ANN.Bull.RILP(1990)No.24, 77-87

A STUDY ON FORMANT SYNTHESIS BY RULE WITH VARIABLE SPEAKING RATE

Satoshi Imaizumi and Shigeru Kiritani

Abstract

This paper proposes a synthesis model of formant trajectories at various speaking rates, and reports on the intelligibility of VCV synthetic speech samples. The model describes the formant trajectories as the summation of temporal functions: a second order delay function which represents vowel-to-vowel transitions, and two first-order delay functions which represent consonant-to-vowel or vowel-to-consonant transitions. The functions are determined from utterances spoken clearly and slowly. Using these functions, the VCV speech samples were synthesized at slow and fast speaking rates, and the intelligibility was tested. The formant model slightly improved the intelligibility of vowels at both speaking rates and of consonants at the slow rate over speech synthesized using formants obtained by analysis. For the consonants at the fast rate, the formant model decreased the intelligibility by about 6%. The model fitted best for vowels and the consonant /b/, and worst for the consonant /g/.

1. Introduction

In order to improve the quality of speech generated by formant synthesizer, several models describing formant trajectories have been proposed. For instance, some studies have used smoothed step functions[1-4], where the step inputs represent putative targets of vowels[2-4] or even of consonants[3]. Some studies propose a linear summation model of the target formant frequencies of vowels and temporal functions representing the effects of adjacent consonants[5,6]. Although these models seem able to describe some phenomena in formant trajectories, for instance the undershoot at fast speaking rates, there have been very few assessment results showing the ability of these models to synthesize high quality speech with variable speaking rates.

On the other hand, as a basic issue in speech research, there are still numerous differences among the conclusions of studies on the effects of speaking rate[7-20]. Some studies[8,9] indicate that increased rates of speech result in systematic deviations in obtained formant values from their putative targets, that is, "vowel reduction". Others[10-12] claim that such "vowel reduction" does not always occur at fast speaking rates. Still other studies claim that adjustments in speaking rate are achieved by strategies which differ among speakers[13,14] and in the carefulness of their articulation[15]. According to electromyographic investigations[16,17], control of the speaking rate is achieved via a reorganization of motor commands.

One approach to this issue is to construct a model, by which we can test if undershoot or reorganization is necessary in generating high quality speech at various speaking rates.

In this paper, we proposed a functional model which described formant transitions as the summation of two kinds of temporal functions: one represented vowel-to-vowel transitions, and the other represented consonant-to-vowel or vowel-to-consonant transitions. The model was assessed via an intelligibility test.

2. Method

2.1 Model of formant transition

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

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

Here,

- $U_n(t)$ is the step response of a second order delay function which represents a vowel-to-vowel transition;
- $C_{np}(t)$ is the a first order delay function which represents the effect of a preceding consonant;
- $C_{nf}(t)$ 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 indicates 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. This numbering is adopted to take into account the continuity in formant trajectories[2,4].

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

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

g_j : time consonant representing transition speed

For transitions from a back vowel to a front vowel, or vice versa, $W_2(t)$ and $W_3(t)$ intersect with each other. Such intersections 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[2,4].

$$U_1= W_1$$
$$U_2= c(W_2W_3)$$
$$U_3= (W_2W_3)/c$$
$$c= e \qquad (3)$$
$$d= (W_2W_2+W_3W_3)/W_2W_3$$
$$e= d-(dd-4(1-kk))/2(1-kk)$$
$$k= 0.2$$

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

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

$$C_{nf,i}(t) = c_{nf,i} \exp\{-(t_{f,i}-t)/g_f\} \quad t_{p,i} < t > t_{f,i} \tag{5}$$

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

2.2 Estimation of model parameters

The details of the recordings and the analyses of the speech material used for the modeling have been reported in other papers[18-20].

For the estimation of the model parameters, we assumed that the following are valid for vowels spoken clearly and slowly.
1) The effects of surrounding consonants upon the vowel formants decrease at vowel midpoint, so we can set $C_{np,i}(t) = C_{nf,i}(t) = 0$.
2) If the vowel segment is long, the formant frequencies $f_n(t)$ obtained by analysis are close to the putative targets, or $R_{i,j}$.

