Formant Frequency Extraction of Speech by a Very Short-Time Spectral Analysis

Formant Frequency Extraction of Speech by a Very Short-Time Spectral Analysis

Yoshiaki MIYOSHI

Abstract

This paper describes a frequency analysis method by which rapid changes of formant frequencies can be precisely tracked. In this method, the time window is reduced to a shorter duration than the pitch period of voiced sounds. Theoretical and experimental studies are carried on the effects of the location and duration of a time window is shorter than the pitch period of the voiced sound. As the result, it appears that if the time window is set so as not to involve the point of time at which the vocal tract is effectively excited by the glottal waves, we can get a smooth frequency spectrum in which the vocal tract transfer function is well preserved. We can also precisely extract formant frequencies by detecting peak frequencies on the spectrum. A formant frequency extraction algorithm is proposed in which the above -mentioned short-time frequency analysis method is utilized. This algorithm was applied to the extraction of formant transitions in synthesized and natural speech sounds.

1. Introduction

The fast Fourier transform (FFT)[1] is generally used with a time window of 20-30ms. By shifting its location by 10ms we make a frequency analysis of speech sound waves and track the temporal change in formant frequencies[2]. In the case of voiced consonants, however, it is necessary to have a time window shorter than one pitch period of the consonantal wave, considered according to the rate of temporal change of the parameters which are used to describe the movement of articulatory organs.

Mathews *et al.* proposed a pitch synchronous spectrum analysis as a frequency analysis method having a time window of about one pitch period[3]. According to this method, however, Fourier analysis is achieved by regarding a waveform of one pitch period as being periodically repeated, so that the frequency spectrum will consequently take on a harmonic structure, and formant frequencies cannot be extracted precisely. Another method was reported, where a short-time frequency spectrum analysis is made on a waveform of one pitch period[4]. It has been suggested that by reducing the time window to a shorter duration than one pitch period, we can derive a smooth frequency spectrum which preserves the vocal tract transfer

Received September 17th, 1991.

function[5),6)]. But since the location of the window affects the short-time frequency spectrum when the time window is made one pitch period in length, it is necessary to choose the location so that the transfer function the vocal tract is well preserved. To solve this problem, detailed researches were made into the location and duration of the time window.

These researches will be examined in detail in this paper, with reference both the theoretical consideration and the computer simulation of synthetic speech sounds. Based on the results, a formant frequency extraction algorithm is proposed which is useful for voiced plosive sounds with very rapid formant frequency changes.

2. Characteristics of Short-Time Frequency Spectra of Voiced Sounds

In this section, we shall examine how peak frequencies on a short-time frequency spectrum change with shifting locations and varying durations of the time window when we make the time window shorter than one pitch period for voiced sounds. For this purpose, we shall regard the short-time frequency spectrum $F(\omega, t_s, T_a)$ for the sound wave $f(t)$ as a function of its starting point t_s and the duration T_a of the time window $w(t)$ and analyze the characteristics of the spectrum by defining it as follows:

$$F(\omega, t_s, T_a)=\int_{t_s}^{t_s+T_a} w(t-t_s)f(t)e^{-jwt}dt, \tag{1}$$

where $w(t)=0$ and $t<0$ or $t>T_a$.

2.1 The Dependence of Peak Frequencies on the Starting Point and Duration of the Time Window

When the sound wave $f(t)$ is voiced, we can describe it as follows:

$$f(t)=\int_0^{\infty} h(\tau)g(t-\tau)d\tau, \tag{2}$$

where $h(t)$: impulse response of the vocal tract, $g(t)$: glottal source waveform, and the radiation characteristic of sound is involved in the characteristic of the vocal tract.

Let a time window $w(t)$, for simplification, be a rectangular window;

$$w(t)=\begin{cases}1 & 0\leq t\leq T_a,\\ 0 & \text{othewise.}\end{cases} \tag{3}$$

From equations (1)-(3), we can derive:

$$F(\omega, t_s, T_a)=\int_0^{\infty} h(\tau)e^{-j\omega\tau}\int_{t_s-\tau}^{t_s+T_a-\tau} g(t)e^{-j\omega t}dtd\tau. \tag{4}$$

Here, let the glottal waveform $g(t)$ be a impulse sequence where the following relation holds:

$$g(t)=\sum_{n=-\infty}^{\infty}\delta(t-nT_0), \tag{5}$$

where $\delta(t)$: delta function, T_0 : period the glottal wave.

