Ann.Bull.RILP(1989)No.23, 109-118

A POLYNOMIAL GLOTTAL SOURCE MODEL FOR THE SYNTHESIS OF VARIOUS VOICE QUALITIES

Satoshi Imaizumi, Shigeru Kiritani,
Hisayuki Fukawa* and Shuzo Saito*

Introduction

Although techniques for speech synthesis by rule have been significantly improved, synthesis of natural sounding speech with various voice qualities still remains as a seemingly unattainable goal. Many researchers have been trying to reach this goal by developing voice source models by which intra- and inter-speaker variability in voice quality can be controled[1-7].

For instance, Fant et al.[2,3] have introduced a four parameter model describing the time derivative of the glottal volume velocity waveform and have tried to synthesize female voice quality with high fidelity. Fujisaki and Ljungqvist[4] have proposed a seven parameter model which might have wider flexibility than other glottal source models. On the other hand, Klatt[5] and Hasegawa et al[6]. have insisted that an additive noise component must be included into the glottal source model to synthesize female voice quality with sufficient naturality. Although these studies provide a fruitful discussion on advanced techniques of Hi-Fi speech synthesis, there are few results reported on how naturally and how variously voice quality can be reproduced by these glottal source models.

In this paper, we examined how naturally and how variously a seven-parameter polynomial model can represent the voice quality of five male and two female speakers.

Method

Data Recording

The following speech materials were recorded and analyzed.

1)Sustained vowels and vowel sequences.
/a/, /i/, /u/, /e/, /o/, /aiueo/, /uoaei/

2)Three sentences consisting only of vowels and semi-vowels
/aoi ei o ou/ (Somebody drives a blue ray away.)
/yayoi wa ayu o ou/ (Yayoi follows a sweetfish.)
/iwao wa yayu o yuu/ (Iwao says meaningless things.)

These materials were recorded from 5 male speakers M_1, M_2, .. M_5, and 2 female speakers F_1 and F_2, who had no laryngeal pathology. Each speaker uttered each item three times at three loudness levels and at three pitch levels. The speech signal was recorded on a PCM Data Recorder through a high quality condenser

* Department of Engineering, Kogakuin University

microphone (B&K2234) whose frequency characteristic was flat (within 1dB) in the range of 10Hz to 10kHz. An electroglottogram (EGG)[8] was also recorded simultaneously to indirectoly observe vocal cord vibration. The EGG signal yielded the glottal closure intervals which were used for a pitch-synchronous covariance LPC analysis[7,8].

The speech material reported here is vowel /a/ uttered at normal pitch and normal loudness for each speaker.

Inverse Filtering

In order to estimate the glottal volume velocity waveform, formants were estimated based on a covariance LPC analysis with pitch synchronous frames corresponding to glottal closure intervals derived from the EGG signal[7,8]. The glottal closure intervals were derived in the same way reported in other sources[7], that is, one interval was determined so as to begin at one positive peak in the EGG time derivative and end at the following negative peak, the length being Ta(n) for the nth pitch period. The beginning of the actual analysis frame was shifted dt later according to the time delay it took for the sound wave to propagate from the glottis to the microphone positioned 15cm away from the lips.

Because the formant trajectories obtained in this manner sometimes revealed cycle by cycle fluctuations especially for the female voice, the formant frequencies and bandwidths were modified manually through an interactive program. This program displayed the speech waveform and its power spectrum, the inverse filtered waveform and its power spectrum, and the EGG time derivative which indicated the glottal closure intervals. The optimal formant frequencies and bandwidths were searched manually so as to minimize ripples in the inverse filtered waveform during the glottal closure intervals and also formant-like peaks in their power spectrum.

The time derivative of the glottal volume velocity waveform was estimated via an inverse filtering in which only one set of the lower five formant frequencies and bandwidths selected from a steady portion of each utterance was used. In other words, the cycle by cycle variation in formant trajectories was avoided.

The Parameter Estimation of the Glottal Source Model.

