A Study on the Extraction and Modeling of Excitation Information from Speech

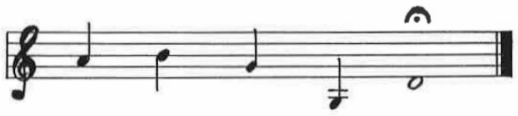
Atsushi Mano[1]

March 3,1988

[1]Department of Electrical Engineering, Graduate School of Science and Engineering, Keio University

Contents

Introduction



—This is the sequence of musical tones used in the film "Close Encounters of the Third Kind" when humans receive an audio signal from outer space. I still vividly remember the climactic scene in which these tones are played in an attempt to establish contact with extraterrestrial beings, culminating in a dramatic close encounter. Perhaps these very tones were, in fact, a form of speech for the extraterrestrials.—

[1] Speech and Speech Signal Processing

Human speech exists for communication. Based on the concepts of information theory proposed by Shannon[1], speech can be represented in terms of message content, or information. The acoustic waveform that carries this message information is the speech signal, and speech communication is accomplished through this signal. In speech production, message information is converted in the brain into a set of neural signals that control the articulatory mechanism, which in turn produces a speech waveform. The resulting acoustic wave propagates through the air and reaches the human ear. It is then conveyed from the auditory organs through the auditory nerve to the listener's brain, where the speech information contained in the acoustic wave is recognized and understood.

Modern society has developed sophisticated scientific and technological means for extending speech communication over distance and time and for enabling its storage and repetition. These technologies are collectively referred to as speech signal processing. Speech signal processing may be considered within a sequence of processes in which: 1. a speech signal is acquired by measuring or observing the speech information source; 2. the acquired speech signal is analyzed and represented; 3. the represented speech signal is transformed; and 4. speech information is extracted and utilized. Within this sequence, speech signal processing can be defined primarily as the representation and transformation of speech signals performed in steps 2 and 3.

The speech information referred to here can be broadly classified into two categories. The first is message information at the cerebral level, represented by phonemes, which constitute a discrete and finite set of symbols. The second consists of information concerning the natural characteristics contained in the continuous speech waveform produced through the articulatory mechanism, including voice quality, intensity, speaker individuality, and emotion. Speech information thus exhibits a considerable degree of diversity, arising from the hierarchical nature of the speech production mechanism and from interactions among psychological, physiological, and physical processes. Consequently, in constructing a speech communication system based on speech signal processing, it is essential, for the transmission, storage, and repetition of speech, to preserve the message information while representing and transforming speech signals in a flexible form that avoids degradation of their naturalness.

[2] The Role of Speech Information in an Advanced Information Society and Speech Information Compression

Since speech is one of the principal means of human communication, speech information plays an extremely important role in society. Particularly in today's increasingly automated office environment, voice communication by telephone, together with numerical and textual information, has become an indispensable means of communication. In recent years, remarkable improvements in the performance and miniaturization of hardware, supported by advances in digital signal processing and LSI technologies, have led to the rapid expansion of digital transmission networks and the digital integration of various services based on concepts such as the ISDN (Integrated Services Digital Network), both in Japan and abroad. The digitalization of such networks offers several important advantages:

1. Digital signal processing permits substantially more sophisticated signal processing than analog techniques and can consequently provide more efficient and economical transmission channels.

2. Digital systems are well suited to information storage, switching, and processing, thereby facilitating more advanced services and the development of new services.

3. By encoding digital signals, accurate transmission can be achieved even over highly noisy channels, resulting in excellent repeater performance.

4. Multiplexing of digital signals makes it possible to integrate transmission and switching functions.

5. Speech, image, and data signals can all be transmitted in a unified format, independently of the type of communication network.

6. Confidentiality of communication, error control, and information protection can be implemented relatively easily.

In digital transmission, the fundamental measure of transmission capacity is the amount of information transmitted per unit time over a communication channel, i.e., the bit rate. Consequently, the early development of information compression technologies is essential for the efficient and economical utilization of networks and for accommodating increasing communication demands. At the present stage of network digitalization, speech signals form a hierarchy based on 64 Kbits/sec per channel, with lower rates of 32 Kbits/sec and 16 Kbits/sec. In contrast, digital signals other than speech, primarily those used for data communication, form a hierarchy based on 9.6 Kbits/sec, with rates of 4.8 Kbits/sec and 2.4 Kbits/sec. Considering that speech, as a means of human communication, is required to provide such functions as immediacy, instruction, and warning, and that it is frequently used in society for administrative communication and operational instructions, the information transmission rate required for speech is considerably higher than that for other forms of information. Reducing the bit rate of speech transmission, and thereby reducing communication costs, is therefore an urgent requirement.

If speech signals can be efficiently encoded at low and medium bit rates of 2.4–9.6 Kbits/sec, medium-quality speech transmission can be achieved in applications such as mobile radio communication.

[3] Efficient Speech Coding and the Position of the Present Study

Efficient speech coding technologies at low and medium bit rates have advanced rapidly in recent years. With the development of signal-processing LSI chips known as Digital Signal Processors (DSPs), coding algorithms previously regarded as too complex for practical implementation can now be implemented in hardware with relative ease. Consequently, extensive research has been conducted with the aim of improving channel utilization and facilitating the early realization of new integrated communication services.

Various efficient speech coding schemes for low and medium bit rates have been proposed[2]; these can broadly be classified into two categories.

The first category comprises waveform coding schemes. These schemes attempt to encode the speech waveform as faithfully as possible within the limits imposed by quantization error and, in principle, can be applied to signals other than speech. Waveform coding schemes provide very high speech quality at high bit rates of 16 Kbits/sec or above. At medium bit rates below 16 Kbits/sec, however, speech quality begins to deteriorate noticeably, and at low bit rates below 9.6 Kbits/sec, the degradation becomes severe.

The second category comprises analysis-synthesis coding schemes, in which speech signals are analyzed on the basis of a speech production model, and the resulting parameters are extracted and encoded. Substantial compression of speech signals requires efficient extraction of speech information from the signal, for which the introduction of a speech production model is essential. A representative and practical analysis-synthesis approach based on such a production model employs a linear prediction model and is generally referred to as LPC (Linear Prediction Coding) analysis-synthesis. Most coding schemes currently considered effective at low and medium bit rates are based on LPC analysis-synthesis and improve coding efficiency by modeling the residual signal corresponding to the excitation in the time domain, or by converting its low-frequency components to baseband and subsequently applying waveform coding.

The present study is concerned with a method for modeling excitation information in an analysis-synthesis system based on a speech production model at low and medium bit rates. Its objectives are to achieve accurate extraction of excitation information in LPC analysis-synthesis, to develop a new method for modeling the excitation based on an excitation-generation model, and, by applying this model, to improve both the speech quality and coding efficiency of LPC analysis-synthesis systems.

[4] Outline of the Present Study

Chapter 1 first provides an overview of efficient speech coding based on speech production models and examines previous studies concerning the extraction and modeling of excitation information in LPC analysis-synthesis systems, clarifying their principal characteristics and limitations.

Chapter 2 describes a method for accurately extracting excitation information in LPC analysis-synthesis systems. The method extracts excitation information from the amplitude-envelope characteristics of the residual signal by applying the principles of LPC analysis-synthesis.

Chapter 3 presents a new method for modeling excitation information in LPC analysis-synthesis systems. It is shown that the amplitude spectrum obtained by pitch-synchronous analysis of the LPC voiced-speech residual exhibits highly distinctive zero characteristics. A modeling method based on inverse LPC analysis of this spectrum is then proposed, and its potential for improving the speech quality of LPC analysis-synthesis systems is discussed. Although modeling the LPC voiced-speech residual has conventionally been considered difficult, it is demonstrated that the introduction of pitch-synchronous analysis makes it possible to assume an excitation-generation model in the frequency domain.

Chapter 4 proposes a pitch-synchronous mel inverse-LSP analysis-synthesis method for efficiently encoding the LPC voiced-speech residual on the basis of the excitation model proposed in Chapter 3. It is shown that, when combined with an LPC analysis-synthesis system, the proposed method enables efficient, high-quality speech coding and synthesis at low bit rates.

Chapter 5 discusses the results obtained in the preceding chapters and presents the conclusions of this study.

Chapter 1

Extraction and Modeling of Excitation Information in Efficient Coding Based on Speech Production Models

1.1 Overview

In the Introduction, the important role of speech information in today's highly developed information society and the necessity of information compression based on efficient speech coding were established.

This chapter first demonstrates the validity of speech information compression based on speech production models and provides an overview of the LPC method as a representative speech coding scheme based on this approach.

In the LPC method, the accuracy with which excitation information is extracted from the residual signal has a major effect on the quality of synthesized speech. Second, therefore, the necessity of extracting and modeling excitation information is discussed. Previous studies in this area are reviewed, their problems are clarified, and the motivation for the present study is introduced.

1.2 Speech Production Models and Information Compression

As stated in the Introduction, speech information is highly diverse, comprising linguistic information, information concerning naturalness, and the interactions between them. None of this information can arise without the speech production mechanism. Therefore, considering that the purpose of speech signal processing in speech coding is to represent and transform speech signals while preserving as much speech information as possible, it is difficult to achieve efficient coding without understanding the speech production mechanism.

During utterance, that is, in the encoding process of speech information, humans produce speech through the integration of articulatory movements. During listening, that is, in the decoding process of speech information, they decode speech by integrating multiple types of speech information on the basis of implicit knowledge of the complex speech production mechanism. The encoding and decoding processes described above contain considerable physical redundancy. An understanding of the speech production mechanism is therefore indispensable for efficiently removing such redundancy and achieving efficient coding.

The speech production process is briefly described below[3]. Broadly speaking, speech production consists of three processes: excitation generation, articulation, and radiation. The speech organs comprise the lung, trachea, larynx, pharynx, nasal cavity, oral cavity, and related structures; the region above the larynx is generally called the vocal tract.

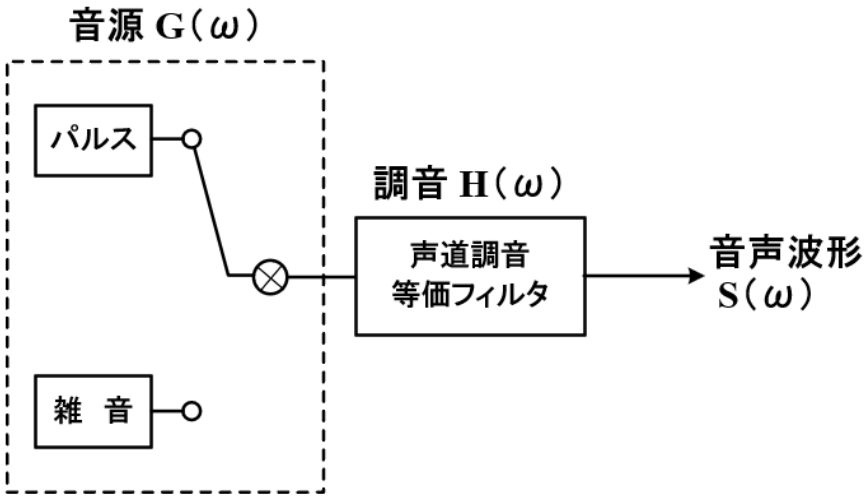


Figure 1.1: The linear separated equivalent circuit model.

Speech production begins with the generation of an excitation, which supplies the energy for speech and may be either a voiced excitation or an unvoiced excitation. The voiced excitation, also called the glottal source, consists of a quasi-periodic train of airflow pulses. It is generated when air expelled from the lungs as the abdominal muscles press the diaphragm passes through the glottis in the larynx, causing the vocal cords to vibrate through the interaction between the restoring force due to muscular elasticity and the Bernoulli force produced by the airflow. The pulse repetition rate is called the fundamental frequency or pitch frequency and is a physical quantity related to vocal-cord tension and subglottal air pressure. Pitch frequency is important speech information that determines such characteristics as speech pitch, accent, and intonation. The unvoiced excitation, in contrast, is generated by turbulence produced when air passes through a narrow constriction formed by the tongue, lips, or other articulators. Airflow generated through these excitation processes resonates in the vocal tract, whose shape is varied by movements of the jaw, tongue, lips, and other articulators, thereby forming phonetic sounds that are radiated from the lips. Speech generated by a voiced excitation is called voiced sound, whereas speech generated by an unvoiced excitation is called unvoiced sound.

The speech production process described above is only an outline and is in reality subject to complex variations, including the effects of the nasal tract. Nevertheless, it incorporates most of the principal characteristics of speech production, and constructing a speech production model on this basis is therefore considered a highly reasonable approach. The purpose of modeling the speech production process is to obtain economical, high-quality synthesized speech by means of a simple and effective model. In view of this objective and on the basis of the speech production process described above, the linear separated equivalent circuit model[4] shown in Fig.1.1 has gained acceptance as the most appropriate model. In this model, the frequency characteristic of the excitation $G(\omega)$ and that of articulation $H(\omega)$ are completely separated, and the frequency characteristic of the speech waveform $S(\omega)$ is generated by connecting their respective electrical equivalent circuits without mutual interaction.

Thus, the frequency characteristic of speech $S(\omega)$ is expressed by Eq.1.1 as follows.

$$S(\omega) = G(\omega)H(\omega) \tag{1.1}$$

In this case, the excitation is ordinarily approximated by pulse and noise sources, whereas articulation is approximated by an all-pole or pole-zero filter. The macroscopic frequency characteristic of the glottal source, together with the radiation characteristic, is incorporated into the articulatory-filter characteristic $H(\omega)$. The excitation characteristic $G(\omega)$ in Eq.1.1 is macroscopically flat.

By introducing such a production model, a speech signal can be accurately described by a small number of parameters, whose temporal variations can be made far slower than those of the speech signal itself. This makes substantial compression of speech information possible. When determining a more specific model on the basis of the linear separated equivalent circuit model described above, the following

conditions must be considered:

1. Model fidelity: how accurately the assumed model can describe the actual speech production process.
2. Model efficiency: how many parameters are required to specify the model.
3. Model analyzability: whether an effective means exists for estimating those parameters from sampled values of an actual speech waveform.

An optimum model can be determined only when these conditions are satisfied.

1.3 Linear Predictive Coding (LPC)

One speech production model based on the linear separated equivalent circuit model described in Section1.2 is the pole-zero model. This model represents the articulatory characteristics of speech most accurately. Numerous methods for estimating pole-zero models on the basis of least-squares estimation have been proposed[5]-[9], and their effectiveness has been confirmed. In least-squares estimation of a pole-zero model, however, solving the equations constitutes a nonlinear problem, requiring least-squares estimation by an iterative approximation method. Furthermore, poles and zeros cannot be estimated entirely independently because they influence one another, making it necessary to select optimum model orders. Because each of these processes requires a large amount of computation, implementation of a speech analysis-synthesis system based on a pole-zero model is costly.

An alternative is linear prediction analysis based on an all-pole model. The concept of linear prediction was first introduced by Wiener[10]; it was first applied to speech analysis-synthesis by Itakura and Saito[11] and by Atal and Schroeder[12], and now forms the theoretical foundation of speech analysis-synthesis.

This analysis method exploits the high correlation among samples of a speech waveform. It represents the current signal S_n as the sum of a predicted value, which is a linear combination of the preceding p samples (approximately ten), and an error signal. The method can be expressed by Eq.1.2.

$$S_n = \sum_{i=1}^{p} a_i S_{n-1} + \epsilon_n \tag{1.2}$$

Here, a_i is called a linear prediction coefficient (LPC coefficient), and ϵ_n is called the prediction-error signal. Determining the LPC coefficients under the condition that the mean-square value of ϵ_n over a prescribed interval is minimized is called Linear Prediction Coding (LPC). To obtain LPC coefficients satisfying this condition, the autocorrelation function r_i of the original speech signal over a prescribed interval is calculated, and the p simultaneous equations called the normal equations, shown in Eq.1.3, are solved using r_i[13][14].

$$\sum_{i=1}^{p} a_i r_{k-i} = -r_k \tag{1.3}$$

Next, the significance of LPC analysis in the frequency domain is considered. Taking the z-transform of both sides of Eq.1.2 gives Eq.1.4.

$$S(z) = \{a_i z^{-1} + \cdots + a_p z^{-p}\} S(z) + E(z) \tag{1.4}$$

Equation1.5 then follows.

$$\begin{aligned} S(z) &= E(z)H(z) \\ H(z) &= \frac{1}{A(z)} \\ A(Z) &= 1 - (a_i \mathrm{z}^{-1} + \cdots + a_p \mathrm{z}^{-p}) \end{aligned} \tag{1.5}$$

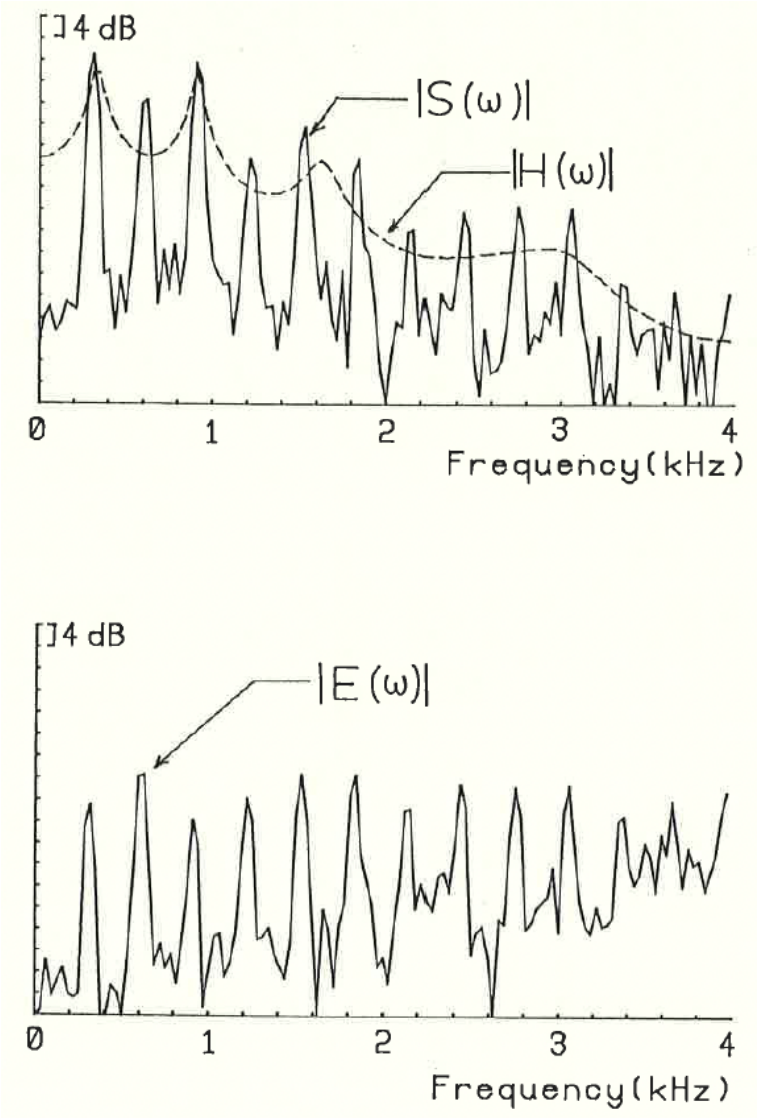


Figure 1.2: Frequency characteristics of $|S(\omega)|$,$|H(\omega)|$,$|E(\omega)|$.

This indicates that $S(z)$ is the output obtained after $E(z)$ is applied to the linear system $H(z)$, and that $H(z)$ is an all-pole model comprising the p poles given by the roots of the pth-order equation $A(z) = 0$, which is determined by the LPC coefficients a_i. Thus, applying LPC analysis to a speech signal is equivalent to assuming a speech production system based on an all-pole model.

When speech is approximated by the all-pole model described above, its poles correspond to the formants of the speech spectrum. The frequency characteristic $H(z)$ in Eq.1.5, determined by the LPC coefficients a_i selected to minimize the prediction error in Eq.1.2, corresponds closely to the frequency-spectrum characteristic of the vocal-tract filter in the linear separated equivalent circuit model of Fig.1.1. The frequency-spectrum characteristic of the excitation in Fig.1.1 corresponds to the frequency-spectrum characteristic $E(z)$ of the prediction-error signal ϵ_n in Eq.1.2, which cannot be approximated by the all-pole model in LPC analysis. Figure1.2 shows the amplitude-spectrum characteristic $|S(\omega)|$ of a speech signal, the amplitude-spectrum characteristic $|H(\omega)|$ of the LPC all-pole approximation filter, and the amplitude-spectrum characteristic $|E(\omega)|$ of the prediction-error signal ϵ_n. As the figure shows, $|H(\omega)|$ obtained by LPC analysis eliminates the effect of the harmonic structure caused by the pitch of the voiced excitation and accurately approximates the formant peaks of the vocal-tract spectrum. Although the amplitude spectrum of the prediction error, $|E(\omega)|$, has an approximately flat envelope, it clearly retains the harmonic-structure components associated with the excitation.

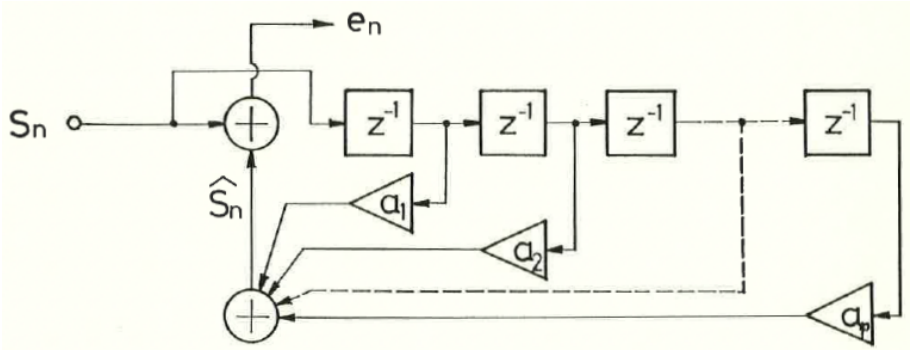


Figure 1.3: The LPC inverse filter.

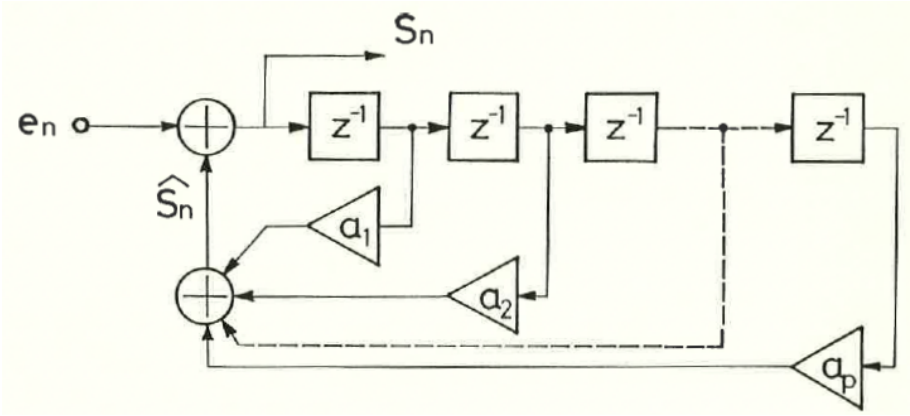


Figure 1.4: The LPC synthesis filter.

1.4 LPC-Based Speech Synthesis and the Extraction and Modeling of Excitation Information from the Residual

The LPC coefficients obtained from the theory described above are system parameters obtained when a speech signal is approximated by an all-pole speech production model over an interval in which the frequency characteristic $S(\omega)$ of the speech signal can be regarded as stationary. These coefficients make it possible to extract articulatory characteristics as speech information, and efficient speech coding can be achieved by encoding them.

Equations1.2 and 1.5 in Section1.3 can be rewritten as Eqs.1.6 and 1.7, respectively.

$$\epsilon_n = S_n - \sum_{i=1}^{p} a_i S_{n-i} \tag{1.6}$$

$$E(z) = \frac{S(z)}{H(z)} = S(z)A(z) \tag{1.7}$$

Thus, as shown in Fig.1.3, the prediction-error signal ϵ_n can be obtained by passing the original speech signal S_n through an inverse filter having the inverse characteristic $A(z)$ of the linear system modeled by LPC analysis.

This signal is also called the residual signal. In addition to its original meaning as the prediction error in LPC analysis, it can be regarded, from Eq.1.5 or Fig.1.4, as the input to the system that produces the output S_n, namely, as the excitation.

The residual signal is, however, a complex waveform containing nearly as much information as the original speech signal, and using it directly therefore makes information compression difficult. Accordingly, the residual signal ϵ_n is modeled under an approximation considered reasonable in light of the excitation-generation process. Its properties are first reviewed below.

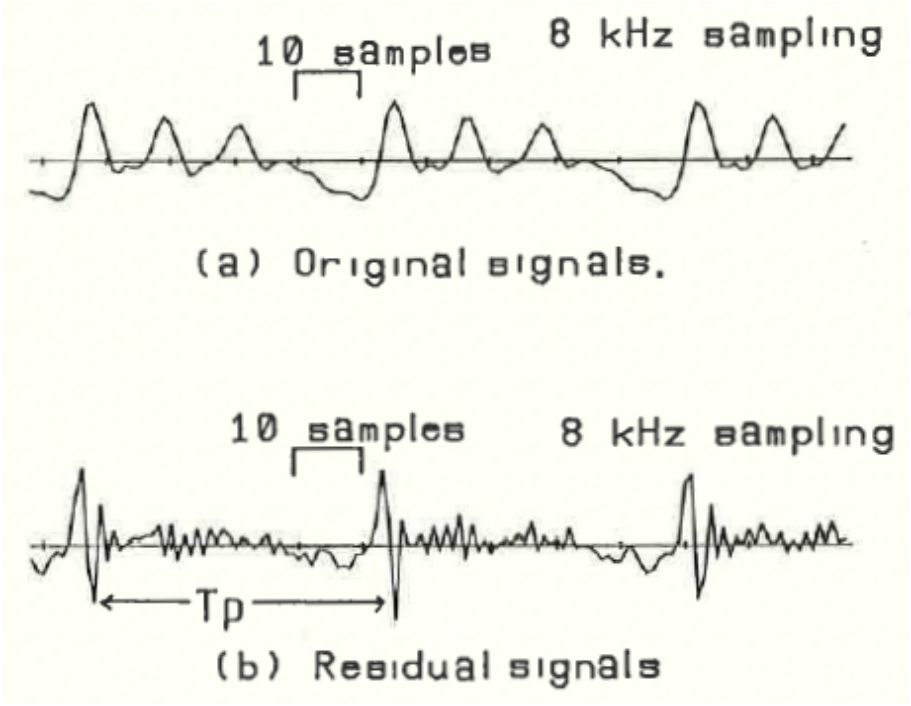


Figure 1.5: Original signals and LPC residual signals.

Figure1.5 shows examples of residual-signal waveforms in the time domain in comparison with the corresponding original speech signals.

If the speech signal were perfectly predicted by LPC analysis, the residual would be a completely random signal with a flat power spectrum. In practice, however, particularly in voiced portions, which account for most speech, the excitation can be regarded as a quasi-periodic pulse train having a period T_p, as described in Section1.2. LPC analysis, in contrast, performs modeling under the assumption that the externally unobservable input (excitation) is a single impulse or white noise. Within the range of the output produced by a single impulse, the residual can therefore be regarded as a random prediction error. When a new impulse is applied at the input, however, the prediction becomes grossly inaccurate and produces a large-amplitude residual signal. Once the input has been applied and the system output enters the interval that can be regarded as the response to that impulse, the residual again becomes the intrinsic prediction error and takes the form of a small-amplitude random signal. This change repeats each time a new pulse produced by vocal-fold vibration is applied to the input. Consequently, as shown in Fig.1.5(b), pulse-like signals occur in the residual with the same period T_p as the pitch pulses. In unvoiced portions, where the input excitation can be approximated by white noise, the residual is also approximately white noise. Thus, the residual contains macroscopic excitation information consisting of periodicity for voiced excitation and white-noise characteristics for unvoiced excitation, and it constitutes a good approximation of the excitation signal. On the basis of this macroscopic information, the simplest effective residual model is the one shown in Fig.1.6, in which voiced portions are approximated by an impulse train with a pitch period and unvoiced portions by white noise.

