VOWEL NORMALIZATION REVISITED: INTEGRATION OF ARTICULATORY, ACOUSTIC, AND PERCEPTUAL MEASUREMENTS

C.-S. Yang and H. Kasuya
Utsunomiya University, Utsunomiya, Japan

ABSTRACT

Vocal tract (VT) area functions were measured from magnetic resonance images (MRI) for five Japanese vowels /i, e, a, o, and u/ across a child, a female adult, and a male adult. Effects of uniform and non-uniform normalization of the area function with respect to the length of three parts of VT, *i.e.* oral, pharyngeal and laryngeal sections, are investigated at articulatory, acoustic, and perceptual levels. Significance of uniform normalization is suggested.

INTRODUCTION

Relationships of formant patterns of vowels among male, female and child speakers are found to be nonuniform [1],[7]. How human beings normalize in the auditory perception a specific class of vowels is a classical but difficult problem. Many efforts have been made to solve the problem [1]-[7]. Kasuya, *et al.* [3] and Fujisaki and Nakamura [4], for example, proposed a coordinate system for the auditory representation of vowel classes based on uniform scaling of the first three formant frequencies in terms of the vocal tract length. Fant attributed the non-uniformity observed in the formant patterns between female and male speakers to that of VT dimensions [1]: ratio of pharynx length to mouth cavity length is greater for males than for females. Nordstrom found that anatomical differences between males and females only explain part of the differences based on the VT shapes predicted from X-ray photographs of midsagittal sections [5].

In this paper, we first measure the area function of three parts of the VT, *i.e.* oral, pharyngeal and laryngeal cavities, from MRI data of the five Japanese vowels using a newly developed image processing method [8] and investigate acoustic and perceptual significance of uniform and nonuniform scaling of the length of the three cavities.

METHOD

Measurement of VT area function

We have developed a method to accurately measure 3 dimensional VT shapes from MRI data, for the acquisition of which a General Electric SIGNA machine (1.5 T) was used [8]. VT data were obtained for five Japanese vowels, /i, e, a, o, and u/, of three subjects, a child, a female, and a male. The VT was divided into three sections as shown in Fig.1: the oral (from the lips to the uvula), pharyngeal(from the uvula to the top of the epiglottis), and laryngeal (from the top of the epiglottis to the glottis) sections. Length of each section was measured along the VT center line which was semiautomatically estimated on the midsagittal section image. Percentage of the length of each VT section to the entire VT length was then calculated for all the vowel data.

Uniform and nonuniform scaling of VT dimensions

Each of the area functions of the child and female subjects was normalized, first by making the entire VT length identical to the male's by using two different methods, *i.e.* uniform and nonuniform scaling, and then by adjusting the areas so that the maximum value becomes identical to that of the male.

In the uniform scaling, an entire VT length was uniformly extended to that of the male following the next equation (see Fig.2(a)):

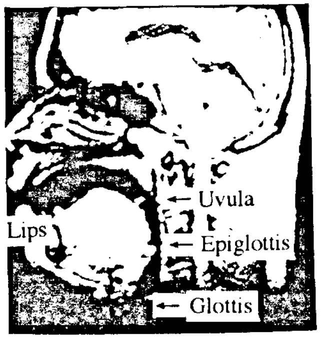

Fig. 1 Midsagittal section of the vocal tract.

$$x' = \alpha_L \times x,$$

where x is an original length from the glottis, α_L is a scale factor and x' is a normalized length.

In the nonuniform scaling, on the other hand, each length of the three VT sections of the child and female subjects was separately extended to that of the male as follows (see Fig.2(b)):

$$x'_i = \alpha_{L_i} \times x_i, \quad \text{for } i=1, 2, 3,$$

where x_i is an original length measured from the upstream end of the i-th section and α_{L_i} is a scale factor of the i-th section.

After scaling the VT lengths, all the values of the area function were adjusted so as to have the same maximum value as that of the male.

Computation of formant frequencies

The first four formant frequencies and bandwidths were computed from the acoustic transfer function [9] for all the VT area functions including the uniformly and non-uniformly normalized ones.

Perceptual experiments

Perceptual similarity experiments were performed on the phonetic quality of vowel sounds synthesized with the formant frequencies that were obtained from the area functions of the female subject.

