A study of high quality speech synthesis based on the analysis of the randomness in speech signals

音声信号におけるランダムネスの解析に基づいた
高品質音声合成に関する研究

A Thesis Submitted for the Degree of
Doctor of Engineering
in the Graduate School of Engineering
Hokkaido University
Sapporo, Japan

by
Naofumi Aoki
March 2000

Abstract

Randomness observed in human speech signals is considered to be a key factor in the naturalness of human speech. This research project has investigated the characteristics of several kinds of randomness observed in human speech signals phonated by normal speakers. Based on the results of the analysis, some advanced techniques for artificially reproducing such randomness were developed with the aim of enhancing the voice quality of synthesized speech. The types of randomness particularly investigated in this project were: (1) amplitude fluctuation, (2) period fluctuation, (3) waveform fluctuation, (4) random fractalness of the source signals obtained by linear predictive analysis, and (5) unvoiced characteristics, namely, aperiodicity observed in voiced consonants. Using their statistical characteristics, a simple model was made for these forms of randomness, and was evaluated how it could contribute to realize high quality speech synthesis systems based on the LPC (linear predictive coding) vocoder.

Normal sustained vowels always contain a cyclic change of maximum peak amplitudes and pitch periods, even at those times when the values seem to be quite stable. This project investigated the statistical characteristics of the fluctuations that were particularly labeled amplitude fluctuation and period fluctuation, respectively. Since the frequency characteristics of these fluctuation sequences appeared to be roughly subject to a $1/f$ power law, the author reached the conclusion that amplitude and period fluctuation could be modeled as $1/f$ fluctuations for a preliminary model. Psychoacoustic experiments performed in this study indicated that the differences in the frequency characteristics of the amplitude and period fluctuation could potentially influence the voice quality of synthesized speech. Compared with $1/f^0$ (white noise), $1/f^2$, and $1/f^3$ fluctuation models, amplitude and period fluctuation modeled as $1/f$ fluctuations could produce voice quality which was more similar to that of human speech phonated by normal speakers.

Normal sustained vowels also always contain a cyclic change of the waveform itself, even during their most steady parts. This project investigated the statistical characteristics of the waveform fluctuations extracted from the residual signals of the LPC vocoder. Since the frequency characteristics of the waveform fluctuations appeared to be subject to a $1/f^2$ power law, the author reached the conclusion that the waveform fluctuations could be modeled as $1/f^2$ fluctuations for a preliminary model. Psychoacoustic experiments performed in this study indicated that the differences in the frequency characteristics of waveform fluctuations could potentially influence the voice quality of synthesized speech. Compared with $1/f^0$ (white noise), $1/f$, and $1/f^3$ fluctuation models, waveform fluctu-

ations modeled as $1/f^2$ fluctuations could produce voice quality which was more similar to that of human speech phonated by normal speakers.

Theoretically, the source signals of the LPC vocoder are defined as being characterized by a spectral -6 dB/oct decay in the frequency domain, when the -12 dB/oct glottal vibration and the $+6$ dB/oct mouth radiation characteristics are taken into consideration simultaneously. Since this frequency characteristic is equivalent to a $1/f^2$ spectral decay, the source signals of the LPC vocoder can be potentially classified as Brownian motion from the viewpoint of the random fractal theory. This project employed a multiresolution analysis method, based on Schauder expansion, in order to statistically investigate the time domain characteristics of the source signals. The results of the analysis indicated that the random fractalness was clearly observed, particularly when a large resolution level was chosen. The author also found that a certain limitation existed in the size of the discontinuity for the source signal waveforms obtained from human speech signals. Based on the results of the analysis, an advanced technique was newly developed with the aim of enhancing the voice quality of synthesized speech produced by the conventional impulse train. This study reached the conclusion that the buzzer-like degraded voice quality resulting from utilizing the impulse train could be improved by removing the extremely large discontinuity of the waveforms from the impulse train. The developed technique also included a method called random fractal interpolation for restoring power in the high frequency region which had been undesirably decreased by removing the sharpness of the impulse train.

The author implemented two applications that exemplified the effectiveness of the techniques developed through this research. One such application was a real-time vocoder system implemented on a digital signal processor (DSP) evaluation module (Texas Instruments, TMS320C62EVM); the other was a Japanese rule-based speech synthesis system implemented on a personal computer (Apple, Macintosh Quadra 840AV). Both applications employed the modified LPC vocoder as their speech synthesizer which fully implemented the features that were investigated in this research. In addition, these applications demonstrated how the voice quality of voiced consonants was enhanced by a MELP (mixed excitation linear prediction) scheme. Since voiced consonants are a mixture of both a periodic component attributed to voiced characteristics and an aperiodic component attributed to unvoiced characteristics, the waveforms of unvoiced consonants – which seem basically periodic due to reflecting the voiced feature – are disturbed in detail by the unvoiced feature. Psychoacoustic experiments conducted in this research clarified that synthesized voiced consonants produced by the conventional LPC vocoder tended

to degrade in voice quality, since such a vocoder completely disregards the incorporation of the unvoiced feature into the voiced consonants. An advanced technique, employing a wavelet transform for processing subband decomposition and reconstruction, was developed as a method for the inclusion of the unvoiced component with the voiced component at desirable bands. It was concluded that synthesized voiced consonants, for which the unvoiced feature was incorporated at high frequency subbands, could be perceived as possessing a more natural voice quality than that of the conventional LPC vocoder.

This project has reached the following two major conclusions: (1) the voice quality of synthesized speech can be enhanced by the inclusion of the randomness that is artificially produced by adequate models, (2) the knowledge acquired through the techniques developed in this project can be applied to the design of LPC-vocoder-based high quality speech synthesis systems that can be expected to produce more realistic human-like natural speech.

Contents

1 Introduction **1**

1.1 Motivation of the project 1

1.2 Overview of the following chapters 3

2 Analysis and perception of amplitude fluctuation and period fluctuation **6**

2.1 Introduction 6

2.2 Speech analysis 7

2.2.1 Speech samples 7

2.2.2 Stationarity of amplitude sequence and period sequence 8

2.2.3 Distributions of the coefficients of variation of amplitude sequence and the standard deviation of period sequence 9

2.2.4 Correlation between amplitude sequence and period sequence 10

2.2.5 Distributions of amplitude sequence and period sequence 10

2.2.6 Frequency characteristics of amplitude sequence and period sequence 11

2.3 Psychoacoustic experiments 12

2.3.1 Stimuli 12

2.3.2 Subjects 16

2.3.3 Procedures 16

2.3.4 Results of conditions 1 and 2 17

2.3.5 Results of conditions 3 and 4 18

2.4 Objective evaluation on the psychoacoustic experiments 19

2.4.1 Results of conditions 1 and 2 20

2.4.2 Results of conditions 3 and 4 21

2.5 Discussion 22

2.6 Conclusions 25

3 Analysis and perception of waveform fluctuation **26**

3.1 Introduction 26

3.2 Speech analysis . . . 26
3.2.1 Speech samples . . . 27
3.2.2 Extraction of waveform fluctuation . . . 27
3.2.3 Average power of the waveform fluctuation . . . 29
3.2.4 Frequency characteristics of waveform fluctuation . . . 30
3.3 Psychoacoustic experiments . . . 30
3.3.1 Stimuli . . . 30
3.3.2 Subjects . . . 34
3.3.3 Procedures . . . 34
3.3.4 Results of conditions 1 and 2 . . . 36
3.3.5 Results of conditions 3 and 4 . . . 37
3.4 Objective evaluation on the psychoacoustic experiments . . . 38
3.4.1 Results of conditions 1 and 2 . . . 40
3.4.2 Results of conditions 3 and 4 . . . 40
3.5 Discussion . . . 41
3.6 Conclusions . . . 43

4 Analysis and perception of the random fractalness in the source signals of LPC vocoder 44
4.1 Introduction . . . 44
4.2 Speech analysis . . . 45
4.2.1 Speech samples . . . 45
4.2.2 Extraction of the source signals of the speech samples . . . 46
4.2.3 Schauder analysis of the source signals . . . 47
4.3 Psychoacoustic experiments . . . 52
4.3.1 Modification technique for impulse train source . . . 52
4.3.2 Stimuli . . . 54
4.3.3 Subjects . . . 56
4.3.4 Procedures . . . 57
4.3.5 Results of the experiments . . . 58
4.4 Discussion . . . 59
4.5 Conclusions . . . 61

5 Development of two speech synthesis applications 62
5.1 Introduction . . . 62
5.2 Implementation of MELP vocoder using lifting wavelet transform . . . 62

5.2.1 Introduction . . . 63
5.2.2 Lifting wavelet transform . . . 64
5.2.3 Modification of triangular pulse by random fractal interpolation . . 66
5.2.4 Models of three fluctuations in the steady part of voiced speech . . 69
5.2.5 Implementation of MELP vocoder . . . 69
5.2.6 Conclusions . . . 71
5.3 Development of a rule-based speech synthesis system for the Japanese language using a MELP vocoder . . . 71
5.3.1 Introduction . . . 72
5.3.2 MELP scheme in the developed system . . . 73
5.3.3 Building a Japanese CV syllable database . . . 74
5.3.4 Three types of fluctuations observed in the steady part of sustained voiced speech . . . 77
5.3.5 Subjective evaluation of synthesized voiced consonants . . . 78
5.3.6 Implementing a Japanese rule-based speech synthesis system . . . 79
5.3.7 Conclusions . . . 79
5.4 Conclusions . . . 80

6 Conclusions **81**
6.1 Summary . . . 81
6.2 Original contributions . . . 83
6.3 Future work . . . 85

A Fractional Brownian motion **87**

B Theoretical definition of the source signals of the LPC vocoder **90**

C C program lists **92**
C.1 Linear predictive analysis . . . 92
C.2 Linear predictive synthesis . . . 96

Acknowledgements **99**

References **100**

Contribution **136**
Paper . . . 136
International conference . . . 137

Technical report 139
Presentation 139

Chapter 1

Introduction

This introductory chapter begins with a description of the background and motivation for this doctoral research project. Afterwards, a brief summary of each subsequent chapter in the thesis is provided.

1.1 Motivation of the project

One of the most significant research topics earnestly investigated in the area of speech signal processing is the establishment of methods for enhancing the voice quality of synthesized speech [474, 475, 476, 477, 478, 479, 481]. Such synthesized speech produced by even state-of-the-art speech synthesizers currently utilized in commercial applications is indicated still not sufficient [118, 259, 260, 480, 482, 483]. Those speech synthesis systems often produce abnormality of voice quality in terms of both acoustical and prosodical features [484, 485]. This abnormality is perceived as being quite different from the voice quality of human speech phonated by normal speakers. This problem increases particularly in the case of speech synthesizers based on a model that simulates the mechanism of human speech production. Because of the employment of inadequate models that oversimplify human speech production, such model-based speech synthesizers often produce degraded voice quality that is generally expressed as buzzer-like or machine-like. Since the precise mechanism of human speech production has not been fully understood due to the high complexity of its physical system, synthesizing natural voice quality by using model-based speech synthesizers is still recognized as a difficult goal to achieve.

This research focused on developing the practical methods for enhancing the naturalness of synthesized speech produced by a linear predictive coding (LPC) vocoder [110, 111, 113, 116, 150, 152, 171]. The LPC vocoder, which is widely used as a typical model-based speech synthesizer, generates synthesized speech based on the source-filter theory of human speech production [119, 120, 121, 122, 123, 182, 183, 184, 197, 198]. Accord-

ing to the source-filter theory, a source signal approximates either a glottal vibration for periodic voiced speech or an aspirated sound in the case of aperiodic unvoiced speech. Furthermore, its filter approximates a vocal tract structure that dominantly characterizes the spectral envelope of speech signals. The source-filter theory assumes that a speech signal is produced by filtering a source signal with a particular vocal tract filter. For convenience in the processing of speech synthesis, the conventional LPC vocoder employs a source signal represented simply by an impulse train for periodic voiced speech and white noise for aperiodic unvoiced speech [143, 145, 152, 156]. The degradation in the voice quality of the LPC vocoder is therefore considered to be caused mainly by such an oversimplification of the source signals [151, 153]. Bearing this in mind, the author attempted to design more realistic source signals of the LPC vocoder in order to enhance the naturalness of the resulting synthesized speech. The primary purpose of this research was to obtain the know-how to implement the LPC-vocoder-based high quality speech synthesis systems that produce more human-like natural speech. Since the LPC-vocoder-based speech synthesizer is widely employed in commercial speech applications, such as mobile telephones or car navigation systems [118, 475, 480, 484, 485, 519, 520], the know-how for enhancing the voice quality of LPC-vocoder-based speech synthesizer obtained from this project would potentially contribute to improving the quality of such promising industrial applications. In addition to this concrete objective, a further motivation of this research is the hope that the meaningful results which it issues may also potentially contribute to developing an advanced model of human speech production.

The author investigated the characteristics of various forms of randomness observed in human speech signals phonated by normal speakers. Even in the case when signals are very steady, they still contain several kinds of randomness [54, 161]. For instance, period fluctuations, which disturb the periodicity of voiced speech signals, are always observed even during the most steady part of sustained vowels in which the speech signals seem to be perfectly periodic [14, 15, 21, 22, 61]. Such random characteristics of human speech signals are considered to be a key factor in the naturalness of human speech. With this in mind, an examination was made to see whether such randomness could also be a factor for enhancing the naturalness of synthesized speech. Specifically, the kinds of randomness investigated in this project were: (1) amplitude fluctuation [2, 4], (2) period fluctuation [1, 2, 4], (3) and waveform fluctuation [1, 3]. The speech materials for the analysis of all three kinds of fluctuations were extracted from the steady portion of sustained voiced speech. In addition to these three fluctuations, this research also investigated: (4) random fractalness of the source signal obtained by linear predictive

analysis [5, 7], and (5) unvoiced characteristics, namely, aperiodicity observed in voiced consonants [6, 8]. Using their statistical characteristics, a simple model was made for these forms of randomness, and was evaluated how it could contribute to realize high quality speech synthesis systems based on the LPC (linear predictive coding) vocoder.

1.2 Overview of the following chapters

This doctoral thesis consists of six chapters, each of which is briefly summarized as follows: Chapter 2 describes the statistical characteristics of the amplitude and period fluctuations, as well as their psychoacoustic effects on influencing the naturalness of synthesized speech. Normal sustained vowels always contain a cyclic change of maximum peak amplitudes and pitch periods, even at those times when the values seem to be quite stable [11, 12, 14, 15, 16, 21, 22]. In this study, the updated values of the maximum peak amplitudes and pitch periods were labeled as amplitude and period sequence, respectively. Based on the statistical characteristics of these fluctuation sequences, the author attempted to develop advanced models of them. Since the frequency characteristics of these fluctuation sequences are a contributing factor that significantly influences the naturalness of sustained vowels, the model of these fluctuations was made which reflected their general frequency characteristics. Psychoacoustic experiments were also performed in order to examine the effectiveness of the developed model. The psychoacoustic experiments particularly investigated whether the differences in the frequency characteristics of amplitude and period fluctuations could influence the voice quality of synthesized speech.

Chapter 3 describes the statistical characteristics of waveform fluctuation, and the influence of psychoacoustic effects of waveform fluctuation on the naturalness of synthesized speech. In addition to the amplitude and period fluctuations described in Chapter 2, normal sustained vowels also always contain cyclic changes in the waveform itself, even during their most steady parts [69, 193, 194, 218]. This study particularly investigated some statistical characteristics of the waveform fluctuations which were extracted from the residual signals obtained by LPC inverse filtering. This was carried out in order to obtain the useful knowledge for the implementation of LPC-vocoder-based high quality speech synthesis systems. Based on the results of the analysis, this study attempted to develop a simple model for generating appropriate waveform fluctuations. Since the frequency characteristics of waveform fluctuations potentially influence the voice quality of synthesized speech, the author produced a model of the waveform fluctuations that reflected their general frequency characteristics. This study also performed psychoacoustic

experiments in order to examine the effectiveness of the developed model. The psychoacoustic experiments investigated whether the differences in the frequency characteristics of the waveform fluctuations could influence the voice quality of synthesized speech.

Chapter 4 describes random fractal characteristics observed in the source signals of the LPC vocoder and its psychoacoustic effects on the naturalness of synthesized speech. Theoretically, the source signals of the LPC vocoder are defined as being characterized by the spectral -6 dB/oct decay in the frequency domain when both -12 dB/oct glottal vibration and $+6$ dB/oct mouth radiation characteristics are simultaneously taken into consideration [110, 111, 116, 120, 121]. From the viewpoint of random fractal theory, since the spectral -6 dB/oct decay is equivalent to the spectral $1/f^2$ characteristics, the source signals of the LPC vocoder are potentially classified as Brownian motions [308, 309, 310, 311, 317, 334, 335, 350, 354]. This study particularly investigated such random fractalness of the source signals in terms of their time domain characteristics by using a method for multiresolution analysis based on Schauder expansion [428, 429, 462]. In addition, using the statistical characteristics of the source signals obtained by the Schauder analysis, this study newly developed a technique for enhancing the degraded voice quality of synthesized speech produced by the conventional impulse train. The psychoacoustic experiments were performed in order to examine the effectiveness of the developed techniques.

Chapter 5 describes two applications that exemplify the effectiveness of the techniques proposed through this project. One is a real-time vocoder system implemented on a digital signal processor (DSP) evaluation module (Texas Instruments, TMS320C62EVM) [299]; the other is a Japanese rule-based speech synthesis system implemented on a personal computer (Apple, Macintosh Quadra 840AV). Both applications employed the modified LPC vocoder as their speech synthesizer that fully implemented the features investigated in the project. This chapter also described the unvoiced characteristics observed in voiced consonants [160, 162, 163]. Since a voiced consonant is a mixture of both a periodic component attributed to voiced characteristics and an aperiodic component attributed to unvoiced characteristics, waveforms of unvoiced consonants – which seem basically periodic due to the voiced feature – are disturbed in detail by the unvoiced feature. Psychoacoustic acoustic experiments performed in this project clarified that synthesized voiced consonants produced by the conventional LPC vocoder tended to degrade in voice quality, since such a vocoder completely disregard the inclusion of the unvoiced feature into the voiced consonants. An advanced technique employing wavelet transforms for subband decomposition and reconstruction was developed as a method for the inclusion of the unvoiced component with the voiced component. Psychoacoustic experiments were

performed in order to examine the effectiveness of the newly developed technique. They investigated whether synthesized voiced consonants, for which the high frequency voiced subbands were replaced with unvoiced characteristics, could be perceived as more natural in voice quality than that of the conventional LPC vocoder.

In Chapter 6, the results obtained from this project are summarized. Additionally, some unique contributions of this research are described. Lastly, some other topics left for future study are also discussed.

Chapter 2

Analysis and perception of amplitude fluctuation and period fluctuation

2.1 Introduction

Synthesized sustained vowels tend to be perceived as possessing a buzzer-like unnatural quality when their maximum amplitudes and pitch periods are kept constant. It has been shown that one effective way to mitigate this undesirable degradation is to artificially incorporate the fluctuations in the maximum amplitudes and pitch periods [11, 23, 30, 31]. Furthermore, it has been indicated that the characteristics of the fluctuations obtained from human sustained vowels depend on the types of voice quality [9, 10, 11, 12, 14, 15, 16, 21, 22, 41]. Taking account of this finding, it can be considered that the synthesized sustained vowels, in which both the fluctuations appropriately reflect the characteristics of normal sustained vowels, could potentially sound similarly to normal sustained vowels.

From earlier investigations on the perception of fluctuations, it is indicated that their size affects the voice quality of synthesized sustained vowels [11, 23, 30, 31]. Since the excessively large size of the fluctuations results in a pathologically rough voice quality, appropriately adjusting the size is necessary for enhancing the voice quality. Another property of the fluctuations which significantly influences the voice quality is their frequency characteristics [1, 2, 3, 4, 70, 72, 74, 81]. Since the frequency characteristics of normal sustained vowels generally differ from those of pathologically rough cases [18, 25, 26, 27, 38, 39, 40, 78, 79, 80], such characteristics should be taken into account in incorporating the fluctuations.

In developing an appropriate model of the amplitude and period fluctuations which can be utilized to enhance the voice quality of synthesized sustained vowels, this study has analyzed several statistical characteristics of both kinds of fluctuations obtained from normal human speakers. The analysis included the frequency characteristics as well as

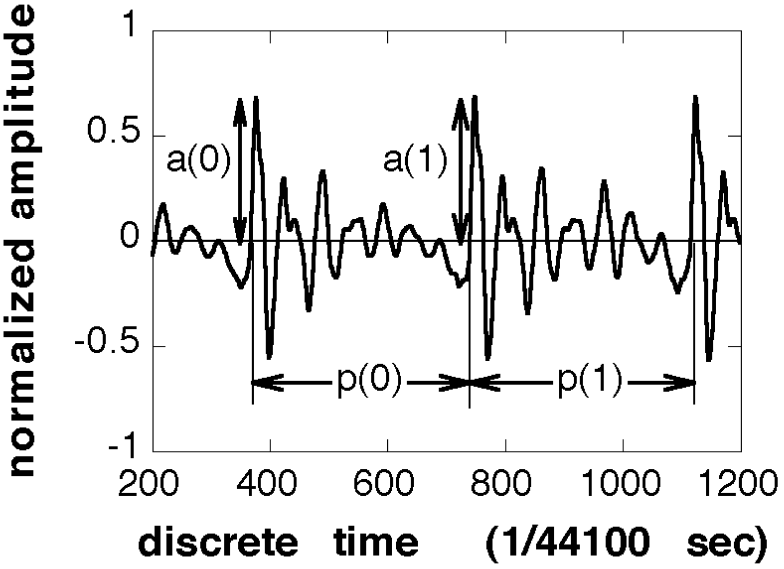

Figure 2.1: Extraction of amplitude sequences $a(n)$ and period sequences $p(n)$

the size of the fluctuations. In order to investigate the effectiveness of modeling the frequency characteristics for enhancing the voice quality of synthesized sustained vowels, psychoacoustic experiments were carried out. The experimental results were discussed with an objective evaluation method newly devised by the author.

2.2 Speech analysis

This section describes several statistical characteristics of the amplitude and period sequences obtained from normal sustained vowels. The speech analysis included the investigation of the size of the standard deviation and the frequency characteristics of both fluctuation sequences. In addition, their stationarity, distribution, and the correlation between the amplitude and period sequence were investigated to obtain the useful information for modeling the fluctuation sequences.

2.2.1 Speech samples

Ten male subjects between 22 and 26 years of age who did not suffer from any laryngeal disorders were selected in order to obtain normal sustained vowels. Each subject was requested to phonate the sustained vowel /a/ as steadily as possible in a soundproof anechoic room (Rion, audiometry room) toward an electret condenser microphone (Audio-

technica, AT822) at a distance of about 15 cm from the mouth. The sustained vowels were directly recorded onto a hard disk of a personal computer (Apple, Macintosh Quadra 800) by way of a microphone mixer (Mackie, microseries 1202-VLZ), a low pass filter (8th order Bessel characteristic), and an analog-to-digital converter (Digidesign, audiomedia II). The sampling rate and quantization level were 44.1 kHz and 16 bits, respectively. The cut-off frequency of the low pass filter was set at 5 kHz. The subjects phonated the vowels at a pitch and loudness that was comfortable. The duration of the phonation was requested to be approximately ten seconds. All records contained a steady portion of at least 512 pitch periods lasting over approximately four seconds, in which the mean pitch period was found to range from 7.6 msec to 9.1 msec. The calculated mean pitch period of all speech samples was 8.3 msec. The sound pressure level (SPL) was also measured by a precision noise meter using the C weighting condition (Brüel & Kjær, type 2209) which was placed about 15 cm from the mouth [83, 84, 412]. Measured SPL ranged from 80 dB to 86 dB for all subjects. The gain of the microphone mixer was adjusted for each subject for an optimal recording level. Twenty speech samples were taken per subject. Two hundred speech samples in total (20 utterances $\times$ 10 subjects) were obtained [5, 20].

Since updated values for each cycle of both fluctuations were required to form fluctuation sequences, each value of maximum amplitude sequence and pitch period sequence was extracted from the digitized speech samples using a peak-picking method and a zero-crossing method, respectively [12, 13, 14, 15, 16, 17]. Figure 2.1 illustrates the definition of the amplitude and period sequence, where the amplitudes of the speech sample were normalized to range from -1 to 1. This normalization corresponds to from -32768 to 32767, the quantization level of 16 bits. Figure 2.2 shows an example of a pair of the amplitude sequence and period sequence taken from one of the speech samples obtained in the sampling session.

2.2.2 Stationarity of amplitude sequence and period sequence

In developing a model of the fluctuation sequences, it is useful to examine whether the size of fluctuation sequences changes according to the duration. In order to clarify this issue, the stationarity of the amplitude and period sequences were investigated. In this study, stationarity was defined as the time-invariance of the mean and the variance of the sequence [108, 109, 386, 409]. A runs-test was employed in order to judge whether a fluctuation sequence was a stationary or nonstationary process. The test examined whether the changes in the short-time mean and variance of a sequence were acceptable as those of a stationary process [108, 386]. The severity of the tests defined by the level

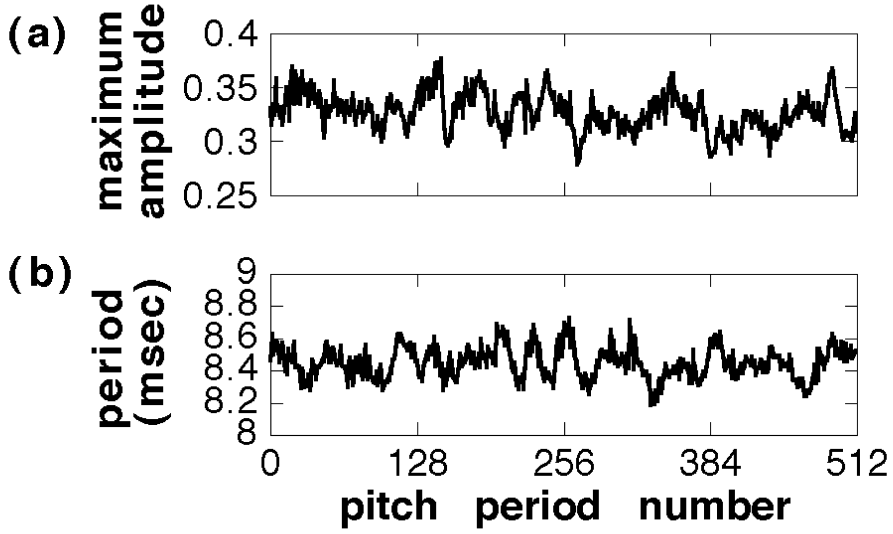

Figure 2.2: Examples of (a) amplitude sequence and (b) period sequence

of significance was chosen to be $\alpha = 0.01$. Since almost all of the amplitude (92 %) and period sequences (96 %) were found to be acceptable as stationary processes, it was concluded that amplitude and period sequences extracted from the steady part of normal sustained vowels could be regarded as stationary processes. This indicates that the size of fluctuation sequences would not change even though their duration changed. This result shows one of the features of the fluctuation sequences to be taken into account in developing a model.

2.2.3 Distributions of the coefficients of variation of amplitude sequence and the standard deviation of period sequence

As mentioned earlier, the size of the amplitude and period sequence is considered to be a significantly important factor that influences the voice quality of sustained vowels. The standard deviations of both fluctuation sequences were statistically analyzed in order to investigate their valid size for normal sustained vowels. Since the gain of the microphone mixer was adjusted depending on the loudness of the subject, the magnitudes of the amplitude sequences were represented by arbitrary units. Therefore, the coefficient of variation (C.V.) was chosen as the measure for the size of the amplitude sequences. C.V. is a measure that represents the standard deviation of a sequence normalized by the mean [19, 409]. The distribution of the C.V. of the amplitude sequences is shown in figure 2.3 (a). It ranged from 2 to 20 % and its mode was found to be around 7.5 %. The

mode (6.68 ± 3.03 %) was within the normative range for normal amplitude sequences (mean ± standard deviation %) reported in a literature [20]. As for the size of the period sequences, the standard deviation (S.D.) was employed as a measure to compare the size of the period sequences. As shown in figure 2.3 (b), the S.D. of the period sequences ranged from 0.03 to 0.13 msec with a mode of around 0.05 msec. In addition, the C.V. of the period sequences was also calculated for a comparison with the normative range reported in the literature [20]. The C.V. of the period sequences ranged from 0.5 % to 1.6 % with a mode around 0.8 %. The mode was within the normative range 1.05 ± 0.40 % for the normal period sequences [20].

The correlation of the size of the amplitude sequence and the period sequence, both of which were obtained from an identical speech sample, was also investigated with the aim of developing a more realistic model of the fluctuation sequences. Although a moderate positive correlation coefficient ($r = 0.64$) was obtained from the scattergram of average tendency, there was such a variety in the combinations that it obscured the meaning of the average correlation coefficient. For example, one of the cases showed that the C.V. of the amplitude sequences was small, while the S.D. of the period sequences was large, and *vice versa*. Such deviations were found among individual speech samples even from the same subject. Since it was difficult to make any meaningful conclusion that the amplitude and period sequences were either correlated or independent in their size, the size of a period sequence for an amplitude sequence or *vice versa* could be chosen rather arbitrarily for the preliminary model.

2.2.4 Correlation between amplitude sequence and period sequence

The correlation between the amplitude and period sequence also influences the model of the fluctuation sequences. Correlation coefficients were calculated from all of the pairs of the amplitude and period sequences [108, 109, 409]. Since individual tendency was not particularly different from subject to subject, a pooled distribution of the correlation coefficients was obtained from all subjects. The result is shown in figure 2.4. The mean and standard deviation of the distribution were -0.12 and 0.32, respectively. Since the correlation coefficients tended to center approximately at zero, both fluctuation sequences are likely to be modeled as independent processes of each other.

2.2.5 Distributions of amplitude sequence and period sequence

The distributions of fluctuation sequences are one of the important features for developing their adequate model. It has been reported that the distributions of the amplitude and

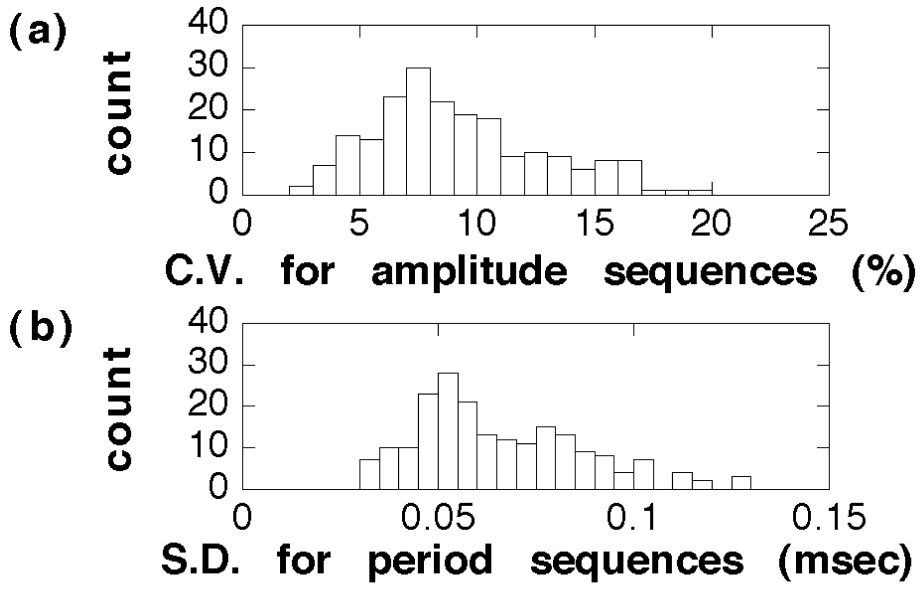

Figure 2.3: Distributions of (a) coefficient of variation (C.V.) of amplitude sequences and (b) standard deviation (S.D.) of period sequences

period sequences are generally regarded as Gaussian [5, 74]. This tendency was statistically reexamined in this study by the chi-squared test. The results indicated that the Gaussian distribution is considered to be one of the possible choices in modeling, since more than half of the distributions of the amplitude sequences (61 %) and the period sequences (60 %) were acceptable as Gaussian in terms of the statistical test. However, the results did not clearly confirm that the Gaussian distribution was always necessary for a model, since it appeared that a number of the distributions were not acceptable as Gaussian. Further investigation of their distribution in order to develop more precise models of the fluctuation sequences was left for future study.

2.2.6 Frequency characteristics of amplitude sequence and period sequence

The frequency characteristics of the amplitude and period sequences were investigated using the 512-point fast Fourier transform (FFT) with a Hamming window [96, 97, 98, 104, 105, 108, 109, 409]. As a result, it was found that the gross approximation of the frequency characteristics was subject to the spectral $1/f^{\beta}$ power law [308, 309, 310, 311, 317, 334, 335, 350, 354], although the details might deviate from this approximation. This tendency was consistently observed among all fluctuation sequences. Figure 2.5 shows examples of the frequency characteristics of the amplitude and period sequences. The value of the exponent β was estimated by the least-squares line fitting method [108, 109, 409],

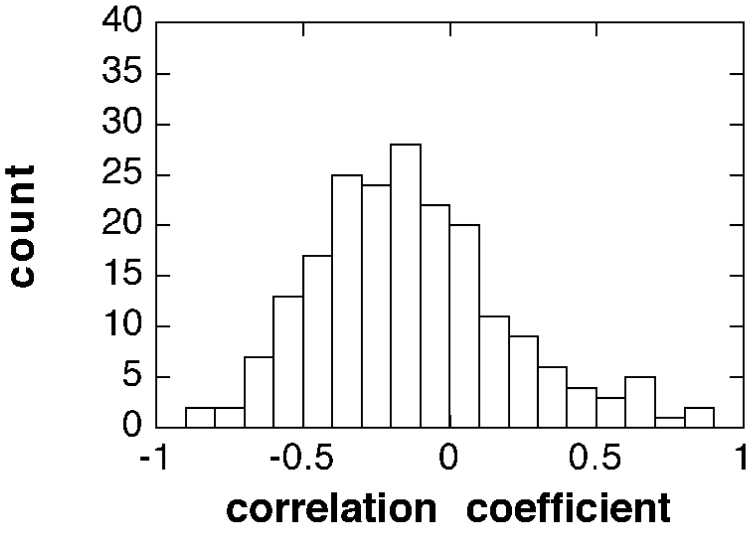

Figure 2.4: Distribution of the correlation coefficients between amplitude sequences and period sequences. The mean and standard deviation of the distribution are -0.12 and 0.32, respectively.

where the exponent β was equivalent to the gradient of the fitted line in the frequency characteristics plotted in the log-log scale. The value of β of this example was 0.99 for the amplitude sequence and 0.96 for the period sequence. Since the value of β tended to center at roughly "1" for all the fluctuation sequences, the results of frequency analysis concluded that both fluctuation sequences could be modeled as spectral $1/f$ processes for a preliminary model.

2.3 Psychoacoustic experiments

In order to explore the influence of the frequency characteristics of fluctuation sequences on speech perception, a series of psychoacoustic experiments were conducted. The purpose of the experiments was to investigate how the differences in their frequency characteristics caused the subjective differences in the voice quality of synthesized sustained vowels.

2.3.1 Stimuli

Stimuli for the psychoacoustic experiments were sustained vowels /a/ produced by a partial autocorrelation (PARCOR) synthesizer [110, 111, 116, 120, 121]. The vowels were characterized by the different combinations of the amplitude and period sequences.

The filter coefficients of the PARCOR synthesizer were derived from one of the speech

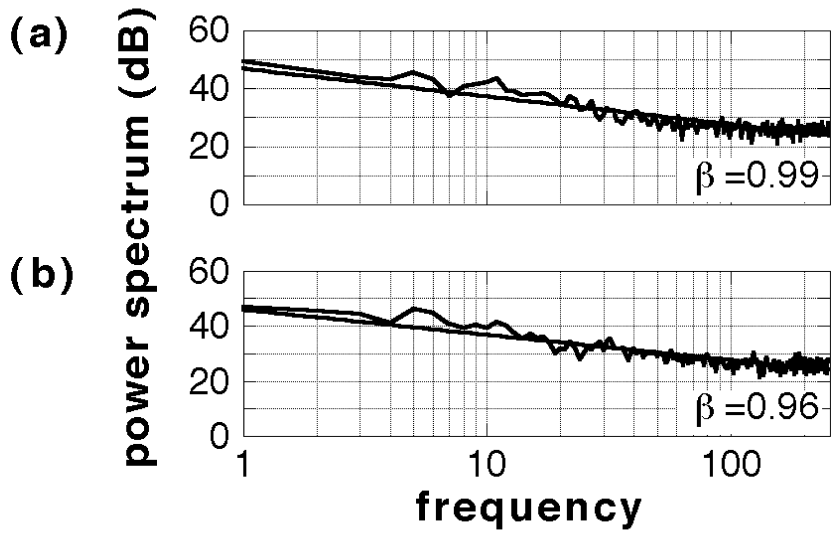

Figure 2.5: Examples of the frequency characteristics of (a) amplitude sequence, and (b) period sequence

samples whose amplitude and period sequence showed the representative characteristics of all the fluctuation sequences investigated in this study. Specifically, the C.V. of the amplitude sequence was 7.5 % and the S.D. of the period sequence was 0.05 msec. Both fluctuation sequences were acceptable as stationary processes. Their distributions were also acceptable as Gaussian. Since the correlation coefficient between the amplitude and period sequence was -0.06, they were considered not to be strongly correlated with each other. Their frequency characteristics were approximated by the spectral $1/f^{\beta}$ power law in which the value of the exponent β was 0.99 for the amplitude sequence and 0.96 for the period sequence. Both frequency characteristics are respectively shown in figure 2.5. Furthermore, the mean pitch period calculated from the period sequence was 8.4 msec, which was close to the average of all speech samples (8.3 msec). Both fluctuation sequences were employed in synthesizing the stimuli.

The filter order of the PARCOR synthesizer was set at 40 for the condition of a 44.1 kHz sampling rate. The filter order was determined by a visual inspection which ascertained that the frequency characteristics of the PARCOR filter in this condition appropriately represented four dominant formants below the cut-off frequency of the low pass filter of 5 kHz. The coefficients of the PARCOR filter were not altered during the synthesis. This condition was based on the assumption that the characteristics of the vocal tract filter for normal sustained vowels do not substantially change during the

phonation. In order to synthesize sustained vowels, normal impulse trains, which are conventionally used for synthesizing voiced speech, were employed as source signals of the PARCOR synthesizer [110, 111, 116, 120, 121]. Period fluctuations in the stimuli were implemented by the pulse position modulation [94, 137]. Period sequences were employed as the modulating signals. In order to guarantee the accuracy of the modulation at a 44.1 kHz sampling rate, impulse trains represented in analog form were passed through a -6 dB/oct low pass filter and sampled at a 44.1 kHz sampling rate. This low pass filter theoretically consisted of both -12 dB/oct glottal characteristics and $+6$ dB/oct radiation characteristics from the mouth [110, 111, 116, 120, 121]. After synthesizing sustained vowels by the PARCOR filter, cyclic gain adjustments defined by amplitude sequences were employed to incorporate amplitude fluctuations.

Each stimulus consisted of 128 pitch periods. Since the mean pitch period was set to be 8.4 msec, the duration of each stimulus was approximately one second. A linearly increasing or decreasing gate function whose duration was 10 msec was employed at the beginning and the end of each stimulus in order to prevent undesirable clicking-like sounds.

The psychoacoustic experiments investigated four different conditions in regard to amplitude and period sequences. In conditions 1 and 2, the frequency characteristics of the amplitude sequences were manipulated, while all of the stimuli employed the period sequence obtained from the speech sample. On the other hand, the period sequences were changed from stimulus to stimulus in conditions 3 and 4, while all of the stimuli employed the amplitude sequence obtained from the speech sample. Thus, conditions 1 and 2 focused on how the differences in the amplitude sequences influenced perception, while conditions 3 and 4 focused on how perception was influenced by the different frequency characteristics of the period sequences.

Fourteen stimuli labeled from "a" to "n" were produced for each condition. Stimulus "a" employed the amplitude sequence and the period sequence obtained from the speech sample. Although stimulus "a" was not the speech sample itself, its voice quality was considered to reflect the characteristics of the amplitude and period sequence of the speech sample. Since stimulus "a" was used as a reference stimulus in evaluating all stimuli, including stimulus "a" itself, it was also labeled the reference stimulus. Comparisons between the reference stimulus and stimulus "a" were the control for the experiment. Stimulus "b" was produced without incorporating amplitude or period fluctuation. The aim of this stimulus was to examine whether amplitude or period fluctuation was an important factor for speech perception. In addition, four stimulus groups were produced, each of which consisted of three stimuli. The three stimuli of each stimulus group employed

the amplitude or period sequences whose frequency characteristics were classified in the same category, while the fluctuation sequences themselves were different from each other, since randomization used for producing the fluctuation sequences was different. The four stimulus groups were labeled as "$\beta 0$", "$\beta 1$", "$\beta 2$", and "$\beta 3$" according to the frequency characteristics of the fluctuation sequences. Stimulus group $\beta 0$, consisting of stimuli "c", "d", and "e", employed spectral $1/f^0$ sequences (white noise). Stimulus group $\beta 1$, consisting of stimuli "f", "g", and "h", employed spectral $1/f$ sequences. Stimulus group $\beta 2$, consisting of stimuli "i", "j", and "k", employed spectral $1/f^2$ sequences. Stimulus group $\beta 3$, consisting of stimuli "l", "m", and "n", employed spectral $1/f^3$ sequences.

The value of the exponent β in the spectral $1/f^\beta$ sequences for each stimulus group was considered to be a rather preliminary choice. The integer values from zero to three were employed, since there was no *a priori* knowledge about the relationship between speech perception and the values of β. Using three stimuli in each stimulus group aimed to examine whether or not the perceptual effects were categorized by the values of β.

