging(MRI) in vocal tract research: clinical application. *127th Meeting of Acoustical Society of America Abstract 1pSP35*.
- [6] Liljencrants, J. and Fant G. (1975) Computer program for VT-resonance frequency calculation. *STL-QPSR*, 4, 15-21.
- [7] Yang, B. 1996. A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonetics*, 24, 1, 245-261.
- [8] Yang, B. 1998. Vowel perception by formant variation. *Proceedings Acoustical Society of America* 98, 3005-3006.
- [9] Stevens, K.N. 1972. The quantal nature of speech: evidence from articulatory-acoustic data. In David, E.E. & Denes, P.B., eds. *Human Communication: a Unified View*, 51-66.
- [10] Wood, S. 1979. A radiographic analysis of constriction locations for vowels. *Journal of Phonetics*, 7, 25-43.