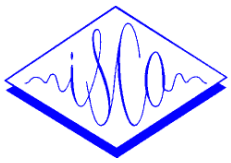

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

A MODEL OF DYNAMIC CHARACTERISTICS OF THE VOICE SOURCE AND FORMANT TRAJECTORIES

Satoshi Imaizumi, Hiroshi Imagawa and Shigeru Kiritani

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

ABSTRACT

This paper describes a model of the voice source and of formant trajectories which can be used in developing a high-fidelity speech synthesizer. A polynomial model was used to generate the glottal source. Formant trajectories are modelled as the sum of two kinds of functions: one represents vowel-to-vowel transitions and the other represents the effects of surrounding consonants upon the formants. The intelligibility and fidelity were tested for the speech synthesized based on the model at slow and fast speaking rates. Compared to speech obtained by an analysis-synthesis method, the model slightly improved the intelligibility of vowels at both speaking rates, and of consonants at slow rate. For consonants at fast rate, the model made the intelligibility decrease by 6%. The polynomial model of the glottal source could reproduce to some extent delicate voice quality differences in vowels uttered at various pitch and loudness. It was found that this model is useful as a high-fidelity synthesizer with variable speaking rate.

I. INTRODUCTION

Synthesis of natural sounding speech with various voice qualities and various speaking styles still remains as a seemingly unattainable goal. Many researchers have been trying to reach this goal by developing voice source models and formant transition models[1-15].

For instance, Fant and his colleagues[2-5] have introduced a four parameter model describing the time derivative of the glottal volume velocity waveform. Fujisaki and Ljungqvist [6] have proposed a seven parameter model which might have wider flexibility than other glottal source models. On the other hand, Klatt [7] and Hasegawa et al [8] stated that an additive noise component must be included into the glottal source model to synthesize female voice quality with sufficient naturality.

Concerning formant trajectories, smoothed step functions are used in some studies [10,11], where step inputs represent formant targets for vowels or consonants. Some studies propose a linear summation model of putative target formant frequencies of vowels and temporal functions representing effects of adjacent consonants [12,13].

Although these models seem to have the capability to describe some phenomena in the glottal source and in the formant trajectories, there has been few assessment results showing the ability of these models to synthesize high quality speech with variable speaking style.

In this paper, we propose a functional model to describe the dynamic characteristics of the glottal source and of the formant transitions. The present model was tested by synthesizing speech at slow and fast speaking rates and performing an intelligibility and fidelity test.

II. METHOD

2.1 Glottal source model

The inverse filtered waveform, or the time derivative of the glottal volume velocity waveform, was approximated in each cycle by the following polynomial function g(t),

$$\begin{aligned} g(t) &= a_1(t-t_1)^3 + a_2(t-t_1)^2 + a_3, && 0 < t < t_1, \\ &= a_3, && t_1 < t < t_2, \\ &= a_4(t-t_1)^4 + a_5(t-t_1)^3 + a_6(t-t_1)^2 + a_3, && t_2 < t < T, \end{aligned} \quad (1)$$

where t=0 is the negative peak in the inverse filtered waveform, and t=T is the duration of one period. Furthermore, g(t) was set to be continuous at t=0, t_1, t_2, T, and $g'(t_1)=0.0$. The parameters t_1, t_2, a_i, (i=1,6) were determined based on the least square error criterion between the actual inverse filtered waveform $g_i(t)$ and the model g(t). One example from a female speaker is shown in Fig. 1.

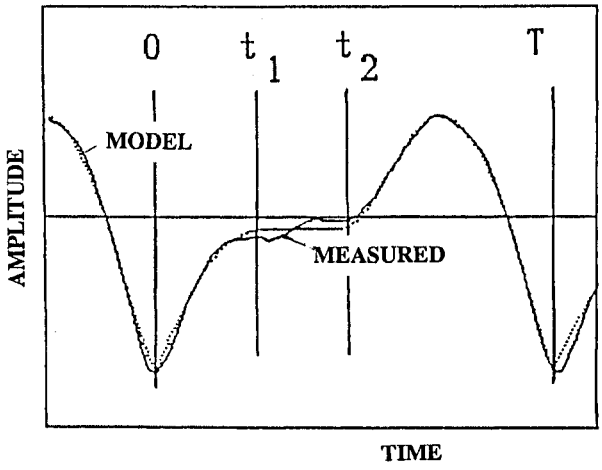

Fig. 1. The polynomial model of the glottal source adapted to an inverse filtered waveform for vowel /a/ uttered by F_2.