Vowel stimuli used were the original vowel sounds that were synthesized from the original area functions of the female (Reference, REF), vowels that were synthesized from the uniformly normalized area functions (Uniform vowel, UV), and those that were synthesized from non-uniformly normalized area functions (Nonuniform vowel, NV). Fundamental frequencies of the REF stimuli were the ones of the vowels spoken by the female and those of the UV and NV stimuli were the same as the ones of the male. Duration of the stimuli were all 600 ms.

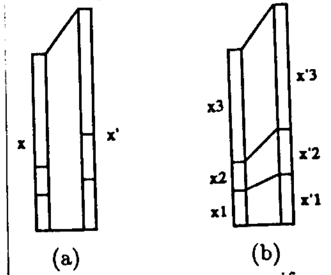

Fig. 2 (a)Uniform and (b)nonuniform normalization of the VT length.

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.

RESULTS AND DISCUSSIONS

VT area functions

Figure 3 illustrates a percentage of each length of the three VT sections to the whole VT length. The male shows smaller oral cavity length but larger laryngeal section length in percentage than the female and child. It is also seen that the percentage of nonuniformity is different from vowel to vowel.

Scalings of VT length

Using the uniform and nonuniform sacling methods described above, the area functions of the female and child were normalized as shown in Fig. 4, where (a) and (b) are for uniform and nonuniform scalings, respectively. It seems from the figure that the normalized area functions depend little on the type of the scalings in all the vowels.

Formant frequencies

The first four formant frequencies were calculated from the original and normalized area functions [9]. Distributions of the first and second formant frequencies of the female are shown in Fig. 5. Differences in the formant values between the two scaling methods were all less than 5 % which is close to the perceptual difference limen (DL) of the formant frequencies [10]. This was the case for the child. These seggest that nonuniformity of the VT dimensions among the child, female and male speakers is only a secondary factor in the normalization process.

Perception of vowel quality

Results of the perceptual similarity tests of vowel quality between REF and UV or NV stimuli were such that REF stimuli were more similar to UVs than NVs in the vowels /i, a, and o/, nearly equally similar

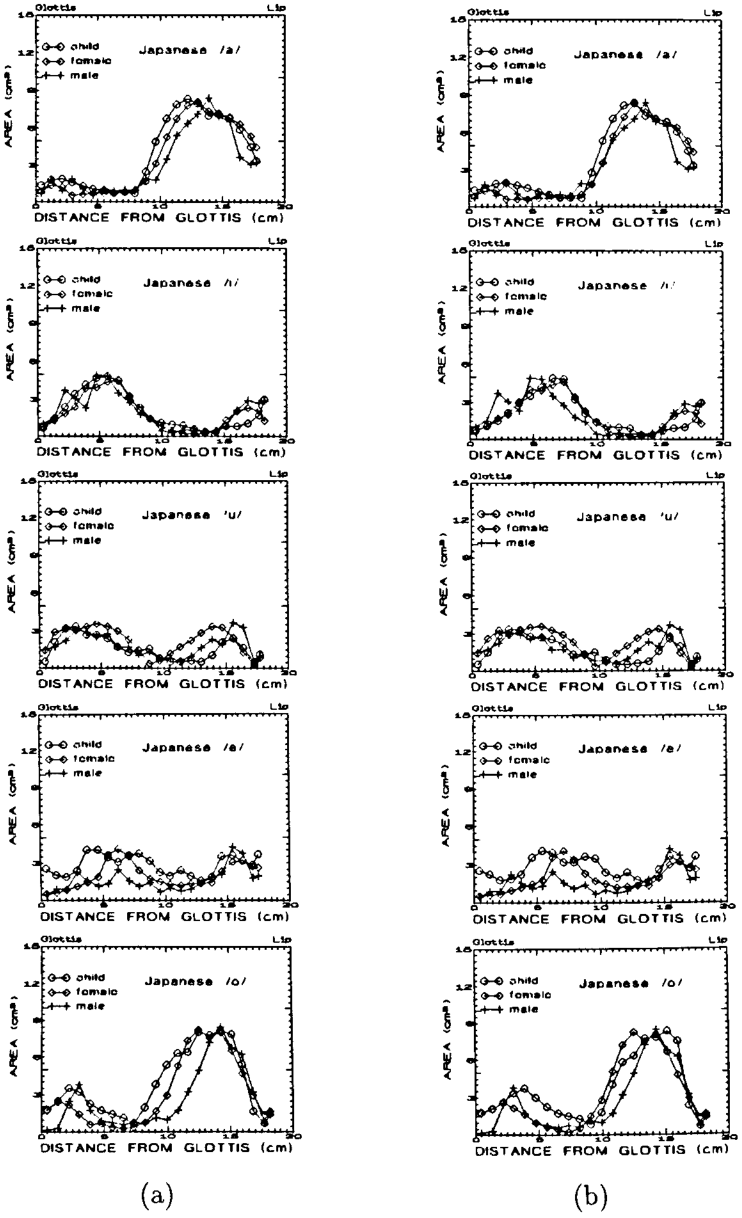

Fig. 4 VT area functions of the child, female and male subjects for the five Japanese vowels, normalized by (a) uniform and (b) nonuniform scaling.

to the two in the vowels /e and u/.

