Ann.Bull.RILP(1988)No.22, 143-152

A STUDY ON DYNAMIC CHARACTERISTICS OF SPEECH SIGNAL

Shuzo Saito* and Kazuhiro Tamaribuchi*

Introduction

Speech signal conveys linguistic or phonemic information of talker's message in its time-varying waveform. From the view point of speech production model, temporal variation of control signal in speech organ is rather slow comparing to the quick variation of vibrating amplitude in speech waveform. As acoustic feature parameters of speech spectrum correspond well to the control signal in speech organ, spectral feature parameters of speech are often used for analysis of speech information bearing element. Parametric analysis based on the all-pole model of speech spectrum is useful for the efficient speech coding, but it seems that there is an inconsistency between stationary stochastic process and a actual waveform of speech signal. In this paper, several studies on temporal variation of speech spectrum are reviewd, then a new method for speech analysis is described accompaning a few examples.

Speech Analysis Based on All-Pole Model of Speech Spectrum

Speech analysis based on the maximum likelihood spectrum estimation method[1] is valid for the extraction of feature parameters of speech signal and is available for the narrow-band speech coding system being a good transmission performance. To extend its application fields, it is needed to improve its output speech quality by use of the speech analysis procedure being well-matched to the dynamic characteristics of speech signal. For such a purpose, effects of windowing width for speech analysis and sampling frequency of parameters for speech synthesis on the decorded speech quality has been measured using the PARCOR speech analysis/synthesis system.[2]

In this study, widths of the Hamming windows used for spectrum estimation were set at 7.5, 10.0, 15.0, 20.0 and 30.0 milliseconds and frequency spectrum for each windowed signal was calculated repeatedly on the shifted window of 2.5 milliseconds of speech signal. The PARCOR coefficients of 12-orders were derived on each windowed signal and used as the control signals of frequency spectrum envelope in the PARCOR synthesizer, whereas the sampling periods of the control signal was not fixed at 2.5 milliseconds, but set at three kinds of values, that is, 2.5, 5.0 and 10.0 milliseconds. It should be noted that the Hamming windowed signal of 30.0 milliseconds width was always

* Kogakuin University

used for analysis of the fundamental frequency. Using these two kinds of control signals, that is, glottal excitation signal and vocal tract filtering characteristics, various kinds of speech-like sounds were synthesized and used for listening test on phonemic information.

Results of listening tests are shown in Fig. 1. Abscissa of Fig.1(a) is the width of Hamming window for speech analysis. Ordinate is the synthesized speech quality represented in the equivalent articulation loss (dB). Parameter in this figure is the sampling period of control signal for frequency spectrum envelope in the PARCOR synthesizer. It seems that there are the optimum widths of Hamming windows in every sampling frequencies of the control signals. Fig.1(b) shows the estimated optimum window widths from results of Fig.1(a). It is found that the optimum window width decreases in accordance with the decrease of the sampling period of the control signal. Results of the listening tests were examined in each phoneme and were found that the unvoiced stop consonants were deteriorated mainly from the reduction of the sampling frequency of the control signal of vocal tract filter. Anexample is shown in Fig.1(c), where the eqivalent articulation losses for unvoiced stop consonants are compared with those of whole sounds in the test condition of sampling period of 5.0 milliseconds.

Similar listening test were also reported in Reference 3), although the width of the Hamming window for fundamental frequency analysis was not fixed, but varied in accordance with that used for frequency spectrum envelope analysis. From these results, it may be concluded that the shorter sampling period brought into the better speech quality in the PARCOR synthesizer, although the shorter sampling period results in the worse efficiency on transmission capacity.

Another approach to investigate the prediction error in the maximum likelihood spectrum estimation method was performed and reported in Reference 4), in which the short analysis interval of about one-third long of the pitch period was used forlinear predictive analysis of speech signal. As the free vibration in the vocal cavity appears only at the closure interval of glottis, the short analysis interval described above was used for precise estimation of vocal tract filter characteristics. Fig. 2 shows the root mean square value of prediction errors for speech signal /a/ uttered by a male talker. Abscissa is the position of short analysis interval of 3.125 milliseconds of speech signal. Ordinate is the root mean square value of the prediction residuals and parameter of the figure is the order of linear prediction analysis. It is seen that there are several error minimum positions at around 2 and 11 milliseconds in the cases of p = 8 and 12, although error value does not reduced to zero. It seems that the error minimum positions correspond to the free vibration interval of vocal cavity. It is possible to estimate the optimum frequency spectrum using these error minimum positions, but it needs a lot of calculation procedures. In the succeeding section, another analysis method of speech signal is