2.2 Formant transition model

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

$$F_n(t)= U_n(t)-C_{nf}(t)-C_{np}(t) \quad (2)$$

Here, $U_n(t)$ 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

$$W_j(t)= R_{1,j}+d_i(t)(R_{i,j}-R_{i-1,j}), \quad (3)$$

$d_i(t)=1-\{1+b_j(t)\}\exp(-b_j(t))u(t-t_i)$,

$b_j(t)=(t-t_i)/g_j$, $u(t-t_i)=1$ $t>t_i$, $=0$ $t<t_i$,

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 [11].

$$U_1=W_1,\ U_2=h(W_2W_3)^{0.5},\ U_3=(W_2W_3)^{0.5}/h, \quad (4)$$

$$h=s^{0.5},\quad q=(W_2W_2+W_3W_3)/W_2W_3,$$

$$s=q-(qq-4(1-kk))^{0.5}/2(1-kk),\quad k=0.2.$$

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.

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

$$C_{nf,i}(t)=c_{nf,i}\exp\{-(t_{f,i}-t)/g_f\},\quad t_{p,i}<t>t_{f,i}, \quad (6)$$

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 model, 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 variable 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.

2.3 Voice source and formant frequency data collection

In order to determine model parameters, the following two kinds of speech materials were recorded and analyzed. 1) Sustained vowels and vowel sequences. 2) $V_1CV_2CV_1$ samples surrounded by a carrier sentence, where C was one of /b, d, g, r/ and V_1,V_2 were /a, i, u/. The speech material 2) was recorded at two speaking rates, fast and slow.

These materials were recorded from 5 male speakers M_1, M_2, .. M_5, and 5 female speakers F_1,F_2,.., F_5, who had no laryngeal pathology. Each speaker uttered each item three times at three loudness levels and at three pitch levels, and of the material 2) five times at the most comfortable level. An electroglottogram (EGG) was also recorded simultaneously.

Formants frequencies were estimated by a covariance LPC analysis with pitch synchronous frames corresponding to glottal closure intervals derived from the EGG signal [9]. The parameters for controlling the voice source was derived by inverse filtering of the recorded speech.

The estimation of the model parameters was carried out based on a least square error method and careful interactive modification using the sustained vowels and the VCVCV samples uttered slowly and clearly. The detailed method was reported on in other place [9,14,15].

2.4. Perceptual assessment of the model.

For the assessment of the model, four kinds of speech samples, (O, I, G and M) were prepared. Here, O indicates the original speech, I the synthesis by analysis speech, G the speech synthesized using formant trajectories obtained by analysis and the polynomial glottal source model, and M the speech synthesized using the formant transition model and the polynomial glottal source model.

Three perceptual experiments were performed to examine how naturally and how variously voice quality could be reproduced by the polynomial model of the glottal source. Hearing subjects were five adults with healthy hearing who were not familiar with this study.

Experiment I was carried out to examine how closely the polynomial model of the glottal source could reproduce the voice quality of the original vowel uttered at five conditions (low, normal, and high pitch at normal loudness, soft, and loud at normal pitch). The test was an ABX test, in which X was one of the original vowels, and A and B were synthetic vowels. The subjects selected A or B which was felt more similar to the original vowel X. The degree of resemblance was calculated as the preference score (%), that is, the percentage of selection in the ABX test.

