TR－A－0096

# A glottal waveform model for high quality speech synthesis

*Seiichi TENPAKU and Tatsuya HIRAHARA*

1990. 12.14

ATR視聴覚機構研究所

〒619-02 京都府相楽郡精華町乾谷　☎07749-5-1411

**ATR Auditory and Visual Perception Research Laboratories**

Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1411
Facsimile: +81-7749-5-1408
Telex: 5452-516 ATR J

# A glottal waveform model for high quality speech synthesis*

Seiichi TENPAKU
and
Tatsuya HIRAHARA

ATR Auditory and Visual Perception Research Laboratories,
Seika-cho, Soraku-gun, Kyoto 619-02, Japan
(e-mail address: tenpaku%atr-hr.atr.co.jp@uunet.uu.net)

## *ABSTRACT*

A new glottal waveform model for high quality speech synthesis is proposed and the results of the perceptual evaluations for synthesized speech using the proposed model and other models are compared. The proposed glottal waveform model consists of two parts; a waveform generator and a spectrum shaping filter. A third order polynomial, whose coefficients are determined by combinations of open quotient (OQ), speed quotient (SQ), amplitude of voicing (AV) and fundamental frequency (F0), is used for the waveform generator. A second order infinite impulse response (IIR) filter, which is designed to control the spectral tilt and the relative amplitudes of lower harmonic components using two parameters, serves as the spectrum shaping filter. Thus, the parameters have a direct effect on the waveform and its spectral shape. Using three kinds of information (F0, power and formant) extracted from the 8 different Japanese words produced by two professional announcers (one male and one female), 80 synthesized speech stimuli were prepared for preference tests. The stimuli were generated by cascade formant synthesizer using 5 different glottal waveform models: the proposed model, Fant's model [Fant, Liljencrants & Lin, 1985], Fujisaki's model [Fujisaki & Ljungqvist, 1986], Klatt's model [Klatt, 1980] and Rosenberg's model [Rosenberg, 1971]. Results of the preference tests with 20 subjects by the proposed model are as good as those of the Fant and Fujisaki models.

PACS number : 43.72.Ja

* J. Acoust. Soc. Am. Suppl. 1, 88, S152.
*(Acoustical Society of America, 120th Meeting, San Diego, 26-30 November 1990)*

## INTRODUCTION

It is now well known that, in order to synthesize high quality speech, models of both the vocal tract filter function and the sound source must be improved. In particular, it has become clear that the source plays an important role in giving desirable qualities to synthesized speech. Several glottal waveform models have been proposed to synthesize more natural speech sound, and recently such models have been used in text-to-speech synthesis [e.g., Carlson, Fant, Gobl, Granström, Karlson & Lin, 1989; Pinto, Childers & Lalwani, 1989; Klatt & Klatt, 1990]. There is however room for improvement, particularly as glottal waveform data covering many speakers and contexts ( speech material, speaking style, etc. ) are now available to test new models [e.g., Holmberg, Hillman & Perkell, 1988; Gobl, 1989]. In this paper, we propose a new glottal waveform model and compare its performance, via perception tests, to that of four major voicing source models.

## 1. BACKGROUND

There are two ways to model glottal flow. One way simulates the mechanics of the vocal cords [e.g., Flanagan & Ishizaka, 1978; Titze, 1984]. However, this way is perhaps too complex for speech synthesis applications. The other way models an idealized time-domain contour of the excitation signal without reference to the mechanical system and produces the glottal flow or the derivative of glottal flow. This way can be put to practical use.

There are two ways of representing speech production [Fant, 1983]. One consists of three parts: a voicing source, a vocal tract transfer function, and a radiation transfer function. The other way consists of just two parts: differentiation of the voicing source and a vocal tract transfer function. Differentiation of the voicing source includes the radiation transfer function specified in the first method. In this way, the derivative of glottal flow corresponds to the differentiation of voicing source. The derivative of glottal flow has been examined in recent studies because it eliminates the dc components found in direct measures of glottal flow and is less affected by different recording conditions. Because of the advantages of working with the derivative of glottal flow, the second representation appears to be the most promising.

In general, the glottal flow or its derivative can be determined by four time-based parameters and three amplitude-based parameters. Figure 1 shows an illustration of glottal flow and its derivative waveform. The four time-based parameters are the pitch period (T0); the open quotient (OQ), which is the ratio of opening time to pitch period [(t1+t2)/T0]; the speed quotient (SQ), which is the ratio of opening to closing time [t1/t2]; and the closing quotient (CQ), which is the ratio of closing time to pitch period [t2/T0]. The three amplitude-based parameters are the peak flow, the dc flow which is the minimum flow during the closed phase, and the ac flow which is calculated as peak flow minus dc flow.

However, frequency-domain properties are also very important to bring out the perceived speech quality ( e.g., breathiness, creakiness, brightness ). Time-based and amplitude-based parameters are mutually related to frequency-domain properties. In particular, the three frequency-domain properties, which are most important, are spectral tilt, the relative amplitudes of lower harmonic components and the irregularity of harmonic components at higher frequency.

Several voicing source models have been used for speech synthesis. These models can be classified into two types. One type is determined with time-based and amplitude-based parameters to simulate natural glottal flow or its derivative. The other type uses frequency-domain properties to affect perceptual qualities of synthesized speech.

An example of the first type of model is Rosenberg's [Rosenberg, 1971], which was defined using time-based and amplitude-based parameters. Rosenberg was concerned with the effect on naturalness of the variation of glottal waveform shapes. He proposed a number of glottal flow models using one amplitude and two time-based parameters: amplitude ($\alpha$), opening time (Tp) and closing time (Tn), as shown in Figure 2. Tp is the portion of the pulse with a positive slope; Tn is the portion of the pulse associated with a negative slope. In his model, the relative opening time Tp/T0 is equal to the OQ, the ratio of Tp/Tn implies the SQ, and the amplitude factor $\alpha$ is the ac flow. One of his proposed models is composed of trigonometric functions with one slope discontinuity at closing. This is the model usually referred to as the "Rosenberg model".

It is the second type of model whose low-pass filtered impulse has been used as a voicing source in synthesizers. Klatt [Klatt,1980] used an impulse train generator followed by a glottal resonator (RGP) and a glottal anti-resonator (RGZ) as a voicing source in his formant synthesizer. Figure 3 shows a block diagram of this model. The RGP works as a low-pass filter since it typically has one pole at the resonance frequency. The filtered impulse train has a spectrum envelope of approximately -12 dB per octave ignoring for the moment the effects of RGZ.

The Rosenberg and Klatt models were relatively simple and had little flexibility, because the Rosenberg model had only two time-based parameters and one amplitude-based parameter, and because Klatt paid attention only to spectral tilt. These limitations affected the quality of the synthesized speech. For example, speech produced using Klatt's model retains a slightly pulse-like sound. This is partly because the slope of the glottal spectra was fixed at -12 dB per octave, rather than being free to vary (e.g., -12 to -18 dB per octave [Monsen & Engebreston, 1977]).

Recently, more complex and realistic glottal waveform models have been proposed as a voicing source. The LF-model [Fant, Liljencrants & Lin, 1985] is one of them. The LF-model is referred to as the "Fant model" in this paper. This model is defined by glottal flow derivative parameters in two stages, as illustrated in Figure 4. The first stage is an opening phase with an exponentially increasing sinusoid function, which reaches the negative value Ee at Te. The second stage is a closing phase with an exponentially decaying function with the time constant Ta. The model involves four parameters in addition to amplitude factor (E0), and pitch period (T0). These four parameters are Ee/Ei, Rk, Rg and Ra. Ee/Ei is the ratio of the negative peak (Ee) to the positive maximum (Ei) in the flow. Rk is defined by Te/Tp-1 and depends on the open quotient. Rg is the ratio of the glottal frequency (Fg) to the fundamental frequency (F0). Ra is the ratio of the time constant Ta to the pitch period T0. In addition, intermediate parameters $\alpha$ and $\varepsilon$ are required to generate the waveform from the four parameters. However, the parameters $\alpha$ and $\varepsilon$ are calculated by arithmetical methods (e.g. Newton-Rapson method), and therefore increase calculation cost considerably.

Fujisaki and Ljungqvist [Fujisaki & Ljungqvist, 1986] proposed a glottal flow model in which the derivative of glottal flow is composed of polynomial segments, as shown in Figure 5. This model has four time-based parameters and three amplitude-based parameters. The four time-based parameters are: pitch period (T), open phase duration (W), pulse skew (S), and the time interval between glottal closure and time of maximum negative flow (D). The three amplitude-based parameters are: waveform slopes at glottal opening (A), prior to closure (B) and following closure (C). As this model has more parameters than the Fant model, it may be more flexible. However, it is more difficult to control the parameters.