CONCLUSION

Nonuniform scaling of the vocal tract dimensions with respect to the length of oral, pharyngeal and laryngeal cavities effects little on the first three formant frequencies and the vowel sounds of the formants normalized uniformly were perceived phonetically equivalent. These findings support the significance of uniform scaling of the VT length.

REFERENCES

[1] Fant, G. (1973), *Speech sound and features,* The MIT Press.

[2] Kasuya, H., Suzuki, S. and Kido, K. (1968), "Changes in pitch and first three formant frequencies of five Japanese vowels with age and sex of speakers," J. Acoust. Soc. Jpn., vol. 24, pp.355-364 (in Japanese).

[3] Kasuya, H., Suzuki, H., and Kido, K.(1968), "On auditory model of vowel perception," Proc. 6th Int. Congr. on Acoustics, B-3-3, Tokyo, Japan.

[4] Fujisaki, H. and Nakamura, N. (1969), "Normalization and recognition of vowels," Ann. Rep. of the Engr. Res. Rept., vol.28, pp.61-66.

[5] Nordstrom, P.-E. (1977), "Female and infants vocal tract simulated from male area functions," J. of Phonetics, vol.5, pp.81-92.

[6] Kent, R.D. and Forner, L.L. (1979), "Developmental study of vowel formant frequencies in an imitation task," J. Acoust. Soc. Am., vol. 65, pp.208-217.

[7] Wakita, H. (1977), "Normalization of vowels by vocal tract length and its application to vowel identification," IEEE Trans. Acoust., Speech & Signal Process., vol. ASSP-25, pp.183-192.

[8] Yang, C.-S. and Kasuya, H. (1994), "Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects," Proc. Int. Conf. on Spoken Language Process., vol. 2, pp.623-626.

[9] Sondhi, M. M. and Shroeter, J. (1987), "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Trans. Acoust., Speech & Signal Process., vol. ASSP-35, pp. 955-967.

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

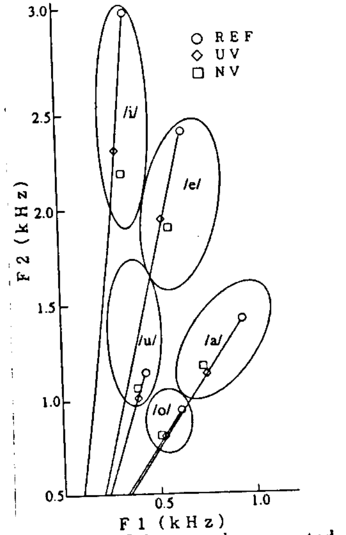

Fig.5 F1 and F2 of the vowels computed from uniform(UV) and nonuniform(NV) scalings.

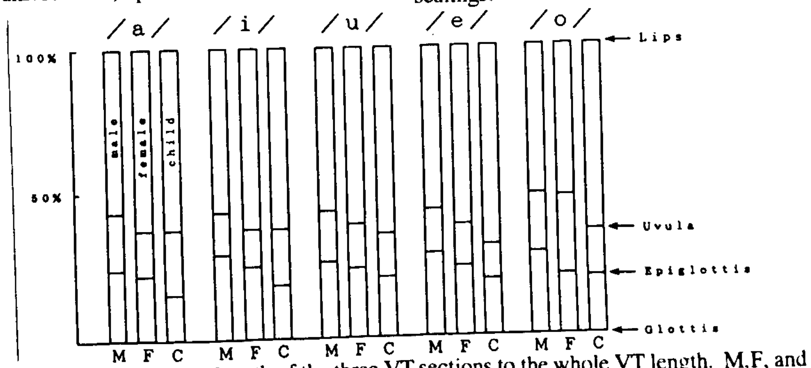

Fig. 3 Percentage of the length of the three VT sections to the whole VT length. M,F, and C are respectively male, female and child subjects.