Experiment II was carried out to examine how closely the polynomial model of the glottal source could reproduce the voice quality of the VCVCV samples uttered at two speaking rates. The test was carried out for two kinds of synthetic speech (I and M), and original speech samples (O). These speech samples were synthesized or recorded at two speaking rates, slow (S) and fast (F). So, the speech tested consisted of the six groups, SO, FO, SI, FI, SM and FM. The rating was performed in a paired comparison method using a scale with 7 successive categories. The comparison was between O and I, and between O and M.

Experiment III was an intelligibility test to examine how well the model could reproduce the intelligibility of the VCVCV samples. The test was carried out for two kinds of synthetic speech (G and M), and the original speech (O). The speech tested consisted of the six groups, SO, FO, SG, FG, SM and FM. Each group consisted of 48 V_1CV_2 samples. 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 SG, FG, SM and FM, $V_1C_pV_2C_fV_3$ was synthesized to simulate effects of articulately undershoot, and then the segments of $V_1C_pV_2$ and $V_2C_fV_3$ were extracted.

III. RESULTS AND DISCUSSION

3.1. Experiment I.

The results of the perceptual judgments on the degree of resemblance between the original vowels and the synthetic vowels are shown in Figure 2. The resemblance degree is

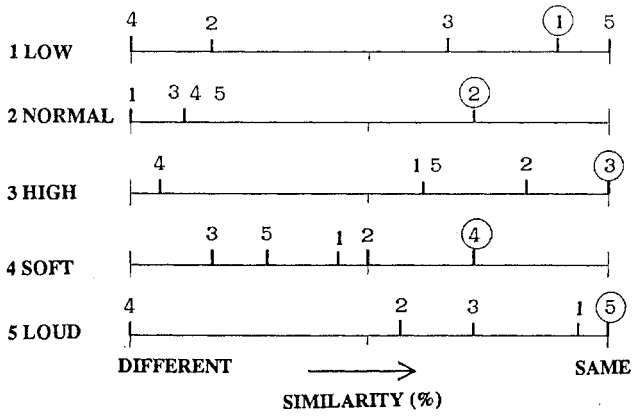

Fig. 2. Degree of resemblance between the original vowels and the synthetic vowels in the preference score(%). Each line indicates the utterance condition, 1:low pitch, 2:normal pitch, 3:high pitch, 4:soft, and 5:loud. The numbers on each line indicate the vowels synthesized with glottal model whose parameters were set to simulate the utterance conditions. The samples used were /a/ uttered by female speaker F_2.

indicated by the preference score (%), that is, the percentage of selection in the ABX test.

Each line in Fig. 2 shows the result for each of the simulated utterance conditions, 1:low pitch, 2:normal pitch, 3:high pitch, 4:soft, and 5:loud. The numbers on each line indicate the vowels synthesized with the glottal model, the parameters of which were set to simulate the utterance conditions. For the synthetic vowels on one line, the formant frequencies and their bandwidths, and the manner of fluctuation in pitch and in amplitude were set as the same as those of the original vowel simulated.

For instance, line No. 5 in Fig. 2 indicates the resemblance degree to the original "loud vowel" for the five synthetic vowels with different glottal source. The resemblance degree was the highest for No.5 synthetic vowel for which glottal parameters were set to simulate "loud" vowel, and was the lowest for No.4 synthetic vowel for which glottal parameters were set to simulate "soft" vowel. For the other sets of glottal parameters, the resemblance degree was somewhere between these extreme values.

These results shown in Fig. 2 indicate that the delicate voice quality differences can be reproduced by the polynomial model of the glottal source.

3.2. Experiment II.

The results of the perceptual judgments on the degree of resemblance between the original vowels and the synthetic vowels (I or M at two speaking rates S:slow and F:fast) are shown in Figure 3.

For the synthesis by analysis speech (SI, FI), the resemblance degrees are scattered between 5:similar, 6:very similar, and 7:perfectly the same. This result indicates that the analysis was carried out successfully especially for the slow utterances.