Now, let the duration of the time window T_a be less than one pitch period ($T_a<T_0$). $F(\omega, t_s, T_a)$ is represented as:

$$F(\omega,\ t_s,\ T_a)=\sum_{n=-\infty}^{n_0} e^{-j\omega nT_0}\int_{t_s-nT_0}^{t_s+T_a-nT_0} h(\tau)\,e^{-j\omega\tau}d\tau, \tag{6}$$

where $n_0=[(t_s+T_a)/T_0]$, and [ ] is Gaussian notation. Here, let the starting point (t_s) of the time window be described as:

$$t_s=mT_0+t_{s0} \tag{7}$$

where m : integer, $|t_{s0}|\leq T_0/2$, $t_{s0}+T_a<T_0$. we can now derive:

$$F(\omega,\ t_s,\ T_a)=e^{-j\omega mT_0}\sum_{k=0}^{\infty}\int_{t_{s0}}^{t_{s0}+T_a} h(t+kT_0)\,e^{-j\omega t}dt. \tag{8}$$

In other words: (1) if the glottal source can be regarded as an inpulse train having the period T_0, (2) if the time window is shorter than the period of the glottal wave ($T_a<T_0$), and (3) if the location of the window is set in such a way that it does not involve the next impulse excitation point, under the these three conditions the short-time frequency spectrum will be the product of the phase rotation ($e^{-j\omega mT_0}$) and the sum of the short-period Fourier transforms. Generally speaking, since the amplitude of $h(t)$ is damped when t is increased, the item in the case of $k=0$ is the dominant component of the summation on the right side of the Eq.(8). And when $t<0$, then $h(t)=0$, so that if we set $t_{s0}<0$, only the integral range of this dominant component will effectively decrease. This value $t_{s0}>0$ is therefore taken as a suitable condition.

Now, for simplification, we shall consider the case where the transfer function of the vocal tract has a single pole. In this case, we have

$$h(t)=\begin{cases} e^{-\alpha t}\sin\omega_0 t & t\geq 0,\\ 0 & t<0,\end{cases} \tag{9}$$

where $\alpha>0$. Let t_{s0} be equal to or greater than zero and substitute (9) into (8). We can then obtain the following:

$$F(\omega,\ t_s,\ T_a)=e^{-j\omega mT_0}\left[\frac{e^{-\{\alpha+j(\omega-\omega_0)\}t_{s0}}[1-e^{-\{\alpha+j(\omega-\omega_0)\}T_a}]}{2j\{\alpha+j(\omega-\omega_0)\}\{1-e^{-(\alpha-j\omega_0)T_0}\}}\right.$$
$$\left.-\frac{e^{-\{\alpha+j(\omega+\omega_0)\}t_{s0}}[1-e^{-\{\alpha+j(\omega+\omega_0)\}T_a}]}{2j\{\alpha+j(\omega+\omega_0)\}\{1-e^{-(\alpha+j\omega_0)T_0}\}}\right]. \tag{10}$$

When ω is in the neighborhood of ω_0, then $|\omega+\omega_0|\gg|\omega-\omega_0|$. Therefore, the second item on the right side of the Eq.(10) can be omitted, thus we have:

$$|F(\omega,\ t_s,\ T_a)|\approx\frac{1}{2}\left|\frac{e^{-\{\alpha+j(\omega-\omega_0)\}t_{s0}}[1-e^{-\{\alpha+j(\omega-\omega_0)\}T_a}]}{\{\alpha+j(\omega-\omega_0)\}\{1-e^{-(\alpha-j\omega_0)T_0}\}}\right|$$
$$=\frac{e^{-\alpha t_{s0}}}{2}\sqrt{\frac{1+e^{-2\alpha T_a}-2e^{-\alpha T_a}\cos(\omega-\omega_0)T_a}{\{\alpha^2+(\omega-\omega_0)^2\}\{1+e^{-2\alpha T_0}-2e^{-\alpha T_0}\cos\omega_0 T_0\}}}. \tag{11}$$

Therefore, if ω satisfies $\partial|F(\omega,\ t_s,\ T_a)|/\partial\omega=0$, that is,

$$D(\omega)=\{\alpha^2+(\omega-\omega_0)^2\}T_a e^{-\alpha T_a}\sin(\omega-\omega_0)T_a - (\omega-\omega_0)\{1+e^{-2\alpha T_a}-2e^{-\alpha T_a}\cos(\omega-\omega_0)T_a\} =0, \tag{12}$$

then $|F(\omega, t_s, T_a)|$ takes the extreme value or has a point of inflection. This equation (12) holds regardless of t_{s0} or T_0. From this we find that if the impulse excitation point is not involved in the time window, the locations of peak frequencies on the short-time frequency spectrum depend only on the duration of the time window, not on the starting point of the window or on the period of the impulse train. And furthermore, since $\omega=\omega_0$ always satisfies the Eq.(12), the location of the peak frequency corresponding to the pole frequency does not depend on the duration of the window either. This is a very important characteristic to extract formant frequencies from peak frequencies.

2.2 The Minimal Time Limit of the Time Window

In order to precisely track rapid temporal changes of formant frequencies, as in the case of voiced plosive sounds, it is desirable that the duration of the time window be as short as possible. However, considering the frequency resolution, the duration can not be less than a certain value. We shall be concerned here with the lowest limit of this duration.