This model effectively describes excitation generation by vocal-fold vibration and turbulence and adequately satisfies model fidelity, one of the conditions for a good model discussed in Section1.2. Modeling the residual in this way also makes it possible to compress a residual signal over a prescribed interval (several hundred samples) into two parameters, pitch period and amplitude, producing a substantial benefit in terms of model efficiency, another condition for a good model.

The microscopic characteristics of the residual are considered next. From Eq.1.5 in Section1.3, the frequency characteristic $S(\mathrm{e}^{j\omega})$ of a speech signal is expressed, as shown in Eq.1.8, as the product of the frequency characteristic $H(\mathrm{e}^{j\omega})$ of the LPC all-pole approximation filter and the frequency characteristic $E(\mathrm{e}^{j\omega})$ of the prediction residual.

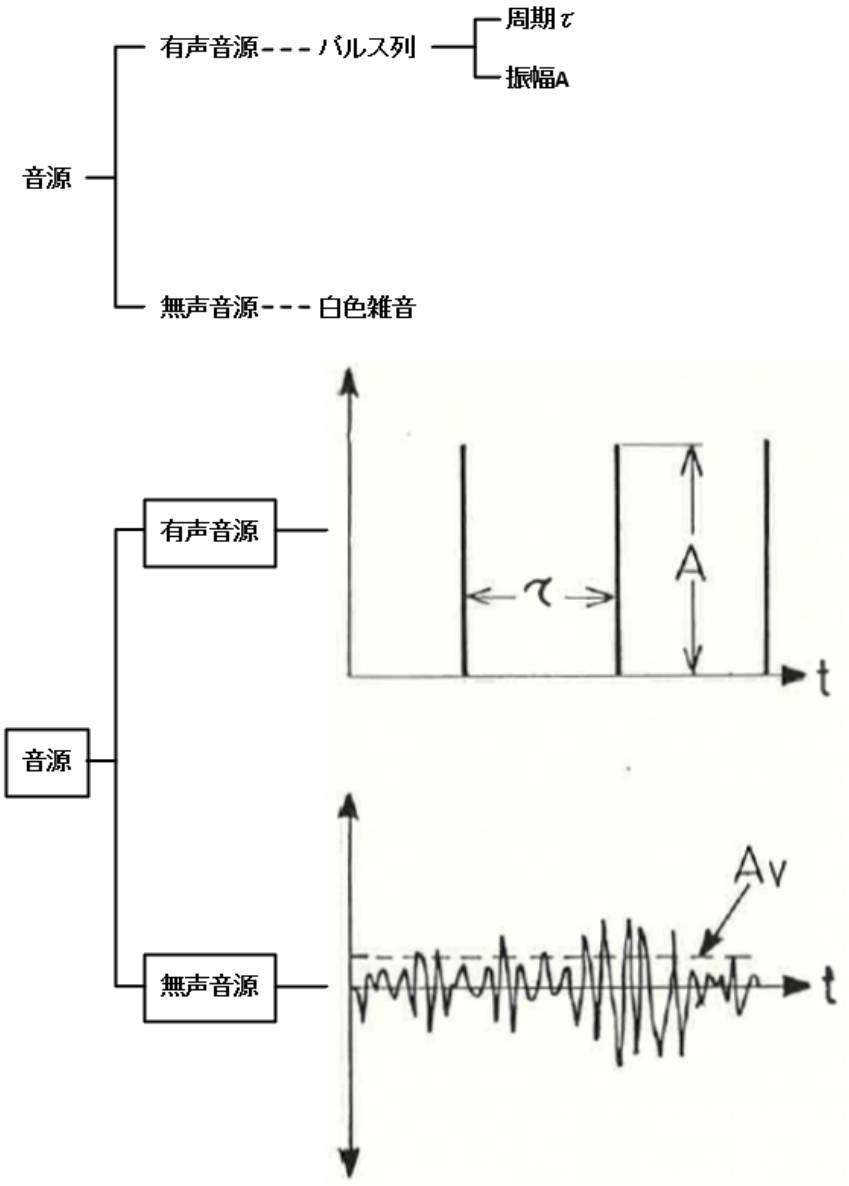


Figure 1.6: Macro-modeling of residual signals.

$$S(\mathrm{e}^{j\omega}) = E(\mathrm{e}^{j\omega})H(\mathrm{e}^{j\omega}) \tag{1.8}$$

Because LPC analysis determines linear prediction coefficients that minimize short-time average residual power, the residual power E_p can be expressed from Eq.1.8 as Eq.1.9, where G is the gain.

$$\begin{aligned} E_p &= \frac{G^2}{2\pi}\int_{-\pi}^{\pi} |E(\mathrm{e}^{j\omega})|^2 \mathrm{d}\omega \\ &= \frac{G^2}{2\pi}\int_{-\pi}^{\pi} \frac{|S(\mathrm{e}^{j\omega})|^2}{|H(\mathrm{e}^{j\omega})|^2} \mathrm{d}\omega \end{aligned} \tag{1.9}$$

Under this error criterion, frequency regions in which $|S(\mathrm{e}^{j\omega})|^2 > |H(\mathrm{e}^{j\omega})|^2$ receive greater weight than regions in which $|S(\mathrm{e}^{j\omega})|^2 < |H(\mathrm{e}^{j\omega})|^2$[15]. Thus, LPC analysis closely approximates formants (peaks) having high power in the speech spectrum. This expresses, in terms of the error criterion, the assumption of an all-pole model in LPC analysis and is also evident from Fig.1.2. The speech spectrum, however, also contains zero characteristics caused by the quasi-periodic opening and closing of the vocal folds and by the articulatory mechanism for nasal sounds[16][17]. As described above, LPC analysis does not closely approximate low-power zeros (valleys). Consequently, although the frequency characteristic $E(\mathrm{e}^{j\omega})$ of the residual $\epsilon(n)$ in Eq.1.8 has an overall flat short-time amplitude spectrum, it also contains, particularly in voiced intervals, zero characteristics representing the difference from the all-pole model. These zero characteristics do not represent the true excitation characteristic of the original speech signal. Nevertheless, if the residual is regarded as the excitation in an analysis-synthesis system based on the linear separated equivalent circuit model, they can be regarded as microscopic excitation information in a broad sense. It should therefore be possible to improve synthesized-speech quality by approximating the residual signal, particularly the voiced residual, with an all-zero model.

The present study aims to improve the quality of speech synthesized by LPC analysis-synthesis systems based on speech production models at low and medium bit rates. It reports the results of research on two principal themes: a method for extracting macroscopic excitation information from LPC residual signals, and a method for extracting microscopic excitation information based on a new model of the residual signal in an LPC analysis-synthesis system.

1.5 Previous Studies and Problems in Excitation Information Extraction and Modeling

This section reviews previous studies of macroscopic and microscopic excitation information extraction, as discussed in Section1.4, and identifies their problems.

It has already been stated that the excitation model shown in Fig.1.6, which is based on the macroscopic excitation information of periodicity in voiced excitation and white-noise characteristics in unvoiced excitation, satisfies the two conditions of model fidelity and model efficiency. To be a good model, however, it must also satisfy the condition of model analyzability described in Section1.2. In other words, the problem of how to extract the modeled excitation information from an actual residual signal must be solved. In particular, extraction of the pitch period, including voiced/unvoiced classification in Fig.1.6, has been an important problem since the invention of speech synthesis, and numerous studies have been conducted[18].

Among the methods examined, the autocorrelation-function method has traditionally received the greatest attention and has been regarded as effective. For a residual signal ϵ_n, the rth-order autocorrelation function R_τ is calculated from Eq.1.10, where N is the analysis interval.

$$R_\tau = \sum_{n=0}^{N-1-\tau} \epsilon_n \epsilon_{n+\tau} \tag{1.10}$$

Because sharp peaks synchronized with the pitch period T_p occur in the residual signal during voiced intervals, as shown in Fig.1.5(b), the autocorrelation function R_τ calculated by Eq.1.10 attains its maximum at $\tau = T_p$, as shown in Fig.1.7.

The basic pitch-extraction procedure based on this principle is as follows.

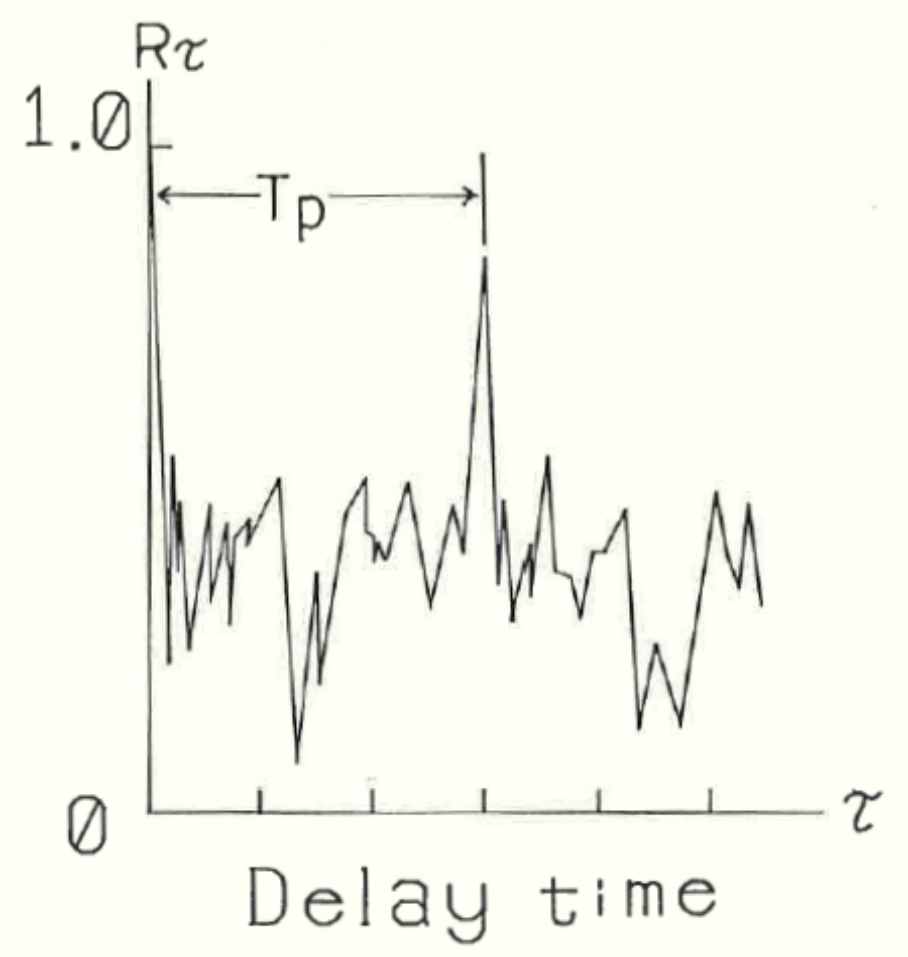


Figure 1.7: The autocorrelation function of LPC residual.

① The excitation energy is taken to be the average power R_0 of the residual waveform.

② Let $\tau = \tau_p$ be the nonzero lag that gives the maximum correlation. If the normalized autocorrelation R_{τ_p}/R_0 is at least 0.2–0.3, the interval is classified as voiced and τ_p is extracted as the pitch period. If it is below this threshold, the interval is classified as unvoiced and R_0 is taken as the unvoiced-speech power.

A practical implementation of this theory compensates for insufficient temporal resolution in calculating the autocorrelation function by applying the above algorithm to a residual signal obtained through LPC analysis after low-pass filtering the speech signal with a cutoff frequency of approximately 900 Hz[13]. Another proposed method uses the Average Magnitude Difference Function (AMDF), expressed as the sum of absolute differences in Eq.1.11, instead of the sum of products in Eq.1.10[19].

$$A_\tau = \sum_{n=0}^{N-1-\tau} |\epsilon_n - \epsilon_{n+\tau}| \tag{1.11}$$

In this method, the pitch period can be extracted because A_τ exhibits a sharp valley when the lag τ equals the pitch period. It also has the advantage of requiring less computation time because no multiplication is involved.

Because these two pitch-extraction methods can be applied not only to residual signals but also to ordinary original speech signals, methods have also been proposed in which the original speech signal is quantized to three levels (+1, 0, -1) to emphasize pitch periodicity and autocorrelation is calculated using logical decisions alone for multiplication among the three levels, thereby reducing the amount of computation[20].

However, methods that extract pitch periods by correlating residual or original speech signals are prone to extraction errors for low-pitched speakers because only a small number of pitch periods fall within the analysis interval, preventing a sufficiently large correlation value from being obtained. Lengthening the analysis interval to address this problem reduces temporal resolution because of changes in pitch frequency, while also making the interval excessively long for high-pitched speakers. The methods therefore involve conflicting requirements.

Another comparatively common conventional approach selects multiple pitch-peak candidates from the speech waveform and determines the pitch period by logical operations. Methods in which multiple such peak detectors operate in parallel and make the final decision by majority vote have also long been proposed[21]. Such methods, however, frequently extract false peaks, consequently requiring complex logical operations and producing relatively poor extraction rates.

For reference, a representative method intended for analysis-synthesis systems other than LPC systems is cepstrum-based pitch extraction through spectral processing[22]. Because the frequency characteristic $S(\omega)$ obtained by Fourier-transforming the speech signal can be approximated by a model in which the articulation $H(\omega)$ and excitation $G(\omega)$ are multiplied, as shown in Eq.1.1 in Section1.3, taking the logarithm of $S(\omega)$ transforms Eq.1.1 into Eq.1.12, separating the articulatory and excitation characteristics.

$$\log S(\omega) = \log H(\omega) + \log G(\omega) \tag{1.12}$$

Computing its inverse Fourier transform gives a time-domain signal called the cepstrum. In voiced intervals, a prominent peak appears at a position proportional to the pitch period, as shown in Fig.1.8, allowing the pitch period to be extracted. This method also has a problem, however: for high-pitched speakers, the number of pitch harmonics in the frequency domain decreases, making the cepstral peak indistinct and increasing pitch-extraction errors.

As described above, conventional pitch-extraction methods have difficulty maintaining pitch-extraction accuracy for both high- and low-pitched speakers. Furthermore, because the temporal resolution of extraction is determined by the analysis-interval length, adequate temporal resolution cannot be obtained. Conventional excitation information extraction also assumes that the excitation characteristic does not change within a prescribed interval (20–30 msec). In reality, even within such a short interval, the pitch period and intensity fluctuate finely at transitions between unvoiced and voiced intervals and at word-final or word-initial portions. Conventional pitch-extraction methods that do not account for these

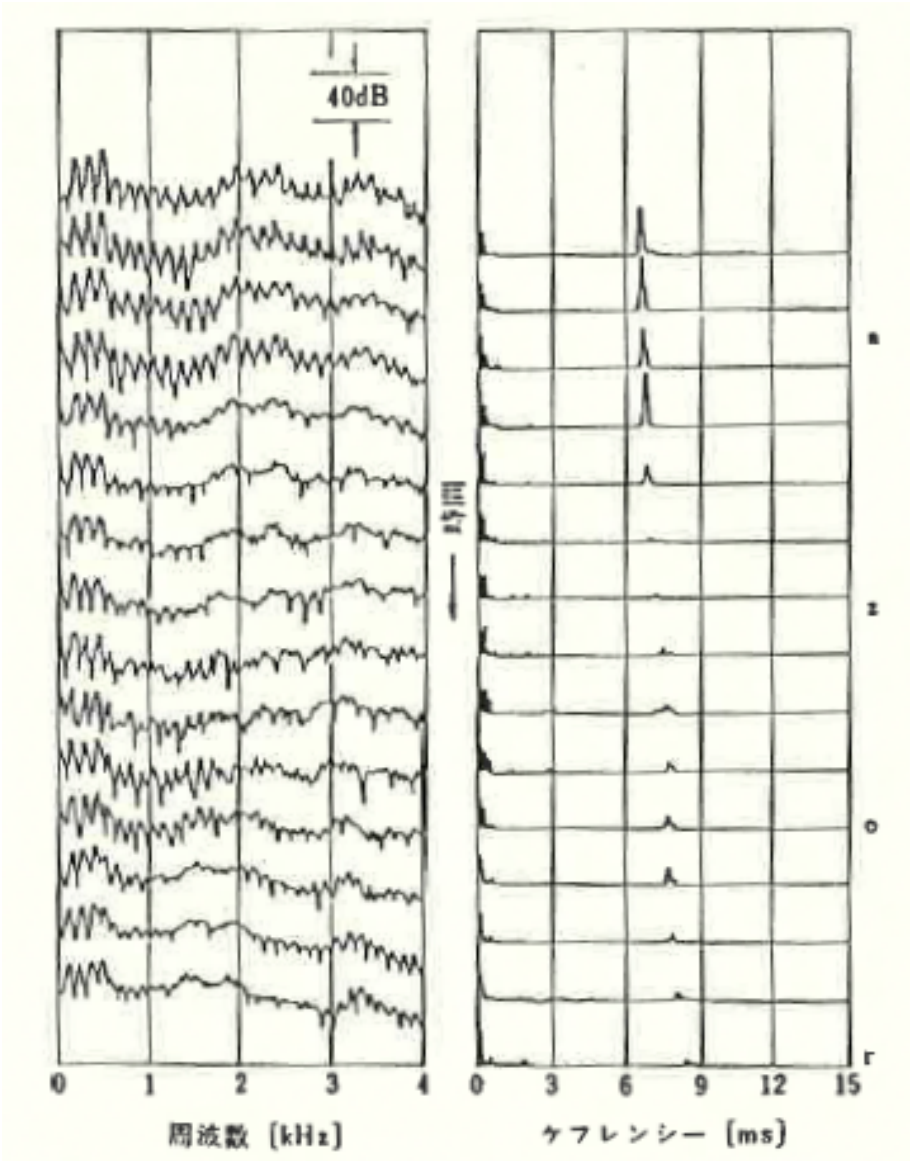


Figure 1.8: Cepstrum signals.

effects average out such fine variations, thereby degrading the quality of speech synthesized from the resulting excitation information.

If macroscopic excitation information can be accurately extracted from the residual, the residual can be modeled using the macroscopic excitation model shown in Fig.1.6; using this model as the excitation for LPC synthesis yields synthesized speech of practical quality. LPC analysis-synthesis schemes using macroscopic excitation models have in fact been improved in the form of PARCOR schemes[23][24], LSP schemes[25], and related methods, achieving highly efficient speech coding over low-bit-rate transmission bands of 1.2–4.8 Kbits/sec. For medium-quality speech transmission in applications such as mobile radio, however, low- and medium-bit-rate speech coding at 4.8–9.6 Kbits/sec is required. When LPC is used in this transmission band, its speech quality saturates near 4.8 Kbits/sec[25] and does not adequately achieve the quality required in practice; improving the quality therefore remains an important problem. The principal cause of this saturation is not poor modeling of speech articulation by LPC analysis-synthesis, but the modeling of the residual solely by the macroscopic excitation model shown in Fig.1.6. This creates the need to extract from the residual not only macroscopic information but also the microscopic excitation information discussed in Section1.4.

As discussed in Section1.4, the microscopic excitation information contained in the residual consists of zero characteristics representing the difference from the all-pole model. Frequency analysis of the residual is required to extract these characteristics. The most common method is the short-time discrete Fourier transform, defined as the Fourier transform of the residual signal multiplied by a 20–30-msec time window. When analyzing voiced residuals in particular, however, this method presents two conflicting problems. First, because the voiced residual is a quasi-periodic signal having a pitch period, lengthening the time window for narrowband analysis causes harmonic components to appear at integer multiples of the pitch frequency owing to the periodic pulse train in the time domain, as shown in Fig.1.2. These components obscure the zero characteristics; that is, the microscopic information is masked by the macroscopic information. Second, shortening the time window for wideband analysis to reduce the influence of the harmonic components blurs the zero characteristics because of convolution with the Fourier transform of the window. The short-time discrete Fourier transform is therefore not optimal for accurately obtaining zero characteristics other than the harmonic components in a voiced residual. One possible remedy is to smooth the harmonic components of the short-time residual spectrum by using cepstral analysis[26][27], which is effective for estimating speech-spectrum envelopes. This measure is not fully satisfactory, however, because the zero characteristics are already obscured by harmonic components in the original short-time spectrum.

Thus, although an all-zero model could be predicted for the microscopic excitation characteristics of the residual, no technique existed for extracting those characteristics from the residual. In other words, the condition of model analyzability described in Section1.2 was not satisfied. Consequently, no conventional coding scheme assumed a microscopic excitation model for the residual and encoded the residual on that basis.

Another speech synthesis scheme uses an impulse-train-equivalent excitation[28], employing as its voiced excitation the impulse-train-equivalent excitation pulses shown in Fig.1.9. This approach exploits the fine structure of the extremely short-time voiced-residual spectrum, in which high- and low-frequency emphasis intervals alternate within each pitch period. The scheme is promising because it uses excitation pulses resembling the microscopic characteristics of the residual, but no means has been presented for optimally determining pulses of the form shown in Fig.1.9 from residual-analysis results. It therefore does not fully satisfy the condition of model analyzability.

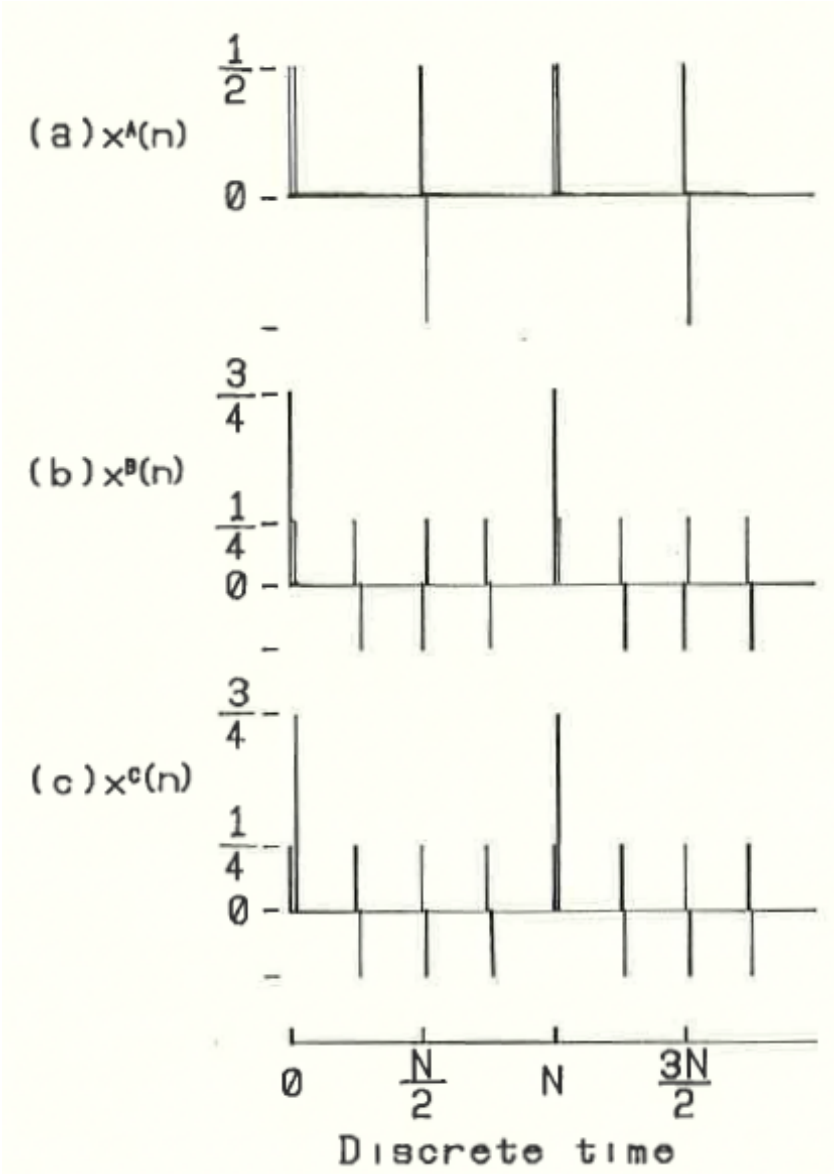


Figure 1.9: Impulse train equivalent excitation signals.

Conventionally, therefore, schemes for coding the residual as faithfully and efficiently as possible have assumed an appropriate residual model not based on a production model, or have used a hybrid of analysis-synthesis coding and waveform coding in which only the residual is waveform-coded and transmitted. These conventional schemes are not optimally efficient because they do not code the residual on the basis of a production model, and they differ from the direction pursued in the present study. Several representative schemes are nevertheless outlined below for reference.

① Multi-Pulse Excited LPC Coding (MPLPC)[29]

In this scheme, the driving excitation in LPC is approximated by multiple pulses rather than by the pulse-and-noise excitation model based on the speech production process described in Section1.4. Basically, as shown in Fig.1.10, iterative computation based on the Ab-s (Analysis by synthesis) technique determines the amplitudes and positions of a finite number of excitation pulses that minimize, in the frequency domain, the perceptually weighted error between synthesized and input speech. Information describing the optimum excitation pulses is transmitted, and LPC synthesis is performed.

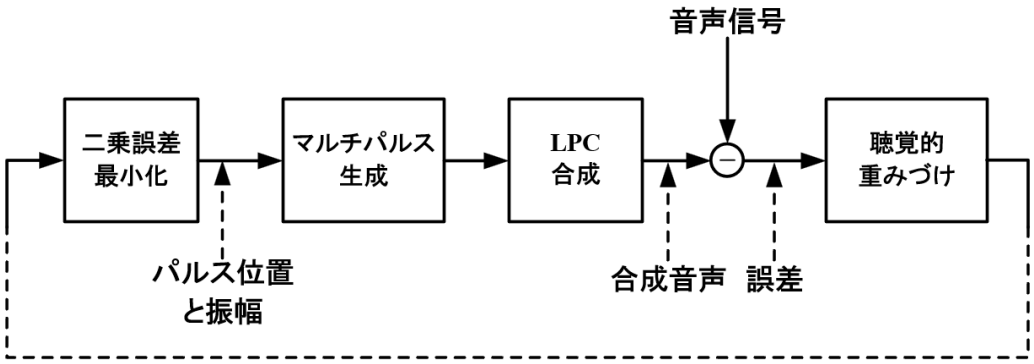


Figure 1.10: LPC coding by multi-pulse excitation.

② Code-Excited Linear Prediction (CELP)[30]

This scheme exploits the fact that the residual becomes white noise if prediction in LPC is perfect. First, the ordinary LPC residual is obtained with an LPC inverse filter. Long-term prediction is then applied to the residual to obtain a waveform from which the pulse-train component has been removed. This waveform is matched against noise sequences of a prescribed length in a prepared codebook, and the selected sequence is used to drive the LPC synthesis filter. As in MPLPC, and as shown in Fig.1.11, the optimum noise sequence is selected from the codebook so as to minimize the error between speech synthesized using the sequence and the input speech. Its code number is transmitted, and LPC synthesis is performed.