The C.V. of all amplitude sequences was set to be 7.5 % for condition 1 and 15 % for condition 2, in which the C.V. of the amplitude sequence obtained from the speech sample was also readjusted to be 7.5 % or 15 %. Condition 2 aimed to examine whether or not the larger C.V. of amplitude sequences influenced the experimental result compared with condition 1. The S.D. of period sequences was set to be 0.05 msec for all stimuli throughout conditions 1 and 2.

On the other hand, the S.D of period sequences was set to be 0.05 msec for condition 3 and 0.10 msec for condition 4. Condition 4 aimed to examine whether or not the larger S.D. of period sequences influenced the experimental result compared with condition 3. The C.V. of amplitude sequence was set to be 7.5 % for all stimuli throughout conditions 3 and 4.

All of the amplitude and period sequences employed in the stimulus groups were fractional Brownian motions artificially produced by the FFT method [311, 317, 318, 319, 334, 350, 354, 358]. Gaussian white noise was first transformed to the frequency domain, then passed through the low pass filter characterized by the spectral $1/f^\beta$ power law. The result was transformed back into the time domain. Although the speech analysis indicated that the distributions of the amplitude and period sequences were not necessarily Gaussian, the fluctuation sequences employed in this study were simply assumed to be Gaussian.

The power spectrum of a spectral $1/f^\beta$ sequence is represented as

$$S_v(f) = |T(f)|^2 S_w(f) \propto |T(f)|^2, \tag{2.1}$$

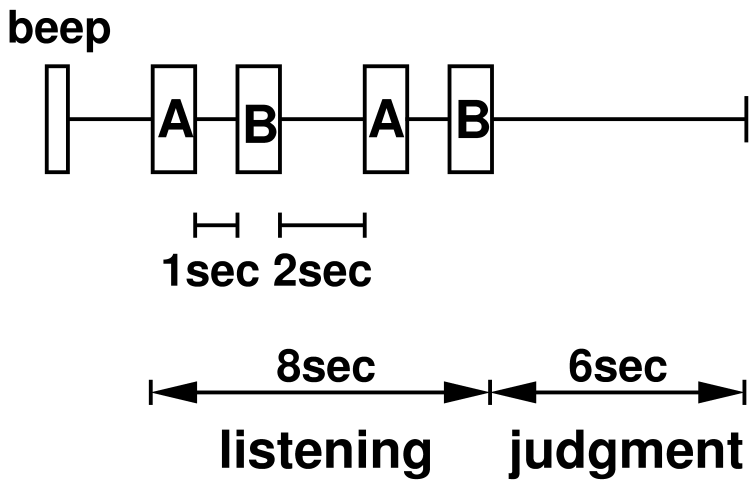

Figure 2.6: Procedure of paired-comparison test

where

$S_v(f)$ is the power spectrum of a spectral $1/f^\beta$ sequence $v(t)$,

$T(f)$ is the frequency characteristics of spectral $1/f^\beta$ filter,

$S_w(f)$ is the power spectrum of Gaussian white noise.

Thus, the spectral $1/f^\beta$ filter is required to be

$$|T(f)| = 1/f^{\beta/2}. \tag{2.2}$$

2.3.2 Subjects

Twenty subjects consisting of twelve males and eight females participated in the experiment. Their age ranged from 20 to 26 years. None of them were experienced in psychoacoustic experiments. All reported having no hearing problems.

2.3.3 Procedures

All stimuli were synthesized using a personal computer (Apple, Macintosh Quadra 800). The stimuli were passed through a digital-to-analog converter (Digidesign, audiomedia II) and then through a low pass filter (8th order Bessel characteristic). The sampling rate and quantization level were 44.1 kHz and 16 bits, respectively. The cut-off frequency of the low pass filter was set at 10 kHz. The stimuli were presented through a monitor speaker (Denon, USC-101), which was attached to a pre-main amplifier (Sansui, AU-α 507XR). The speaker was placed 1.5 meters in front of the subject in a soundproof anechoic room (Rion, audiometry room). The SPL of the stimuli was set to be 65 dB upon presentation.

Each subject took part individually in the experiment under all four conditions. Each condition consisted of fourteen paired-comparison trials to evaluate the similarity between the preceding stimulus A and the succeeding stimulus B. The reference stimulus, namely stimulus "a", was always chosen stimulus A, while stimulus B was one of the fourteen stimuli produced for each condition. The order of the presentation in regard to stimulus B was randomized.

Stimuli A and B were presented to the subject twice in the form of an AB pair as illustrated in figure 2.6. There was a one-second silent interval between stimuli A and B, and a two-second silent interval between the first and second AB pair. For the judgment, a six-second interval was given to the subject after listening to the two AB pairs.

The subject was asked to judge whether or not the voice quality of stimulus B was perceived as the same as that of stimulus A. They were forced to select one of the following three choices: (1) "same", (2) "undecided", or (3) "different". The three-point scale aimed to examine whether or not the subject could correctly distinguish the voice quality between a stimulus created from artificially produced fluctuation sequences and one created from fluctuation sequences obtained from the speech sample. For the above reason, the five or seven-point scale, which is conventionally used to grade the differences in voice quality, was not employed [402, 403, 404]. In order to compare the experimental results, the three choices were translated into a numerical measure named similarity, which was defined as (1) 100 %, (2) 50 %, and (3) 0 %, corresponding to the three choices.

2.3.4 Results of conditions 1 and 2

The results of conditions 1 and 2 are summarized in figure 2.7 (a) and (b), respectively. The similarity of each stimulus is represented by an open circle which shows the average of the results over all subjects. It appeared that the control stimulus "a" and the stimulus group $\beta 1$ tended to be evaluated as more similar to the reference stimulus than the other stimuli throughout conditions 1 and 2. These results indicated that most of the subjects found it difficult to distinguish the voice quality of the stimulus group $\beta 1$ from that of the reference stimulus, namely stimulus "a". The larger C.V. of the amplitude sequences examined in condition 2 did not substantially influence this tendency.

As for the other stimuli, most of the subjects reported that the voice quality of the stimulus group $\beta 0$ was rougher than that of the reference stimulus. Particular changes in the loudness were perceived in the stimulus groups $\beta 2$ and $\beta 3$, while such features were not perceived in the reference stimulus. Some of the subjects also reported that the changes in loudness of stimulus "b" were perceived as being flat compared with the

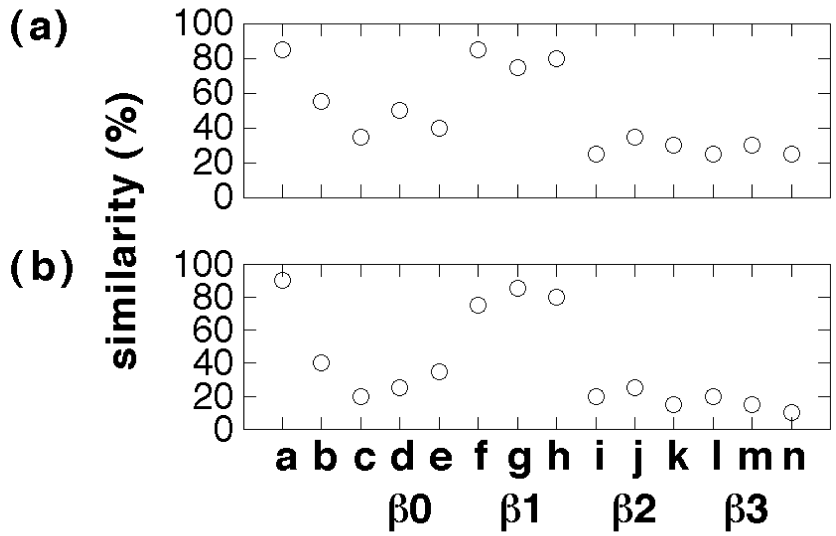

Figure 2.7: (a) Results of experiment 1: C.V. of amplitude sequence was set at 7.5 %. (b) Results of experiment 2: C.V. of amplitude sequence was set at 15 %.

reference stimulus. These differences could be a subjective clue to discriminate between the reference stimulus and these stimuli.

2.3.5 Results of conditions 3 and 4

The results of conditions 3 and 4 are summarized in figure 2.8 (a) and (b), respectively. It appeared that the control stimulus "a" and the stimulus group $\beta 1$ tended to be evaluated as more similar to the reference stimulus than the other stimuli throughout conditions 3 and 4. These results indicated that most of the subjects found it difficult to distinguish the voice quality of the stimulus group $\beta 1$ from that of the reference stimulus, namely stimulus "a". The larger S.D. of the period sequences examined in condition 4 did not substantially influence this tendency.

As for the other stimuli, most of the subjects reported that the voice quality of the stimulus group $\beta 0$ was rougher than that of the reference stimulus. Unstable changes in the pitch were perceived in the stimulus groups $\beta 2$ and $\beta 3$, while such features were not perceived in the reference stimulus. Furthermore, most of the subjects reported that stimulus "b" was perceived as buzzer-like compared with the reference stimulus. These differences could be a subjective clue to discriminate between the reference stimulus and these stimuli.

Since the similarity of stimulus "b" commonly tended to be judged as low throughout

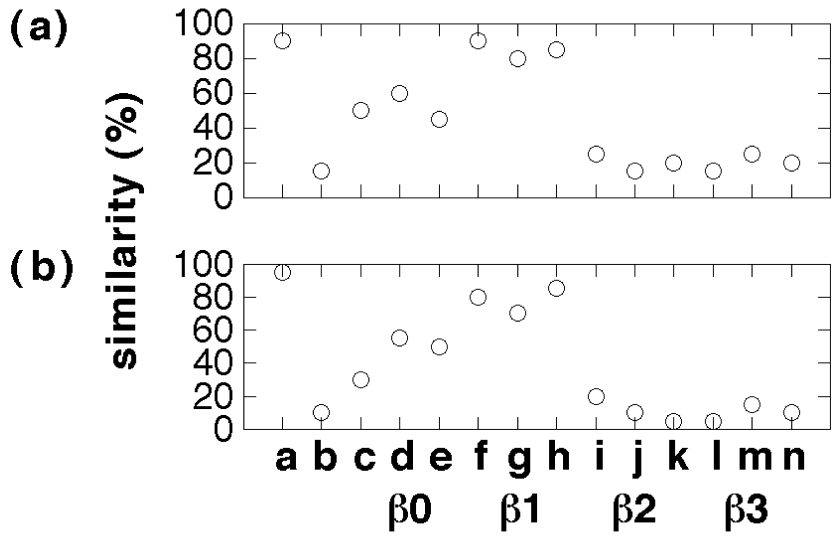

Figure 2.8: (a) Results of experiment 3: S.D. of period sequence was set at 0.05 msec. (b) Results of experiment 4: S.D. of period sequence was set at 0.10 msec.

all the conditions, it can be concluded that the amplitude and period fluctuation play significant roles in the speech perception of sustained vowels as mentioned in the previous literature [11, 23, 30, 31, 61]. The results of the other stimuli suggest that the differences in the frequency characteristics of the amplitude and period sequence can significantly influence speech perception. High similarity between the stimulus group $\beta 1$ and the reference stimulus can be attributable to the similarity in the frequency characteristics of the fluctuation sequences between stimulus group $\beta 1$ and the reference stimulus.

Since there were no large differences in the similarity among all the stimuli of the stimulus group $\beta 1$, the randomization for producing the fluctuation sequences could have little effect on speech perception. In addition, it seemed that the similarity of the stimulus group $\beta 1$ could not be significantly influenced by the differences in the size of the fluctuation sequences. The similarity of the stimulus group $\beta 1$ was consistently evaluated high throughout conditions 1 and 2 as well as during conditions 3 and 4, in spite that the size of the fluctuation sequences varied.

2.4 Objective evaluation on the psychoacoustic experiments

An objective evaluation of the stimuli, all of which were employed in the psychoacoustic experiments, was performed to judge the validity of the subjective evaluation. As earlier studies indicate, the voice quality can be objectively classified by the index which measures

the ratio of the inharmonic spectra energy to the harmonic spectral energy [42, 43, 59, 60]. Such an index becomes large in the case of purely periodic speech signals that may be perceived as quite buzzer-like. On the other hand, the index becomes small in the case of aperiodic speech signals that may be perceived as being a rough voice quality.

This study devised a new objective index named Inharmonic Energy Ratio (IHER) for investigating the voice quality of the stimuli utilized in the psychoacoustic experiments. As defined in equation 2.3, in which f denotes frequency and $S(f)$ the power spectrum, the proposed index measured the ratio of the energy obtained from inharmonic spectral regions to that of the harmonic spectral region. The definition of IHER is also illustrated in figure 2.9, in which h_j represents jth harmonic frequency, $a = (h_{j-1} + h_j)/2$, $b = (h_j + h_{j+1})/2$, $c = (a + h_j)/2$, and $d = (h_j + b)/2$. The inharmonic spectral regions were defined as the bands which met the conditions $a \leq f < c$ or $d \leq f < b$, while the other region represented as $c \leq f < d$ was defined as a harmonic region. This definition came from the fact that it was hard to accurately estimate the purely harmonic frequencies when the speech signals contained a certain aperiodicity, such as the amplitude and period fluctuation. In that situation, the separation of the harmonic spectral components was quite difficult, especially in the high frequency region. Although IHER does not require the estimation of harmonic frequencies, it was still difficult to separate the harmonic spectral regions from the inharmonic spectral regions in the high frequency region. Therefore, this study applied the proposed index to only the low frequency region under 10th harmonic frequency. The power spectra utilized in the calculation of the index was estimated by 4096-point FFT.

$$\mathrm{IHER}(j) = 10 \log \frac{\sum_{f \geq c}^{f<d} S(f)}{\sum_{f \geq a}^{f<c} S(f) + \sum_{f \geq d}^{f<b} S(f)} \tag{2.3}$$

2.4.1 Results of conditions 1 and 2

The calculated indices for all the stimuli used in conditions 1 and 2 are summarized in figure 2.10 (a) and (b), respectively. It appeared that the indices of both the control stimulus "a" and the stimulus group $\beta 1$ tended to be similar throughout conditions 1 and 2. On the other hand, the indices of stimulus "b", the stimulus group $\beta 2$, and the stimulus group $\beta 3$ were larger than that of the control stimulus "a". As for the stimulus group $\beta 0$, the indices were smaller than that of the control stimulus "a". The larger C.V. of the amplitude sequences examined in condition 2 did not substantially influence this tendency.

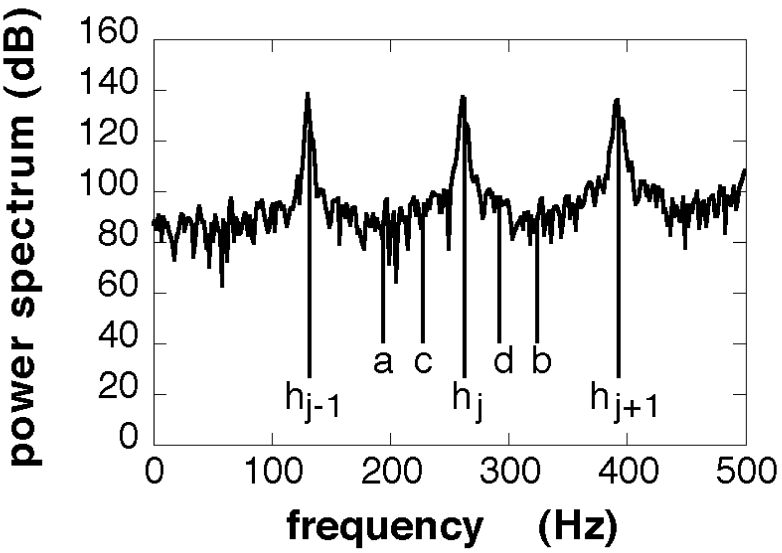

Figure 2.9: Definition of Inharmonic Energy Ratio (IHER)

2.4.2 Results of conditions 3 and 4

The calculated indices for all the stimuli used in conditions 3 and 4 are summarized in figure 2.11 (a) and (b), respectively. It appeared that the indices of both the control stimulus "a" and stimulus group $\beta 1$ tended to be similar throughout conditions 1 and 2. On the other hand, the indices of stimulus "b", the stimulus group $\beta 2$, and the stimulus group $\beta 3$ were larger than that of the control stimulus "a". As for the stimulus group $\beta 0$, the indices were smaller than that of the control stimulus "a". The larger S.D. of the period sequences examined in condition 4 did not substantially influence this tendency.

Since the indices for the stimulus group $\beta 1$ and the control stimulus "a" fell within a similar range, these stimuli could be categorized as belonging to the same group in terms of their spectral characteristics. Since the indices of stimulus "b", the stimulus group $\beta 2$, and the stimulus group $\beta 3$ were larger than that of the control stimulus "a" through all the conditions, the harmonic spectral components for these stimuli were considered to be more dominant in their spectral structures. On the other hand, it was considered that the inharmonic spectral components were dominant for the stimulus group $\beta 0$. It can be concluded that the subjective evaluation obtained in the psychoacoustic experiments may have reflected these differences in the spectral structures of the stimuli.

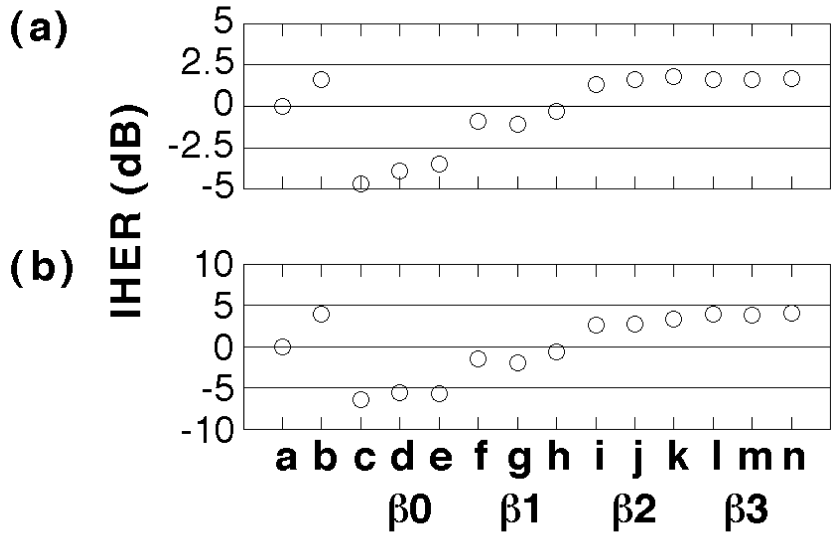

Figure 2.10: (a) IHERs of the stimuli used in the psychoacoustic experiment 1, (b) IHERs of the stimuli used in the psychoacoustic experiment 2

2.5 Discussion

The experimental results suggest that the frequency characteristics of the fluctuation sequences could be a significant factor that influences the voice quality of sustained vowels. Compared with other stimulus groups, the stimuli which were characterized by spectral $1/f$ sequences tended to be evaluated as more similar to the reference stimuli. This result could be attributable to the fact that the frequency characteristics of the spectral $1/f$ sequences were similar to those of the amplitude or period sequences employed in the reference stimuli.

The spectral $1/f$ power law is often observed in a variety of natural phenomena, including fluctuation sequences obtained from biomedical signals, such as heart rate fluctuation [311, 335, 350, 354, 360]. Although the spectral $1/f^{\beta}$ sequences are generally classified as nonstationary processes under the condition of $1 < \beta$, quasi-stationarity is observed in spectral $1/f$ sequences [308, 309, 310, 311, 317, 334, 335, 350, 354]. The mean and the mean square value of spectral $1/f^{\beta}$ sequences are subject to the following relationship under the condition of $1 < \beta < 3$ and $\beta = 2H + 1$ [308, 334, 350].

$$\begin{aligned} \langle v(rt) \rangle &= r^{H} \langle v(t) \rangle \\ \langle v^2(rt) \rangle &= r^{2H} \langle v^2(t) \rangle, \end{aligned} \tag{2.4}$$

where

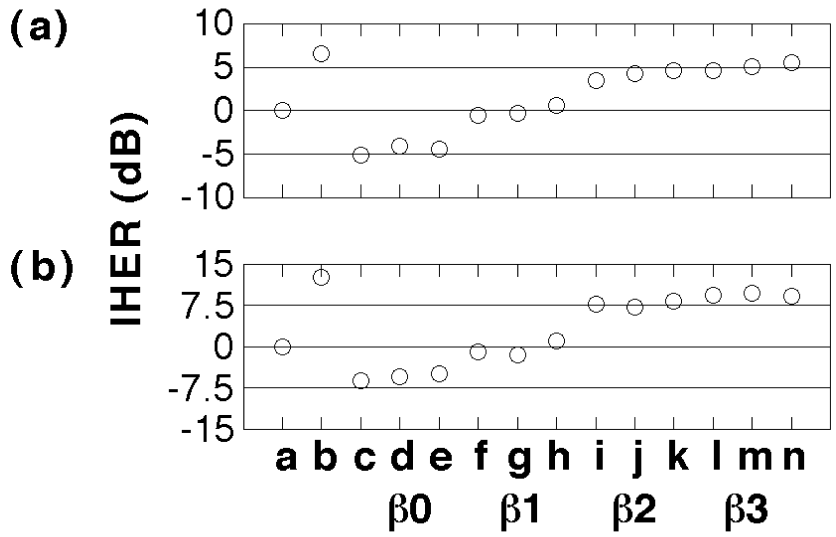

Figure 2.11: (a) IHERs of the stimuli used in the psychoacoustic experiment 3, (b) IHERs of the stimuli used in the psychoacoustic experiment 4

$v(t)$ is a spectral $1/f^{\beta}$ sequence,

r is the resolution factor of the sequence,

H is the Hurst exponent,

$\langle \cdot \rangle$ is the expectation operator.

Since the mean and the variance derived from equation 2.4 are considered to change as the resolution factor r changes, $1/f^{\beta}$ fluctuations are classified as nonstationary processes with the condition of $1 < \beta$. However, the mean and the variance of $1/f$ sequences derived from equation 2.4 with the condition of $\beta \to 1$, namely $H \to 0$, are statistically invariant even though the resolution factor r changes. This nature of spectral $1/f$ sequences is known as self-similarity which guarantees the quasi-stationarity of the sequences [308, 334, 350].

Taking these discussion into consideration, the amplitude and period sequences, modeled as either spectral $1/f^2$ or $1/f^3$ sequences, are classified as nonstationary processes. The nonstationary changes in the loudness and the pitch by spectral $1/f^2$ or $1/f^3$ sequences are considered to have caused differences in the voice quality from the reference stimulus which was characterized by the stationary amplitude and period sequence. Taking account of the result of the speech analysis in which almost all fluctuation sequences would be acceptable as stationary processes, spectral $1/f^2$ and $1/f^3$ sequences could not be appropriate models for the amplitude and period sequences of normal sustained vowels.

Compared with the shortcoming of these nonstationary sequences, spectral $1/f$ sequences are considered to be more appropriate models for the amplitude and period sequence due to their quasi-stationarity. The psychoacoustic experiment also subjectively supported the validity of the models. Considering the result of the speech analysis in which it was suggested that the fluctuation sequences be approximated as spectral $1/f$ processes, it could be concluded that the models of spectral $1/f$ sequences are potentially useful choices for the amplitude and period sequences of normal sustained vowels.

In this study, gross approximation of the frequency characteristics of fluctuation sequences was the only focus. However, the result of the frequency analysis suggested the possibility of much finer models. For example, it was found that the frequency characteristics in the high frequency region tended to increase to a greater level than in the case of the spectral $1/f$ power law, as shown in figure 2.5. It will be of interest to investigate how finer models influence speech perception. In addition, it will also be of interest to examine how the perceptual differences are caused by the value of β as it gradually changes from one. These issues are currently being investigated by the authors for developing more detailed models of the fluctuation sequences.

Some previous studies have also developed a model of the period sequences for normal sustained vowels from the viewpoint of its frequency characteristics [74, 81]. These models have represented the frequency characteristics of period sequences by auto-regressive (AR) [81] or by an auto-regressive moving-average (ARMA) [74] form of a digital filter. Similar to the results presented in this paper, these studies also indicate that the frequency characteristics of period sequences would be a key factor in determining the voice quality of sustained vowels. It can be suggested that the gradual decreasing characteristics found in the high frequency region are a feature of the period sequences which could be related to the voice quality of normal sustained vowels.

The decreasing characteristics of amplitude as well as period sequences were also pointed out by other previous studies [18, 70, 71, 78, 79]. A speech analysis of normal sustained vowels showed decreases of high frequency components of both amplitude and period sequences [78, 79]. These fluctuation sequences were extracted from three Japanese vowels /a/, /i/, and /u/ phonated by one adult male speaker. Such decreasing characteristics were also suggested by the speech analysis of period sequences obtained from normal sustained vowels /a/ [18]. Comparisons with pathological cases also indicated that the decreasing characteristics of amplitude and period sequences might be one of the features of normal sustained vowels [70, 71, 79].

From the above two viewpoints, namely, (1) speech perception caused by fluctuation

sequences and (2) speech analysis of fluctuation sequences, it is conceivable that the decreasing characteristics are one of the features of the fluctuation sequences obtained from normal sustained vowels. Such decreasing characteristics could be a significant factor in the voice quality of normal sustained vowels.

2.6 Conclusions

The present study statistically showed several aspects of the amplitude and period sequences of normal sustained vowels. The speech analysis indicated that both fluctuation sequences would be approximated as spectral $1/f$ sequences in terms of their frequency characteristics.

In addition, the psychoacoustic experiments indicated that voice quality of sustained vowels appeared to be influenced by the frequency characteristics of fluctuation sequences. The results of the present study may provide useful information for exploring the speech perception caused by fluctuation sequences as well as for developing their appropriate models with the aim of enhancing the voice quality of synthesized sustained vowels.

Chapter 3

Analysis and perception of waveform fluctuation

3.1 Introduction

In addition to the amplitude and period fluctuation described in Chapter 2, normal sustained vowels also always contain cyclic changes of the waveform itself even during their most steady parts. Although the size of such waveform fluctuation obtained from normal sustained vowels is quite small, earlier investigations indicate that the waveform fluctuation is also considered to be a key factor in the naturalness of human speech [24, 69, 193, 194, 218]. It can be therefore assumed that incorporating the waveform fluctuation, which is artificially generated with an appropriate model of the fluctuation, may potentially contribute to the enhancement of voice quality of synthesized speech.

In order to obtain the know-how to implement LPC-vocoder-based high quality speech synthesis systems, this study particularly investigated some statistical characteristics of the waveform fluctuations that were extracted from the residual signals obtained by LPC inverse filtering. Based on the results of the analysis, the author attempted to newly develop an advanced model for appropriately generating the waveform fluctuations that could be incorporated into the source signals of synthesized speech for the enhancement of its voice quality. For the evaluation of the effectiveness of the developed model, a series of psychoacoustic experiments were carried out. Not only were the experimental results subjectively discussed, but they were also discussed with an objective evaluation method devised in this study.

3.2 Speech analysis

This section describes the method that was employed for the extraction of the waveform fluctuations from the residual signals obtained by LPC inverse filtering. Some statisti-

cal characteristics of the waveform fluctuations obtained from normal sustained vowels were investigated to obtain useful information for developing an appropriate model of the waveform fluctuation.

3.2.1 Speech samples

Ten male subjects between 22 and 26 years of age who did not suffer from any laryngeal disorders were selected in order to obtain normal sustained vowels. Each subject was requested to phonate the sustained vowel /a/ as steadily as possible in a soundproof anechoic room (Rion, audiometry room) toward an electret condenser microphone (Audio-technica, AT822) at a distance of about 15 cm from the mouth. The sustained vowels were directly recorded onto a hard disk of a personal computer (Apple, Macintosh Quadra 800) by way of a microphone mixer (Mackie, microseries 1202-VLZ), a low pass filter (8th order Bessel characteristic), and an analog-to-digital converter (Digidesign, audiomedia II). The sampling rate and quantization level were 44.1 kHz and 16 bits, respectively. The cut-off frequency of the low pass filter was set to be 5 kHz. The subjects phonated the vowels at the pitch and loudness that was comfortable. The duration of the phonation was requested to be approximately ten seconds. All records contained a steady portion of at least 512 pitch periods lasting over approximately four seconds, in which the mean pitch period was found to range from 7.6 msec to 9.1 msec. The calculated mean pitch period of all speech samples was 8.3 msec. The sound pressure level (SPL) was also measured by a precision noise meter using the C weighting condition (Brüel & Kjær, type 2209), which was placed about 15 cm from the mouth [83, 84, 412]. Measured SPL ranged from 80 dB to 86 dB for all subjects. The gain of the microphone mixer was adjusted for each subject for an optimal recording level. Twenty speech samples were taken per subject. Two hundred speech samples in total (20 utterances $\times$ 10 subjects) were obtained [5, 20].

3.2.2 Extraction of waveform fluctuation

At first, the residual signals of the speech samples were obtained by the LPC inverse filtering [110, 158, 186, 212, 213, 214, 215, 225, 226, 232]. The order of the filter was set at 40 for the condition of a 44.1 kHz sampling rate. The filter order was determined by a visual inspection which ascertained that the frequency characteristics of the filter in this condition appropriately represented four dominant formants below the cut-off frequency of the low pass filter of 5 kHz. The coefficients of the LPC inverse filter were not altered during the analysis. This condition was based on the assumption that the characteristics of the vocal tract filter for normal sustained vowels do not substantially change during

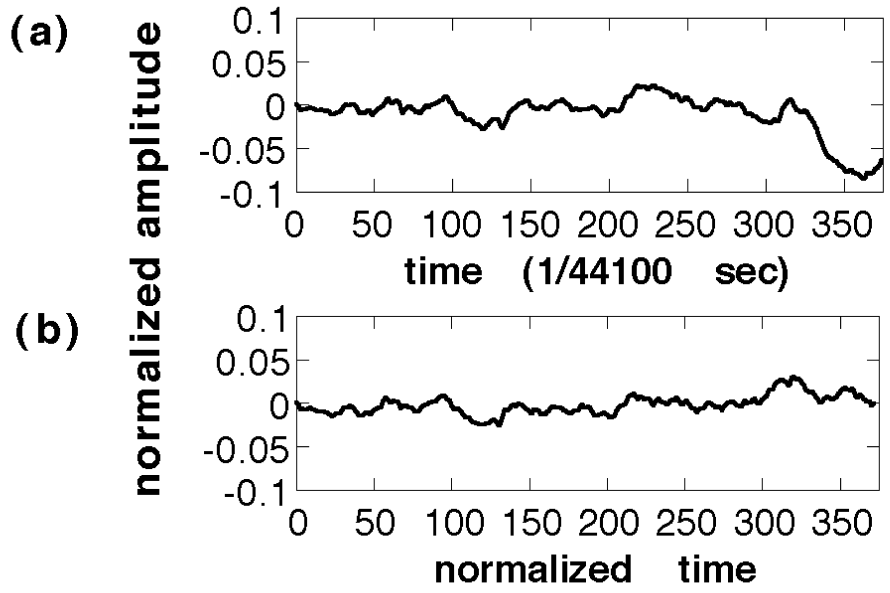

Figure 3.1: (a) Waveform fluctuation extracted without linear matching process: Adjacent pitch periods for the comparison were 372 and 376 points, respectively. (b) Waveform fluctuation obtained by using linear matching process: Pitch periods were normalized to be 376 points.

the phonation. From the resulting residual signals, waveform fluctuations were simply extracted as the differences in waveform between two adjacent pitch periods. In order to improve the accuracy of the matching process between the two adjacent pitch periods, the residual signals were in advance processed with a -12 dB/oct low pass filter. This low-pass-filtered signals were labeled as source signals. Since the updated pitch period length observed in the sustained vowels always changed even for the two adjacent pitch periods due to their inherent period fluctuations, the magnitude of the waveform fluctuation, which was extracted without an appropriate normalization process, unnecessarily increased at around the end point of the pitch period [46, 47]. Figure 3.1 (a) exemplifies this problem with the waveform fluctuation extracted without a particular normalization process. This was the case when the initial point was only used as an anchor point in the comparison of the two adjacent pitch periods. In order to mitigate this problem, a linear matching process was simply employed as a pilot method for preprocessing the extraction of the waveform fluctuations. As shown in figure 3.1 (b), the waveform fluctuation obtained with the linear matching process could successfully reduce the spurious trend often observed at around the end point of the pitch period as shown in figure 3.1 (a).

For the actual extraction process, this study normalized the period length of the source

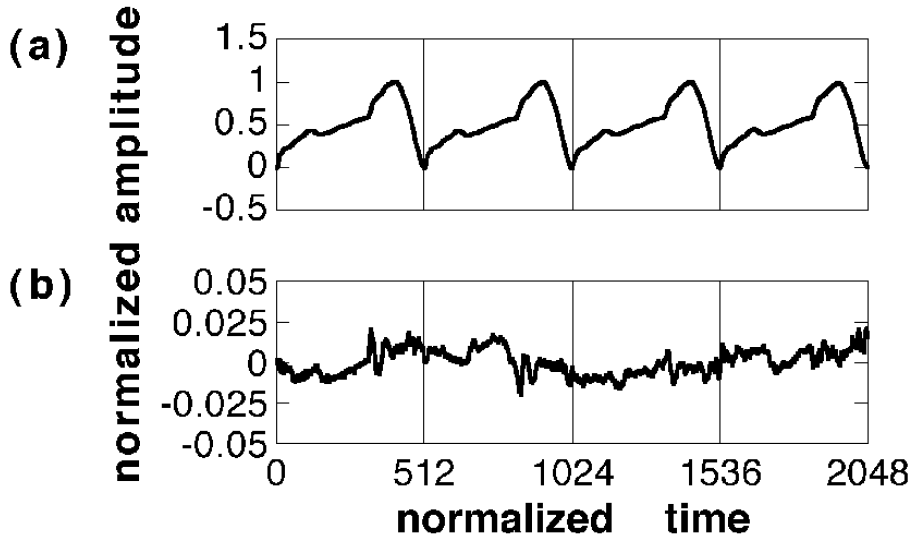

Figure 3.2: (a) Example of normalized source signal of a sustained vowel /a/, and (b) its waveform fluctuation

signals to 512 points by oversampling technique. Figure3.2 (a) illustrates the first four pitch periods of such a normalized source signal, which can be denoted as $e[512m+n], m = 0, 1, \ldots, 127, n = 0, 1, \ldots, 511$ for 128 pitch periods. Waveform fluctuations extracted from this normalized source signal are shown in figure 3.2 (b). The definition of the waveform fluctuation can be represented in equation 3.3.

$$\begin{aligned} f[512m+n] &= e[512(m+1)+n] - e[512m+n] \\ m &= 0, 1, \ldots, 126, \\ n &= 0, 1, \ldots, 511. \end{aligned} \tag{3.1}$$

3.2.3 Average power of the waveform fluctuation

The size of the waveform fluctuation is considered to be a significant factor that influences the voice quality of sustained vowels. An excessively large waveform fluctuation may result in rough voice quality [24]. The power of the waveform fluctuation to the source signals was statistically analyzed in order to investigate its valid size for normal sustained vowels. The results showed that the calculated power of all waveform fluctuations fell in similar values, with an average power of approximately -25 dB.

3.2.4 Frequency characteristics of waveform fluctuation

Although the waveform fluctuations appeared to be complicated random signals, their complexity was not apparently the same degree as in the case of white noise as inferred from figure 3.2 (b). In order to investigate their general characteristics for normal sustained vowels, the frequency analysis was performed using the 512-point fast Fourier transform (FFT) with a Hamming window [96, 97, 98, 104, 105, 108, 109, 409]. The results indicated that the gross approximation of the frequency characteristics was subject to the spectral $1/f^{\beta}$ power law that met the condition of $\beta \simeq 2$ [308, 309, 310, 311, 317, 334, 335, 350, 354], although the details might deviate from this approximation. This tendency was consistently observed among all waveform fluctuations. Figure 3.3 shows an example of the frequency characteristic of waveform fluctuations, whereas the average frequency characteristic among all waveform fluctuations is shown in figure 3.4. The value of the exponent β was estimated by the least-squares line fitting method [108, 109, 409], where the exponent β was equivalent to the gradient of the fitted line in the frequency characteristics plotted in the log-log scale. Since the value of β tended to center at approximately two for all the waveform fluctuations, the results of frequency analysis concluded that waveform fluctuations could be modeled as spectral $1/f^2$ processes for a preliminary model.

3.3 Psychoacoustic experiments

In order to explore the influence of the frequency characteristics of waveform fluctuations on speech perception, a series of psychoacoustic experiments were conducted. The purpose of the experiments was to investigate how the differences in their frequency characteristics caused the subjective differences in the voice quality of synthesized sustained vowels.

3.3.1 Stimuli

Stimuli for the psychoacoustic experiments consisted of sustained vowels /a/ produced by a partial autocorrelation (PARCOR) synthesizer [110, 111, 116, 120, 121]. They were individually characterized by the different types of waveform fluctuations. The filter coefficients of the PARCOR synthesizer were derived from two typical speech samples whose waveform fluctuations showed the representative characteristics of all the cases investigated in this study. For the extraction of the set of PARCOR coefficients, conditions 1 and 2 in the psychoacoustic experiments employed the speech sample obtained from subject A.H., whereas conditions 3 and 4 in the psychoacoustic experiments employed the

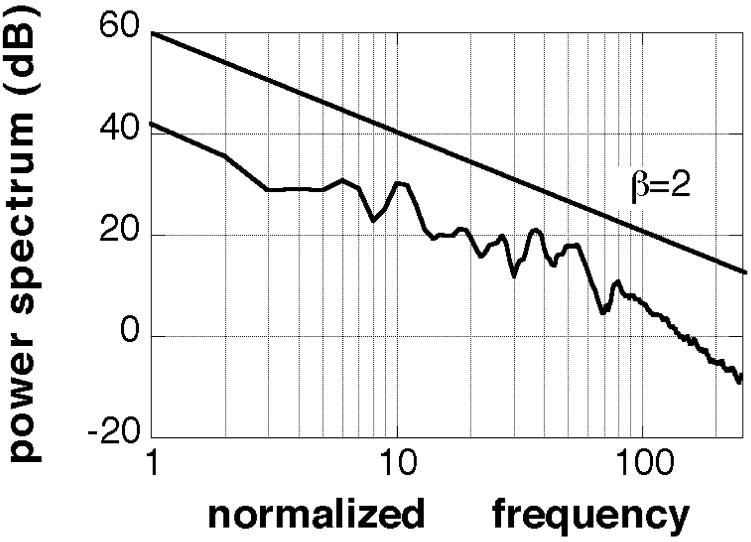

Figure 3.3: Example of the frequency characteristics of waveform fluctuation

speech sample obtained from subject H.S. The amplitude and period fluctuations derived from the speech samples were incorporated into the stimuli utilized in conditions 2 and 4 in order to examine how these fluctuations influenced the evaluation. On the other hand, conditions 1 and 3 did not utilize the amplitude and period fluctuations.

The filter order of the PARCOR synthesizer was set at 40 for the condition of a 44.1 kHz sampling rate. The filter order was determined by a visual inspection which ascertained that the frequency characteristics of the PARCOR filter in this condition appropriately represented four dominant formants below the cut-off frequency of the low pass filter of 5 kHz. The coefficients of the PARCOR filter were not altered during the synthesis. This condition was based on the assumption that the characteristics of the vocal tract filter for normal sustained vowels do not substantially change during phonation. Each stimulus consisted of 128 pitch periods. The duration of each stimulus was approximately one second. A linearly increasing or decreasing gate function whose duration was 10 msec was employed at the beginning and the end of each stimulus in order to prevent undesirable clicking-like sounds.

Fourteen stimuli labeled from "a" to "n" were produced for each condition. Stimulus "a" employed the waveform fluctuations originally obtained from the speech samples. Although stimulus "a" was not the speech sample itself, its voice quality was considered

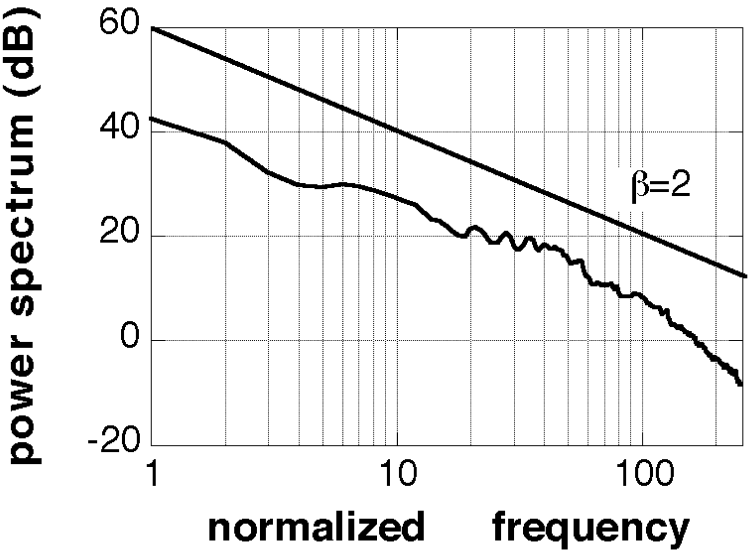

Figure 3.4: Mean frequency characteristics of waveform fluctuations obtained from all speech samples

to reflect the characteristics of the waveform fluctuations of the speech sample. Since stimulus "a" was used as a reference stimulus in evaluating all stimuli, including stimulus "a" itself, it was also labeled the reference stimulus. Comparisons between the reference stimulus and stimulus "a" constituted the control for the experiment. Stimulus "b" was produced without incorporating waveform fluctuations. The objective for using this stimulus was to examine whether waveform fluctuation was an important factor for speech perception. In addition, four stimulus groups were produced, each consisting of three stimuli. The three stimuli of each stimulus group employed the waveform fluctuations whose frequency characteristics were classified in the same category, while the waveform fluctuations themselves were different from each other. This is because randomization used for producing the waveform fluctuations was different. The four stimulus groups were labeled as "$\beta 0$", "$\beta 1$", "$\beta 2$", and "$\beta 3$", according to the frequency characteristics of the fluctuation sequences. Stimulus group $\beta 0$, consisting of stimulus "c", "d", and "e", employed spectral $1/f^0$ waveform fluctuations (white noises). Stimulus group $\beta 1$, which consisted of stimulus "f", "g", and "h", employed spectral $1/f$ waveform fluctuations. Stimulus group $\beta 2$, which consisted of stimulus "i", "j", and "k", employed spectral $1/f^2$ waveform fluctuations. Stimulus group $\beta 3$, consisting of stimulus "l", "m", and

"n", employed spectral $1/f^3$ waveform fluctuations. The mean power of the waveform fluctuations was set at -25 dB to the source signals.