When the impulse response of the vocal tract is represented as the summation of two damped sinusoidal waves, it follows that:

$$h(t)=\begin{cases} e^{-\alpha t}\sin\omega_1 t+e^{-\alpha t}\sin\omega_2 t & t\geq 0, \\ 0 & t<0. \end{cases} \tag{13}$$

Referring now to Eq.(11), in the case of a single pole, we have

$$|F(\omega, t_s, T_a)|\approx\frac{1}{2}\left|\frac{1-e^{-\{\alpha+j(\omega-\omega_1)\}T_a}}{\alpha+j(\omega-\omega_1)}+\frac{1-e^{-\{\alpha+j(\omega-\omega_2)\}T_a}}{\alpha+j(\omega-\omega_2)}\right|, \tag{14}$$

where $t_{s0}=0$ and $T_0=\infty$, for simplification. Under the condition that on the frequency spectrum there should be the peaks corresponding to $\omega=\omega_1$ and $\omega=\omega_2$, the lowest limit for the window duration is thought to be the shortest duration of the time window where the frequency spectrum is minimum when $\omega=(\omega_1+\omega_2)/2$. Therefore, the lowest limit will be the smallest value of T_a which satisfies $\partial^2|F(\omega, t_s, T_a)|/\partial\omega^2>0$ when $\omega=(\omega_1+\omega_2)/2$. That is to say:

$$\begin{aligned} &-ST_a(\alpha T_a+2)e^{-\alpha T_a}\Delta\omega^5+2\{(S^2+1)e^{-2\alpha T_a}+C(\alpha^2T_a^2-2)e^{-\alpha T_a}+1\}\Delta\omega^4 \\ &-8\alpha Se^{-\alpha T_a}(2Ce^{-\alpha T_a}+\alpha^2T_a^2+2\alpha T_a-2)\Delta\omega^3 \\ &+16\alpha^2\{(3C^2-1)e^{-2\alpha T_a}+C(\alpha^2T_a^2-4)e^{-\alpha T_a}+2\}\Delta\omega^2 \\ &+16\alpha^3Se^{-\alpha T_a}(4Ce^{-\alpha T_a}-\alpha^2T_a^2-2\alpha T_a-4)\Delta\omega \\ &-32\alpha^4\{C^2e^{-2\alpha T_a}-C(\alpha^2T_a^2+2)e^{-\alpha T_a}+1\}>0, \end{aligned} \tag{15}$$

where $\Delta\omega=|\omega_1-\omega_2|$, $S=\sin(\Delta\omega T_a/2)$, $C=\cos(\Delta\omega T_a/2)$.

From Eq.(15), even if $T_a=\infty$, there exist no peaks corresponding to $\omega=\omega_1$ and $\omega=\omega_2$ respectively on the frequency spectrum unless $\Delta\omega > 0.972\alpha$. Figure 1 shows the relation between $\Delta\omega$ and the smallest T_a value which satisfies the Eq. (15).

From Fig. 1, we find that as $\Delta\omega$ approaches 0.972α (for instance, $\Delta\omega/2\pi=145.8$Hz in the case of $\alpha=300\pi$), T_a rapidly increases, while in the case of $\Delta\omega/2\pi\geq 500$Hz, the following relation obtains between $\Delta\omega$ and T_a:

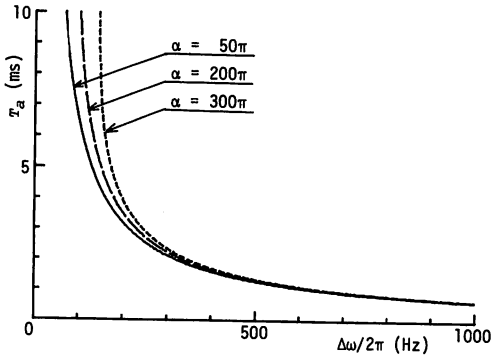

Fig. 1 The relation between $\Delta\omega$ and the smallest T_a value which satisfies Eq.(15).

$$\Delta\omega T_a=1.3\pi, \tag{16}$$

regardless of a value in the range $50\pi\leq\alpha\leq 300\pi$.

3. Examination of the Characteristic of the Short-Time Frequency Spectrum of a Voiced Sound by the Use of a Synthesized Vowel

In Chapter 2, we considered the case where the glottal source is an impulse train and the transfer function of the vocal tract has one or two poles. In the case of natural voices, however, the glottal source is not an impulse train and the transfer function of the vocal tract has several poles. Since it is difficult to consider such cases theoretically, in this section we shall be concerned with the simulation of synthetic vowels.

A vowel /a/ was synthesized by the use of terminal-analogue type digital simulation. The conditions were as follows. The source was a glottal wave having a period T_0 of 8ms, as is shown in Fig. 2(a)[7]. The formant frequency $F_1=700$Hz, $F_2=1,300$Hz, $F_3=2,500$Hz and the formant bandwidths $B_i=50\{1+F_i^2/(6\times10^6)\}$Hz. A high boost of 12dB/oct was used for higher-pole correction and the radiation characteristic. Employing this synthesized /a/, we shall examine the effects of the starting point and the duration of the time window, and the pitch period of the /a/.