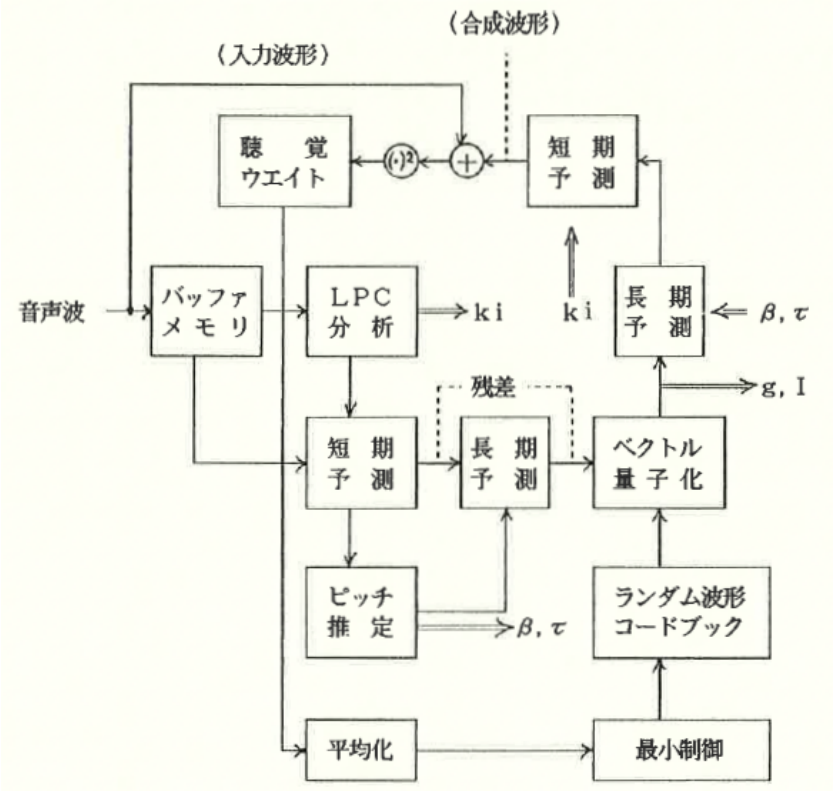


Figure 1.11: The code-excited LPC coding.

③ Residual-Excited Linear Prediction (RELP)[31]

As shown in Fig.1.12, low-frequency components of the LPC residual, for example those below 800 Hz, are downsampled, waveform-coded, and transmitted. At the receiver, the high-frequency compo-

nents are approximately regenerated from the low-frequency components by nonlinear processing such as rectification and clipping, and LPC synthesis is performed.

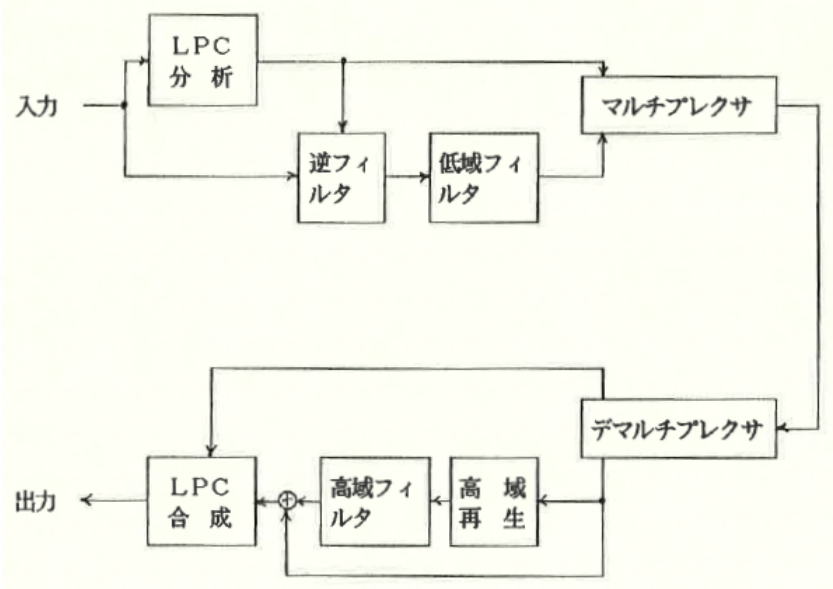


Figure 1.12: The residual-excited LPC coding.

Chapter 2

Excitation Information Extraction from the Amplitude-Envelope Characteristics of Residual Signals Using LPC Analysis

2.1 Overview

Chapter1 clarified the importance of extracting excitation information from residual signals in LPC and demonstrated the need for accurate extraction of macroscopic excitation information based on the pitch peaks appearing in the residual. This chapter focuses on the ability of LPC analysis to model formant peaks appropriately from a waveform with large variations in the frequency domain. A method is proposed in which a time-domain residual signal is regarded as a frequency-domain spectrum, pitch peaks are made to correspond to formant peaks, and the peaks are approximated by an all-pole model using LPC analysis. To make pitch extraction possible in nasal portions as well, a method is also proposed for identifying nasalized speech and applying the above processing to the original speech signal in such portions. Using the proposed system, the accuracy of pitch-peak extraction is evaluated, and actual speech is synthesized with the excitation information extracted by the method to confirm its effectiveness.

2.2 Excitation Information Extraction System

Figure2.1 shows the configuration of the excitation information extraction system proposed in Chapter2[33]. First, the input original speech signal S_1 is subjected to PARCOR analysis, an improved form of LPC analysis, and an inverse-characteristic filter is constructed using the resulting PARCOR coefficients[32]. This inverse-characteristic filter has the inverse transfer characteristic of the vocal-tract spectral characteristic of the input speech signal. Passing S_1 through the filter produces the residual signal S_2 corresponding to the original speech signal. Next, the mean-amplitude ratio between the residual and original speech signals and the ratio of the maximum value to the mean amplitude of the residual signal are calculated. The results determine whether excitation information is to be extracted from the original speech signal or from the residual signal. The selected signal is squared (if S_1 is selected, it is first half-wave rectified), and the resulting signal S_3 is regarded as a pseudo-frequency-domain amplitude spectrum. Signal S_3 is squared again and regarded as a power spectrum; it is then made even-symmetric, and the autocorrelation function r_i corresponding to the power spectrum is calculated by inverse FFT. LPC analysis is subsequently performed using the autocorrelation function r_i to obtain the LPC coefficients a_i. Envelope calculation by FFT is performed using these LPC coefficients a_i and the previously obtained r_i, producing an envelope signal S_4 corresponding to the squared signal S_3. The square root of S_4 is then calculated to obtain the envelope signal S_5 of the time-domain signal S_1 or S_2 regarded

as a pseudo-spectrum. Peaks are extracted from S_5 and error correction is performed, after which the pitch-peak positions P_j and their amplitudes A_{mj} are extracted as voiced excitation information V. If the interval is judged to be unvoiced during peak extraction, the mean amplitude A_{me} of residual signal S_2 is calculated and used as unvoiced excitation information UV. The foregoing describes the overall configuration of the excitation information extraction system proposed in this chapter. Its operation and theory are described in detail below.

2.3 Modeling of Excitation Information in the Proposed Method

In LPC-based speech analysis-synthesis, as described in Section1.4, transmitting the residual signal itself permits reproduction of the original speech, and the prediction errors in the residual signal are synchronized with pitch pulses in voiced speech. On the basis of these facts, directly modeling the positions and amplitudes of residual-signal peaks as shown in Fig. 2-2 should make it possible to follow fine variations in the excitation adequately. There remains, however, the question of whether the behavior of the excitation pitch pulses and that of the prediction errors in the residual signal are actually synchronized correctly. Consider the speech signal shown in Fig.2.3, which presents the speech waveform of /ma/ and its residual signal. Although /ma/ is voiced, almost no residual appears in the /m/ portion. This is because a phoneme such as the nasal /m/ is close to a sinusoidal signal containing almost no harmonics, so its spectral characteristic is predicted almost completely by LPC analysis as a vocal-tract characteristic. If excitation information is extracted from the residual of such speech, the portion is judged unvoiced because it has no periodicity, and speech synthesis is consequently performed with white noise. Although a signal having a similar spectrum can be synthesized because the spectral information for /m/ is retained in the LPC coefficients, the LPC coefficients contain no phase information. The phase component representing pitch is therefore lost, degrading the synthesized speech and causing its pitch phase to disagree with that of the following vowel /a/. If pitch is also extracted in nasal portions (sinusoidal speech) and synthesis is performed with a pitch-impulse train having the same amplitude as the residual signal in that portion, the phase representing pitch should be reproduced accurately.

2.4 Decision Logic for Residual and Original Speech Signals

As described in Section2.3, when the original speech signal is close to a sinusoid, as in a nasal sound such as /m/, excitation information cannot be extracted from the residual. Such speech portions must therefore be identified. Two parameters are used in the decision logic. Figure2.4 first shows the original and residual signals for an unvoiced portion, a nasal portion (/m/), and a vowel portion (/a/). The figure shows that, in the unvoiced and vowel portions, the relative amplitudes of the original and residual signals agree, whereas in the nasal portion the original speech signal has high power and the residual has low power. The ratio between the powers of the original and residual signals is therefore calculated; a high ratio indicates a nasal or sinusoidal-speech portion. Here, the mean-amplitude ratio TH_{amp} over a prescribed interval is used as one decision parameter. The second parameter is TH_{mav}, the ratio between the maximum and mean amplitudes of the residual signal over a prescribed interval. As Fig.2.4 shows, this value is small in unvoiced and nasal (sinusoidal-speech) portions. Using the two parameters TH_{amp} and TH_{mav}, an interval is judged to contain sinusoidal speech, and excitation information is extracted from the original speech signal, if the mean-amplitude ratio of the original to residual signal is at least a prescribed value and the ratio of the maximum to mean residual amplitude is no greater than a prescribed value. In all other cases, excitation information is extracted from the residual signal.

2.5 Amplitude-Envelope Characteristics of Residual and Original Speech Signals

If the decision described in Section2.4 determines that the speech interval to be analyzed is not sinusoidal, excitation information is extracted from the residual signal. As shown in Fig.2.2, a residual signal varies comparatively rapidly. The problem is therefore how to suppress the influence of other

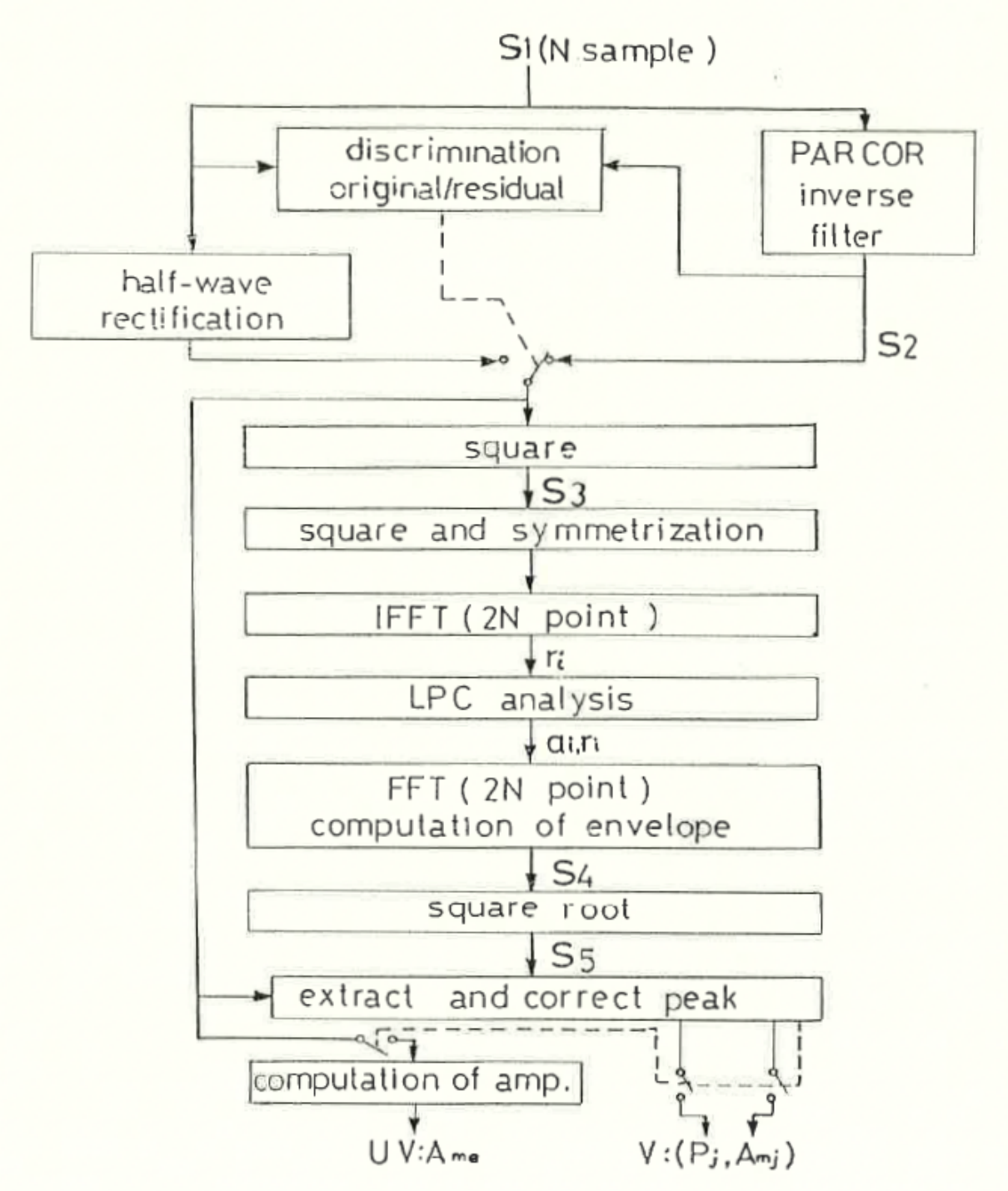


Figure 2.1: Block diagram of excitation extraction system.

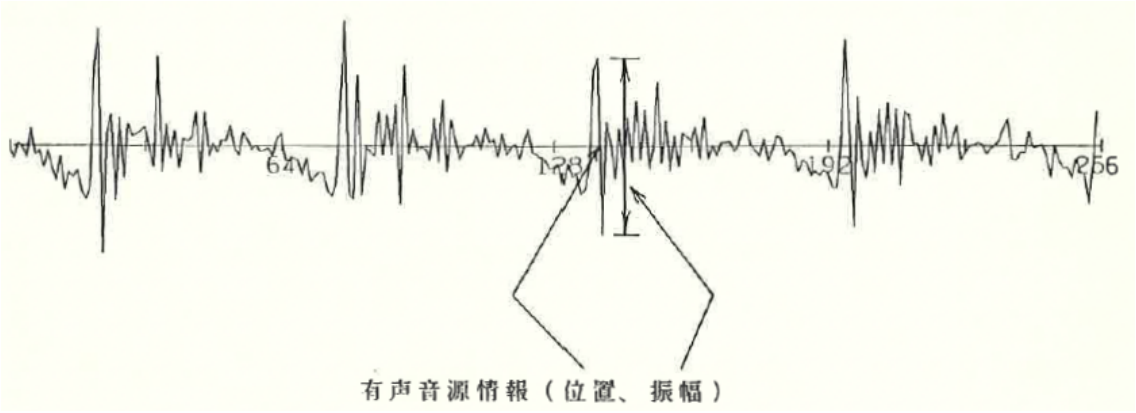


Figure 2.2: New modeling scheme of voiced-speech residual.

portions when extracting the positions and amplitudes of pitch peaks. For this purpose, the proposed method uses an algorithm that applies LPC analysis[32]. As described in Section1.3, LPC analysis optimally approximates the vocal-tract spectrum of a time-domain speech signal S_1 with an all-pole model. Using the autocorrelation function r_i and LPC coefficients a_i obtained from S_1, Eq.2.1 can be calculated to obtain E_i, the vocal-tract spectral amplitude-envelope characteristic $|H(\mathrm{e}^{j\omega})|$, as shown in Fig.2.5(a), thereby accurately approximating the formant peaks[34].

$$\begin{aligned} |H(\mathrm{e}^{j\omega})| &= \frac{G}{|1+\sum_{i=1}^{p} a_i \mathrm{e}^{-j\omega * i}|} \\ &= r_0 + \frac{\sum * i = 1^p a_i r_i}{|1+\sum_{i=1}^{p} a_i \mathrm{e}^{-j\omega_i}|} \end{aligned} \tag{2.1}$$

In other words, this analysis method removes the rapidly varying components of the frequency-amplitude spectral characteristic F_i of the original speech signal S_1 and obtains its envelope characteristic E_i. The time-domain residual signal S_2 is therefore made to correspond to the frequency-amplitude spectral characteristic F_i, as shown in Fig.2.5(b) (this is called a pseudo-frequency-amplitude spectrum), and an envelope characteristic S_5 corresponding to E_i is obtained. Pitch peaks can then be extracted by treating them as formant peaks. The algorithm is described below. Because LPC analysis requires an autocorrelation function, the autocorrelation function corresponding to the pseudo-frequency-amplitude spectrum is needed. As shown in Fig.2.1, the N-sample input residual S_2 is squared to emphasize changes in its peak portions, while the resulting signal S_3 is regarded as a pseudo-frequency-amplitude spectrum in which the first sample represents the dc component and the (N+1)th sample represents the highest frequency. Signal S_3 is squared again to form a power spectrum, made even-symmetric, and subjected to inverse FFT, yielding the autocorrelation function r_i corresponding to S_3[35]. Conventional LPC analysis is applied to r_i to obtain the LPC coefficients a_i according to Eq.1.3, described in Section 1.3. Equation2.1 is then calculated to obtain the envelope signal S_4 of the pseudo-frequency spectrum S_3. This calculation can be performed rapidly using the FFT[36]. Taking the square root of this envelope signal yields the amplitude-envelope characteristic S_5 of the time-domain residual signal S_2. As shown in Fig.2.6, the above processing removes the rapidly varying components from residual signal S_2 and produces an envelope signal S_5 that closely approximates the pitch peaks. The relationship between S_2 and S_5 corresponds closely to that between F_i and E_i in Fig. 2-5. The same processing can be applied when the decision in Section2.4 identifies a sinusoidal-speech interval and excitation information is extracted from the original speech signal. Because sinusoidal speech contains little information other than pitch periodicity, as shown by S_1' in Fig.2.7, it is first half-wave rectified so that the intervals between adjacent peaks equal the pitch interval. Applying exactly the same processing to this signal produces the amplitude-envelope signal of the original speech signal S_1'. Because the resulting amplitude-envelope signal is represented by the LPC all-pole model, it has sharp poles at the pitch peaks, as shown by S_5' in Fig.2.7, enabling those peaks to be extracted.

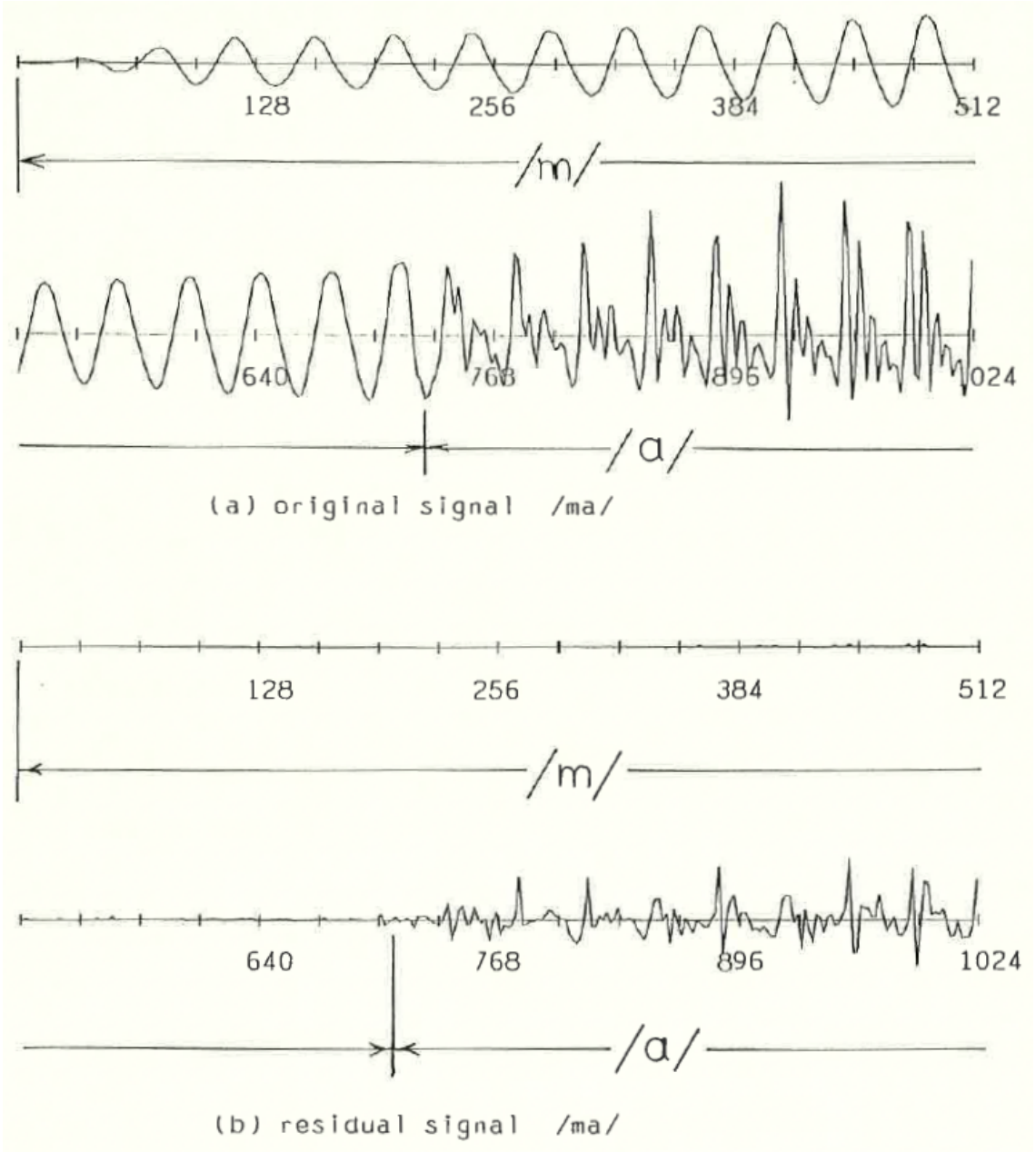


Figure 2.3: Original and residual signals /ma/.

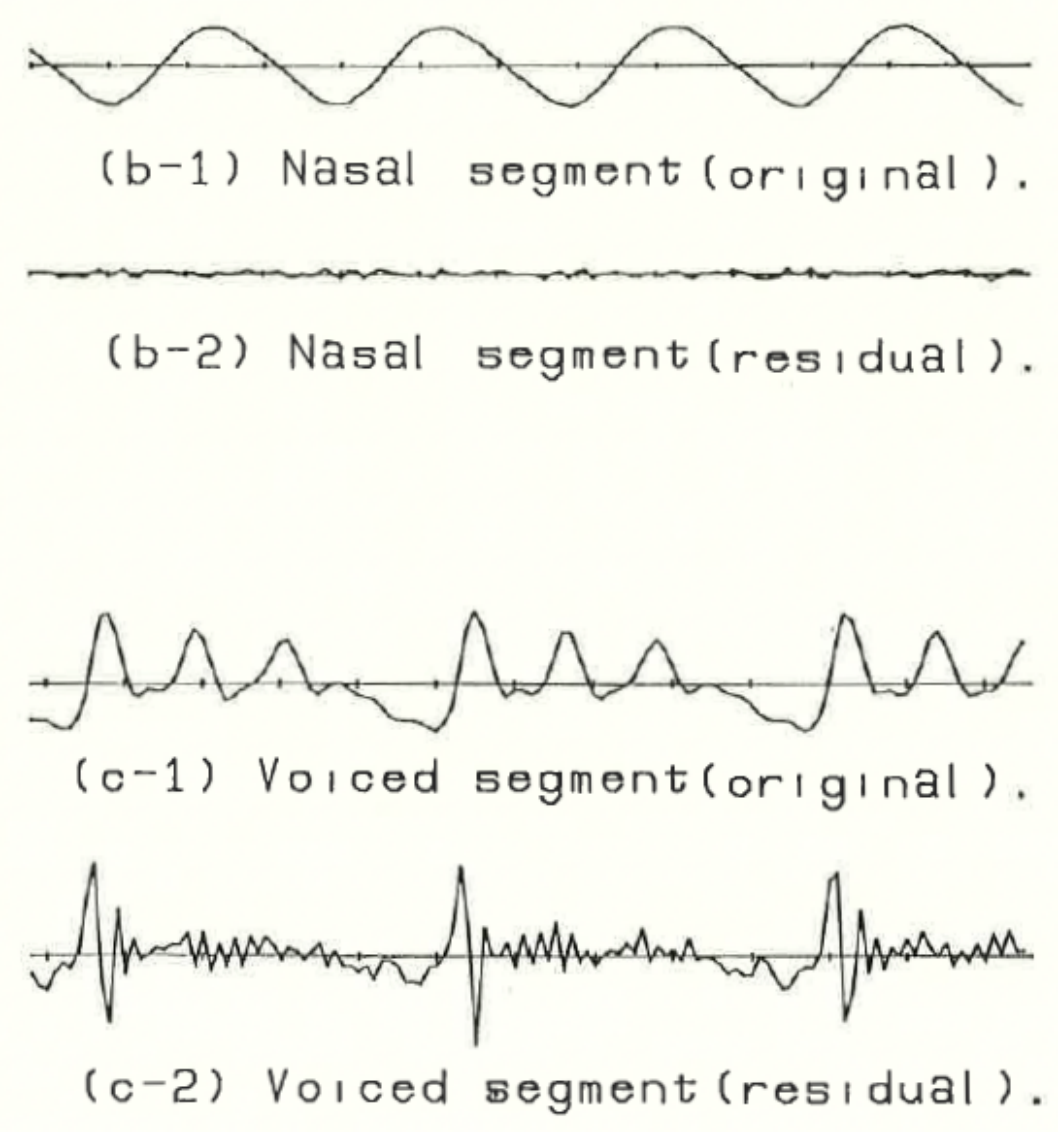
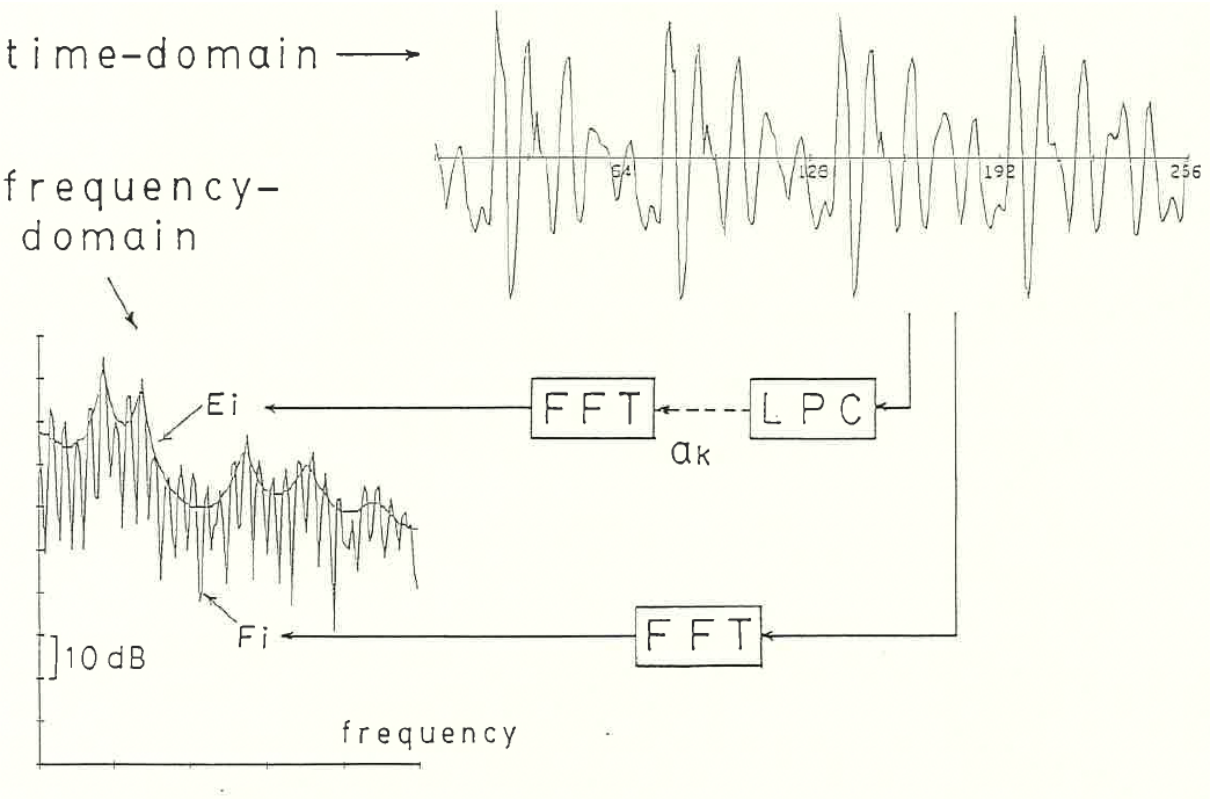
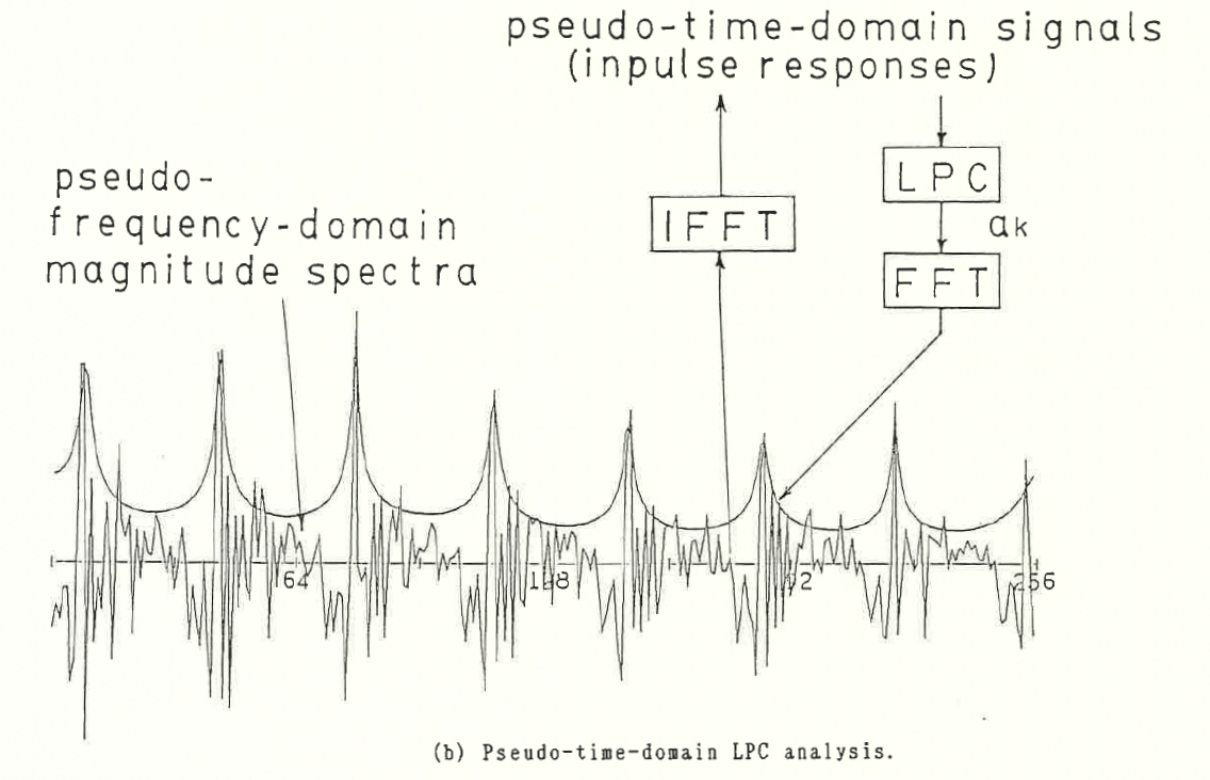