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

$$\begin{aligned} g(t) &= a(t-t_1) + b && 0 < t \leq t_1, \\ &= b && t_1 < t \leq t_2, \\ &= c(t-t_1)^3 + d(t-t_1)^2 + e(t-t_1) + b && t_2 < t \leq T, \end{aligned} \qquad (1)$$

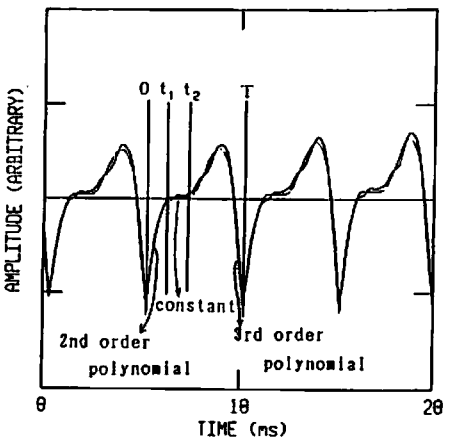

Fig. 1. The polynomial model of the glottal source adapted to a measured glottal source waveform obtained by inverse filtering /a/ uttered by F_2.

where t=0 is the negative peak in the inverse filtered waveform, and t=T is the duration of one pitch. The parameters t_1, t_2, a, b, c, d, and e 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.

Perceptual Experiments

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. Subjects were 6 students with normal hearing capacity.

Experiment I was carried out to examine how closely the voice quality of the original vowel was reproduced by the polynomial model of the glottal source. The subjects rated the degree of resemblance between the original vowel and the vowel synthesized using the polynomial model. For the sake of comparison, they also rated the resemblance between the original vowel and the vowel synthesized using Rosenberg's Type B model of the glottal source[1]. The rating was performed in a paired comparison method using a scale with 7 successive categories, 1:completly different, 2:very different, 3:different, 4:neutral, 5:similar, 6:very similar, 7:perfectly the same.

Experiment II was carried out to examine how variously the voice quality of vowels uttered by five male speakers M_1, ..., M_5 were reproduced by the glottal model using a multi-dimensional scaling method[9,10].

Five vowel samples of 0.5 s in length, O_1, O_2, .. , O_5, corresponding to the five male speakers M_1, M_2, .. , M_5, were

resynthesized using one pitch interval extracted from the inverse filtered waveform. Then, using one pitch interval from the polynomial model of the glottal source adapted to each vowel, five vowels G_1, G_2, .. , G_5 having a length of 0.5 s were synthesized. The pitch and its fluctuation were the same for all samples as those observed from /a/ uttered by M_1. The constant intervals corresponding to the glottal closure periods were lengthened or shortened to align the pitch for all samples.

The listening subjects rated the dissimilarity in voice quality for each of all possible pairs of O_1, O_2, ..., and G_1, G_2, .. , G_5. The ratings on dissimilarity were then analyzed by the multidimensional scaling method INDSCAL included in the ALSCAL program[10], and the similarity among these 10 vowel samples was represented by mutual distance in a two-dimensional space.

Experiment III was carried out to examine the perceptual effects of fluctuations in the waveform (W), pitch (P) and amplitude (A) of the glottal source upon the natualness of the synthetic vowels. Five kinds of synthetic vowels -- P_1, P_2, ..., P_5 -- were generated to contain various fluctuations observed in the original vowel P_0. P_1 contained W+P+A; P_2:P+A; P_3:P; P_4:A; and P_5:no fluctuation. Here, waveform variation means the cycle to cycle variation in the modeled glottal voice source. The pitch fluctuation was the cycle to cycle variation in the intervals between negative peaks in the inverse filtered vowel waveform. The amplitude fluctuation was the ycle to cycle variation in the amplitude of the negative peaks in the inverse filtered vowel waveform.

All possible pairs of P_0, P_1, .. , and P_5 were made and presented to the listeners in random orders. Each listener selected one member of each pair felt to be more natural than the other.

Results and Discussion

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 1. The samples used were /a/ uttered by five male speakers and 2 female speakers. The symbol G indicates the vowel synthesized using the polynomial model, and R represents the one synthesized with Rosenberg's glottal source model.

As shown in Fig.2, for all speakers the ratings for the synthetic vowels with the polynomial model of the glottal source (G) are higher than those for the vowels synthesized with Rosenberg's model (R). This result shows that the polynomial model of the glottal source is better than Rosenberg's model at reproducing the voice quality of the vowels for which glottal source models are adapted.