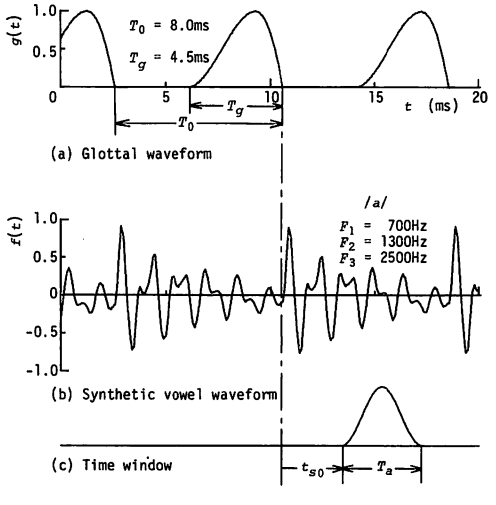

Fig. 2 Glottal waveform and synthetic vowel waveform produced by a terminal analog synthesizer and definitions of the parameters t_{s0} and T_a.

3.1 Peak Frequencies and the starting point of the time window

To investigate the effect of the starting point of the time window, we chose for the fiducial point ($t_{s0}=0$) the closing point of the glottal

wave. That is, the effective excitation point during the pitch period of the glottal waveform. Figure 3 shows that how the peak frequency's ($\boldsymbol{F}^p=\{F_1^p, F_2^p, F_3^p, \cdots\}$), frequencies corresponding to maximal values on a short-time frequency spectrum derived by FFT, are related to the starting point t_{s0} when the duration of the time window is fixed at 4ms ($T_a=T_0/2$). Figure 4 shows examples of the short-time frequency spectra which are the results of FFT (sampling frequency f_s : 10kHz, sampling point total : 512) on the waveform cut out by the Hanning window with the starting point t_s and the window duration T_a. Since $T_a f_s<512$ here, zeros are appended to the windowed segment of speech samples[8].

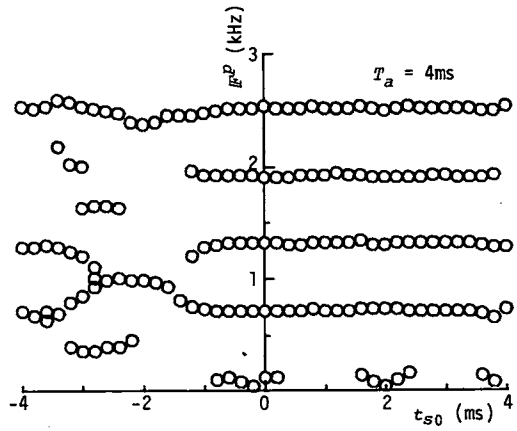

Fig. 3 Dependence of the peak frequency's $\boldsymbol{F}^p$ on the starting point t_{s0} of the time window.

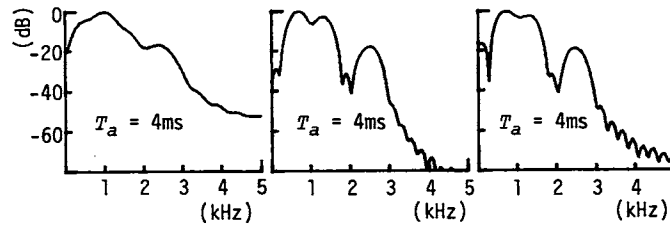

Fig. 4 Examples of short-time frequency spectra.

Figure 3 shows that there is hardly any effect of the starting point of the time window on the location of the peak frequency on the frequency ordinate, except the ones below about 200Hz in the range $-0.6\text{ms}\leq t_{s0}\leq 4\text{ms}$. That is, the range where the glottis closing point is not effectively involved in the time window since a Hanning window, such as is shown in Fig. 2(c), is used here. From Fig. 4, we find that if t_{s0} is equal to or greater than zero, we are able to derive a smooth frequency spectrum envelope which preserves the transfer function of the vocal tract. The error E_i of formant frequency extraction by the peak picking on this frequency spectrum is defined as:

Table 1 The error E_i of formant frequency extraction (T_a=4 ms).

t_{s0}(ms)	E_1(%)	E_2(%)	E_3(%)
0.0	0.4	0.7	0.8
2.0	3.2	−0.8	−0.8

$$E_i=\frac{F_{ci}^p-F_i}{F_i}, \tag{17}$$

where F_{ci}^p is the peak frequency corresponding to F_i.

Table 1 shows E_i in the cases t_{s0}=0ms and 2ms. Table 1 indicates that the simple peak picking method is capable of precise formant frequency extraction because we can get a smooth frequency spectrum in which the vocal tract transfer function is well preserved.

3.2 Peak Frequencies and the Duration of the Time Window

Figure 5 shows the relationship between the time window duration T_a and the peak frequency F^p in the case of the synthetic /a/ shown in Fig. 2, and when t_{s0}=0ms. Figure 6 exemplifies the short-time frequency spectra. From Fig. 5, it can be said that the peak frequencies which correspond to the formant frequencies F_i hardly change their locations on the frequency ordinate in the range $T_a \geq 3.4$ms, once they appear as peaks in the process of T_a increasing. In other words, the peak frequencies which do not change their locations on the frequency ordinate are those corresponing to the formant frequencies. This is an important feature of the automatic extraction of formant frequencies. Table 2 shows E_i when T_a=3.0ms, 3.5ms, 4.0ms, and 5.0ms.