Figure 2.4: Original signals and residual signals of unvoiced, nasal, and voiced sounds.



(a) Time-domain LPC analysis.



(b) Pseudo-time-domain LPC analysis.

Figure 2.5: The principal of extracting time-domain amplitude envelope signals.

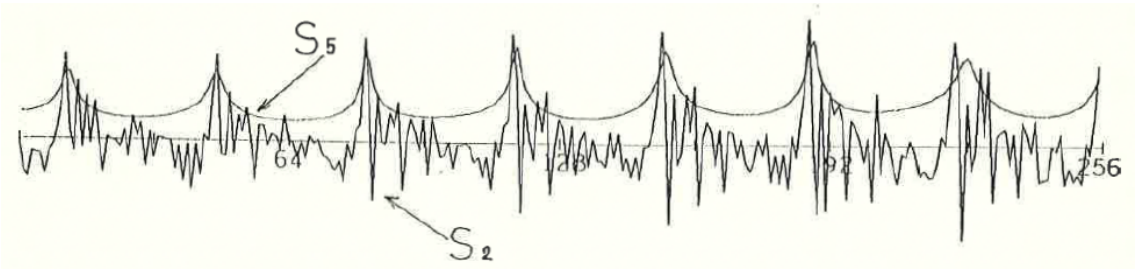


Figure 2.6: Residual signalws and the amplitude envelope signals.

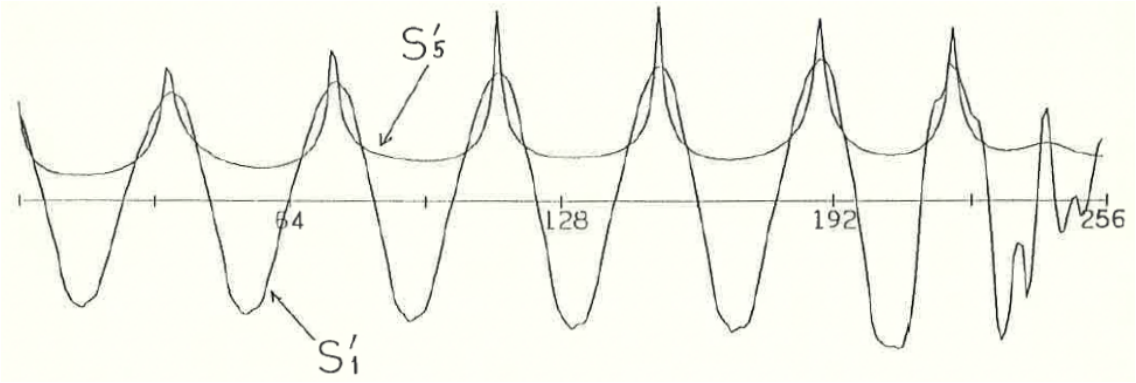


Figure 2.7: Original signalws and the amplitude envelope signals.

2.6 Peak Extraction and Correction, and Excitation Information Extraction

The amplitude-envelope signal S_5 (S_5') obtained in Section2.5 suppresses other influences to a minimum by modeling pitch peaks as poles. The sample positions and amplitudes at which S_5 (S_5') takes local maxima can therefore be extracted easily as pitch-peak candidates. It has already been stated that, in the LPC all-pole model, the linear system $H(z)$ representing the model has p poles given by the roots of the pth-order equation $A(z) = 0$ in Eq.1.5 of Section1.3. Because one formant in the frequency characteristic of speech approximated by an all-pole model can be represented by one pair of complex poles, an analysis order approximately twice the number of formant peaks is optimal[37]. In the system of Fig.2.1, described in Section2.2, the time-domain residual S_2, regarded as a pseudo-frequency-amplitude spectrum, contains at most approximately twelve pitch peaks when N is 256 samples for speech sampled at 10 KHz, because the highest pitch frequency is approximately 500 Hz. The LPC analysis order in this system is therefore 24. When the amplitude envelope is calculated and peaks are extracted in this way, the number of peaks within one analysis interval corresponds appropriately to the analysis order for high-pitched voiced speech, such as that of a female speaker, and the correct peaks are obtained. For low-pitched voiced speech, such as that of a male speaker, however, the number of peaks within one analysis interval is small and the analysis order is too high, so spurious peaks may appear in addition to the true peaks. The same tendency occurs at transitions between unvoiced and voiced speech. The correction algorithm for removing spurious peaks in such cases is described below with examples. It consists broadly of two routines: an amplitude-decision routine and an interval-decision routine. I. Amplitude-decision routine

(i) First, small-amplitude peaks that are clearly not pitch peaks are removed using a small amplitude threshold TH_{cut}.

(ii) The ratio between the amplitude A_p of the peak currently being examined and the amplitude A_v

of the immediately preceding valley is calculated from Eq.2.2.

$$r_p = \frac{A_p}{E_v} \tag{2.2}$$

If r_p is no greater than a prescribed value TH_r, the peak is removed as spurious. This decision routine operates when the interval from the preceding peak position is at least one analysis interval (256 samples), when only a few (approximately four) subsequent true peaks have been obtained, or when the interval-decision routine described below cannot make a decision.

For a comparatively high-power unvoiced-speech residual such as that in Fig.2.8(a), peaks are obtained from the envelope characteristic as indicated by the dashed arrows. Because the value of r_p calculated from Eq.2.2 is below the threshold for every peak, all are removed as spurious. The routine also accurately extracts peaks at the beginning of voiced speech, as shown in Fig.2.8(b). Ⅱ. Interval-decision routine After the first several pitch peaks at the beginning of voiced speech have been extracted by amplitude-decision routine I-(ii), the mean of several preceding pitch intervals (four to eight intervals) is calculated for the peak currently under examination. If the interval between the current and preceding peaks is within a prescribed error of this mean (within 10–20% of the mean), the current peak is judged to be true. Otherwise, if the interval between the peak following the current one and the preceding true peak is within the prescribed error of the mean, the current peak is judged spurious and removed. If neither condition applies, a decision is made by amplitude-decision routine I-(ii). If the interval from the peak thus judged true is at least twice the preceding mean, speech is judged to have been interrupted in that portion, and the following peak is extracted using only amplitude-decision routine I-(ii). That peak interval is not included in calculation of the next mean. The interval-decision routine is based on the fact that, even when a spurious peak occurs between true peaks, the order constraint imposed when approximating pitch peaks by an all-pole model permits at most one such peak. When there are many true peaks, as in the female residual signal in Fig.2.8(c), this routine classifies all of them as true. When there are few true peaks (solid arrows), as in the male residual in Fig.2.8(d), consider, for example, the current peak PK_w. Comparing the interval T_{now} between PK_w and the immediately preceding true peak PK_t with the interval T_{next} between the following peak PK_n and PK_t, T_{next} is closer to the mean T_{av} of the several immediately preceding pitch intervals. The current peak PK_w is therefore removed as spurious. Even if the interval-decision routine cannot decide, the peak can be removed because r_p, calculated by amplitude-decision routine I-(ii), is below the threshold. When peak PK_{next} is examined and the interval-decision routine cannot decide, it is classified as true because r_p, calculated by the subsequent amplitude-decision routine I-(ii), exceeds the threshold, as shown in the figure. Excitation information is extracted from the decisions of these routines as follows.

(I) If the routines extract no peak within one analysis interval (256 samples), the interval is regarded as unvoiced. The mean amplitude A_{me} of the residual in the interval is calculated and output as unvoiced excitation information UV (Fig.2.1).

(II) If the decision described in Section2.4 causes excitation information to be extracted from the residual and the routines identify true peaks, their positions $P_j (1 \leqq P_j \leqq 256)$ and amplitudes A_{m_j} are output as voiced excitation information V (Fig.2.1).

(III) If the decision described in Section2.4 causes excitation information to be extracted from the original speech signal and the routines identify true peaks, their positions P_j and the mean amplitude of the corresponding residual signal near each position P_j (approximately 10–20 samples) are output as voiced excitation information V, with the latter used as the extracted peak amplitude A_{m_j} (Fig.2.1).

In this way, excitation information based on the model described in Section2.3 is extracted. Experiments on the proposed system shown in Fig.2.1 and their results are discussed below.

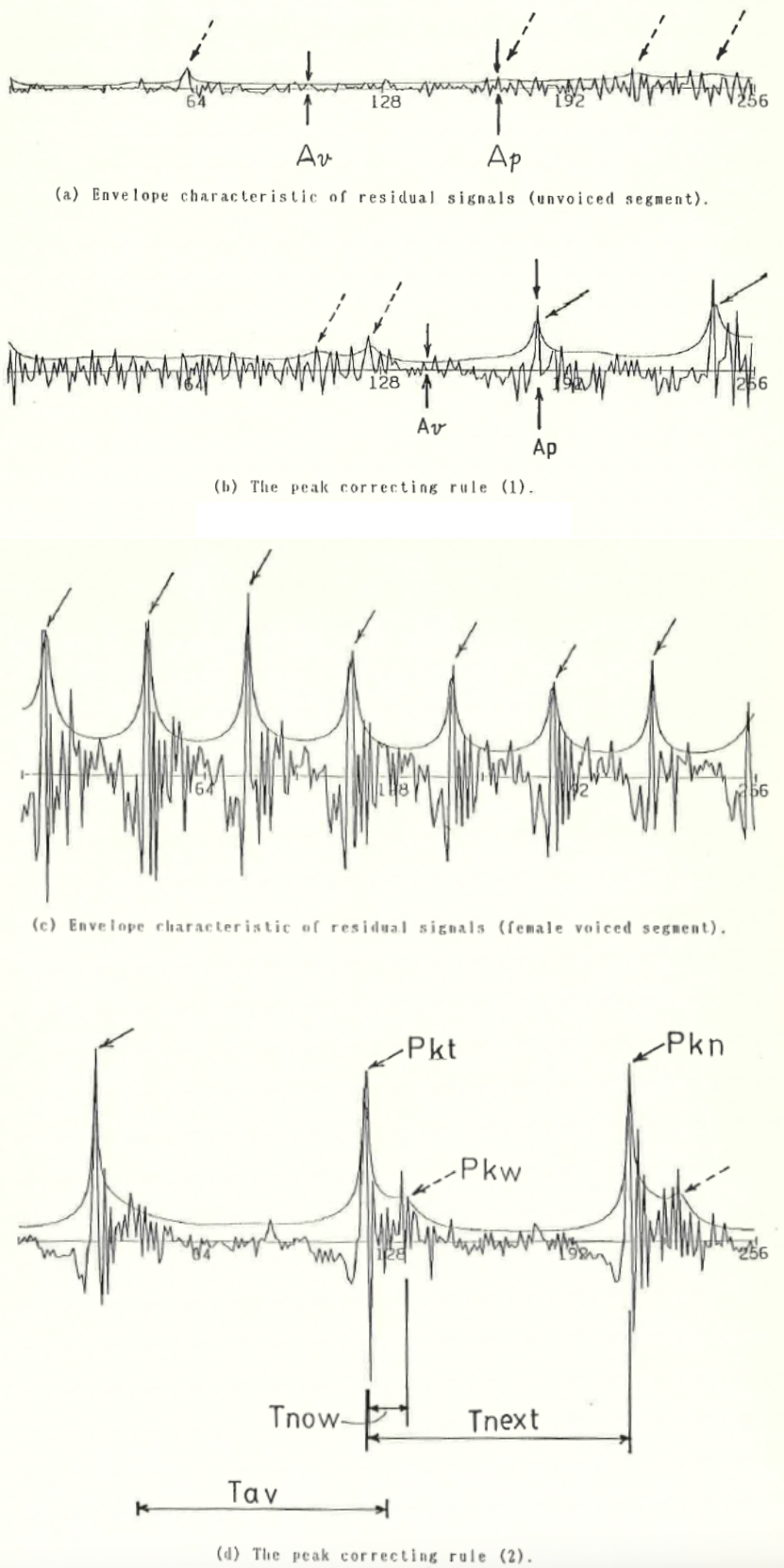


Figure 2.8: Peak extraction and peak correction.

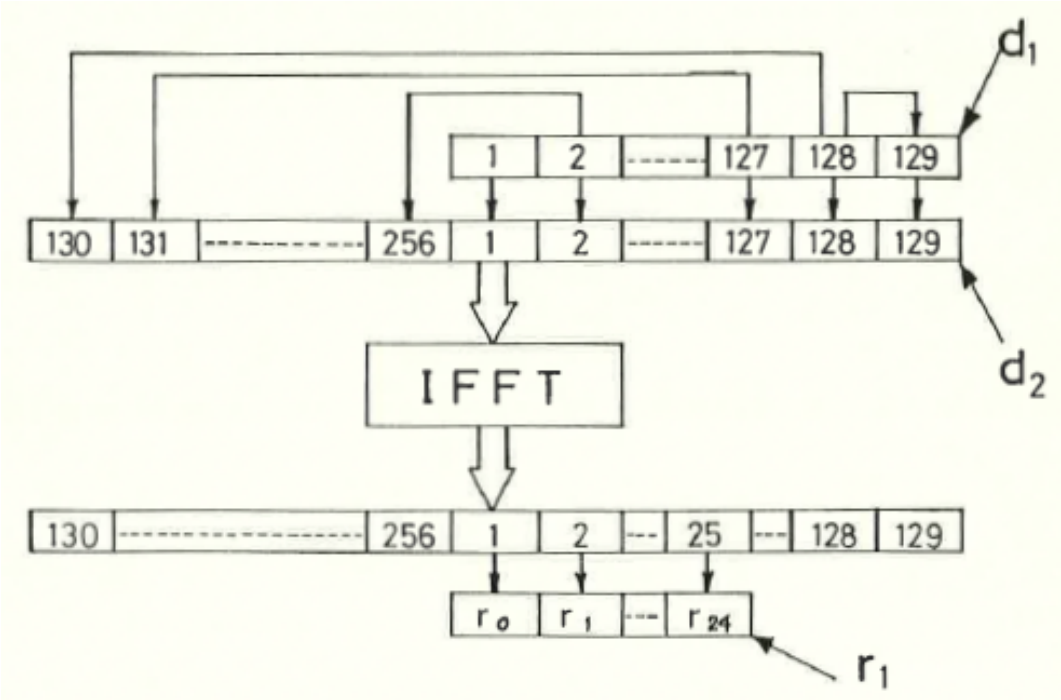


Figure 2.9: Data symmetrization.

2.7 Experimental Results

2.7.1 Experimental Conditions for Excitation Information Extraction

A PDP-11 equipped with 12-bit A/D and D/A converters was used as the main experimental system. Speech data from an open-reel tape deck were passed through a 5-KHz low-pass filter and A/D-converted at a sampling rate of 10 KHz. An IF800 microcomputer was used as the graphic display for the PDP-11, together with an XY plotter. A 5-KHz low-pass filter, amplifier, and loudspeaker were connected to the D/A output. The speech used in the experiments was sampled at 10 KHz and quantized with 12-bit precision. All processing was performed on speech-data files containing 20 Kwords (20,480 samples), corresponding to approximately two seconds. Excitation information was first extracted from actual speech using the proposed system (Fig.2.1). In this system, the inverse-characteristic filter for PARCOR analysis performs residual extraction using a PARCOR lattice filter. One PARCOR analysis frame contains 256 samples, and the analysis order is ten. The PARCOR coefficients are obtained by the Durbin-Levinson-Itakura method[14]. For the residual/original-signal decision logic described in Section2.4 and the amplitude-envelope algorithm described in Section2.5, one frame contains 128 samples. Six samples overlap at each end, and the data are updated by 116 samples to eliminate end effects. Figure2.9 shows the specific algorithm for the squaring (power-spectrum conversion), even symmetrization, and inverse FFT portions of Fig.2.1. Here, d_1 is the squared residual after conversion to a power spectrum, and d_2 is the real input to the inverse FFT; all imaginary input values are zero. The inverse FFT produces the autocorrelation function r_i as its real output. The Durbin-Levinson-Itakura method is also used for the LPC analysis in Fig.2.1. As stated in Section2.6, an order approximately twice the number of peaks is optimal. Assuming that at most approximately six peaks (corresponding to a 500-Hz pitch) appear in one 128-sample frame, the order is set to twelve. The autocorrelation functions $r_0 - r_{12}$ in Fig.2.9 are therefore used. Figure2.10 shows the specific algorithm for the FFT and envelope calculation in Fig.2.1. Here, d_3 represents the LPC coefficients $a_0 - a_{12}$ obtained by LPC analysis, and d_4 is the real input to the FFT; all imaginary input values are zero. Of the 256 real outputs X_i and imaginary outputs Y_i produced by the FFT, the first 128 are processed in operation (a) to calculate their sum of squares. Operation (b) is then applied to the resulting signal $|A_i|^2$, thereby calculating Eq.2.1 and obtaining the 128-sample envelope signal S_4 of the squared residual S_3 in Fig.2.1. Taking the square root of S_4, as

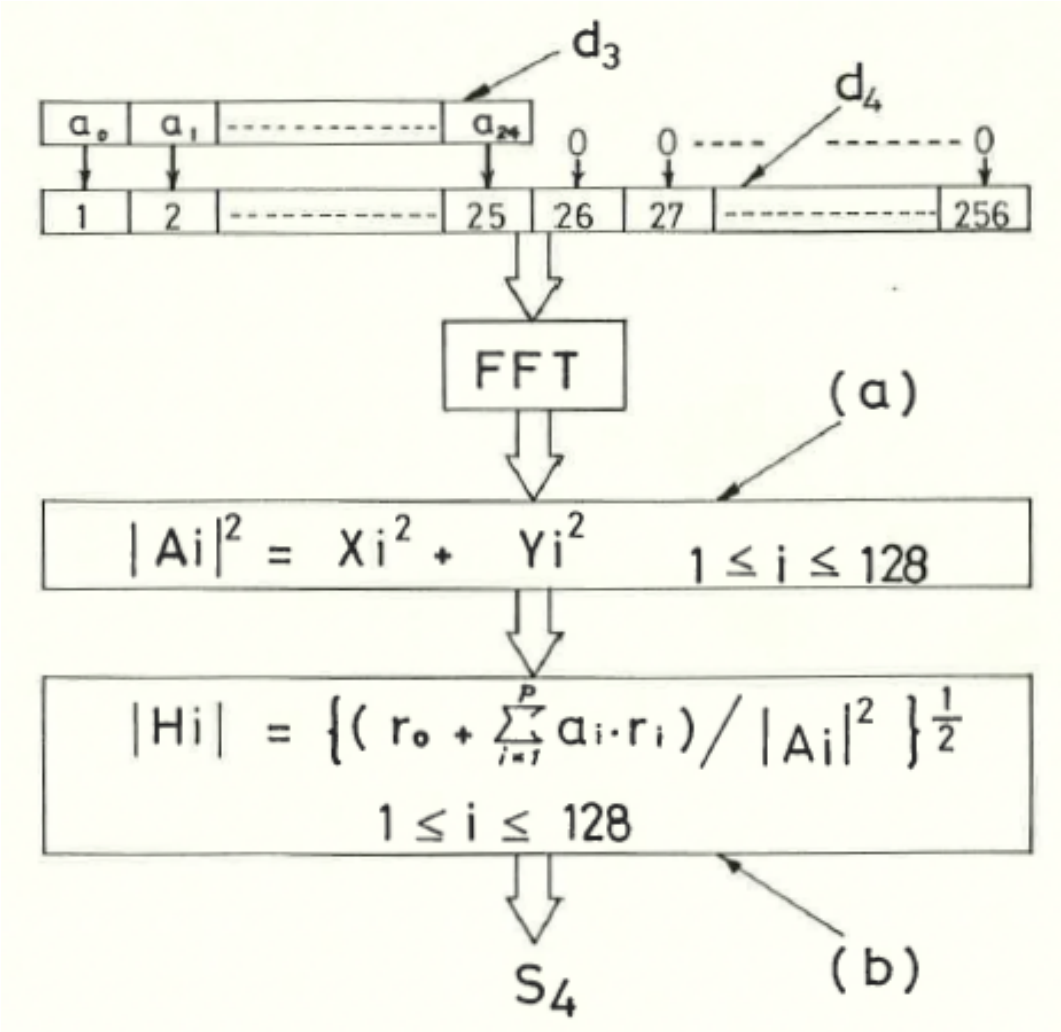


Figure 2.10: Envelope computation.

shown in Fig.2.1, yields the envelope signal S_5 of the original residual. For the central 116 samples of S_5, peaks can be extracted from the positions and amplitudes of local maxima. Because the numerator in operation (b) of Fig.2.10 is constant within one analysis interval, actual peak extraction first determines peak positions from the positions of local minima in the output $|A_i|^2$ of operation (a). Operation (b) and the square-root calculation (Fig.2.1) are performed only for samples at those peak positions to calculate their amplitudes. This eliminates the need to perform operation (b) and the square-root calculation for the entire envelope, reducing the amount of computation. The subsequent peak-correction process (Section2.6) uses a 256-sample analysis (correction) frame, unlike the frame length used for the amplitude-envelope calculation. For voiced speech, peak positions are extracted as positions among the 256 samples; for unvoiced speech, the mean amplitude of the 256 samples is extracted. The envelope-analysis frame was not set to 256 samples because, as discussed below, processing 128 samples twice requires fewer FFT, inverse-FFT, and LPC-analysis computations than processing 256 samples once. If the frame is made too short, however, analysis (peak-extraction) accuracy deteriorates. Preliminary experiments therefore led to selection of 128 samples as the shortest frame length that did not reduce accuracy. The experimental results obtained under these conditions are described below. Ten speech files, five for male and five for female speakers, were used.

No.1(DATA07.DAT)	“Is June in autumn?”	(Male)
No.2(DATA07.DAT)	“ ”	(Female)
No.3(DATA18.DAT)	“In which month is Christmas?”	(Male)
No.4(DATA18.DAT)	“ ”	(Female)
No.5(DATA28.DAT)	“In which prefecture is Yokohama?”	(Male)
No.6(DATA28.DAT)	“ ”	(Female)
No.7(XVOI01.DAT)	“It will probably rain tomorrow.”	(Male)
No.8(XVOI03.DAT)	“ ”	(Female)
No.9(XVOI02.DAT)	“Can one ski in a tropical country?”	(Male)
No.10(XVOI04.DAT)	“ ”	(Female)

Each file comprises blocks 0–79, with 256 samples per block (blk).

2.7.2 Decision Between Residual and Original Speech Signals

A preliminary experiment was first performed to determine the values used by the decision logic described in Section2.4. The optimum values of the mean-amplitude ratio TH_{amp} between the original and residual signals and the maximum-to-mean amplitude ratio TH_{mav} of the residual were found to be

$TH_{amp} = 7.0$

$TH_{mav} = 6.0$

The ten speech files were then classified using these values and 116-sample analysis frames. Examples are shown in Fig.2.11(a), (b), and (c). Each region delimited by vertical lines is one analysis frame (116 samples), and frames marked $\times$ were judged to contain sinusoidal speech. In the frames marked $\times$ in (a) and (b), the original signal is nearly sinusoidal, so the periodicity of its residual is indistinct and its power is low. Switching to the original signal was therefore decided very successfully. In the example in (c), some periodicity is visible in the residual even in the frame marked $\times$. This degree of periodicity nevertheless permits accurate excitation extraction from the original signal and has almost no effect on the extraction accuracy discussed below. The algorithm was thus found to identify accurately sinusoidal-speech intervals in which almost no periodicity appears in the residual. Although 7.0 and 6.0 were obtained as the optimum thresholds TH_{amp} and TH_{mav}, respectively, these values imply that the mean-amplitude ratio between the original and residual signals of an ordinary voiced vowel is no greater than seven and that residual periodicity is present if the ratio between a pitch peak and the mean residual amplitude is at least six. The results in Fig.2.11 confirm these assumptions. Because the method is very simple, it may classify a signal as nonperiodic even when some periodicity is present, as in (c). This decision need not be extremely strict: even if excitation information is mistakenly extracted from the original signal when the residual is periodic, the subsequent envelope-based extraction algorithm is intrinsically effective for both signals, so excitation information can still be extracted. The decision remains useful because extraction from the residual is more accurate. The experiments showed the method to be sufficiently simple and effective.