According to Assumption 2), $R_{i,j}$ is set to the formant frequencies $f_n(t)$ obtained by analysis at the midpoint t_i of vowel V_i. $U_n(t)$ is calculated using equations (2) and (3), then, the temporal function $X_n(t) = U_n(t) - f_n(t)$ can be calculated.

As shown in Fig.1, $X_n(t)$ is large at the initial point of vowel V_i and decreases rapidly to zero. After the midpoint, it increases to its maximum at the endpoint of V_i. Thus, $X_n(t)$ can be approximated by two first-order functions $C_{np,i}(t)$ and $C_{nf,i}(t)$. $C_{np,i}(t)$ is large at the beginning of V_i and then decreases exponentially, while $C_{nf,i}(t)$ is small at the midpoint and increases exponentially to its maximum at the end of V_i.

To determine $C_{np,i}(t)$ and $C_{nf,i}(t)$, $t_{p,i}$, the initial point of V_i is set at the point where the intraoral pressure starts to drop rapidly, or the release point of the consonant, and $t_{f,i}$, the final point of V_i, is set where the intraoral pressure starts to rise due to the vocal tract closure for the stop consonant. $C_{np,i}, g_p$ is adjusted so as to minimize the square error between $X_n(t)$ and $C_{np,i}(t)$ for the initial half of the segment V_i, and $C_{nf,i}$, g_f is adjusted so as to minimize the square error between $X_n(t)$ and $C_{nf,i}$ for the final half of the segment V_i.

Fig.1 (a) shows an example of the model functions estimated for /abiba/ spoken slowly and clearly. The uppermost curve is the original speech waveform, the second curve is the intraoral pressure measured simultaneously[18-20]. $X_n(t)$ is shown by three kinds of dotted lines according to formant number. In the bottom section, the formant frequencies $f_n(t)$ obtained by analysis (ooo), and the model formant frequencies (ooo) are shown together with $U_n(t)$ (....). During the vocal tract closure, $F_n(t_{f,i-1})$ and $F_n(t_{p,i})$ are linearly interpolated. Fig.1 (a) shows that $F_n(t)$ fits well with $f_n(t)$.

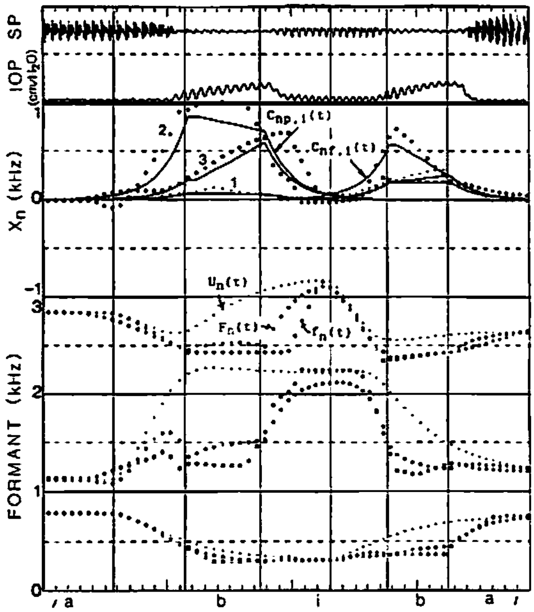

Fig. 1(a). Model formant trajectories $F_n(t)$, and those obtained by analysis $f_n(t)$ for a slow utterance of /abiba/. See text for details.