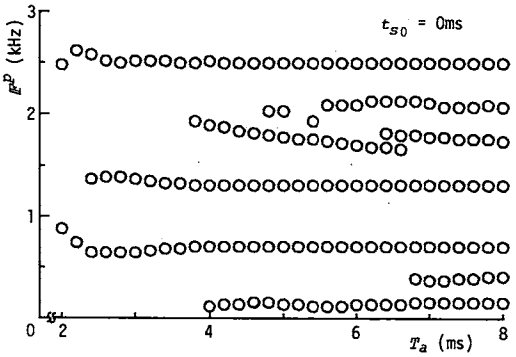

Fig. 5 Dependence of the peak frequency's F^p on the time window duration T_a.

In Fig. 6, we find that when T_a is small, peaks corresponding to F_1 and F_2 do not separate but combine to form one big peak, and they form two peaks only when $T_a \geq 2.3$ms. When the first and the second formant frequencies and band widths are respectively 700Hz, 1,300Hz, 54.1Hz, and 64.1Hz, $\Delta\omega$ and α in Eq.(15) are thought to be 600Hz and 64.1π respectively, then the shortest duration of the time window for the peaks corresponding to F_1 and F_2 to exist is about 1.1ms, according to Eq.(15). Theses frequency spectra in Fig. 6 are obtained by using the Hanning window as a time window[9]. The effective duration T_{ae} of a Hanning window having the window duration T_a is supposed to be $0.67T_a$, that is, $T_{ae}=0.67T_a$[10]. Thus, a Hanning window having the duration of 2.3ms effectively corresponds to the duration of 1.54ms. Furthermore, the first and the second formants of synthetic sounds are different in their amplitude. Taking these factors into consideration, the fact that, in the range $T_a \geq 2.3$ms, there exist peaks corresponding to $F_1$and F_2 is almost in accord with the result of the Section 2.2.

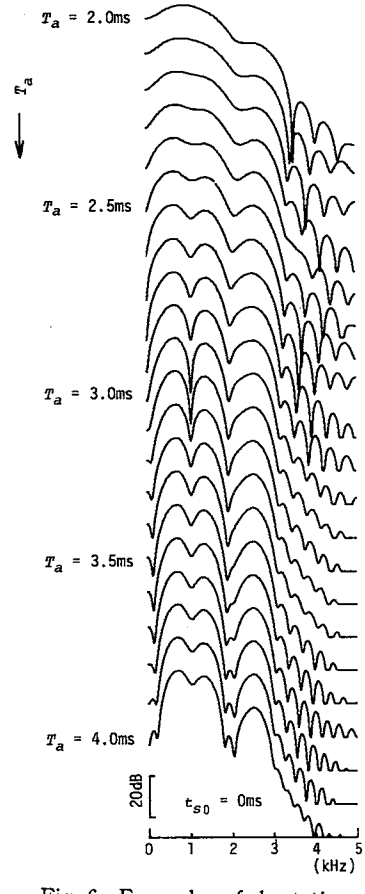

Fig. 6 Examples of short-time frequency spectra.

3.3 Peak Frequencies and the Period of a Voiced Sound

Figure 7 shows the effect of the period T_0 on the peak frequency F^p, in the digital simulation of which the opening coefficient (T_g/T_0, see Fig. 2(a)) is fixed and the period T_0 is changed. And the sampling condition is t_{s0}=0ms and T_a=4ms.

As a result of this we can conclude that if $T_0 \geq 4.6$ms, the period T_0 has hardly any effect on the location of peak frequencies along the frequency ordinate except that a peak appears in the neighborhood of 150Hz half way along the abscissa.

Table 2 The error E_i of formant frequency extraction ($t_{s0}=0$ ms).

T_a(ms)	E_1(%)	E_2(%)	E_3(%)
3.0	−7.9	5.2	0.8
3.5	−2.3	2.2	0.8
4.0	0.4	0.7	0.8
5.0	0.4	0.7	0.0

Based on the above-mentioned discussions, we find that there are the same kind of characteristics in the short-time frequency spectrum of the synthetic vowel as the result of the analysis in the previous Chapter 2. Also ; (1) if the location of the window does not involve the point of time when the vocal tract is effectively excited by the glottal source (for instance, at the closing point of the glottis in the case of the glottal waveform shown in Fig. 2(a)), and (2) if the time window is made shorter in duration than the pitch period, we can then derive a smooth envelope of the frequency spectrum which preserves the transfer function of the vocal tract. It is also possible to precisely extract formant frequencies from the peak frequencies of this envelope. This method will be called a FRAPS (FRActional Period Spectral analysis) method, hereafter.

Fig. 7 Dependence of the peak frequency's F^p on the pitch period T_0.

4. Algorithm for Formant Frequency Extraction by the FRAPS Method

As the result of many computer simulations, it is shown that peak frequencies corresponding to formant frequencies on the short-time frequency spectrum obtained by the use of the FRAPS method have the following important features.