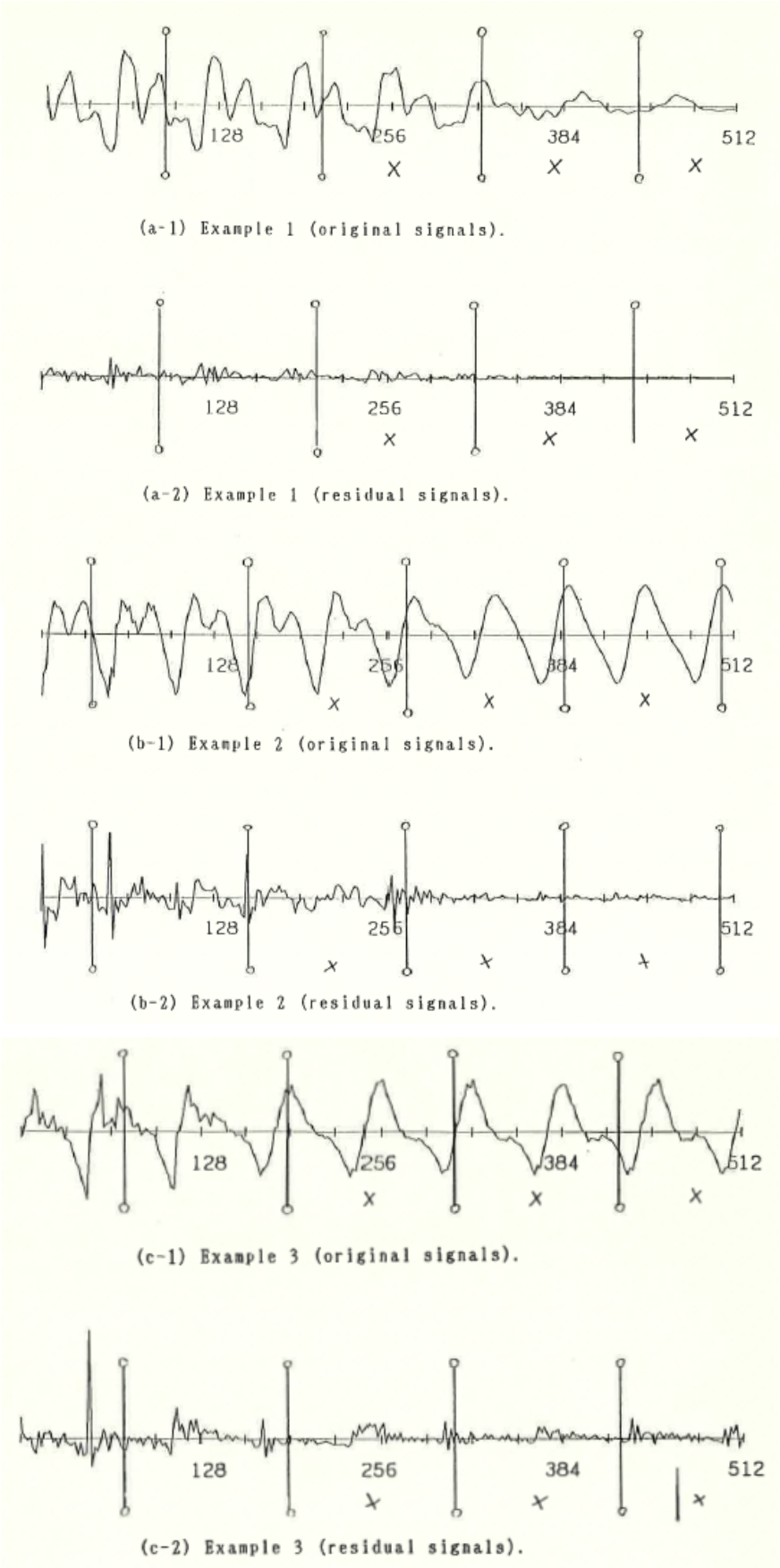


Figure 2.11: Decision of residual signals or original signals.

2.7.3 Amplitude-Envelope Characteristics

Following the above decision, the amplitude-envelope characteristics calculated for the residual or original speech signal using the algorithm in Fig.2.1, described in Section2.5, and the procedures in Figs.2.9 and 2.10 are presented. Figure2.8(c), used in the explanation in Section2.6, is an example of the amplitude-envelope characteristic of the residual in speech file No.9 (XVOI02.DAT). It models the position and amplitude of each pitch peak very accurately. Figure2.8(b) is an example for speech file No.3 (DATA18.DAT). It corresponds to /ku/ at the beginning of "kurisumasu wa..." and is a transition from unvoiced to voiced speech; here too, the positions and amplitudes of the pitch peaks are closely approximated. Figure2.8(d) shows a residual from speech file No.5 (DATA28.DAT), in which spurious peaks occur in addition to pitch peaks. Figure2.8(a) shows another residual portion from file No.3: the initial unvoiced portion of /ku/ in "kurisumasu wa...", where meaningless spurious peaks appear. Figure2.12(a) is an example of the amplitude-envelope characteristic of the original signal in speech file No.9. It too corresponds accurately to each pitch-peak position, demonstrating that the envelope algorithm can be applied directly to a sinusoidal original speech signal. Although the peak amplitudes are somewhat displaced, this poses no problem because, for an original signal, the mean amplitude of the residual near the peak position is used. Figure2.12(b) shows the original signal from another portion of file No.5; closely spaced spurious peaks appear. Figure2.12(c) shows the original signal from another portion of file No.9. This is the beginning of /ma/ in "...masu ka"; despite the presence of a spurious peak, the pitch peaks are closely approximated. Thus, for both residual and original signals, the calculated amplitude-envelope characteristic minimizes the occurrence of spurious peaks while approximating pitch peaks very accurately. The plots in Fig.2.1 were calculated from negative-amplitude portions to match the peak characteristics of the residual in the following vowel. Approximating the amplitude envelope of a time-domain speech signal by an all-pole model is the principal feature of the proposed method. As Fig.2.6 shows, the influence of portions other than pitch pulses is removed very effectively, while the pitch-pulse portions are approximated very accurately. This is because an all-pole amplitude-envelope model can effectively approximate the sharp pitch-pulse peaks. An important issue is selection of the model order. One interval of the time-domain residual is regarded as a pseudo-spectrum extending from dc to the highest frequency. Because the optimum all-pole order is twice the number of spectral peaks, as discussed in Section2.6, the optimum order is twice the number of pitch peaks in one interval. The number of peaks varies, however, so the order was selected to approximate the maximum number. Peaks can therefore appear outside the pitch-peak portions, as shown in Fig.2.8(d). Experiments showed that at most approximately one spurious peak appears between true peaks (as can also be inferred from the order constraint), allowing the subsequent correction algorithm to remain very simple. If the order is too low, other components influence the result and the peaks are not approximated accurately; a higher order was therefore deliberately used while allowing spurious peaks. Such peaks frequently approximate formant effects and consequently tend to occur immediately after true peaks, a property that also facilitates construction of the correction algorithm.

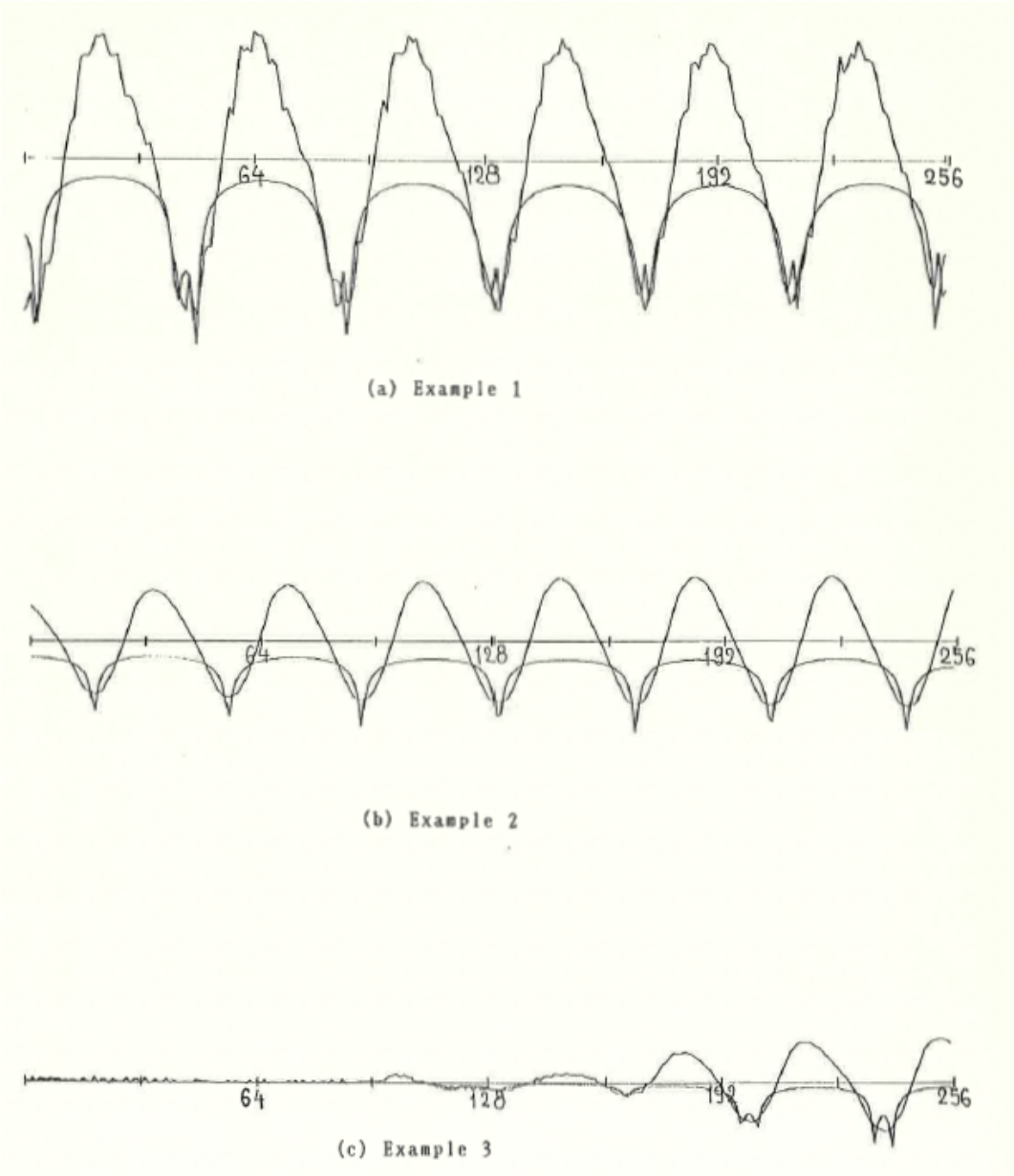


Figure 2.12: Amplitude envelope characteristic of original signals.

The same concept was successfully applied to original speech signals. As shown in Fig.2.12, the pole peaks appear appropriately at waveform peaks and may be regarded as pitch peaks. Furthermore, the preceding decision algorithm restricts the original signals to be analyzed to signals close to sinusoids. They therefore contain little influence other than pitch, enabling successful peak extraction from their amplitude envelopes. It is important in this case to half-wave rectify the original signal first.

2.7.4 Peak Correction and Peak-Extraction Accuracy

Peak extraction and correction were performed from the calculated amplitude-envelope characteristics according to the algorithm in Section2.6, and excitation information was extracted. Preliminary experiments determined the correction thresholds as follows. For amplitude-decision routine I-(i), $TH_{cut} = 60$ for residual signals and $TH_{cut} = 150$ for original signals. For amplitude-decision routine I-(ii), $TH_r = 2.0$ for residual signals and $TH_r = 4.0$ for original signals. The permissible error relative to the mean in the interval-decision routine was $\pm 15\%$. Using these parameters, excitation information was extracted from

Table 2.1: Accuracy of peak extraction.

File number	Extraction accuracy (%)
1	98.6
2	97.2
3	97.5
4	97.1
5	99.2
6	94.7
7	98.3
8	96.5
9	98.1
10	100.0

Table 2.2: Cause of peak extraction errors.

(Unit: number of occurrences)

Cause	Omissions		False extractions	
Signal analyzed	Original waveform	Residual waveform	Original waveform	Residual waveform
Number	29	24	0	5

the ten speech files with the system in Fig.2.1. The results were evaluated against extraction performed by visually inspecting the original and residual waveforms on a graphic display. Table2.1 gives the results. Here, extraction accuracy is the percentage of extracted peak positions agreeing with visually determined positions to within ± 2 samples. Every speech file exhibited very high accuracy, although file No.6 was somewhat less accurate. File No.10 achieved 100% accuracy. Table2.2 summarizes the causes of errors. Most errors were omissions in which no peak appeared in the amplitude envelope at a pitch-peak position. In some word-final cases, no peak followed the current peak; when the current peak was a formant-induced spurious peak, it could not be decided by the interval routine and was then misclassified and extracted by the amplitude routine. Most omissions occurred in portions resembling sinusoidal speech. As can be seen from Fig.2.8 and related results, the extracted peak amplitudes differed by only approximately 10% from the actual residual peak amplitudes in every case. Peak correction makes effective use of the properties of the amplitude envelope. The interval-decision routine exploits the continuous variation of pitch intervals (periods). The experiments assume that adjacent pitch intervals cannot differ by more than 15%, and the results confirmed this assumption. This alone cannot identify correct peaks at speech onset or after an erroneous pitch interval has been extracted, so the amplitude-decision routine is used in such cases. As Fig.2.8 shows, this routine exploits the envelope property that the amplitude difference between a true peak and the valley immediately following it is large. Because spurious peaks occur close to true peaks, as noted above, their amplitude difference from the preceding valley is small, allowing effective discrimination. The same reasoning applies to original speech signals such as those in Fig.2.12. The values of TH_{cut} in amplitude-decision routine I-(i) were set to 60 for residual signals and 150 for original signals because portions below these amplitudes were regarded as ineffective speech. These thresholds assume that the input speech has been adjusted by an AGC (automatic gain controller) or similar device so that its maximum approaches 2048, the maximum for 12-bit quantization. As Table2.1 shows, high pitch-extraction accuracy was obtained. Because this accuracy is the proportion of correct results among all peaks, it is extremely high and demonstrates the effectiveness of modeling pitch-peak portions with an all-pole model. Table2.2 shows that errors occur most often when the original speech is sinusoidal. This is thought to occur when a residual having little periodicity is classified as periodic by the decision in Section2.7.2, causing extraction to be performed on the residual and preventing a peak from appearing in the envelope. Even when extraction is performed on the original signal, a peak sometimes fails to appear because formant peaks in addition to pitch peaks make the analysis order relatively low for the total number of peaks, preventing accurate modeling. When extraction is performed from a residual, the amplitude-decision routine may operate

Table 2.3: Relations between frame length,analysis order,and numbers of multiplications.

Analysis-frame length	LPC order	Number of multiplications
256	24	7560
128	12	6504
64	6	5592
32	3	4752

at a word-final peak and classify a spurious peak as true. Such errors were nevertheless very infrequent, and the system can be considered to have excellent overall performance.

2.7.5 Evaluation of Synthesized-Speech Quality

Speech synthesis was next performed using the excitation information extraction system. Its conditions were identical to those in Section2.7.1, and the system output voiced excitation information $V(P_j, A_{mj})$ or unvoiced excitation information $UV(A_{me})$. The synthesis system was a tenth-order PARCOR synthesis system with an analysis-frame length of 256 samples; the PARCOR coefficients were linearly interpolated at intervals of 16 samples. For voiced information, a residual was synthesized for each 256 samples by placing an impulse of amplitude A_{m_j} at each pitch position P_j and was used as the PARCOR synthesis-filter excitation. For unvoiced information, 256 samples of white noise with amplitude A_{me} were used. Paired-comparison listening tests were conducted against a method that extracted a mean pitch period for each 256-sample frame by visual inspection. Files No.8 and No.9 were used. Six listeners heard each pair three times in random order. The proportion preferring the proposed method was 61.1%, confirming its effectiveness. Listener comments were as follows.

(i) The comparison method sounds more muffled; the proposed method sounds clearer.

(ii) A small amount of clicking noise, apparently caused by peak-extraction errors, can be heard with the proposed method.

(iii) The comparison method produces synthesized speech with less contrast and definition.

These results demonstrate improved synthesized-speech quality. The greater clarity of the proposed method is attributed to its accurate modeling of pitch-peak positions and amplitudes. The residual pitch-impulse train varies subtly even within a 256-sample frame, leaving frequency characteristics not fully removed by the inverse filter. Conventional methods cannot model these variations. Accurate extraction in nasal and other difficult portions also improves fidelity to the original speech. Listeners also found the /ki/ and /ka/ portions of "sukii wa nangoku de dekimasu ka" clear, suggesting successful modeling at unvoiced-to-voiced transitions. The system was therefore highly effective for speech synthesis, although clicking caused by extraction errors could not yet be eliminated fully automatically.

2.7.6 Analysis-Frame Length and Number of Multiplications

If the analysis-frame length in Fig.2.1 is N samples and the LPC order is P, the required multiplications are[36][38]:

Squaring (twice)	:N $\times 2$
Inverse FFT	:$\frac{1}{2} \times \frac{2N}{4} \times \log_2 \frac{2N}{4} \times 4$
(Even-symmetric real data permit an FFT one-quarter this size.)	
LPC analysis	: $P^2 + 3P$
FFT	: $\frac{1}{2} \times \frac{2N}{2} \times \log_2 \frac{2N}{2} \times 4$
Sum of squares (Fig.2.10(a))	:2N

The relationships required to calculate an envelope for peak extraction from 256 samples are shown in Table2.3. Shorter frames reduce multiplication counts but increase frames without pitch peaks and thus spurious extraction. Preliminary tests found nearly equal accuracy for 128- and 256-sample frames, so 128 samples is preferable. Total processing is comparable to or lower than that of conventional methods, making implementation feasible as FFT and LPC hardware becomes available.

2.8 Summary

Chapter2 proposed a new model of macroscopic excitation information. A method was proposed in which a time-domain signal is regarded as a frequency spectrum and its amplitude envelope is obtained as an LPC-based all-pole model. A method for identifying and appropriately processing nasal portions was also proposed. These methods accurately determine pitch-peak positions and amplitudes. Applied to speech synthesis, they accurately model unvoiced-to-voiced transitions and subtle pitch-period variations, thereby improving synthesized-speech quality.

Chapter 3

Excitation-Pulse Model with Zeros Based on Pitch-Synchronous Analysis of the LPC Voiced-Speech Residual

3.1 Overview

The method proposed in Chapter2 achieved accurate extraction of macroscopic excitation information for an LPC speech analysis-synthesis system. This suggested that sufficiently efficient speech coding could be achieved in the low-bit-rate transmission range of 1.2–4.8 kbits/sec. Chapter1, on the other hand, showed that using LPC for low- and medium-bit-rate speech coding at 4.8–9.6 kbits/sec requires the extraction of microscopic excitation information, particularly from the voiced-speech residual, in addition to macroscopic information. It was also shown that the microscopic information contained in the residual consists of zero characteristics representing the difference from the all-pole model used in LPC, and that modeling these characteristics is an urgent requirement. This chapter demonstrates that pitch-synchronous analysis of the LPC voiced-speech residual is an effective and feasible means of extracting these zero characteristics as microscopic excitation information, and that the resulting pitch-synchronous amplitude spectrum contains highly distinctive zeros. As a voiced excitation model based on this analysis, an excitation pulse with zeros (ZEP) is proposed. It is obtained by modeling the pitch-synchronous amplitude spectrum of the voiced-speech residual by inverse LPC analysis and resynthesizing the spectrum from the resulting coefficients. The voiced-speech residual is then modeled experimentally, and the effectiveness of the method is confirmed through subjective evaluation of speech synthesized by LPC using the resynthesized residual as the voiced excitation[39].

3.2 Pitch-Synchronous Analysis of the Voiced-Speech Residual

3.2.1 Pitch-Synchronous Discrete Fourier Transform of the Voiced-Speech Residual

From the discussion in Section1.4, the voiced-speech residual $\epsilon(n)$ can be regarded as a quasi-periodic pulse train containing zero characteristics that represent the difference from the all-pole filter in LPC analysis. The voiced-speech residual is therefore defined by Eq.3.1.

$$\epsilon(n) = \sum_{r=-\infty}^{\infty} h_e(n+rN_p) \qquad (-\infty < r < +\infty) \tag{3.1}$$

Here, h_e is an impulse response having the above characteristics, and N_p is the pitch period. The Fourier transform $X(\mathrm{e}^{j\omega})$ of h_e is defined by Eq.3.2.

$$X(\mathrm{e}^{j\omega}) = \sum_{m=-\infty}^{\infty} h_e(m) e^{-j\omega m} \tag{3.2}$$

When a rectangular window is applied to $\epsilon(n)$ over $0 \leqq n \leqq N_p - 1$ to extract one pitch period, its Fourier transform $\tilde{X}(k)$ is calculated by Eq.3.3.

$$\tilde{X}(k) = \sum_{n=0}^{N_p-1} \epsilon(n) \mathrm{e}^{-j\frac{2\pi}{N_p}kn} \tag{3.3}$$

Substitution of Eq.3.1 into Eq.3.3 gives Eq.3.4.

$$\begin{aligned} \tilde{X}(k) &= \sum_{n=0}^{N_p-1} [\sum_{r=-\infty}^{\infty} h_e(n + rN_p)] \mathrm{e}^{-j\frac{2\pi}{N_p}kn} \\ &= \sum_{r=-\infty}^{\infty} [\sum_{n=0}^{N_p-1} h_e(n + rN_p)] \mathrm{e}^{-j\frac{2\pi}{N_p}k(n+rN_p)} \end{aligned} \tag{3.4}$$

Setting $m = n + rN_p (-\infty < m < +\infty)$ in Eq.3.4 gives Eq.3.5.

$$\tilde{X}(k) = \sum_{m=-\infty}^{\infty} h_e(m) \mathrm{e}^{-j\frac{2\pi}{N_p}km} \tag{3.5}$$

Equations3.2 and 3.5 yield Eq.3.6.

$$\tilde{X}(k) = X(\mathrm{e}^{j\frac{2\pi}{N_p}k}) \tag{3.6}$$

Thus, the pitch-synchronous discrete Fourier transform of $\epsilon(n)$ is equivalent to sampling $X(\mathrm{e}^{j\omega})$ at $\omega = 2\pi k/N_p$. The component at each sampling point accurately represents the frequency spectrum of the impulse response h_e, allowing its zero characteristics to be observed accurately without the influence of the pitch harmonic structure.

3.2.2 Determination of the Starting Point of Each Pitch Period in the Voiced-Speech Residual

Pitch-synchronous analysis of an original voiced-speech signal requires the starting point of each pitch period to be determined. In general, however, it is difficult to determine these points in an original speech signal or to make their amplitudes small, as shown in Fig.3.1. Consequently, applying pitch-synchronous analysis in an analysis-synthesis system has conventionally been difficult[40][41].

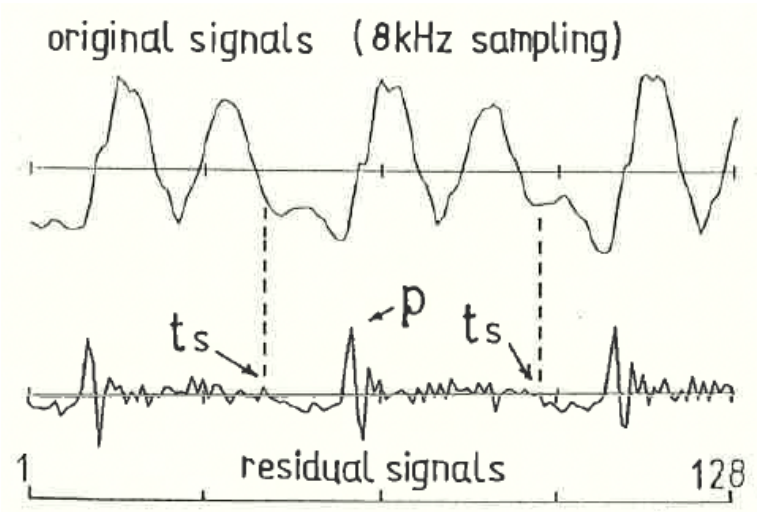


Figure 3.1: Voiced original and residual signals.

In the voiced-speech residual, by contrast, removal of the articulatory characteristics by the all-pole model makes periodicity clear. Its peaks p can be extracted automatically by the method described in Chapter2 or similar methods, and the starting point t_s of each pitch period can also be identified comparatively easily by visual inspection. As discussed later in Section3.2.4, the relationship between each peak position and pitch starting point has a statistical tendency dependent on the vocal-fold closing speed. This relationship can therefore be used to determine the starting point of each pitch period from the voiced-speech residual shown in Fig.3.1. Moreover, because residual power has decayed sufficiently near a starting point, pitch-synchronous analysis with a rectangular window minimizes frame-end effects and has low sensitivity to errors in the extraction position. In addition to the effect described in Section3.2.1, these facts permit automatic analysis of the voiced-speech residual and demonstrate the feasibility of an analysis-synthesis system using it. In the present method, each pitch period is first extracted by visual inspection to permit detailed observation of the pitch-synchronous amplitude spectrum. The starting points are subsequently determined from statistical analysis of peak positions and pitch periods, and the resulting effects are investigated.

3.2.3 Pitch-Synchronous Spectrum of the Voiced-Speech Residual

Each pitch period of the voiced-speech residual is first extracted by visual inspection as described in Section3.2.2, and its pitch-synchronous amplitude spectrum is calculated by the method in Section3.2.1. According to that method, the sampling positions on the frequency axis depend on pitch period N_p, as shown in Eq.3.6. For statistical analysis or modeling, however, a constant frequency-axis sampling interval is preferable. The N_p samples of each extracted pitch period are therefore inserted into the first N_p real-input points of a 256-point FFT, with zeros inserted into the remaining real and all imaginary inputs. This produces a uniform sampling interval $\Delta\omega = 2\pi/256$. Strictly speaking, the resulting spectrum differs from sampling $X(\mathrm{e}^{j\omega})$ in Eq.3.2 at $\omega = 2\pi/256$. It nevertheless performs approximate interpolation equivalently with little practical error, is computationally advantageous because it uses the FFT, and enables automation of pitch-synchronous analysis. Figure3.2 shows typical pitch-synchronous residual amplitude spectra of the five vowels for male and female speakers. Unlike the conventional short-time spectra shown in Fig.1.2, these spectra contain highly distinctive zeros. Although the bandwidths of the zeros are affected by all-pole modeling and vary with the phoneme, their frequencies are relatively insensitive to the phoneme and tend to differ between speakers. This result is consistent with the discussion in Section1.4: the zero characteristics of the voiced-speech residual arise from the vocal folds and the articulatory mechanism of nasal sounds and are given as the difference from the all-pole model. Their frequency characteristics are examined further in modeling the voiced-speech residual in Section3.3.

