A FUNCTIONAL MODEL OF DYNAMIC CHARACTERISTICS OF FORMANT TRAJECTORIES

S. Imaizumi, H. Imagawa and S. Kiritani

Research Institute of Logopedics and Phoniatrics
Faculty of Medicine, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 JAPAN

ABSTRACT

This paper describes a functional model to generate dynamic characteristics of formant trajectories at various speaking rates. The formant trajectories are modelled as the sum of two kinds of temporal functions: a second order delay function which represents vowel-to-vowel transitions, and two first order delay functions which represent the effects of surrounding consonants on the vowel formant trajectories. Using this model, VCV speech samples were synthesized at slow and fast speaking rates, and their intelligibility tested. Results suggested that the model works very well, although some additional strategies are needed to improve the intelligibility of the consonants at fast speaking rates.

1. INTRODUCTION

There are still numerous differences among the conclusions of studies on the effects of speaking rate[3, 5, 6, 7, 8, 10, 11]. Some studies indicate the occurence of "vowel reduction"[10]. Some others claim that such "vowel reduction" does not always occur at fast speaking rates[5, 11]. Other studies claim that adjustments in speaking rate are achieved by strategies which differ among speakers, and on the carefulness of their articulation[3, 6, 8].

One approach to arrive at correct conclusion for this issue is to construct a model, by which we can test if undershoot or reorganization is necessary or not to generate high quality speech at various speaking rates. Although some models have been proposed to generate formant transitions[1,2,7,9],there are still problems remaining to be solved to generate natural formant trajectories.

In this paper, we proposed a functional model which describes the formant transitions as the sum of two kinds of temporal functions: one represents vowel-to-vowel transitions, and the otherrepresents CV or VC transitions. The model was assessed via an intelligibility test.

2. METHOD

2.1 Formant Transition Model

The trajectory of nth formant, $F_n(t)$, in a vowel segment is assumed to be expressed as

$$F_n(t)=U_n(t)-C_{nf}(t)-C_{np}(t) \quad (1).$$

Here, Un(t) is the step response of a second order delay function which represents vowel-to-vowel transition, $C_{np}(t)$ is the first order delay function which represents the effect of a preceding consonant, and $C_{nf}(t)$ is the first order delay function which represents the effect of a following consonant.

To generate $U_n(t)$, the putative target frequency $R_{i,j}$ of each vowel in the sequence $V_1C_pV_2C_fV_3$, (i=1, 2, 3, j=1, 2, 3) is assumed to be set at t_i as a step input. The suffix i represents vowel number, j formant number. For the back vowels /a,u,o/, j represents jth lower formant frequency.

For the front vowels /i,e/, $R_{i,1}$ is the lowest, $R_{i,2}$ the third, and $R_{i,3}$ the second.

Let $W_j(t)$ represent the step response of a second order delay function, $W_j(t)$ can be expressed as

$$W_j(t)=R_{1,j}+a_i(t)(R_{i,j}-R_{i-1,j}), \quad (2)$$
$$a_i(t)=1-\{1+b_j(t)\}\exp(-b_j(t))u(t-t_i),$$
$$b_j(t)=(t-t_i)/g_j,$$
$$u(t-t_i)=1 \text{ for } t>t_i, =0 \text{ for } t<t_i,$$

g_j: time constant representing transition speed.

For transitions from a back vowel to a front vowel or vice versa, $W_2(t)$ and $W_3(t)$ intersect each other. Such intersection never occur in actual speech due to the coupling between two resonance frequencies. Therefore the resonance frequencies $W_j(t)$ are modified accounting for the coupling between $W_2(t)$ and $W_3(t)$ as follows [4].

$$U_1=W_1, U_2=c\sqrt{(W_2W_3)}, U_3=\sqrt{(W_2W_3)}/c,$$
$$c=\sqrt{e}, d=(W_2W_2+W_3W_3)/W_2W_3, \quad (3)$$
$$e=d-(dd-4(1-kk))/2(1-kk), \quad k=0.2.$$

Two functions representing the effect of a preceding consonant $C_{np,i}(t)$ and of the effect of a following consonant $C_{nf,i}(t)$ upon the formant trajectories in the segment V_i are assumed as follows.

$$C_{npi}(t)=c_{np,i}\exp\{(t-t_{p,i})/g_p\}, \text{ for } t_p<t>t_{f,i}, \quad (4)$$
$$C_{nf,i}(t)=c_{nf,i}\exp\{-(t_{f,i}-t)/g_f\}, \text{ for } t_{p,i}<t>t_{f,i}, \quad (5)$$