The value of the exponent β in the spectral $1/f^\beta$ noises for each stimulus group was considered to be a rather preliminary choice. The integer values from zero to three were employed, since there was no *a priori* knowledge about the relationship between speech perception and the values of β. Using three stimuli in each stimulus group aimed to examine whether or not the perceptual effects were categorized by the values of β.

The stimulus group $\beta 0$ employed normal Gaussian white noises as its waveform fluctuations. On the other hand, all of the waveform fluctuations employed for synthesizing the stimulus group $\beta 1$, $\beta 2$, and $\beta 3$ were artificial fractional Brownian motions generated by the mid point displacement method [310, 311, 334]. The procedure for generating the waveform fluctuation is illustrated in figure 3.5. In this figure, $p[m], m = 0, 1, 2$ denotes pitch periods. The period length was chosen to be 512 points. In each pitch period, the mid point displacement method recursively generated spectral $1/f^2$ waveform fluctuations. The algorithm of the mid point displacement method is defined in equation 3.2. It is a kind of a wavelet reconstruction process for which a triangular hat function is chosen the scaling function [318, 319, 334, 428].

$$\begin{aligned} c_{j+1}(2k) &= c_j(k) \\ c_{j+1}(2k+1) &= (c_j(k) + c_j(k+1))/2 + d_j(k) \\ k &= 0, 1, \ldots, 2^j, \end{aligned} \tag{3.2}$$

where

$c_j(k)$ is the scaling function coefficients at level j,

$d_j(k)$ is the wavelet coefficients at level j.

Gaussian white noise is simply chosen the wavelet coefficients $d_j(k)$ for generating fractional Brownian motions by the mid point displacement method. It is necessary that the mean and standard deviation of $d_j(k)$ should be zero and $2^{-jH}\sigma_0$, respectively, where σ_0 represents the standard deviation of the wavelet coefficients at level 0. Since fractional Brownian motion is required to meet both the conditions $0 < H < 1$ and $2H = \beta - 1$, the Hurst exponent H must be $1/2$ for generating random fractals characterized by $\beta = 2$ [317, 334, 350].

Figure 3.6 (b) shows an example of the waveform fluctuation produced by the mid point displacement method. Figure 3.6 (a) shows an example of the source signal in which the waveform fluctuations shown in figure 3.6 (b) were incorporated. In this case, the source signal was defined as the sum of $\bar{e}(n)$ and $g(n)$, where $\bar{e}(n)$ is the average

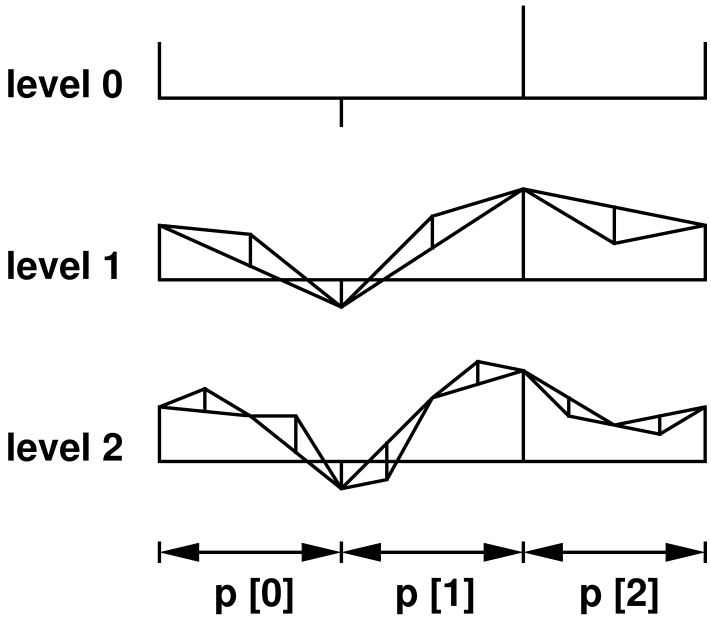

Figure 3.5: Procedure for generating waveform fluctuation using the mid point displacement method

waveform through 128 pitch periods of the speech sample. The waveform fluctuations were, therefore, represented as,

$$\begin{aligned} h[512m+n] &= g[512(m+1)+n] - g[512m+n] \\ m &= 0, 1, \ldots, 126, \\ n &= 0, 1, \ldots, 511. \end{aligned} \tag{3.3}$$

It appeared that the mean frequency characteristic of the waveform fluctuation shown in figure 3.6 (b) was almost subject to the spectral $1/f^2$ power law as illustrated in figure 3.7. Due to the adequate frequency characteristics, the degree of the complexity visually inspected in the artificial waveform fluctuations was similar to that of the waveform fluctuations obtained from the human speech samples.

3.3.2 Subjects

Twenty subjects consisting of twelve males and eight females participated in the experiment. Their age ranged from 20 to 26 years. None of them were experienced in psychoacoustic experiments. All reported having no hearing problems.

3.3.3 Procedures

All stimuli were synthesized using a personal computer (Apple, Macintosh Quadra 800). The stimuli were passed through a digital-to-analog converter (Digidesign, audiomedia

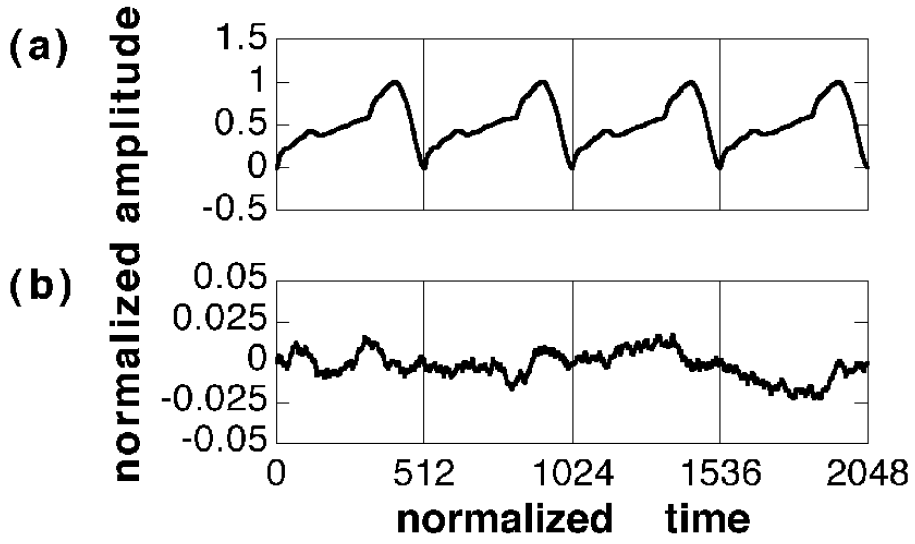

Figure 3.6: (a) Normalized source signal which contains waveform fluctuation generated by the proposed method, and (b) its waveform fluctuation

II), followed by a low pass filter (8th order Bessel characteristic). The sampling rate and quantization level were 44.1 kHz and 16 bits, respectively. The cut-off frequency of the low pass filter was set to be 10 kHz. The stimuli were presented through a monitor speaker (Denon, USC-101), which was attached to a pre-main amplifier (Sansui, AU-α 507XR). The speaker was placed 1.5 meters in front of the subject in a soundproof anechoic room (Rion, audiometry room). The SPL of the stimuli was set to be 65 dB upon presentation.

Each subject took part in all four conditions of the experiment individually. Each condition consisted of fourteen paired-comparison trials to evaluate the similarity between the preceding stimulus A and the succeeding stimulus B. The reference stimulus, namely stimulus "a", was always chosen stimulus A, while stimulus B was one of the fourteen stimuli produced for each condition. The order of the presentation in regard to stimulus B was randomized.

Stimulus A and B were presented to the subject twice in the form of an AB pair, as illustrated in figure 3.8. There was a one-second silent interval between stimulus A and B and a two-second silent interval between the first and second AB pair. For the judgment period, a six-second interval was given to the subject after listening to the two AB pairs.

The subject was asked to judge whether or not the voice quality of stimulus B was perceived as the same as that of stimulus A. They were forced to select one of the following three choices, (1) "same", (2) "undecided", or (3) "different". The three-point scale aimed

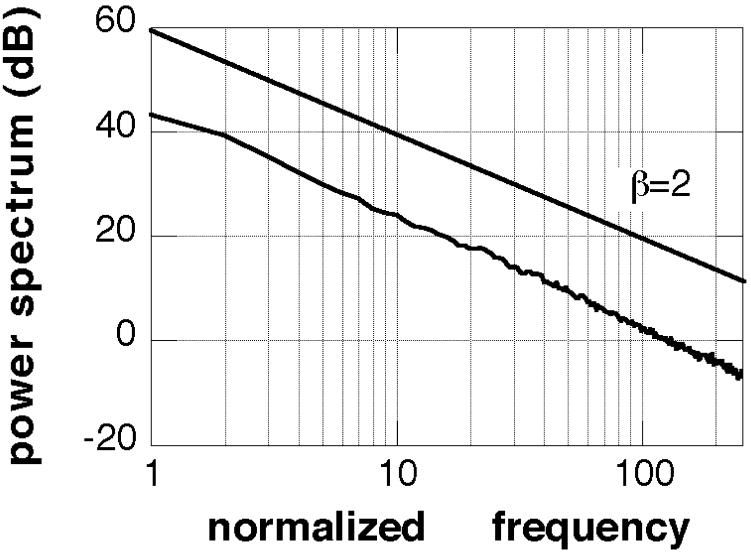

Figure 3.7: Frequency characteristics of waveform fluctuation shown in Fig.6 (b)

to examine whether or not the subject could correctly distinguish the voice quality between a stimulus of the artificially produced waveform fluctuation and one from the waveform fluctuation obtained from the speech sample. For the above reason, the five or seven-point scale, which is conventionally used to grade the differences in the voice quality, was not employed [402, 403, 404]. In order to compare the experimental results, the three choices were translated into the numerical measure named similarity. This measure was defined as (1) 100 %, (2) 50 %, and (3) 0 % corresponding to the three choices.

3.3.4 Results of conditions 1 and 2

The results of conditions 1 and 2 are summarized in figure 3.9 (a) and (b), respectively. The similarity of each stimulus is represented by an open circle which shows the average of the results over all subjects. It appeared that the control stimulus "a" and the stimulus group $\beta 2$ tended to be evaluated as more similar to the reference stimulus than the other stimuli throughout conditions 1 and 2. These results indicated that most of the subjects found it difficult to distinguish the voice quality of the stimulus group $\beta 2$ from that of the reference stimulus, namely stimulus "a". The effect of amplitude and period fluctuation examined in condition 2 did not substantially influence this tendency.

As for the other stimuli, most of the subjects reported that hissing noises were clearly

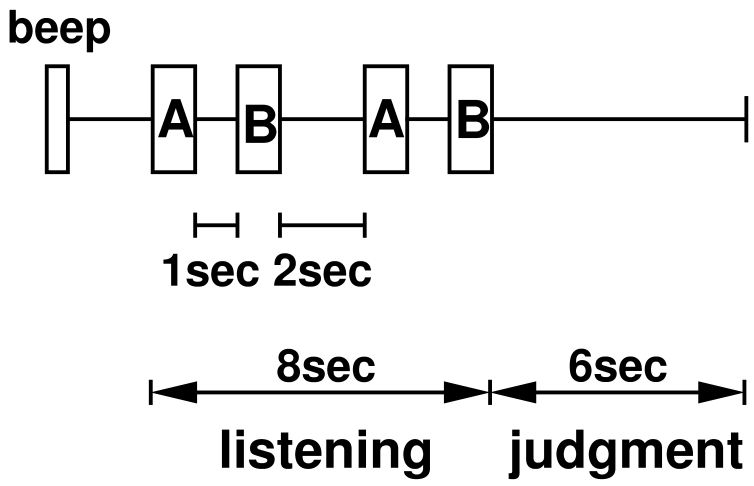

Figure 3.8: Procedure of paired-comparison test

perceived in the stimulus groups $\beta 0$ and $\beta 1$, while a buzzer-like quality was perceived under stimulus "b" and the stimulus group $\beta 3$. Such features were not, however, perceived in the case of the reference stimulus. Therefore, these differences could be a subjective clue to discriminate between the reference stimulus and these stimuli.

3.3.5 Results of conditions 3 and 4

The results of conditions 3 and 4 are summarized in figure 3.10 (a) and (b), respectively. The similarity of each stimulus is represented by an open circle which shows the average of the results over all subjects. It appeared that control stimulus "a" and the stimulus group $\beta 2$ tended to be evaluated as more similar to the reference stimulus than the other stimuli throughout conditions 3 and 4. These results indicated that most of the subjects found it difficult to distinguish the voice quality of the stimulus group $\beta 2$ from that of the reference stimulus, namely stimulus "a". The effect of the amplitude and period fluctuations examined in condition 4 did not substantially influence this tendency.

As for the other stimuli, most of the subjects reported that hissing noises were clearly perceived in the stimulus group $\beta 0$ and $\beta 1$, while the buzzer-like quality sound was perceived under stimulus "b" and for the stimulus group $\beta 3$. Such features were not, however, perceived in the case of the reference stimulus. Therefore, these differences could be a subjective clue in the discrimination process between the reference stimulus and these stimuli.

Since the similarity of stimulus "b" commonly tended to be judged as low throughout all the conditions, it can be consequently concluded that the waveform fluctuations play significant roles in the speech perception of sustained vowels as mentioned in the previ-

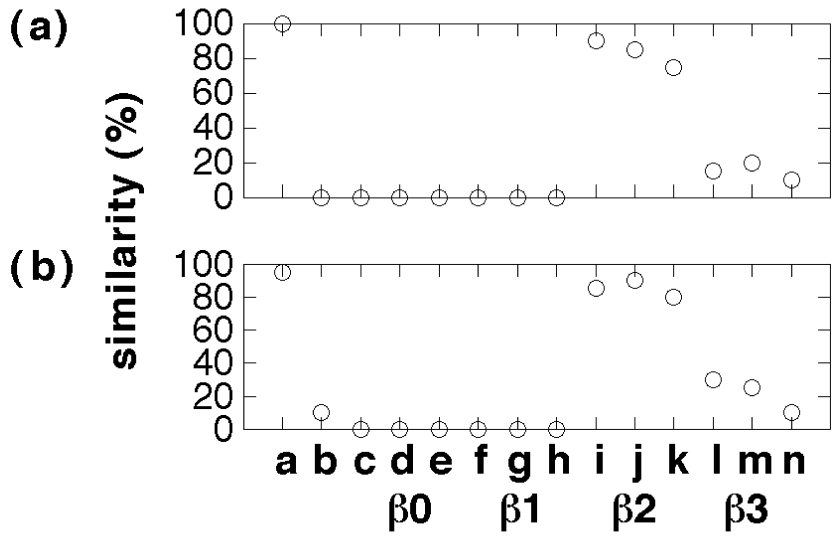

Figure 3.9: Results of (a) condition 1, and (b) condition 2: One of the speech samples obtained from subject A.H. was used for the stimuli. In condition 1, each stimulus contained neither pitch period fluctuation nor amplitude fluctuation. In condition 2, these fluctuations were incorporated into each stimulus.

ous literature [24, 69, 193, 194, 218]. The results of the other stimuli suggest that the differences in the frequency characteristics of the waveform fluctuations can significantly influence speech perception. High similarity between the stimulus group $\beta 2$ and the reference stimulus can be attributable to the similarity in the frequency characteristics of the fluctuation sequences between stimulus group $\beta 2$ and the reference stimulus.

Since there were no large differences in similarity among all the stimuli of the stimulus group $\beta 2$, the randomization for producing the fluctuation sequences could have little effect on speech perception. In addition, it seemed that the similarity of the stimulus group $\beta 2$ could not be significantly influenced by the amplitude and period fluctuations. The similarity of the stimulus group $\beta 2$ was consistently evaluated high throughout conditions 1 and 2 as well as during conditions 3 and 4.

3.4 Objective evaluation on the psychoacoustic experiments

An objective evaluation of the stimuli, all of which were employed in the psychoacoustic experiments, was performed to judge the validity of the subjective evaluation. As earlier studies indicate, the voice quality can be objectively classified by the index which measures the ratio of the inharmonic spectra energy to the harmonic spectral energy [42, 43, 59, 60].

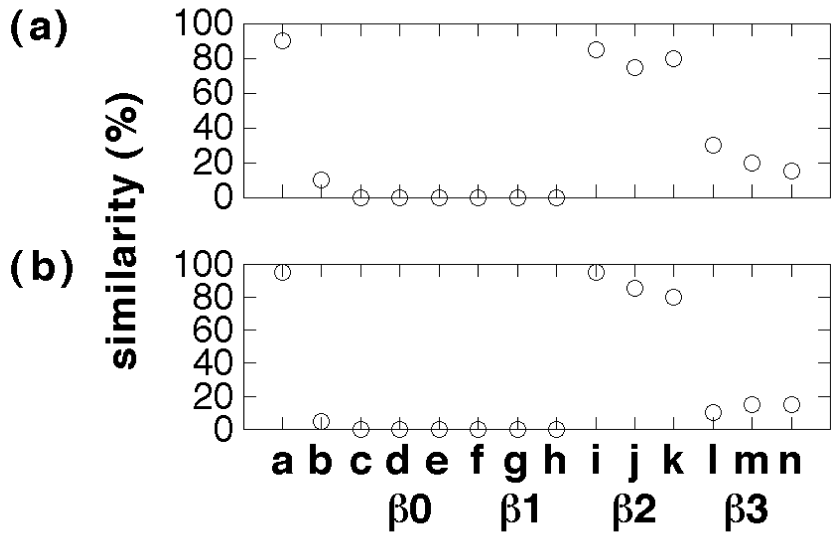

Figure 3.10: Results of (a) condition 3, and (b) condition 4: One of the speech samples obtained from subject H.S. was used for the stimuli. In experiment 3, each stimulus contained neither pitch period fluctuation nor amplitude fluctuation. In experiment 4, these fluctuations were incorporated into each stimulus.

Such an index becomes large in the case of purely periodic speech signals that may be perceived as quite buzzer-like. On the other hand, the index becomes small in the case of aperiodic speech signals that may be perceived as being a rough voice quality.

This study devised a new objective index named Inharmonic Energy Ratio (IHER) for investigating the voice quality of the stimuli utilized in the psychoacoustic experiments. As defined in equation 3.4, in which f denotes frequency and $S(f)$ the power spectrum, the proposed index measured the ratio of the energy obtained from inharmonic spectral regions to that of the harmonic spectral region. The definition of IHER is also illustrated in figure 3.11, in which h_j represents jth harmonic frequency, $a = (h_{j-1} + h_j)/2$, $b = (h_j + h_{j+1})/2$, $c = (a + h_j)/2$, and $d = (h_j + b)/2$. The inharmonic spectral regions were defined as the bands which met the conditions $a \leq f < c$ or $d \leq f < b$, while the other region represented as $c \leq f < d$ was defined as a harmonic region. This definition came from the fact that it was hard to accurately estimate the purely harmonic frequencies when the speech signals contained a certain aperiodicity, such as the amplitude and period fluctuation. In that situation, the separation of the harmonic spectral components was quite difficult, especially in the high frequency region. Although IHER does not require the estimation of harmonic frequencies, it was still difficult to separate the harmonic spectral

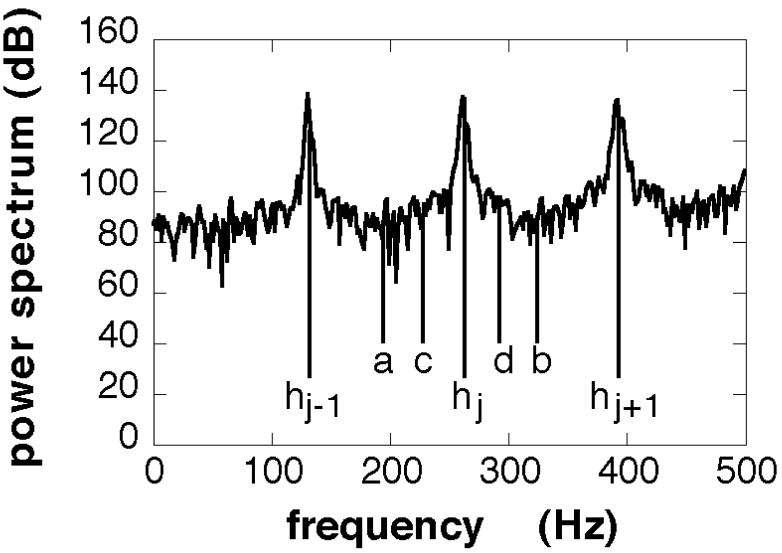

Figure 3.11: Definition of Inharmonic Energy Ratio (IHER)

regions from the inharmonic spectral regions in the high frequency region. Therefore, this study applied the proposed index to only the low frequency region under 10th harmonic frequency. The power spectra utilized in the calculation of the index was estimated by 4096-point FFT.

$$\mathrm{IHER}(j) = 10\log\frac{\sum_{f\geq c}^{f<d} S(f)}{\sum_{f\geq a}^{f<c} S(f) + \sum_{f\geq d}^{f<b} S(f)} \tag{3.4}$$

3.4.1 Results of conditions 1 and 2

The calculated indices for all the stimuli used in conditions 1 and 2 are summarized in figure 3.12 (a) and (b), respectively. It appeared that the indices of both the control stimulus "a" and the stimulus group $\beta 2$ tended to be similar throughout conditions 1 and 2. On the other hand, the indices of stimulus "b" and the stimulus group $\beta 3$ were larger than that of the control stimulus "a". As for the stimulus group $\beta 0$ and $\beta 1$, the indices were smaller than that of the control stimulus "a". Incorporating the amplitude and period fluctuations examined in condition 2 did not substantially influence this tendency.

3.4.2 Results of conditions 3 and 4

The calculated indices for all the stimuli used in conditions 3 and 4 are summarized in figure 3.13 (a) and (b), respectively. It appeared that the indices of both the control stimulus "a" and the stimulus group $\beta 2$ tended to be similar throughout conditions 1 and

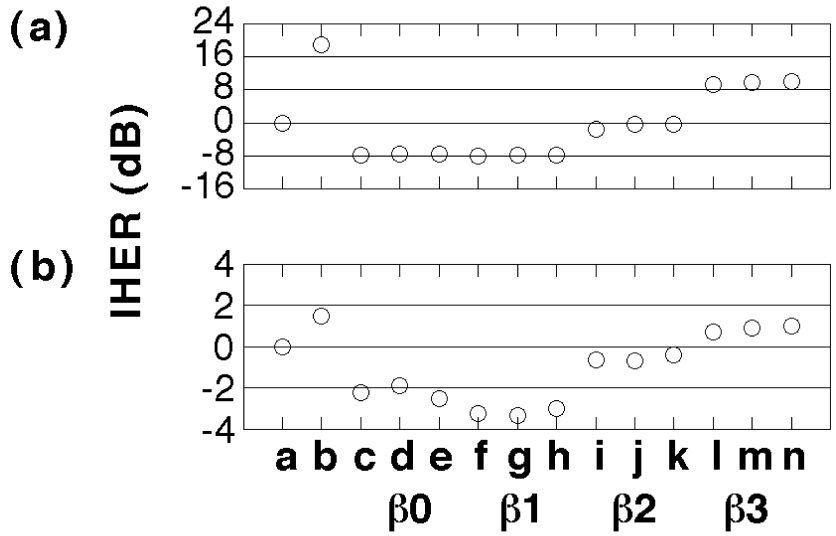

Figure 3.12: (a) IHERs of the stimuli used in the psychoacoustic experiment 1. (b) IHERs of the stimuli used in the psychoacoustic experiment 2

2. On the other hand, the indices of stimulus "b" and the stimulus group $\beta 3$ were larger than that of the control stimulus "a". As for the stimulus group $\beta 0$ and $\beta 1$, the indices were smaller than that of the control stimulus "a". Incorporating the amplitude and period fluctuations examined in condition 4 did not substantially influence this tendency.

Since the indices ranged at similar values for stimulus group $\beta 2$ and the control stimulus "a", it would appear that these stimuli were categorized in the same group in terms of their spectral characteristics. Since the indices of stimulus "b" and the stimulus group $\beta 3$ were larger than that of the control stimulus "a" throughout all the conditions, the harmonic spectral components for these stimuli were considered to be more dominant in their spectral structures. On the other hand, results suggested that the inharmonic spectral components were dominant for the stimulus groups $\beta 0$ and $\beta 1$. It can be concluded that the subjective evaluation obtained in the psychoacoustic experiments may have reflected the differences in these spectral structures of the stimuli.

3.5 Discussion

The results of the psychoacoustic experiments indicate that the stimuli, which incorporated $1/f^2$ waveform fluctuations generated by the mid point displacement method, were evaluated as having almost the same voice quality as the reference stimulus. On the other hand, the other stimulus groups as well as stimulus "b" did not show such a tendency.

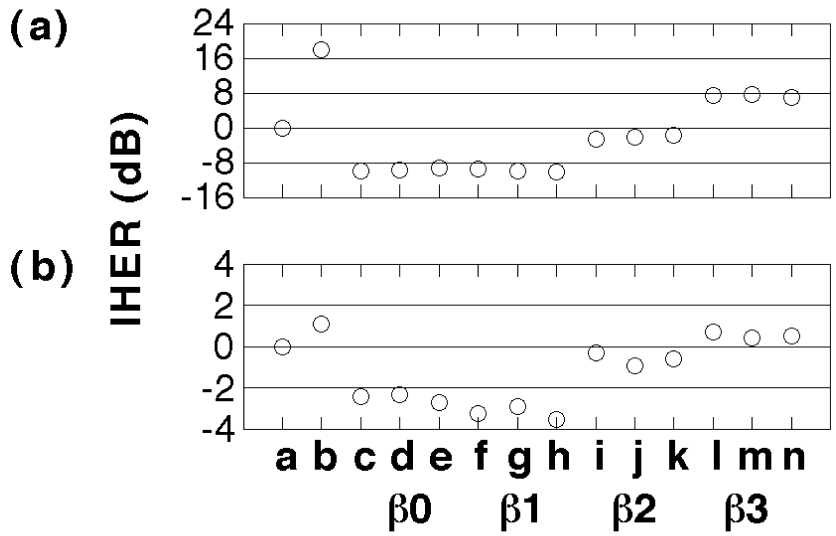

Figure 3.13: (a) IHERs of the stimuli used in the psychoacoustic experiment 3. (b) IHERs of the stimuli used in the psychoacoustic experiment 4

This study therefore reached the conclusion that the waveform fluctuations, which was appropriately incorporated into synthesized sustained vowels, would significantly enhance the voice quality of synthesized speech. The frequency characteristics of the waveform fluctuations are considered to be a key factor in the speech perception.

Earlier studies also investigated the characteristics of the waveform fluctuations in terms of the mechanism of speech production [193, 194, 218]. It is reported that the naturalness of synthesized speech was enhanced when the frequency characteristics of the waveform fluctuations were modeled as being subject to spectral -6 dB/oct $(1/f^2)$ decay [64, 65, 66, 83, 84, 193, 194, 218]. Since the present study particularly analyzed the spatial frequency of normalized waveform fluctuations, its results can not be directly compared with those of the earlier studies even though both the studies resulted similar conclusions with regard to the frequency characteristics of waveform fluctuations. However, it can at least be concluded that the present study as well as the earlier studies indicates that the frequency characteristics of the waveform fluctuations should be adequately designed for enhancing the voice quality of synthesized sustained vowels

As illustrated in the results of conditions 2 and 4 in the psychoacoustic experiments, this study also reached the conclusion that waveform fluctuations would be a contributing factor in the enhancement of the voice quality of synthesized speech even though there were no amplitude or period fluctuations. This conclusion supports the findings of several

earlier studies which reported that waveform fluctuation was a dominant factor which significantly influenced the naturalness of stimuli even though period fluctuations were not incorporated [24, 69, 162, 163].

3.6 Conclusions

This study investigated how modeling the waveform fluctuations in terms of their frequency characteristics was effective for enhancing the naturalness of synthesized sustained vowels. The psychoacoustic experiments investigated how the differences in the frequency characteristics of the waveform fluctuations influenced the voice quality of synthesized speech. Based on the statistical characteristics of waveform fluctuations, this study developed a practical technique for generating waveform fluctuations. The proposed technique may be an important component that can be used for the development of LPC-vocoder-based high quality speech synthesis systems.

Chapter 4

Analysis and perception of the random fractalness in the source signals of LPC vocoder

4.1 Introduction

Random fractal modeling of natural objects is known as a powerful method to artificially mimicking their naturalness. An example is found in mountainous random fractal terrain generated by a computer graphics technology [350, 310]. As inferred from this typical example, since the random fractal feature of natural objects is considered to be a key factor in their naturalness, measuring their random fractalness can potentially be a useful viewpoint in objectively investigating their naturalness, which is often discussed with only a certain subjective method [367, 309].

The authors have investigated several types of fluctuations observed in sustained vowels phonated by normal speakers in terms of their random fractalness [1, 3]. Until now, period fluctuation, amplitude fluctuation, and waveform fluctuation have been the focuses. The analyses performed in our study indicate that random fractalness is a property that is inherently observed in the above three fluctuations. From a series of psychoacoustic experiments, the study has also reached a conclusion that the incorporation of the fluctuations, which are modeled as being random fractals, is effective for enhancing the naturalness of sustained vowels synthesized by a source-filter speech production model, such as an LPC (linear predictive coding) vocoder [116, 110]. It can be indicated that the random fractalness is potentially a key factor in the naturalness of human speech.

Besides the inclusion of these fluctuations, several earlier studies have indicated that the waveforms themselves of source signals utilized in an LPC vocoder must be appropriately modeled for enhancing the voice quality of synthesized vowels [201, 256, 208, 193, 207, 206]. In order to increase the efficiency of the process of speech synthesis, the con-

ventional LPC vocoder often employs oversimplified source signals for synthesizing vowel speech. For this purpose, an impulse train is most commonly used [116, 110]. The degradation of the voice quality, which is perceived as buzzer-like or machine-like unnatural voice quality, is considered to be caused mainly by such an oversimplification of the source signal [250, 249, 206, 217]. In order to obtain the know-how for designing more realistic source signals for developing LPC-vocoder-based high quality speech synthesis systems, this study statistically investigated the differences between the impulse train source and the source signals obtained from human speech, in terms of the random fractal feature of the source signals.

Theoretically, the source signals of an LPC vocoder are defined as being characterized by the spectral -6 dB/oct decay in the frequency domain when both -12 dB/oct glottal vibration characteristics and $+6$ dB/oct mouth radiation characteristics are simultaneously taken into consideration [116, 110]. From the viewpoint of random fractal theory, since the spectral -6 dB/oct decay is equivalent to the spectral $1/f^2$ characteristics, the source signals of an LPC vocoder are potentially classified as Brownian motions [350, 367, 330]. This study particularly investigated such random fractalness of the source signals with regard to their time domain characteristics by using a multi-resolution analysis method based on Schauder expansion [334, 462]. In addition, using the statistical characteristics of the source signals obtained by the Schauder analysis, this study newly developed a technique for enhancing the degraded voice quality of the synthesized speech produced by an LPC vocoder using a conventional impulse train. The psychoacoustic experiments were performed in order to examine the effectiveness of the proposed technique.

4.2 Speech analysis

This section describes the method that was employed for extracting the source signals of an LPC vocoder. Some statistical characteristics of the source signals were investigated by a multi-resolution analysis method based on Schauder expansion.

4.2.1 Speech samples

Ten male subjects between 22 and 26 years of age who did not suffer from any laryngeal disorders were selected in order to collect normal sustained vowels. Each subject was requested to phonate the sustained vowel /a/ as steadily as possible in a soundproof anechoic room (Rion, audiometry room) toward an electret condenser microphone (Audio-technica, AT822) at a distance of about 15 cm from the mouth. The sustained vowels were

directly recorded onto a hard disk of a personal computer (Apple, Macintosh Quadra 800) by way of a microphone mixer (Mackie, microseries 1202-VLZ), a low pass filter (8th order Bessel characteristic), and an analog-to-digital converter (Digidesign, audiomedia II). The sampling rate and quantization level were set at 44.1 kHz and 16 bits, respectively. The cut-off frequency of the low pass filter was set to be 5 kHz. The subjects phonated the sustained vowels at the pitch and loudness that were comfortable. The duration of the phonation was requested to be approximately ten seconds. All records contained a steady portion of at least 128 pitch periods lasting over approximately four seconds, in which the mean pitch period was found to range from 7.6 ms to 9.1 ms. The calculated mean pitch period of all speech samples was 8.3 ms. The sound pressure level (SPL) was also measured by a precision noise meter using the C weighting condition (Brüel & Kjær, type 2209), which was placed about 15 cm from the mouth. Measured SPL ranged from 80 dB to 86 dB for all subjects. The gain of the microphone mixer was adjusted for each subject for an optimal recording level. Twenty speech samples were taken per subject. Two hundred speech samples in total (20 utterances $\times$ 10 subjects) were obtained.

4.2.2 Extraction of the source signals of the speech samples

Residual signals of the speech samples were obtained by an LPC inverse filtering technique [116, 110]. A Hamming window was adopted in the calculation. The window length was set to be the same as the duration of the 128-pitch-period steady portion of each speech sample. The order of the filter was set at 40 for the condition of a 44.1 kHz sampling rate. The filter order was determined by a visual inspection which ascertained that the frequency characteristics of the filter in this condition appropriately represented four dominant formants below the cut-off frequency of the low pass filter of 5 kHz. The coefficients of the LPC inverse filter were not altered during the calculation. This condition was based on the assumption that the characteristics of the vocal tract filter for normal sustained vowels do not substantially change during the phonation.

Theoretically, the source signals of an LPC vocoder are defined as being characterized by the spectral -6 dB/oct decay in the frequency domain [116, 110]. This study, therefore, extracted the source signals of the speech samples by filtering the resulting residual signals, for which the spectral envelope was approximately flat, with a -6 dB/oct low-pass filter represented as,

$$H(z) = \frac{1}{1 - 0.98z^{-1}}. \tag{4.1}$$

Figure 4.1 shows examples of: (a) a speech sample and (b) its source signal obtained by the above described procedure.

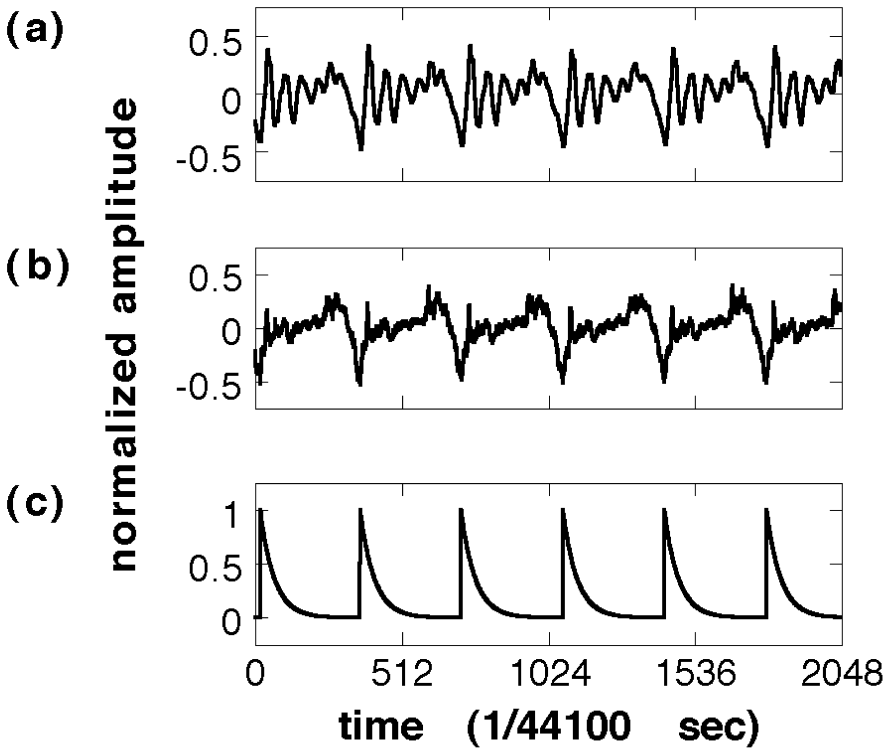

Figure 4.1: Example of (a) a speech sample, (b) its source signal, and (c) impulse train source

As mentioned in the previous section, it is preferred for the conventional LPC vocoder to employ oversimplified source signals instead of the actual ones for the convenience of efficiently performing the process of speech synthesis [116, 110]. Figure 4.1 (c) shows an example of such an oversimplified source signal that was made from an impulse train. In order to meet the conditions of the spectral -6 dB/oct characteristics for the source signals of the LPC vocoder, the impulse train, for which the spectral envelope is characterized as flat, was processed with the low-pass-filter defined in equation 4.1. The resultant signal was labeled the impulse train source in this study.

4.2.3 Schauder analysis of the source signals

Random processes, for which the spectral characteristics are represented as $1/f^{\beta}$ types, can be classified as random fractal under the condition of $1 < \beta < 3$ [350, 330]. Since spectral -3β dB/oct decay is equivalent to a spectral $1/f^{\beta}$ characteristic, the source signals characterized by the spectral -6 dB/oct decay can be potentially classified as spectral $1/f^{2}$ random fractals. The signals that meet this condition are called Brownian motion, which is known as a typical random fractal [350, 330]. This study particularly investigated

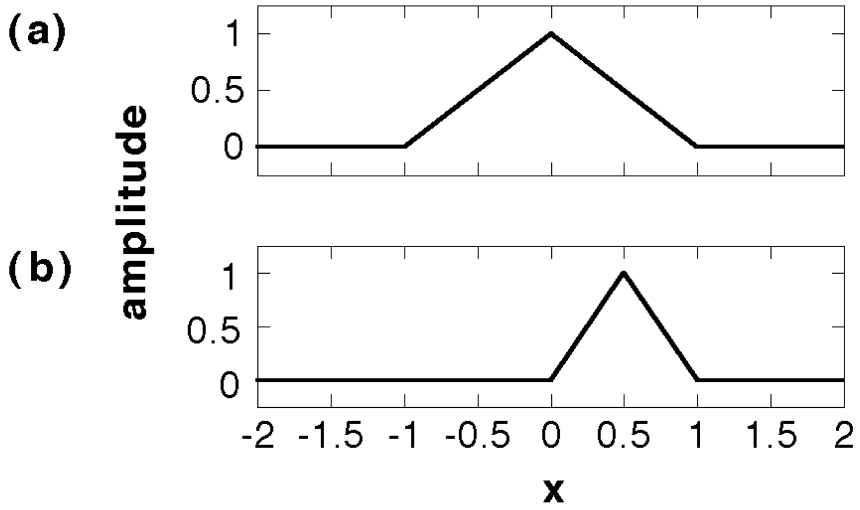

Figure 4.2: (a) Hat functions: $\phi(x)$ and (b) $\psi(x)$.

such random fractalness of the source signals in terms of their time domain characteristics by using a multi-resolution analysis method based on Schauder expansion [462]. This is because time domain analysis is also required for validating the random fractalness of a signal even if its frequency characteristics meet the condition of random fractals [329]. The Schauder analysis is a method to investigate the time domain characteristics of random fractals. It is known that the reconstruction procedure of the Schauder expansion is a well-known algorithm called mid point displacement method for generating artificial random fractals [428, 334].

The major feature of the time domain characteristics of random fractals is statistical self-similarity [334, 309, 310]. Schauder analysis can quantitatively examine the self-similarity of random fractals by comparing the standard deviations of the Schauder coefficients between two adjacent resolution levels. Random fractals are particularly characterized by having a constant scaling factor with regard to the standard deviations. In other words, the statistical size of the Schauder coefficients constantly varies when the resolution level changes; the larger the resolution level, the smaller the Schauder coefficients, and *vice versa*. It is known that the scaling factor is determined according to the value of the exponent β [428, 334]. Taking advantage of this property, the mid point displacement method generates random fractals.

Schauder expansion can be defined by its recursive form represented as in equation

4.2, in which j denotes the resolution level for representing the signal $f(x)$.

$$f_j(x) = f_{j-1}(x) + g_{j-1}(x). \tag{4.2}$$

$f_j(x)$ and $g_j(x)$ are represented as,

$$\begin{aligned} f_j(x) &= \sum_k c_j(k)\phi(2^j x - k), \\ g_j(x) &= \sum_k d_j(k)\psi(2^j x - k), \end{aligned} \tag{4.3}$$

where

$c_j(k)$ is the coefficients of the scaling function $\phi(x)$ at level j,

$d_j(k)$ is Schauder coefficients at level j.

The scaling function $\phi(x)$ is a hat function defined as,

$$\phi(x) = \begin{cases} 1+x & , -1 \leq x < 0 \\ 1-x & , 0 \leq x \leq 1 \\ 0 & , \text{otherwise.} \end{cases} \tag{4.4}$$

The following relationship exists between $\phi(x)$ and $\psi(x)$ where k denotes an integer value.

$$\begin{aligned} \phi(x) &= \sum_k p(k)\phi(2x - k) \\ \psi(x) &= \sum_k q(k)\phi(2x - k) \end{aligned} \tag{4.5}$$

where $p(-1) = 1/2$, $p(0) = 1$, $p(1) = 1/2$, $q(1) = 1$, and the other all $p(k)$ and $q(k)$ are required to be zero [428]. Figure 4.2 shows the waveforms of $\phi(x)$ and $\psi(x)$.

