音楽情報科学 19－14
(1997. 2. 21)

母音の不変性と個人性：調音的と音響的観察に基づく証拠

楊 長盛、 粕谷 英樹

宇都宮大学工学部
〒321 宇都宮市石井町 2753
E-mail: yang@klab.ishii.utsunomiya-u.ac.jp

あらまし 本研究ではMR画像 (Magnetic Resonance Image) により計測した男性、女性の声道形状の個人性を検討し、声道形状の各部分の違いが音響的な影響及び知覚との関係について検討する。母音の類似性に関する聴覚実験の結果から、女性から男性への母音の正規化は主に声道長に依存することが分かった。同じ音質の母音を発声する時の声道形状と音響特徴を比較した結果、成人男性は調音的近似よりも音響的パラメータ（聴覚的に重要であるF1, F2, F3）を「不変量」として発声している示唆が得られた。また、高次フォルマント（F4,F5）は声道の変動に対して安定な個人性特徴であることを示した。

キーワード MR画像、声道形状、ホルマント周波数、不変性、個人性

Invariance and Individuality of the Vowel: Evidence from Articulatory and Acoustic Observations

Chang-Sheng Yang and *Hideki Kasuya*

Faculty of Engineering, Utsunomiya University,
2753 Ishii-machi, Utsunomiya 321, Japan
E-mail: yang@klab.ishii.utsunomiya-u.ac.jp

Abstract This paper discusses individualities of the vocal tract (VT) shape of vowels measured from MR images of males and females. Differences in VT dimensions of the subjects and their effects on acoustic characteristics are investigated. Perceptual similarity tests of vowel quality showed that normalization of vowels from females to males could be made by relying largely on the VT length. Vowels of the males were carefully compared at the articulatory and acoustic levels. The result suggests that males most probably produce an identical vowel by keeping the aoustic parameters (F1, F2 and F3) nearly invariant. The higher formant frequencies (F4 and F5) indicate stable speaker individualities.

key words MR image, vocal tract shape, formant frequencies, invariance, individuality

I. INTRODUCTION

Speech perception is achieved based on, therefore the invariance contained in speech signal is represented by articulatory[1] or acoustic[2], is the fundamental problem for speech research.

For the purpose of automatic speech recognition, many researchers have devoted to find the "invariant" so as to normalize differences among individual speakers. Most of the studies are mainly based on uniform scaling of the formant frequencies in terms of the vocal tract (VT) length[3-7].

In order to find a reasonable solution to the problem of invariant nature of the vowel, we have measured vocal tract (VT) shape of three males (MHK, MMM and MSH) and three females (FKK, FMS and FMSu) and an 11 years old boy (MKK) from MR images that were taken during sustained phonation of the five Japanese vowels /a, i, u, e and o/[8-10].

This paper discusses nonuniform differences in VT dimensions, individualities of VT shapes of the subjects, and relationship between perceptual effects of acoustic properties and VT variations from female to male.

Vowels are carefully compared among the male subjects under the condition that phonetic qualities are the same. Invariant factors and speaker individualities of vowels of the male subjects are discussed at the articulatory and acoustic levels.

II. VT DIMENSIONS

A. Measurement of VT Dimensions

To see dimensional individualities in the VT shapes of the subjects, a VT length was divided into three sections: the oral section (from the lips to the top of the uvula), the pharyngeal section (from the top of the uvula to the top of the epiglottis) and the laryngeal section (from the top of the epiglottis to the glottis). A length of each of the three sections was measured from the mid-sagittal MR image. Then a percentage of each section to the whole VT length was calculated. Figure 1 gives a schematic illustration to measure the lengths.

B. Results and discussions

Percentages of each of the three VT section lengths to the whole VT length are shown for the subjects in Table 1, where each item represents the average for the five Japanese vowels. In Table 1, we see that variations of the percentage among the males are relatively larger than those of the females and boy. The values of the oral section of the males MHK and MMM are relatively smaller in percentage and those of the laryngeal section are relatively larger as compared to the female and boy subjects. But for MSH, the values of the oral section are close to that of the female subjects. This suggests that non-uniform differences of the oral cavity length is not necessarily a significant feature to discriminate males from females.