The more recent glottal waveform models, such as Fant and Fujisaki models, were formulated using time-domain parameters. There are, of course, close relationships between time-domain entities and frequency-domain properties — e.g., the effect of waveform changes on its spectrum [Fant & Lin, 1988]. However, these relationships are indirect and complicated. In order to fit the shape of actual glottal flow, the number of time-domain parameters will increase. Increasing the number of parameters makes them more difficult to control. Thus, time-domain and frequency-domain properties should be treated separately, as much as possible. On the other hand, T0, OQ, SQ and ac flow are time-domain parameters which are indispensable in characterizing the shape of the glottal flow or its derivative. The spectral tilt and the relative amplitudes of lower harmonic components are basic properties in the frequency-domain. Since both time-domain and frequency-domain properties affect the perceived quality of synthesized speech, we propose a new glottal waveform model that controls these properties directly and easily.

## 2. NEW GLOTTAL WAVEFORM MODEL

We propose a new glottal waveform model consisting of two parts: a waveform generator and a spectrum shaping filter. The model is designed to allow direct control of the OQ, SQ, spectrum tilt and relative amplitudes of lower harmonic components. Figure 6 shows the block diagram of the model and the waveform generated by the proposed model.

The waveform generator produces a waveform S(t) using a third order polynomial equation, whose three coefficients are determined by four parameters: the fundamental frequency (F0), the amplitude of voicing (AV), the OQ and the SQ. The third order polynomial equation is as follows:

$$S(t)=\begin{cases} t(2a-3bt+4ct^2) & 0\le t\le T_0\times OQ \\ 0 & T_0\times OQ<t\le T_0 \end{cases} \qquad (1.1)$$

where

$$a=cyz \qquad (1.2)$$

$$b=c(y+z) \qquad (1.3)$$

$$c=\frac{AV\times T_0}{x^2(y-x)(z-x)} \qquad (1.4)$$

$$x=\frac{SQ}{SQ+1}T_0\times OQ \qquad (1.5)$$

$$y=T_0\times OQ \qquad (1.6)$$

$$z=x\frac{3y-4x}{2y-3x} \qquad (1.7)$$

$$T_0=1/F_0 \qquad (1.8)$$

S(t) is fed to the spectrum shaping filter. The filter, a second order infinite impulse response (IIR) filter, is designed to manipulate the spectral tilt as well as the relative amplitudes of lower harmonic components using two parameters α and γ. The transfer function of the filter is:

$$H(z)=\frac{1-\beta}{2}\cdot\frac{(1+z^{-1})(1-\alpha z^{-1})}{1-\beta z^{-1}} \qquad (2.1)$$

where

$$\beta=\frac{\varepsilon-1}{\varepsilon+1} \qquad (2.2)$$

$$\varepsilon=\frac{1}{tan(2\pi F_c/F_s)} \qquad (2.3)$$

$$F_c=\gamma F_0; \qquad \text{Lower cut - off frequency} \qquad (2.4)$$

$$F_s=\text{Sampling Frequency [Hz]}$$

Spectrum tilt is determined by the value of α, which is between 0.0 and 1.0. The higher frequency energy increases as the value of α approaches 1. The value of γ is positive and determines the lower cut-off frequency (Fc). Fc determines the relative amplitudes of lower harmonic components. When the value of $\gamma$ is less than 2.0, Fc is set lower than the second harmonic component 2F0. As a result, the level of the F0 component L(F0) is relatively enhanced since the level of the second harmonic component L(2F0) is attenuated by the filter. The waveform shape of the filter output is similar to a derivative glottal waveform E(t).

Figure 7 (a)-(e) shows output waveforms, their spectra and phase characteristics for five conditions. F0, AV and Fs were fixed at 200 Hz, 90 dB and 20 kHz, respectively. For each condition, the top row shows the model output waveform for one pitch period; the middle row shows the spectrum of the waveform; and the bottom row shows the phase characteristics.

As can be seen in the figure, the proposed model can directly control the time-domain parameters ( i.e. OQ and SQ ) and the frequency-domain.properties ( i.e. $\alpha$ and $\gamma$ ). Thus, comparing (a) with (b), opening time increases as a direct function of OQ. Similarly, comparing (a) with (c), SQ directly affects the skew of waveform shape. Comparing (a) with (d), as the value of $\alpha$ is increased, spectrum tilt is reduced. Comparing (a) with (e), when the value of $\gamma$ is decreased, the lower cut-off frequency (Fc) decreases. Hence, the F0 component energy is relatively enhanced. Thus, OQ, SQ, $\alpha$ and $\gamma$ have direct effects on the waveform shape and the spectrum of the waveform.

The model has simple structure because the waveform generator is determined by the third order polynomial equation and the spectrum shaping filter is determined by the second order IIR filter. Both the polynomial equation and the IIR filter can be easily implemented. Furthermore, the parameters they control are non-dimensional and, therefore, conveniently treated.

## 3. COMPARSION OF 5 MODELS

In Figure 8 examples of waveform shapes and spectra are shown for the five source models ( Rosenberg, Klatt, Fant, Fujisaki, and the proposed model ). Table 1 contains the parameters for these models. We tried to choose parameter values that would optimize the performance of each model. The figure shows the derivative glottal waveforms (top) and their spectra (bottom). The waveform shape resembles that of the Fant and Fujisaki models. Spectrum tilt for these models ranges between -6 dB per octave and -9 dB per octave. The proposed model is the only one that has a zero-pole at the Nyquist frequency. The first harmonic is higher than the second harmonic in the Rosenberg and Fant models.

Figure 9 shows execution times for the five models. **R**, **K**, **L**, **F**, and **T** represent execution times for the Rosenberg, Klatt, Fant, Fujisaki, and proposed models, respectively. Execution times were calculated on a SUN3/470GX with FPA+ (Sun Microsystems, Inc.) from the average of three measurement tests using the *time* command, which is a built-in UNIX command. The Fant model was by far the slowest of the five, because it used arithmetic iteration methods. Among the remainder, whose speeds are more acceptable, the Fujisaki model was the fastest. Execution times for the proposed model, **T**, was about the same as that of the Klatt model.

## 4 EVALUATION TEST

### *4.1. Stimuli*

In order to evaluate the performance of the proposed model, synthesized speech generated by the five models was compared for naturalness. Using Japanese words (8) produced by two professional announces (1 male and 1 female), synthesized speech samples of each word were generated by each model (5). This resulted in a total of 16 natural and 80 synthesized stimuli to be used in the preference test.

The eight Japanese words used in this experiment were made up of only vowels; /ai/ [love], /au/ [to meet], /aoi/ [blue], /iu/ [to say], /ie/ [house], /ue/ [hunger], /oi/ [nephew], and /ou/ [to run after]. They were recorded in a sound-proof room and then sampled at 20 kHz with 16 bit accuracy.

Fundamental frequency, formant frequency and bandwidth, and power were extracted from the 16 natural speech utterances (8 words × 2 speakers). The analysis conditions are listed in Table 2. Fundamental frequency (F0) was extracted by correlating coefficients of the polarity of the speech signal. Linear predictive analysis (LPC) was used to extract formant information and power. All extracted data were examined visually, and extraction errors were corrected manually.

Using the extracted formant data, cascade type formant filters were designed for each word. Nine formants were used for male speech materials and eight formants were used for female speech. While traced values were used for the lower four formants, averaged values over a word were used for the higher formants.

Using the extracted F0 data, power data and model parameters shown in Table 1, source waveforms were generated using each source model. The model parameters were fixed across the word. The amplitude of voicing (AV) was calculated by subtracting the designed formant filter gain contour from the extracted power contour. Figure 10 shows the formant frequency, F0, and AV used to synthesize /ai/ of male stimuli.

### *4.2 Procedure*

For each word, we prepared 30 stimulus pairs covering all possible combinations of stimuli produced in the five synthesized and one natural speech conditions. Then 240 paired stimuli of eight words were arranged in several quasi-random orders. They were recorded with a Digital Audio Tape-recorder (DAT; SONY DTC-1000ES) with an inter-stimulus interval (ISI) of 1 second and an inter-trial interval (ITI) of 3 seconds. Male and female stimuli were treated separately. These stimuli were presented to subjects through headphones (STAX SR-Lambda Pro.), in a sound-proof room at a level of 70 dB SPL.