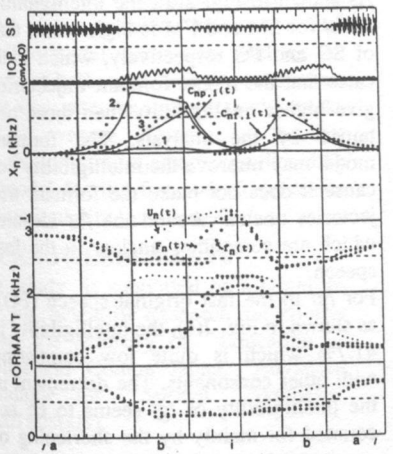

Fig.1(a). An example of formant trajectories Fn(t), and those obtained by analysis fn(t) for a slow utterance of /abiba/.

t_{pi}: initial time of vowel V_i,
t_{fi}: final time of V_i,
g_p,g_f:time constant representing the decay speed.

In this paper, only the temporal parameters, t_i:onset time of the targets for vowel V_i, t_{pi}: initial time of V_i and t_{fi}: final time of V_i are changeable depending on the speaking rate. This means that this model does not take account of possible changes or "reorganization" in the vowel targets or other parameters such as g_p and g_f.

2.2 ModelParameters Estimation

The speech material used here consisted of $V_1C_pV_2C_fV_3$ samples, where V_1, V_2 were one of /a, i, u,/, $V_1=V_2$, and C_p or C_f was one of /b, d, g, p, t, k, r/. Those samples were recorded from two male speakers spoken at slow (S) and fast (F) speaking rates. The original utterances were /korewa $V_1C_pV_2C_fV_3$ desu/, that means, "this is $V_1C_pV_2C_fV_3$". Among five Japanese vowels, only three vowels /a, i, u/ were used as the typical examples of low-back, high-front, and intermidiate vowels.

The details of recording, analyses, parameter estimation method were reported

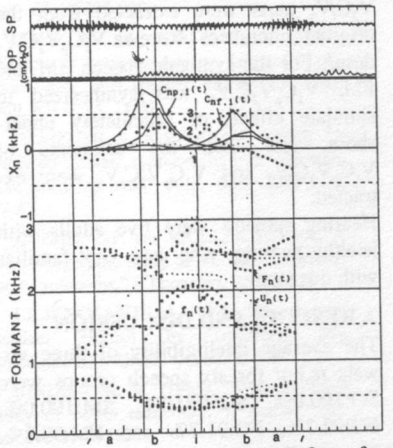

Fig.1(b). Same as Fig.1(a), but for a fast utterance of /abiba/.

in other place [6,7]. The formant trajectories were estimated mainly using the closed phase LPC analysis.The estimation of the model parametrs was carried out based on a least square error method and interactive modification using only speech uttered slowly and clearly.

2.3 Intelligibility Test

To assess how well the model could generate formant trajectories, an intelligibility test was carried out for two kinds of synthetic speech (G and M), and the original speech samples (O) from which model parameters were extracted. Here, synthetic speech samples G were generated using the formant frequencies f(t) obtained from SO by the analysis and the glottal source obtained from the polynomial glottal source model. Synthetic speech samples M were generated using the model formant trajectories F(t) and the model glottal source. The speech samples were synthesized or recorded at two speaking rates, slow (S) and fast (F). Each group consisted of 84 V_1CV_2 samples, where V_1, V_2 were one of /a, i, u,/, V_1=V_2, and C was one of /b, d, g, p, t, k, r/. For SO and FO, $V_1C_pV_2$ and $V_2C_fV_3$ parts were extracted from the original utterances /korewa $V_1C_pV_2C_fV_3$ desu/. For the synthetic speech SM and FM. $V_1C_pV_2C_fV_3$ was synthesized to simulate effects of articulately undershoot, and then the segments of $V_1C_pV_2C_fV_3$ and $V_1C_pV_2C_fV_3$ were extracted.

Hearing subjects were five adults with healthy hearing who were not familiar with this study.

3. RESULTS AND DISCUSSION

The average intelligibility of three vowels /a,i,u/ for six speech groups were SO:100.0%, SG:100.0%, SM:100.0%, FO:92.7%, FG:91.7% and FM:93.8%. The intelligibility of FM is 93.8% which is better than those of FO and FG. Concerning the vowels, it is suggested that the formant model maintains or even slightly improve the intelligibility compared to the original speech in slow and fast speaking rates.

On the other hands, the average intelligibility of the consonants /b,d,g,r/ were SO:91.7%, SG:81.3%, SM:83.3%, FO:79.2%, FG:68.8% and FM:62.5%. For the consonants, the use of model voicing source without plosion decreases the intelligibility about 10%. The use of the formant model slightly increases the intelligibility by about 2%(SM-SG). For the consonants in the slow speech, the formant model works well on average, and even slightly improve the intelligibility comparing with that of SG. However, for the fast speech, the formant trajectories predicted by the model for the fast speech decreases the intelligibility by about 6%.