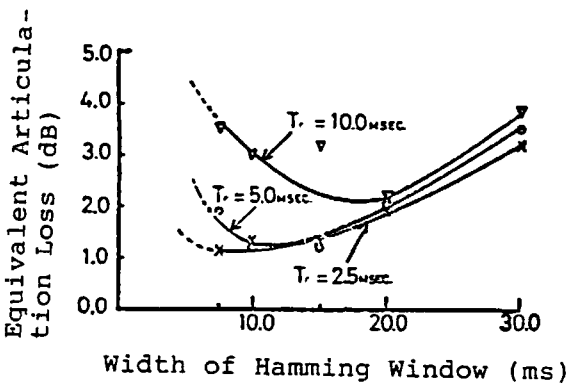

Fig. 1 (a)
Effect of window width on equivalent articulation loss (dB)

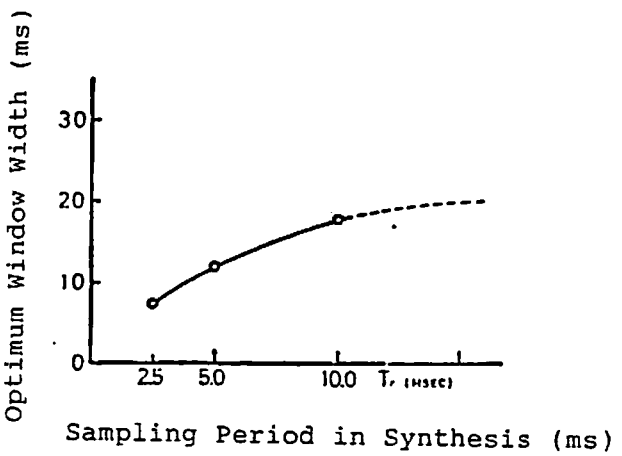

Fig. 1 (b)
Relation between optimum window width and sampling period in synthesis

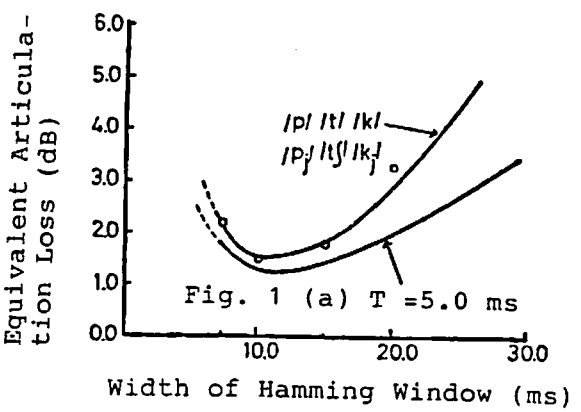

Fig. 1 (c)
Effect of window width on equivalent articulation loss of unvoiced stops

explained, in which a simple algebraic calculation is used for acoustic signal analysis instead of the statistical methods.