Subjects were asked to judge which stimulus in each presented pair was more natural. They were allowed to judge equal naturalness, "equal response", when it was difficult to judge the preference. Two different series of stimuli pairs were presented to each subject. Ten male and ten female subjects participated in the preference tests. They all had normal hearing ability and none of them had previous experience with synthesized speech. The responses were collected and preference scores were calculated. In the calculation, the "equal response" was eliminated.

### *4.3 Results*

Since there were no differences between male/female subjects' response, responses of the two subject groups were combined. Figures 11 and 12 show the preference scores for the male speech and the female speech, respectively. In each figure, the preference scores for each speech stimulus category are plotted. **O** represents the natural speech. **R**, **K**, **L**, **F**, and **T** represent the speech synthesized with the Rosenberg, Klatt, Fant, Fujisaki, and proposed models, respectively. A box represents the ±25% range of the population score and a thick line in the box indicates the median value of the preference scores

With regard to male speech stimuli, the median value of the preference scores of natural stimuli was 20% to 30% higher than that of synthesized stimuli. Preference scores of the natural stimuli were generally higher than those of synthesized stimuli. However, in some words some subjects judged that synthesized stimuli were more natural than the natural stimuli. Among the five types of synthesized stimuli, there was little difference in the median values of the preference scores. Stimuli synthesized with the Fujisaki model (F) had the highest preference score and the smallest score deviation. Those with the proposed model came next, then those of the Fant model (L) and the Rosenberg model (R). Stimuli synthesized with the Klatt model (K) had the worst preference scores and the largest deviation.

With regard to female speech stimuli, in contrast to the results for male speech stimuli, the median value of the preference score of the natural stimuli was 99%. That is, all subjects judged the natural stimuli of female speech to be the most natural across all words. Furthermore, preference scores for the five types of synthesized stimuli were only 10% to 35%. Synthesized stimuli generated with the Fant model (L) had the highest preference score. Stimuli synthesized with the Rosenberg model (R) came next, then those of the proposed model (T) and the Fujisaki model (F). Again, stimuli synthesized with the Klatt model (K) had the worst preference score. Finally, the preference score deviations among words and subjects were very small compared to those of male speech stimuli.

## 5. DISCUSSION

In actual speech, glottal waveform parameters such as OQ and SQ vary continually during production; and the pattern of that variability is context specific — e.g., speaker, word. However, in this study, all parameter values except F0 and AV were kept constant. Fixing these source parameters undoubtedly affects the perceived quality of synthesized speech. In addition, we used parameter values for the Fant, Fujisaki and proposed models that are similar to those reported for synthesis of male speech [Carlson et al., 1989]. Thus, the synthesized stimuli driven by these source models might not be appropriate for female speech. Furthermore, there were power contour differences between the synthesized speech and the natural speech. Figure 13 shows an example of this difference, which might influence the judgement of naturalness.

There were relatively small differences in preference scores among the five types of synthesized stimuli. Namely, the perceived naturalness of synthesized stimuli was much the same regardless of the source waveform. This result suggests that, when all formant information and the intonation of actual speech are preserved, the small source waveform and/or spectrum differences do not play important roles in the perception of naturalness.

It is interesting that the preference scores of the synthesized stimuli using a simple source waveform, such as that provided by the Rosenberg and Klatt models, were as good as the more realistic waveforms generated by the more complex models. However, there is little room to improve the synthesized speech quality for these two simple models. Thus, the preference scores obtained for the Rosenberg and Klatt models might be at their upper limit. On the other hand, we believe there is ample room to improve the synthesized speech quality of the other models. We would expect the Fant, Fujisaki, and proposed models to perform much better if the model parameters were assigned more flexibly and appropriately (especially for female speech). For example, the glottal waveform parameters should be determined using analysis-by-synthesis for each word.

There are, of course, many other ways in which model performance can be compared. For example, instead of fixing parameters such as OQ and SQ while preserving the formant and intonation patterns of the original speech sample, the quality of synthesized speech could be evaluated using artificially determined formant, power and/or F0 information.

Finally, how we assess the quality of synthesized speech is very important. "Naturalness" is the most commonly used index of synthesized speech quality, since it is believed to represent the total quality of the stimuli. However, naturalness is one of the most difficult subjective judgements to interpret, because it is a vague concept composed of many unknown elements that probably vary from person to person. Although we have at present no alternative to the naturalness criterion, we would like to find other criteria for assessing the quality of synthesized speech.

## SUMMARY

In this report, a new glottal waveform model was proposed and its performance was compared to that of four other source waveform models via preference tests. Using three kinds of information, (F0, power and formant) extracted from natural speech, synthesized speech stimuli were generated for each of the five source models.

Results of the preference tests showed that when all formant information and the intonation of natural speech are preserved, the source waveform has little influence upon the perceived naturalness of the synthesized speech. Thus, some other method should be used to evaluate the performance of the source waveform in future. One example would be to evaluate the quality of synthesized speech using artificially determined formant information and/or intonation.

Despite the similarity in perceived performance, the new glottal waveform model has certain inherent advantages over some of the models tested. For example, waveform shape and the spectrum of the waveform can be manipulated easily and directly. Also, the proposed model is more flexible than the simpler source models proposed by Rosenberg and Klatt. In order to compare our model to those of Fant and Fujisaki, more realistic test criteria are needed such as flexible assignment of appropriately determined parameter values. In future, we hope to show that our model can improve the quality of synthesized speech.

## ACKNOWLEDGMENTS

Authors are very grateful to Dr. Eric Vatikiotis-Bateson for giving them many helpful comments on writing this paper.

## REFERENCE

Flanagan, J.L. and Ishizaka, K. (1978). "Computer Model to Characterize the Air Volume Displaced by the Vibrating Vocal Cords", J. Acoust. Soc. Am. 63, 1559-1565

Carlson, R., Fant, G., Gobl, C., Granström, B., Karlson, I., and Lin, Q. (1989). "Voice Source Rules for Text-to-Speech Synthesis", Proc. Int. Conf. Acoust. Speech Signal Process. ICASSP-89, 223-226

Fant, G. (1983). "The Voice Source-Theory and Acoustic Modeling", in *Vocal Fold Physiology* edited by Titze and Scherer, 453-464

Fant, G., Liljencrants, J. and Lin, Q. (1985). "A Four-Parameter Model of Glottal Flow", Speech Trans. Lab. Q. Prog. Stat. Rep. 4, Royal Institute of Technology, Stockholm, 1-13

Fant, G. and Lin, Q. (1988). "Frequency Domain Interpretation and Derivation of Glottal Flow Parameters", Speech Trans. Lab. Q. Prog. Stat. Rep. 2-3, Royal Institute of Technology, Stockholm, 1-21

Fujisaki, H. and Ljungqvist, M. (1986). "Proposal and Evaluation of Models for the Glottal Source Waveform", Proc. Int. Conf. Acoust. Speech Signal Process. ICASSP-86, 1605-1608

Gobl, C. (1989). "A Preliminary Study of Acoustic Voice Quality Correlates", Speech Trans. Lab. Q. Prog. Stat. Rep. 4, Royal Institute of Technology, Stockholm, 9-22

Holmberg, E. B., Hillman, R.E. and Perkell J. S. (1988). "Glottal Airflow and Transglottal Air Pressure Measurements for Male and Female Speakers in Soft, Normal, and Loud Voice", J. Acoust. Soc. Am. 84, 511-529

Klatt, D. H. (1980). "Software for a Cascade/Parallel Formant Synthesizer", J. Acoust. Soc. Am. 67, 971-995

Klatt, D. H. and Klatt, L. C. (1990). "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers", J. Acoust. Soc. Am. 87, 971-995

Monsen, R. B. and Engebretson A. M. (1977). "Study of Variations in the Male and Female Glottal Wave", J. Acoust. Soc. Am. 62, 981-993

Pinto, N. B., Childers, D. G., and Lalwani, A. L. (1989). "Formant Speech Synthesis: Improving Production Quality", IEEE Trans.. Acoust. Speech Sig. Proc. 37, 1870-1887

Rosenberg, A. E. (1971). "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", J. Acoust. Soc. Am. 49, 583-590

Titze, I. R. (1984). "Parameterization of the Glottal Area, Glottal Flow, and Vocal Fold Contact Area", J. Acoust. Soc. Am. 75, 570-580

Table 1. Lists of parameters for 5 source models.

| | |
|---|---|
| Rosenberg model | (Tp+Tn)/T0=0.5,Tp/Tn=1.0 |
| Klatt model | FGP=F0,BGP=2F0 |
| Fant model | Ee/Ei=2.5,Rk=0.4,Rg=1.0,Ra=0.025 |
| Fujisaki model | W/T=0.5,R/F=1.8,D/B=0.1,A=0.0 |
| Proposed model | OQ=0.5,SQ=1.8,$\alpha$=0.0,$\gamma$=5.0 |