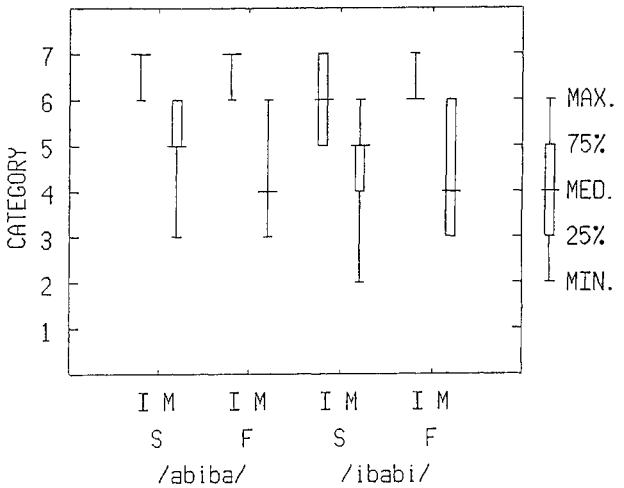

Fig. 3. Degree of resemblance between the original utterances of /abiba/, /ibabi/ and the synthetic samples. Symbols used are S: slow rate, F:fast rate, O:original, I:synthesis by analysis speech, M:model synthetic speech. Categories are 1:completly different, 2:very different, 3:different, 4:neutral, 5:similar, 6:very similar, 7:perfectly the same.

For the speech synthesized using the formant transition model and the polynomial glottal source model, the medians of the resemblance degrees scattered between 4:neutral and 5:similar. The resemblance degree was lower for the fast speech than for the slow speech.

This result indicates that the model works fairly well, although some additional strategies are needed to control parameters especially at fast speaking rates.

3.3. Experiment III.

Fig. 4(a) shows the average intelligibility of three vowels /a, i, u/ for each hearing subjects for the six speech groups. The box-whisker 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 hearing subjects. Fig. 4(b) shows the average intelligibility of four consonants /b, d, g, r/ for the six speech groups.

As shown in Fig. 4(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. 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. It should be noted that the intelligibility of the vowels at the fast rate was somewhat low because of the very short vowel segments.

On the other hands, as shown in Fig. 4(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%. For the consonants in the slow speech, the use of model voicing source without any plosive noise source decreases the intelligibility by 10%(=SG-SO). The use of the formant model slightly increases the intelligibility by 2%(=SM-SG). The formant

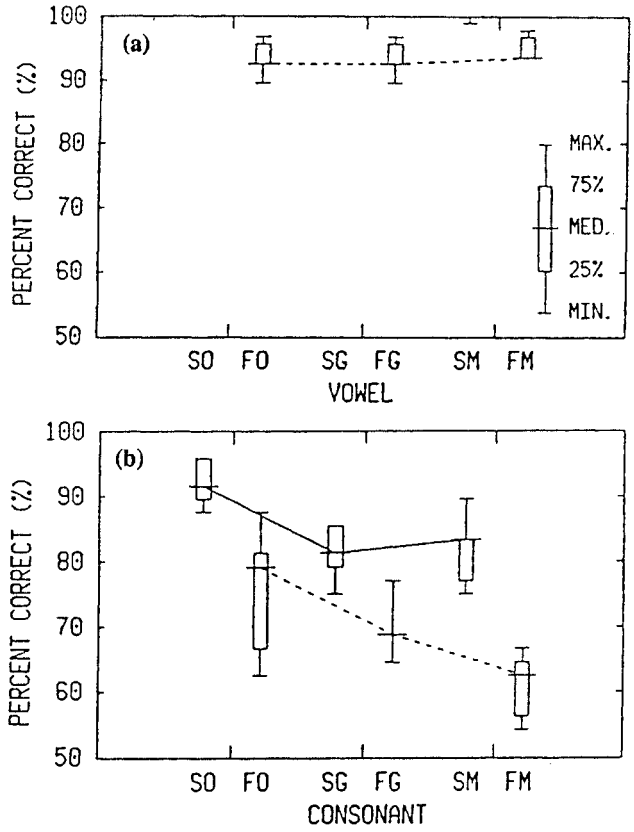

Fig. 4. The box-whisker graph of the intelligibility scores for the three vowels (a) and for the four consonants (b).

model works well on average, and even slightly improve the intelligibility comparing with that of SG. For the consonants in the fast speech, the formant trajectories predicted by the model decreases the intelligibility by 6% (=FG-FM).