(A) Peak frequencies corresponding to formant frequencies are not dependent on the duration of the time window.

(B) Peak frequencies corresponding to formant frequencies have a large basewidth B_i^p defined by the Eq.(18) than the other peak frequencies.

$$B_i^p = F_i^v - F_{i-1}^v,\ i=1, 2, 3,\cdots, \tag{18}$$

where F_i^v : frequency which is minimal on the short-time frequency spectrum (but $F_0^v=0$). That is, B_i^p is a parameter which represents the stretch of the base of a frequency spectrum envelope having F_i^p as maximal values(cf. Fig. 8).

The use of these feature leads to an algorithm for formant frequency extraction as is shown in Fig. 8. The procedures are as follows:

(a) Set the starting point of the time window

It is not necessary to take the effective excitation point of the glottal wave precisely as the starting point of the time window (cf. Fig. 3 and explanation in the section 3.1), but it is desirable to have the starting point its neighborhood as far as possible. For practical purposes, we will choose the point where the amplitude is 20% of the greatest amplitude immediately before the greatest amplitude during the pitch period of the waveform.

(b) Set the initial values

The initial value of the time window duration T_a is set at T_{a0} and n (a parameter that is an integer, to be mentioned later) is made to be 1.

(c) Find the peak frequency $\boldsymbol{F}^p$

The next procedure is to find the peak frequency $\boldsymbol{F}^p$ on the short-time frequency spectrum envelope, the time window of which is T_a in duration.

(d) Find $\boldsymbol{F}^0$

The next is to find $\boldsymbol{F}^0$. This is composed of F_i^p where Δf_i, defined by the Eq.(19), is less than a threshold value f_θ.

$$\Delta f_i = \min_j \{ | F_j^p(T_a - \Delta T_a) - F_i^p(T_a) | \}, \quad (19)$$

where $i=1, 2, 3, \cdots$, $\dim(F^p(T_a))$, $j=1, 2, 3, \cdots$, $\dim(F^p(T_a-\Delta T_a))$, and $F_i^p(*)$ represents peak frequency on the short-time frequency spectrum whose time window is $*$ in duration. This procedure is, in other words, to pick out formant frequency candidates from the components of $\boldsymbol{F}^p$, using the above-mentioned feature (A).

(e) In order to extract F_1, F_2 and F_3, increase T_a by ΔT_a and return to (c), if the order of $\boldsymbol{F}^0$ ($\dim(\boldsymbol{F}^0)$) is less than 3.

(f) Find the candidate formant frequency $\boldsymbol{F}^c$

This is to pick out from the first to the third in decreasing order of basewidth B_i^0 from $\{F_1^0, F_2^0, F_3^0, \cdots\}$. We then define

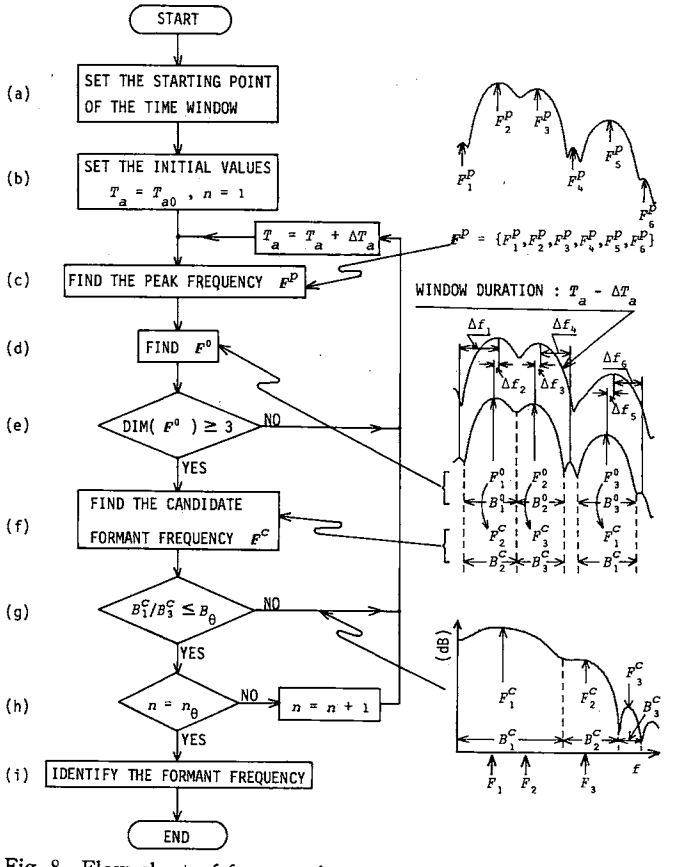

Fig. 8 Flow chart of formant frequency extraction by the FRAPS method.

them as $\boldsymbol{F}^c=\{F_1^c, F_2^c, F_3^c\}$ and their basewidths as $\boldsymbol{B}^c=\{B_1^c, B_2^c, B_3^c\}$. Therefore, the components of $\boldsymbol{B}^c$ are in decreasing order, that is, $B_1^c \geq B_2^c \geq B_3^c$. In this procedure, three formant frequency candidates are selected by making use of the above-mentioned feature (B).