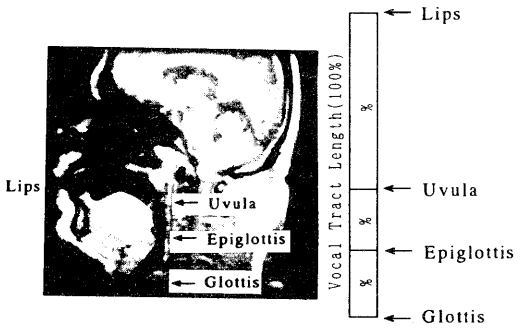

Figure 1: Measurement of VT dimensions.

Table 1: Averaged percentages of the three VT section length to the whole VT length of the subjects.

Sub.	$\frac{L_1}{L}$(%)	$\frac{L_2}{L}$(%)	$\frac{L_3}{L}$(%)	L(mm)
MHK	27.1	16.7	56.2	179
MMM	25.3	16.1	58.6	159
MSH	24.2	15.4	60.4	165
FKK	22.9	16.8	60.3	142
FMS	21.6	18.5	59.9	136
FMSu	23.2	14.6	62.2	140
MKK	19.3	16.9	63.8	139

In Table 1, variations of the percentage are relatively larger among the males than the females. The main reason for this is that there is a big difference of the laryngeal height among the male subjects. This can be observed from the mid-sagittal MR images.

Table 1 also suggests that the anatomical difference of the oral section length or the laryngeal and pharyngeal sections between males and females is non-uniform in the sense of group mean, but individual lengths of the vocal tract may

continuously distribute from females to males.

III. DIFFERENCE OF FEMALE/MALE

A. Normalization of VT Shapes

In this section, we use the measured area functions and detailed information about VT dimensions to investigate relationships between formant differences and VT variations from female to male.

Three normalization methods were applied with respect to the VT length for each of the vowels:

(1) A VT length was uniformly scaled to that of the reference.

(2) Lengths of the three VT sections were individually scaled to those of the reference using measured ratios of the corresponding sections.

(3) A VT length was first uniformly scaled and cavity volumes of the three VT sections were then adjusted so as to be the same as that of the reference.

Formant frequencies were computed from the normalized area functions and compared with those of the reference.

A difference of formant frequency is calculated as follows.

$$F_{\mathrm{Diff}}(i) = \frac{F_{\mathrm{SUB}}(i) - F_{\mathrm{MHK}}(i)}{F_{\mathrm{MHK}}(i)} \times 100\%, i = 1, 2, \cdots$$

Where, $F_{\mathrm{Diff}}(i)$ is a difference of the ith formant frequency, $F_{\mathrm{MHK}}(i)$ is the ith formant frequency of a vowel of MHK, and $F_{\mathrm{SUB}}(i)$ that of a subject. In the experiment, area functions of MHK were used as the reference. Those of FKK were normalized by the described methods.

B. Acoustic Result

Table 2 lists the first three formant frequencies of the five Japanese vowels that were computed from the original and the normalized area functions of FKK. Differences between the formants of MHK and those of the normalized ones are also given in the table.

From Table 2, we see that a large part of the formant differences is eliminated by using method (1) which means that the VT length distributed the most parts of the differences.

The most important thing is that, in Table 2, differences of the first three formant frequencies between the uniform and nonuniform methods are all less than 5% which is close to the perceptual difference limen (DL) of the formant frequencies of vowels reported by Flanagan[11], except for F3 of /i and u/. This gives us a strong hint that non-uniformity among the VT dimensions of the male, female is only a secondary factor in the normalization process. It is suggested, therefore, that the non-uniformity in the vocal tract dimensions is not an essentially factor to the nonuniformity of the formant patterns.

Table 2: Formant frequencies computed from the original and normalized area functions of FKK. The Diff's are compared with those of MHK.