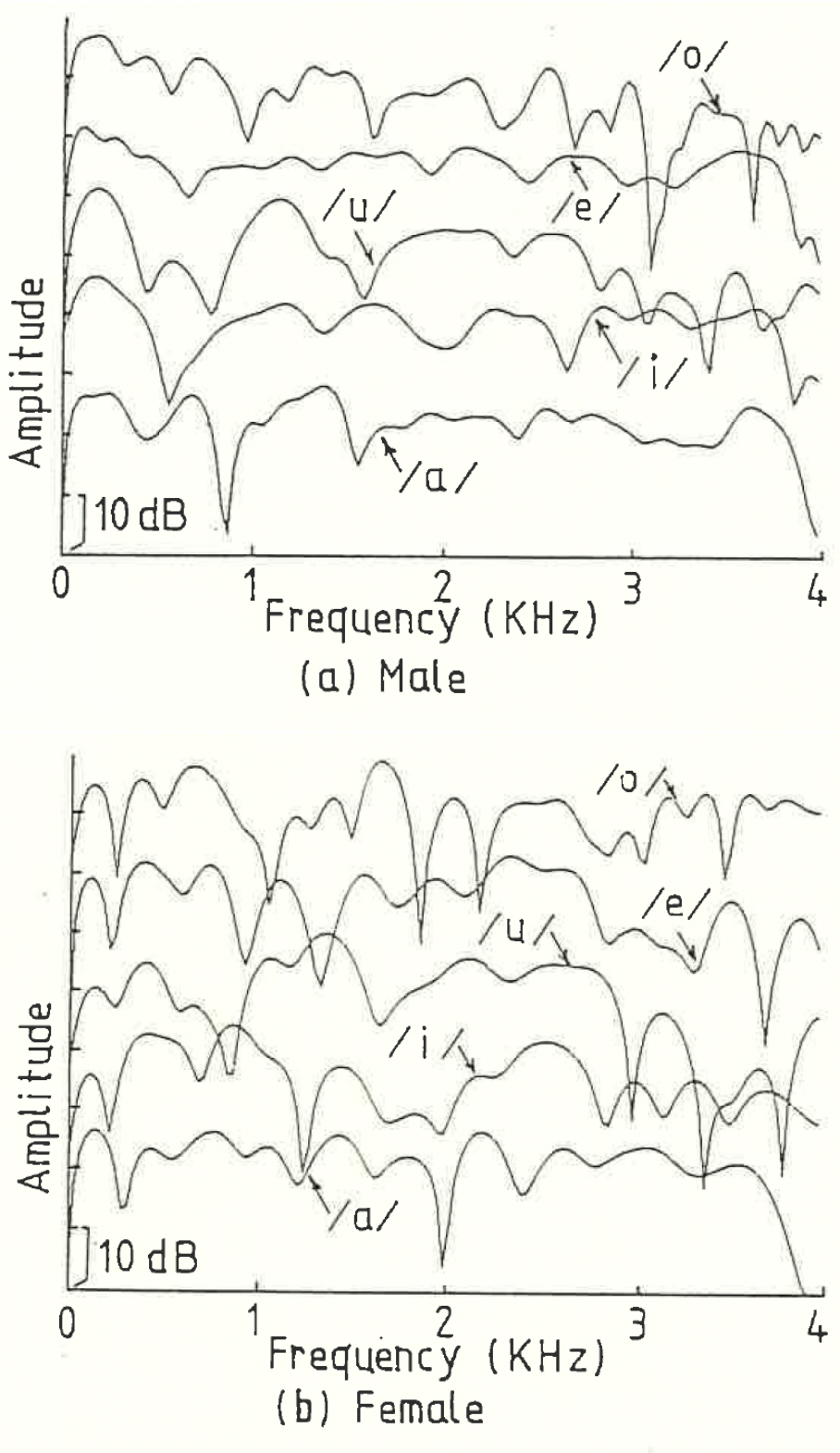


Figure 3.2: FFT-computed pitch synchronous amplitude spectrum of residual signals of five vowels.

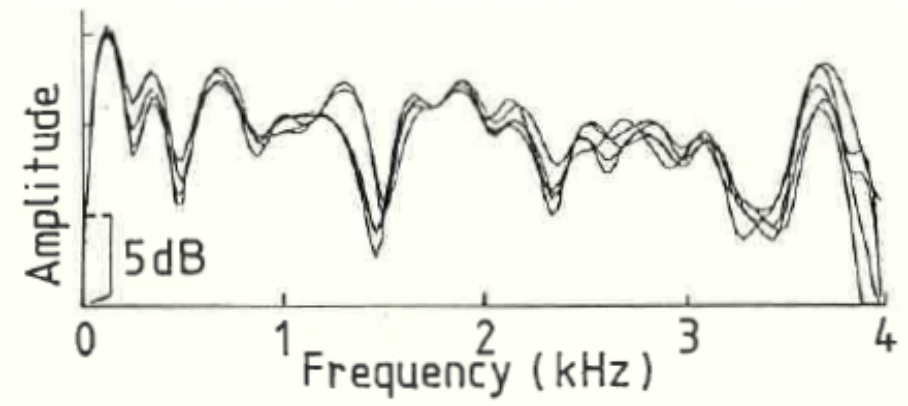


Figure 3.3: Pitch synchronous amplitude spectrum of residual of continuous pitch periods.

Figure3.3 shows pitch-synchronous amplitude spectra for five consecutive pitch periods of the residual of the voiced vowel /a/. Quasi-stationary zeros with high Q appear between consecutive periods, particularly in the low- to mid-frequency range.

The constancy described above is next analyzed quantitatively. Let frame m consist of J consecutive pitch periods of the voiced-speech residual, and let $Z_k(m,j)$ dB denote the pitch-synchronous log-amplitude spectrum obtained by a 256-point FFT for pitch period $j(1 \leqq j \leqq J)$ in frame m, where k denotes frequency and $1 \leqq k \leqq 129$. The mean spectrum $\mu_k(m)$ dB of $Z_k(m,j)$ in frame m is defined by Eq.3.7.

$$\mu_k(m) = \frac{1}{J}\sum_{j=1}^{J} Z_k(m,j) \qquad \text{dB} \tag{3.7}$$

For frame m, the spectral distortion (squared deviation) $Z(m,j)$ dB^2 from $\mu_k(m)$ for each pitch period and its standard deviation $\sigma(m)$ dB are defined by Eqs.3.8 and 3.9, respectively.

$$Z(m,j) = \frac{1}{128}\sum_{k=1}^{128}\{Z_k(m,j) - \mu_k(m)\}^2 \qquad \text{dB}^2 \tag{3.8}$$

$$\sigma(m) = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J} Z(m,j)} \qquad \text{dB} \tag{3.9}$$

This standard deviation $\sigma(m)$ is then averaged over all M frames obtained by updating one pitch period at a time. The mean standard deviation $\bar{\sigma}$ dB is defined by Eq.3.10.

$$\bar{\sigma} = \frac{1}{M}\sum_{m=1}^{M} \sigma(m) \qquad \text{dB} \tag{3.10}$$

The mean standard deviation $\bar{\sigma}$ expresses, as the standard deviation of spectral distortion, the degree to which the pitch-synchronous amplitude spectrum varies over J consecutive pitch periods and can therefore be regarded as a measure of its stationarity. Table3.2 gives $\bar{\sigma}$ for J=5 for the pitch-synchronous residual amplitude spectra of continuous /aiueo/ utterances by one male and one female speaker.

Table 3.1: Constancy of pitch synchronous spectrum of residual.

Table 3.2: Constancy of pitch synchronous spectrum of residual.

Sex	Mean standard deviation σ
Male	2.31 dB
Female	3.15 dB

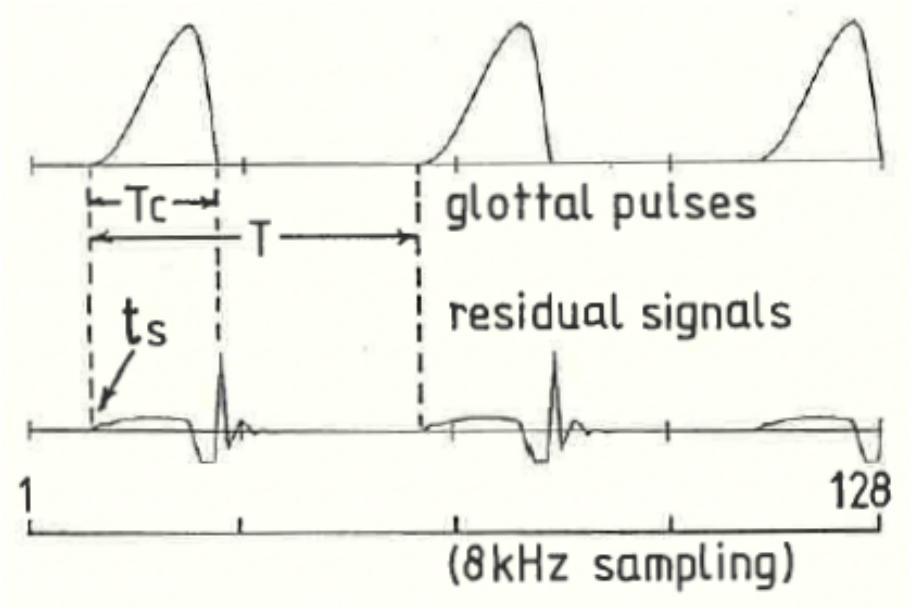


Figure 3.4: Relation between glottal pulses and residual signals.

The pitch-synchronous amplitude spectrum changes little over approximately five consecutive pitch periods (about 30 msec) and is therefore quasi-stationary. This indicates that model parameters can be extracted stably and suggests that parameter interpolation over 5–30 msec will be possible. The results in Figs.3.2 and 3.3 and Table3.2 suggest that synthesized-speech quality can be improved by modeling the pitch-synchronous amplitude spectrum of the voiced-speech residual with particular emphasis on its zeros.

3.2.4 Relationship Between the Peak and Starting Point of Each Pitch Period

The vocal-fold vibration waveform (glottal pulse) of a voiced excitation can generally be represented by the model shown in Fig.3.4[41]. In LPC analysis, the overall frequency characteristic (envelope) of the glottal pulse is absorbed into the LPC coefficients together with the vocal-tract characteristic. The residual waveform obtained by flattening the glottal-pulse frequency characteristic with an LPC inverse filter closely matches the ordinary voiced-speech residual waveform in LPC analysis, as shown in the figure. The phase relationship with the glottal pulse shows that the sharp peak occurs at the instant of rapid vocal-fold closure. The duty factor (T_c/T), the ratio of the interval T_c during which the vocal folds are open to one pitch period T in Fig.3.4, is known to have a statistical tendency related to pitch period T and vocal stress[42]. Because a voiced-residual peak is closely related to the glottal-pulse waveform, the duty factor in the voiced-speech residual should have a similar statistical tendency. If that tendency is known, the pitch starting point t_s can be determined approximately from the peak position. Table3.3 gives the mean and standard deviation of the ratio of peak position to pitch period, calculated from peak positions and pitch starting points determined by visual inspection for continuous /aiueo/ utterances (approximately 320 pitch periods each) by the male and female speakers used in Section3.2.3. For these

Table 3.3: The duty factor of voiced-speech residual.

Sex	Mean	Standard deviation
Male	42.29 %	0.87 %
Female	42.53 %	0.85 %

Table 3.4: Mean spectral distortion when duty factor was fixed.

Sex	Mean spectral distortion $\bar{SD}$
Male	2.42 dB
Female	2.43 dB

utterances, the peak occurred at approximately 42% of the pitch period from its starting point. The effect of determining pitch starting points from peak positions using this duty factor was then examined. Let PZ_k be a pitch-synchronous log-amplitude spectrum calculated by a 256-point FFT using starting points determined visually, as in Section3.2.3, and let PZ'_k be the corresponding spectrum calculated using starting points determined from peak positions and the values in Table3.3. Spectral distortion SD was calculated by Eq.3.11 and averaged over all pitch periods to obtain the mean spectral distortion $\bar{SD}$.

$$\bar{SD} = \frac{1}{128}\sum_{k=1}^{128}(PZ_k - PZ'_k)^2 \tag{3.11}$$

Table3.4 gives $\bar{SD}$ for the same speech used to calculate Table3.3. Considering that human hearing is comparatively insensitive to zeros, the effect of pitch-period extraction errors on the pitch-synchronous amplitude spectrum is sufficiently small. This can be attributed to the substantial decay of residual power near the pitch starting point, as discussed in Section3.2.3.

3.3 Modeling of the Voiced-Speech Residual – Zero Excitation Pulse Model

The pitch-synchronous amplitude spectrum of the voiced-speech residual is next modeled on the basis of the analyses in Sections3.2.3 and 3.2.4. As shown in Fig.3.2, this smooth spectrum contains highly distinctive zeros. Although various zero-weighted models are possible, an inverse linear-prediction model is used here. When the input spectrum is all-zero, it is inverted to form an all-pole spectrum, LPC analysis is applied, and the resulting LPC coefficients are used as model parameters for the input spectrum. The spectrum is reproduced at synthesis by reversing these operations.

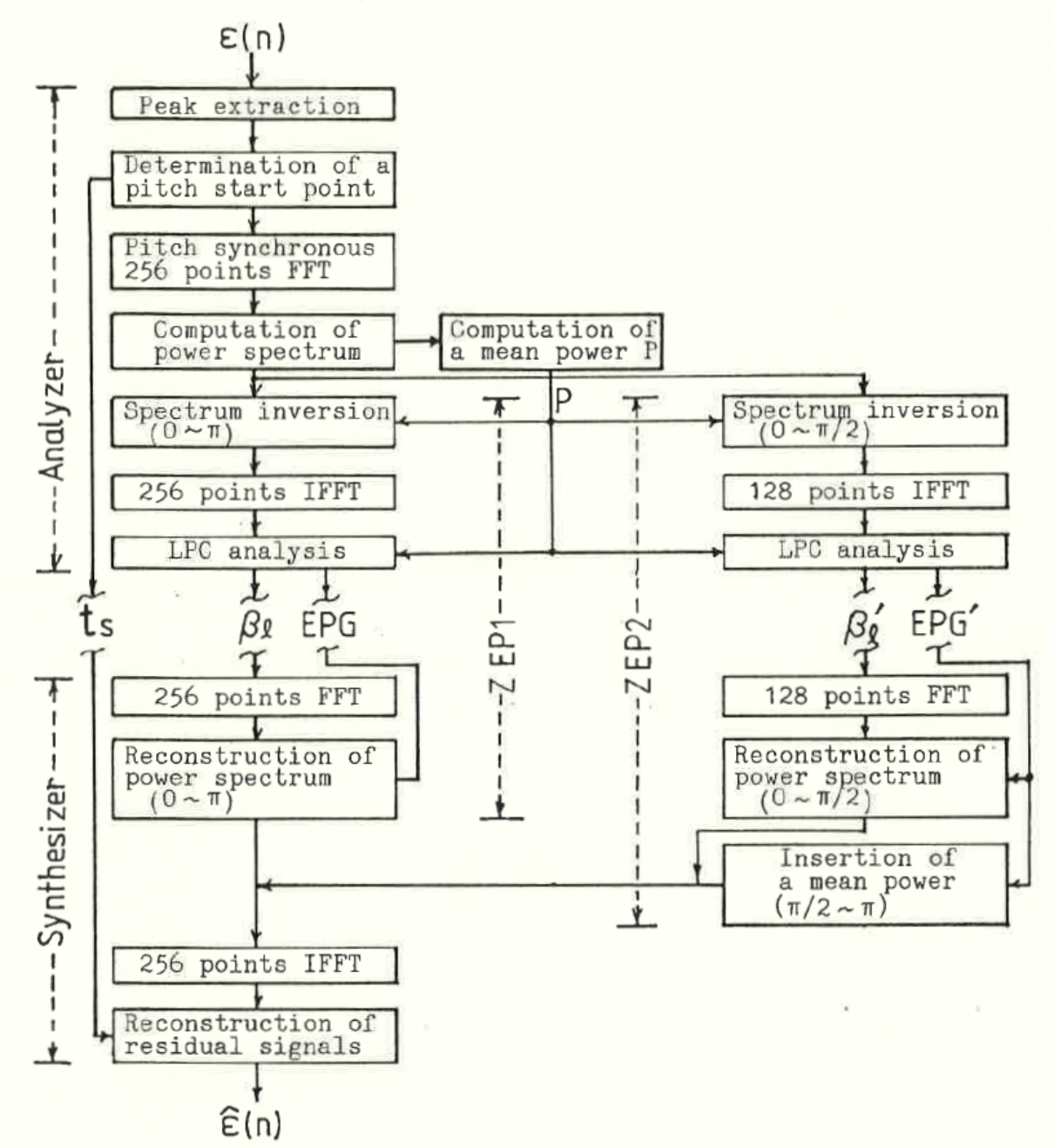


Figure 3.5: Block diagram of the analysis synthesis system using ZEP.

This method normally requires preprocessing to remove pole effects from the input speech and spectral smoothing to remove the pitch harmonic structure[15]. With conventional short-time Fourier spectra, however, zero characteristics are already blurred, as discussed in Section1.5, making zero-weighted spectral smoothing difficult. In the pitch-synchronous amplitude spectrum of the voiced-speech residual, by contrast, pole effects have been removed by the LPC inverse filter and the influence of pitch harmonic structure has been removed by pitch-synchronous analysis. The zero characteristics therefore appear clearly, allowing the inverse linear-prediction model to be applied readily. The detailed modeling procedure is described below. Figure3.5 is its block diagram. Two models are proposed; Model 1 is described first. Peaks are extracted from each pitch period of the voiced-speech residual, and the duty factor obtained in Table3.3 is used to determine the pitch starting points. Pitch-synchronous analysis is then performed by a 256-point FFT as in Section3.2.3. Let the resulting pitch-synchronous power spectrum be $|\hat{X}(k)|^2 \quad (0 \leqq k \leqq 256)$. After its mean power P is calculated, the inverse spectrum $|\hat{Y}(k)|^2$ is obtained from Eq.3.12.

$$|\hat{Y}(k)|^2 = P \cdot \frac{P}{|\hat{X}(k)|^2} = \frac{P^2}{|\hat{X}(k)|^2} \tag{3.12}$$

The inverse spectrum is modeled by the all-pole spectrum in Eq.3.13.

$$|\hat{Y}(k)|^2 = \frac{G^2}{|1 - \sum_{l=1}^{q} \beta_l \mathrm{e}^{-j\frac{2\pi}{256}kl}|^2} \tag{3.13}$$

Equations3.12 and 3.13 give Eq.3.14.

$$\begin{aligned} |\hat{X}(k)|^2 &= \frac{P^2}{G^2}|1 - \sum_{l=1}^{q} \beta_l \mathrm{e}^{-j\frac{2\pi}{256}kl}|^2 \\ &= EPG \cdot |1 - \sum_{l=1}^{q} \beta_l \mathrm{e}^{-j\frac{2\pi}{256}kl}|^2 \end{aligned} \tag{3.14}$$

The LPC coefficients β_l and gain G are obtained as follows. From the inverse spectrum calculated by Eq.3.12, the autocorrelation function r_l defined by the inverse DFT in Eq.3.15 is calculated by inverse FFT, and LPC analysis is then performed using it.

$$r_l = \frac{1}{N} \sum_{k=1}^{\frac{N}{2}} |\hat{Y}(k)|^2 \cos(\frac{2\pi}{N}kl) \qquad 0 \leqq l \leqq q \tag{3.15}$$

The LPC coefficients β_l and normalized power EPG obtained by this analysis are output as the voiced excitation model parameters. At synthesis, the inverse spectrum represented by the second factor on the right-hand side of Eq.3.14 is obtained by a 256-point FFT of β_l, and the power spectrum is reproduced using that equation. A zero-phase 256-point inverse FFT yields an impulse response, which is aligned with each peak position to reconstruct the voiced-speech residual. As shown in Eq.3.1, each 256-point impulse response may overlap adjacent pitch intervals. The voiced excitation model consisting of an excitation pulse with zeros obtained for each pitch period is called ZEP1 (zero excitation pulse 1). In ZEP1, the order of the inverse LPC coefficients β_l must be large enough to approximate zeros over the full frequency band. Selecting the optimum order is difficult because peaks other than zeros increase particularly in the middle and upper portions of the amplitude spectrum. Model 2 therefore models zeros only in the high-power region $0 \leqq \omega \leqq \pi/2$ and assumes a flat characteristic equal to normalized power EPG in $\pi/2 \leqq \omega \leqq \pi$. As shown in Fig.3.5, the FFT size for LPC analysis is then half that required for ZEP1. The voiced excitation model obtained for each pitch period by this procedure is called ZEP2.

3.4 Experiments on Voiced-Residual Modeling

3.4.1 Modeling Characteristics of ZEP

Figure3.6 shows an example of the amplitude spectrum of ZEP1 obtained by the above procedure. The order q of the LPC coefficients β_l was set to 24 in preliminary experiments. Because the region

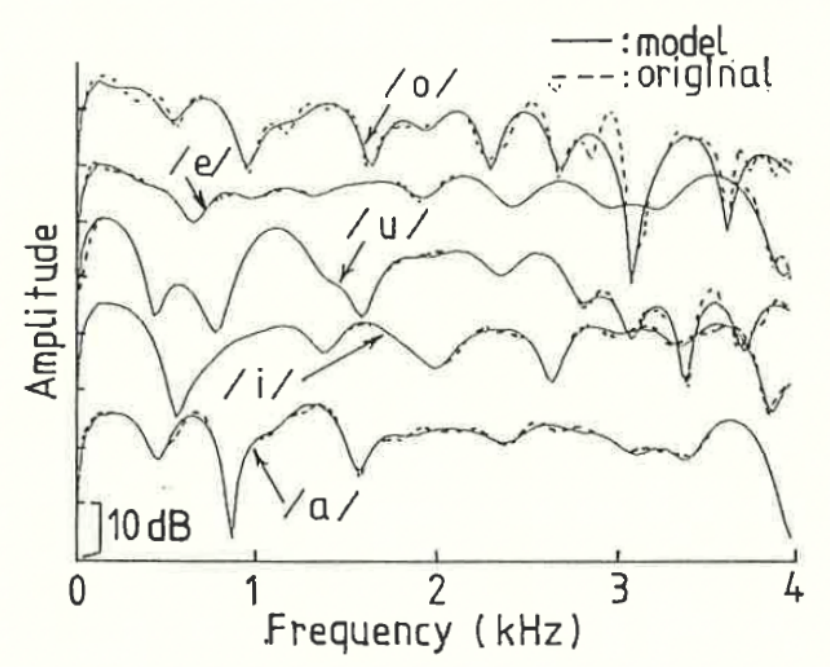


Figure 3.6: Amplitude spectrum of ZEP1 correspondingto fig.3.2(a).

near dc is difficult to approximate, dc power was set to one hundredth of its value at f=125 Hz, and the intervening region was reconstructed by linear interpolation in the power domain. Figure3.6 shows that the ZEP1 amplitude spectrum closely approximates the zero characteristics of the pitch-synchronous amplitude spectrum of the voiced-speech residual. Figure3.7 shows histograms of local minimum frequencies extracted from the amplitude spectrum of ZEP1 for each pitch period of continuous /aiueo/ utterances by the male and female speakers used in Table3.2. The results show that ZEP1 models zeros very stably and extracts them accurately. Zero frequencies have markedly nonuniform distributions largely independent of the phoneme, and the distributions differ between speakers.

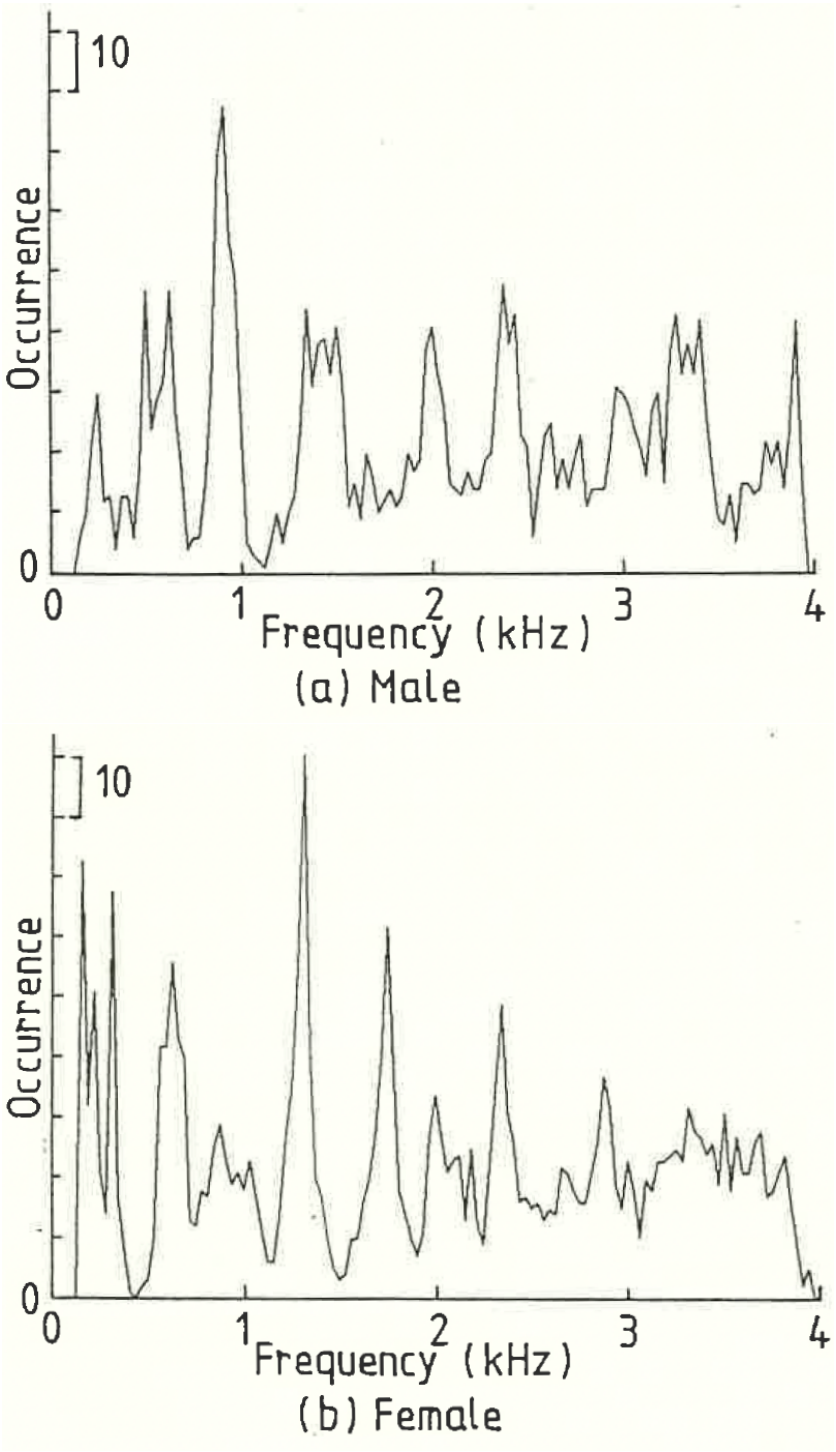


Figure 3.7: Distribution of zero frequency.