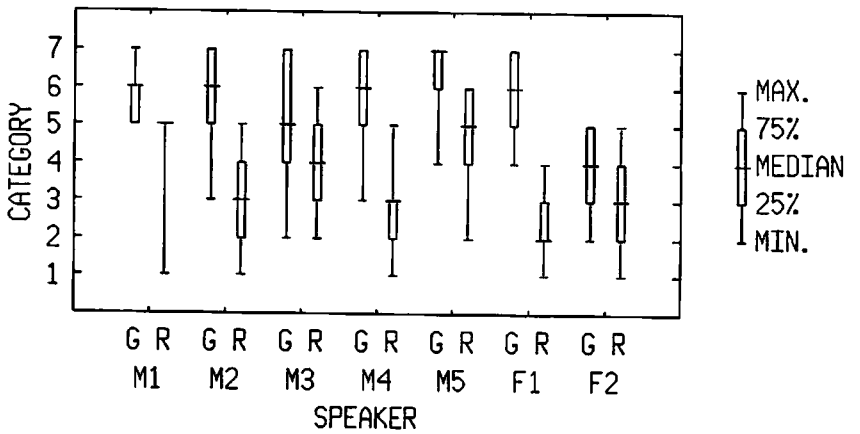

Fig. 2. The results of perceptual judgments on the degree of resemblance between original vowels and synthetic vowels with the polynomial model of the glottal source (G), and that between original vowels and synthetic vowels with Rosenberg's voice source (R). Category 7 represents the greatest possible resemblance.

For the vowels uttered by the male speakers, M_1, M_2, ..., M_5, the medians of the ratings for the polynomial model scatter between 5:similar and 7:perfect the same. Those for Rosenberg's model lie between 3:different and 5:similar. This result indicates that the polynomial model of the glottal source can reproduce the voice quality of the male speakers analyzed here.

For the female speaker, F_1, the median of the rating scores for the polynomial model is 6:very similar, although the median of the ratings for Rosenberg's model is 2:very different. On the other hand, for the female speaker, F_2, the median of the ratings for the polynomial model is 4:neutral, and the median for Rosenberg's model is 3:different. These results indicate that the polynomial model of the glottal source can reproduce some female voice qualities.

Figures 3(a) and 4(a) show the inverse filtered waveform and its model representation for F1 and F2 respectively. Figures 3(b) and 4(b) show their power spectra. The polynomial model of the glottal source for F_1 reproduces the voice quality of the original vowel very well, while that for F_2 does not.

In Fig. 4(a), the inverse filtered waveform or the measured glottal source have positive main lobes which skew right, and this characteristic is not represented well enough in the model. The intervals which are approximated by constant b in the model contain waveform fluctuation in the measured glottal source. The negative peaks in the model source are too sharp compared to those of the measured glottal source. In Fig. 4(b), harmonics higher than 2kHz in the power spectrum of the measured glottal source are not clear. On the other hand, the model shows clear

harmonics for a higher range than 2 kHz. These discrepancies are not so large in F_1 as shown in Fig. 3, although the waveform fluctuation in the intervals which are approximated by constant b in the model are not approximated well.

The skewing and waveform fluctuation observed in Fig. 4(a) might be effects of the source-tract interaction[11-13]. The disappearance of harmonics higher than 2 kHz might be due to the turbulence noise. These effects are not approximated in the polynomial model of the glottal source, thus the voice quality of F_2 which reveals these effects clearly can not be reproduced with high fidelity.