The algorithms of the Schauder decomposition and reconstruction are defined in equation 4.6 and 4.7, respectively. Figure 4.3 illustrates these algorithms.

$$\begin{aligned} c_{j-1}(k) &= c_j(2k), \\ d_{j-1}(k) &= c_j(2k+1) \\ &- (c_j(2k) + c_j(2k+2))/2. \end{aligned} \tag{4.6}$$

$$\begin{aligned} c_j(2k) &= c_{j-1}(k), \\ c_j(2k+1) &= d_{j-1}(k) \\ &+ (c_{j-1}(k) + c_{j-1}(k+1))/2. \end{aligned} \tag{4.7}$$

As mentioned earlier, the mid point displacement method is equivalent to the reconstruction algorithm defined in equation 4.7. Random fractals can be generated by the algorithm under the condition in which the standard deviation of $d_j(k)$, denoted

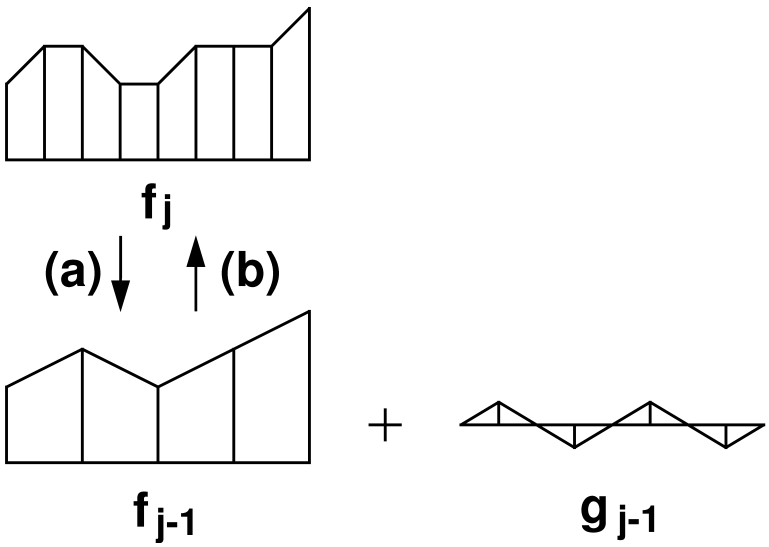

Figure 4.3: Schauder expansion: (a) decomposition algorithm and (b) reconstruction algorithm

as σ_j, is subject to the following relationship with the conditions of $0 < H < 1$ and $2H = \beta - 1$ [334].

$$\sigma_j = 2^{-H}\sigma_{j-1}. \tag{4.8}$$

Schauder decompositions of the source signals shown in figure 4.1 (b) and (c) were calculated. The resulting Schauder coefficients are shown in figure 4.4 and 4.5, respectively. The highest resolution level was 9 due to 512-point Schauder decompositions. Thus, the highest Schauder level was determined to be 8. In figure 4.4 and 4.5, the Schauder coefficients of resolution level j are represented in the location of $2^j \leq x < 2^{j+1}$.

As shown in figure 4.4, the Schauder coefficients of the source signal obtained from the speech samples were found to be tend to decrease in the size when the resolution level became larger. On the other hand, as shown in figure 4.5, large coefficients still existed even at large resolution levels in the case of the impulse train source .

Some statistical characteristics of the Schauder coefficients were analyzed. Figure 4.6 shows the maximum absolute value of the Schauder coefficients calculated at each resolution level. There were substantial differences at large resolution levels between the source signals obtained from the speech samples and the impulse train source due to the large Schauder coefficients observed in the impulse train source. From this result, it can be indicated that the discontinuities of the waveforms are large particularly for the impulse train source.

Figure 4.7 shows the standard deviation of the Schauder coefficients calculated at each resolution level, where the extremely large coefficients (5 % of the distributions) were

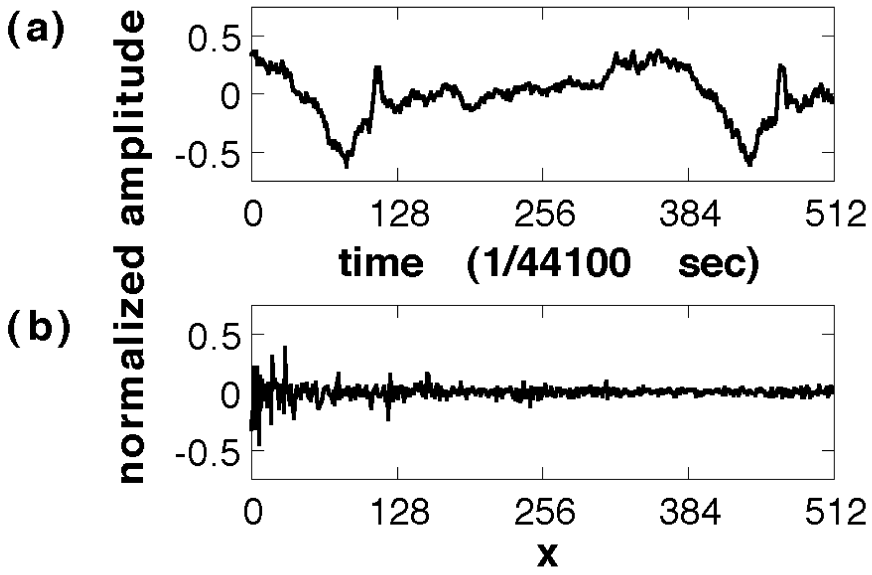

Figure 4.4: Example of Schauder decomposition: (a) source signal obtained from a speech sample and (b) its Schauder decomposition

removed from the calculation in order to reduce the influence of such large coefficients on the estimation of the standard deviation. The value of the exponent β in each level, which was related to the scaling factor 2^{-H} under the condition of $2H = \beta - 1$, was also obtained using equation 4.9 deduced from equation 4.8.

$$\beta_j = 1 - 2\log_2(\sigma_j/\sigma_{j-1}) \tag{4.9}$$

As indicated in figure 4.8, the source signals obtained from the speech samples were found to be potentially classified as Brownian motion at large resolution levels, since the values of the exponent β_j ranged around two. However, this tendency was not observed at small resolution levels. Unlike actual Brownian motion, there was a certain limitation in the range of the Schauder coefficients at the small resolution levels for the source signals due to their periodicity. This property of the source signals resulted in rather small values of the exponent β at small resolution levels.

As regards the impulse train source, it did not meet the time domain condition of Brownian motion even at large resolution levels, in spite that its frequency characteristic is subject to the condition of Brownian motion, namely the spectral $1/f^2$ decay. As shown in figure 4.8, the value of the exponent β calculated from the impulse train source varied in an unstable manner and had a range up to approximately eight. Consequently, the Schauder analysis performed in this study reached a conclusion that the source signals

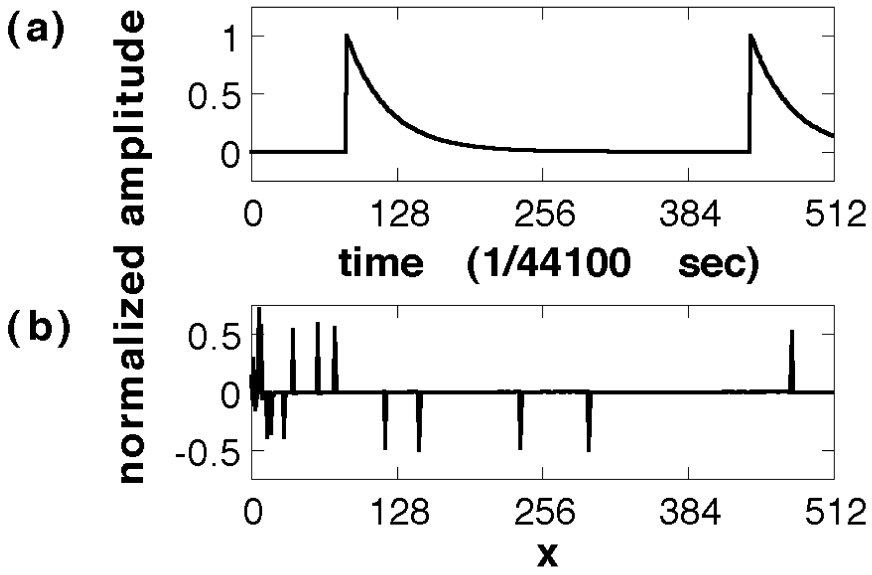

Figure 4.5: Example of Schauder decomposition: (a) impulse train source and (b) its Schauder decomposition

obtained from human vowel speech can potentially be modeled as Brownian motion at large resolution levels, whereas the impulse train source can not be classified as Brownian motion in terms of its time domain characteristics.

4.3 Psychoacoustic experiments

In order to investigate the influence of the differences in the above-described time domain characteristics of the source signals on speech perception, a series of psychoacoustic experiments were carried out. The purpose of the experiments was to investigate how the differences in the time domain characteristics caused the subjective differences in the voice quality of synthesized speech.

4.3.1 Modification technique for impulse train source

Some of the stimuli employed in the psychoacoustic experiments were synthesized with a technique that was newly developed for the modification of the impulse train source. Based on the statistical characteristics of the maximum values of the Schauder coefficients described in the previous section, the extremely large Schauder coefficients found in the impulse train source at high resolution levels were appropriately clipped in order to reduce the large discontinuities of the waveform. Since this process was basically the same as low pass filtering, it had an undesirable side-effect of decreasing the high frequency

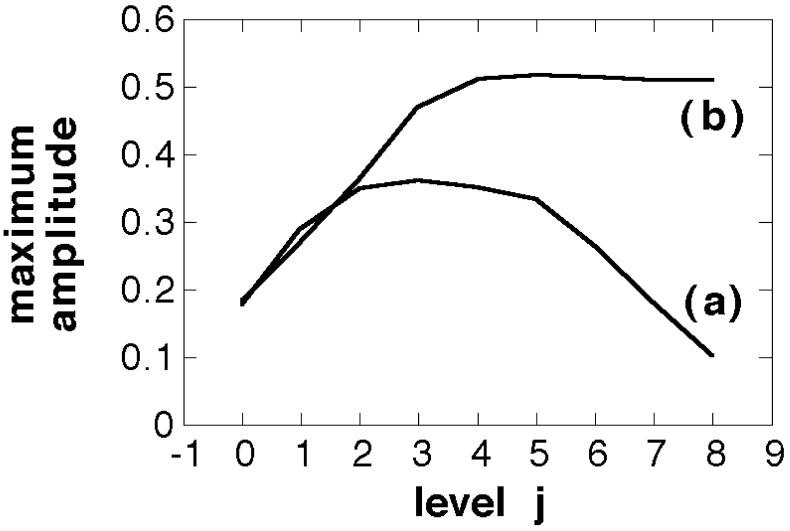

Figure 4.6: Maximum absolute value of the Schauder coefficients at each level: (a) source signals of speech samples, (b) impulse train source

power of the impulse train source, which must be subject to the -6 dB/oct decay (a spectral $1/f^2$ characteristic). Figure 4.9 shows an impulse train source after clipping the large Schauder coefficients and its frequency characteristic calculated by a 512-point fast Fourier transform (FFT) with a Hamming window. In order to eliminate this shortcoming, the proposed technique employed a method called random fractal interpolation [320]. Specifically, this technique incorporated Brownian motion, which was artificially generated by the mid point displacement method, into the impulse train source at high resolution levels for improving the frequency characteristics. Figure 4.10 (a) shows an example of impulse train source processed by the random fractal interpolation. As shown in figure 4.10 (b), the random fractal interpolation could almost restore the -6 dB/oct decay (a spectral $1/f^2$ characteristic) required for the source signals of an LPC vocoder.

Figure 4.11 shows how the discontinuities of the impulse train source varies depending on the process of the proposed technique. Each of the three figures represents the first-order differential of a pitch period of the impulse train source. These were calculated from figure 4.5 (a), figure 4.9 (a), and figure 4.10 (a), respectively. The comparison between figure 4.11 (a) and figure 4.11 (b) indicates that the discontinuities of the impulse train source was remarkably softened by the clipping procedure of the proposed technique. Furthermore, figure 4.11 (c) shows that the high frequency components are artificially incorporated by the random fractal interpolation.

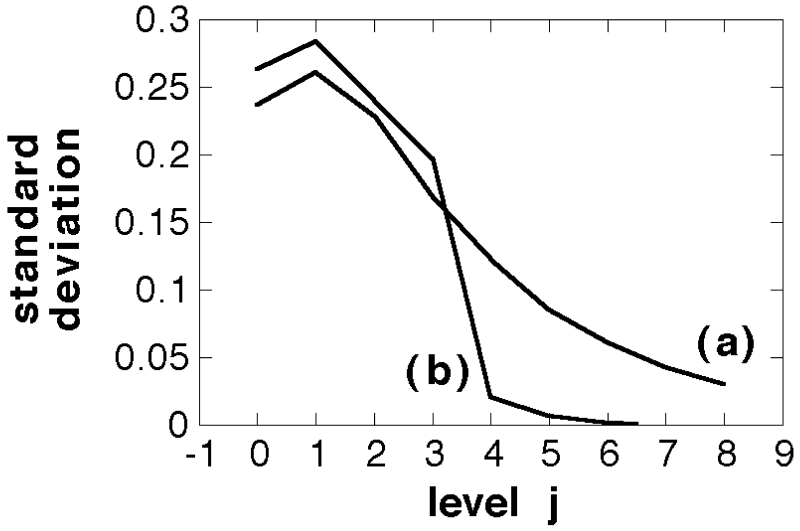

Figure 4.7: Standard deviation of the Schauder coefficients at each level: (a) source signals of speech samples, (b) impulse train source

4.3.2 Stimuli

The psychoacoustic experiments consisted of three different conditions. Five stimuli labeled from "a" to "e" were produced for each condition. They were independently characterized by the different types of source signals. Stimulus "a" utilized the original source signal obtained from a speech sample. Stimulus "b" utilized the conventional impulse train source. Stimulus "c" utilized the impulse train source modified only by the clipping procedure of the proposed technique. Stimuli "d" and "e" utilized the impulse train source modified by the full use of the proposed technique, including the random fractal interpolation. Two stimuli "d" and "e" were generated in order to investigate how the difference in the randomization for the random fractal interpolation influenced the experimental results.

Stimuli used in condition 1 were sustained vowels /a/ produced by a partial autocorrelation (PARCOR) synthesizer [116, 110]. One of the speech samples utilized in the Schauder analysis was selected as a speech sample, from which the essential parameters for the speech synthesis were taken. The filter order of the PARCOR synthesizer was set at 40 for the condition of a 44.1 kHz sampling rate. The filter order was determined by a visual inspection which ascertained that the frequency characteristics of the PARCOR filter in this condition appropriately represented four dominant formants below the cut-

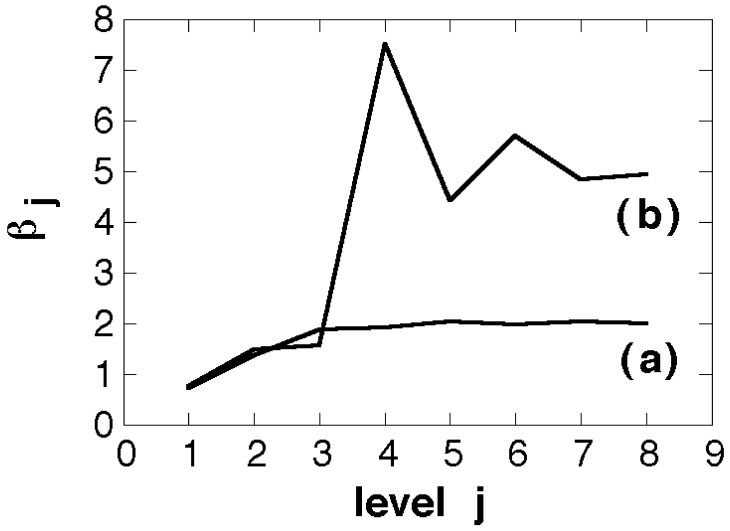

Figure 4.8: β_j estimated from the changes of standard deviations between contiguous two levels: (a) source signals of speech samples, (b) impulse train source

off frequency of the low pass filter of 5 kHz. The coefficients of the PARCOR filter were not altered during the synthesis. This condition was based on the assumption that the characteristics of the vocal tract filter for normal sustained vowels do not substantially change during the phonation. Any types of fluctuations that are found in the maximum amplitude, period length, and waveform itself were not incorporated into the stimuli [1, 3], since these fluctuations can potentially influence the evaluation with regard to the naturalness of the synthesized sustained vowels. For each of the stimuli, the waveform of a pitch period was cyclically concatenated to form the 128-pitch-period source signal, which was equivalent to approximately one second. The period length was set at 366 points for all the stimuli. Stimulus "a" particularly used the waveform of a pitch period taken from the actual source signal. It is known that this stimulus possesses the human-like natural voice quality [253].

Stimuli used in condition 2 consisted of a continuous male speech: "The US one hundred dollar bill was recently replaced with a new version." Stimuli used in condition 3 consisted of a continuous female speech: "I know a lot of students are planning to travel to the States this summer." These two sentences were speech samples recorded by a male and a female native English speaker [142]. The duration of each sentence was approximately 4 seconds. A PARCOR synthesizer was utilized for producing all the stimuli of each condition except for stimuli "a", for which the original speech samples

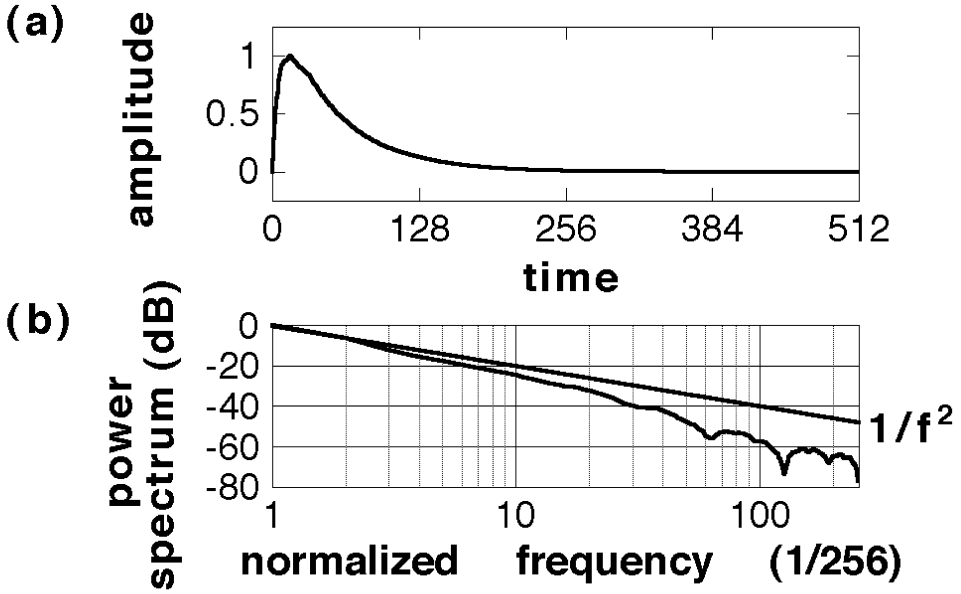

Figure 4.9: (a) Impulse train source after clipping the large Schauder coefficients and (b) its frequency characteristic calculated by a 512-point FFT

themselves were simply employed. The sampling rate and quantization level of these sentences were chosen to be 11.025 kHz and 16 bits, respectively. The speech samples, which were originally recorded on an audio CD (compact disk) at a 44.1 kHz sampling rate, were downsampled by an editing program (Macromedia, Sound Edit 16). For the analysis and synthesis procedure of the PARCOR synthesizer, the frame length was set at 23.2 ms (256 sampling points) and the overlap between two adjacent frames was set at 5.8 ms (64 sampling points). The filter order of the PARCOR synthesizer was set at fourteen for the condition of an 11.025 kHz sampling rate. In the analysis stage of the PARCOR synthesis procedure, the autocorrelation function was calculated in each frame for the determination of a voiced/unvoiced decision. For voiced parts, the waveform of a pitch period, for which the period length was set to be 256 sampling points, was concatenated to form source signals with the technique of a zero-padding or a truncation according to the actual period length of each frame. With regard to the source signals of unvoiced parts, Brownian motion was adopted, which was made from Gaussian white noise with the -6 dB/oct low-pass-filter defined in equation 4.1.

4.3.3 Subjects

Ten male subjects participated in the experiment. Their age ranged from 23 to 26 years. None of them were experienced in psychoacoustic experiments. All reported having no

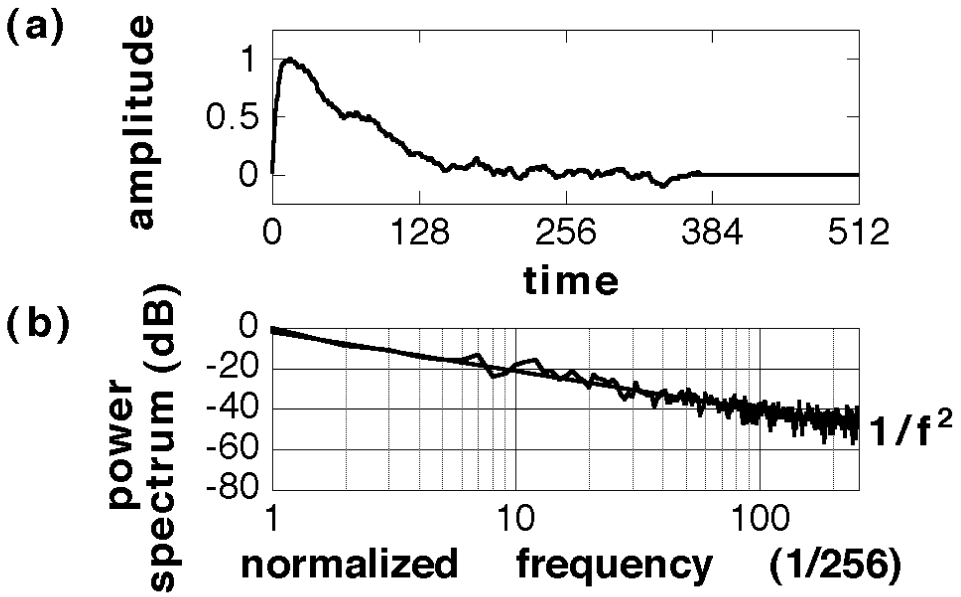

Figure 4.10: (a) Impulse train source after the random fractal interpolation and (b) its frequency characteristic calculated by a 512-point FFT

hearing problems.

4.3.4 Procedures

All stimuli were generated by a personal computer (Apple, PowerBook 5300c). The stimuli were presented through a monitor headphone (Sony, MDR-CD370). The SPL of the stimuli was set at a comfortable level for each subject upon presentation.

Each subject took part in the experiment individually, under all three conditions. Each condition consisted of twenty paired-comparison trials to evaluate the naturalness between the preceding stimulus A and the succeeding stimulus B. All combinations were investigated, including the reverse order of the five stimuli produced for each condition. The order of the presentation of the in-total twenty combinations was randomized. The subject was asked to answer which stimulus in each paired-comparison trial was perceived as having a more human-like natural voice quality. They were forced to make a judgment by selecting either stimulus A or B.

Stimulus A and B were presented to the subject twice in the form of an AB pair as illustrated in figure 4.12. There was a one-second silent interval between stimulus A and B, and a two-second silent interval between the first and second AB pair. For the case of condition 1, the listening phase lasted approximately eight seconds, due to using the one-second stimuli, while approximately twenty seconds for condition 2 and 3 due to the

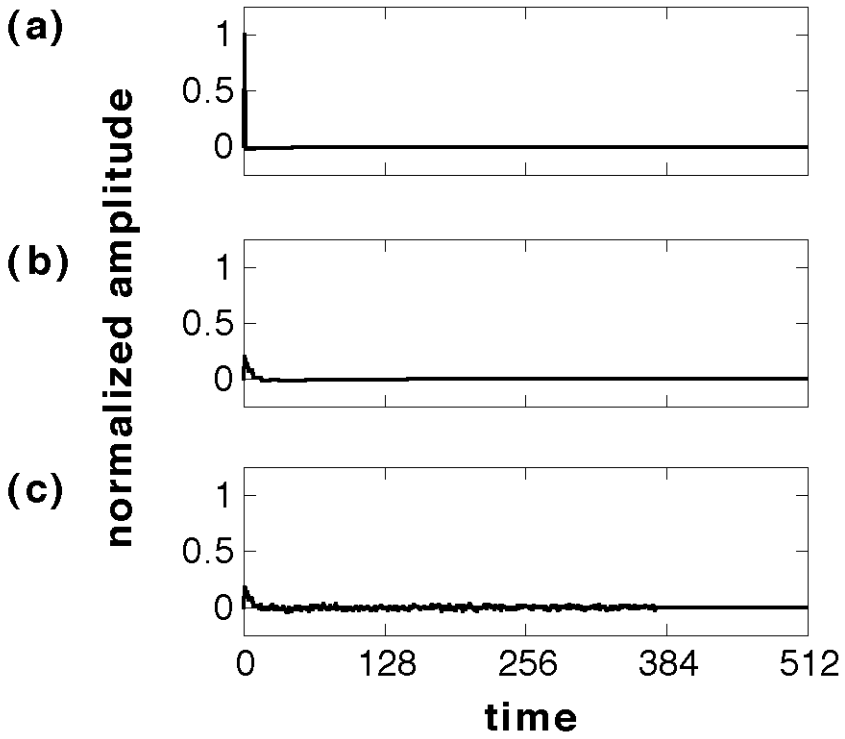

Figure 4.11: First-order differential of the impulse train: (a) original waveform shown in Fig. 5 (a), (b) a modified waveform by the clipping procedure shown in Fig. 9 (a), and (c) a modified waveform by the full use of the proposed technique shown in Fig. 10 (a)

four-second stimuli. In order to make a judgment, a six-second interval was given to the subject after listening to the two AB pairs.

4.3.5 Results of the experiments

Thurstone's case V was applied to the experimental results to convert them into a one-dimensional evaluation scale [395]. Figure 4.3 shows the preference orders of the stimuli with regard to their naturalness. It was found that stimulus "a" was evaluated the most natural stimulus. As indicated from the experimental results, the clipping procedure of the proposed technique appeared to contribute to the enhancement of the voice quality, since stimulus "c" was evaluated higher than stimulus "b" that employed the normal impulse train source. In spite of the fact that the voice quality of stimulus "b" was more bassy due to the clipping procedure, most of the subjects reported that stimulus "c" sounded more natural than stimulus "b", since the buzzer-like quality perceived in stimulus "b" was not exhibited in stimulus "c". This could be a subjective clue in the discrimination between stimuli "b" and "c".

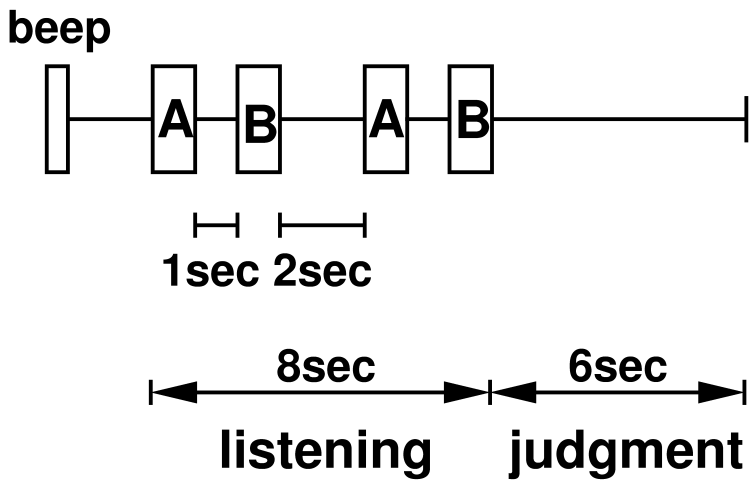

Figure 4.12: Procedure of paired-comparison test

As regards stimuli "d" and "e", the evaluation was much higher than that of stimulus "c". Similar to the case of stimulus "c", most of the subjects reported that the buzzer quality was not perceived in stimuli "d" and "e". In addition, the voice quality of these stimuli was much clearer than that of stimulus "c", since the high frequency characteristics were improved by the random fractal interpolation. This could be a subjective clue in discriminating these stimuli from stimulus "c". Since there were no large differences between stimuli "d" and "e", the randomization for the random fractal interpolation could have little effect on speech perception.

As indicated from the above experimental results, it can be concluded that the full implementation of the developed technique potentially contributes to the improvement of the normal impulse train source. Since a similar tendency was obtained through the three conditions despite the use of different types of stimuli, it may be concluded that the proposed technique can be effectively applied to various kinds of speech.

4.4 Discussion

Discontinuities in the source signals is considered to be a factor in significantly influencing the voice quality of synthesized speech. This finding of this study is also pointed out from several earlier studies which have attempted to develop advanced parametric models of glottal vibration waveforms [201, 256, 208, 193, 207, 206]. These models basically reflected the temporal characteristics of the glottal vibration, which consists largely of a gradual opening and a sudden closing. Since the abruptness of the closing phase is mainly responsible for causing the buzzer-like quality [250, 249, 206, 217], developing the

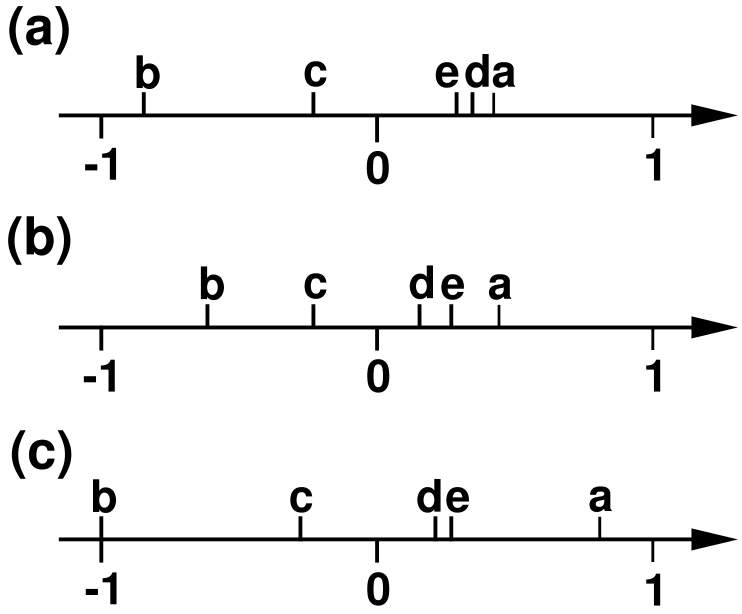

Figure 4.13: Results of psychoacoustic experiments: (a) experiment 1 (sustained vowel), (b) experiment 2 (continuous male speech), and (c) experiment 3 (continuous female speech)

models of the glottal vibration waveform which can adequately control the degree of the discontinuity in the closing phase is found to be effective for enhancing the voice quality of synthesized vowels [201, 256]. The present study also reached the same conclusion obtained from the above studies; namely that the voice quality of synthesized vowel speech can be enhanced by removing the large discontinuities of the conventional impulse train source. In addition to this improvement of the time domain characteristics of the impulse train source, this study also succeeded to devise a practical technique to compensate the high frequency characteristics of the modified impulse train source, which were decreased by removing the large discontinuity of the normal impulse train source. The random fractal interpolation was indicated to be a convenient method that could improve the high frequency characteristics simply by restoring the random fractalness, which was inherently observed in the source signals obtained from the speech samples.

Although the proposed technique successfully improved the conventional impulse train, the voice quality was still not equal to that of the actual speech samples, as indicated particularly in conditions 2 and 3. It appears that the factors investigated in this study are not sufficient enough to produce true human-like natural voice quality. Other significant factors, such as the fluctuations observed in human speech [1, 3], the influence of the differences in glottal vibration waveforms [253, 207], and nonlinear interaction between a

source signal and a vocal tract filter [230], seem necessary to be taken into consideration in order to further enhance the voice quality of synthesized speech.

4.5 Conclusions

This study statistically investigated the random fractalness observed in the source signals of an LPC vocoder by means of the Schauder analysis. The results showed that a certain random fractalness was particularly observed in the source signals obtained from the speech samples, while the impulse train source did not show such a feature. As inferred from the other cases in which random fractalness is considered to be a property of naturalness [350, 367, 309, 310], it might be considered that the finding of this study shows that the naturalness of human speech can objectively discussed in terms of its random fractalness.

Moreover, for enhancing the naturalness of synthesized vowels, an advanced technique for improving the conventional impulse train source of an LPC vocoder was developed in this study, which was based on the results of the Schauder analysis. The psychoacoustic experiments indicated that the large discontinuities of the source signals, a characteristic particularly observed in the impulse train source, could be a factor in degrading the voice quality of synthesized vowel speech. This study showed that undesirable buzzer-like quality by the impulse train source was eliminated by clipping the large Schauder coefficients from the impulse train source. Furthermore, the random fractal interpolation, which was developed by taking advantage of the random fractalness of the source signals, was found to be a useful technique for restoring the high frequency characteristics that were decreased by the clipping procedure. The proposed technique may be used as a practical method for improving the conventional impulse train source for realizing more natural voice quality through the use of LPC-vocoder-based speech synthesizers. Prospective applications should include a low bit-rate speech coder and a text-to-speech system based on a rule-based speech synthesis using an LPC vocoder [6, 8].

Chapter 5

Development of two speech synthesis applications

5.1 Introduction

This chapter describes two speech synthesis applications that were implemented in order to exemplify the effectiveness of the several techniques developed in this project. One is a real-time vocoder system made on an evaluation module of a digital signal processor (DSP) (Texas Instruments, TMS320C62EVM) [299]. The other is a Japanese rule-based speech synthesis system implemented on a personal computer (Apple, Macintosh Quadra 840AV). Both applications employed the modified LPC (linear predictive coding) vocoder as their speech synthesizer that fully implemented the features investigated in this project.

In addition, the unvoiced characteristics observed in voiced consonants are also investigated in this chapter. Since voiced consonants are a mixture of both a periodic and an aperiodic component attributed to unvoiced characteristics, waveforms of unvoiced consonants – which seem basically periodic due to the voiced feature – are influenced in detail by the unvoiced feature that disturbs the periodicity of voiced speech. Based on this characteristics of voice consonants, this chapter describes an advanced technique developed for enhancing the naturalness of synthesized voiced consonants. The technique was adopted as a method of enhancing the voice quality of voiced consonants in the two speech synthesis applications.

5.2 Implementation of MELP vocoder using lifting wavelet transform

This section describes an implementation of a MELP (mixed excitation linear prediction) vocoder [162, 163, 165, 166]. Unlike the conventional MELP vocoder, subband division necessary for implementing the MELP vocoder was performed particularly by lifting

wavelet transforms for gaining the computational efficiency that was, in turn, necessary for implementing real-time vocoder systems [423, 428, 451]. A new method for generating an appropriate glottal waveform was devised. In addition, three kinds of fluctuations observed in the steady parts of purely voiced speech were incorporated in order to enhance the naturalness of synthesized speech.

5.2.1 Introduction

A MELP vocoder is an advantageous speech codec when the data rate is limited [155, 157, 159, 162, 163, 164, 167]. The voice quality of a MELP vocoder is potentially comparable to a 4.8 kbps CELP (code excitation linear prediction) vocoder [145, 171, 172, 173], even if it runs at a 2.4 kbps data rate [162]. The main concept of a MELP vocoder is to perform a voiced/unvoiced decision in each subband and to generate excitation signals based on that decision. Mixing periodic and aperiodic components in excitation signals contributes to the enhancement of the naturalness of synthesized speech.

This study implemented a real-time vocoder system that applied the above feature of a MELP vocoder. Wavelet transform was employed to perform the subband division during the implementation [430, 431, 432, 449, 450]. Since computational efficiency was required for real-time implementation, a lifting scheme [423, 428, 451] was employed as a fast algorithm to perform wavelet transforms in order to reduce the computational redundancy.

One of the factors which significantly influences voice quality is the characteristics of glottal waveforms [193, 194, 195, 199, 200, 201, 206, 207, 208, 210, 217]. A new method to modify a triangular pulse was devised that is often used as a model of glottal waveforms. In order to improve the frequency characteristics of triangular pulses, which tend to degrade in the high frequency region, the proposed method reproduces the random fractalness that is observed in source signals obtained by LPC inverse filtering [5].

In an acoustically clean environment, a MELP vocoder may potentially work as a normal LPC vocoder, since every subband is likely to judge voiced speech as "voiced" due to the high signal-to-noise ratio. In such cases, synthesized speech tends to elicit a buzzer-like unnatural voice quality. In order to mitigate this degradation, three kinds of fluctuations which are always observed in the steady parts of voiced speech were incorporated into the developed MELP vocoder [1, 2, 3, 4].

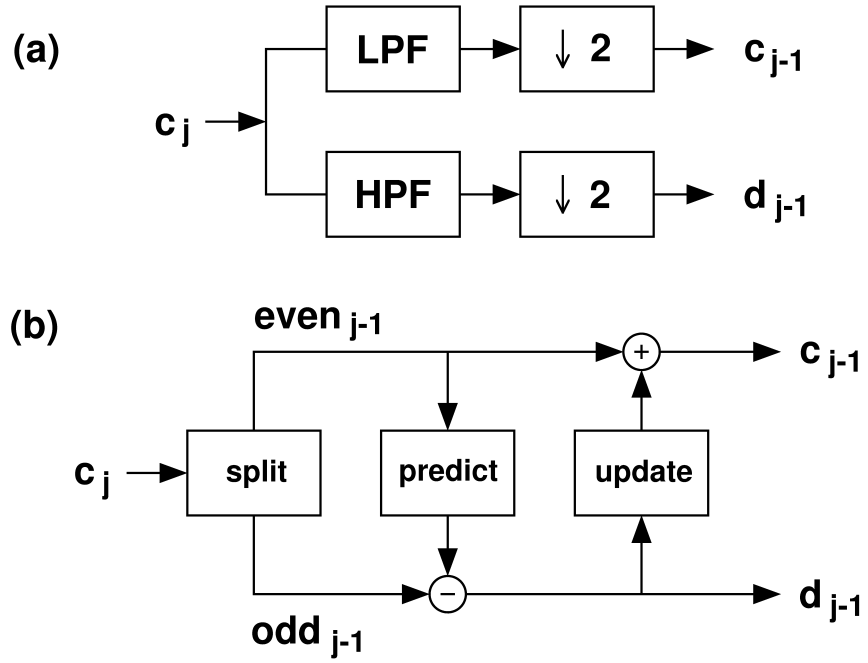

Figure 5.1: Wavelet transform: (a) classical implementation, (b) lifting scheme

5.2.2 Lifting wavelet transform

A lifting scheme is a fast algorithm of wavelet transforms [423, 428, 451]. It avoids classical wavelet transform procedure that is based on conventional low-pass and high-pass filtering by convolution. As shown in figure 5.1, the lifting scheme consists of: (1) splitting an original discrete sequence into an even and odd sequence, (2) subtracting the prediction estimated by the even sequence from the odd sequence, and (3) updating the even sequence in order to avoid alias effects [423, 428, 451]. These procedures consequently render the even sequence the scaling function coefficients and the odd sequence the wavelet coefficients. In this way, a lifting scheme can reduce the computational redundancy which is inevitable for classical implementation of the wavelet transform. In addition, in-place calculation can be performed to save extra memory which is required to store the results of the convolution of the classical implementation.

The scaling function and wavelet used in the subband division in the pilot case is shown in figure 5.2. The prediction utilizes four adjacent even samples whose weights are defined $-1/16$, $9/16$, $9/16$, and $-1/16$, respectively. This condition is equivalent to a low-pass filter for which the coefficients are set to be $1/64$, 0, $-1/8$, $1/4$, $46/64$, $1/4$, $-1/8$, 0, $1/64$. This is referred to in the literature as cubic interpolating wavelet transforms [428]. The lifting wavelet transform in this condition only requires 9 floating operations (6 additions and 3 multiplications) for calculating a coefficient of scaling function and a

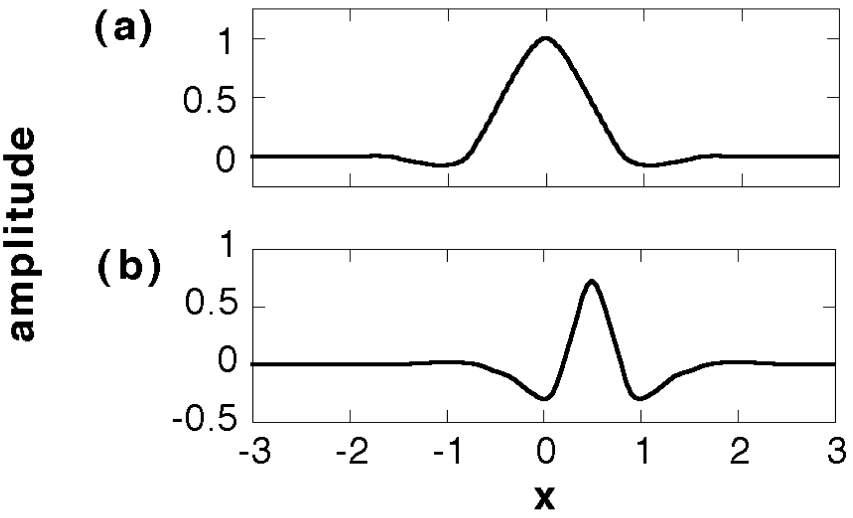

Figure 5.2: Cubic interpolating wavelet transform: (a) scaling function, (b) wavelet

wavelet coefficient, while classical implementation actually requires 16 (10 additions and 6 multiplications).