(g) If B_1^c/B_3^c is greater than a threshold value B_θ, the procedure is to increase T_a by ΔT_a and to return to (c), assuming that the peak frequencies corresponding to $F_i (i=1, 2, 3)$ have not appeared yet. That is, either F_1 and F_2 or F_2 and F_3 form one peak and cannot be separated.

(h) If the number n of the times when both the conditions (e) and (g) are satisfied is not equal to a threshold n_θ, add one to n and return to (c), increasing T_a by ΔT_a. If we set n_θ be equal to or greater than 2, the procedure (c)-(g) will be repeated to increase the reliability of formant frequency candidates $\boldsymbol{F}^c$.

(i) Identify the formant frequency

The candidate formant frequencies $\{F_1^c, F_2^c, F_3^c\}$ which have satisfied all the above procedures will be rearranged in increasing order of frequency, and be identified as extracted formant frequencies $\{F_1, F_2, F_3\}$. Here ends this algorithm.

5. Tracking of Formant Frequencies of a Voiced Plosive by the FRAPS Method

The FRAPS method uses a time window which is much shorter in duration than the usual ones (namely, 20-30ms)[11]. This feature is especially useful for tracking rapid formant frequency changes during the transition sections of speech sounds. As an example, it will show to track the formant frequencies of a voiced plosive.

5.1 Syntesized Voiced Plosive

By the same digital simulation as the previous Chapter 3, a voiced plosive was synthesized in which the formant frequencies change temporally as shown by the solid lines in Fig. 9. Formant frequencies extracted by means of the FRAPS method are shown by small circles in the same figure. The triangles in it, shown for comparison, illustrate formant frequencies obtained by peak picking from the frequency spectrum envelope which is calculated by the cepstrum method[12], in which a Hanning window with a duration of 30ms and the center of the window is set at the center of that of the FRAPS method. This cepstrum method will be called a SE (Spectral Envelope) method, in which for low-pass filtering on the cepstrum using a weighting function $w_e(t)$, shown by the Eq.(20), where $\tau_1=\Delta\tau=1.5$ms.

$$w_e(t)=\begin{cases} 1 & t<\tau_1, \\ 0.5\{1+\cos(\pi(t-\tau_1)/\Delta\tau\} & \tau_1 \leq t<\tau_1+\Delta\tau, \\ 0 & t\geq\tau_1+\Delta\tau. \end{cases} \tag{20}$$

From Fig. 9, the followings are found. (1) The FRAPS method is more precise in extractive formant frequencies from the transition part of speech sounds than the SE method.

Table 3 The error E_i of formant frequency extraction at on-set frequencies of formants.

VOICED PLOSIVE	FRAPS method			SE method		
	E_1(%)	E_2(%)	E_3(%)	E_1(%)	E_2(%)	E_3(%)
/ba/	2.2	−0.7	0.2	36.2	13.8	3.8
/da/	2.2	0.7	−0.2	30.5	− 2.4	−4.9
/ga/	2.2	0.3	−0.7	30.5	−10.4	3.8

(2) In the extraction of on-set frequencies of formants, which are essential for the recognition of voiced plosives[13], the extraction error of the SE method has been reduced to the order of ten by the use of the FRAPS method (see Table 3).

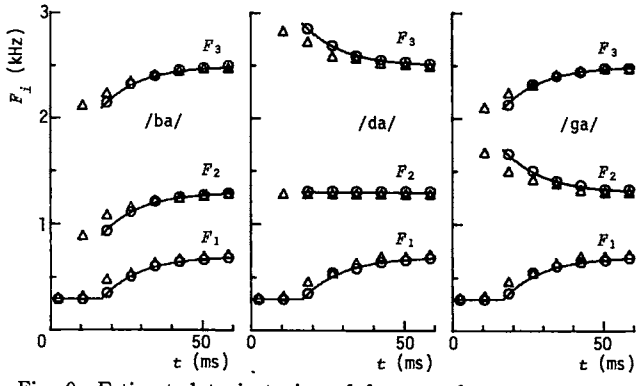

Fig. 9 Estimated trajectories of formant frequencies of synthetic voiced plosives. The solid lines : formant trajectories of synthetic voiced plosive, the circles : estimated formant frequencies by the FRAPS method, the triangles : estimated formant frequencies by peak picking method on the cepstrally-smoothed spectrum.

5.2 A Natural Voiced Plosive

Figure 10 shows a comparison between the two method, regarding their tracking of the formant frequencies of /ga/ uttered by a male. From this result, we can say the follows. (1) The first formant frequencies extracted by the FRAPS method are temporally continuous, while there is an unnatural discontinuity in those extracted by the SE method. (2) The SE method has shown a peak frequency in the neighborhood of 1,600Hz, which may possibly be identified as the third formant, but this must be unnatural, judging from the temporal continuity of formant frequencies. On the contrary, the FRAPS method shows no such peaks. This means that the SE method produces a number of spurious peaks in the transition part of analyzed speech sound, which the FRAPS method does not.