Table 2. The parameters for acoustic-analysis.

| | |
|---|---|
| SAMPLING | 20 kHz, 16 bit |
| WINDOWING (HANNING) | 30 ms |
| FRAME PERIOD | 2.5 ms |
| ORDER OF LPC | 30 |
| DFT POINTS | 4096 |
| PRE-EMPHASIS FACTOR | 0.98 |

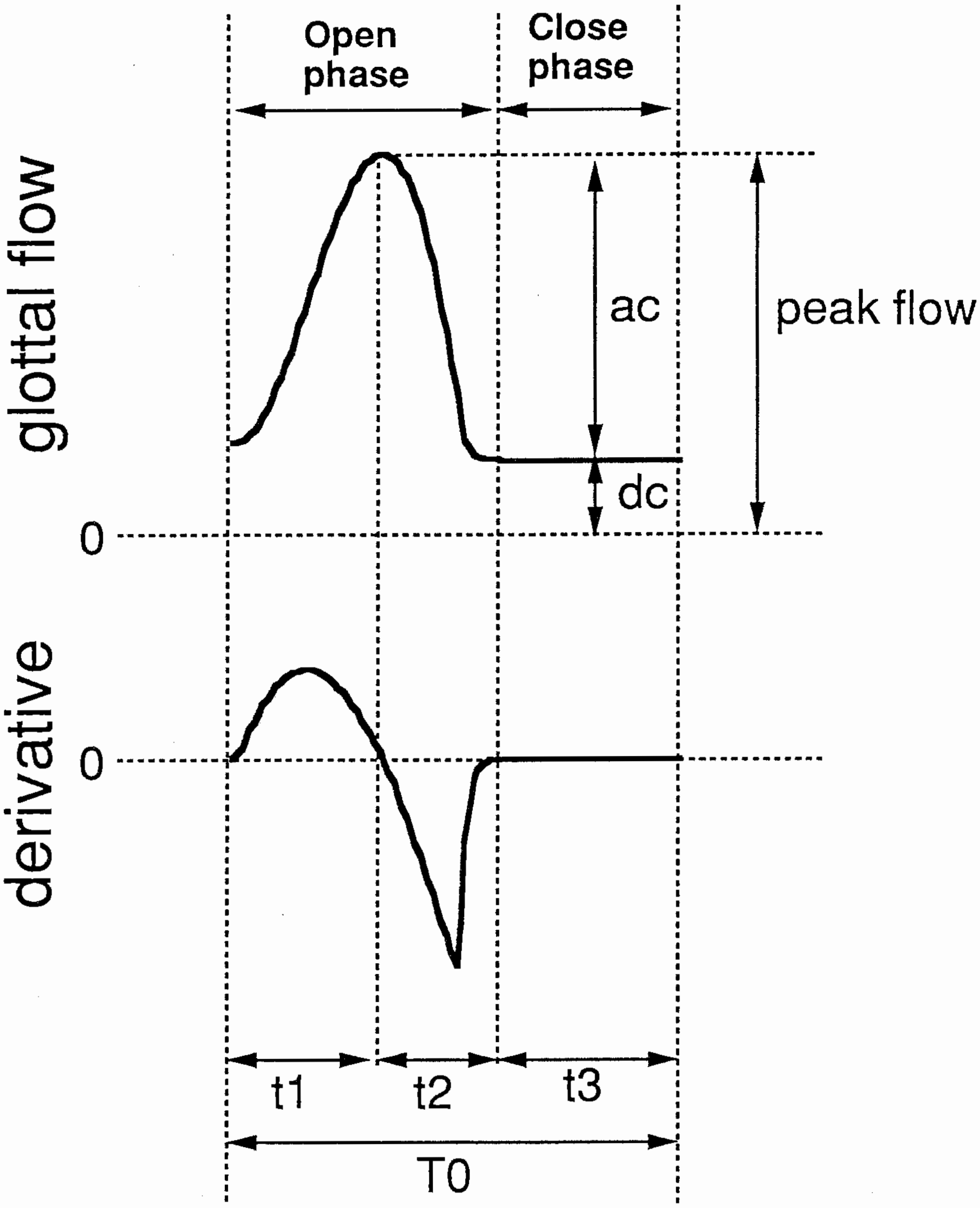


**Fig.1.** Illustration of glottal flow and its derivative. The glottal flow or its derivative can be essentially determined by four time-based parameters and three amplitude-based parameters. The four time-based parameters are the T0 (pitch period), the OQ (open quotient) which is the ratio of opening time to pitch period [(t1+t2)/T0], the SQ (speed quotient) which is the ratio of opening to closing time [t1/t2], and the CQ (closing quotient) which is the ratio of closing time to pitch period [t2/T0]. The three amplitude-based parameters are the peak flow, the dc flow which is the minimum flow during the closed phase and the ac flow which is calculated as peak flow minus dc flow.