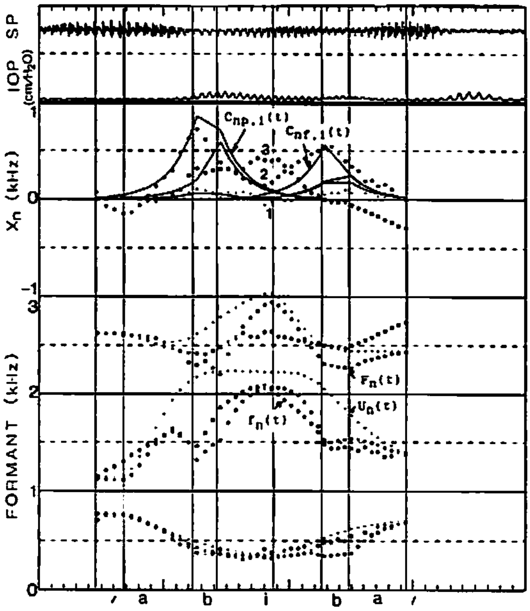

Fig. 1(b). Same as Fig. 1(a), but for a fast utterance of /abiba/.

2.3 Speech synthesis by rule at variable speaking rate

For the synthesis of speech at various speaking rates, rules for generating t_i, $t_{p,i}$, $t_{f,i}$ should be constructed. We do not discuss here such rules. Instead, we discuss how well such a model predicts the formant trajectories observed in actual fast speech. For the assessment of the model proposed here, we compared speech samples actually uttered at a fast rate (FO), which was twice as faster as the slow rate examined, with synthetic speech generated based on a model where t_i, $t_{p,i}$, $t_{f,i}$ were adapted to a fast speech FO. The other parameters were set to the same values obtained from the slow utterances (SO) from which the model parameters were estimated.

Fig. 1(b) shows one example of a fast /abiba/ uttered by the same speaker as in Fig.1(a). Here, t_i, $t_{p,i}$, $t_{f,i}$ are adapted to the actual utterance of /abiba/. $R_{i,j}$ for the vowel V_1 (initial /a/ in this case) are set to the actual average values of f_1, f_2 and f_3 obtained by the analysis, because the vowel reduction for V_1 can not fully be estimated without the preceding phonemes. Fig.1(b) shows that the model formant trajectories $F_n(t)$ for fast /abiba/ predict some gross characteristics in the f_1 and the f_2

transitions well, but fail to represent a large downward shift in f_3.

2.4 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 also for original speech samples (O) from which model parameters were extracted. These speech samples were synthesized or recorded at two speaking rates, slow(S) and fast (F). The speech samples tested consisted of the following six groups.

SO: Original speech samples uttered slowly and clearly, from which the model parameters were extracted.

FO: Original speech samples uttered fast, from which the temporal parameters for the synthetic fast speech (FM, FG) were extracted.

SG: Synthetic slow speech, generated using the formant frequencies $f_n(t)$ obtained from SO by analysis and the glottal source obtained from the polynomial glottal source model[20].

FG: Synthetic fast speech generated in the same way as SG.

SM: Synthetic slow speech generated using the model formant trajectories $F_n(t)$ and the model glottal source.

FM: Synthetic fast speech generated in the same way as SM.

Each group of consisted of 48 V_1CV_2 samples, where V_1, V_2 were one of /a, i, u,/, V_1=V_2, and C was /b, d, g, r/. For SO and FO, $V_1C_pV_2$ and $V_2C_fV_3$ were extracted from the original utterances /korewa $V_1C_pV_2C_fV_3$ desu/. For the synthetic speech SG, FG, SM and FM, $V_1C_pV_2C_fV_3$ was synthesized to simulate the effects of articulatory undershoot, and then the segments of $V_1C_pV_2$ and $V_2C_fV_3$ were extracted.

The subjects for the listening test were five adults with normal hearing who were not familiar with the purpose of this study, two phoneticians, one speech pathologist, one speech scientist and one graduate student majoring in speech science. The listening test was carried out only once to avoid possible adaptation to the synthetic speech. Each subject was instructed to transcribe each speech sample in phonetic symbols or in the Roman alphabet.