/V/ Sub	F1	Diff. (%)	F2	Diff. (%)	F3	Diff (%)
/a/ MHK	686		1121		2501	
FKK	933	36.0	1413	26.0	3215	28.5
(1)	759	10.6	1146	2.2	2589	3.5
(2)	737	7.4	1181	5.4	2504	0.1
(3)	688	0.3	1151	2.7	2532	1.2
/i/ MHK	328		2098		2584	
FKK	368	12.2	3004	43.2	3390	31.2
(1)	314	-4.3	2284	8.9	2596	0.5
(2)	313	-4.6	2181	4.0	2742	6.1
(3)	311	-5.2	2245	7.0	2600	0.6
/u/ MHK	430		1313		2296	
FKK	424	-1.4	1148	-12.6	3049	32.8
(1)	375	-12.8	989	-24.7	2487	8.3
(2)	375	-12.8	1051	-20.0	2303	0.3
(3)	391	-9.1	988	-24.8	2494	8.6
/e/ MHK	543		1724		2314	
FKK	633	16.6	2409	39.7	3059	32.2
(1)	530	-2.4	1957	13.5	2492	7.7
(2)	547	0.7	1927	11.8	2419	4.5
(3)	555	2.2	1809	4.9	2543	9.9
/o/ MHK	522		874		2558	
FKK	601	15.1	936	7.1	3389	32.5
(1)	520	-0.4	800	-8.5	2767	8.2
(2)	500	-4.2	813	-7.0	2864	12.0
(3)	525	0.6	860	-1.6	2581	0.9

C. Perceptual Effect

To examine perceptual effect of the nonuniformity in the VT dimensions, perceptual similarity experiment was performed on the phonetic quality of synthesized vowel sounds.

Vowel stimuli used were the original vowel sounds that were synthesized from the original area functions of FKK as references (REF), vowels that were synthesized from the uniformly normalized area functions (Uniform Vowel, UV), and those that were synthesized from nonuniformly normalized area functions (Nonuni-

form Vowel, NV). Fundamental frequencies of the REF stimuli were the ones of the vowels spoken by FKK and those of the UV and NV stimuli were the same as the ones of male subject MHK. Duration of the stimuli were all 600 ms. Distributions of the first and second formant frequencies of the vowel stimuli are shown in Fig. 2.

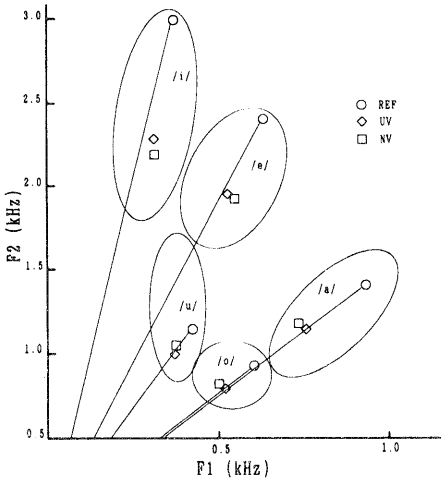

Figure 2: F1 and F2 distribution of the vowels computed from uniform (UV) and nonuniform (NV) scalings.

A triad consisting of either REF, UV and NV, or REF, NV and UV was presented to a speech scientist who was trained to make a phonetic judgment of vowel sounds. An interval between the stimuli within the triad was 800 ms and time for the judgment was 3.5 seconds. All the triads of the five vowels were randomly presented to the subject who was required to make a judgment on which is phonetically more similar to the REF.

Result of the perceptual similarity tests of vowel quality between REF and UV or NV stimuli showed that REF stimuli were more similar to UVs than NVs in the vowels /i, a, and o/, nearly equally similar to the two in the vowels /e and u/. This means the vowel sounds of the formants normalized uniformly were perceived phonetically equivalent.

These results suggest that normalization of Japanese vowels from females to males could be made by relying largely on uniform scaling of VT length.

IV. DIFFERENCES AMONG MALES

Table 3: The parameters used for modifying vocal tract shapes. All the parameters are changed slightly and are controlled independently while kepping the vocal tract length constant.

X_c	position of the constriction (changed by 1 minimeter per-step)
A_{back}	ratio of the back cavity (from the glottis to the X_c)
A_{LP}	ratio of the laryngeal-pharyngeal cavity (from the glottis to the uvula)
A_{oral}	ratio of the oral cavity (from the uvula to the inside of the lips)
A_{lip}	ratio of the lip opening

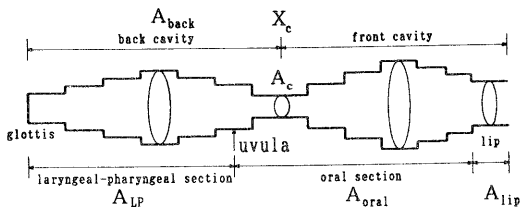

Figure 3: A model for area function modification.

In [10] we have indicated that not only the VT length but also the back and front cavity volumes are very important in considering speaker individualities of vowel articulation.