Figure 5.3 shows the frequency characteristics of five subbands divided by the lifting wavelet transform. Since the prediction employs only four even samples, each subband is not well separated in the frequency domain. This problem will be alleviated by increasing the number of samples employed in the prediction. However, this leads to an increase in the computation by a factor of two [428]. In this pilot case, computational efficiency took priority over the precise subband division for the sake of real-time implementation.

Figure 5.4 shows examples of reconstructed residual signals in each subband. Voiced/unvoiced decisions were made in each subband by evaluating the magnitude of normalized autocorrelation functions of the wavelet coefficients (level W1 to W4) or the scaling function coefficients (level V4) at the lag of the estimated pitch period [120, 121, 162, 163]. Since the number of the samples that were employed in calculating autocorrelation functions could be small, especially at low levels (e.g. 16 in the case of W4 and V4 when the analysis frame consists of 256 points), this also contributed to increasing the computational efficiency by a factor of two. In order to improve the decision making process, normalized autocorrelation functions were also evaluated after rectifying and smoothing the coefficients [162, 163]. The smoothing was performed by a -6 dB/oct low-pass filter implemented as a first-order auto-regressive (AR) digital filter. The pitch period was estimated by investigating the periodicity of the autocorrelation function of

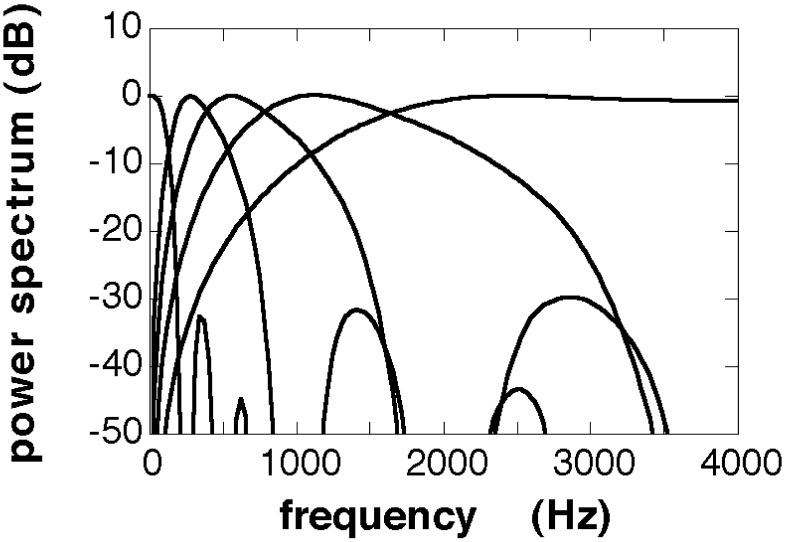

Figure 5.3: Frequency characteristics of five subbands: W1 (highest subband), W2, W3, W4, and V4 (lowest subband)

residual signals smoothed by the same −6 dB/oct low-pass filter [110, 111, 116, 120, 121].

5.2.3 Modification of triangular pulse by random fractal interpolation

In this implementation, excitation signals of the MELP vocoder were defined as spectral −6 dB/oct low-pass characteristics in the frequency domain, which included −12 dB/oct glottal and +6 dB/oct radiation characteristics from the mouth [110, 111, 116, 120, 121, 198].

As implied in the literature [162, 163], excitation signals that employ a triangular pulse are considered to be more proper for synthesizing a human-like natural voice quality, since it can adequately remove the discontinuities which are observed in conventionally used impulse excitation. However, the frequency characteristics of a triangular pulse do not necessarily guarantee the appropriate frequency characteristics that are required for the source signals of the vocoder, especially in the high frequency region. The graph (a) in the upper panel of figure 5.5 shows an example of a triangular pulse for which the main excitation ratio (ER) is set to be 32/512, where the main excitation is defined as the largest and fastest change in an excitation signal. The frequency characteristic of the triangular pulse is shown by graph (a) in the lower panel of figure 5.5. Apparently, the frequency characteristic decreases faster than a spectral −6 dB/oct decay.

In order to mitigate this problem, a technique called random fractal interpolation was

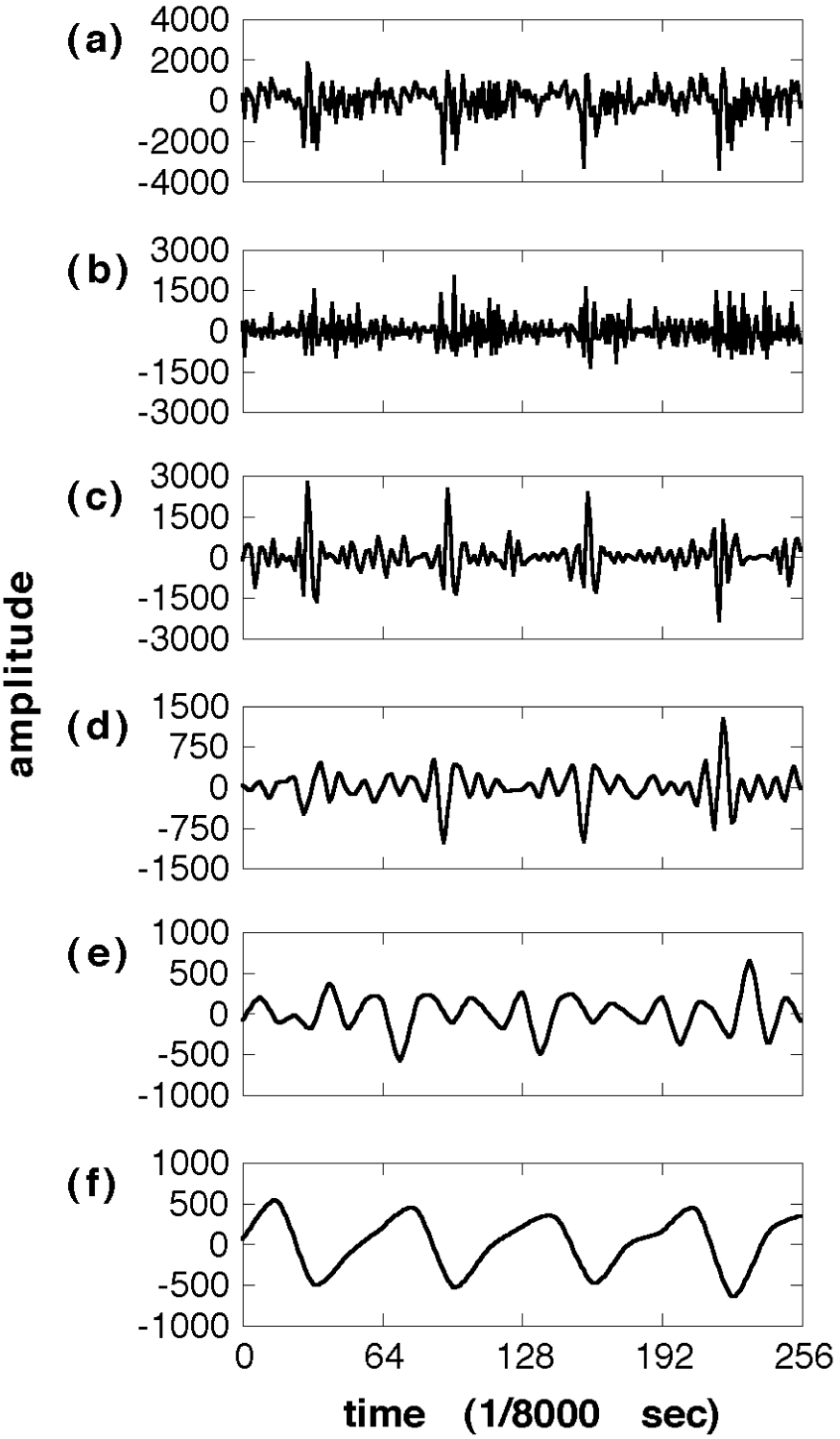

Figure 5.4: (a) Residual signal, and (b)–(f) reconstructed signals in the five subbands: (b) W1 (highest subband), (c) W2, (d) W3, (e) W4, and (f) V4 (lowest subband)

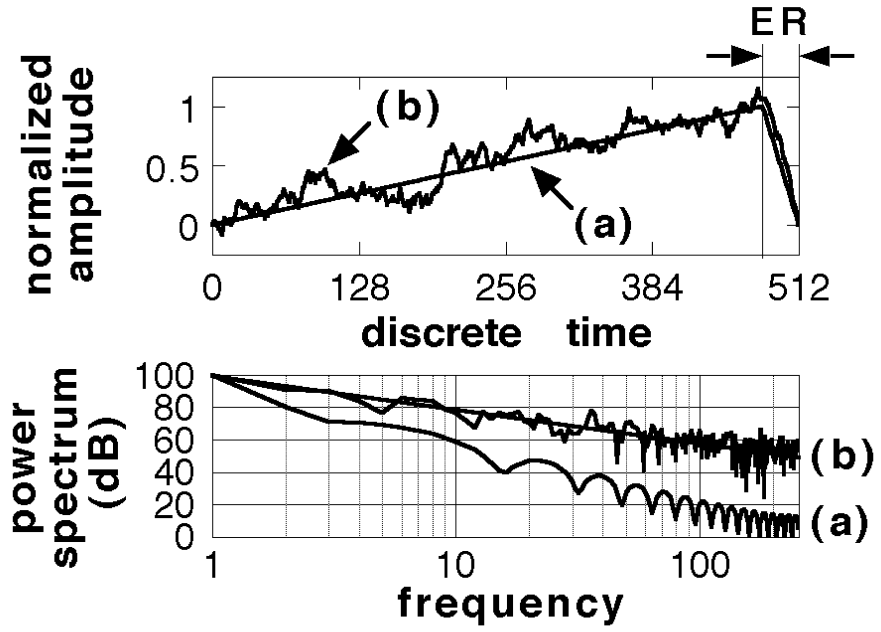

Figure 5.5: Temporal and frequency characteristics of (a) triangular pulse, (b) result of random fractal interpolation

applied [5, 7, 320]. It reproduces the random fractalness that is observed in the source signals obtained by LPC inverse filtering. Since the frequency and temporal characteristics of the -6 dB/oct source signals are equivalent to those of Brownian motion except that the source signals are characterized by a certain periodicity, there exists possibility that the source signals can be modeled as Brownian motion in detail. Our previous study investigating this hypothesis proved that the random fractalness becomes dominant as the resolution increases [5, 7]. The random fractal interpolation, based on this finding, can recover the power at the high frequency region.

The curve (b) in the upper panel of figure 5.5 shows an example of such triangular pulses that are modified by the proposed method. The lower panel of figure 5.5 shows that the frequency characteristics of the modified triangular pulse become more appropriate for the source signals of the vocoder, since it approximates -6 dB/oct spectral decay. With regard to the temporal characteristic, it appears that the modified triangular pulse still maintains almost the same excitation ratio of the original triangular pulse (32/512), even after the frequency characteristics are changed.

Voice quality of synthesized speech changes as the excitation ratio changes [250]. In general, the smaller the excitation ratio, the clearer the voice quality. On the other hand, the larger the excitation ratio, the softer the voice quality. During implementation, the excitation ratio was changed according to the peakiness value p defined in equation

5.1 [162, 250].

$$p = \frac{\frac{1}{N}\sum_{n=0}^{N-1} s_n^2}{(\frac{1}{N}\sum_{n=0}^{N-1} |s_n|)^2} \tag{5.1}$$

5.2.4 Models of three fluctuations in the steady part of voiced speech

Even during the steadiest phase of voiced speech, speech signals are not completely periodic. Pitch period and maximum amplitude fluctuations are always observed [1, 2, 4]. In addition, waveform itself changes slightly from pitch period to pitch period [1, 3]. These fluctuations are considered to be a contributing factor in the naturalness of voiced speech [1, 2, 3, 4].

The earlier study in this project indicates that pitch period and maximum amplitude fluctuation can be modeled as a $1/f$ fluctuation [1, 2, 4]. Examples of pitch period and maximum fluctuation for 512 pitch periods are shown in figure 5.6. Their frequency characteristics are also shown in figure 5.7. Psychoacoustic experiments performed in this study confirmed that the frequency characteristics of both fluctuations significantly influenced the naturalness of sustained vowels. For instance, pitch period and maximum amplitude fluctuations modeled as white noise would cause an unnatural pathologically rough voice quality [1, 2, 4]. The pitch period and maximum amplitude fluctuations modeled by a $1/f$ power law were incorporated into the implementation for the aim of enhancing the voice quality of synthesized voiced speech. The standard deviation of the pitch period fluctuation was set at 0.08 msec while the coefficient of variation of maximum amplitude fluctuation was set at 8 %.

A simple model for waveform fluctuation can be white noise when the excitation signals are defined as -6 dB/oct frequency characteristics [3, 193, 218]. For the implementation, the average power of waveform fluctuation to the modified triangular pulse was set at -25 dB for the implementation [3].

The implementation of our MELP vocoder also incorporated the aperiodic flag which was decided by the peakiness value calculated by equation 5.1 [162, 163]. When the aperiodic flag was true, the standard deviation of the pitch period fluctuation was set at 0.1 msec.

5.2.5 Implementation of MELP vocoder

Figure 5.8 shows the block diagrams for the analysis and synthesis stage of the developed MELP vocoder. For the voiced speech, the modified triangular pulse was repeatedly concatenated to generate periodic excitation signals. For realizing the MELP scheme, wavelet

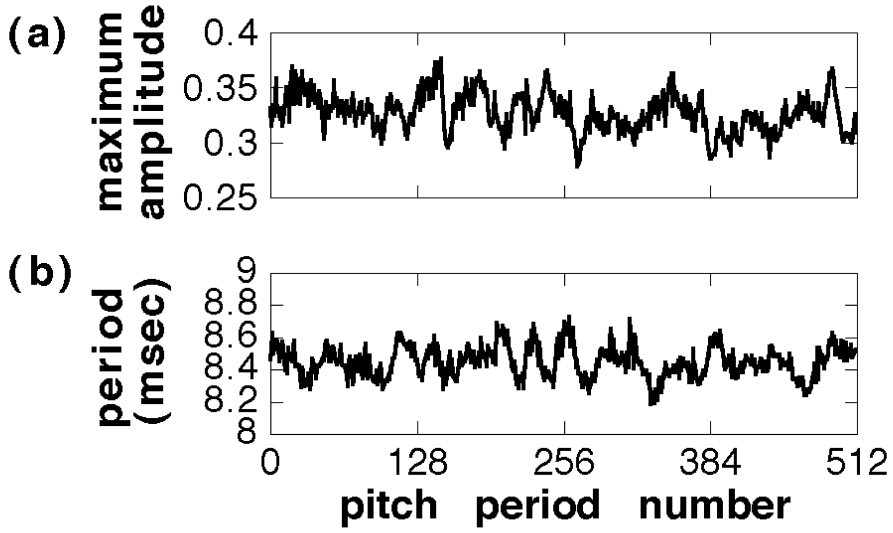

Figure 5.6: Examples of fluctuation sequence: (a) pitch period fluctuation, (b) maximum amplitude fluctuation

transform of the excitation signals was then performed. Using the voiced/unvoiced information for each subband obtained in the analysis stage, wavelet coefficients at unvoiced subbands were replaced with random wavelet coefficients characterized by Gaussian white noise.

This MELP vocoder was implemented on both a personal computer (Apple, Macintosh PowerBook 5300cs) as a prototype and a DSP evaluation module (TMS320C62 [299]) in order to investigate the feasibility of real-time implementation. In spite that the perfect optimization on the C source program was not performed, the DSP evaluation module was able to run the analysis and synthesis stage simultaneously within approximately 80 % of the capacity of the DSP processor.

Informal evaluation showed that the voice quality was perceived as more natural compared with a normal LPC vocoder. The buzzer-like voice quality was considerably eliminated. Three kinds of fluctuations incorporated into this MELP vocoder contributed to enhancing the naturalness of speech, even when all subbands were determined to be voiced, such as in the case of the steady part of sustained vowels. In particular, the naturalness of female speech, which often tends to degrade with the normal LPC vocoder, was remarkably improved. These successful results seemed to be attributable to adopting both the MELP scheme and the techniques of improving source signals by exploiting their random characteristics.

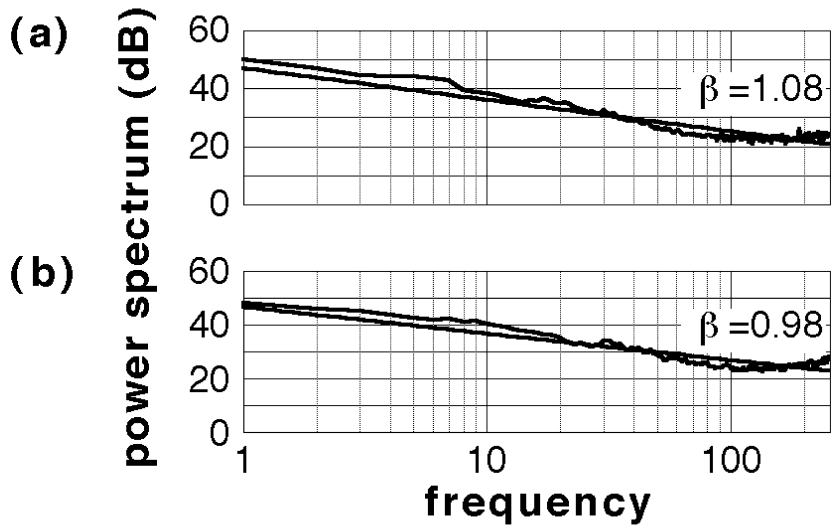

Figure 5.7: Frequency characteristics of fluctuation sequence: (a) pitch period fluctuation, (b) maximum amplitude fluctuation

5.2.6 Conclusions

This real-time implementation of the MELP vocoder indicated that mixing periodic and aperiodic components in excitation signals, which is the central concept of a MELP vocoder, could considerably enhance the naturalness of synthesized speech. In addition, the appropriately modeled randomness incorporated in the source signals contributed to enhancing the naturalness of the voice quality. Further evaluation, including formal psychoacoustic tests, is planned in order to investigate the validity of employing the developed MELP vocoder for the actual case of low-bit-rate telecommunications.

5.3 Development of a rule-based speech synthesis system for the Japanese language using a MELP vocoder

A Japanese rule-based speech synthesis system was implemented, which employs a MELP vocoder as its speech synthesizer. This section especially describes the speech synthesis techniques utilized in the system. Since the MELP vocoder developed in this study could effectively increase the naturalness of voiced consonants as well as purely voiced speech, the implemented system could successfully enhance the voice quality of synthesized speech compared with a system which employed a conventional LPC vocoder.

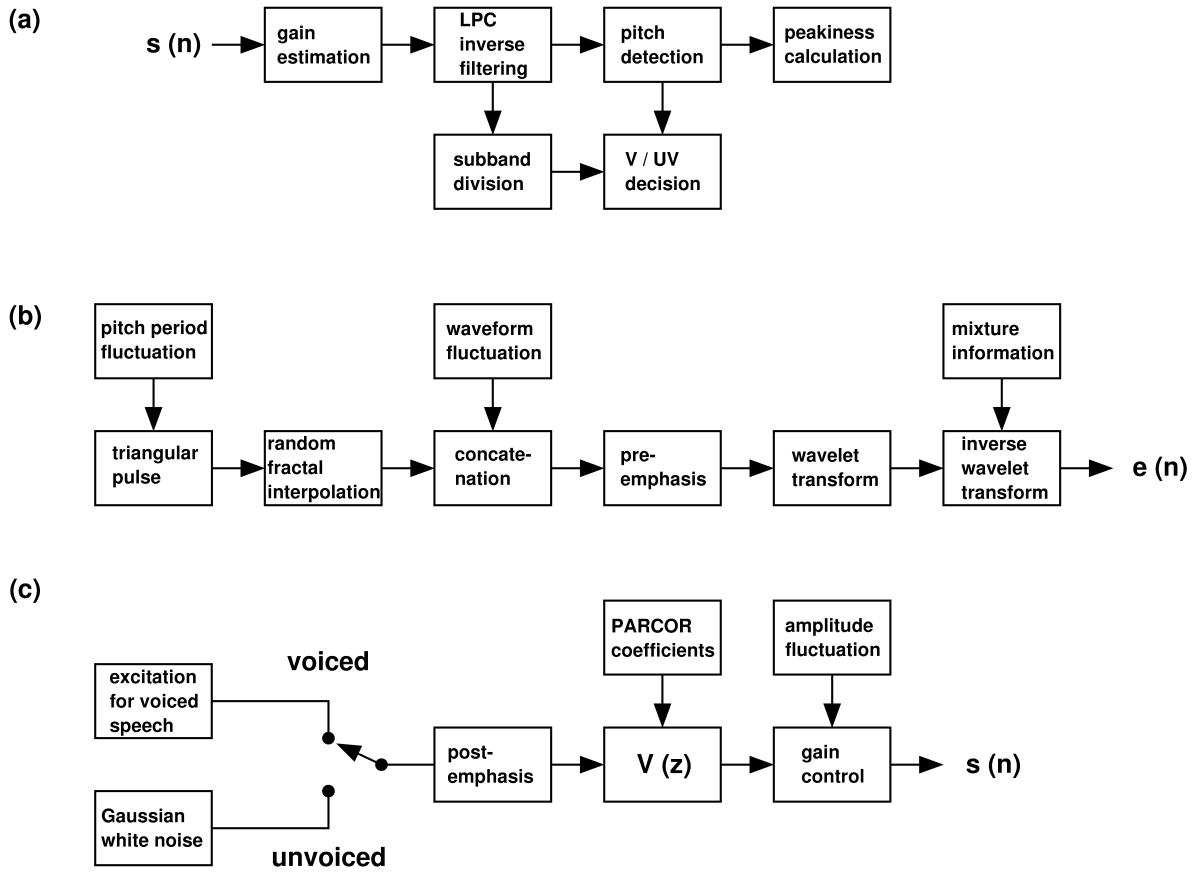

Figure 5.8: Procedure of the analysis and synthesis stage of our MELP vocoder: (a) speech analysis (b) generation of excitation signals for voiced speech, (c) speech synthesis

5.3.1 Introduction

Recently, a rule-based speech synthesis technology has been used intensively for the realization of a new style multimedia communication system that employs humanoid agents called "avatars" as its human interface [517, 518]. Since the degradation of synthesized speech phonated by the avatars often causes discomfort to human users in such communication systems, the development of more advanced techniques for enhancing the voice quality of the synthesized speech is strongly required. A high quality speech synthesis technology is considered to contribute greatly to realizing much more fluent speech communication between the avatars and human users [475, 519, 520]. This study has developed several speech synthesis techniques that can effectively improve the voice quality of the rule-based speech synthesis systems particularly for the Japanese language.

The model-based speech synthesizer employed in the rule-based speech synthesis systems conventionally utilizes oversimplified source signals represented such as an impulse train for voiced speech and white noise for unvoiced speech [110, 111, 116]. Although these

simple source signal models are convenient to use for processing speech synthesis, they inevitably result in degradation of the voice quality of the resulting synthesized speech. This is considered to be a side effect resulting from the oversimplification of source signals. In particular, a buzzer-like voice quality perceived in the voiced parts of synthesized speech is recognized as a major problem [155, 157, 159, 162, 163, 206, 207]. This undesirable degradation is distinctly perceived in synthesized voiced consonants. The reason for the buzzer quality of synthesized voiced consonants can be attributed mainly to the insufficiency of the binary source signal model which switches exclusively between either the impulse train or the white noise. Since voiced consonants phonated by normal human speakers do not necessarily show perfect periodicity in all frequency bands [54, 161], synthesized voiced consonants which are produced only by an impulse train may sound differently from the human voiced consonants.

A MELP vocoder takes advantage of these human speech characteristics for synthesizing more realistic speech [6, 162, 163]. It employs a mixture of both the impulse train and the white noise as its source signal for synthesizing a voiced consonant. The voice quality of the synthesized voiced consonant, therefore, may potentially be more similar to human speech than that produced by a conventional LPC vocoder that simply employs the binary switching model for producing its source signals [110, 111, 116]. The primary purpose of this study was to investigate the effectiveness of the MELP scheme for enhancing the voice quality of synthesized voiced consonants.

Even during the steadiest parts of sustained vowels phonated by normal human speakers, speech signals are not completely periodic [1, 2, 3, 4]. Since the buzzer-like quality of synthesized voiced speech can be reduced by appropriately disturbing the periodicity of speech signals, such aperiodicity is considered to be a potential factor in the naturalness of human voiced speech. In order to improve the voice quality of synthesized voiced speech, the MELP vocoder developed in this study also incorporated three kinds of fluctuations which are always observed in human voiced speech. This technique is especially aimed at enhancing the naturalness of the purely voiced parts of synthesized speech for which the MELP scheme is ineffective. This is because in most cases the source signals of the MELP scheme for purely voiced speech are often the same as normal impulse trains themselves.

5.3.2 MELP scheme in the developed system

In the synthesis stage, the MELP vocoder generates source signals for its synthesis filter based on the binary information of voiced/unvoiced decisions obtained from its subbands in the analysis stage [6, 162, 163]. Since the wavelet transforms can efficiently perform

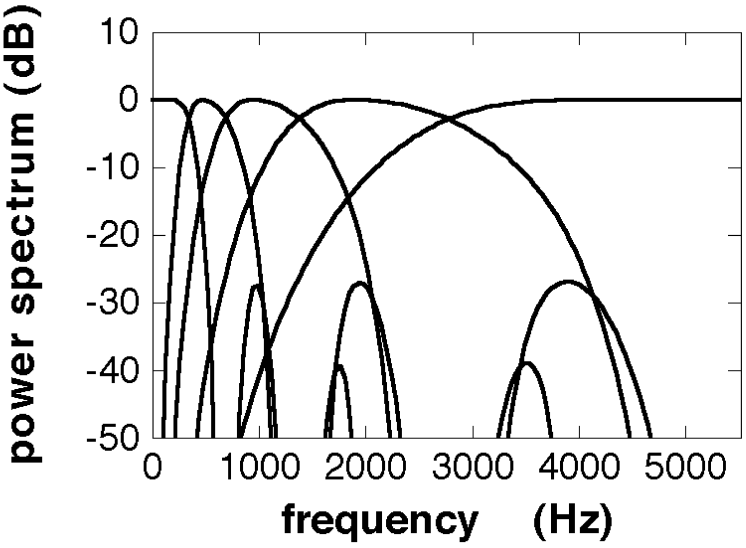

Figure 5.9: Frequency characteristics of five subbands: W1 (highest subband), W2, W3, W4, and V4 (lowest subband)

subband reconstruction as well as subband division, the MELP vocoder developed in this study employed the wavelet transforms for mixing periodic and aperiodic components in its source signals. As a pilot case, a twelve tap Daubechies' wavelet transform was utilized [423, 430, 431, 432, 449, 450, 452, 453, 454]. The frequency characteristics of five subbands, which are the results of the subband division by the Daubechies' wavelet transform, are shown in figure 5.9.

Figure 5.10 shows the block diagrams of the analysis and the synthesis stage of the MELP vocoder. Prior to the synthesis stage, the voiced/unvoiced decision was made for each subband in the analysis stage by evaluating the magnitude of the normalized autocorrelation function at the lag of the estimated pitch period. In the synthesis stage, to incorporate the aperiodicity into periodic source signals of voiced speech, random values characterized by Gaussian white noise were added to the wavelet coefficients of the subbands that were classified as unvoiced. The source signals in the time domain were finally obtained by the inverse wavelet transforms of the resultant signals.

5.3.3 Building a Japanese CV syllable database

The Japanese language basically uses five V (vowel) syllables and ninety-six CV (consonant + vowel) syllables for its pronunciation [123, 124, 125]. This study attempted to build a Japanese CV syllable database, including the five V syllables as well, in order to

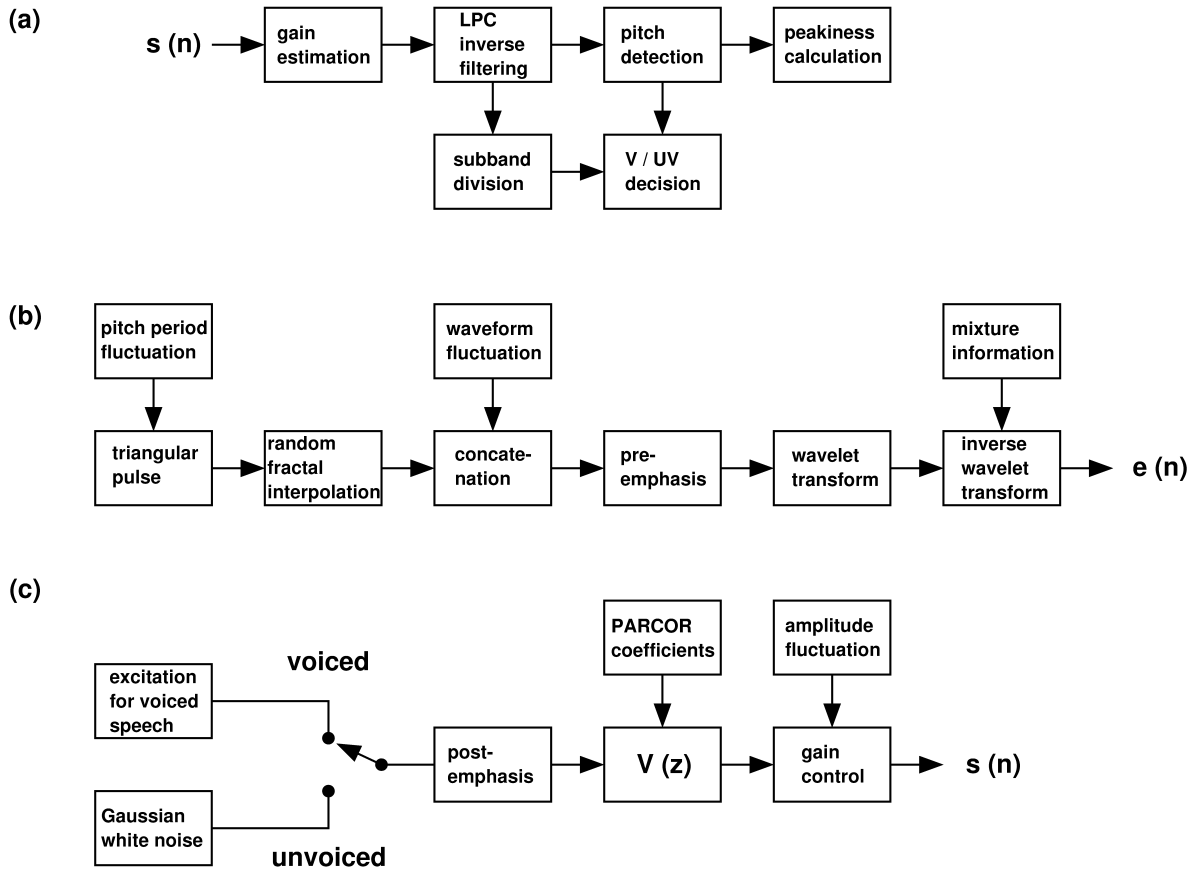

Figure 5.10: Block diagrams of (a) the analysis stage, (b), the procedure for generating excitation signals for voiced speech, and (c) the synthesis stage of the developed MELP vocoder

implement a Japanese rule-based speech synthesis system. The database had the unique feature of including the results of voiced/unvoiced decisions estimated in the five subbands for the MELP scheme. For building a complete CV syllable database of Japanese 101 syllables, speech samples were obtained under the sampling condition where the sampling frequency was 11.025 kHz and the quantization level 16 bits. Each syllable was phonated by a normal male speaker. From a frame which consisted of 256 samples of digitized speech, the following parameters were calculated: (1) fourteen PARCOR (partial autocorrelation) coefficients, (2) an overall voiced/unvoiced decision, (3) voiced/unvoiced decisions of the five subbands, (4) a normalized gain, and (5) a peakiness value of the residual signal. As illustrated in figure 5.5, to reduce the buzzer quality of synthesized voiced speech, a triangular pulse was employed during the synthesis process instead of a normal impulse train. The peakiness value obtained in the analysis stage was used for adjusting the gradient of the triangular pulse in order to control the softness of the voice quality [162, 250]. The CV syllable database recorded the above calculated parameters

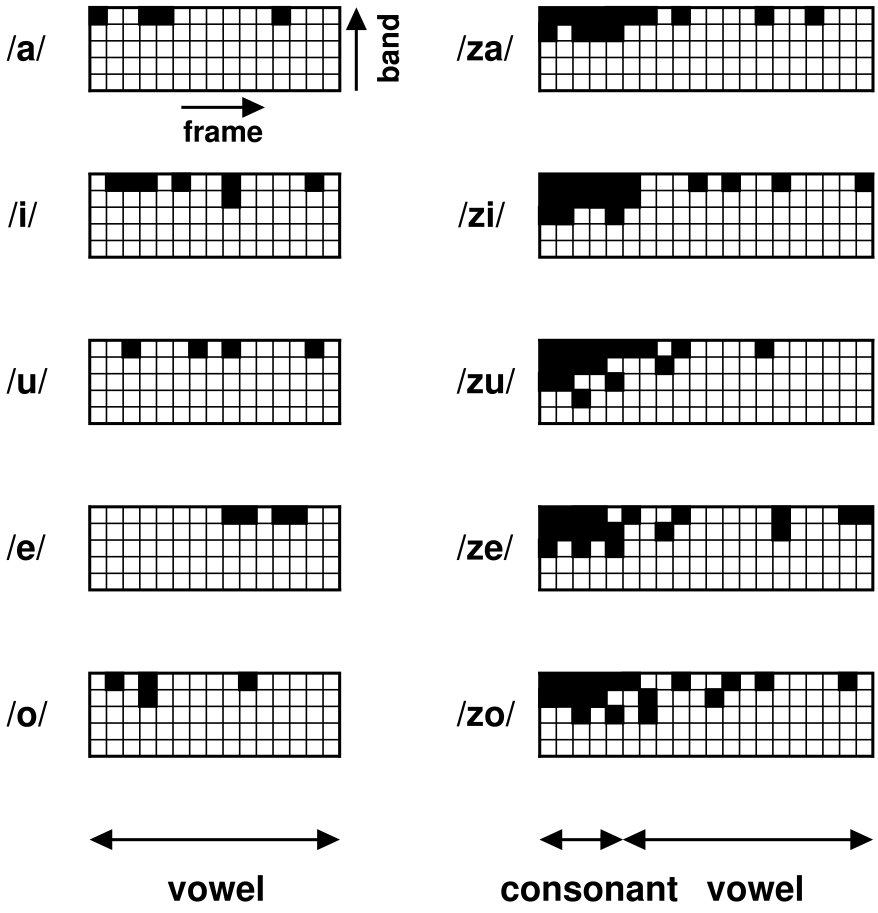

Figure 5.11: Examples of the voiced/unvoiced decisions in the five subbands: a filled square represents an unvoiced subband and a blank one represents a voiced subband.

for the duration of more than fifteen consecutive frames of each syllable.

Figure 5.11 shows examples of the voiced/unvoiced decisions in the five subbands obtained from the five Japanese vowels /a/, /i/, /u/, /e/, /o/ and voiced fricative consonants /za/, /zi/, /zu/, /ze/, /zo/. A filled square represents an unvoiced subband and a blank one represents a voiced subband. These decisions were recorded in the CV syllable database. Although the overall voiced/unvoiced decisions indicated that all the syllables could be classified as voiced speech, high frequency subbands tended to be classified as unvoiced at around the initial frame particularly for the voiced fricative consonants. At around their time of onset, the voiced fricative consonants are, therefore, considered to be featured by unvoiced as well as voiced characteristics. On the other hand, the five vowels did not show such a tendency. They are considered to be almost purely voiced speech, since they are classified as voiced in almost all the five subbands.

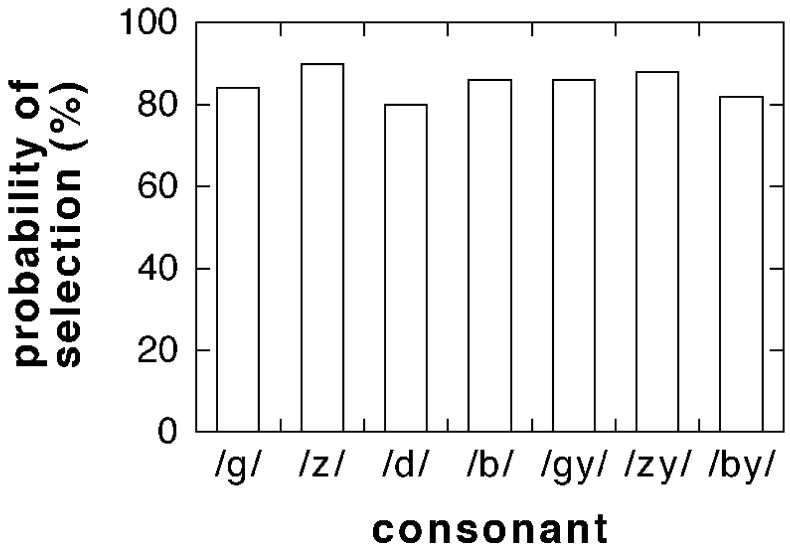

Figure 5.12: Subjective evaluation of the voice quality of Japanese voiced consonants synthesized by the developed MELP vocoder. The experimental results shows the probability of selecting the stimuli synthesized by the MELP vocoder compared with the stimuli synthesized by the conventional LPC vocoder.

5.3.4 Three types of fluctuations observed in the steady part of sustained voiced speech

Even during the steadiest parts of voiced speech, speech signals are not completely periodic. Fluctuations in pitch period and in maximum amplitude are always observed. In addition, waveform itself changes slightly from pitch period to pitch period. These three types of fluctuations are considered to be potential factors in enhancing the naturalness of synthesized voiced speech [1, 2, 3, 4]. Our earlier study indicates that the pitch period fluctuation as well as the maximum amplitude fluctuation can be modeled as a $1/f$ fluctuation [1, 2, 4]. A simple model for waveform fluctuations is Gaussian white noise when the source signals are defined as spectral -6 dB/oct signals [1, 3]. These fluctuations were employed in the MELP vocoder as illustrated in figure 5.10. The standard deviation of the pitch period fluctuation was set to be 0.08 msec and the coefficient of variation of the maximum amplitude fluctuation was set to be 8 %. The average power of the waveform fluctuation to source signals was set at -25 dB.

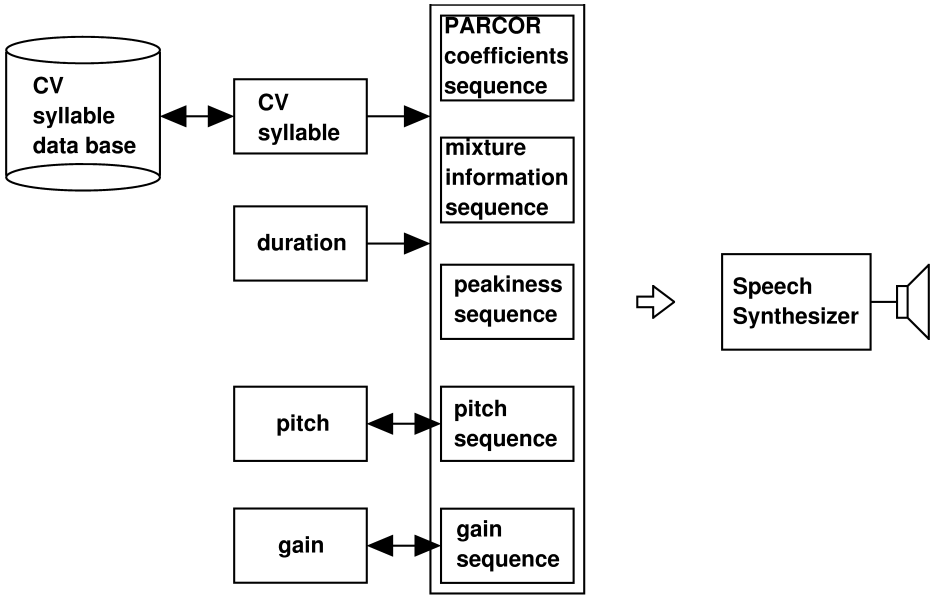

Figure 5.13: Parameters utilized in the developed Japanese rule-based speech synthesis system. These parameters were modified by a dictionary file which defined the proper pitch period, gain, and duration of each syllable contained in a word to be synthesized.

5.3.5 Subjective evaluation of synthesized voiced consonants

Subjective evaluation was performed with regard to the voice quality of the Japanese voiced consonants synthesized by the developed MELP scheme. Ten subjects participated in the psychoacoustic experiments. The subjects were asked to choose a stimulus that was perceived as having a more human-like natural voice quality in each paired-comparison test. The same voiced consonants synthesized by a conventional LPC vocoder were compared with the stimuli synthesized by the MELP vocoder.

The experimental results are summarized in figure 5.12. The results indicated that the MELP vocoder was able to produce a more natural voice quality than the LPC vocoder. The buzzer-like quality was considerably reduced when the MELP vocoder was employed. The subjective evaluation, as a result, indicated that the MELP vocoder would potentially contribute to enhancing the voice quality of a rule-based speech synthesis system. In addition, three kinds of fluctuations incorporated into synthesized speech also contributed to the production of more realistic voice quality. In particular, the pitch period fluctuation effectively reduced the buzzer-like quality perceived in the purely voiced parts of voiced consonants.

```
/*
(Dictionary File Format)
*word
syllable <pitch period> <gain> <
duration>
*/

*nihoNgo
ni    95    200    8
ho    70    250    8
N     72    150    8
go    74    200    8
```

Figure 5.14: Example of the dictionary file: This example defines the parameters of the syllables contained in a Japanese word "nihoNgo".