2.5 Four factors accounting for intelligibility

The following four factors were used to interpret the intelligibility of the six groups of speech as shown in Fig. 2.

a1: the decrement in percent due to the shortening of segmental duration in fast speech

a2: the decrement due to the omission of reorganization when the model is applied to fast speech

(-a2:the increment due to reorganization in fast speech)

a3: the decrement due to the lack of plosive source for the stop consonants for the slow speech

a3': same as a3, but for fast speech

a4: the decrement due to the formant model mismatch

The factor a1 accounts for the decrement between the intelligibility of SO and that of FO. Because the speech samples with FO have shorter duration, smaller formant transitions and also a larger undershoot or vowel reduction than those with SO, the intelligibility of FO may decrease largely compared to that of SO. However, if the speaker reorganize the articulation for fast speech to increase intelligibility against the disadvantages mentioned above, such a factor (reorganization:-a2) might raise the intelligibility. As a result, the difference between FO from SO should be a1-a2.

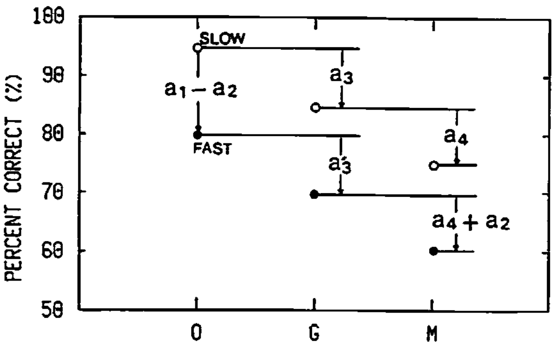

Fig. 2. Four factors accounting for the intelligibility.

Table 1. Four factors estimated from the medians of the intelligibility. V:3 vowels, C:for 4 consonants.

Factor	V	C	/b/	/d/	/g/	/r/
a 1	5.2	20.8	16.6	8.3	50.0	16.6
a 2	-2.1	8.3	8.2	-8.3	-8.2	8.3
a 3	0.0	10.4	25.0	16.7	0.0	0.0
a 3'	1.0	10.4	0.0	33.4	0.0	8.4
a 4	0.0	-2.0	-16.6	8.3	16.6	8.4
a 1-a 2	7.3	12.5	8.4	0.0	42.8	8.3
a 4+a 2	-2.1	6.0	-8.4	0.0	8.4	16.7

For SG, the intelligibility may be a3 lower than that of SO, because SG is synthesized without a plosive source. And, for the same reason, the intelligibility of FG may be a3' lower than that of FO. a3' may be different from a3 because the original fast speech may have only weak plosion or even no plosion.

The intelligibility of SM may be a4 lower than that of SG due to a mismatch of the formant model. The intelligibility of FM is assumed to be a4+a2 lower than FG, where a4 represents the effect of the failure of the model to adapt for slow speech, and a2 represents the effect of the failure of the model to predict the formant trajectories in fast speech since it was devised based on slow speech. In other words, the factor a2 represents the fact that the model does not take into account possible changes in articulation between two speaking rates, that is, reorganization.

3. Results and discussion

3.1 The intelligibility of vowels and consonants on average

Fig. 3(a) shows the average intelligibility of the three vowels /a,i,u/ for each subjects in the six speech groups. The box-whisper graph in this figure shows the minimum, 25%-tile, median, 75%-tile and the maximum of the intelligibility scores averaged for three vowels in reference to each of the five subjects. Fig. 3(b) shows the average intelligibility of the four consonants /b,d,g,r/ for the six speech groups. Table 1 shows the four factors estimated from the results shown in Fig. 3 based on the relationships shown in Fig. 2.

As shown in Fig. 3(a), the medians of the intelligibility for the six speech groups are 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.

This result indicates that the use of a formant model with a model voicing source does not decrease intelligibility (a4, a3, a3'=0.0, as shown in Table 1). The disregard of reorganization in the formant model slightly increases the intelligibility (a2=-2.1). Concerning the vowels, it can be suggested that the formant model maintains or even slightly improves the intelligibility compared to the original speech at the slow and fast speaking rates.