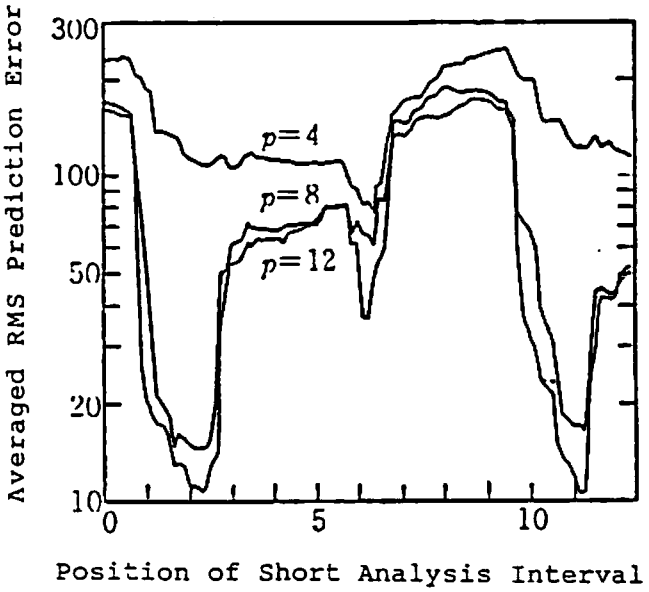

Fig. 2 RMS Prediction Errors of Vowel /a/ Uttered by a Male Speaker

An Analysis Method for Acoustic Signal Represented by Composite Cosine Waves

Acoustic signal may be regarded as the combination of several sinusoidal waves, in which a sinusoidal wave is fully represented by its frequency, amplitude and initial phase. An analysis method to estimate frequency components of input signal was proposed several years ago by S. Sagayama et al[5], in which any input signal could be analyzed into its frequency components of definite frequencies and amplitudes, excepting their initial phases, by the use of autocorrelation function represented in a discrete form. It seems that the use of autocorrelation function in a discrete representation may bring about some errors. An arbitrary wave, however, composed of finite number of components m, may be described uniquely using independent variables of 3m from the point of view of the degree of freedom. Here, a method to determine the frequency components of an arbitray acoustic signal is described by use of 3m discrete sequential data of the signal, provided that input signal is composed of m frequency components.[6]

Function f(r) is defined as a signal composed of m kinds of cosine waves at time rT, where T is the discrete sampling interval.

$$f(r)=\sum_{n=0}^{m-1} a_n \cos(r\omega_n T+\phi_n)$$

$$(r=0,\ 1,\ \cdots,\ 3m) \qquad (1)$$

where a_n, ω_n and ϕ_n are an amplitude, an angular frequency and an initial phase of the n-th frequency component, respectively.

To determine ω_n, function $F_s(r-s)$ is defined as follows,

$$F_s(r)\equiv\frac{1}{2}(1-\delta_{s0})\{F_{s-1}(r)+F_{s-1}(r+2)\}+\delta_{s0}f(r)$$

$$(s=0,\ 1,\ \cdots,\ [(3m+1)/2]-1)$$

$$(r=0,\ 1,\ 2,\ \cdots,\ 3m-2s-1) \qquad (2)$$

where δ_{s0} is Kronecker's delta. $F_s(r-s)$ could be calculated from the 3m sequentially sampled data $f(r)$ $(r = 0,1,\ldots,3m-1)$ being in the ranges of $0 \leq s \leq [(3m+1)/2]-1$ and $0 \leq r \leq 3m-s-1$. The symbol "[]" is the Gauss's symbol. Refering to the equations (1) and (2), the following eauation (3) is derived.

$$F_s(r)=\sum_{n=0}^{m-1} a_n \cos(r\omega_n T+\phi_n)\cos^s(\omega_n T) \qquad (3)$$

As $F_s(r-s)$ is the function of $\cos^s(\omega_n T)$, the following equations are the m-order algebraic equations of $\cos(\omega_n T)$.

$$F_m(r-m)+\sum_{s=0}^{m-1} b_s F_s(r-s)=0$$

$$(r=m,\ m+1,\ \cdots,\ 2m-1) \qquad (4)$$

where b_s $(s = 0,1,\ldots,m-1)$ are constants defined by the equation (4). Then the roots X_n $(n = 0,1,\ldots,m-1)$ of the following m-order equation

$$x^m+\sum_{s=1}^{m-1} b_s x^s=0 \qquad (5)$$

is equal to $\cos(\omega_n T)$, that is,

$$x_n=\cos(\omega_n T) \qquad (n=0,\ 1,\ \cdots,\ m-1) \qquad (6)$$

As $F_s(r-s)$ and $\cos(\omega_n T)$ are known as shown in equations (2) and (6), respectively, $a_n \cos(r\omega_n T + \phi_n)$ ($n = 0,1,\ldots,m-1$; $r = m-1,m,\ldots,2m$)may be solved after equation (3). Let $A_n(r)$ be the solution of $a_n \cos(r\omega_n + \phi_n)$, the phase ϕ_n is determined by use of $A_n(r_1)$ and $A_n(r_2)$ ($r_1 \neq r_2$; $m-1 \leq r_1, r_2 \leq 2m$) as follows,

$$\phi_n = \tan^{-1}\left(\frac{C_n(r_1, r_2)}{S_n(r_1, r_2)}\right) \tag{7}$$

where

$$C_n(r_1, r_2) = A_n(r_1)\cos(r_2\omega_n T) - A_n(r_2)\cos(r_1\omega_n T) \tag{8}$$

$$S_n(r_1, r_2) = A_n(r_1)\sin(r_2\omega_n T) - A_n(r_2)\sin(r_1\omega_n T) \tag{9}$$

Finally, the amplitude a_n is determined by the following equation (10),

$$a_n = \frac{A_n(r_1)}{\cos(r_1\omega_n T + \phi_n)} \tag{10}$$

Summarizing the calculation procedure described above, frequency of input signal is determined using equations (4), (5) and (6), phase is after equation (7) and amplitude is using equation (10).