5.3.6 Implementing a Japanese rule-based speech synthesis system

A rule-based speech synthesis system for the Japanese language was finally implemented by using the MELP vocoder. The parameters recorded in the CV syllable database were utilized in the speech synthesizer as illustrated in figure 5.13. These parameters were modified by a dictionary file which defined the proper pitch period, gain, and duration of each syllable contained in a word to be synthesized [112]. Figure 5.14 shows an example of the dictionary file. In the example, the Japanese word "nihoNgo", meaning "the Japanese language" *per se*, is decomposed into four syllables /ni/, /ho/, /N/, and /go/. The dictionary file defines the properties of each syllable: pitch period, gain, and duration. Approximately 100 Japanese words are currently contained in the dictionary file. Concatenation processes for the modified parameters between the adjacent syllables were then performed with appropriate interpolation rules described in the literature [124].

Informal evaluation has indicated that the developed system developed by the author can successfully produce more naturally sounding synthesized speech than a system that uses a conventional LPC vocoder. As implied from the results of the psychoacoustic experiments with regard to synthesized voiced consonants, this success is considered to be attributable to adopting both the MELP scheme and the techniques of incorporating the three fluctuations into synthesized speech.

5.3.7 Conclusions

A Japanese rule-based speech synthesis system was developed by using the MELP scheme. This new system could more successfully enhance the voice quality of synthesized speech

than one that employed a conventional LPC vocoder.

Currently, a speech dialogue system for the realization of intelligent communication with avatars is under development, in collaboration with a venture company of 3-D animation. The developed speech synthesis system is supposed to be employed in the project.

5.4 Conclusions

Two speech synthesis applications developed in this project could successfully enhance the voice quality of synthesized speech, especially when compared with a system that employed a conventional LPC vocoder. This research shows that all of the techniques developed in this project should provide useful know-how for implementing the LPC-vocoder-based high quality speech synthesis systems that are expected to produce more realistic human-like natural speech. In order to investigate the feasibility of employing these new speech synthesis systems for actual industrial applications, further evaluation of those systems is currently underway, including formal psychoacoustic tests.

Chapter 6

Conclusions

6.1 Summary

This research project has investigated the characteristics of several kinds of randomness observed in human speech signals phonated by normal speakers. Based on the results of the analysis, several advanced techniques for artificially reproducing such randomness were developed with the aim of enhancing the voice quality of synthesized speech. The types of randomness particularly investigated in this project were: (1) amplitude fluctuation, (2) period fluctuation, (3) waveform fluctuation, (4) random fractalness of the source signals obtained by linear predictive analysis, and (5) unvoiced characteristics, namely, aperiodicity observed in voiced consonants. Modeling these forms of randomness, based on their statistical characteristics, was performed to obtain the know-how for the realization of LPC-vocoder-based high quality speech synthesis systems.

Chapter 2 described a study of the analysis and perception of the amplitude and period fluctuations. Since the frequency characteristics of these fluctuation sequences appeared to be roughly subject to a $1/f$ power law, this study reached the conclusion that amplitude and period fluctuation could be modeled as $1/f$ fluctuations for a preliminary model. Psychoacoustic experiments performed in this study indicated that the differences in the frequency characteristics of the amplitude and period fluctuations could potentially influence the voice quality of synthesized speech. Compared with $1/f^0$ (white noise), $1/f^2$, and $1/f^3$ fluctuation models, amplitude and period fluctuations modeled as $1/f$ fluctuations could produce a voice quality which was more similar to that of human speech phonated by normal speakers.

Chapter 3 described a study of the analysis and perception of the waveform fluctuations extracted from the residual signals of the LPC vocoder. Since the frequency characteristics of the waveform fluctuations appeared to be roughly subject to a $1/f^2$ power law, this study reached the conclusion that the waveform fluctuations could be modeled as $1/f^2$

fluctuations for a preliminary model. Psychoacoustic experiments performed in this study indicated that the differences in the frequency characteristics of waveform fluctuations could potentially influence the voice quality of synthesized speech. Compared with $1/f^0$ (white noise), $1/f$, and $1/f^3$ fluctuation models, waveform fluctuations modeled as $1/f^2$ fluctuations could produce voice quality which was more similar to that of human speech phonated by normal speakers.

Chapter 4 described a study of the analysis and perception of the random fractalness observed in the source signals of the LPC vocoder obtained from human speech samples. This study employed a multiresolution analysis method, based on Schauder expansion, in order to statistically investigate the time domain characteristics of the source signals. The results of the analysis indicated that the random fractalness was observed, particularly when a large resolution level was chosen. This study also found that a certain limitation existed with regard to the size of the discontinuity for the source signal waveforms. Based on the results of the analysis, an advanced technique was newly developed with the aim of enhancing the voice quality of synthesized speech produced by the normal impulse train. This study reached the conclusion that the buzzer-like degraded voice quality resulting from utilizing the impulse train could be mitigated by removing the extremely large discontinuity of the waveforms of the impulse train. The technique devised by the author also led to the creation of a method called random fractal interpolation. The purpose of this method was to restore the power in the high frequency region that had been undesirably decreased when the sharpness of the impulse train was reduced.

Chapter 5 described two applications that exemplified the effectiveness of the techniques developed in this project. One was a real-time vocoder system implemented on a digital signal processor (DSP) evaluation module (Texas Instruments, TMS320C62). The other was a Japanese rule-based speech synthesis system implemented on a personal computer (Apple, Macintosh Quadra 840AV). Both applications employed the modified LPC vocoder as their speech synthesizer that fully implemented the features investigated in this project. In addition, this chapter also described a study of the analysis and perception of the unvoiced characteristics observed in voiced consonants. Since voiced consonants are a mixture of both a periodic component attributed to voiced characteristics and an aperiodic component attributed to unvoiced characteristics, the waveforms of unvoiced consonants – which seem basically periodic due to reflecting the voiced feature – are disturbed in detail by the unvoiced feature. Psychoacoustic experiments conducted in this study clarified that synthesized voiced consonants produced by the conventional LPC vocoder tended to degrade in voice quality, since this method completely disregarded the

incorporation of the unvoiced feature into voiced consonants. An advanced technique, employing a wavelet transform for processing subband decomposition and reconstruction, was developed as a method for the inclusion of the unvoiced component with the voiced component. This study reached the conclusion that synthesized voiced consonants, for which the unvoiced feature was incorporated at high frequency subbands, could be perceived as having a more natural voice quality than that produced by the conventional LPC vocoder.

This project has achieved the following two major conclusions: (1) the voice quality of synthesized speech can be enhanced by the inclusion of the various types of randomness that are artificially produced by their adequate models, (2) The techniques developed in this project can be used as know-how towards the realization of LPC-vocoder-based high quality speech synthesis systems that are expected to produce more realistic human-like natural speech.

6.2 Original contributions

Synthesizing human-like natural voice quality by using the model-based speech synthesizers is still recognized as a difficult goal to achieve. Since an in-depth understanding of the mechanism of human speech production seems essential for the creation of natural voice quality by such model-based speech synthesizers, it would appear that considerable time is still required to reach the goal of perfecting the development of a realistic model of human speech production due to the high complexity of its physical system [259, 260, 480, 482, 483]. Although such a precise model is expected to contribute greatly to the enhancement of synthesized speech, some updated industrial speech applications including a mobile telephone and a car navigation system require more immediate techniques that are effective enough to produce a more natural voice quality in the given systems. This research project focused on developing such practical techniques for enhancing the naturalness of synthesized speech in terms of appropriately incorporating the randomness observed in human speech signals. The research has quantitatively demonstrated the specific procedures which address the question of how to artificially produce randomness, particularly for the LPC vocoder, which is not sufficiently described in some earlier studies [193, 218]. Since the LPC-vocoder-based speech synthesizer is widely employed in the above-mentioned commercial speech applications, the know-how for enhancing the voice quality of synthesized speech obtained from this project would potentially contribute to improving the quality of such promising industrial applications.

In addition, the results obtained from this project may also potentially stimulate the study of developing an advanced model of human speech production, which is expected to clarify how the randomness is generated in the actual physical system.

It should be emphasized that several types of randomness investigated by the author inherently exhibited the random fractal characteristics. This project uniquely took advantage of such random fractalness for developing advanced techniques to artificially producing such randomness. Since the naturalness of synthesized speech was successfully improved by the developed techniques, the conclusion was reached that such random fractalness could be a key factor for developing advanced models of the randomness. Earlier studies explain that random fractalness is also found in a variety of natural phenomena, including fluctuation obtained from biomedical signals such as heart-rate signals [308, 309, 310, 311, 317, 334, 335, 350, 354, 360]. As demonstrated typically in the heart-rate case, many biomedical systems particularly show that a spectral $1/f$ characteristic is frequently observed in the fluctuation sequences obtained from normal subjects. On the other hand, such a tendency does not appear in the fluctuation sequences of abnormal subjects. The spectral $1/f$ characteristic is, therefore, considered to be an inherent feature which reflects the normality of the biomedical systems. It is of interest that the experimental results obtained in the study of the amplitude and period fluctuations also showed strong agreement with the above-described tendency. Specifically, amplitude and period fluctuations characterized by spectral $1/f$ decay could produce normal voice quality, whereas, fluctuations characterized by a spectral $1/f^0$ decay resulted in a pathological rough voice quality. This project also theoretically discussed how the $1/f$ fluctuations could serve as an adequate model for representing these fluctuations in terms of their stationarity.

Not only has this project demonstrated the effectiveness of the new techniques, but the author has also shown the capability of developing commercial applications based on exploiting the advantages of the proposed techniques. Specifically, two speech synthesis applications were implemented in this research in order to exemplify the effectiveness of the developed techniques. One was a real-time vocoder system implemented on a digital signal processor (DSP) evaluation module (Texas Instruments, TMS320C62). The other was a Japanese rule-based speech synthesis system implemented on a personal computer (Apple, Macintosh Quadra 840AV). It was shown that both systems were potentially applicable to actual industrial applications. Recently, a rule-based speech synthesis technology has been used intensively in the development of a new style multimedia communication system that employs humanoid agents called "avatars" as its human interface. A high

quality speech synthesis technology is considered to contribute greatly to more fluent speech communication between the avatars and human users [517, 518, 519, 520]. Currently, based on the system developed in this research project, a speech dialogue system for intelligent communications with the avatars is under development in collaboration with a venture company of 3-D animation. In addition, the developed techniques are also planned to be employed for implementing a more effective speech codec system that makes use of a Voice over IP (internet protocol) technology for the Internet telecommunications.

6.3 Future work

This project has consisted of merely a pilot study to simply develop preliminary models of several types of the randomness. In order to improve the developed models, it is worth investigating how more precise models could contribute to further enhancement in the voice quality of synthesized speech. Especially for the three types of fluctuations, one of the most intriguing topics would be the analysis of how the characteristics of the fluctuations observed in the non-steady part of speech signals differ from the findings obtained in this research. Since the author only investigated the characteristics of the fluctuations obtained from the steady part of sustained vowels, a certain limitation remains in the utilization of these results as know-how for enhancing the voice quality of the non-steady part of synthesized speech. Although the two applications developed in this project successfully produced fairly natural voice quality in spite of not taking the above consideration into account, a more complete model can be expected to be generally accepted as a realistic model of the randomness.

Several topics are left for future work with regard to the further improvement of the two applications developed in this project. As regards the MELP vocoder, more accurate method for the pitch period estimation is strongly desirable, since erroneous decision making still considerably degrades the voice quality of synthesized speech even when the MELP scheme is employed instead of using the normal LPC vocoder. In addition, the MELP vocoder is also limited by the difficulty in making reasonable voiced/unvoiced decisions in each subband. In the implementation of this project, an autocorrelation technique was simply utilized as a preliminary method for the voiced/unvoiced decisions. However, due to the non-steady characteristics of actual human speech signals, the voiced/unvoiced decisions made by the autocorrelation technique often failed. Some other algorithms are under consideration by the author in order to perform the voiced/unvoiced decision making more appropriately for gaining the robustness of the MELP scheme.

The text-to-speech technology is now intensively employed in several speech applications such as car navigation systems, which have recently emerged as promising commercial applications of speech synthesis. At the present moment, the motion picture experts group (MPEG) is working on a standardization for various types of the multimedia communications as MPEG-4, anticipating that text-to-speech applications will be a key technology for uniquely identifying its standard [171, 172, 173]. Although a particular speech synthesizer has not been chosen for the MPEG-4 standard, the draft of the MPEG-4 defines several control parameters for realizing text-to-speech systems. The author is currently planning to implement an MPEG-4-compatible text-to-speech system based on the rule-based speech synthesis system developed in this project. Not only the voice quality of text-to-speech systems reflect the particular types of speech synthesis algorithm, but it also considerably depends on the control parameters that define the prosodical information of synthesized speech. This research project, therefore, anticipates that the study on the text-to-parameter conversion should be included in its future work for realizing high quality rule-based speech synthesis systems.

Appendix A

Fractional Brownian motion

Mandelbrot and Wallis originally introduced the concept of fractional Brownian motion as a model for investigating the characteristics of wide-band random processes [308, 309, 310, 311, 317, 334, 335, 350, 354, 425]. It is basically a theoretical extension of the conventional Brownian motion.

A wide-band random process $v(t)$ is characterized by its spectral densities $S_v(f)$. There introduced a simple power law which defines that $S_v(f)$ is proportional to spectral $1/f^{\beta}$ characteristics through all frequency bands with the conditions of $1 < \beta < 3$, $\beta = 2H + 1$ and $\beta = 5 - 2D$ [308, 309, 310, 311, 317, 334, 335, 350, 354, 425]. This definition provides useful relationship among the spectral decay β, the Hurst exponent H, and the fractal dimension D. The random processes which meet the condition of the above power laws are called fractional Brownian motion [308, 309, 310, 311, 317, 334, 335, 350, 354, 425].

The frequency characteristics of a fractional Brownian motion give information of its time correlations. When $S_v(f)$ increases steeply at low frequency, $v(t)$ varies more slowly, and *vice versa*. In addition, it is known that the mean and the mean square value of a fractional Brownian motion are subject to the following statistical relationship [334, 311, 425].

$$\begin{aligned} \langle v(rt) \rangle &= r^H \langle v(t) \rangle \\ \langle v^2(rt) \rangle &= r^{2H} \langle v^2(t) \rangle, \end{aligned} \tag{A.1}$$

where

$v(t)$ is a spectral $1/f^{\beta}$ sequence,

r is the resolution factor of the sequence,

$\langle \cdot \rangle$ is the expectation operator.

Since the mean and the variance derived from equation A.1 are considered to change as the resolution factor r changes, $1/f^{\beta}$ fluctuations are classified as nonstationary processes

with the condition of $1 < \beta$. However, the mean and the variance of $1/f$ sequences, which are derived from equation A.1 with the condition of $\beta \to 1$, namely $H \to 0$, are statistically invariant even though the resolution factor r changes. This nature of spectral $1/f$ sequences is known as self-similarity which guarantees the quasi-stationarity of the sequences [308, 309, 310, 311, 317, 334, 335, 350, 354, 425]. Spectral $1/f$ sequences are often observed in the many physical systems [335, 350, 354, 360, 361, 362, 363, 364, 365, 366, 367]. They are found in the noises of almost all electronic components from simple carbon resistors to advanced semiconducting devices. Also in the biomedical systems, spectral $1/f$ sequences are inherently found from the nerve membrane level to the human behavioral level [360, 361, 362, 363, 364, 365, 366, 367]. Why this kind of random processes can be commonly observed in a variety of physical systems is not fully answered yet [335, 350, 352, 354, 360].

Fast Fourier transform (FFT) is a useful tool for generating fractional Brownian motion [310, 311, 334, 318]. Gaussian white noise was first transformed to the frequency domain, then passed through the low pass filter characterized by a spectral $1/f^{\beta}$ power law. Finally, the result was transformed back into the time domain. The power spectrum of a spectral $1/f^{\beta}$ sequence is represented as

$$S_v(f) = |T(f)|^2 S_w(f) \propto |T(f)|^2, \tag{A.2}$$

where

$S_v(f)$ is the power spectrum of a spectral $1/f^{\beta}$ sequence $v(t)$,

$T(f)$ is the frequency characteristics of spectral $1/f^{\beta}$ filter,

$S_w(f)$ is the power spectrum of Gaussian white noise.

Thus, the spectral $1/f^{\beta}$ filter is required to be

$$T(f) \propto 1/f^{\beta/2}. \tag{A.3}$$

The typical sequences produced by this method are shown in figure A.1. These are the examples of spectral $1/f^0$ (white noise), $1/f$, $1/f^2$, and $1/f^3$ sequences, respectively. It can be seen that the smoothness of the sequences increases as the value of β increases. Simultaneously, the sequences show to be nonstationary processes. These are attributable to the dominance of the low frequency components in the sequences for a larger β. The examples of spectral $1/f^2$ and $1/f^3$ sequences shown in figure A.1 (c) and (d) particularly exemplify the non-stationarity of fractional Brownian motions in the case of a larger β. Compared with spectral $1/f^0$ and $1/f$ sequences shown in figure A.1 (a) and (b), nonstationary changes in the short-time mean of spectral $1/f^2$ and $1/f^3$ sequences are easily detected even by visual inspection.

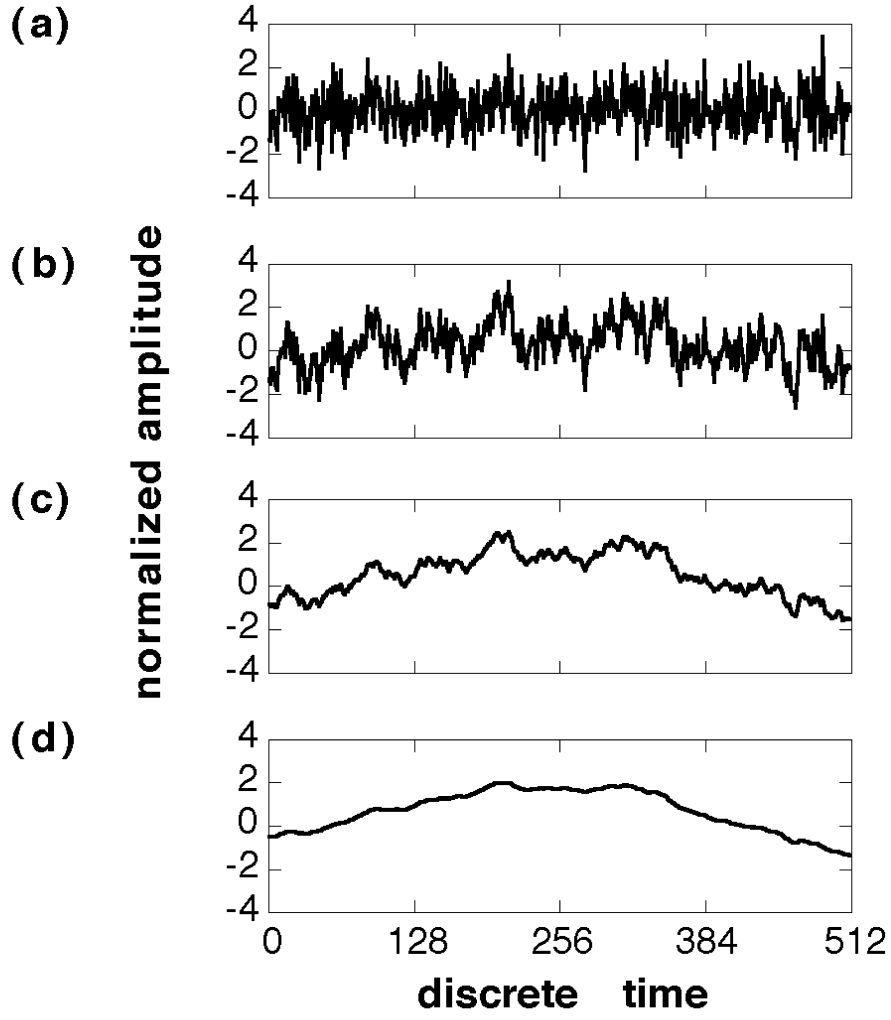

Figure A.1: Variations of $1/f^{\beta}$ sequence. The values of the exponent β are (a) zero, (b) one, (c) two, and (d) three. The mean and the standard deviation are normalized to be zero and one. Compared with spectral $1/f^0$ and $1/f$ sequences shown in (a) and (b), nonstationary changes in the short-time mean of spectral $1/f^2$ and $1/f^3$ sequences shown in (c) and (d) are easily detected even by visual inspection.

Appendix B

Theoretical definition of the source signals of the LPC vocoder

Theoretically, the source signals of the LPC vocoder is defined as being characterized by a spectral -6 dB/oct decay in the frequency domain [110, 111, 116, 120, 121, 198]. The LPC analysis assumes that the spectral envelope of its residual signals is approximately represented as being flat with the conditions of -12 dB/oct glottal vibration characteristics and $+6$ dB/oct mouth radiation characteristics. Taking these factors into consideration, the source signals of the LPC vocoder can be represented as spectral -6 dB/oct signals.

Figure B.1(a) illustrates the three stages of the speech synthesis procedure performed in the LPC vocoder [110, 111, 116, 120, 121, 198]. These three stages are individually represented as digital filters labeled as a glottal vibration filter $G(z)$, a vocal tract filter $V(z)$, and a mouth radiation filter $R(z)$ in the exhibited system. The vocal tract filter $G(z)$ defines the temporal and frequency characteristics of glottal vibration. As is mentioned above, the frequency characteristics of the glottal filter $G(z)$ are largely represented as being the -12 dB/oct decay. The vocal tract filter $V(z)$, which is characterized by LPC coefficients, particularly represents dominant peaks in the spectral envelope called formants. The mouth radiation filter $R(z)$, which represents an impedance mismatching between the inside and outside of the mouth, is simply characterized as a $+6$ dB/oct high-pass-filter. Synthesized speech denoted as $s(n)$ in the figure is produced by filtering an excitation signal $e(n)$ with all the three filters.

Since the above-described filters are theoretically defined as being linearly separated, the order of the filters can be changed without influencing the resulting synthesized speech. Figure B.1(b) shows an instance in which the order of the $V(z)$ and $R(z)$ are reversed. In this case, the excitation signal $e(n)$ is processed by both the filter $G(z)$ and $R(z)$ to be a source signal of the LPC synthesis filter $V(z)$. The frequency characteristic of the source signal is, therefore, defined as being a spectral -6 dB/oct decay due to combining

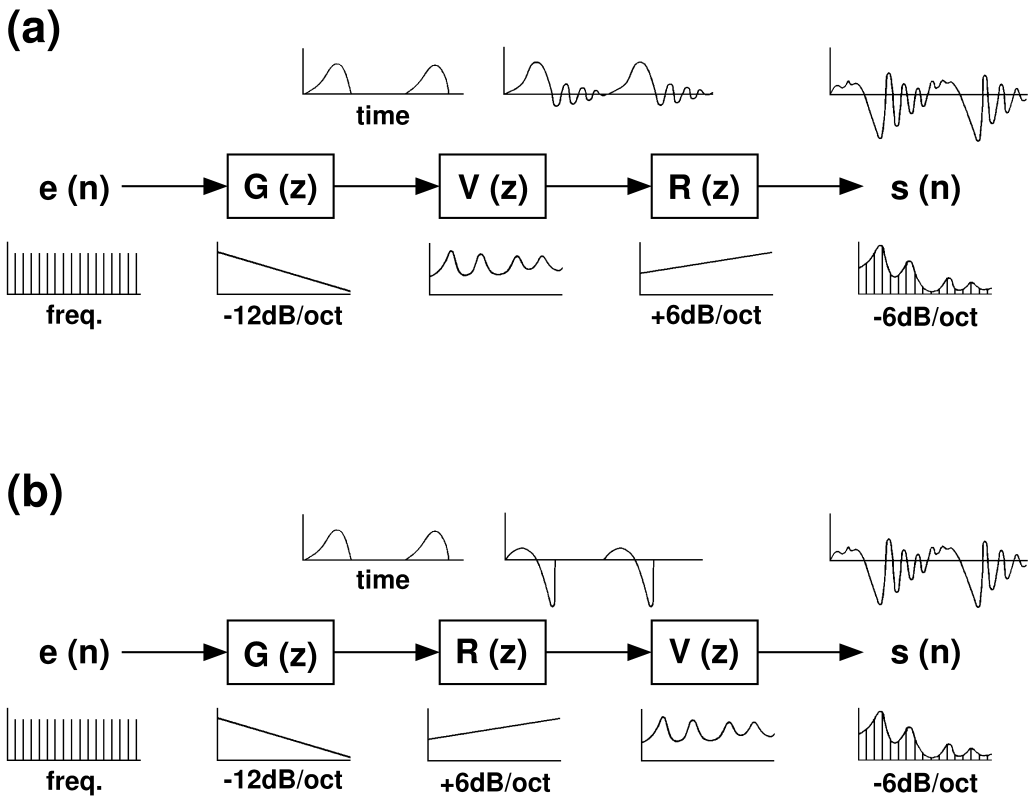

Figure B.1: (a) LPC vocoder synthesizes speech by processing the three stages: a glottal vibration filter $G(z)$, a vocal tract filter $V(z)$, and a mouth radiation filter $R(z)$. (b) The filters $V(z)$ and $R(z)$ are exchanged each other. Excitation signal $e(n)$ is processed by both the filter $G(z)$ and $R(z)$ to be a source signal of the vocal tract filter $V(z)$.

together the filter $G(z)$ and $R(z)$.

A first-order auto-regressive (AR) system defined in equation B.1 is usually employed as a -6 dB/oct low-pass-filter for generating source signals from excitation signals $e(n)$.

$$H(z) = \frac{1}{1 - 0.98z^{-1}} \tag{B.1}$$

The source signals for synthesizing voiced speech are, therefore, represented as the exponentially decreasing waveforms as shown in figure 4.1(c), when normal impulse trains are chosen excitation signals. As for unvoiced speech, Brownian motion, which is a wide-band random process for which the frequency characteristics are particularly characterized by the spectral -6 dB/oct decay, becomes the source signals when conventional Gaussian white noise is chosen excitation signals.

Appendix C

C program lists

The speech synthesizer developed in this project required a linear predictive analysis for the extraction of PARCOR coefficients from a speech signal and a linear predictive synthesis for remaking the analyzed speech. The C programs for these two routines are listed below.

C.1 Linear predictive analysis

```
/*
    This program performs linear predictive analysis
    of speech signal using auto-correlation method.

    s[n]    speech signal,n=0,...,N-1
    e[n]    residual signal,n=0,...,N-1
    a[m]    LPC coefficients,m=0,...,M
    gamma[m]    PARCOR coefficients,m=0,...,M

    last revised in Sep.20, 1999

    programmed by Naofumi AOKI
*/

#include <stdio.h>
#include <stdlib.h>
#include <math.h>

#define M 10
#define N 256
#define PI 3.1415926535

void LPC(float s[],float a[],float gamma[]);
void HammingWindow(float s[]);
void PARCOR_a(float s[],float e[],float gamma[]);

void main(void)
{
    FILE *fp;
    char fn[20];
    int i;
    float *s,*e,*a,*gamma,dummy=0;
```

```
    s=(float *)calloc(N+1,sizeof(float));
    e=(float *)calloc(N,sizeof(float));
    a=(float *)calloc(M+1,sizeof(float));
    gamma=(float *)calloc(M+1,sizeof(float));

/* input speech file */

    printf("Enter Speech File Name (input file) ->");
    scanf("%s",fn);
    fp=fopen(fn,"r");
    for(i=0;i<N+1;i++)
    {
        fscanf(fp,"%f",&dummy);
        s[i]=dummy;
    }
    fclose(fp);

/* +6 dB/oct pre-emphasis */

    for(i=0;i<N;i++)
    {
        s[i]=s[i+1]-0.98*s[i];
    }

/* Hamming window */

    HammingWindow(s);

/* LPC analysis */

    LPC(s,a,gamma);

/* residual calculation */

    PARCOR_a(s,e,gamma);

    printf("Enter LPC Coefficients File Name (output file) ->");
    scanf("%s",fn);
    fp=fopen(fn,"w");
    for(i=0;i<=M;i++)
    {
        fprintf(fp,"%f\n",a[i]);
    }
    fclose(fp);

    printf("Enter PARCOR Coefficients File Name (output file) ->");
    scanf("%s",fn);
    fp=fopen(fn,"w");
    for(i=0;i<=M;i++)
    {
        fprintf(fp,"%f\n",gamma[i]);
    }
    fclose(fp);

    printf("Enter Residual File Name (output file) ->");
    scanf("%s",fn);
    fp=fopen(fn,"w");
```

```
    for(i=0;i<N;i++)
    {
        fprintf(fp,"%f\n",e[i]);
    }
    fclose(fp);
}

void HammingWindow(float s[])
{
    int i;

    for(i=0;i<N;i++)
    {
        s[i]=(0.54-0.46*cos(2*PI*i/(N-1)))*s[i];
    }
}

void LPC(float s[],float a[],float gamma[])
{
    int i,j,m;
    float *r,*b,power,dummy;

    r=(float *)calloc(M+1,sizeof(float));
    b=(float *)calloc(M+1,sizeof(float));

    for(i=0;i<=M;i++)
    {
        r[i]=0;

        for(j=0;j<N-i;j++)
        {
            r[i]=r[i]+s[j]*s[j+i];
        }
    }

    gamma[0]=0;
    a[0]=1;
    power=r[0];

    gamma[1]=-r[1]/r[0];
    a[1]=gamma[1];
    power=power*(1-gamma[1]*gamma[1]);

    m=2;

    while(power>0 && m<=M)
    {
        for(i=1;i<m;i++)
        {
            b[i]=a[m-i];
        }

        dummy=0;

        for(i=0;i<m;i++)
        {
            dummy=dummy+a[i]*r[m-i];
        }
```

```
        gamma[m]=-dummy/power;

        for(j=1;j<m;j++)
        {
            a[j]=a[j]+gamma[m]*b[j];
        }

        a[m]=gamma[m];
        power=power*(1-gamma[m]*gamma[m]);

        m++;
    }

    free(r);
    free(b);
}

void PARCOR_a(float s[],float e[],float gamma[])
{
    int i,n;
    float *f,*b;

    f=(float *)calloc(M+1,sizeof(float));
    b=(float *)calloc(M+1,sizeof(float));

    for(n=0;n<N;n++)
    {
        if(n==0)
        {
            f[0]=s[n];
            b[0]=s[n];

            for(i=1;i<=M;i++)
            {
                f[i]=s[n];
                b[i]=gamma[i]*f[i];
            }

            e[0]=f[M];
        }

        if(n>0)
        {
            f[0]=s[n];

            for(i=1;i<=M;i++)
            {
                f[i]=f[i-1]+gamma[i]*b[i-1];
            }

            for(i=M;i>=1;i--)
            {
                b[i]=b[i-1]+gamma[i]*f[i-1];
            }

            b[0]=s[n];
            e[n]=f[M];
```

```
        }
    }

    free(f);
    free(b);
}
```

C.2 Linear predictive synthesis

```
/*
    This program performs linear predictive synthesis
    of speech signal using PARCOR method.

    s[n]    speech signal,n=0,...,N-1
    e[n]    residual signal,n=0,...,N-1
    a[m]    LPC coefficients,m=0,...,M
    gamma[m]    PARCOR coefficients,m=0,...,M

    last revised in Sep.20, 1999

    programmed by Naofumi AOKI
*/

#include <stdio.h>
#include <stdlib.h>
#include <math.h>

#define M 10
#define N 256

void PARCOR_s(float s[],float e[],float gamma[]);

void main(void)
{
    FILE *fp;
    char fn[20];
    long i;
    float *s,*e,*gamma,dummy=0;

    s=(float *)calloc(N+1,sizeof(float));
    e=(float *)calloc(N,sizeof(float));
    gamma=(float *)calloc(M+1,sizeof(float));

/* PARCOR coefficients file */

    printf("Enter PARCOR Coefficients File Name (input file) ->");
    scanf("%s",fn);
    fp=fopen(fn,"r");
    for(i=0;i<=M;i++)
    {
        fscanf(fp,"%f",&dummy);
        gamma[i]=dummy;
    }
    fclose(fp);

/* excitation file */
```

```
    printf("Enter excitation File Name (input file) ->");
    scanf("%s",fn);
    fp=fopen(fn,"r");
    for(i=0;i<N;i++)
    {
        fscanf(fp,"%f",&dummy);
        e[i]=dummy;
    }
    fclose(fp);

/* PARCOR synthesis */

    PARCOR_s(s,e,gamma);

/* -6dB/oct post-emphasis */

    for(i=0;i<N;i++)
    {
        s[i+1]=s[i+1]+0.98*s[i];
    }

    printf("Enter Sound File Name (output file) ->");
    scanf("%s",fn);
    fp=fopen(fn,"w");
    for(i=0;i<N;i++)
    {
        fprintf(fp,"%f\n",s[i]);
    }
    fclose(fp);
}


void PARCOR_s(float s[],float e[],float gamma[])
{
    long i,n;
    float *f,*b;

    f=(float *)calloc(M+1,sizeof(float));
    b=(float *)calloc(M+1,sizeof(float));

    for(n=0;n<N;n++)
    {
        if(n==0)
        {
            f[M]=e[n];

            for(i=M-1;i>=0;i--)
            {
                f[i]=f[i+1];
            }

            b[0]=f[0];

            for(i=0;i<=M-1;i++)
            {
                b[i+1]=gamma[i+1]*f[i];
            }
```

```
            s[0]=f[0];
        }

        if(n>0)
        {
            f[M]=e[n];

            for(i=M-1;i>=0;i--)
            {
                f[i]=f[i+1]-gamma[i+1]*b[i];
            }

            for(i=M-1;i>=0;i--)
            {
                b[i+1]=b[i]+gamma[i+1]*f[i];
            }

            b[0]=f[0];
            s[n]=f[0];
        }
    }

    free(f);
    free(b);
}
```

Acknowledgements

This doctoral thesis is the summary of the research project which has been conducted for five years since 1995 at the laboratory of Sensory Information Engineering at Research Institute for Electronic Science, Hokkaido University.

There are many people who have been involved in this project. Without their support, the project would never have been realized. I am indebted to my research supervisor, Professor Tohru Ifukube from the Research Institute for Electronic Science, Hokkaido University. He inspired me to get involved in this quite interesting research subject and also provided me with the opportunity to continue the project for five whole years. I would like to express my sincere gratitude to him.

I would like to thank Professor Yoshinao Aoki from the Graduate School of Engineering, Hokkaido University. He has financially supported me as a caring father in addition to offering me valuable advice about how to perform excellent research.

I would like to thank Associate Professor Nobuhiro Miki from the Graduate School of Engineering, Hokkaido University. Both his graduate students and he have helped to expand my technical knowledge which was necessary to accomplish the project.

I would like to thank Professor Kunio Takaya from the Department of Electrical Engineering, University of Saskatchewan, Canada, and professors, staff members, and students from the Telecommunications Research Laboratories (TRLabs), Saskatchewan, Canada for the many educational experiences and the excellent working atmosphere that I was able to enjoy during my stay in Saskatoon, Canada as a visiting researcher from 1998 to 1999.

I would like to express my gratitude to two scholarship organizations, Nihon Ikueikai from 1995 to 1998 and the Japan Society for the Promotion of Science for Young Scientists from 1998 to the present for their financial support. I would also like to express my gratitude to the Ministry of Education, Culture and Science of Japan for providing a grant toward this research project. These financial aids have served to strongly promote my research activities, such that I was able to obtain meaningful results from the project.

All of my friends helped me by creating a very inspiring and pleasant atmosphere. Without hesitation, they have been willing to serve as subjects in the psychoacoustic experiments. I sincerely appreciate all of their help.

Most importantly, I would like to thank my family for all the support that I have received. No single statement could express my sincere gratitude for their help.

Bibliography

[1] N. Aoki and T. Ifukube, "Fractal modeling of fluctuations in the steady part of sustained vowels for high quality speech synthesis," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol.E81-A, no.9, pp.1803–1810, 1998.

[2] 青木　直史, 伊福部　達, "持続発声母音における振幅ゆらぎ及びピッチゆらぎの周波数特性とその音響心理的効果," 電子情報通信学会論文誌 A, vol.J82-A, no.5, pp.649–657, 1999.

[3] 青木　直史, 伊福部　達, "合成持続発声母音の自然性改善を目的とした音源波形揺らぎの生成とその主観的および客観的評価," 電子情報通信学会論文誌 D-II, vol.J82-D-II, no.5, pp.843–852, 1999.

[4] N. Aoki and T. Ifukube, "Analysis and perception of spectral 1/f characteristics of amplitude and period fluctuations of normal sustained vowels," Journal of the Acoustical Society of America, vol.106, no.1, pp.423–433, 1999.

[5] 青木　直史, 伊福部　達, "母音音源信号のランダムフラクタル性を利用した合成音声の自然性改善," 電子情報通信学会論文誌 A, vol.J82-A, no.9, pp.1437–1445, 1999.

[6] N. Aoki, T. Ifukube, and K. Takaya, "Implementation of MELP vocoder using lifting wavelet transform," IEEE Region 10 Conf. TENCON, Cheju, South Korea, pp.194–197, Sep.15–17, 1999.

[7] N. Aoki and T. Ifukube, "Enhancing the naturalness of synthesized speech using the random fractalness of vowel source signals," Electronics and Communications in Japan: Part 3. (to be published soon)

[8] N. Aoki, "Development of a rule-based speech synthesis system for the Japanese language using a MELP vocoder," European Signal Processing Conference, Tampere, Finland, pp.1–4, Sep.5–8, 2000.

[9] R.J. Baken, "Irregularity of vocal period and amplitude: A first approach to the fractal analysis of voice," J. voice, vol.4, no.3, pp.185–197, 1990.

[10] H. Herzel, D. Berry, I.R. Titze, and M. Saleh, "Analysis of vocal disorders with methods from nonlinear dynamics," J. Speech and Hearing Research, vol.37, pp.1008–1019, 1994.

[11] A.J. Rozsypal and B.F. Millar, "Perception of jitter and shimmer in synthetic vowels," J. Phon., vol.7, pp.343–355, 1979.

[12] H. Hollien, J. Michel, and E.T. Doherty, "A method for analyzing vocal jitter in sustained phonation," J. Speech and Hearing Research, vol.31, pp.485–490, 1988.

[13] E.T. Doherty and T. Shipp, "Tape recorder effects on jitter and shimmer extraction," J. Phon., vol.1, pp.85–91, 1975.

[14] Y. Horii, "Some statistical characteristics of voice fundamental frequency," J. Speech and Hearing Research, vol.18, pp.192–201, 1975.

[15] Y. Horii, "Fundamental frequency perturbation observed in sustained phonation," J. Speech and Hearing Research, vol.22, pp.5–19, 1979.

[16] Y. Horii, "Vocal shimmer in sustained phonation," J. Speech and Hearing Research, vol.23, pp.202–209, 1980.

[17] I.R. Titze, Y. Horii, and R.C. Scherer, "Some technical considerations in voice perturbation measurements," J. Speech and Hearing Research, vol.30, pp.252–260, 1987.

[18] I.R. Titze and H. Liang, "Comparison of F0 extraction methods for high-precision voice perturbation measurements," J. Speech and Hearing Research, vol.36, pp.1120–1133, 1993.

[19] N.B. Pinto and I.R. Titze, "Unification of perturbation measures in speech signals," J. Acoust. Soc. Am., vol.87, no.3, pp.1278–1289, 1990.

[20] R.C. Scherer, V.J. Vail, and C.G. Guo, "Required number of tokens to determine representative voice perturbation values," J. Speech and Hearing Research, vol.38, pp.1260–1269, 1995.

[21] P. Lieberman, "Perturbations in vocal pitch", J. Acoust. Soc. Am., vol.33, no.5, pp.597–603, 1961.

[22] P. Lieberman, "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges," J. Acoust. Soc. Am., vol.35, no.3, pp. 344–353, 1963.

[23] R.F. Coleman, "Effect of median frequency levels upon the roughness of jittered stimuli," J. Speech and Hearing Research, vol.12, pp.330–336, 1969.

[24] R.F. Coleman, "Effect of waveform changes upon roughness perception," Folia phoniat., vol.23, pp.314–322, 1971.

[25] Y. Koike, "Vowel amplitude modulations in patients with laryngeal diseases", J. Acoust. Soc. Am., vol.45, no.4, pp. 839–844, 1969.

[26] Y. Koike, "Application of some acoustic measures for the evaluation of laryngeal dysfunction," Studia Phonologica, vol.7, pp.17–23, 1973.

[27] H. von Leden and Y. Koike, "Detection of laryngeal disease by computer technique," Arch. Otolaryng, vol.91, pp.3–10, 1970.

[28] H. von Leden, P. Moore, and R. Timcke, "Laryngeal vibrations: Measurements of the glottic wave," A. M. A. Archives of Otolaryngology, Vol.71, pp.26–45, 1960.

[29] K. Kitajima and W.J. Gould, "Vocal shimmer in sustained phonation of normal and pathologic voice," Ann. Otol. Rhinol. Laryngol., vol.85, pp.377–381, 1976.

[30] R.W. Wendahl, "Laryngeal analog synthesis of harsh voice quality," Folia phoniat., vol.15, pp. 241–250, 1963.

[31] R.W. Wendahl, "Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness," Folia phoniat., vol.18, pp.98–108, 1966.

[32] M.H.L. Hecker and E.J. Kreul, "Descriptions of the speech of patients with cancer of the vocal folds. Part I: Measures of fundamental frequency," J. Acoust. Soc. Am., vol.49, no.4, pp.1275–1282, 1971.

[33] E.J. Kreul and M.H.L. Hecker, "Descriptions of the speech of patients with cancer of the vocal folds. Part II: Judgments of age and voice quality," J. Acoust. Soc. Am., vol.49, no.4. pp.1283–1286, 1971.

[34] S. Bielamowicz, J. Kreiman, B.R. Gerratt, M.S. Dauer, and G.S. Berke, "Comparison of voice analysis systems for perturbation measurement," J. Speech and Hearing Research, vol.39, pp.126–134, 1996.