On the other hand, as shown in Fig. 3(b), the medians for the consonants are SO:91.7%, SG:81.3%, SM:83.3%, FO:79.2%, FG:68.8% and FM:62.5%. The four factors accounting for the intelligibility are a1=20.8, a2=8.3, a3=10.4, a3'=10.4 and a4=-2.0, as shown in Table 1.

For the consonants, the use of a model voicing source without plosion decreases the intelligibility by about 10%. The use of the formant model slightly increases the intelligibility by about 2% (a4=-2.0). The disregard of reorganization may reduce the intelligibility by about 8%. Also, a1 is estimated at 20.8, which means that the decrement in the intelligibility due to speed or shortening in the fast speech is large. For the consonants in the slow speech, the formant model works well on average and even slightly improves the intelligibility compared to SG. However, for the fast speech the formant trajectories predicted by the model decrease the intelligibility by about 6%.

3.2 The intelligibility of the individual consonants

In Figs. 4 (a) to (d) the average intelligibility is shown for the individual consonants in the same way as in Fig. 3. Table 1 shows the four factors accounting for the intelligibility. Since these factors are estimated from medians, the average of each factor for the four consonants is not the same as the value listed in column C.

As shown in Fig. 4(a), there is a 25% difference in the intelligibility of /b/ between SO and SG, while there is no difference between FO and FG. The intelligibilites of SM and FM are higher than those of SG and FG, respectively, which indicates that the model formant trajectories give higher intelligibility than those obtained by the analysis. Because the plosion for /b/ becomes unclear in the fast original speech, the disregard of the plosive source does not decrease the intelligibility. The for-

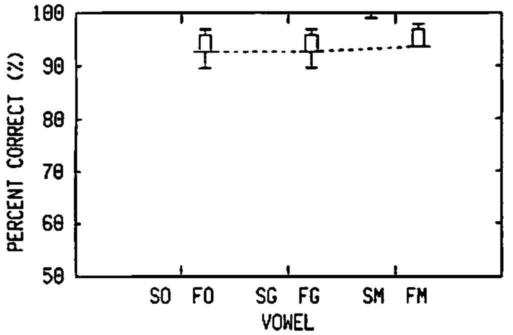

Fig. 3(a). The average intelligibility of the three vowels.

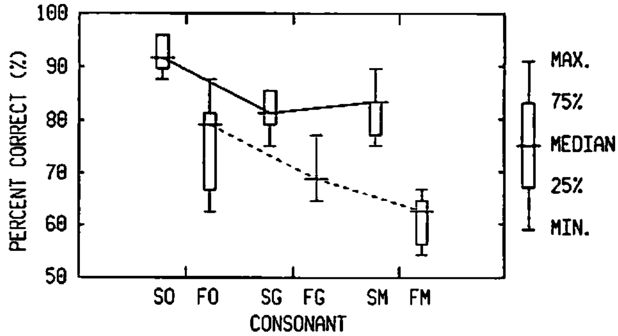

Fig. 3(b). The average intelligibility of the four consonants.

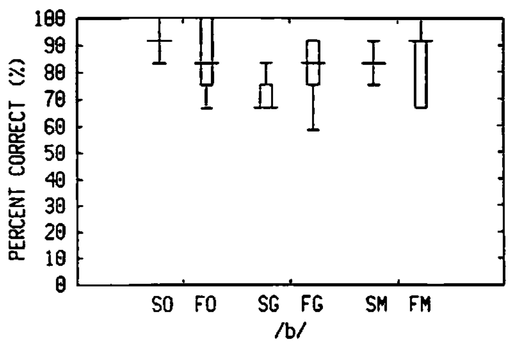

Fig. 4(a). The intelligibility of /b/ in 6 cases.