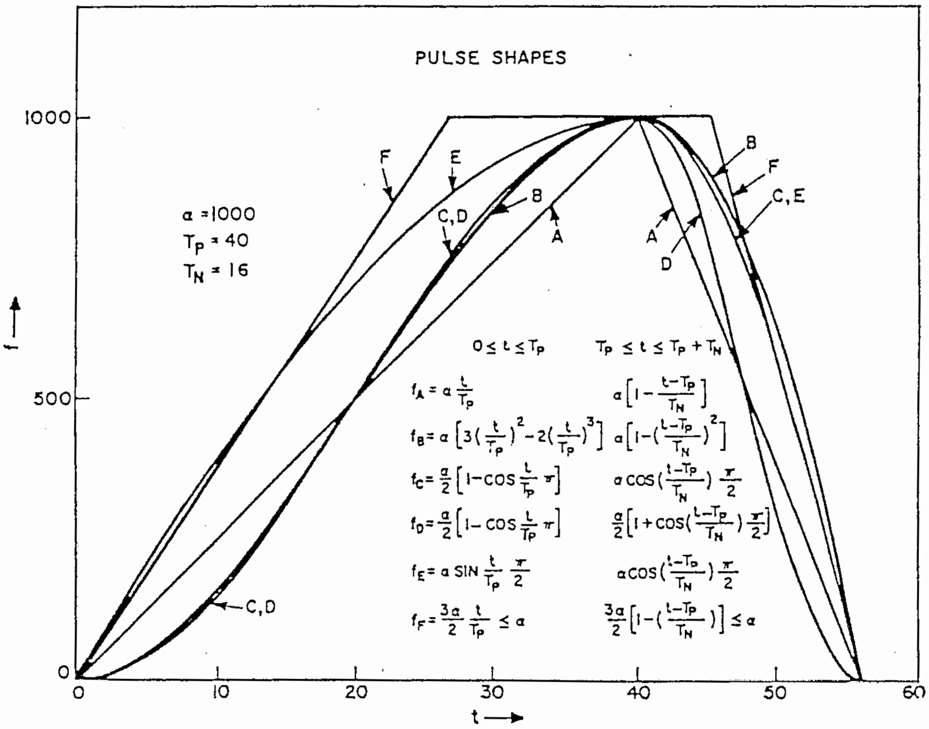


**Fig.2.** Shapes of the Rosenberg model from Rosenberg (1971, p.585)

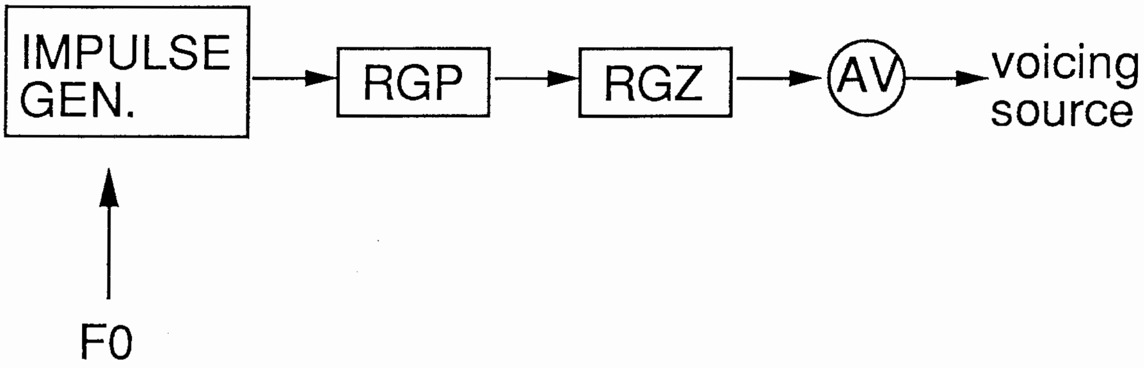
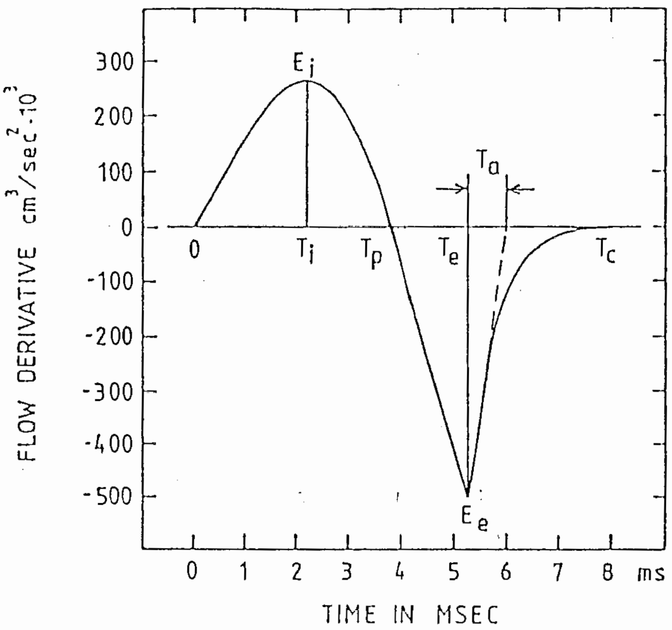


**Fig.3.** Block diagram of the Klatt model from Klatt (1980, p.975). RGP is a glottal resonator and RGZ is a glottal anti-resonator.



$$E(t) = E_0 e^{\alpha t} \sin \omega_g t$$

$$(t < T_e)$$

$$E(t) = \frac{-E_e}{\varepsilon T_a} \cdot \left[ e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_c-T_e)} \right]$$

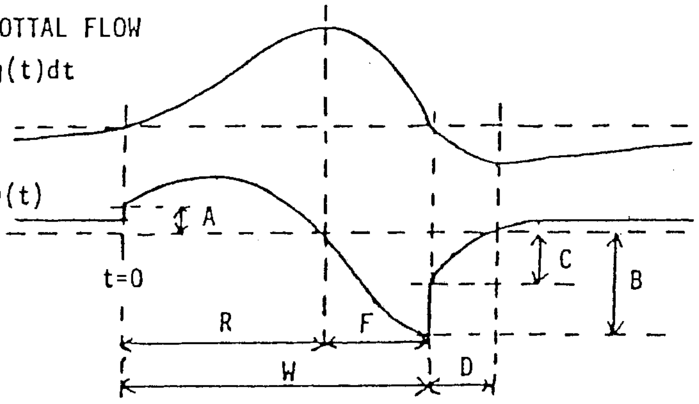
$$(T_e < t < T_c)$$

$$\omega_g = 2\pi F_g \qquad F_g = \frac{1}{2T_p} \qquad T_c = T0 = \frac{1}{F0}$$

$$R_g = \frac{F_g}{F0} \qquad R_k = \frac{T_e}{T_p} - 1 \qquad R_a = \frac{T_a}{T0}$$

$$O_q = \frac{T_e + T_a}{T0} \qquad O_q' = \frac{T_e}{T0} \qquad F_a = \frac{1}{2\pi T_a}$$

**Fig.4.** Differentiated glottal flow with the Fant model from Fant & Lin (1988, p.2).



GLOTTAL PARAMETERS

W - PULSE WIDTH (R+F)
S - PULSE SKEW (R+F)/(R-F)
D - GLOTTAL CLOSURE TIMING

A - SLOPE AT GLOTTAL OPENING
B - SLOPE PRIOR TO CLOSURE
C - SLOPE FOLLOWING CLOSURE

$$g(t) = \begin{cases} A - \frac{2A+R\alpha}{R}t + \frac{A+R\alpha}{R^2}t^2, & 0<t\le R, \\ \alpha(t-R) + \frac{3B-2F\alpha}{F^2}(t-R)^2 - \frac{2B-F\alpha}{F^3}(t-R)^3, & R<t\le W, \\ C - \frac{2(C-\beta)}{D}(t-W) + \frac{C-\beta}{D^2}(t-W)^2 & W<t\le W+D, \\ \beta & W+D<t\le T, \end{cases}$$

$$\text{where} \quad \alpha = \frac{4AR-6FB}{F^2-2R^2} \quad \text{and} \quad \beta = \frac{CD}{D-3(T-W)},$$