[35] D. Martin, J. Fitch, and V. Wolfe, "Pathologic voice type and the acoustic prediction of severity," J. Speech and Hearing Research, vol.38, pp.765–771, 1995.

[36] G. de Krom, "Some spectral correlates of pathological breathy and rough voice quality for different types of vowel fragments," J. Speech and Hearing Research, vol.38, pp.794–811, 1995.

[37] M.P. Gelfer, "Fundamental frequency, intensity, and vowel selection: Effects on measures of phonatory stability," J. Speech and Hearing Research, vol.38, pp.1189–1198, 1995.

[38] H. Kasuya, S. Ogawa, Y. Kikuchi, and S. Ebihara, "An acoustic analysis of pathological voice and its application to the evaluation of laryngeal pathology," Speech Commun., vol.5, pp.171–181, 1986.

[39] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," J. Acoust. Soc. Am., vol.80, no.5, pp.1329–1334, 1986.

[40] H. Kasuya, Y. Kobayashi, T. Kobayashi, and S. Ebihara, "Characteristics of pitch and amplitude perturbations in pathologic voice," Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, pp.1372–1375, 1983.

[41] C.R. Howard, "Pitch perturbation detection," IEEE Transactions on Audio, vol.AU-13, pp.9–14, 1965.

[42] E. Yumoto and W.J. Gould, "Harmonics-to-noise ratio as an index of the degree of hoarseness," J. Acoust. Soc. Am., vol.71, no.6, pp.1544–1550, 1982.

[43] N. Hiraoka, Y. Kitazoe, H. Ueta, S. Tanaka, and M. Tanabe, "Harmonic-intensity analysis of normal and hoarse voices," J. Acoust. Soc. Am., vol.76, No.6, pp.1648–1651, 1984.

[44] H. Kojima, W.J. Gould, A. Lambiase, and N. Isshiki, "Computer analysis of hoarseness," Acta Otolaryngol, vol.89, pp.547–554, 1980.

[45] N. Yanagihara, "Significance of harmonic changes and noise components in hoarseness," J. Speech and Hearing Research, vol.10, pp.531–541, 1967.

[46] Y. Qi, "Time normalization in voice analysis," J. Acoust. Soc. Am., vol.92, no.5, pp. 2569–2576, 1992.

[47] Y. Qi, B. Weinberg, N. Bi, and W.J. Hess, "Minimizing the effect of period determination on the computation of amplitude perturbation in voice," J. Acoust. Soc. Am., vol.97, no.4, pp.2525–2532, 1995.

[48] J. Hillenbrand, R.A. Cleveland, and R.L. Erickson, "Acoustic correlates of breathy vocal quality," J. Speech and Hearing Research, vol.37, pp.769–778, 1994.

[49] J. Hillenbrand, "A Methodological study of perturbation and additive noise in synthetically generated voice signals," J. Speech and Hearing Research, vol.30, pp.448–461, 1987.

[50] T. Baer, "Effect of single-motor-unit firings on fundamental frequency of phonation," J. Acoust. Soc. Am., vol.64, Suppl., no.1, p.S90, 1978.

[51] F. Klingholtz, "Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels," J. Acoust. Soc. Am., vol.87, no.5, pp.2218–2224, 1990.

[52] A.G. Askenfelt and B. Hammarberg, "Speech waveform perturbation analysis: A perceptual-acoustical comparison of seven measures," J. Speech and Hearing Research, vol.29, pp.50–64, 1986.

[53] M.N. Vieira, F.R. McInnes, and M.A. Jack, "Analysis of the effects of electroglottographic baseline fluctuation on the F0 estimation in pathological voices," J. Acoust. Soc. Am., vol.99, no.5, pp.3171–3178, 1996.

[54] L. Dolansky and P. Tjernlund, "On certain irregularities of voiced-speech waveforms," IEEE Transactions on Audio and Electroacoustics, vol.AU-16, no.1, pp.51–56, 1968.

[55] L. Eskenazi, D.G. Childers, and D.M. Hicks, "Acoustic correlates of vocal quality," J. Speech and Hearing Research, vol.33, pp.298–306, 1990.

[56] I. Pollack, "Amplitude and time jitter thresholds for rectangular-wave trains," J. Acoust. Soc. Am., vol.50, no.4, pp.1133–1142, 1971.

[57] E. Ozimek, J. Konieczny, 鈴木　陽一, 曽根　敏夫, "振幅と周波数が不規則に変化する AM 音の変化の知覚," 日本音響学会聴覚研究会資料, vol.H-96-41, 1996.

[58] M.A. Toner, F.W. Emanuel, and D. Parker, “Relationship of spectral noise levels to psychophysical scaling of vowel roughness,” J. Speech and Hearing Research, vol.33, pp.238–244, 1990.

[59] M.A. Lively and F.W. Emanuel, “Spectral noise levels and roughness severity ratings for normal and simulated rough vowels produced by adult females,” J. Speech and Hearing Research, vol.13, pp.503–517, 1970.

[60] F.E. Sansone, Jr. and F.W. Emanuel, “Spectral noise levels and roughness severity ratings for normal and simulated rough vowels produced by adult males,” J. Speech and Hearing Research, vol.13, pp.489–502, 1970.

[61] S. Hiki, K. Sugawara, and J. Oizumi, “On the rapid fluctuation of voice pitch,” 日本音響学会誌, vol.22, pp.290–291, 1966.

[62] H.H. Heller and S.E. Widnall, “Sound radiation from rigid flow spoilers correlated with fluctuations forces,” J. Acoust. Soc. Am., vol.47, no.3, pp.924–936, 1970.

[63] H.M. Hanson, “Glottal characteristics of female speakers: Acoustic correlates,” J. Acoust. Soc. Am., vol.101, no.1, pp.466–481, 1997.

[64] K.N. Stevens, “Air flow and turbulence noise for fricative and stop consonants: Static considerations,” J. Acoust. Soc. Am., vol.50, no.4, pp.1180–1192, 1971.

[65] C.H. Shadle, “The effect of geometry on source mechanisms of fricative consonants,” J. Phon, vol.19, pp.409–424, 1991.

[66] R.E. Hillman, E. Oesterle, and L.L. Feth, “Characteristics of the glottal turbulent noise source,” J. Acoust. Soc. Am., vol.74, no.3, pp.691–694, 1983.

[67] S. Imaizumi, “Acoustic measures of roughness in pathological voice,” J. Phon., vol.14, pp.457–462, 1986.

[68] 今泉　敏, 比企　静雄, 平野　実, 松下　英明, “サウンド・スペクトログラフによる病的音声の分析,” 日本音響学会誌, vol.36, no.1, pp.9–16, 1980.

[69] 伊福部　達, 橋場　参生, 松島　純一, “母音の自然性における「波形ゆらぎ」の役割,” 日本音響学会誌, vol.47, no.12, pp.903–910, 1991.

[70] 遠藤　康男, 粕谷　英樹, “発声の非定常性を考慮した基本周期および振幅ゆらぎの分析,” 音声言語医学, vol.34, no.4, pp.349–353, 1993.

[71] 遠藤　康男, 粕谷　英樹, “病的音声の分析・変換・合成システムとその応用,” 信学技報, vol.SP95-2, pp.9–14, 1995.

[72] 遠藤　康男, 粕谷　英樹, “基本周期ゆらぎのスペクトル分析とそのモデル化,” 音声言語医学, vol.35, no.2, pp.193–198, 1994.

[73] 小室　修, 粕谷　英樹, “病的母音波形の高調波ゆらぎの分析,” 日本音響学会誌, vol.47, no.2, pp.85–91, 1991.

[74] 小室　修, 粕谷　英樹, “基本周期のゆらぎの性質とそのモデル化に関する検討,” 日本音響学会誌, vol.47, no.12, pp.928–934, 1991.

[75] 粕谷　英樹, “声の音響的評価,” 音声言語医学, vol.31, no.3, pp.331–334, 1990.

[76] 粕谷　英樹, “音響分析による音声の評価,” 音声言語医学, vol.29, no.2, pp.194–199, 1988.

[77] 粕谷　英樹, “音声言語医学領域の用語とその解説 (I),” 音声言語医学, vol.31, pp.344–348, 1990.

[78] 垣田　有紀, 平間　淳司, 浜谷　清也, 大谷　盛人, 鈴木　浩恭, “音声波の基本周波数と最大振幅のゆらぎ,” 日本音響学会音声研究会資料, vol.S85-103, pp.805–812, 1986.

[79] 平間　淳司, 垣田　有紀, “ゆらぎのパワースペクトルに着目した典型的な Rough の病的音声 1 例の特徴,” 音声言語医学, vol.30, no.3, pp.225–230, 1989.

[80] 平間　淳司, 垣田　有紀, “F0 のゆらぎが病的音声のカテゴリ (rough, voice tremor) に与える影響,” 音声言語医学, vol.33, no.1, pp.2–10, 1992.

[81] 小林　哲則, 関根　英敏, “合成音の自然性に対する基本周期の揺らぎの役割,” 日本音響学会誌, vol.47, no.8, pp.539–544, 1991.

[82] 加藤　比呂子, 河原　英紀, “統計的時系列モデルを用いた基本周波数のゆらぎに関する解析,” 日本音響学会聴覚研究会資料, vol.H95-93, pp.95–100, 1995.

[83] 日本音声言語医学会編, “声の検査法, 基礎編,” 医歯薬出版, 1979.

[84] 日本音声言語医学会編, “声の検査法, 臨床編,” 医歯薬出版, 1979.

[85] 田中　信三, 平野　実, “声帯振動の検査,” 音声言語医学, vol.31, no.3, pp.309–315, 1996.

[86] 北嶋　和智, “声門上下圧差の基本周波数におよぼす影響,” 音声言語医学, vol.34, no.4, pp.374–379, 1993.

[87] 岩田　重信, 竹内　健二, 岩田　義弘, 小島　秀嗣, 大山　俊廣, 高須　昭彦, “空気力学的に見た発声機序,” 音声言語医学, vol.36, no.3, pp.338–349, 1995.

[88] 切替　一郎, 野村　恭也, “新耳鼻咽喉科学,” 南山堂, 1967.

[89] E.C. Ifeachor and B.W. Jervis, “Digital signal processing, a practical approach,” Addison-Wesley, 1993.

[90] C.E. Reid and T.B. Passin, “Signal processing in C,” John Wiley & Sons, 1992.

[91] M.H. Hayes, “Statistical digital signal processing and modeling,” John Wiley & Sons, 1996.

[92] S.K. Mitra and J.F. Kaiser, “Handbook for digital signal processing,” John Wiley & Sons, 1993.

[93] R.N. Bracewell, “Two-dimensional Imaging,” Prentice-Hall, 1995.

[94] B.P. Lathi, 山中　惣之助/宇佐美　興一訳, “通信方式,” マグロウヒル, 1977.

[95] 小川　吉彦, “信号処理の基礎,” 朝倉書店, 1991.

[96] 小池　慎一, “C による科学技術計算,” CQ 出版, 1994.

[97] 辻井　重男, 鎌田　一雄, “ディジタル信号処理,” 昭晃堂, 1990.

[98] 佐川　雅彦, 貴家　仁志, “高速フーリエ変換とその応用,” 昭晃堂, 1993.

[99] 三谷　政昭, “ディジタルフィルタデザイン,” 昭晃堂, 1987.

[100] 三上　直樹, “ディジタル信号処理プログラミング入門,” CQ 出版, 1993.

[101] 石田　義久, 鎌田　弘之, “ディジタル信号処理のポイント,” 産業図書, 1989.

[102] 石田　義久, 鎌田　弘之, “DSP 活用のポイント,” 産業図書, 1990.

[103] C. Marven, G. Ewers, 山口　博久訳, “ディジタル信号処理の基礎,” 丸善, 1995.

[104] 越川　常治, “信号解析入門,” 近代科学社, 1992.

[105] 日野　幹雄, “スペクトル解析,” 朝倉書店, 1977.

[106] S. ヘイキン, 武部　幹訳, “適応フィルタ入門,” 現代工学社, 1987.

[107] A. パポーリス, 町田　東一/村田　忠夫訳, “アナログとディジタルの信号解析,” 現代工学社, 1982.

[108] J.S. ベンダット, A.G. ピアソル, 得丸　英勝訳, “ランダムデータの統計的処理,” 培風館, 1976.

[109] S. ブラント, 吉城　肇/高橋　秀知/小柳　義夫訳, “データ解析の方法,” みすず書房, 1976.

[110] L.R. Rabiner and R.W. Schafer, “Digital processing of speech signals,” Prentice Hall, 1978.

[111] J.D. Markel and A.H. Gray, Jr., 鈴木　久喜訳, “音声の線形予測,” コロナ社, 1980.

[112] C. Wheddon and R. Linggard, “Speech and language processing,” Chapman and Hall, 1990.

[113] F.J. Owens, “Signal processing of speech,” Macmillan New Electronics, 1993.

[114] P. Foster and T. Schalk, “Speech recognition,” Telecom Library, Inc, 1993.

[115] H.F. Silverman and D.P. Morgan, “The application of dynamic programming to connected speech recognition,” IEEE ASSP Magazine, pp.6–25, Jul., 1990.

[116] A.M. Kondoz, “Digital speech,” John Wiley & Sons, 1994.

[117] J.W.S. Rayleigh, “The theory of sound, vol.1 & 2,” Dover Publications, 1945.

[118] “特集, 圧縮/認識/合成を追及する音声処理の徹底研究,” インタフェース, CQ 出版, Aug., 1998.

[119] 比企　静雄, “音声情報処理,” 東京大学出版会, 1973.

[120] 古井　貞熙, “音響・音声工学,” 近代科学社, 1992.

[121] 古井　貞熙, “ディジタル音声処理,” 東海大学出版会, 1985.

[122] 甘利　俊一, 中川　聖一, 鹿野　清宏, 東倉　洋一, “音声・聴覚と神経回路網モデル,” オーム社, 1990.

[123] 中田　和男, “音声,” コロナ社, 1977.

[124] 今井　聖, “音声信号処理,” 森北出版, 1996.

[125] 今井　聖, “音声認識,” 共立出版, 1995.

[126] 新美　康永, “音声認識,” 共立出版, 1979.

[127] 永田　邦一, “電子音響工学,” 朝倉書店, 1987.

[128] 三浦　種敏, “聴覚と音声,” 電子情報通信学会, 1980.

[129] 早坂　寿雄, “音の歴史,” 電子情報通信学会, 1989

[130] 早坂　寿雄, “楽器の科学,” 電子情報通信学会, 1992.

[131] 中島　博美, “聴感と音声,” 放送技術双書, 日本放送協会編, 1960.

[132] 藤村　靖, “音声科学,” 東京大学出版会, 1972.

[133] 天外　伺朗, “ディジタル・オーディオの謎を解く,” ブルーバックス, 講談社, 1987.

[134] 中島　平太郎, “オーディオに強くなる,” ブルーバックス, 講談社, 1973.

[135] 橋本　尚, “楽器の科学,” ブルーバックス, 講談社, 1979.

[136] F.R. コナー, 関口　利男/辻井　重男訳, “フィルタ回路入門,” 森北出版, 1990.

[137] F.R. コナー, 関口　利男/辻井　重男/高原　幹夫訳, “変調入門,” 森北出版, 1985.

[138] 藤井　信生, “なっとくする電子回路,” 講談社, 1994.

[139] 藤岡　繁夫, “PA 音響システム,” 工学図書, 1996.

[140] ドン. ディビス, キャロライン. ディビス, 進藤　武男訳, “サウンドシステムエンジニアリング,” 誠文堂新光社, 1992.

[141] D. Mellor, “How to set up a home recording studio,” PC Publishing, 1990.

[142] 大杉　正明, “What’s New?,” NHK ラジオ英会話リスニング CD, 株式会社ディーエイチシー, 東京, 1997.

[143] T.E. Tremain, “The government standard linear predictive coding algorithm: LPC-10,” Speech Technology, pp.40–49, 1982.

[144] M. Nelson, “The data compression book,” M & T Publishing, Inc., 1992.

[145] M.R. Schroeder and B.S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.937–940, 1985.

[146] D.W. Griffin and J.S. Lim, "Multiband excitation vocoder," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.36, no.8, pp.1223–1235, 1988.

[147] R.J. McAuley and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-34, no.4, pp.744–754, 1986.

[148] R. Gonzalez, "Hypermedia data modeling, coding, and semiotics," Proceedings of the IEEE, vol.85, No.7, pp.1111–1140, Jul., 1997.

[149] W.B. Kleijn and J. Haagen, "A speech coder based on decomposition of characteristic waveforms," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.508–511, 1995.

[150] B.S. Atal, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., vol.47, p.65 (A), 1970.

[151] B.S. Atal and N. David, "On synthesizing natural-sounding speech by linear prediction," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.44–47, 1979.

[152] B.S. Atal and S.L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am., vol.50, no.2, pp.637–655, 1971.

[153] B.S. Atal and J.R. Remede, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.614–617, 1982.

[154] L.R. Rabiner, B.S. Atal, and M.R. Sambur, "LPC prediction error-analysis of its variation with the position of the analysis frame," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-25, no.5, pp. 434–442, 1977.

[155] J. Makhoul, R. Viswanathan, R. Schwartz, and A.W.F. Huggins, "A mixed-source model for speech compression and synthesis," J. Acoust. Soc. Am., vol.64, no.6, pp.1577–1581, 1978.

[156] J. Makhoul, "Linear prediction: A tutorial review," Proceedings of the IEEE, vol.63, no.4, pp.561–580, 1975.

[157] S.Y. Kwon and A.J. Goldberg, “An enhanced LPC vocoder with no voiced/unvoiced switch,” IEEE transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-32, no.4, pp.851–858, 1984.

[158] J.D. Markel, “Digital inverse filtering-A new tool for formant trajectory estimation,” IEEE transactions on Audio and Electroacoustics, vol.AU-20, no. 2, pp.129–137, 1972.

[159] G.S. Kang and S.S. Everett, “Improvement of the excitation source in the narrow-band linear prediction vocoder,” IEEE transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-33, no.2, pp.377–386, 1985.

[160] L.J. Siegel and A.C. Bessey, “Voiced/unvoiced/mixed excitation classification of speech,” IEEE transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-30, no.3, pp.451–460, 1982.

[161] O. Fujimura, “An approximation to voice aperiodicity,” IEEE transactions on Audio and Electroacoustics, vol.AU-16, no.1, pp.68–72, 1968.

[162] A.V. McCree and T.P. Barnwell III, “A mixed excitation LPC vocoder model for low bit rate speech coding,” IEEE transactions on Speech and Audio Processing, vol.3, no.4, pp.242–250, 1995.

[163] A.V. McCree and T.P. Barnwell III, “Improving the performance of a mixed excitation LPC vocoder in acoustic noise,” Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, vol.II, pp.137–140, 1992.

[164] A.S. -Nielsen, “Characterizing human ability to discriminate talkers over low data rate voice coders,” J. Acoust. Soc. Am., vol.100, no.4, pt.2, Oct., p.2762, 1996.

[165] United States Department of Defense Digital Voice Processing Consortium, “http://www.plh.af.mil/ddvpc/implemnt.html,” 1998.

[166] United States Department of Defense Digital Voice Processing Consortium, “http://www.plh.af.mil/ddvpc/melp.htm,” 1998.

[167] 古井　貞熙, 守谷　健弘, 匂坂　芳典, 中川　聖一, “1992 年音響・音声・信号処理国際会議 (ICASSP-92) 報告,” 日本音響学会誌, vol.48, no.7, pp.524–525, 1992.

[168] 小澤　一範, 荒関　卓, 小野　茂, “マルチパルス駆動型音声符号化法の検討,” 信学技報, vol.CS82-161, 1983.

[169] 小澤　一範, 小野　茂, 荒関　卓, “マルチパルス駆動型音声符号化法の品質改善,” 音声研資, vol.S83-78, 1984.

[170] 西口　正之, “いかにして高品質な音をコンパクトに表すか,” 日本音響学会誌, vol.54, no.7, pp.527–532, 1998.

[171] 守谷　健弘, “音声符号化,” 電子情報通信学会, 1998.

[172] 三木　弼一, “MPEG-4 のすべて,” 工業調査会, 1998.

[173] 安田　浩, “マルチメディア符号化の国際標準,” 丸善, 1991.

[174] 安田　浩, 渡辺　裕, “ディジタル画像圧縮の基礎,” 日経 BP 出版センター, 1996.

[175] “マルチメディア標準テキストブック, 基礎要素技術/システム編,” 画像情報教育振興協会, 1997.

[176] B. Gold, “Computer program for pitch extraction,” J. Acoust. Soc. Am. vol.34, no.7, pp.916–921, 1962.

[177] C.M. Harris and M.R. Weiss, “Pitch extraction by computer processing of high-resolution Fourier analysis data,” J. Acoust. Soc. Am., vol.35, pp.339–343, 1963.

[178] M.R. Schroeder, “Period histogram and product spectrum: New methods for fundamental-frequency measurement,” J. Acoust. Soc. Am., vol.43, no.4, pp.829–834, 1968.

[179] K. Steiglitz, “On the simultaneous estimation of poles and zeros in speech analysis,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-25, no.3, pp.229–234, 1997.

[180] L.B. Jackson, “Noncausal ARMA modeling of voiced speech,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.37, no.10, pp.1606–1608, 1989.

[181] W.R. Gardner and B.D. Rao, “Noncausal all-pole modeling of voiced speech,” IEEE Transactions on Speech and Audio Processing, vol.5, no.1, pp.1–10, 1997.

[182] H. Wakita, “New methods of analysis in speech acoustics,” Phonetica, vol.37, pp.87–108, 1980.

[183] G. Fant, “The relations between area functions and the acoustic signal,” Phonetica, vol.37, pp.55–86, 1980.

[184] O. Fujimura, "Modern methods of investigation in speech production," Phonetica, vol.37, pp.38–54, 1980.

[185] L. Pakula and S. Kay, "Simple proofs of the minimum phase property of the prediction error filter," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-31, no.2, p.501, 1983.

[186] M. Nakatsui and J. Suzuki, "Method of observation of glottal-source wave using digital inverse filtering in time domain," J. Acoust. Soc. Am., pp.664–665, 1970.

[187] G.E. Kopec, A.V. Oppenheim, and J.M. Tribolet, "Speech analysis by homomorphic prediction," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-25, no.1, pp.40–49, 1977.

[188] L.B. Almeida and J.M. Tribolet, "Nonstationary spectral modeling of voiced speech," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-31, no.3, pp.664–678, 1983.

[189] S. David and B. Ramamurthi, "Two-sided filters for frame-based prediction," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.39, no.4, pp.789–794, 1991.

[190] F. Grandori, P. Pinelli, P. Ravazzani, F. Ceriani, G. Miscio, F. Pisano, R. Colombo, S. Insalaco, G. Tognola, "Multiparametric analysis of speech production mechanisms," IEEE Engineering in Medicine and Biology, April/May, pp.203–209, 1994.

[191] J.L. Flanagan and M.G. Saslow, "Pitch discrimination for synthetic vowels," J. Acoust. Soc. Am., vol.30, no.5, pp.435–442, 1958.

[192] A.E. Rosenberg, "Effect of pitch averaging on the quality of natural vowels," J. Acoust. Soc. Am., vol.44, no.6, pp.1592–1595, 1968.

[193] D.H. Klatt and L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am., vol.87, no.2, pp.820–857, 1990.

[194] D.H. Klatt, "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am., vol.67, no.3, pp.971–995, 1980.

[195] D.H. Klatt, "Review of text-to-speech conversion for English," J. Acoust. Soc. Am., vol.82, no.3, pp.737–793, 1987.

[196] B.R. Fink and R.J. Demarest, "Laryngeal biomechanics," Harvard University Press, 1978.

[197] G.E. Peterson and J.E. Shoup, "A physiological theory of phonetics," J. Speech and Hearing Research, vol.9, pp.5–67, 1966.

[198] G. Fant, "The voice source-theory and acoustic modeling," Vocal Fold Physiology, pp.453–464, Edited by Titze and Scherer, Published by the Denver Center for the Performing Arts, 1983.

[199] G. Fant and Q. Lin, "Frequency domain interpretation and derivation of glottal flow parameters," STL-QPSR, vol.2, no.3, pp.1–21, 1988.

[200] G. Fant and K. Gustafson, "LF-frequency domain analysis," TMH-QPSR, vol.2, pp.135–138, 1996.

[201] G. Fant, J. Liljencrants, and Q.G. Lin, "A four-parameter model of glottal flow," STL-QPSR, vol.4, no.3, pp.1–13, 1985.

[202] K.N. Stevens, "Physics of laryngeal behavior and larynx modes," Phonetica, vol.34, pp.264–279, 1977.

[203] K.N. Stevens and A. S. House, "An acoustical theory of vowel production and some of its implications," J. Speech and Hearing Research, vol.4, no.4, pp.303–320, 1961.

[204] K.N. Stevens, "On the quantal nature of speech," J. Phon., vol.17, pp.3–45, 1989.

[205] H. Dudley, "Remaking speech," J. Acoust. Soc. Am., vol.11, no. 2, pp.169–177, 1939.

[206] M.R. Sambur, A.E. Rosenberg, L.R. Rabiner, and C.A. McGonegal, "On reducing the buzz in LPC synthesis," J. Acoust. Soc. Am., vol.63, no.3, pp.918–924, 1978.

[207] A.E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am., vol.49, no.2, pp.583–590, 1971.

[208] P. Hedelin, "A glottal LPC-vocoder," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.1.6.1–1.6.4, 1984.

[209] H.K. Dunn, J.L. Flanagan, and P.J. Gestrin, "Complex zeros of a triangular approximation to the glottal wave," J. Acoust. Soc. Am., vol.34, pp.1977–1978 (A), 1962.

[210] K.E. Cummings and M.A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," J. Acoust. Soc. Am., vol.98, no.1, pp.88–98, 1995.

[211] M.V. Mathews, J.E. Miller, and E.E. David, Jr., "Pitch synchronous analysis of voiced sounds," J. Acoust. Soc. Am., vol.33, no.2, pp.179–186, 1961.

[212] R.B. Monsen and A.M. Engebretson, "Study of variations in the male and female glottal wave," J. Acoust. Soc. Am., vol.62, no.4, pp.981–993, 1977.

[213] H.W. Strube, "Determination of the instant of glottal closure from the speech wave," J. Acoust. Soc. Am., vol.56, no.5, pp.1625–1629, 1974.

[214] D.Y. Wong and J.D. Markel, "An excitation function for LPC synthesis which retains the human glottal phase characteristics," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.171–174, 1978.

[215] D.Y. Wong, J.D. Markel and A.H. Gray, Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," IEEE transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-27, no.4, pp.350–355, 1979.

[216] A.M. Sulter and H.P. Wit, "Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age," J. Acoust. Soc. Am., vol.100, no.5, pp.3360–3373, 1996.

[217] M.R. Schroeder, "Waveforms and radiation patterns," In Number Theory in Science and Communication, Springer-Verlag, pp.289–300, 1997.

[218] D.G. Childers and C.K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am., vol.90, no.5, pp.2394–2410, 1991.

[219] D.G. Childers and K. Wu, "Quality of speech produced by analysis-synthesis," Speech Communication, vol.9, pp.97–117, 1990.

[220] D.G. Childers and H.T. Hu, "Speech synthesis by glottal excited linear prediction," J. Acoust. Soc. Am., vol.96, no.4, pp.2026–2036, 1994.

[221] D.G. Childers, D.M. Hicks, G.P. Moore, L. Eskenazi, and A.L. Lalwani, "Electroglottography and vocal fold physiology," J. Speech and Hearing Research, vol.33, pp.245–254, 1990.

[222] N.B. Pinto, D.G. Childers, and A.L. Lalwani, "Formant speech synthesis: Improving production quality," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.37, no.12, pp.1870–1887, 1989.

[223] D.B. Pisoni, H.C. Nusbaum, and B.G. Greene, "Perception of synthetic speech generated by rule," Proceedings of the IEEE, vol.73, no.11, pp.1665–1676, 1985.

[224] P. Alku and E. Vilkman, "Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering," J. Acoust. Soc. Am., vol.98, no.2, pt.1, pp.763–767, 1995.

[225] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," Speech Communication, vol.11, pp.109–118, 1992.

[226] H.R. Javkin, N.A. -Barroso, and I. Maddieson, "Digital inverse filtering for linguistic research," J. Speech and Hearing Research, vol.30, pp.122–129, 1987.

[227] M. Rothenberg and J.J. Mahshie, "Monitoring vocal fold abduction through vocal fold contact area," J. Speech and Hearing Research, vol.31, pp.338–351, 1988.

[228] M. Rothenberg, "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," J. Acoust. Soc. Am., vol.53, no.6, pp.1632–1645, 1973.

[229] M. Rothenberg, "Measurement of airflow in speech," J. Speech and Hearing Research, vol.20, pp.155–176, 1977.

[230] M. Rothenberg, "Source-tract acoustic interaction in breathy voice," Vocal Fold Physiology, Edited by Titze and Scherer, Published by the Denver Center for the Performing Arts, pp.465–481, 1983.

[231] P. Milenkovic, "Least mean square measures of voice perturbation," J. Speech and Hearing Research, vol.30, pp.529–538, 1987.

[232] P. Milenkovic, "Glottal inverse filtering by joint estimation of an AR system with a linear input model," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-34, no.1, pp.28–42, 1986.

[233] P.H. Milenkovic, "Voice source model for continuous control of pitch period," J. Acoust. Soc. Am., vol.93, no.2, pp.1087–1096, 1993.

[234] M.M. Thomson and B.J. Guillemin, "Design and performance of an algorithm for estimating vocal system parameters," IEEE Transactions on Speech and Audio Processing, vol.2, no.4, pp.531–536, 1994.

[235] R.L. Miller, "Nature of the vocal cord wave," J. Acoust. Soc. Am., vol.31, no.6, pp.667–677, 1959.

[236] J.N. Holmes, "Formant excitation before and after glottal closure," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.39–42, 1976.

[237] J.N. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," IEEE transactions on Audio and Electroacoustics, vol.AU-21, no.3, pp.298–305, 1973.

[238] I.R. Titze, "The human vocal cords: A mathematical model, part I," Phonetica, vol.28, pp.129–170, 1973.

[239] I.R. Titze, "The human vocal cords: A mathematical model, part II," Phonetica vol.29, pp.1–21, 1974.

[240] W.A. Conrad and D.M. McQeen, "Two-mass model of the larynx: Vocal fold vibration as a negative differential resistance oscillation," J. Acoust. Soc. Am. Suppl.1, vol.79, Spring, p.S82, 1986.

[241] D.R. Allen and W.J. Strong, "A model for the synthesis of natural sounding vowels," J. Acoust. Soc. Am., vol.78, no.1, pp.58–69, 1985.

[242] X. Pelorson, A. Hirschberg, R.R. van Hassel, and A.P.J. Wijnands, "Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model," J. Acoust. Soc. Am., vol.96, no.6, pp.3416–3431, 1994.

[243] J.L. Flanagan, "Some properties of the glottal sound source," J. Speech and Hearing Research, vol.1, pp.99–116, 1958.

[244] J.L. Flanagan, K. Ishizaka, and K.L. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," Bell Syst. Tech. J., vol.54, no.3, pp.485–506, 1975.

[245] J.L. Flanagan, "Source-system interaction in the vocal tract," Ann N.Y. Acad. Sci. vol.155, pp.9–17, 1968.

[246] J.L. Flanagan and L.L. Landgraf, “Self-oscillating source for vocal-tract synthesizers,” IEEE transactions on Audio and Electroacoustics, vol.AU-16, no.1, pp.57–64, 1968.

[247] K. Ishizaka and J.L. Flanagan, “Synthesis of voiced sounds from a two-mass model of the vocal cords,” Bell Syst. Tech. J., vol.51, no.6, pp.1233–1268, 1972.

[248] 石坂　謙三, ジェームズ L. フラナガン, “声帯の自励振動モデル,” 日本音響学会誌, vol.34, no.3, pp.122–131, 1978.

[249] 河原　英紀, “聴覚の情景分析が生み出した高品質 VOCODER: STRAIGHT,” 日本音響学会誌, vol.54, no.7, pp.521–526, 1998.

[250] 濱上　知樹, “音源波形形状を高調波位相により制御する音声合成方式,” 日本音響学会誌, vol.54, no.9, pp.623–631, 1998.

[251] 大村　浩, “声帯振動による非線形性を考慮した振幅制御型音声合成方式,” 信学技報, vol.SP95-78, pp.39–46, 1995.

[252] 武田　昌一, 淺川　吉章, 市川　熹, “残差音源利用分析合成方式とマルチパルス法の基本特性の比較検討,” 信学論誌 A, vol.J73-A, no.11, pp.1735–1742, 1990.

[253] 櫛木　好明, 今井　良彦, 高井　紀代, 新居　康彦, 浮穴　浩二, 古屋　正久, “PARCOR 形音声合成 LSI における駆動波の考察,” 信学技報, vol.S81-40, 1981.

[254] 赤嶺　政巳, 籠嶋　岳彦, 土屋　勝美, “高品質 LPC 分析残差駆動合成器,” 音声言語情報処理, no.17-16, pp.91–96, 1997.

[255] 藤崎　博也, マッツ. ユンクヴィスト, “声帯音源波形の新しいモデルとその音声分析への応用,” 信学論誌 D-II, vol.J72-D-II, no.8, pp.1109–1117, 1989.

[256] H. Fujisaki and M. Ljungqvist, “Proposal and evaluation of models for the glottal source waveform,” Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.1605–1608, 1986.

[257] 丁　文, 粕谷　英樹, “ARX 音声生成モデルと声道・音源パラメータ推定,” 信学技報, vol.SP95-3, pp.15–22, 1995.

[258] 松下　貴光, 丁　文, 楊　長盛, 粕谷　英樹, “フォルマント型音声合成における音源パラメータの制御法,” 日本音響学会講演論文集, pp.353–354, Sep, 1995.

[259] 矢頭　隆, “テキスト音声合成技術の最新状況,” インタフェース, CQ 出版, pp.161–168, Dec., 1996.

[260] 矢頭　隆, “テキスト音声合成システム作成の実際,” インタフェース, CQ 出版, pp.197–204, Jan., 1997.

[261] 本多　清志, “音声言語の生成,” BME, vol.7, no.8, pp.2–10, 1993.

[262] D.F. プロクター, 原田　康夫訳, “呼吸，発声，歌唱,” 西村書店, 1995.

[263] レイ D. ケント, チャールズ. リード, 荒井　隆行/菅原　勉訳, “音声の音響分析,” 海文堂, 1996.

[264] T. Kientzle, “A programmer’s guide to sound,” Addison-Wesley Developers Press, 1998.

[265] R. Cook, “Real-time sound processing,” Dr. Dobb’s Journal, vol.23, no.10, pp.50–55, oct., 1998.

[266] M. Andrews, “The incredible sound machine,” Addison Wesley, 1993.

[267] S.H. Baker and D. Mark, “Macworld, Mac programming FAQs,” IDGBooks Worldwide, Inc., 1996.

[268] S.T. Pope and G. van Rossum, “Machine tongues XVIII: A child’s garden of sound file formats,” Computer Music Journal, vol.19, no.1, pp.25–63, 1995.

[269] “Macintosh, Developer’s Journal, No.5,” 技術評論社, 1993.

[270] “Macintosh, Developer’s Journal, No.16,” 技術評論社, 1995.

[271] “Macintosh, Developer’s Journal, No.17,” 技術評論社, 1996.

[272] “Macintosh, Developer’s Journal, No.18,” 技術評論社, 1996.

[273] “Macintosh, Developer’s Journal, No.20,” 技術評論社, 1996.

[274] “Macintosh, Developer’s Journal, No.22,” 技術評論社, 1996.

[275] “C Magazine,” ソフトバンク, Dec., 1996.

[276] “C Magazine,” ソフトバンク, Oct., 1997.

[277] デーブ. マーク, 滝沢　徹/牧野　祐子訳, “Macintosh プロフェッショナルプログラミング,” トッパン, 1995.

[278] 赤松　正行, “マルチメディアプログラミング,” ビー・エヌ・エヌ, 1993.

[279] D. マーク, C. リード, 藤村　行俊/笠　和子訳, “Macintosh C プログラミング,” アジソンウェスレイ・トッパン, 1993.

[280] P.M. Embree, B. Kimble, “C language algorithms for digital signal processing,” Prentice-Hall, 1991.

[281] P.M. Embree, “C algorithms for real-time DSP,” Prentice-Hall, 1995.

[282] R. Chassaing, “Digital signal processing with C and the TMS320C30,” John Wiley & Sons, 1992.

[283] R. Chassaing and D.W. Horning, “Digital signal processing with the TMS320C25,” John Wiley & Sons, 1990.

[284] H.V. Sorensen and J. Chen, “A digital signal processing laboratory using the TMS320C30,” Prentice-Hall, 1997.

[285] H.M. Ahmed and R.B. Kline, “Recent advances in DSP systems,” IEEE Communications Magazine, pp.32–45, May, 1991.

[286] J.R. Rice, “Computational science and the future of computing research,” IEEE Computational Science & Engineering, pp.35–41, Winter, 1995.

[287] 瀬谷　啓介, “DSP プログラミング入門,” 日刊工業新聞社, 1996.

[288] 磯部　俊夫, “C 言語とアセンブラ,” 工学図書株式会社, 1989.

[289] 楠　菊信, “マイクロプロセッサ,” 丸善, 1994.

[290] 村瀬　康治, “はじめて読むマシン語,” アスキー出版局, 1983.

[291] “テクノロジーがいっぱい, AC’97 パソコンサウンドのディジタル化へ向けて,” ASCII, vol.20, no.8, pp.308–312, 1996.

[292] “トランジスタ技術 SPECIAL, No.21,” CQ 出版, 1990.

[293] “トランジスタ技術 SPECIAL, No.41,” CQ 出版, 1993.

[294] “トランジスタ技術 SPECIAL, No.44,” CQ 出版, 1994.

[295] “インタフェース,” CQ 出版, Jul, 1994.

[296] “電子技術,” 日刊工業新聞社, Dec., 1994.

[297] “電子技術,” 日刊工業新聞社, Nov., 1995.

[298] “電子技術,” 日刊工業新聞社, Nov., 1996.

[299] Texas Instruments, “http://www.ti.com/sc/c62xevm,” 1998.

[300] Texas Instruments, “TMS320C62x/C67x CPU and instruction set reference guide,” 1998.

[301] Texas Instruments, “TMS320C62x/C67x programmer’s guide,” 1998.

[302] Texas Instruments, “TMS320C6201/C6701 peripherals reference guide,” 1998.

[303] Texas Instruments, “TMS320C6x assembly language tools user’s guide,” 1998.

[304] Texas Instruments, “TMS320C6x C source debugger user’s guide,” 1998.

[305] Texas Instruments, “TMS320C6x peripheral support library programmer’s reference,” 1998.

[306] Texas Instruments, “TMS320C6x evaluation module reference guide,” 1998.

[307] Texas Instruments, “TMS320C6x optimizing C compiler user’s guide,” 1998.

[308] R.F. Voss and J. Clarke, “‘$1/f$ noise’ in music and speech,” Nature, vol.258, pp.317–318, 1975.

[309] R.F. Voss and J. Clarke, “ “$1/f$ noise” in music: Music from $1/f$ noise,” J. Acoust. Soc. Am., vol.63, no.1, pp.258–263, 1978.

[310] R.F. Voss, “Random fractals: Self-affinity in noise, music, mountains, and clouds,” Physica D, vol.38, pp.362–371, 1989.

[311] R.F. Voss, “Fractals in nature: From characterization to simulation,” In H.-O. Peitgen and D. Saupe, eds., The Science of Fractals Images, Springer-Verlag, New York, pp.21–70, 1988.

[312] H.F. Olson and H. Belar, “Aid to music composition employing a random probability system,” J. Acoust. Soc. Am., vol.33, no.9, pp.1163–1170, 1961.

[313] C.A. Pickover, “Fractal horizons,” St. Martin’s Press, New York, 1996.

[314] C.A. Pickover, "On the use of symmetrized dot patterns for the visual characterization of speech waveforms and other sampled data," J. Acoust. Soc. Am., vol.80, no.3, pp.955–960, 1986.

[315] D.S. Marzel and M.H. Hayes, "Using iterated function systems to model discrete sequences," IEEE Transactions on Signal Processing, vol.40, no.7, pp.1724–1734, 1992.

[316] A.E. Jacquin, "Image coding based on a fractal theory of iterated contractive image transformations," IEEE Transactions on Image Processing, vol.1, no.1, pp.18–30, 1992.

[317] H.-O. Peitgen, H. Jürgens, and D. Saupe, "Fractals for the classroom, part one," Springer-Verlag, New York, 1992.

[318] S.P.V. Pallati and E.A. Yfantis, "A fast Fourier method for mountain generation," In E.A. Yfantis, ed., Intelligent Systems, pp.885–895, 1995.

[319] A. Watt and M. Watt, "Advanced animation and rendering techniques, theory and practice," Addison-Wesley, 1992.

[320] G. Monro, "Fractal interpolation waveforms," Computer Music Journal, vol.19, no.1, pp.88–98, 1995.

[321] A.A. Verveen and H.E. Derksen, "Fluctuation phenomena in nerve membrane," Proceedings of the IEEE, vol.56, no.6, pp.906-916, 1968.

[322] W.E. Thain, Jr. and J.A. Connelly, "f^{α} noise source streamlines SPICE simulations," IEEE Circuits and Devices, pp.22–27, May, 1996.

[323] S. Oishi and H. Inoue, "Pseudo-random number generators and chaos," Transactions of the IECE of Japan, vol.E65, no.9, pp.534–541, 1982.

[324] Y. Okabe and Y. Nakano, "The theory of KM_2O-Langevin equations and its applications to data analysis (I): Stationary analysis," Hokkaido Mathematical Journal, vol.20, pp.45–90, 1991.