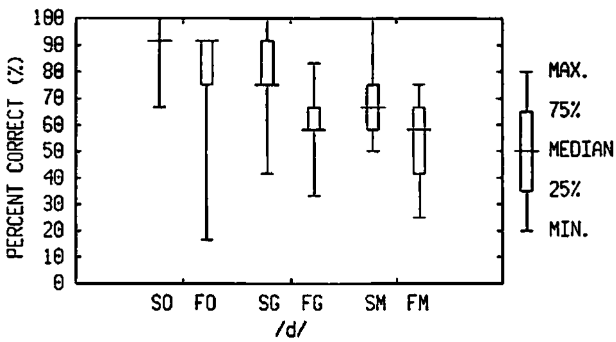

Fig. 4(b). The intelligibili- of /d/.

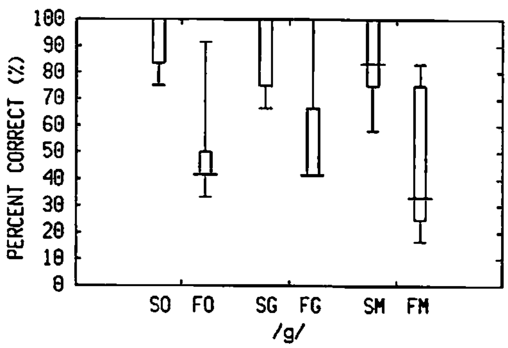

Fig. 4(c). The intelligibility of /g/.

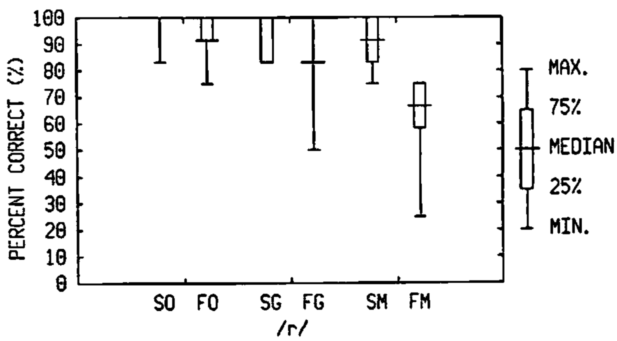

Fig. 4(d). The intelligibility of /r/.

mant model may improve the intelligibility, because it does not make the formant trajectories unclear around the /b/ closure, which are sometimes unclear in fast speech in the same positions.

For /d/ shown in Fig. 4(b), the use of a model voicing source without plosion largely decreases the intelligibility, to 16.7% for the slow speech (SO-SG), and to 33.4% for the fast speech (FO-FG). Also the formant model decreases the intelligibility, to 8.3% for the slow speech. This indicates that the plosive source is important for /d/, especially in fast speech. Thus, the formant model should also be modified at least for /d/, which is not easy to produce or to perceive in Japanese because of phonological constraints. These syllables are used only in words of foreign origin and are usually pronounced as /dzi/ or /dzu/, respectively. This constraint may affect our results.

For /g/ in the fast original speech (FO), as shown in Fig. 4 (c), the intelligibility is 41.7%, which is quite low compared with that of the other consonants. For this consonant, the use of a model voicing source without plosion does not decrease the intelligibility, but the formant model decreases it to about 16.6%, as represented by the differences between SM and SO or SG. For /g/, the decrement in the intelligibility seems to be accounted for mainly by the shortening of the segments due to the speaking rate. Since the formant transition of /g/ is slower than that of the other stops, the rate change may greatly affect the intelligibility of /g/.

For the /r/ shown in Fig. 4 (d), the use of the formant model decreases the intelligibility to about 8.4%, and the use of a model voicing source without plosion decreases the intelligibility to 8.4% only for the fast rate. The largest decrement is due to the effect of speed, or the shortening in segmental duration in fast speech, as indicated by a1, which is 16.6%. The effect of reorganization is estimated as 8.3%, which means that the discrepancy between the model formant trajectories and the those obtained by the analysis is not negligible for fast speech.