It can be concluded from this result that the FRAPS method is superior to the SE method in tracking the formant frequencies of the transition part of voiced sounds.

6. Conclusions

For a method of frequency analysis of speech sound, the FRAPS method was proposed which can precisely track the rapid temporal change of formant frequencies by having the time window duration shorter than the pitch period of voiced sound. The effects of the starting point and duration of the time window on short-time frequency spectra were examined. It was found that (1) if the duration of the time window is shorter than the pitch period of the voiced

sound and (2) if the location of the window is set so as not to involve the effective excitation point during the pitch period of the glottal source, we can obtain a smooth frequency spectrum envelope which preserves the transfer function of the vocal tract. From peak frequencies of the envelope we can then extract formant frequencies with precision.

The use of this method was made to form an agorithm for extracting formant frequencies. This algorithm was applied to the analysis of synthesized and natural voiced plosives and proved to be useful.

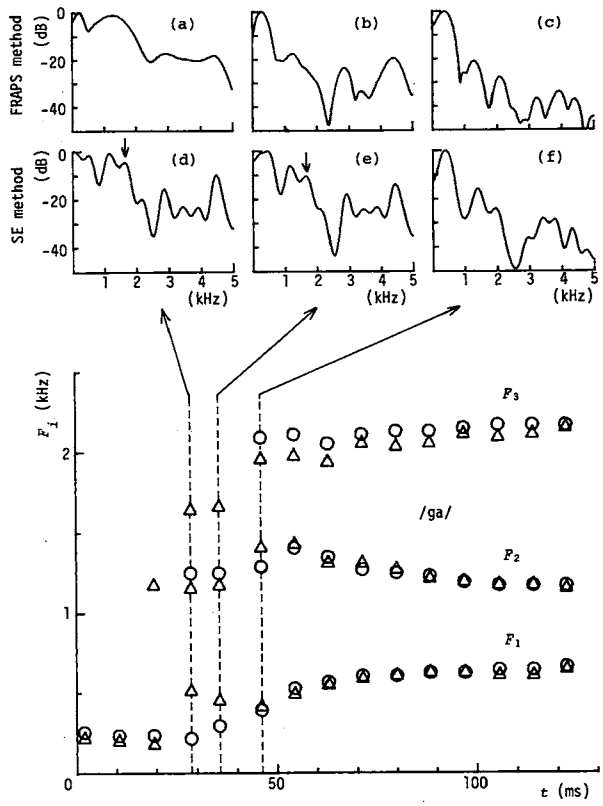

Fig. 10 Estimated trajectories of formant frequencies of natural voiced plosive /ga/ and examples of short-time frequency spectra. The circles : estimated formant frequencies by the FRAPS method, the triangles : estimated formant frequencies by the SE method.

References

(1) W. T. Cochram : "What is the fast Fourier transform ?", IEEE Trans, **AU-15**, 6, p. 45(1967).

(2) R. W. Schafer and L. R. Rabiner : "System for automatic formant analysis of voiced speech", J. Acoust. Soc. Amer. **47**, 2, p. 634(1970).

(3) M. V. Mathews, J. E. Miller and E. E. David : "Pitch synchronous analysis of voiced sounds", J. Acoust. Soc. Amer. **33**, 2, p. 179(1961).

(4) W. J. Hess : "A pitch-synchronous digital feature extraction system for phonemic recognition of speech", IEEE Trans. Acoust., Speech, Signal Processing, **ASSP-24**, 1, p. 14(1976).

(5) R. W. Schafer and L. R. Rabiner : "Digital representations of speech signals", Proc. IEEE, **63**, 4, p. 662(1975).

(6) J. B. Allen and L. R. Rabiner : "A unified approach to short-time fourier analysis and synthesis", Proc. IEEE, **65**, 2, p. 583(1977).

(7) A. E. Rosenberg : "Effect of glottal pulse shape on the quality of natural vowels", J. Acoust. Soc. Amer. **49**, 2, p. 583(1971).

(8) K. Okada : "A remak on the direct Fourier transform method for computing power spectrum", Trans. IECE Japan, **J66-A**, 3, p. 288(1983).

(9) K. Toraichi, M. Kamada, S. Itahashi and R. Mori : "A series of window functions obtained by convolution integrals of rectangle window with itself", Trans. IEICE Japan, **J70-A**, 3, p. 481(1987).

(10) R. B. Blackman and J. W. Tukey : "The measurement of power spectra", p. 20, Dover, New York(1959).

(11) T. Nakajima and T. Suzuki : "Power spectrum envelope (PSE) speech analysis-synthesis system", J. Acoust. Soc. Jpn. **44**, 11, p. 824(1988).

(12) A. V. Oppenheim : "Speech analysis-synthesis based homomorphic filtering", J. Acoust. Soc. Amer. **45**, 2, p. 458(1969).

(13) H. Nakashima and O. Kakusho : "Effect of formant transition time on the perception of the voiced stop consonants", J. Acoust. Soc. Jpn. **39**, 1, p. 52(1983).