Slight differences in the phonetic quality were observed between the males for the vowel samples. Some method must be devised to modify the area functions measured from MR images so that the area functions result in the same phonetic quality, in order to make more accurate comparisons at the articulatory and formant levels. Since we need to modify area functions measured from vocal tract shapes of nature phonations, a simplified model is sufficent to specify variations of the area functions. Figure 3 illustrates a model used for modifying vocal tract shapes. Parameters used in this model are shown in Table 3.

A parameter X_c is used for the constricton position which is moved by scaling the back and front cavity lengths respectively. Size changes in the vocal tract shape affected by X_c moving are ignored in this procedure. It can be covered by other parameters.

It is considered that the back cavity volume of a phonation has relatively freedom to be con-

Table 4: Parameters of area function modification for MMM.

	/a/	/i/	/u/	/e/	/o/
X_c(mm)	0	0	-3	0	0
A_{back}	1.0	1.2	1.0	1.0	0.85
A_{LP}	1.0	1.1	1.0	1.0	1.0
A_{oral}	1.15	1.5	1.65	1.55	1.15
A_{lip}	0.85	1.0	1.5	1.0	1.25

Table 5: Parameters of area function modification for MSH.

	/a/	/i/	/u/	/e/	/o/
X_c(mm)	-5	-20	3	0	2
A_{back}	0.7	1.0	1.0	1.0	0.75
A_{LP}	1.0	1.0	0.85	0.85	1.0
A_{oral}	1.35	1.35	1.55	1.25	1.2
A_{lip}	1.0	1.0	1.2	1.0	1.0

troled. As an independent parameter, A_{back} is used for modifying the cross areas of the back cavity.

We use two parameters A_{LP} and A_{oral} to describe changes of the size of the laryngeal-pharyngeal cavity and the size of the oral cavity. In genernal these parameters and A_{back} are not independent. They are mainly concerned with tongue root movement, tongue shape changing and jaw opening. From Harshman *et al.*[12], these movements of the vocal organs can be divided into two sections by the uvula. Considering a total effect on the vocal tract shapes which is how to control volumes of the two sections, it is plausible to describe them independently in a limited range.

A_{lip} is used for lip opening which has relatively large flexibility.

All the A-parameters are used as a ratio to the original cross areas of the related sections.

A. Experiment

In this experiment, area functions of MHK are used as references. Those of MMM and MSH are modified. Formants are calculated from the modified area functions. Vowels are synthesized and carefully compared with those of MHK's. To diminish factors affected by voice source, vowels are synthesized by using the same voicing source parameters. and the respective formants. The sampling frequency is 10kHz. If phonetic quality is different, then the parameters are changed. This procedure is repeated for many times until the perceived phonetic quality is very close.

B. Results and discussion

Table 4 showes the parameters of area function modification for subject MMM. In this table we see that in most cases only the parameters of the oral cavity and the lips are modified. This means that MMM seems to adjust the size of the lips and the oral cavity to phonate vowels which very close to MHK.

In Table 4, we see that the parameter A_{oral} of vowels /i, u and e/ have big values. This means that the cross areas of the oral cavity are scaled by more than 150%. Because MMM has rather small oral cavities for these vowels, it can be considered that these are posible change ranges to articulate the vowels. We have confirmed that it is posible to produce those vowels from the imitation utterances of MMM.

Table 6 lists the formant frequencies of MMM calculated from the modified area functions for the five Japanese vowels. We find that the first three formant frequencies of MMM' of each vowel are very close to those of MHK. For vowel /i/, we find that difference of F2 between MHK and MMM' is big. When F2 (2797Hz) which obtained by LPC analysis from the original utterance /i/ of MHK, the vowel quality was considerably near the modified /i/ of MMM'.

Table 5 lists the parameters of modification for subject MSH. From the table we see that mainly the parameters of the oral cavity and the back cavity are modified. This means MSH seems to adjust the sizes of the the oral cavity and the back cavity to produce the same vowels of MHK. For the formant frequencies, the same results similar to MMM are obtained. The first three formant frequencies of MHK and MSH' of each vowel are very close.