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 that model. The intelligibility of vowels on average was 100% at a slow speaking rate and was about 93% at a fast rate, which was about twice as fast as the slow rate. The intelligibility of the consonant, was 83% for the slow rate and about 63% for the fast rate. It was found that the formant model slightly improves the intelligibility of vowels at both speaking rates and that of consonants at the slow rate compared with the speech synthesized using formant trajectories obtained by analysis. However, for the consonants in the fast speech, the formant model decreases the intelligibility by about 6%. The use of a model voicing source without plosion decreases the intelligibility by about 10%, and the disregard of reorganization was estimated to reduce the intelligibility by about 8%

for the consonants. It was also found that the model fitted best for vowels and the consonant /b/ and worst for the consonant /g/.

Acknowledgement

This study was 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).

References

1) J. Liljencrants :Speech synthesizer control by smoothed step functions, STL-QPSR 4/1969, 43-50 (1970).
2) H. Fujisaki, M. Yoshida, Y. Sato, Y. Tanabe: Automatic recognition of connected vowels using a functional model of the coarticulatory process, J. Acoust. Soc. Jpn, 29, 636-638(1974).
3) S. Yokoyama and S. Itahashi: Approximation of formant trajectory by second order system with applications to consonants, Proc. Acoust. Soc. Japan (1975.5), 89-90 (1975).
4) Y. Sato and H. Fujisaki:Formulation of the process of coarticulation in terms of formant frequencies and its application to automatic speech recognition, J. Acoust. Soc. Jpn, 34,3, 177-185 (1978).
5) D.J. Broad, R.H. Fertig:Formant-frequency trajectories in selected CVC utterances, J. Acoust. Soc. Am.,47,1572-1582(1970).
6) D.J. Broad and F. Clermont: A methodology for modeling vowel formant contours in CVC context, J. Acoust. Soc. Am., 81(1) 155-165 (1987).
7) J.L.Miller:Effects of speaking rate on segmental distinctions, (Perspectives on the study of speech. P.D.Eimas and J.L.Miller Eds., Lawrence Erlbaum Associates, New Jersey), 39-74 (1981).
8) B. Lindblom: Spectrographic study of vowel reduction, J. Acoust. Soc. America, 35(11), 1773-1781 (1963).
9) T. Gay:Effect of speaking rate on diphthong formant movements, J. Acoust. Soc. Am, 44, 1570-1573 (1968).
10) R.R. Rerbrugge and D. Shankweiler: Prosodic information for vowel identity, J. Acoust. Soc. Am, 61, S39 (1977).
11) T. Gay:Effect of speaking rate on vowel formant movements, J. Acoust. Soc. Am., 63(1), 223-230 (1978).
12) D. O'Shaughnessy:The effects of speaking rate on formant transitions in French synthesis-by-rule, Proc. 1986 IEEE-IECEJ-ASJ, Tokyo, 2027-2039 (1986).
13) D.P. Kuehn and K.L. Moll:A cineradiographic study of VC and CV articulatory velocities, J. Phonetics, 4, 303-320 (1976).
14) Y. Sonoda:Effects of speaking rate on articulatory dynamics and motor event, J. Phonetics, 15, 145-156 (1987).
15) J.E. Flege: Effects of speaking rate on tongue position and velocity of movement in vowel production, J. Acoust. Soc. Am., 84(3), 901-916 (1988).
16) K. Harris: Mechanisms of duration change, (in Speech Communication 2, G. Fant Ed., Almqvist & Wiksell), 299-305 (1974).
17) T. Gay, T. Ushijima, H. Hirose, and F. Cooper:Effect of speaking rate on labial consonant-vowel articulation, J. Phonetics,

2, 47-63 (1974).
18) S. Imaizumi, S. Kiritani, H. Hirose, S. Togami, K.Shirai:Preliminary report on the effects of speaking rate upon formant trajectories, Ann. Bull. RILP(1987), 21, 147-151 (1987).
19) S. Imaizumi, S. Kiritani:Effects of speaking rate on formant trajectories and inter-speaker variations, Ann. Bull. RILP (1989), 23, 27-37 (1989).
20) S. Imaizumi, S. Kiritani: Perceptual evaluation of a glottal source model for voice quality control, Proc. 6th Vocal Fold Physiology Conference, Stockholm, 1-10 (1989).