$$T = \text{fundamental period}.$$

**Fig.5.** Parameters and formulas for the Fujisaki model from Fujisaki & Ljungqvist (1986, p.1607).

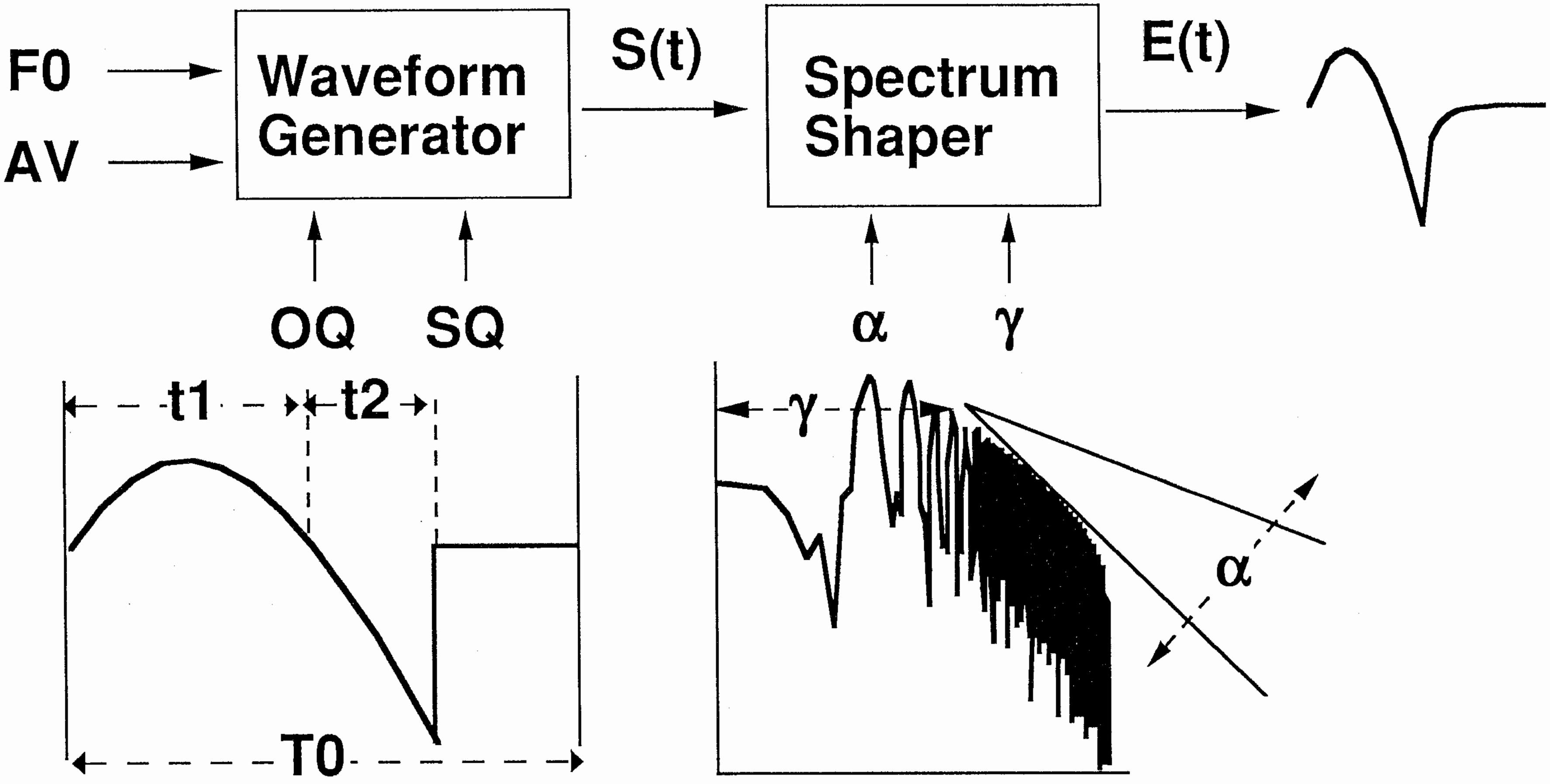
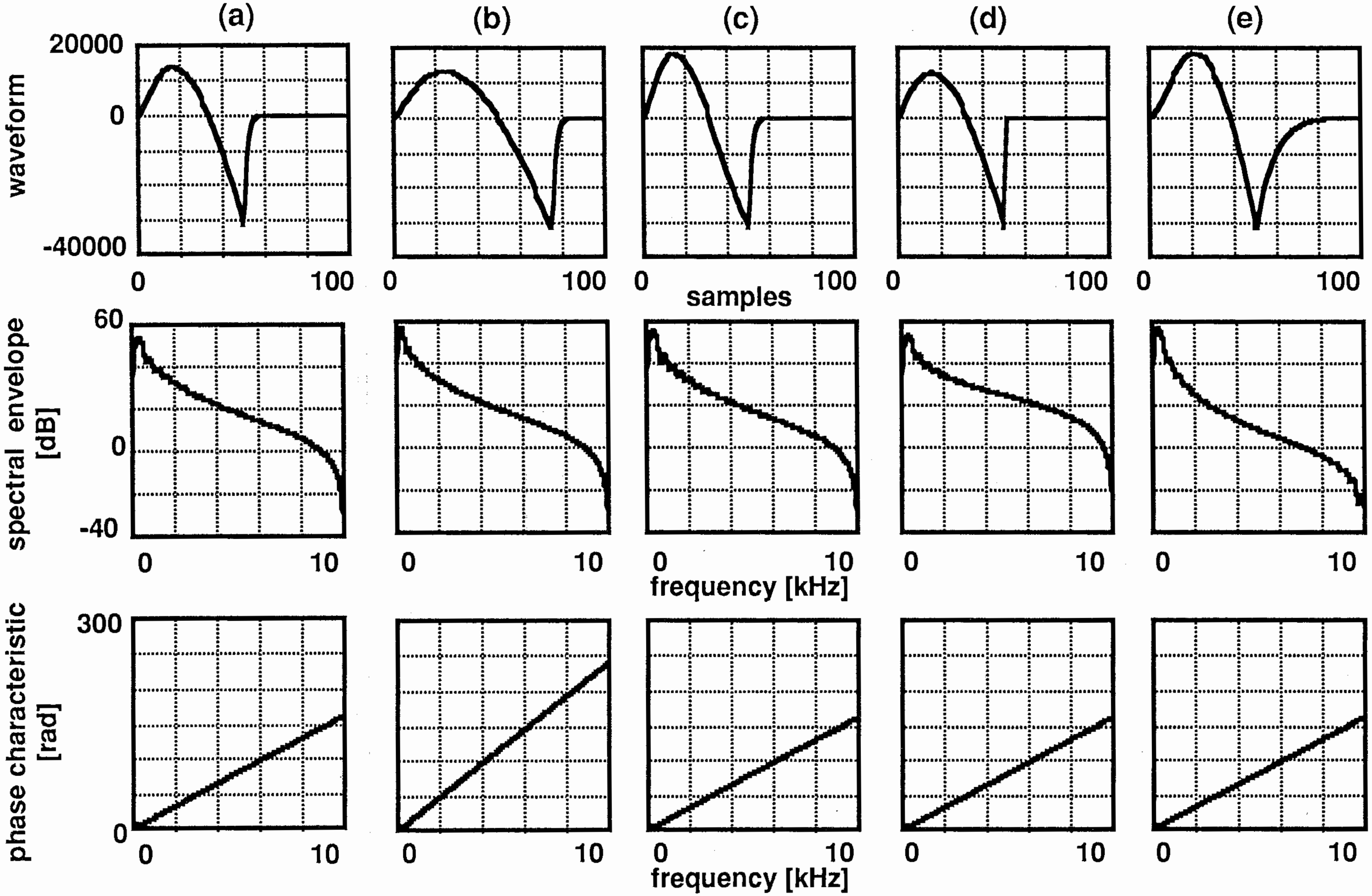


**Fig. 6.** Block diagram of the proposed glottal waveform model. The proposed model consists of two parts: waveform generator and spectrum shaper. The waveform generator is determined by four parameters: the fundamental frequency (F0), the amplitude of voicing (AV), the open quotient (OQ) and the speed quotient (SQ). The spectrum shaper is a second IIR filter to manupilate the spectal tilt (α) and the relative amplitude of lower harmonic componets (γ).



| OQ | 0.50 | 0.75 | 0.50 | 0.50 | 0.50 |
|---|---|---|---|---|---|
| SQ | 1.80 | 1.80 | 1.50 | 1.80 | 1.80 |
| α | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 |
| γ | 5.0 | 5.0 | 5.0 | 5.0 | 1.0 |

Fig.7. Waveform shapes, spectra and phase characteristics of the proposed model with five conditions.

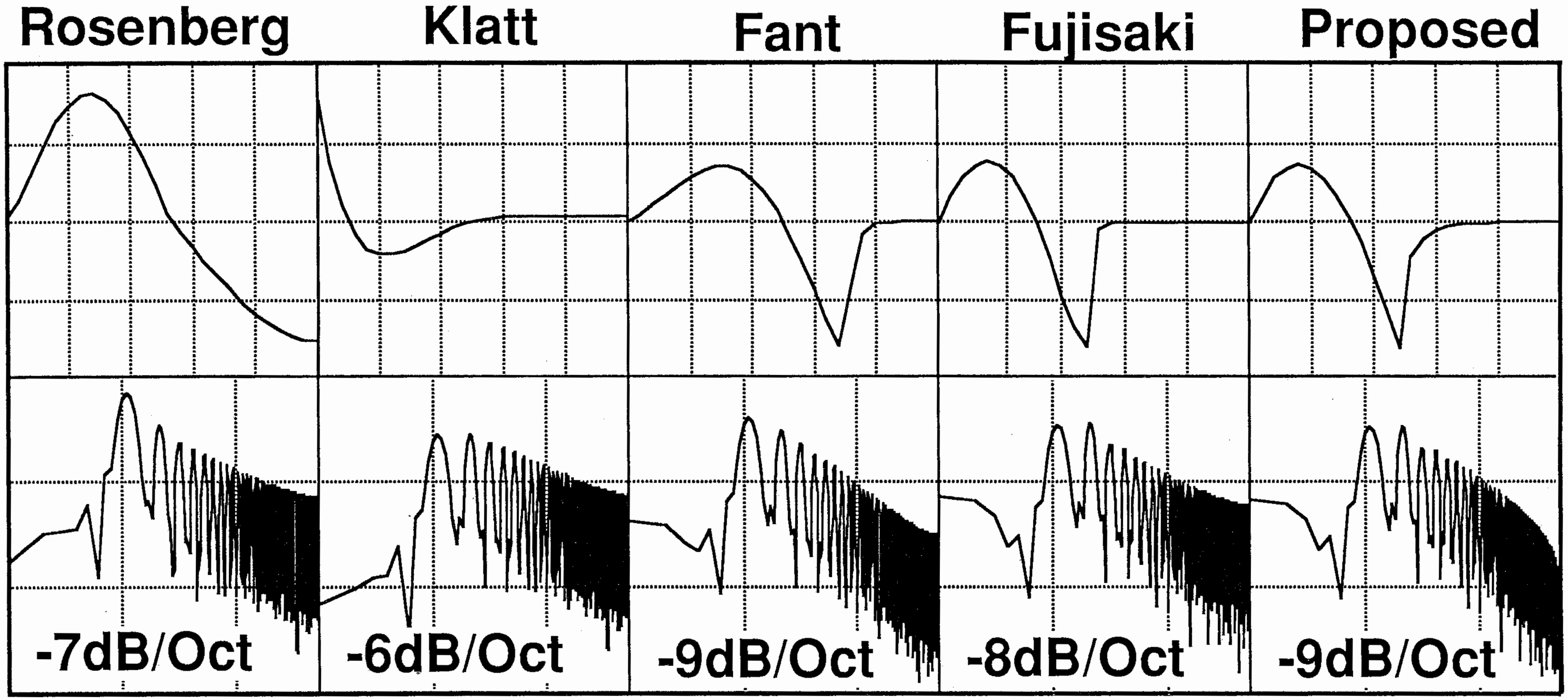


**Fig.8.** Examples of the waveform shapes and its log scale spectra with five models: Rosenberg, Klatt, Fant, Fujisaki and the proposed model.