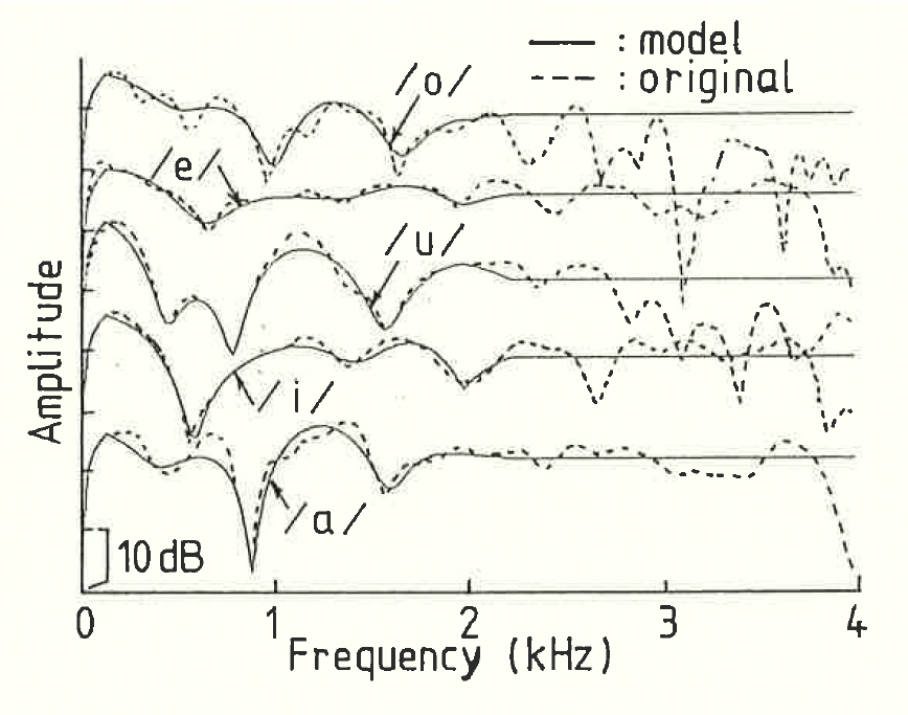


Figure 3.8: Amplitude spectrum of ZEP2 correspondingto fig.3.2(a).

Figure3.8 shows an example of the ZEP2 amplitude spectrum. Preliminary experiments showed that an order q of approximately eight is sufficient. The region near dc is processed as in ZEP1. Linear interpolation in the power domain is also performed so that the power spectrum connects smoothly in the logarithmic domain near $\omega = \pi/2$. Figure3.8 shows that the zero characteristics are approximated sufficiently well in the frequency band below $\omega = \pi/2$. Figure3.9 shows examples of the time-domain waveforms of ZEP1 and ZEP2.

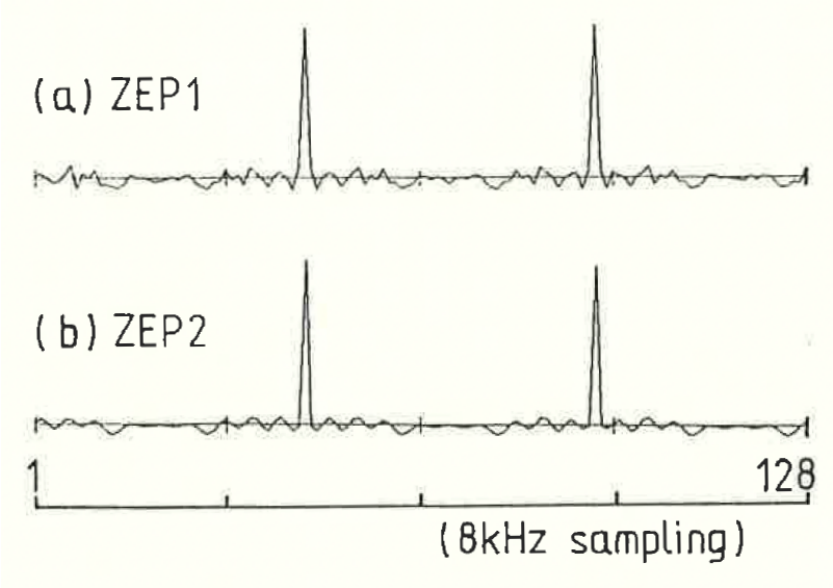


Figure 3.9: Examples of the time-domain waveform of ZEP.

3.4.2 Evaluation of LPC-Synthesized Speech Using ZEP

To evaluate speech synthesized using ZEP obtained by pitch-synchronous analysis of the voiced-speech residual, paired-comparison listening tests were performed on the following five types of speech. Each pair was presented four times with its order reversed. The original signals were continuous /aiueo/ utterances by one male and one female speaker, sampled at 8 KHz and quantized with 12-bit precision. Speech 1 was eight-bit log-PCM speech for each original signal. Speech 2–5 were obtained by applying the following models to the voiced-speech residual produced by tenth-order LSP analysis of each original signal with a 32-msec frame and a 5-msec update period, and then resynthesizing the speech. No perceptible quality difference was found between synthesis using pitch starting points selected visually and synthesis using points determined from peak positions according to Table3.3 in Section3.2.4.

Speech 2: For each pitch period, the peak position was extracted visually and the starting point determined from it using Table3.3. The impulse response was then obtained directly at zero phase from the pitch-synchronous amplitude spectrum calculated by the method of Section3.2.3, and the voiced-speech residual was resynthesized.

Speech 3: voiced-speech residual obtained with ZEP1 (q=24).

Speech 4: voiced-speech residual obtained with ZEP2 (q=8).

Speech 5: voiced-speech residual resynthesized by placing an impulse at each peak position extracted visually for every pitch period.

Because the objective was to evaluate directly the quality of ZEP1 and ZEP2 produced by pitch-synchronous analysis, their parameters β_l and EPG, and the LSP coefficients used in speech synthesis, were not quantized; peak positions were also determined visually. The subjects were nine men and one woman. Preference scores were obtained while they attended to richness, resonance, smoothness, clarity, and related qualities. Figure3.10 shows the results.

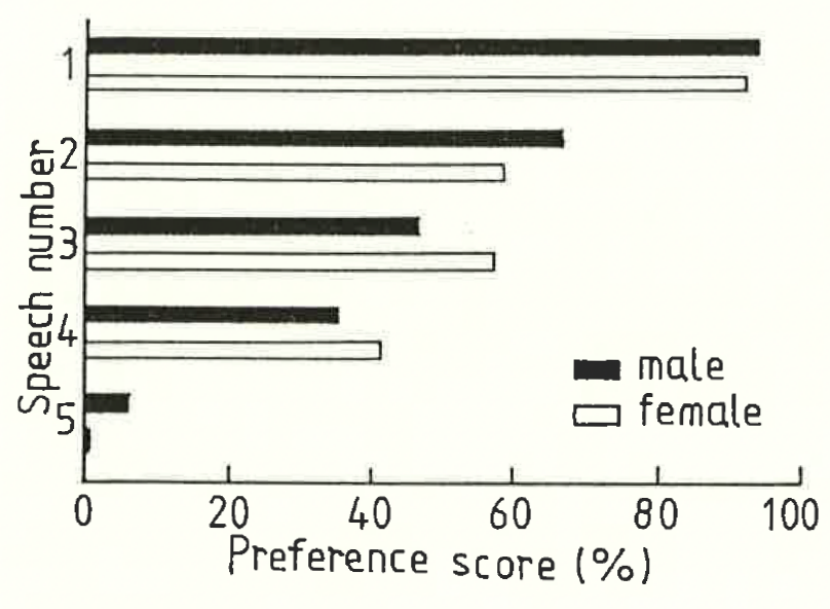


Figure 3.10: Preference score of each synthesized speech.

Although Speech 2, 3, and 4 did not equal the eight-bit log-PCM Speech 1, the synthesized speech using excitation with pitch-synchronous amplitude characteristics scored considerably higher than Speech 5, which used conventional impulse-train excitation. Quality differences among Speech 2, 3, and 4 were small, especially for the female voice, demonstrating effective ZEP modeling. Differences between Speech 3 and 4 were also small for both speakers, showing that zero characteristics below the middle frequency range contribute substantially to quality. Thus, a ZEP2 voiced excitation pulse can provide synthesized speech of adequate quality. Subjects reported that Speech 2 was smooth and approached the quality of Speech 1. Compared with Speech 5, Speech 2, 3, and 4 sounded richer or fuller, reproduced resonance—especially nasal resonance—more effectively, and substantially improved the lack of clarity of the phoneme /o/ that was conspicuous in Speech 5, particularly for the female voice. Few differences among Speech 2, 3, and 4 were perceived. These results suggest that zero characteristics are important in expressing speaker individuality and improve phonemic clarity in nasalized vowels. In smoothness, however, Speech 3 and 4 fell slightly short of Speech 1 and retained a somewhat discontinuous quality, particularly for the female voice. In addition to modeling error, this was probably caused by incomplete decay of the residual within each pitch period during pitch-synchronous analysis, which introduced error into the resulting spectrum. Because the pitch-synchronous amplitude spectrum was confirmed to be quasi-stationary over approximately 30 msec, substantial information compression should be possible by extracting a representative pitch period every 5–30 msec, performing pitch-synchronous analysis and modeling, and interpolating between frames. For ZEP2, the information required to model the residual consists only of approximately eighth-order LPC (or LSP) coefficients and power information. The method can therefore be implemented with approximately 50% more information than a conventional impulse-excited LPC vocoder. Because only one representative pitch interval need be extracted per frame, a simpler peak-extraction method than that proposed in Chapter 2 may also be used. The specific implementation and its performance are described in detail in Chapter4. Furthermore, the results in Fig.3.7 suggest that zero characteristics obtained by pitch-synchronous analysis of the voiced-speech residual contain information about speaker individuality. A speaker-identification method using zeros in the mean residual envelope has been reported[43]. Because the present pitch-synchronous analysis effectively avoids harmonic-component effects and clearly observes zero characteristics, its application to speaker recognition also appears promising and remains a topic for future work.

3.5 Summary

Chapter3 achieved accurate extraction of zero characteristics representing the difference from the all-pole model by applying pitch-synchronous analysis to the voiced-speech residual obtained by LPC analysis. Pitch-synchronous analysis was shown to be easier for the residual than for original speech, and the resulting amplitude spectrum was shown to be quasi-stationary over intervals of approximately 30 msec. The analysis can be automated and extracts zeros far more readily than pole-zero analysis methods such as Ab-s. On the basis of this analysis, ZEP, an excitation pulse with zeros, was proposed as a voiced excitation model and shown to model the zero characteristics of the voiced-speech residual accurately. The voiced-speech residual was modeled with ZEP and synthesized-speech quality was evaluated subjectively. The results showed that it produces higher-quality speech than conventional impulse-excited LPC. In particular, ZEP2, which models only zero characteristics below the middle frequency range, also produced speech of sufficiently high quality and is applicable to low-bit-rate vocoders.

Chapter 4

Pitch-Synchronous Mel-Inverse-LSP Analysis-Synthesis of the LPC Voiced-Speech Residual

4.1 Overview

Chapter3 showed that the pitch-synchronous amplitude spectrum obtained by pitch-synchronous analysis of the LPC voiced-speech residual contains highly distinctive zero characteristics. A method was proposed in which parameters obtained by inverse LPC analysis of this spectrum are transmitted as voiced excitation information; at the synthesis side, an excitation pulse with zeros (ZEP) is synthesized from these parameters and used as the voiced excitation for LPC synthesis. Although the full frequency range of the pitch-synchronous amplitude spectrum can be approximated accurately by approximately 24 LPC coefficients, an analysis-synthesis system that reduces the coded information without degrading quality must address the following issues: (1) Because the spectrum is quasi-stationary over several consecutive pitch periods, compression is possible by subsampling pitch periods for analysis and interpolating between them, but model parameters with good interpolation characteristics must be determined. (2) A direct time-domain filter construction method is required to facilitate interpolation of model parameters and synthesis of excitation pulses. (3) The analysis order must be reduced while minimizing quality degradation to reduce the coded information. This chapter addresses the first two issues by using LSP coefficients as model parameters in inverse LPC analysis and showing that an LSP inverse filter constructed from them can synthesize ZEP directly in the time domain. The third issue is addressed by mel-transforming the pitch-synchronous amplitude spectrum of the LPC voiced-speech residual and applying inverse LSP analysis. The resulting mel-LSP coefficients provide good modeling at an order of approximately twelve, and a direct construction method for an LSP inverse filter using these coefficients is proposed. For voiced-speech residuals synthesized by this analysis-synthesis system, the modeling order, quantization characteristics, and interpolation characteristics of mel-inverse LSP are examined and optimized using mean spectral distortion of the pitch-synchronous amplitude spectrum in the mel domain as the evaluation measure. Subjective evaluation of speech synthesized by the resulting pitch-synchronous mel-inverse-LSP analysis-synthesis system demonstrates that high-quality speech can be produced at approximately 4.8 Kbits/sec and that the method is applicable to a low-bit-rate vocoder[44].

4.2 Pitch-Synchronous Mel-Inverse-LSP Analysis-Synthesis System for the Voiced-Speech Residual

4.2.1 Synthesis of the Voiced-Speech Residual Based on Pitch-Synchronous Inverse LPC

Section3.3 of Chapter3 showed that the pitch-synchronous spectrum $|\hat{X}(k)|^2$ of the voiced-speech residual, which contains distinctive zero characteristics, can be modeled by LPC coefficients β_l and normalized power EPG. These are obtained by calculating the inverse spectrum $|\hat{Y}(k)|^2$ from Eq.3.12, calculating its autocorrelation function r_l from Eq.3.15, and applying LPC analysis. In principle, a voiced-speech residual can be synthesized from β_l and EPG by reversing these operations: calculating $|\hat{X}(k)|^2$ and obtaining its impulse response by inverse FFT. In fact, however, the spectrum is represented by the all-pole model using β_l and EPG in Eq.3.14. Because the second factor on the right-hand side of that equation is itself the frequency characteristic of the LPC inverse filter used to obtain a residual, its impulse response can be calculated from β_l and EPG using the LPC inverse filter of Eq.1.6, as expressed by Eq.4.1, and aligned with each peak position to resynthesize the voiced-speech residual.

$$\hat{h}_e(m) = \delta(m) - \sum_{i=1}^{q} \beta_i \delta(m-i) \tag{4.1}$$

$$\text{where} \begin{cases} \delta(m) = EPG & \text{for } m = 0 \\ \delta(m) = 0 & \text{for } m \neq 0 \end{cases}$$

Here, the impulse response $h_e(m)$ has a minimum-phase characteristic close to the phase characteristic of the original residual[15].

4.2.2 LSP Representation of the Inverse LPC Analysis-Synthesis System

When LPC coefficients obtained by inverse LPC analysis are encoded and transmitted, coefficient quantization and interpolation are required. Because LPC coefficients have relatively poor coding characteristics, they can instead be converted into LSP coefficients, which have good interpolation and quantization characteristics, for transmission and synthesis[25]. When LSP coefficients are introduced into pitch-synchronous inverse LPC analysis, they are calculated at the analysis side from the LPC coefficients obtained by inverse LPC analysis. At the synthesis side, an LSP inverse filter is constructed, and the voiced-speech residual for each pitch period can be synthesized directly by applying an impulse of amplitude EPG. The transfer function $A_q(z)$ of the LSP inverse filter is given by Eq.4.2[25].

$$\begin{aligned} A_q(z) = \frac{z^{-1}}{2}[& \sum_{i=2(i=even)}^{q} (c_i + z^{-1}) \prod_{j=0(j=even)}^{i-2} (1 + c_j z^{-1} + z^{-2}) \\ & - \prod_{j=2(j=even)}^{q} (1 + c_j z^{-1} + z^{-2}) \\ & + \sum_{i=1(i=odd)}^{q-1} (c_i + z^{-1}) \prod_{j=-1(j=odd)}^{i-2} (1 + c_j z^{-1} + z^{-2}) \\ & - \prod_{j=1(j=odd)}^{q-1} (1 + c_j z^{-1} + z^{-2})] + 1 \end{aligned}$$

$$\text{where} \begin{cases} c_i = -2\cos(2\pi f_i) & 1 \leqq i \leqq q \\ c_0 = c_{-1} = -z^{-1} \end{cases} \quad f_i(\text{Hz}) \text{ is an LSP}$$

4.2.3 Mel-LSP Representation of Inverse LPC Analysis

To synthesize a pitch-synchronous voiced-speech residual in the time domain, the pitch-synchronous amplitude spectrum in Eq.3.14 must be modeled over the full frequency range. As shown in Section3.4.1 of Chapter3, an order of approximately 24 is required for accurate full-band modeling. The order is therefore reduced by exploiting the mel-scale frequency characteristic of human hearing and representing

the LPC coefficients in inverse LPC analysis by mel-LSP coefficients. Human hearing is known to become less sensitive at higher frequencies[45], and the psychoacoustic mel scale represents this characteristic well. The mel scale can be obtained by selecting a=0.31 at 8 KHz in the frequency transformation from a linear to a nonlinear scale using the all-pass filter expressed by Eqs.4.2, 4.3, and 4.4[46][47].

$$\tilde{z}^{-1} = \frac{(z^{-1} - a)}{(1 - az^{-1})} \tag{4.2}$$

$$\tilde{\omega} = \omega + 2\tan^{-1}\frac{a\sin\omega}{1 - a\cos\omega} \tag{4.3}$$

$$\frac{\mathrm{d}\tilde{\omega}}{\mathrm{d}\omega} = \frac{1 - a^2}{1 + a^2 - 2a\cos\omega} \tag{4.4}$$

This transformation is used to mel-transform the pitch-synchronous inverse LPC analysis of the voiced-speech residual described in Section3.3. The inverse DFT used to calculate the autocorrelation function r_l in Eq.3.15 is replaced by the mel-domain inverse DFT[46] in Eqs.4.5 and 4.6; conventional LPC and LSP analyses are then applied to the resulting mel autocorrelation function $\tilde{r}_l$ to obtain the mel-LSP coefficients.

$$\tilde{r}_l = \frac{1}{N}\sum_{k=0}^{\frac{N}{2}} |\hat{Y}(k)|^2 U_{lk} \qquad \text{where } 0 \leqq l \leqq q \tag{4.5}$$

$$U_{lk} = \frac{\mathrm{d}\tilde{\omega}_k}{\mathrm{d}\omega_k}\cos(l\tilde{\omega}_k) \tag{4.6}$$

Substituting $\omega_k = 2\pi k/N$ and Eqs.4.3 and 4.4 into Eq.4.6 gives Eq.4.7.

$$U_{lk} = \frac{1 - a^2}{1 + a^2 - 2a\cos\frac{2\pi}{N}k}\cos(l(\frac{2\pi}{N}k + 2\tan^{-1}\frac{a\sin\frac{2\pi}{N}k}{1 - a\cos\frac{2\pi}{N}k})) \tag{4.7}$$

Equation4.5 can be calculated rapidly by calculating Eq.4.7 for $k = 0$–$N/2$ and $l = 0$–q in advance and storing the results in a table.

4.2.4 Mel Transformation of the LSP Inverse Filter

To resynthesize the voiced-speech residual for each pitch period from LSP coefficients obtained by inverse LPC analysis, the impulse response of the LSP inverse filter having the transfer function in Eq.4.2 is required, as described in Section4.2.2. A mel-LSP inverse filter is therefore constructed using directly the mel-LSP coefficients obtained by the process in Section4.2.3. Because the transfer function of the LSP inverse filter contains no feedback loop, as shown in Eq.4.2, z^{-1} in each component $z^{-1}/2$, $c_i + z^{-1}$, and $1 + c_j z^{-1} + z^{-2}$ can simply be replaced by $\tilde{z}^{-1}$ in Eq.4.2. Let $H0(z) = \tilde{z}^{-1}/2$, $HA_i(z) = \tilde{c}_i + \tilde{z}^{-1}$, $HB_j(z) = 1 + \tilde{c}_j\tilde{z}^{-1} + \tilde{z}^{-2}$, $\tilde{c}_i = -2\cos(2\pi\tilde{f}_i)$, where $\tilde{f}_i$ (Hz) is a mel-LSP coefficient, and $\tilde{c}_0 = \tilde{c}_{-1} = -\tilde{z}^{-1}$. Equation4.2 then gives Eqs.4.8, 4.9, and 4.10.

$$H0(z) = \frac{\tilde{z}^{-1}}{2} = \frac{1}{2}\frac{-a + z^{-1}}{1 - az^{-1}} \tag{4.8}$$

$$HA_i(z) = \tilde{c}_i + \tilde{z}^{-1} = \frac{(\tilde{c}_i - a) + (1 - \tilde{c}_i a)z^{-1}}{1 - az^{-1}} \tag{4.9}$$

$$HB_j(z) = 1 + \tilde{c}_j + \tilde{z}^{-1} + \tilde{z}^{-2} = \frac{(1 - \tilde{c}_j a + a^2) + (\tilde{c}_j - 4a + \tilde{c}_j a^2)z^{-1} + (1 - \tilde{c}_j a + a^2)z^{-2}}{1 - 2az^{-1} + az^{-2}} \tag{4.10}$$

From these three equations and Eq.4.2, the transfer function $A_q(z)$ of the mel-LSP inverse filter is determined by Eq.4.11.

$$\begin{aligned} A_q(z) = H0(z)[&\sum_{i=2(i=even)}^{q} HA_i(z)\prod_{j=0(j=even)}^{i-2} HB_j(z) - \prod_{j=2(j=even)}^{q} HB_j(z) \\ &+ \sum_{i=1(i=odd)}^{q-1} HA_i(z)\prod_{j=-1(j=odd)}^{i-2} HB_j(z) - \prod_{j=-1(j=odd)}^{q-1} HB_j(z)] + 1 \end{aligned} \qquad \text{where } HB_0 = HB_{-1} = 0 \tag{4.11}$$

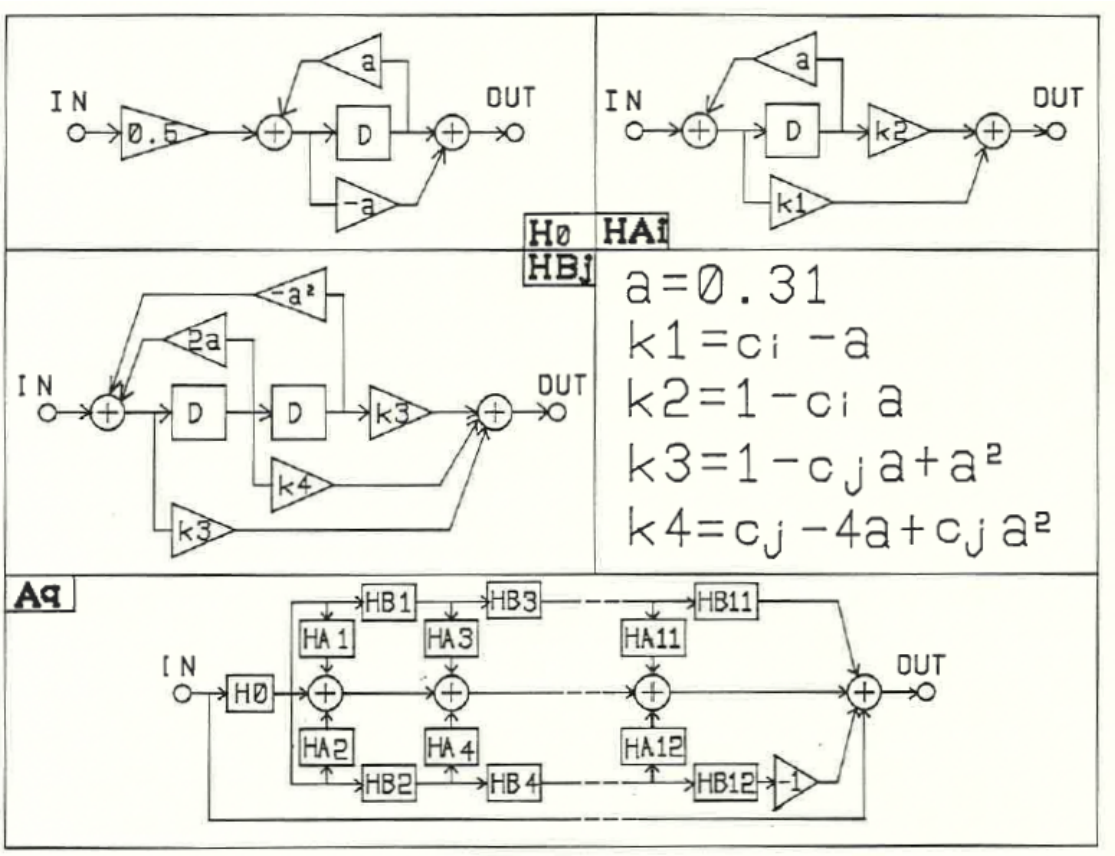


Figure 4.1: Mel LSP inverse filter.

Figure4.1 shows the mel-LSP inverse filter for an 8-KHz sampling frequency and an even analysis order.

4.2.5 Pitch-Synchronous Mel-Inverse-LSP Analysis-Synthesis Procedure

The procedures for pitch-synchronous mel-inverse-LSP analysis and synthesis of the voiced-speech residual based on the preceding principles are summarized below.

(Analysis procedure)

(1) Extract each pitch interval from the voiced-speech residual ϵ_n, append zeros to obtain N points, and calculate the pitch-synchronous power spectrum $|\hat{X}(k)|^2 \quad (0 \leqq k \leqq N)$ and its mean power P by an N-point FFT. Because the pitch interval of speech sampled at 8 KHz rarely exceeds 128 samples, N=128 is sufficient.

(2) Calculate the inverse spectrum $|\hat{Y}(k)|^2$ using Eq. (3-12) in Section 3.3 of Chapter 3; obtain the mel autocorrelation function $\tilde{r}_l \quad (0 \leqq l \leqq q)$ by the mel inverse DFT; and calculate the mel-LSP coefficients $\tilde{f}_l \quad (1 \leqq l \leqq q)$ by LPC and LSP analysis.

(3) From mean power P and gain G obtained during LPC analysis (Eq.3.13 in Section3.3), calculate the square root $\sqrt{EPG}$ of normalized power EPG from Eq.3.14 as amplitude information.

(4) Output the mel-LSP coefficients $\tilde{f}_l$, amplitude information $\sqrt{EPG}$ for each frame, and the peak position of each pitch period.

(Synthesis procedure)

(5) Use the transmitted mel-LSP coefficients $\tilde{f}_l$ to construct the mel-LSP inverse filter of Eqs.4.8–4.11 (Fig.4.1). Calculate the impulse response for each pitch period by applying an impulse of amplitude $\sqrt{EPG}$, and align it with each peak position to synthesize the voiced-speech residual $\hat{\epsilon}_n$.

4.3 Characteristics of Pitch-Synchronous Mel-Inverse LSP

This section examines the modeling order, quantization characteristics, and interpolation characteristics of mel-inverse LSP.

4.3.1 Modeling Order of Mel-Inverse LSP

Figure4.2 shows the modeling characteristics at each analysis order in the system of Section4.2.

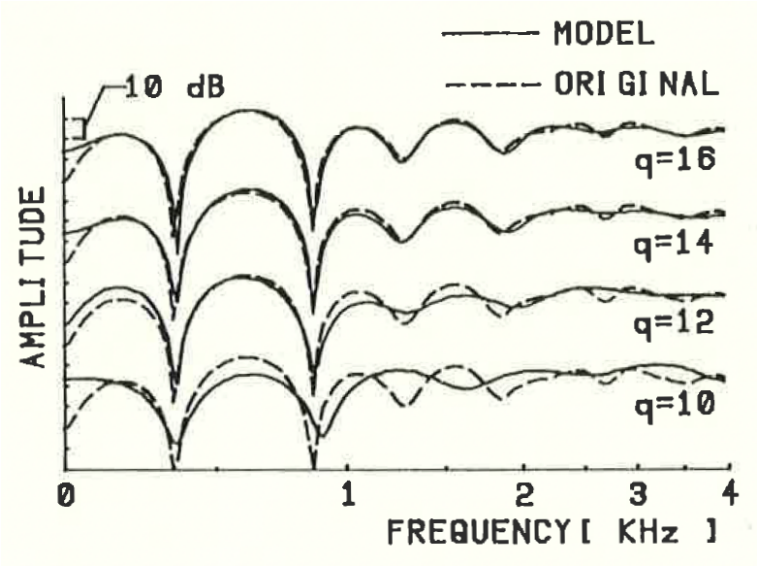


Figure 4.2: Characteristic of modeling at each order.