Applying this analysis procedure to amplitude quantized sampled data of speech signal, it is needed to assume that a number of frequency components of speech signal is m, and also that each frequency component is stationary within the time interval of 3m sequential data of speech signal. It is ensured experimentally that the number of frequency components set at 15 is appropriate for this analysis procedure. As speech signal contains the inherent time varying characteristics, it is needed to apply a kinds of analysis-by-synthesis procedure on signal waveform, inwhich the allowable error threshold is set for the reconstructed signal amplitude derived from the analyzed frequency components.

Results of Experiments

The analysis method of acoustic signal described above is applied to the telephone signal and also to speech signal to evaluate its performance experimentally. The telephone signal

	FREQUENCY [Hz]	PHASE [rad]	AMPLITUDE
TRUE	1209.0000000 697.0000000	0.128830418 0.744939420	1.0000000000 1.0000000000
DOUBLE	1209.0000000 697.0000000	0.128830418 0.744939420	1.0000000000 1.0000000000
16bits	1209.0222398 696.9944905	0.128704374 0.744998646	0.9999842601 0.9999770195
12bits	1209.0392902 696.9851871	0.128489109 0.745154821	0.9999958683 0.9999220908

Table 1(a).

	FREQUENCY [Hz]	PHASE [rad]	AMPLITUDE
TRUE	1209.0000000 697.0000000	0.142636245 0.703521939	1.0000000000 0.2000000000
DOUBLE	1209.0000000 697.0000000	0.142636245 0.703521939	1.0000000000 0.2000000000
16bits	1208.9991538 697.0324581	0.142650739 0.703405925	0.9999998615 0.2000051620
12bits	1208.9369781 696.6809930	0.142654324 0.705055985	0.9996142396 0.1999445766

Table 1(b).

	FREQUENCY [Hz]	PHASE [rad]	AMPLITUDE
TRUE	1209.0000000 697.0000000	4.031373416 2.389079203	0.2000000000 1.0000000000
DOUBLE	1209.0000000 697.0000000	4.031373416 2.389079203	0.2000000000 1.0000000000
16bits	1208.9869810 697.0024374	4.031427638 2.389069309	0.1999967897 1.0000005691
12bits	1212.4481302 695.4234739	4.022572017 2.394008924	0.1964944542 0.9970312356

Table 1(c).

TRUE : True values of components of input signal
DOUBLE: Results for 64 bits double precision data
16bits: Results for 16 bits quantized data
12bits: Results for 12 bits quantized data

Table 1 Results of Frequency Components Analysis for Signals. Amplitude Ratios of Two Components are 1:1, 5:1 and 1:5 in Tables 1(a), 1(b) and 1(c), respectively.