Table 6 and the results of MSH showed that males can produce a vowel of the same lower formant frequencies (F1, F2 and F3) by controling front-to-back cavity volume ratio and place of the constriction although they have different vocal tract lengths. These results indicated that vowels of very close F1, F2 and F3 have perceptually equivalence instead of similar articula-

Table 6: Formant frequencies of MMM before and after modification. MMM's are formant frequencies after modification

V	SUB	F1	F2	F3	F4	F5
/a/	MHK	686	1121	2501	3461	4444
	MMM'	688	1144	2483	3951	4937
	MMM	677	1205	2466	3987	4990
/i/	MHK	328	2098	2584	3431	4156
	MMM'	334	2161	2899	3582	
	MMM	321	2245	3206	3688	
/u/	MHK	430	1313	2296	3357	4168
	MMM'	415	1323	2376	3807	4488
	MMM	347	1391	2408	3813	4573
/e/	MHK	543	1724	2314	3450	4461
	MMM'	557	1813	2352	3718	
	MMM	493	1954	2521	3756	
/o/	MHK	522	874	2558	3692	4675
	MMM'	511	881	2527	3779	4982
	MMM	469	865	2524	3824	4982

tions. This suggested that the *target* of a vowel phonation may mainly depend on acoustic parameters.

From Table 6 and the results of MSH, there are only a few shifts of F4 and F5 are observed after modification. These results show that it is difficult to control the higher formants. The higher formant frequencies (F4, F5) can be considered as stable factors of speaker individuality.

V. SUMMARY

By using VT shapes measured from MR images of vowels, individual differences and their effects on acoustic properties and on perception are investigated.

The results showed that the anatomical dimensions of the vocal tract distributed continuously from females to males, and nonuniformity in the VT dimensions effects little on the first three formant frequencies.

The perceptual similarity tests of vowel quality showed that normalization of vowels from females to males could be made relying largely on the vocal tract length.

Area functions of the male subjects and their formant frequencies were carefully compared under the condition that the perceived vowel qualities were the same. The result showed that, males most probably produce an identical vowel by keeping the aoustic parameters (F1, F2 and F3) nearly invariant which are perceptually important. On the other hand, the higher formant frequencies (F4 and F5) were stable against perturbation of the vocal tract shape and indicated speaker individualities.

ACKNOWLEDGEMENTS

This work was partly supported by Grant-in-Aid for Scientific Research (08650420) from the Ministry of Education, Science and Culture of Japan.

References

[1] Liberman, A. M. and Mattingly, I. G.: "The motor theory of speech perception revised," Cognition **21**, pp.1-36(1985).

[2] Stevens, K. N.: "Toward and model for speech recognition," JASA, Vol.32, pp.47-55(1960).

[3] Peterson, G. E. and Barney, H. L.: "Control methods used in a study of the vowels," JASA, **24**, pp.175-194, 1952.

[4] Kasuya, H., *et al.*: "Changes in pitch and first three formant frequencies of five Japanese vowels with age and sex of speakers," JASJ, **24**, pp.355-364(1968, in Japanese).

[5] Fujisaki, H. and Nakamura, N.: "Normalization and recognition of vowels," Annual Report of the Engineering Research Institute, **28**, pp.61-66(1969).

[6] Wakita, H.: "Normalization of vowels by vocal-tract length and its application to vowel identification," IEEE Trans. ASSP, **25**, pp.183-192(1977).

[7] Kent, R. D. and Forner, L. L.: "Developmental study of vowel fromant frequencies in an imitation task," JASA, **65**, pp.208-217(1979).

[8] Yang, C.-S., *et al.*: "Considerations on measurement method of vocal tract shapes using magnetic resonance imaging," IEICE Trans. Fundamentals, **77-A**, pp.1327-1335(1994, in Japanese).

[9] Yang, C.-S. and Kasuya, H.: "Uniform and non-uniform normalization of vocal tracts measured by MRI across male, female and child subjects," IEICE Trans. On Inf. & Syst., **E78-D**, pp.732-737(1995).

[10] Yang, C.-S. and Kasuya, H.: "Speaker individualities of vocal tract shapes of Japanese vowels measured by magnetic resonance images," Proc. ICSLP96, pp.949-952(1996).

[11] Flanagan, J. L.: *Speech analysis, synthesis and perception*, (2nd ed. Springer-Verlag, New York, 1972).

[12] Harshman, R., Ladefoged, P. and Goldstein, L.: "Factor analysis of tongue shapes," JASA, **62**, pp.693-707(1977).