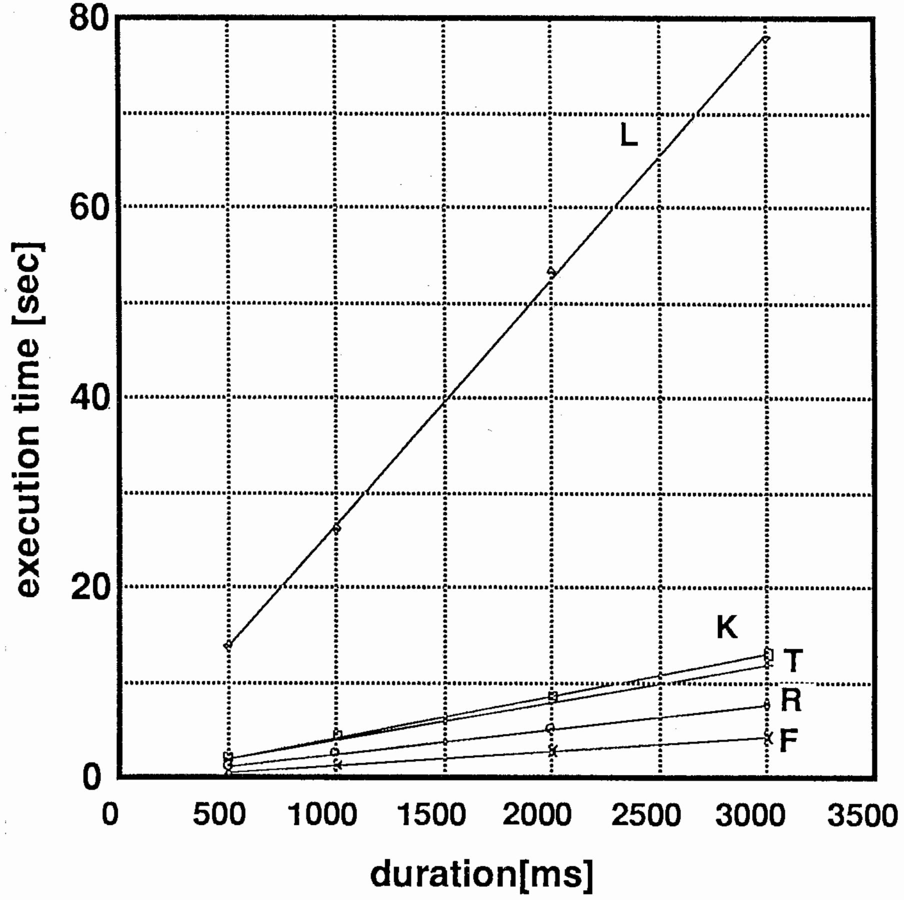


**Fig. 9.** Execution time for the Rosenberg (R), Klatt (K), Fant (L), Fujisaki (F), and the proposed (T) models.

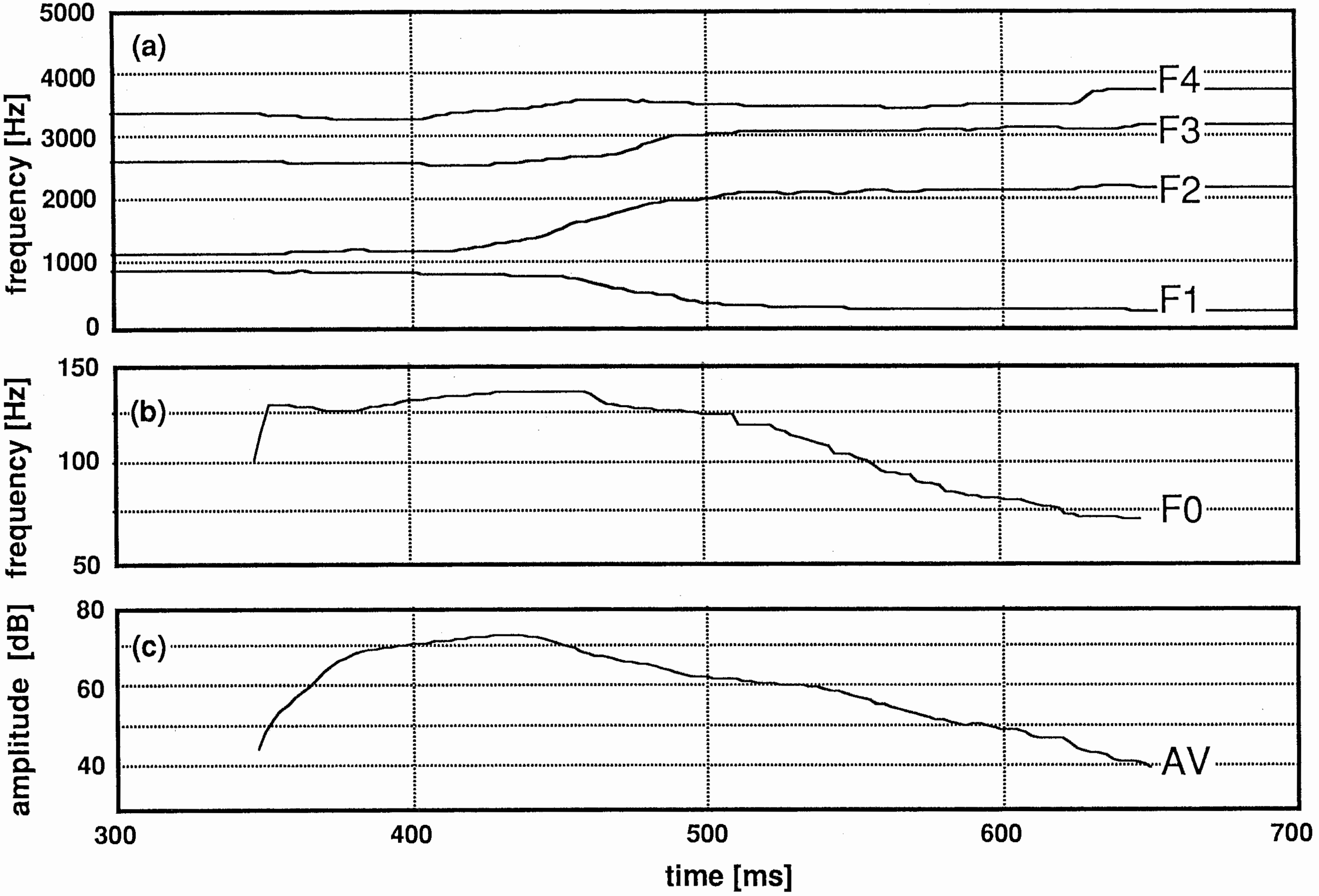


**Fig.10.** Example of parameters to synthesize the sound /ai/ for male voice; (a) formant frequency, (b) fundamental frequency and (c) amplitude of voicing.