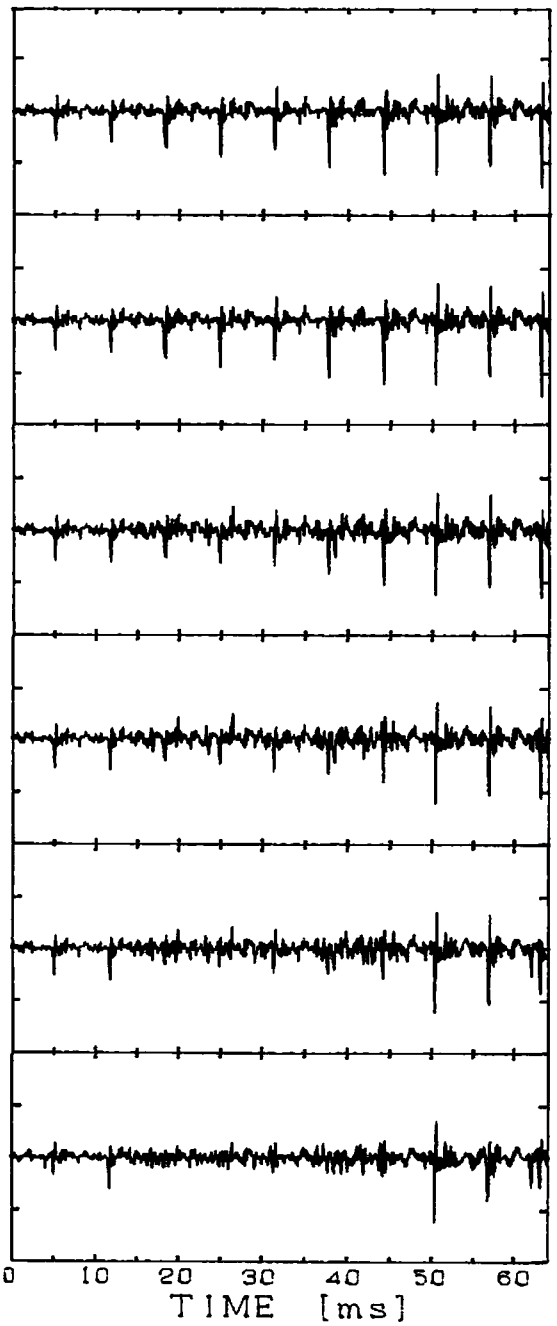

Fig. 3 (a)
Original Prediction Residual of Vowel /i/ Uttered by a Male Speaker

Fig. 3 (b)
Reconstracted Residual Using 15 Components Analyzed

Fig. 3 (c)
Reconstructed Residual Using 9 Components Analyzed

Fig. 3 (d)
Reconstructed Residual Using 7 Components Analyzed

Fig. 3 (e)
Reconstructed Residual Using 5 Components Analyzed

Fig. 3 (f)
Reconstructed Residual Using 3 Components Analyzed

tested is composed of two sinusoidal waves of 697 Hz and 1209 Hz, their initial phases are set arbitrary and their amplitudes are adjusted at three levels of 1:1,5:1 and 1:5. These signals are sampled at 10 kHz and quantized in three kinds of modes, that is, 64 bits double precision, 16 bits and 12 bits. Then sequentially sampled and quantized data are used for analysis. In the case of 64 bits double precision, sequential six data of signal are used for analysis. In 16 and 12 bits quantizations, sequential forty-five data of signal are used. Results are shown in Table 1(a), (b) and (c) for three kinds of amplitude ratios of 1:1, 5:1 and 1:5, respectively. It is seen that analyzed results are perfect in 64 bits double precision, but a few errors appear in 16 and 12 bits quantizations and also in the frequency component of lower amplitude.

This analysis method is also applied to speech signal, especially to the prediction residuals of PARCOR analysis. Speech signal is sampled at 10 kHz and quantized in 12 bits, then PARCOR analysis is executed and residual wave is extracted as shown in Fig. 3(a). This is an example for vowel /i/ predeeded by consonants /s/ uttered by a male speaker. The sequentially sampled data of 45 points are used for analysis of frequency components. By the use of 45 sequential samples, 15 kinds of frequency components are analyzed and it is possible to reconstruct the input signal from analyzed results. Result of reconstruction of residuals are shown in Fig. 3 (b). It is found that Fig. 3 (b) is much identical to Fig. 3 (a). Similar reconstructions of residuals are examined by the reduced frequency components of 9, 7, 5 and 3, and shown in Fig. 3 (c), (d), (e) and (f), respectively. Reduction of frequency components are made under the principle to truncate smaller amplitude component. It seems that the residuals reconstructed with less than 5 components contain several distortions, and such a visual impression coincides well with auditory quality of the synthesized speech, in which the vocal tract filter is excited directly by the reconstructed residuals described above.

Conclusion

Several studies on dynamic characteristics of speech signal are reviewed and an analysis method based on algebraic calculation is described. Validity of this method is confirmed from several experiments. Reduction of amount of information extracted is the future problem to be studied.

References

1) Itakura, F. and S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies," Trans. Inst. Electro. Comm. Eng. Japan, 53-A, 35-42, 1970

2) Tohkura, Y. and S. Saito, "Effects of the Window Width in

PARCOR Analysis," Report of the 1974 Autumn Meeting of the Acoust. Soc. Japan, 439-440, 1974

3) Tohkura, Y. and S. Saito, "Effects of the Sampling Frequency of Feature Parameters in Synthesized Speech," Report of the 1974 Spring Meeting of the Acoust. Soc. Japan, 371-372, 1974

4) Kawahara, H.,K. Tochinai and K. Nagata,"On the Linear Predictive Analysis using a Small Analysis Segment and its Error Evaluation," J. Acoust. Soc. Japan, 33, 470-479, 1977

5) Sagayama, S. and F. Itakura, "Composite Sinusoid Modeling applied to Spectrum Analysis of Speech," Trans Committee on Speech Res., Acoust. Soc. Japan, S79-06, 41-48, 1979

6) Tamaribuchi, K. and S. Saito, "Proposal of a New Analysis Method for Acoustic Waves Composed by Cosine Waves," Report of the Autumn Meeting of the Acoust. Soc. Japan, 247-248, 1987