[325] Y. Okabe, "Nonlinear time series analysis based upon the fluctuation-dissipation theorem," Nonlinear Analysis, Theory, Methods & Applications, vol.30, no.4, pp.2249–2260, 1997.

[326] K. Murano, T. Kohda, K. Noda, and M. Yanase, "$1/f$ noise generator using logarithmic and antilogarithmic amplifiers," IEEE transactions on Circuits and Systems-I: Fundamental Theory and Applications, vol.39, no.10, pp.851–853, 1992.

[327] I. Procaccia and H. Schuster, "Functional renormalization-group theory of universal $1/f$ noise in dynamical systems," Physical Review A, vol.28, no.2, pp.1210–1212, 1983.

[328] J. Jimenez, J.A. Moreno, and G.J. Ruggeri, "Forecasting on chaotic time series: A local optimal linear-reconstruction method," Physical Review A, vol.45, no.6, pp.3553–3558, 1992.

[329] Y. Yamamoto and R.L. Hughson, "Extracting fractal components from time series," Physica D, vol.68, pp.250–264, 1993.

[330] H.E. Schepers, J.H.G.M. van Beek, and J.B. Bassingthwaighte, "Four methods to estimate the fractal dimension from self-affine signals," IEEE Engineering in Medicine and Biology, pp.57–64, and 71, 1992.

[331] M. Schroeder, "Fractals, chaos, power laws," W.H. Freeman and Company, 1991.

[332] Y. Shi, "Correlations of pitches in music," Fractals, vol.4, no.4, pp.547–553, 1996.

[333] A.R. Osborne and A. Provenzale, "Finite correlation dimension for stochastic systems with power-law spectra," Physica D, vol.35, pp.357–381, 1989.

[334] D. Saupe, "Algorithms for random fractals," In H.-O. Peitgen and D. Saupe, eds., The Science of Fractals Images, Springer-Verlag, New York, pp.71–136, 1988.

[335] M.S. Keshner, "$1/f$ noise," Proceedings of the IEEE, vol.70, no.3, pp.212–218, 1982.

[336] M. Sato, Y. Murakami, and K. Joe, "Learning chaotic dynamics by recurrent neural networks," Proceedings of the International Conference on Fuzzy Logic and Neural Networks (Iizuka, Japan, July 20–24, 1990), pp.601–604, 1990.

[337] I. Tokuda, Y. Hirai, and R. Tokunaga, "Back-propagation learning of an infinite-dimensional dynamical system," Proceedings of 1993 International Joint Conference on Neural Networks, pp.2271–2275, 1993.

[338] S.S. Narayanan and A.A. Alwan, "A nonlinear dynamics systems analysis of fricative consonants," J. Acoust. Soc. Am., vol.97, No.4, pp.2511–2524, 1995.

[339] 山口　達也, 中川　匡弘, “音声のフラクタル性とその評価法,” 信学技法, vol.SP93-74, DSP93-75, pp.79–86, 1993.

[340] S. Sabanal and M. Nakagawa, “The fractal properties of vocal sounds and their application in the speech recognition model,” Chaos, Solitons & Fractals, vol.7, no.11, pp.1825–1843, 1996.

[341] S. Sabanal and M. Nakagawa, “A study of time-dependent fractal dimensions of vocal sounds,” Journal of the Physical Society of Japan, vol.64, no.9, pp.3226–3238, 1995.

[342] A.Z.R. Langi, K. Soemintapura, and W. Kinsner, “Multi-fractal processing of speech signals,” Proc. IEEE Int. Conf. Information, Communication, and Signal Processing, Singapore, pp.527–531, Sept., 1997.

[343] M.R. Schroeder, “Auditory paradox based on fractal waveform,” J. Acoust. Soc. Am., vol.79, No.1, pp.186–189, 1986.

[344] J.-C. Risset, “Pitch and rhythm paradoxes: Comments on “Auditory paradox based on fractal waveform” [J. Acoust. Soc. Am. 79, 186–189 (1986)],” J. Acoust. Soc. Am., vol.80, No.3, pp.961–962, 1986.

[345] B. Picinbono, “Fast algorithm for Brownian matrices,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol.ASSP-31, no.2, pp.512–514, 1983.

[346] B.B. Mandelbrot and J.W. van Ness, “Fractional Brownian motions, fractional noises and applications,” SIAM Review, vol.10, no.4, pp.422–437, 1968.

[347] B.B. Mandelbrot and J.R. Wallis, “Computer experiments with fractional Gaussian noises. Part1, averages and variances,” Water Resource Research, vol.5, no.1, pp.228–241, 1969.

[348] B.B. Mandelbrot and J.R. Wallis, “Computer experiments with fractional Gaussian noises. Part2, rescaled ranges and spectra,” Water Resource Research, vol.5, no.1, pp.242–259, 1969.

[349] B.B. Mandelbrot and J.R. Wallis, “Computer experiments with fractional Gaussian noises. Part3, mathematical appendix,” Water Resource Research, vol.5, no.1, pp.260–267, 1969.

[350] ベンワー. マンデルブロ, 広中　平祐訳, “フラクタル幾何学,” 日経サイエンス社, 1985.

[351] 安斎　利洋, “ベンワー・マンデルブロ「フラクタル幾何学」,” InterCommunication, no.17, p.144, NTT 出版, 1996.

[352] マーチン. ガードナー, 一松　信訳, “フラクタル音楽,” 丸善, 1996.

[353] 徳永　隆治, “フラクタル,” ジャストシステム, 1993.

[354] 高安　秀樹, “フラクタル,” 朝倉書店, 1986.

[355] J. ブリッグス, F.D. ピート, 高安　秀樹/高安　美佐子訳, “鏡の伝説,” ダイヤモンド社, 1991.

[356] 石村　貞夫, 石村　園子, “フラクタル数学,” 東京図書, 1990.

[357] 荒川　賢一, “フラクタルで絵を描く，絵を見る,” 電子情報通信学会誌, vol.80, no.11, pp.1162–1166, 1997.

[358] A.K. デュードニー, 山崎　秀記訳, “フラクタル山脈, フラクタル植物などコンピュータ・グラフィックス,” 別冊サイエンス 92, コンピュータレクリエーション II, 遊びの探索, 日経サイエンス, pp.115–120, 1989.

[359] “フラクタル,” コンピュータグラフィックス, 技術系 CG 標準テキストブック, 財団法人画像情報処理教育振興協会, pp.128–133, 1995.

[360] 武者　利光, “ゆらぎの世界,” ブルーバックス, 講談社, 1980.

[361] 武者　利光, “ゆらぎの科学 3,” 森北出版, 1993.

[362] 武者　利光, “ゆらぎの科学 6,” 森北出版, 1996.

[363] 武者　利光, “ゆらぎの科学 7,” 森北出版, 1997.

[364] 武者　利光, “ゆらぎの科学 8,” 森北出版, 1998.

[365] 武者　利光, “ゆらぎの発想,” NHK 出版, 1994.

[366] 武者　利光, “$1/f$ 雑音,” 別冊・数理科学, ゆらぎ・カオス・フラクタル, サイエンス社, pp.27–31, 1994.

[367] T. Musha and M. Yamamoto, “$1/f$-like fluctuations of biological rhythm,” Int. Conf. Noise in Physical Systems, pp.22–31, 1995.

[368] 米沢　富美子, “ブラウン運動,” 共立出版, 1986.

[369] 佐藤　浩, “乱流,” 共立出版, 1982.

[370] 奥野　治雄, 高橋　静昭, 小林　敏孝, 辻　陽一, “創造の科学,” コロナ社, 1996.

[371] 長町　三生, “快適科学,” 海文堂, 1992.

[372] 三宅　浩次, 高橋　延昭, 神山　昭男, 大友　詔雄, “生物リズムの構造,” 富士書院, 1992.

[373] 高安　秀樹, “生体におけるカオス，フラクタルの意味,” バイオメカニズム学会誌, vol.19, no.2, pp.83–89, 1995.

[374] 吉田　倫幸, “脳波におけるカオスと心理現象,” バイオメカニズム学会誌, vol.19, no.2, pp.97–104, 1995.

[375] 吉田　倫幸, “脳波のゆらぎ計測と快適評価,” 日本音響学会誌, vol.46, no.11, pp.914–919, 1990.

[376] “脳と情報, ニューロサイエンス,” 別冊・数理科学, サイエンス社, 1989.

[377] 佐野　雅己, “時系列データからのカオスの診断法をめぐって,” バイオメカニズム学会誌, vol.19, no.2, pp.90–96, 1995.

[378] 斎藤　信彦, “カオス: 科学に与えたインパクト,” 日経サイエンス, no.3, pp.18–25, Mar., 1992.

[379] 合原　一幸, “ひろがる工学への応用,” 日経サイエンス, no.3, pp.26–33, Mar., 1992.

[380] 合原　一幸, 五百旗頭　正, “カオス応用システム,” 朝倉書店, 1995.

[381] 合原　一幸, 徳永　隆治, “カオス応用戦略,” オーム社, 1993.

[382] 小林　考次郎, “アルゴリズムとカオス!?,” Computer Today, vol.7, no.32, pp.17–22, 1989.

[383] 津田　一郎, “脳の情報処理とカオス,” Computer Today, vol.7, no.32, pp.23–28, 1989.

[384] 西江　弘, “生命現象を説明するカオス,” 日経サイエンス, no.3, pp.34–39, 1992.

[385] 山本　光璋, “生体 $1/f$ ゆらぎ研究の現状,” BME, vol.8, no.10, pp.1–4, 1994.

[386] 野崎　大地, 山本　義春, “生体の $1/f^{\beta}$ ゆらぎとその解析法,” BME, vol.8, no.10, pp.5–12, 1994.

[387] 河原　剛一, “生体リズムゆらぎの機能的意義と $1/f$ ゆらぎの個体発生,” BME, vol.8, no.10, pp.22–28, 1994.

[388] 中尾　光之, 山本　光璋, “生体 $1/f$ ゆらぎ現象とそのメカニズム,” 信学会誌, vol.79, no.1, pp.62–64, 1996.

[389] 香田　徹, “カオスの間接的時系列解析法とその応用,” システム/制御/情報, vol.37, no.11, pp.661–668, 1993.

[390] 長島　知正, “生体における自己組織化とカオス,” システム/制御/情報, vol.37, no.11, pp.647–653, 1993.

[391] 中川　雅文, 市川　銀一郎, “朗読音声の $1/f$ ゆらぎの検討,” 医用電子と生体工学, vol.35, no.1, pp.1–6, 1997.

[392] “神はカオスに宿りたもう,” ASCII, vol.20, no.9, pp.300–307, 1996.

[393] T.P. Barnwell, III, “Objective measures for speech quality testing,” J. Acoust. Soc. Am., vol.66, no.6, pp.1658–1663, 1979.

[394] R.E. Remez, P.E. Rubin, D.B. Pisoni, and T.D. Carrell, “Speech perception without traditional speech cues,” Science, vol.212, no.22, pp. 947–950, 1981.

[395] L.L. Thurstone, “A law of comparative judgment,” Psychol. Rev., vol.34, pp.273–286, 1927.

[396] F. Mosteller, “Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed,” Psychometrika, vol.16, no.2, pp.207–218, 1951.

[397] E.H. Rothauser, G.E. Urbanek, and W.P. Pachl, “Isopreference method for speech evaluation,” J. Acoust. Soc. Am., vol.44, no.2, pp.408–418, 1968.

[398] E.H. Rothauser, G.E. Urbanek, and W.P. Pachl, “A comparison of preference measurement methods,” J. Acoust. Soc. Am., vol.49, no.2, pp.1297–1308, 1971.

[399] W.D. Voiers, “Individual differences in evaluation of perceived speech qualities,” J. Acoust. Soc. Am., vol.62, Suppl, no.1, Fall, p.S5, 1977.

[400] B.J. McDermott, “Multidimensional analysis of circuit quality judgements,” J. Acoust. Soc. Am., vol.45, no.3, pp.774–781, 1968.

[401] S. Singh and T. Murry, “multidimensional classification of normal voice qualities,” J. Acoust. Soc. Am., vol.64, no.1, pp.81–87, 1978.

[402] J. Kreiman, B.R. Gerratt, G.B. Kempster, A. Erman, and G.S. Berke, “Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research,” J. Speech and Hearing Research, vol.36, pp.21–40, 1993.

[403] B.R. Gerratt, J. Kreiman, N.A. -Barroso, and G.S. Berke, “Comparing internal and external standards in voice quality judgments,” J. Speech and Hearing Research, vol.36, pp.14–20, 1993.

[404] 中山　剛, 三浦　種敏, “音質評価の方法論について,” 日本音響学会誌, vol.22, pp.319–331, 1966.

[405] 粕谷　英樹, “「声の音響学的評価」報告,” 音声言語医学, vol.33, no.1, p.27, 1992.

[406] 曽根　敏夫, “計量心理学の音響学への応用小特集号に寄せて,” 日本音響学会誌, vol.42, no.10, pp.785–786, 1986.

[407] 江端　正直, “聴覚実験における計量心理学の応用,” 日本音響学会誌, vol.42, no.10, pp.794–800, 1986.

[408] 今泉　敏, “声質の計量心理学的評価,” 日本音響学会誌, vol.42, no.10 号, pp.828–833, 1986.

[409] 東京大学教養部統計学教室, “統計学入門,” 東京大学出版会, 1991.

[410] 東京大学教養部統計学教室, “人文・社会科学の統計学,” 東京大学出版会, 1994.

[411] 東京大学教養部統計学教室, “自然科学の統計学,” 東京大学出版会, 1992.

[412] 通商産業省立地公害局, “公害防止の技術と法規,” 公害防止の技術と法規編集委員会編.

[413] 大串　健吾, 中山　剛, 福田　忠彦, “画質と音質の評価技術,”　昭晃堂, 1991.

[414] 佐藤　信, “統計的官能検査法,” 日科技連, 1985.

[415] 倉智　佐一, 山上　暁, “要説心理統計法,” 北大路書房, 1991.

[416] 田中　良久, “心理学研究法 16, 尺度構成,” 東京大学出版会, 1973.

[417] 難波　精一郎, “音色の測定・評価法とその適用例,” 応用技術出版, 1992.

[418] “心理学がわかる,” AERA Mook3, 朝日新聞社, 1994.

[419] 蓑谷　千凰彦, “推定と検定のはなし,” 東京図書, 1988.

[420] 石居　進, “生物統計学入門,” 培風館, 1975.

[421] 市原　清志, “バイオサイエンスの統計学,” 南江堂, 1990.

[422] B.C.J. ムーア, 大串　健吾訳, “聴覚心理学概論,” 誠信書房, 1994.

[423] C.S. Burrus, “Introduction to wavelets and wavelet transforms, a primer,” Prentice Hall, 1998.

[424] C.S. Burrus, J.H. McClellan, A.V. Oppenheim, T.W. Parks, R.W. Schafer, and H.W. Schuessler, “Computer-based exercises for signal processing using MATLAB,” Prentice Hall, 1994.

[425] G.W. Wornell, “Signal processing with fractals, a wavelet-based approach,” Prentice Hall, 1996.

[426] R.K. Young, “Wavelet theory and its application,” Kluwer Academic Publishers, 1993.

[427] R.J. Moorhead II and Z. Zhu, “Signal processing aspects of scientific visualization,” IEEE Signal Processing Magazine, pp.20–41, Sep., 1995.

[428] A. Fournier, “Wavelets and their application to computer graphics,” Siggraph Course Notes 26, 22nd International Conference on Computer Graphics and Interactive Techniques, 1995.

[429] E.J. Stollnitz, T.D. Derose, and D.H. Salesin, “Wavelets for computer graphics,” Morgan Kaufmann Publishers, 1996.

[430] G. Kaiser, “A friendly guide to wavelets,” Birkhäuser, 1994.

[431] O. Rioul and M, Vetterli, “Wavelets and signal processing,” IEEE Signal Processing Magazine, pp.14–38, Oct., 1991.

[432] M. Vetterli and C. Herley, “Wavelets and filter banks: Theory and design,” IEEE Transactions on Signal Processing, vol.40, no.9, pp.2207–2231, 1992.

[433] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense," IEEE Transactions on Image Processing, vol.2, no.2, pp.160–175, 1993.

[434] Z. Xiong, K. Ramchandran, and M.T. Orchard, "Space-frequency quantization for wavelet image coding," IEEE Transactions on Image Processing, vol.6, no.5, pp.677–693, 1997.

[435] A. Manduca, "Compressing images with wavelet/subband coding," IEEE Engineering in Medicine and Biology, pp.639–646, Sep./Oct., 1995.

[436] W.B. Richardson, Jr., "Applying wavelets to mammograms," IEEE Engineering in Medicine and Biology, pp.551–560, Sep./Oct., 1995.

[437] J. Wexler and S. Raz, "Discrete Gabor expansions," Signal Processing, vol.21, pp.207–220, 1990.

[438] G.C.-H. Chuang and C.-C.J. Kuo, "Wavelet descriptor of planar curves: Theory and applications," IEEE Transactions on Image Processing, vol.5, no.1, pp.56–70, 1996.

[439] A.S. Lewis and G. Knowles, "Image compression using the 2-D wavelet transform," IEEE Transactions on Image Processing, vol.1, no.2, pp.244–250, 1992.

[440] M. Antonini, M. Barland, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," IEEE Transactions on Image Processing, vol.1, no.2, pp.205–220, 1992.

[441] M.H. Gross, O.G. Staadt, and R. Gatti, "Efficient triangular surface approximations using wavelets and quad tree data structures," IEEE Transactions on Visualization and Computer Graphics, vol.2, no.2, pp.130–143, 1996.

[442] P. Flandrin, "On the spectrum of fractional Brownian motions," IEEE Transactions on Information Theory, vol.35, no.1, pp.197–199, 1989.

[443] P. Flandrin, "Wavelet analysis and synthesis of fractional Brownian motion," IEEE Transactions on Information Theory, vol.38, no.2, pp.910–917, 1992.

[444] A.H. Tewfik and M. Kim, "Correlation structure of the discrete wavelet coefficients of fractional Brownian motion," IEEE Transactions on Information Theory, vol.38, no.2, pp.904–909, 1992.

[445] S. Kadambe and G.F.B. -Bartels, “Application of the wavelet transform for pitch detection of speech signals,” IEEE Transactions on Information Theory, vol.38, no.2, pp.917–924, 1992.

[446] R.E. Crandall, “Projects in scientific computation,” Springer-Verlag, New York, 1994.

[447] “数値演算, 基礎と計測, 制御での応用,” インタフェース, pp.85–87, CQ 出版, May, 1995.

[448] “特集, ウェーブレットの産業応用,” エレクトロニクス, オーム社, pp.19–43, Nov., 1995.

[449] 榊原　進, “ウェーブレットビギナーズガイド,” 東京電機大学出版局, 1995.

[450] チャールズ K. チュウイ, 桜井　明/新井　勉訳, “ウェーブレット応用, 信号解析のための数学的手法,” 東京電機大学出版局, 1997.

[451] G. ストラング, T. グエン, 高橋　進一/池原　雅章訳, “ウェーブレット解析とフィルタバンク I & II,” 培風館, 1999.

[452] 貴家　仁志, “マルチレート信号処理,” 昭晃堂, 1995.

[453] 貴家　仁志, “よくわかるディジタル画像処理,” CQ 出版, 1996.

[454] 貴家　仁志, 村松　正吾, “マルチメディア技術の基礎, DCT 入門,” CQ 出版, 1997.

[455] 中野　宏毅, 山本　鎭男, 吉田　靖夫, “ウェーブレットによる信号処理と画像処理,” 共立出版, 1999.

[456] 桜井　明, “スプライン関数入門,” 東京電機大学出版局, 1981.

[457] 市田　浩三, 吉本　富士市, “スプライン関数とその応用,” 教育出版, 1979.

[458] 佐藤　雅昭, “ウェーブレット理論の数学的基礎,” 日本音響学会誌, vol.47, no.6, pp.405–423, 1991.

[459] 河原　英紀, “ウェーブレット解析の聴覚研究への応用,” 日本音響学会誌, vol.47, no.6, pp.424–429, 1991.

[460] 寅市　和男, “フルーエンシ解析からのウェーブレット, フラクタル, カオス,” 数理科学, no.363, pp.8–12, Sep., 1993.

[461] 高木　隆司, “形の数理,” 朝倉書店, 1992.

[462] 山口　昌哉, 畑　政義, 木上　淳, “フラクタルの数理,” 岩波書店, 岩波講座応用数学 [対象 7], 1993.

[463] D. Gabor, “Acoustical quanta and the theory of hearing,” Nature, no.4044, pp.591–594, 1947.

[464] C. Roads, “Automated granular synthesis of sound,” Computer Music Journal vol.2, no.2, pp.61–62, 1978.

[465] C. Roads, “Introduction to granular synthesis,” Computer Music Journal vol.12, no.2, pp.11–13, 1988.

[466] B. Truax, “Discovering inner complexity: Time shifting and transposition with a real-time granulation technique,” Computer Music Journal vol.18, no.2, pp.38–48, 1994.

[467] B. Truax, “Real time granular synthesis with a digital signal processor,” Computer Music Journal vol.12, no.2, pp.14–26, 1988.

[468] D.L. Jones and T.W. Parks, “Generation and combination of grains for music synthesis,” Computer Music Journal vol.12, no.2, pp.27–34, 1988.

[469] 莱　孝之, “音楽とコンピュータ,” InterCommunication, no.9, pp.148–159, NTT 出版, 1994.

[470] 岩竹　徹, “コンピュータミュージック,” オーム社, 1994.

[471] H.A. Deutsch, 梯　郁太郎訳, “シンセサイザー,” パイパーズ, 1979.

[472] “Macworld 音楽大全,” 扶桑社, 1993.

[473] 持田　康典, 青木　栄一郎, “楽器とコンピュータ,” 共立出版, 1994.

[474] G. Fant, “What can basic research contribute to speech synthesis?,” J. Phon., vol.19, pp.75–90, 1991.

[475] R. Cole, L. Hirschman, L. Atlas, M. Beckman, A. Biermann, M. Bush, M. Clements, J. Cohen, O. Garcia, B. Hanson, H. Hermansky, S. Levinson, K. McKeown, N. Morgan, D.G. Novick, M. Ostendorf, S. Oviatt, P. Price, H. Silverman, J. Spitz, A. Waibel, C. Weinstein, S. Zahorian, and V. Zue, “The challenge of spoken language

systems: Research directions for the nineties," IEEE Transactions on Speech and Audio Processing, vol.3, no.1, pp.1–21, 1995.

[476] 樽松　明, "自動翻訳電話の基礎研究," ATR Journal, pp.31–38, 1996.

[477] 山崎　泰弘, "高度音声翻訳通信技術の基礎研究," ATR Journal, pp.63–70, 1996.

[478] 藤村　靖, "音声研究の動向," 日本音響学会誌, vol.22, pp.113–118, 1966.

[479] 中田　和男, "音声の分析合成技術の最近の動向," 日本音響学会誌, vol.38, no.10, pp.663–668, 1982.

[480] 中田　和男, "音声合成と符号化技術," 電子情報通信学会誌, vol.78, no.11, pp.1119–1124, 1995.

[481] 藤崎　博也, "音声研究の現状と将来," 日本音響学会誌, vol.34, no.3, pp.117–121, 1978.

[482] 広瀬　啓吉, "音声の出力に関する研究の現状と将来," 日本音響学会誌, vol.52, no.11, pp.857–861, 1996.

[483] 広瀬　啓吉, "音声の規則合成," 電子情報通信学会誌, vol.79, no.6, pp.612–615, 1996.

[484] 管村　昇, "音声インターフェース," InterCommunication, no.12, pp.138–142, NTT 出版, 1992.

[485] 竹林　洋一, "音声インタフェース," bit 別冊ビジュアルインタフェース, 共立出版, pp.131–141, 1996.

[486] 難波　精一郎, "音の科学," 朝倉書店, 1989.

[487] 吉本　千禎, "指で聴く," 北海道大学図書刊行会, 1979.

[488] 吉本　千禎, "人の感性, 機械の感性," 日本経済新聞社, 1981.

[489] 伊福部　達, "音声タイプライタの設計," CQ 出版, 1983.

[490] 伊福部　達, "感覚代行研究から人工現実感技術へ," コンピュータビジョン 95-5, pp. 65–71, 1995.

[491] 伊福部　達, "心理物理実験によるマン・マシン系の評価法," バイオメカニズム学会誌, vol.14, no.3, pp.131–137, 1990.

[492] 船久保　熙康, 初山　泰弘, "福祉工学," 産業図書, 1995.

[493] 本庄　巖, “人口内耳,” 中山書店, 1994.

[494] 大江　精三, “感覚の数学的構造,” 創文出版, 1955.

[495] “実用段階に入った話速変換技術，音声出力機器の基本機能へ,” 日経エレクトロニクス, no.622, pp.93–98, 1994.

[496] 禰寝　義人, “聴覚補助のための話速変換装置,” 電子情報通信学会誌, vol.78, no.5, pp.462–465, 1995.

[497] 野口　春美, “21 世紀のサウンド・キッチン,” InterCommunication, no.9, pp.136–147, NTT 出版, 1994.

[498] J. Allen, “Natural language understanding,” The Benjamin/Cummings Publishing Company, Inc, 1995.

[499] N.A. Lassen and B. Larsen, “Cortical activity in the left and right hemispheres during language-related brain functions,” Phonetica, vol.37, pp.27–37, 1980.

[500] A.R. Damasio and H. Damasio, “Brain and language,” Scientific American, sep., pp.87–95, 1992.

[501] J. Nolte, “The human brain,” The C.V. Mosby Company, 1988.

[502] D.E. ルーメルハート, 御領　謙訳, “人間の情報処理,” サイエンス社, 1979.

[503] 川人　光男, “脳の仕組み,” 読売新聞社, 1992.

[504] ノーバート・ウィーナー, 鎮目　恭夫訳, “科学と神,” みすず書房, 1965.

[505] リチャード. リーキー, 馬場　悠男訳, “ヒトはいつから人間になったか,” 草思社, 1996.

[506] 北原　靖子, 渡辺　千歳, 加藤　知佳子, “ヒトらしさとは何か, バーチャルリアリティ時代の心理学,” 北大路書房, 1996.

[507] 亀田　和夫, “声とことばのしくみ,” 口腔保健協会, 1986.

[508] G.J. Borden and K.S. Harris, 廣瀬　肇訳, “ことばの科学入門,” メディカルリサーチセンター, 1984.

[509] P.B. Denes and E.N. Pinson, 切替　一郎/藤村　靖/神山　五郎/戸塚　元吉訳, “話しことばの科学,” 東京大学出版会.

[510] 東京大学公開講座 9, “言語,” 東京大学出版会, 1967.

[511] 東京大学公開講座 37, “ことば,” 東京大学出版会, 1983.

[512] 一松　信, 村岡　洋一, “感性と情報処理,” 共立出版, 1993.

[513] L. バス, J. クータ, 廣田　年亮訳, “ユーザ・インタフェースのソフトウェア開発,” アジソンウェスレイ・トッパン, 1992.

[514] 岡田　美智男, “口ごもるコンピュータ,” 共立出版, 1995.

[515] ハワード. ラインゴールド, 沢田　博訳, “バーチャル・リアリティ,” ソフトバンク, 1992.

[516] S. オークスタカルニス, D. ブラットナー, 安村　道晃/伊賀　聡一郎/織畑　涼子訳, “シリコン・ミラージュ, 仮想現実の科学と芸術,” トッパン, 1994.

[517] B. Damer, “Avatars!,” Peachpit Press, 1998.

[518] D.S. Ebert, “Texturing and modeling, A procedural approach,” Academic Press, 1994.

[519] N.O. Bernsen, H. Dybkjær, and L. Dybkjær, “What should your speech system say?,” IEEE Computer, pp.25–31, Dec., 1997.

[520] C. シュマント, 石川　泰訳, “コンピュータとのヴォイスコミュニケーション,” サイエンス社, 1995.

Contribution

Paper

1 Naofumi Aoki and Tohru Ifukube, "Fractal modeling of fluctuations in the steady part of sustained vowels for high quality speech synthesis," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol.E81-A, no.9, pp.1803–1810, 1998.

2 青木　直史, 伊福部　達, "持続発声母音における振幅ゆらぎ及びピッチゆらぎの周波数特性とその音響心理的効果," 電子情報通信学会論文誌 A, vol.J82-A, no.5, pp.649–657, 1999.

3 青木　直史, 伊福部　達, "合成持続発声母音の自然性改善を目的とした音源波形揺らぎの生成とその主観的および客観的評価," 電子情報通信学会論文誌 D-II, vol.J82-D-II, no.5, pp.843–852, 1999.

4 Naofumi Aoki and Tohru Ifukube, "Analysis and perception of spectral 1/f characteristics of amplitude and period fluctuations of normal sustained vowels," Journal of the Acoustical Society of America, vol.106, no.1, pp.423–433, 1999.

5 青木　直史, 伊福部　達, "母音音源信号のランダムフラクタル性を利用した合成音声の自然性改善," 電子情報通信学会論文誌 A, vol.J82-A, no.9, pp.1437–1445, 1999.

6 Naofumi Aoki and Tohru Ifukube, "Enhancing the naturalness of synthesized speech using the random fractalness of vowel source signals," Electronics and Communications in Japan: Part 3. (to be published)

7 Naofumi Aoki and Tohru Ifukube, "Development of a rule-based speech synthesis system for the Japanese language using a MELP vocoder," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. (submitted)

8 Alexander Burger, Yoshinao Aoki, Naofumi Aoki, Atsushi Hashimoto, and Junsoek Kim, "Development of an editing environment for implementing an avatar communication system," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences. (submitted)

International conference

9 Naofumi Aoki and Tohru Ifukube, "Two 1/f fluctuations in sustained phonation and their roles on naturalness of synthetic voice," IEEE Int. Conf. Electronics, Circuits, and Systems, Rodos, Greece, pp.311–314, Oct.13–16, 1996.

10 Naofumi Aoki and Tohru Ifukube, "An analysis of fluctuation on sustained vowel waveform from the viewpoint of 1/f shimmers," The Acoustical Society of America and the Acoustical Society of Japan Joint Meeting, Honolulu, U.S.A., pp.913–918, Dec.2–6, 1996.

11 Naofumi Aoki and Tohru Ifukube, "The generation of waveform fluctuation for the enhancement of the quality of sustained vowels," IEEE Int. Conf. Information, Communications, and Signal Processing, Singapore, pp.998–1002, Sep.9–12, 1997.

12 Naofumi Aoki and Tohru Ifukube, "Fractal modeling of fluctuations in sustained vowels for high quality speech synthesis," IEICE Int. Sympo. Nonlinear Theory and its Applications, Honolulu, U.S.A., pp.1185–1188, Nov.29–Dec.2, 1997. (Invited Paper)

13 Naofumi Aoki and Tohru Ifukube, "Fractal modeling of glottal waveform for high quality speech synthesis," 16th Int. Congress on Acoustics and 135th Meeting of the Acoustical Society of America, Seattle, U.S.A., pp.245–246, Jun.20–26, 1998.

14 Naofumi Aoki and Tohru Ifukube, "Fractal interpolation for the modification of pulse source signal in PARCOR synthesizer," Fourth Int. Conf. Signal Processing, Beijing, China, pp.650–653, Oct.12–16, 1998.

15 Yohji Kudoh, Yoshinao Aoki, and Naofumi Aoki, "Investigation of acoustical ray-tracing method for measurement reflective structure in acoustical archaeology," Fourth Int. Conf. Signal Processing, Beijing, China, pp.1399–1402, Oct.12–16, 1998.

16 Kunio Takaya, Li Ding, and Naofumi Aoki, "H.263 facial video image coding with the MELP voice codec," IEEE Int. Workshop on Soft Computing in Industry, Muroran, Japan, pp.228–233, Jun.16–18, 1999.

17 Naofumi Aoki, Tohru Ifukube, and Kunio Takaya, "Modeling the randomness of speech signals for the enhancement of the voice quality of synthesized speech," 15th Int. Conf. Noise in Physical Systems and 1/f Fluctuations, Hong Kong, China, pp.199–202, Aug.23–26, 1999.

18 Naofumi Aoki and Kunio Takaya, "MELP vocoder using lifting wavelet transform," IEEE Pacific Rim Conf. Communications, Computers and Signal Processing, Victoria, Canada, pp.392–395, Aug.22–24, 1999.

19 Naofumi Aoki, Tohru Ifukube, and Kunio Takaya, "Implementation of MELP vocoder using lifting wavelet transform," IEEE Region 10 Conf. TENCON, Cheju, South Korea, pp.194–197, Sep.15–17, 1999.

20 Naofumi Aoki and Kunio Takaya, "Implementation of MELP vocoder using lifting wavelet transform," IEEE Int. Sympo. on Intelligent Signal Processing and Systems, Phuket, Thailand, pp.753–756, Dec.8–10, 1999.

21 Naofumi Aoki, Yoshinao Aoki, Yoji Kudo, Norifumi Tonouchi, and Masafumi Takase, "Acoustical ray tracing system for visualizing the simulated sound reflection from a plane rock wall in acoustical archaeology," 25th Int. Acoustical Imaging Sympo, Bristol, U.k., Mar.19–22, 2000.

22 Kunio Takaya and Naofumi Aoki, "Low bit-rate video by creation of avatars for a certain class of video images – a VRML/Java approach," Int. Conf. Communications, Control and Signal Processing, Jul.25–28, 2000.

23 Naofumi Aoki, "Development of a rule-based speech synthesis system for the Japanese language using a MELP vocoder," European Signal Processing Conference, Tampere, Finland, Sep.5–8, 2000.

24 Naofumi Aoki and Kunio Takaya, "Development of a rule-based speech synthesis system for the Japanese language using a MELP vocoder," 2000 IEEE Int. Sympo, In-

telligent Signal Processing and Communication Systems, Honolulu, U.S.A., Nov.5–8, 2000.

Technical report

25 青木　直史, 伊福部　達, “音源信号の振幅ゆらぎによる合成音の自然性,” 電子情報通信学会技術研究報告音声研究会資料, vol.SP95-137, pp.15–21, Mar.7–8, 1996.

26 青木　直史, 伊福部　達, “1/f ピッチゆらぎによる合成母音の自然性の向上,” 日本音響学会聴覚研究会資料, vol.H-96-42, pp.1–8, Jun.13–14, 1996.

27 青木　直史, 伊福部　達, “1/f ピッチゆらぎおよび 1/f 振幅ゆらぎの母音合成における有効性の検討,” 電子情報通信学会技術研究報告音声研究会資料, vol.SP96-125, pp.25–32, Mar.6–7, 1997.

28 青木　直史, 伊福部　達, “母音における波形ゆらぎの生成とその知覚特性,” 日本音響学会聴覚研究会資料, vol.H-97-43, pp.1–8, Jun.19–20, 1997.

29 青木　直史, 伊福部　達, “合成持続発声母音の自然性改善を目的とした波形ゆらぎの生成とその主観的および客観的評価,” 電子情報通信学会技術研究報告音声研究会資料, vol.SP97-52, pp.9–16, Oct.23, 1997.

30 青木　直史, 伊福部　達, “PARCOR 分析合成システムにおける母音音源波形のウェーブレット圧縮に関する考察,” 電子情報通信学会技術研究報告音声研究会資料, vol.SP97-126, pp.73–80, Mar.5–6, 1998.

31 青木　直史, 伊福部　達, “PARCOR 分析合成における音源信号のウェーブレット解析とそれに基づいたパルス列音源の改善,” 日本音響学会聴覚研究会資料, vol.H-98-49, pp.1–8, Jun.11–12, 1998.

Presentation

32 青木　直史, 伊福部　達, “音源波形ゆらぎに関する一考察,” 電気関係学会北海道支部連合大会, 北海道工業大学 (札幌), Oct.21–22, 1995.

33 青木　直史, 伊福部　達, “合成音声における音源信号の振幅ゆらぎの役割,” 電子情報通信学会総合大会, 東京工業大学 (東京), Mar.28–31, 1996.

34 青木　直史, 伊福部　達, “音源信号の振幅ゆらぎによるバズ音質の改善,” 電子情報通信学会総合大会, 東京工業大学 (東京), Mar.28–31, 1996.

35 青木　直史, 伊福部　達, “1/f ピッチゆらぎによる合成音声の自然性,” 情報処理北海道シンポジウム’96, 北海道大学 (札幌), Apr.18–19, 1996.

36 青木　直史, “持続発声母音におけるピッチゆらぎのモデリング,” Workshop on Speech Production’96, 大滝セミナーハウス (大滝), Jun.17–19, 1996.

37 青木　直史, 伊福部　達, “1/f ゆらぎを用いた Mixed Excitation 音声合成方式の提案,” 電子情報通信学会ソサイエティ大会, 金沢大学 (金沢), Sep.18–21, 1996.

38 青木　直史, 伊福部　達, “持続発声母音合成における 1/f ゆらぎ現象の応用,” 日本音響学会秋季大会, 岡山大学 (岡山), Sep.25–27, 1996. (スペシャルセッション)

39 青木　直史, 伊福部　達, “持続発声母音における 1/f ゆらぎ現象の解析,” 第 10 回ゆらぎ現象研究会, かながわサイエンスパーク (川崎), Nov.15–16, 1996.

40 青木　直史, 伊福部　達, “母音合成と 1/f ゆらぎ,” インターメディアシンポジウム’97, 札幌パークホテル (札幌), Mar.12–13, 1997.

41 青木　直史, 伊福部　達, “1/f ゆらぎによる持続発声母音における波形ゆらぎのモデリング,” 日本音響学会春期大会, 同志社大学 (京都), Mar.17–19, 1997.

42 青木　直史, 伊福部　達, “持続発声母音における波形ゆらぎに関する一考察,” 情報処理北海道シンポジウム’97, 北海道大学 (札幌), May 7–8, 1997.

43 青木　直史, 伊福部　達, “合成持続発声母音の自然性改善を目的とした波形ゆらぎの生成とその知覚特性,” 日本音響学会秋季大会, 北海道大学 (札幌), Sep.17–19, 1997. (スペシャルセッション)

44 青木　直史, 伊福部　達, “合成母音の自然性改善のための波形ゆらぎの生成とその客観的評価,” 電気関係学会北海道支部連合大会, 北見工業大学 (北見), Oct.18–19, 1997.

45 青木　直史, 伊福部　達, “ウェーブレット分析・合成による音源信号の表現に関する一考察,” 電気関係学会北海道支部連合大会, 北見工業大学 (北見), Oct.18–19, 1997.

46 青木　直史, 伊福部　達, “高品質音声合成のための波形ゆらぎのモデリング,” インターメディアシンポジウム’98, 第百生命札幌ビル (札幌), Feb.3–4, 1998.

47 青木　直史, 伊福部　達, “音源波形のウェーブレット圧縮とフラクタル補間による波形復元,” 日本音響学会春期大会, 慶応義塾大学 (横浜), Mar.17–19, 1998.

48 青木　直史, 伊福部　達, “PARCOR 音声合成における音源信号の設計に関する一考察,” 情報処理北海道シンポジウム’98, 北海道大学 (札幌), May 14–15, 1998.

49 Naofumi Aoki and Kunio Takaya, “LPC vocoder working at 2.4 kbps,” Telecommunications Research Laboratories (TRLabs) Technology Forum, Edmonton, Canada, Oct.27–28, 1998.

50 Naofumi Aoki and Tohru Ifukube, “Fractal interpolation for the modification of pulse source signal in PARCOR synthesizer,” IEEE North Saskatchewan Section Annual Symposium, Saskatoon, Canada, Nov.17, 1998.

51 Naofumi Aoki, Tohru Ifukube, and Kunio Takaya, “PARCOR synthesizer with random fractal modification,” The College Research Day of the Department of Engineering of the University of Saskatchewan, Saskatoon, Canada, Nov.27, 1998.

52 青木　直史, 伊福部　達, 高谷　邦夫, “DSP 評価ボード TMS320C62 を用いた MELP ボコーダの実現,” 電子情報通信学会総合大会, 慶応義塾大学 (横浜), Mar.25–28, 1999.

53 岡田　一秀, 青木　直史, 谷山　一彦, 加藤　光彦, “実音声信号を用いたハードウェアによる自動シミュレーションの検討,” 電子情報通信学会総合大会, 慶応義塾大学 (横浜), Mar.25–28, 1999.

54 青木　直史, 高谷　邦夫, 伊福部　達, “MELP ボコーダのリアルタイム実現,” 情報処理北海道シンポジウム’99, 北海道大学 (札幌), May 12–13, 1999.

55 青木　直史, “DSP 評価ボード TMS320C62 を用いた MELP ボコーダの実現,” 第 2 回 Aoki Digital Media Academy, 札幌アスペンホテル (札幌), Jun.24, 1999.
(http://www.media.eng.hokudai.ac.jp/~aoki/adma)

56 青木　直史, 伊福部　達, “MELP 方式による音声の規則合成に関する一考察,” 電子情報通信学会ソサイエティ大会, 日本大学 (船橋), Sep.7–10, 1999.

57 青木　直史, 伊福部　達, “MELP 方式による音声の規則合成システムの構築,” 電気関係学会北海道支部連合大会, 室蘭工業大学 (室蘭), Oct.23–24, 1999.

58 青木　直史, 青木　由直, “ギター演奏の支援を目的とした運指呈示システムの構築,” 電気関係学会北海道支部連合大会, 室蘭工業大学 (室蘭), Oct.23–24, 1999.

59 青木　直史, 青木　由直, “ギター演奏支援のためのタブ譜作成自動化に関する研究,” 電子情報通信学会総合大会, 広島大学 (広島), Mar.28–31, 1999.

60 青木　直史, “テキストスピーチの開発と音声信号処理教育,” 2000 年 PC カンファレンス, 北海道大学 (札幌), Aug.2–4, 2000.