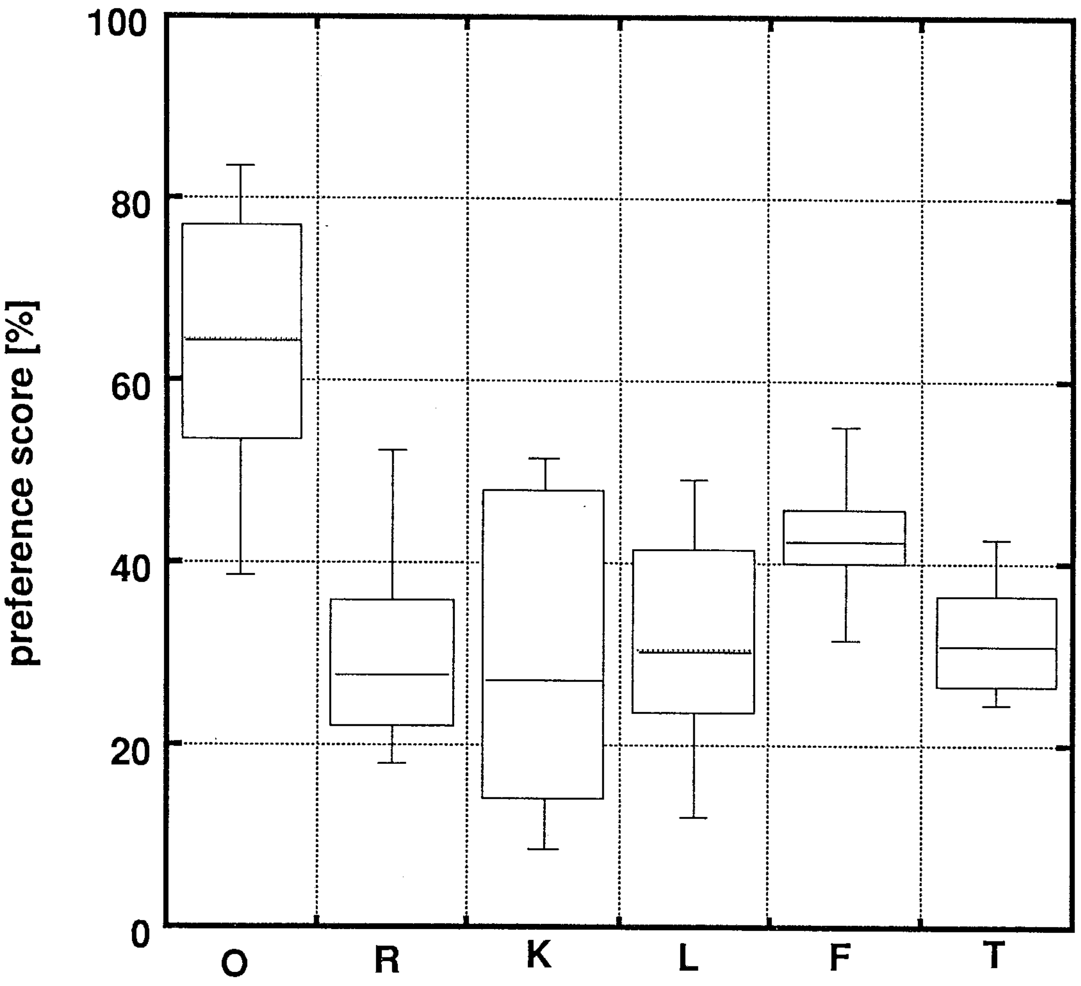


**Fig. 11.** Preference score for male speech stimuli. The preference scores for each speech stimuli category are plotted. **O** represents the natural speech, and **R, K, L, F, T** represent the speech synthesized with the Rosenberg, Klatt, Fant, Fujisaki, and the proposed models, respectively. A box represents the ±25% range of the population score and a thick line in the box indicates the median value of the preference scores .

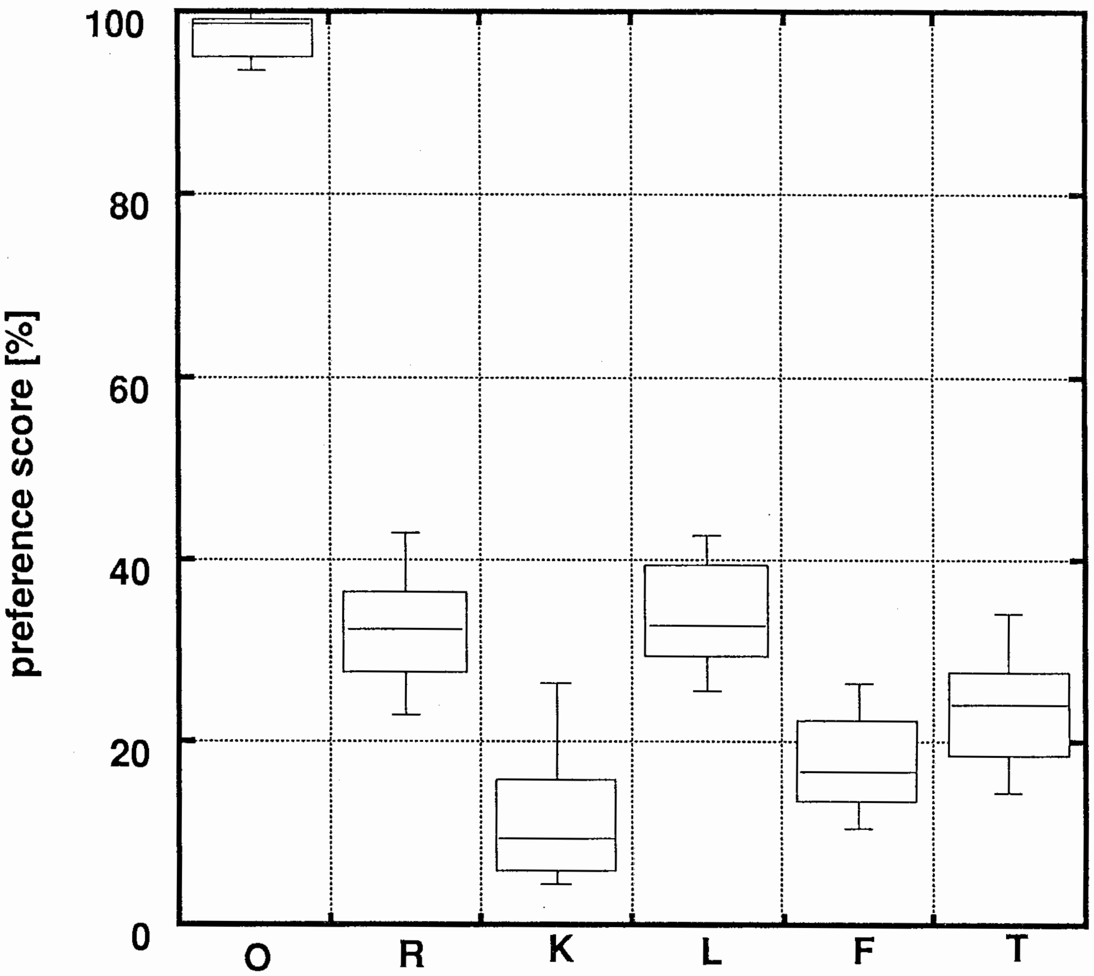


**Fig. 12.** Preference score for female speech stimuli. The preference scores for each speech stimuli category are plotted. **O** represents the natural speech, and **R, K, L, F, T** represent the speech synthesized with the Rosenberg, Klatt, Fant, Fujisaki, and the proposed models, respectively. A box represents the ±25% range of the population score and a thick line in the box indicates the median value of the preference scores .

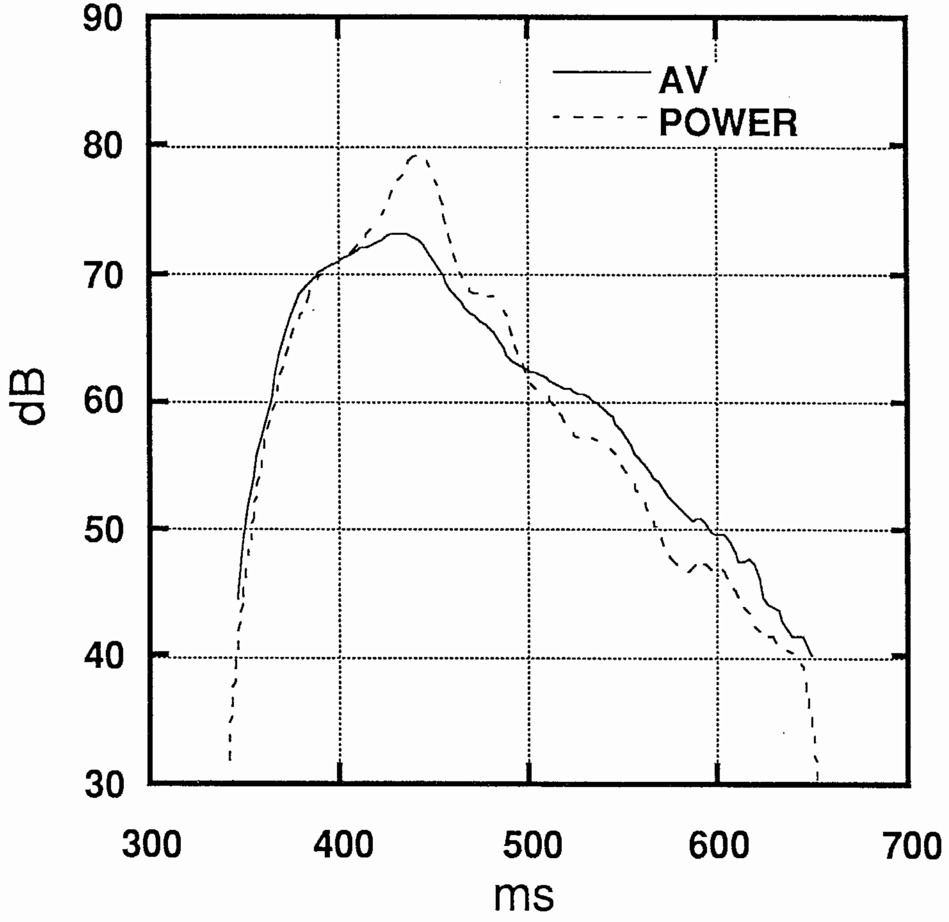


**Fig 13.** Example of the difference between the amplitude of voicing AV and the power of the synthesized speech with the proposed model for male stimuli of /ai/.