There were, however, some differences among the intelligibility of the consonants, or the goodness of the model. Fig. 2 shows the intelligibility of the consonants /b,g/ for the six speech groups. The box-whisker graph in this figure shows the minimum, 25%-tile, median, 75%-tile, and maximum of the intelligibility scores.

As shown in Fig. 2(a), the intelligibility of /b/ for SM and FM is higher than that of SG and FG respectively, which indicates that the model formant trajectories give higher intelligibility than those obtained by the analysis. The formant model may improve the intelligibility because it does not make the formant trajectories unclear around the /b/ closure, which are sometimes unclear in the fast speech.

For /g/ in the fast original speech (FO), as shown in Fig. 2(c), the intelligibility is 41.7% which is quite low comparing with other consonants. The decrement in the intelligibility of /g/ seems to be accounted for mainly by the shortening of segments due to the speaking rate. Because the formant transition of /g/ is slower than that of other stops, the speaking rate change may affect largely the intelligibility of /g/.

In this model, only the timing parameters are changeable depending on the speaking rate. It shoud be noted that the intelligibility of the consonants at fast speaking rates might be improved by taking account of possible changes in some parameters such as those representing transient speed which are set constant in this report.

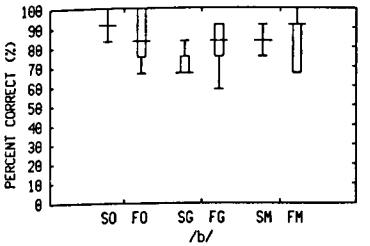

Fig.2(a) The intelligibility of /b/.

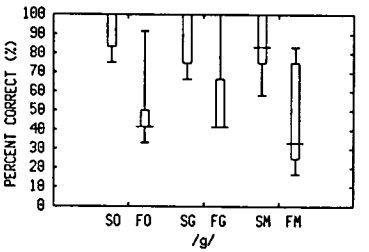

Fig.2(b) The intelligibility of /g/.

4. CONCLUSION

This paper proposes a model of formant trajectories at various speaking rates, and reports on the intelligibility of VCV speech samples synthesized based on the model at two speaking rates, slow and fast. The model works very well for the vowels at both rates. For the consonants at the slow rate, the formant model works well on average. However, at the fast rate, the formant trajectories predicted by the model for the fast speech decreases the intelligibility by about 6%.

5.ACKNOWLEDGMENTS

This study is supported by a Grant-in-Aid for Scientific Research on Priority Areas, the Ministry of Education, Science and Culture, Japan, "Advanced Technique for Speech Synthesis" (No.01608003).

6. References

[1] BROAD, D. J. & FERTIG, R. H. (1970),"Formant-frequency trajectories in selected CVC utterances," J. Acoust. Soc. Am.,47, 1572-1582.

[2] BROAD, D. J. & CLERMONT, F. (1987),"A methodology for modeling vowel formant contours in CVC context," J. Acoust. Soc. Am., 81(1) 155-165.

[3] FLEGE, J.E.(1988),"Effects of speaking rate on tongue position and velocity of movement in vowel production, J. Acoust. Soc. Am., 84(3), 901-916.

[4] FUJISAKI, H., YOSHIDA, M. SATO, Y. and TANABE, Y. (1974),"Automatic recognition of connected vowels using a functional model of the coarticulatory process," J. Acoust. Soc. Jpn, 29, 636-638.

[5] GAY, T.(1978),"Effect of speaking rate on vowel formant movements," J. Acoust. Soc. Am., 63(1), 223-230.

[6] IMAIZUMI, S., KIRITANI, S. (1989), "Effects of speaking rate on formant trajectories and inter-speaker variations," Ann. Bull. RILP, 23, 27-37.

[7] IMAIZUMI, S., KIRITANI, S. (1990), "A study on formant synthesis by rule with variable speaking rate," Ann. Bull. RILP, 23, 77-87.

[8] KUEHN, D.P. and MOLL, K.L.(1976),"A cineradiographic study of VC and CV articulatory velocities," J. Phonetics, 4, 303-320.

[9] LILJENCRANTS, J. (1970),"Speech synthesizer control by smoothed step functions,"STL-QPSR 4/1969, 43-50.

[10] LINDBLOM, B. (1963), "Spectrographic study of vowel reduction," J. Acoust. Soc. Am, 44, 1773-1781.

[11] van SON, R.J.J.H. & POLS, L.C.W. (1990), "Formant frequencies of Dutch vowels in a text, read at normal and fast rate,"J. Acoust. Soc. Am, 44, 1683-1693.