In this model, only the timing parameters are variable depending on the speaking rate. However, detailed inspection on the formant trajectories suggests that there might be some changes in the vowel targets and the parameters representing transient speed depending on the speaking rate. Although the present model works rather well, it should 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 and also by proper implementation of plosive noise source for the stop consonants.

IV. CONCLUSIONS

The present study obtained the following results.

1)The polynomial model of the glottal source can reproduce to some extent the delicate voice quality differences in vowels uttered at various pitch and loudness. The model also works fairly well for VCVCV utterances, although some additional strategies are needed to control parameters especially at fast speaking rates.

2)The formant model slightly improves the intelligibility of vowels in both speaking rates and that of consonants in the slow rate compared with the speech synthesized using the formant trajectories obtained by the analysis. However, for the consonants in the fast speech, the formant model decreases the intelligibility by 6%. These results suggest that the model works well, although it should be noted that some additional strategies to control parameters such as transient speed, which are fixed at the present model, might improve the intelligibility of the consonants especially at fast speaking rates.

These results indicate that the model to generate dynamic characteristics of the voice source and of formant trajectories can be useful in developing a high-fidelity speech synthesizer.

ACKNOWLEDGEMENT

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

REFERENCES

[1] A. E. Rosenberg, ":Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. America., 49(2), 583-598, 1971.

[2] G. Fant, J. Liljencrants and Q. Lin, "A four-parameter model of glottal flow," STL-QPSR, 4/1985, (KTH, Stocholm), 1-13, 1986.

[3] G. Fant and Q. Lin:, " Frequency domain interpretation and derivation of glottal flow parameters. STL-QPSR, 2-3/1988, (KTH, Stocholm), 1-21, 1988.

[4] C. Gobl and A. N. Chasaide, "The effects of adjacent voiced/voiceless consonants on the vowel voice source: A cross language study. STL-QPSR, 2-3/1988, (KTH, Stocholm), 23-59, 1988.

[5] I. Karlsson, "Glottal waveform parameters for different speaker types.STL-QPSR, 2-3/1988, (KTH, Stocholm), 61-67, 1988.

[6] H. Fujisaki, M. Ljungqvist, "A comparative study of glottal waveform models," IEICE Technical Report (EA85-58), 23-29, 1985.

[7] D. H. Klatt, "Acoustic correlates of breathiness: First harmonic amplitude, turbulent noise, and tracheal coupling," J. Acoust. Soc. America, 82(SI), S91, 1987.

[8] K. Hasegawa, T. Sakamoto, H. Kasuya, " Effects of glottal noise on the quality of synthetic speech,"Proceedings of ASJ Spring Meeting (March 1987), 205-206, 1987(In Japanese).

[9] S. Imaizumi and S. Kiritani,"Perceptual evaluation of a glottal source model for voice quality control,"Proc. 6th Vocal Fold Physiology Conference, Stockholm, 1-10, 1989.

[10] J. Liljencrants,"Speech synthesizer control by smoothed step functions,"STL-QPSR 4/1969, 43-50, 1970.

[11] H. Fujisaki, M. Yoshida, Y. Sato, and Y. Tanabe,"Automatic recognition of connected vowels using a functional model of the coarticulatory process," J. Acoust. Soc. Jpn, 29, 636-638, 1974.

[12] D. J. Broad & R. H. Fertig,"Formant-frequency trajectories in selected CVC utterances," J. Acoust. Soc. Am.,47, 1572-1582, 1970.

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

[14] S. Imaizumi, S. Kiritani, H. Hirose, S. Togami, K. Shirai,"Preliminary report on the effects of speaking rate upon formant trajectories," Ann. Bull. RILP, 21, 147-151, 1987.

[15] S. Imaizumi and S. Kiritani,"Effects of speaking rate on formant trajectories and inter-speaker variations," Ann. Bull. RILP, 23, 27-37, 1989.