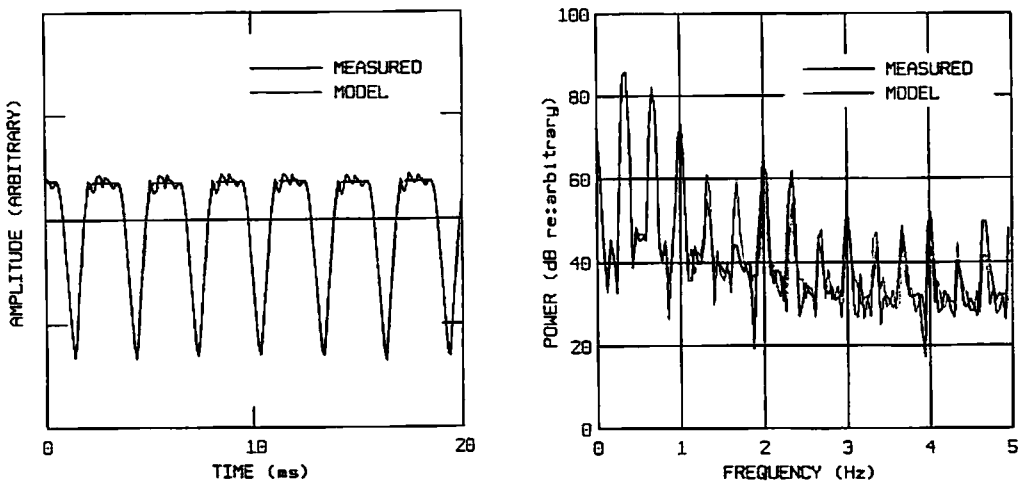

Fig. 3. The measured glottal source waveform and its model representation (a), and their power spectra (b). Female speaker F_1.

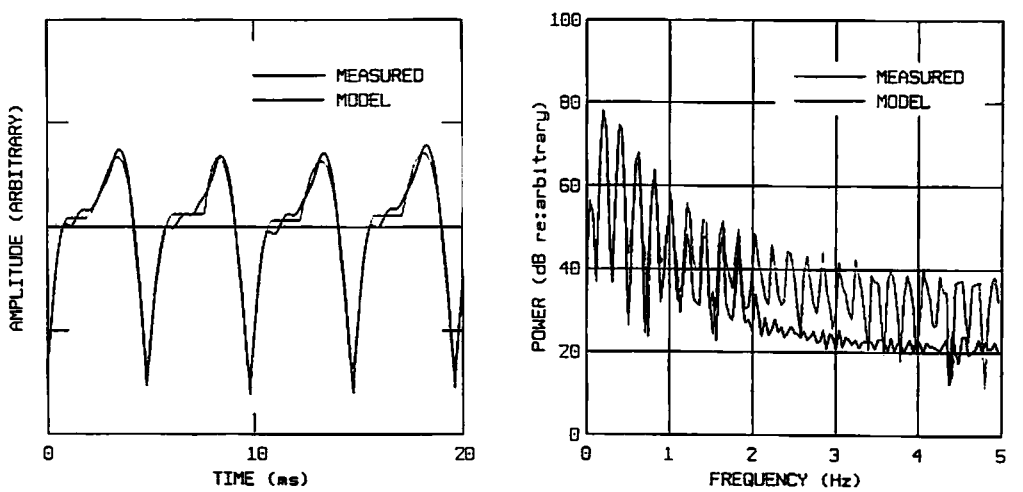

Fig. 4. Same as in Fig. 3, but for female speaker F_2.

Experiment II

The result of Experiment II is shown in Fig. 5. In this figure, the similarity among the 10 synthetic samples were represented by mutual distance in a two-dimensional space.

Fig. 5 shows that there are three types of similarity between O_n synthesized from the inverse filtered waveform and G_n synthesized from the model. Here, n indicates the speaker number. Type 1: for M_1, O and G are relatively close on. Type 2: for M_2, M_4 and M_5, O and G are close on Dimension D_1, but distant on D_2. Type 3: for M_3, O and G are distant on D_1, but close on D_2. This result indicates that the voice quality of each speaker has various aspects, some of which can be reproduced by the polynomial model of the glottal source, and some of which can not.

Fig. 5 also shows that the voice samples O_n resynthesized from the inverse filtered waveform scatter in two-dimensional space, while G_n resynthesized from the model scatter in a one-dimensional manner on the line S_1 and separate into two groups G_2, G_3 and G_4 versus G_1 and G_5. In other words, the two-dimensional variability of the voice quality is maintained in O_n, but is reduced to one dimension in G_n.

These results must be interpreted through an examination of the acoustical and perceptual meanings of dimensions D_1 and D_2, or S_1 and S_2. According to our preliminary examination, S_1 may indicate the contrast between "strained" versus "asthenic" voice quality, or in another definition, a "hyper-fuctional/tense" versus "hypo-functional/lax" quality. G_1 and G_5 have stronger

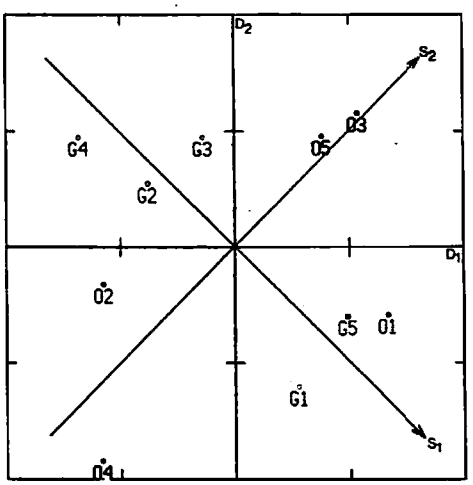

Fig. 5. Two dimensional representation of the similarity among vowels resynthesized from the inverse filtered waveform O_n and those from the polynomial model G_n. Here, n indicates the speaker number M_n, n=1,2, .., 5. D_1 and D_2 are the dimensions extracted by the INDSCAL analysis, while S_1 and S_2 are their rotated version to interprete the configuration.

harmonics in the high frequency range than the others. On the other hand, S_2 may indicate a "breathy/noisy" versus "rough" quality. These results indicate that the polynomial model of the glottal source can reproduce the voice quality represented by S_1, but not that represented by S_2.

Experiment III

Figure 6 shows the results of Experiment III, which was carried out to examine the perceptual effects of fluctuations in the waveform (W), pitch (P) and amplitude (A) of the glottal source upon the natualness of the synthetic vowel. In this experiment, five synthetic vowels -- P_1, P_2, ..., P_5 -- were generated to contain various fluctuations observed in the original vowel /a/, P_0, uttered by F_1. Then, all possible pairs among P_0, P_1, .. , P_5 were presented to four listeners in random orders. Each listener selected one from each pair which was felt to be more natural than the other. The selection rate for the six samples is shown in Fig. 6.

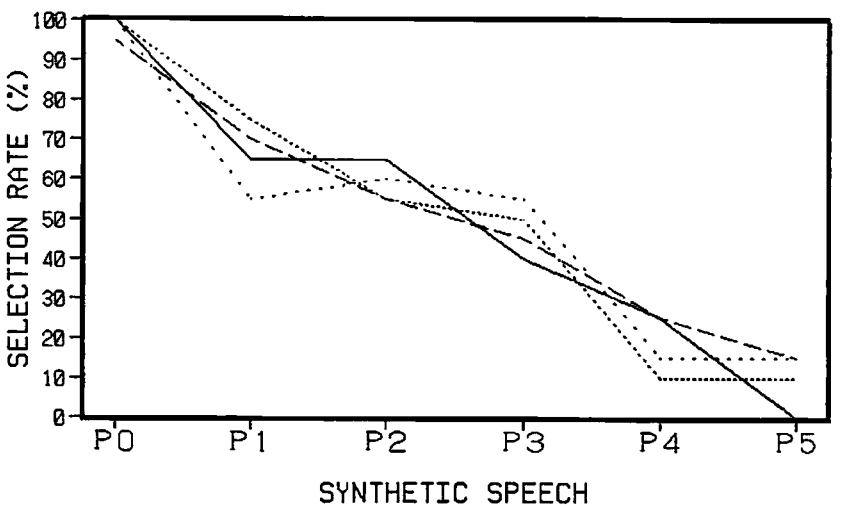

Fig. 6. The rate of selection as the more natural vowel in paired comparisons by four subjects. The tested fluctuations were waveform variation (W), pitch fluctuation (P) and amplitude fluctuation (A). P_0: original vowel /a/ uttered by F_1; P_1:synthetic vowel which contained W+P+A; P_2:P+A; P_3:P; P_4:A; P_5:no fluctuation.

As shown in Fig. 6, the original voice sample P_0 was selected as most natural. Although there were slight differences among the listeing subjects, P_1 which contained fluctuation in waveform, pitch and amplitude was selected as second in naturalness. P_2, which possessed fluctuation in pitch and amplitude, had almost the same selection rate as P_1. Although P_3, containing only pitch fluctuation, had lower selection rate than P_1 and P_2, it showed a higher rate than P_4, which possessed only amplitude fluctuation and P_5 which had no fluctuation.

This result indicates that fluctuation in pitch, amplitude and waveform affects the naturalness of synthetic vowels in this order. Proper modeling of the pitch fluctuation is quite important, because synthetic vowels without any pitch fluctuation here sound quite unnatural. On the other hand, waveform fluctuation in the glottal source did not largely affect naturalness compared to pitch fluctuation in this study. However, the effect of waveform fluctuation on naturalness might have been underestimated in this study, because a cycle by cycle estimation of the model parameters sometimes emphasizes waveform variation, which may generated a hoarse-like voice quality.

Conclusions

The present study obtained the following results.

1) For male voices, the polynomial model of the glottal source can reproduce to some extent the voice quality of original vowels for which the model parameters are adapted. In a simple paired comparison based on a succsesive category method, Experiment I, the degree of resemblance between an original vowel and a synthetic one with the model was quite high. However, a detailed examination of the voice quality based on the multi-dimensional scaling method, Experiment II, showed that some aspects of voice quality are still remain unrepresented in the model.

2) For voices which contain turbulence noise in the high frequency range, and those which contain waveform perturbation and skewing possibly caused by source-tract interaction, the polynomial model fails to reproduce good voice quality.

3) Proper modeling of pitch fluctuation is important for the naturalness of the synthetic voice.

Acknowledgement

This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas, the Ministry of Education, Science and Culture, Japan.

Refferences

1) Rosenberg, A. E.:Effect of glottal pulse shape on the quality of natural vowels. J. Acoust. Soc. America., 49(2), 583-598, 1971.
2) Fant, G., J. Liljencrants and G.Lin: A four-parameter model of glottal flow. STL-QPSR, 4/1985, (KTH, Stocholm), 1-13, 1986.
3) Fant, G. and Q. Lin: Frequency domain interpretation and derivation of glottal flow parameters. STL-QPSR, 2-3/1988, (KTH, Stocholm), 1-21, 1988.
4) Fujisaki, H., M. Ljungqvist: A comparative study of glottal waveform models. IEICE Technical Report (EA85-58), 23-29,

1985.
5) Klatt, D. H.: Acoustic correlates of breathiness: First harmonic amplitude, turbulent noise, and tracheal coupling. J. Acoust. Soc. America, 82(S1), S91, 1987.
6) Hasegawa, K., 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).
7) Imaizumi, S., and S. Kiritani: A study of formant trajectories and voice source characteristics based on the closed phase analysis. Preprints of the Second Symposium on Advanced Man-Machine Interface Through Spoken Language, 1988.
8) Childers, D. and J. Larar: Electro-glottography for laryngeal function assessment and speech analysis. IEEE Trans. BME-31, 12, 807-817, 1984.
9) Kruskal, J. B.: Nonmetric multidimensional scaling: A numerical method. Psychometrika, 29, 115-219, 1964.
10) Takane, Y., F.W. Young and J. de Leeuw: Nonmetric individual differences multidimentional scaling: an alternating least squares method with optimal scaling features. Psychometrika, 42, 7-67, 1977.
11) Rothenberg, M.: Acoustic interactions between the glottal source and vocal tract. in Vocal Fold Physiology, Ed. K.N. Stevens and M. Hirano, (Univ. Tokyo Press, Tokyo), 305-328, 1981.
12) Ananthapadmanabha, T., and G. Fant: Calculation of true glottal flow and its components. STL-QPSR 1/1982, (KTH, Stockholm), 1-30, 1982.
13) Koizumi, T., S. Taniguchi, and S. Hiromitsu:Two-mass models of the vocal cords for natural sounding voice synthesis. J. Acoust. Soc. Amereca, 82(4), 1179-1192, 1987.