In this figure, the dashed line is the pitch-synchronous amplitude spectrum of one pitch period of a voiced-speech residual transformed to the mel-frequency scale, and the solid lines are the mel-scale amplitude spectra of impulse responses synthesized at each analysis order by the preceding system. At q=10, the zeros are not approximated adequately. At q=12, the high-Q zeros in the low-frequency region are approximated well, and at q=14 and 16, good approximation is obtained over the entire range. Figure4.3 shows a synthesized voiced-speech residual for q=14.

Figure 4.3: Original and synthesized voiced-speech residuals(q=14).

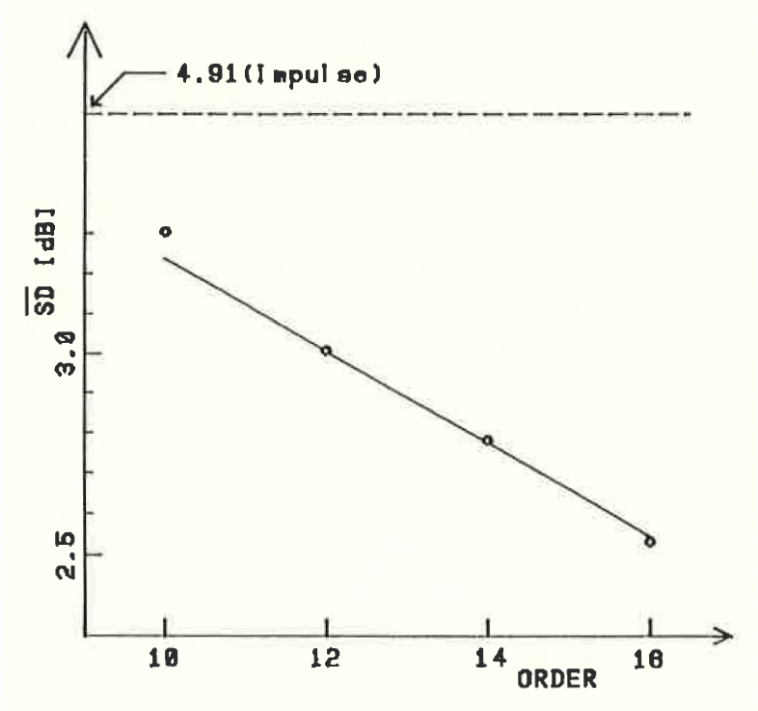


Figure 4.4: Relationship between mean spectral distortion and modeling order.

Figure4.4 shows mean spectral distortion versus analysis order. The approximately two-second sentence "In which prefecture is Yokohama?" spoken by one male and one female speaker was sampled at 8 KHz and subjected to tenth-order LSP analysis with a 5-msec frame period. Each pitch period of the resulting voiced-speech residual was extracted and mel-transformed. Let $PZ_0(k)$ be its mel log-amplitude spectrum calculated by a 128-point FFT, and let $PZ_1(k)$ be the corresponding spectrum of the impulse response synthesized for each pitch period by the preceding system. Spectral distortion SD

was calculated from Eq.4.11 and averaged over all pitch periods for each analysis order.

$$SD = \sqrt{\frac{1}{128}\sum_{k=1}^{64}(PZ_o(k) - PZ_1(k))^2} \tag{4.11}$$

No parameters were quantized or interpolated. Each pitch interval was extracted after its peak position was determined visually, using a point 42.3% of the interval between adjacent peaks as the pitch starting point[44]. $PZ_0(k)$ and $PZ_1(k)$ were normalized to equal mean amplitude. The dashed line in Fig.4.4 shows SD when no modeling was performed for individual pitch periods and a flat spectrum was used. SD decreases linearly as order increases, although this linearity is somewhat lost at order ten. Considering the low-frequency emphasis of human hearing[45], Figs.4.2 and 4.4 show that the pitch-synchronous amplitude spectrum of the voiced-speech residual can be approximated well at an order of approximately 12–14. This reduces the required order by approximately ten compared with the order of approximately 24 required by the inverse LPC analysis described in the preceding chapter.

4.3.2 Quantization Characteristics of Mel-Inverse LSP

For an analysis order of twelve, Fig.4.5 shows the ranges and frequency distributions of the mel-LSP coefficients obtained by pitch-synchronous mel-inverse-LSP analysis of the voiced-speech residual under the same conditions as Fig.4.4.

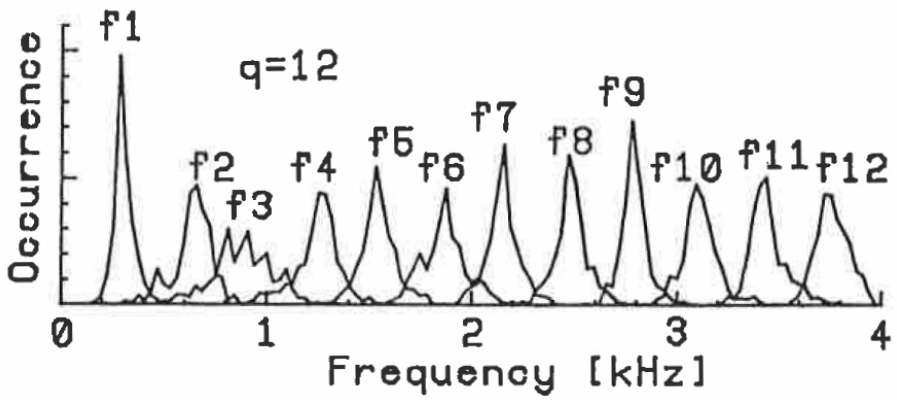


Figure 4.5: Distribution of mel LSP coefficiences(q=12).

The parameter ranges at individual orders are nearly equally spaced, with a maximum width of approximately 800 Hz. Efficient quantization can therefore be performed by quantizing each order within its own range[25]. For analysis order twelve, Fig.4.6 shows mean spectral distortion $\overline{SD}$ versus the number of quantization bits per order. Under the same conditions as Fig.4.4, $PZ_1(k)$ was calculated from impulse responses synthesized using unquantized mel-LSP coefficients, whereas $PZ_2(K)$ was calculated after the coefficients at every order were quantized with the same number of bits within the ranges in Fig.4.5. Mean spectral distortion was calculated as for Fig.4.4.

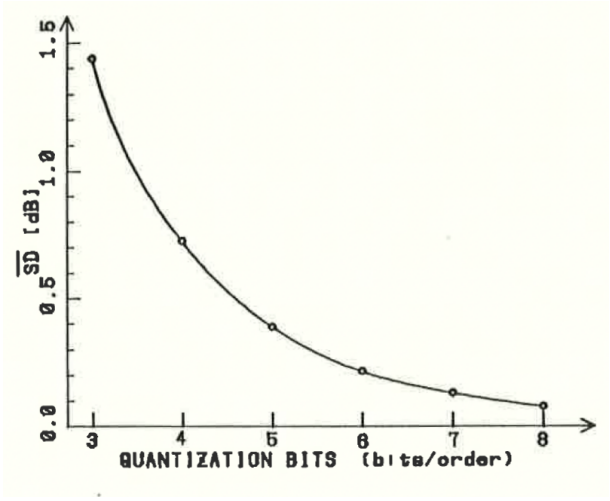


Figure 4.6: Relationship between mean spectral distortion and bit rate per order(q=12).

The figure shows that $\bar{SD}$ becomes sufficiently small when approximately four bits are used per order.

4.3.3 Interpolation Characteristics of Mel-Inverse LSP

As described in Chapter3, the pitch-synchronous amplitude spectrum is quasi-stationary over consecutive pitch periods. Coded information can therefore be compressed by extracting a representative pitch interval at prescribed frame intervals, subsampling pitch periods for analysis and coding, and encoding the mean pitch period of each frame. At synthesis, an impulse train is generated from the mean pitch period, the mel-LSP coefficients are interpolated between frames, and coefficients at each impulse position are calculated. This possibility is examined through the interpolation characteristics of mel-inverse LSP. The mel log-amplitude spectrum for every pitch period of the voiced-speech residual is used as the reference. Pitch periods are subsampled at prescribed intervals, spectra for the intervening periods are obtained by interpolation, and mean spectral distortion from the reference is defined as the interpolation characteristic. Normally, a spectrum obtained with a sufficiently short analysis interval is used as the reference. Because pitch-synchronous analysis cannot use an interval shorter than one pitch period, $PZ_1(k)$, obtained by mel-inverse-LSP analysis-synthesis of every pitch period under the same conditions as Fig.4.4, is used. Subsampling for interpolation must likewise be performed in pitch-period units. For each frame corresponding to an interpolation interval, the pitch period closest to the frame center is extracted and subjected to mel-inverse-LSP analysis; the resulting coefficients are used as the representative values for that frame. Linear interpolation between frames yields the coefficients corresponding to each reference peak position. Synthesis then gives the mel log-amplitude spectrum $PZ_3(k)$. Mean spectral distortion $\bar{SD}$ between $PZ_1(k)$ and $PZ_3(k)$ is calculated as for Fig.4.4 and used as the interpolation characteristic at that interval.

Figure4.7 gives $\bar{SD}$ for analysis order twelve and interpolation (frame) intervals of 10, 20, and 30 msec.

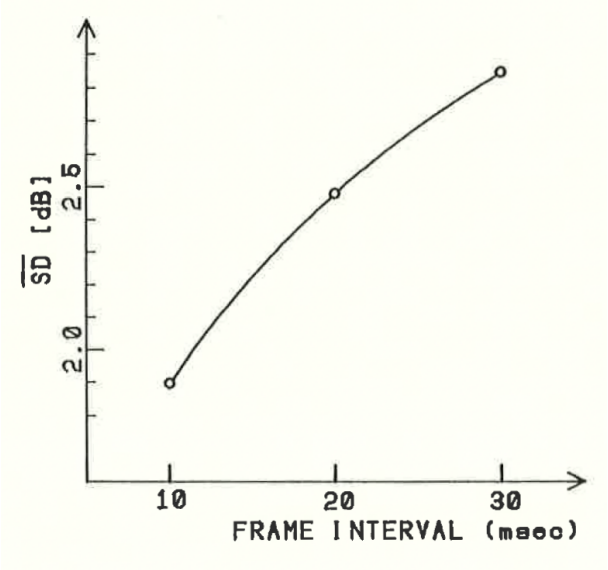


Figure 4.7: Relationship between mean spectral distortion and frame interval of interpolation(q=11).

Rather than representing only the interpolation characteristics of the mel-LSP coefficients, these results indicate a form of stationarity of the pitch-synchronous voiced-residual amplitude spectrum: the degree to which interpolation can reproduce the characteristics of pitch intervals omitted from analysis. In the preceding method, a mean spectral distortion of approximately 2.4 dB caused by errors in pitch extraction produced no perceptible quality difference. In light of that result, Fig.4.7 suggests that efficient coding can be achieved with an interpolation interval of approximately 10 or 20 msec.

4.4 Pitch-Synchronous Mel-Inverse-LSP Analysis-Synthesis System

Figure4.8 shows the pitch-synchronous mel-inverse-LSP analysis-synthesis system for the voiced-speech residual, based on the principles in Section4.2 and the experimental results in Section4.3. First, the mean pitch interval $T_p(m)$ is extracted at each prescribed frame interval (10–20 msec). From the voiced-speech residual $\epsilon_n(m)$ in frame m, the pitch interval nearest the center is extracted as the representative interval. By the procedure in Section4.2.5, the mel-LSP coefficients $\tilde{f}_l(m) \quad (0 \leqq l \leqq q)$ are calculated through the pitch-synchronous power spectrum $|\hat{X}_k(m)|^2$, inverse spectrum $|\hat{Y}_k(m)|^2$, and mel autocorrelation function $\tilde{r}_l(m)$. To improve interpolation between frames, the amplitude information is taken not as $\sqrt{EPG}$ but as $AMP(m)$, the square root of the mean power of $\epsilon_n(m)$ in each frame. The analysis side encodes $\tilde{f}_l(m)$, $AMP(m)$, and $T_p(m)$ for every frame. At the synthesis side, the transmitted information is decoded. An impulse train $\zeta_n(m)$ is generated for each frame while interpolating the mean pitch interval $T'_p(m)$ and amplitude $AMP'(m)$ between frames. The mel-LSP coefficients are then interpolated to obtain coefficients at each impulse position, the mel-LSP inverse filter is constructed (Fig.4.1), and $\zeta_n(m)$ is applied to synthesize the voiced-speech residual $\hat{\epsilon}_n(m)$.

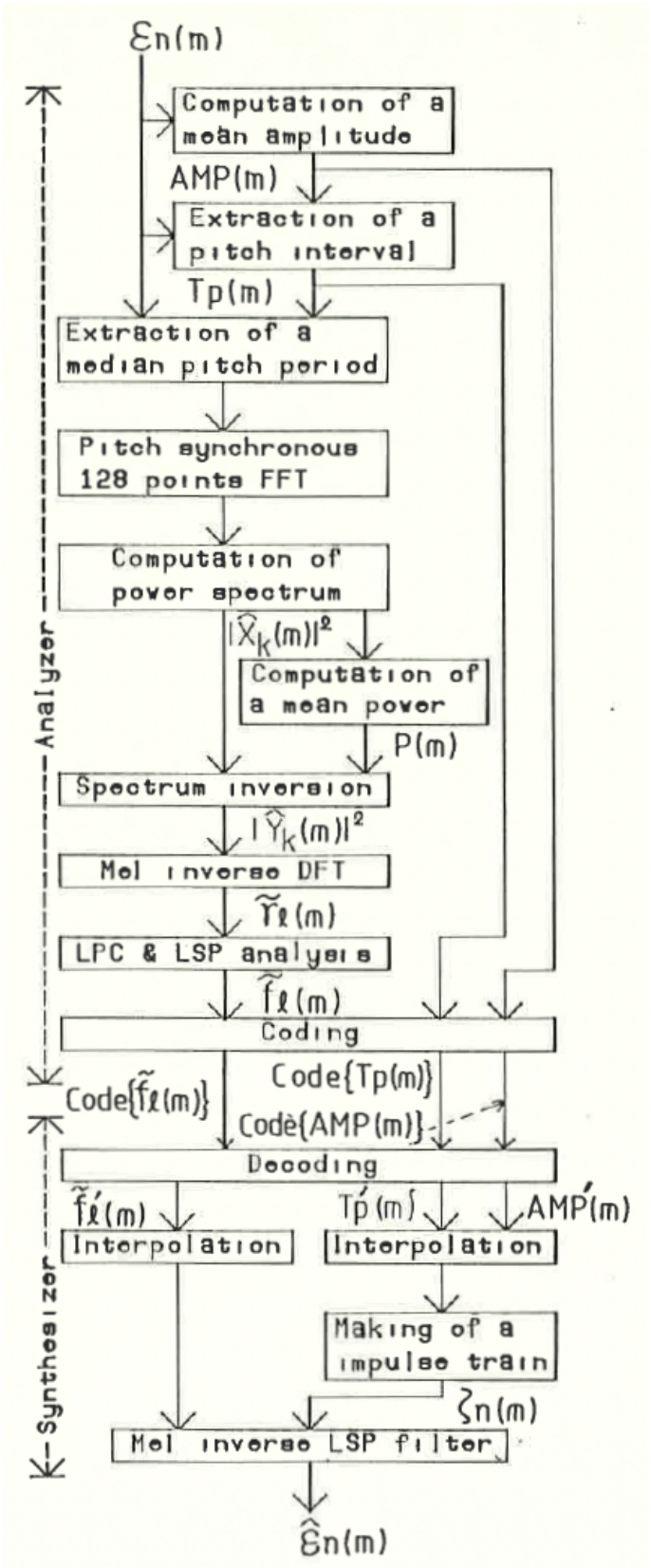


Figure 4.8: Block diagram of the pitchsynchronous mel inverse LSP analysis-synthesis system.

Table 4.1: Synthesized voices used for sound quality estimation.

Speech		1 (Proposed)	2	3
Vocal-tract characteristics		10th-order LSP 37 bits/frame		log PCM
Voiced excitation	Pitch	7 bits/frame		
	Amplitude	7 bits/frame		
	Coefficients	12th-order mel-LSP coefficients 45 bits/frame	–	
Unvoiced excitation		White noise		
Update period (msec)		20	5	
Bit rate (Kbits/s)		4.8	10.2	

Table 4.2: Result of sound quality estimation by MOS values.

Speech sample	MOS	σ
Proposed method (4.8 Kbits/s)	1.4	0.6
Conventional method (10.2 Kbits/s)	0.5	0.4
8-bit log PCM	3.0	0.4
7-bit log PCM	2.5	0.6
6-bit log PCM	1.9	0.5
5-bit log PCM	1.3	0.5
4-bit log PCM	0.5	0.3

4.5 Evaluation of Synthesized-Speech Quality

To evaluate speech synthesized using the voiced-speech residual produced by the system in Fig.4.8, opinion listening tests based on mean opinion scores (MOS)[48] were conducted for the speech signals in Table4.1. The original speech was the approximately two-second sentence "In which prefecture is Yokohama?" spoken by one male and one female speaker, sampled at 8 KHz and quantized with 12-bit precision. Speech 1 in Table4.1 was synthesized by the proposed method using the voiced-speech residual produced by the system in Fig.4.8. The references were LSP-synthesized speech using an impulse train as voiced excitation (Speech 2) and four- to eight-bit log-PCM speech (Speech 3). The subjects were ten men and women. Table4.2 gives the MOS and standard deviation σ for each signal. The tests showed that speech synthesized by the proposed method at 4.8 Kbits/sec had higher quality than LSP-synthesized speech using impulse-train excitation at slightly above 9.6 Kbits/sec and was equivalent in quality to five-bit log-PCM speech. Because five-bit log PCM has a bit rate of 40 Kbits/sec, the proposed method achieves a compression ratio of approximately one tenth, demonstrating substantial information compression. It has been reported that the quality of impulse-excited LSP synthesis saturates above 4.8 Kbits/sec[25]; the proposed method can therefore be regarded as raising the quality-saturation limit of conventional LSP synthesis. Listeners reported reduced noisiness, richer quality, and improved intelligibility compared with impulse-excited LSP speech. Some, however, found unvoiced portions and transitions between unvoiced and voiced portions difficult to understand. This degradation is attributed to the vocoder configuration, which performs voiced/unvoiced decisions and excites unvoiced portions with white noise as in conventional methods, and remains a topic for future work.

4.6 Summary

Chapter4 proposed pitch-synchronous mel-inverse-LSP analysis-synthesis as a new coding method for the LPC voiced-speech residual. The pitch-synchronous amplitude spectrum of the voiced-speech residual, including its zero characteristics, was shown to be encoded efficiently at an order of approximately twelve using mel-LSP coefficients obtained by inverse LSP analysis in the mel domain. A mel-LSP inverse filter that synthesizes the voiced-speech residual directly in the time domain from these coefficients was also proposed, demonstrating straightforward residual synthesis. The modeling order, quantization characteristics, and interpolation characteristics of mel-inverse LSP were then examined and optimized

by evaluating mean spectral distortion of the pitch-synchronous amplitude spectrum of the synthesized voiced-speech residual in the mel domain. Finally, subjective quality evaluation showed that high-quality speech can be synthesized at a bit rate of approximately 4.8 Kbits/sec and demonstrated the method's applicability to low-bit-rate vocoders.

Chapter 5

Conclusion

The objective of this study was to achieve accurate extraction of excitation information in LPC speech analysis-synthesis systems based on speech production models at low and medium bit rates, to develop a new method for modeling the excitation on the basis of an excitation-generation model, and thereby to improve the speech quality and coding efficiency of analysis-synthesis systems. To this end, a method for accurately determining the positions and amplitudes of pitch peaks in the LPC residual was investigated as a means of extracting macroscopic excitation information. In addition, the extraction and modeling of microscopic excitation information from the LPC voiced-speech residual were investigated on the basis of an excitation-generation model.

The Introduction discussed the significance of efficient speech coding, the position of the present study, and its objectives and outline, establishing that this study concerns the extraction and modeling of excitation information in speech analysis-synthesis systems.

Chapter1 reviewed speech synthesis based on speech production models, particularly the LPC method, as well as the extraction and modeling of excitation information from the residual. Previous studies of excitation information extraction and modeling and their associated problems were examined, thereby clarifying the significance of the present study.

Chapter2 addressed the extraction of macroscopic excitation information from the residual—that is, accurate extraction of individual pitch peaks—which is the most fundamental problem in excitation information extraction for LPC systems. In particular, a method was presented for accurately extracting pitch periodicity from a residual containing components other than pitch peaks. A time-domain signal was regarded as a frequency spectrum, and its amplitude-envelope characteristic was obtained as an all-pole model by applying LPC analysis. Nasal portions were also identified and processed appropriately. These procedures made it possible to determine the positions and amplitudes of pitch peaks accurately. When the excitation information obtained from this system was applied to a speech analysis-synthesis system, information at transitions between unvoiced and voiced speech and subtle variations in pitch period could be modeled accurately, thereby improving the quality of synthesized speech.

Chapter3 proposed a method for extracting microscopic excitation information, in addition to macroscopic information, from the voiced-speech residual in LPC. Chapter2 achieved accurate extraction of macroscopic excitation information for an LPC speech analysis-synthesis system and suggested that sufficiently efficient speech coding could be achieved in the low-bit-rate transmission range of 1.2–4.8 kbits/sec. Chapter1, on the other hand, showed that when LPC is used for low- and medium-bit-rate speech coding at 4.8–9.6 kbits/sec, zero characteristics representing the difference from the all-pole model must be extracted from the voiced-speech residual and modeled as microscopic excitation information. The method proposed here addresses this requirement. Pitch-synchronous analysis was shown to be effective for this purpose, and the resulting pitch-synchronous amplitude spectrum was shown to contain highly distinctive zero characteristics. This analysis was also shown to be easier for the residual than for the original speech signal. The pitch-synchronous amplitude spectrum exhibited quasi-stationarity over intervals of approximately 30 msec, and automatic analysis was shown to be possible. On the basis of this analysis, the excitation pulse with zeros, ZEP, was proposed and its effectiveness confirmed. The voiced-speech residual was then modeled using ZEP, and subjective evaluation of speech synthesized with this residual demonstrated higher quality than conventional impulse-excited LPC speech.

Chapter4 proposed pitch-synchronous mel-inverse-LSP analysis-synthesis as a new method for coding the pitch-synchronous amplitude spectrum of the voiced-speech residual, based on the pitch-synchronous analysis proposed in Chapter3, for application to low-bit-rate speech coding at approximately 4.8 kbits/sec. The spectrum, with its distinctive zero characteristics, was shown to be modeled accurately at an analysis order of approximately twelve using mel-LSP coefficients obtained by inverse LSP analysis in the mel domain. A mel-inverse-LSP filter that synthesizes the voiced-speech residual directly in the time domain from the mel-LSP coefficients was also proposed, enabling straightforward synthesis of the residual. The modeling order, quantization characteristics, and interpolation characteristics of mel-inverse LSP were then examined and optimized. Finally, subjective evaluation of speech synthesized by the system demonstrated that high-quality speech could be produced at approximately 4.8 kbits/sec and established the applicability of the proposed microscopic excitation modeling system to low-bit-rate vocoders.

As described above, accurate extraction and modeling of macroscopic excitation information—namely, pitch periodicity—is the most important issue in LPC analysis-synthesis systems based on speech production models. The method proposed in Chapter2 achieved this objective. Furthermore, Chapter3 proposed a method for extracting microscopic excitation information from the voiced-speech residual, and Chapter4 proposed a method for modeling that information. These methods improve the limiting performance of LPC and demonstrate the feasibility of practical low- and medium-bit-rate speech coding systems required by an advanced information society. The objectives of the present study have thus been achieved.

Acknowledgments

This research was conducted while the author was enrolled in the doctoral program of the Graduate School of Science and Engineering at Keio University, under the supervision of Dr. Shinsaku Mori and Dr. Shinji Ozawa of the Faculty of Science and Technology. The author wishes to express his profound gratitude to Dr. Mori for his invaluable guidance and encouragement throughout the course of this research. The author is also deeply indebted to Dr. Ozawa for his constant guidance and encouragement concerning every detail of the work, for his valuable advice, and for directing the research along an appropriate course.

The author is sincerely grateful to Professor Taro Shimogo and Associate Professor Masao Nakagawa of the Faculty of Science and Technology at Keio University, and to Professor Kenichi Kido, Director of the Research Center for Applied Information Sciences at Tohoku University, for their valuable suggestions concerning the contents of this dissertation.

The author received enthusiastic and dedicated guidance from Dr. Shin Miyauchi, Dr. Ken'ichi Onda, and Dr. Minami Miyauchi in carrying out this research. The author also wishes to acknowledge the generous cooperation of Yoshihiko Horio, Hidenobu Harasaki, Takao Mitsuishi, Hideshi Mizuno, Heito Zen, Shuichi Arai, and all the other members of the Ozawa and Mori Laboratories.

Finally, the author wishes to express his gratitude to his parents and to his elder brother and sister-in-law, whose loving support and encouragement sustained him quietly throughout this research. Above all, he offers his heartfelt thanks to his beloved wife, Keiko, who remained his emotional support and continued to encourage him throughout the completion of this research and dissertation.

Bibliography

[1] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell System Technical Journal*, Vol. 27, pp. 379–423 and 623–656, July and October 1948.

[2] Yotaro Yatsuzuka, “Medium- and High-Bit-Rate Speech Coding,” *Journal of the Institute of Electronics, Information and Communication Engineers*, Vol. 70, No. 4, pp. 392–400 (1987), in Japanese.

[3] Miyagawa, Kido, et al., *Digital Signal Processing*, IECE, pp. 43–45 (1983), in Japanese.

[4] H. Dudley, “The Vocoder,” *Bell Laboratories Record*, Vol. 18, No. 4, pp. 122–126 (1939).

[5] 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).

[6] Fukabayashi, Suzuki, ”Speech analysis using pole-zero linear model.”, IEC, J58-A, 5, pp. 270-277(1977).

[7] Ishizaki, ”Identification of order of pole-zero model in speech analysis.”, IEC, J60-A, 4, pp. 423-424(1977).

[8] Morikawa, Fujisaki, ”Speech analysis using the simultaneous estimation method of ARMA parameters.”, ASJ, Speech study group documentation, S76-53(1976).

[9] Sagayama, Furui, ”Maximum likelihood estimation of speech spectrum using pole-zero model.”, ASJ, Speech study group documentation, S76-37(1977).

[10] Wiener, N., ”Extrapolation, interpolation, and smoothing of stationary time series.”, MIT Press, Cambridge, Massachusetts(1966).

[11] Itakura, F. and Saito, S., ”Analysis synthesis telephony based on the maximum likelihood method.”, Reports of the 6th Int. Cong. Acoust., C-5-5 (1968).

[12] Atal, B. S. and Schroeder, M. R., ”Predictive coding of speech signals.”, Reports of 6th Int. Cong. Acoust., C-5-4(1968).

[13] 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); and “The SIFT Algorithm for Fundamental Frequency Estimation,” Vol. AU-20, No. 5, pp. 367–377 (1972).

[14] J. Makhoul, ”Linear Prediction-A Tutorial Review.”, Proc. IEEE, Vol. 63, pp. 561-580(1975).

[15] J. Makhoul, ”Spectral Linear Prediction:Properties and Applications.”, IEEE Trans. Acoust. Speech, Signal Processing, ASSP-23, 3, pp. 283-296(1975).

[16] Supervision by Tanetoshi Miura, ”Hearing and Speech”, IECE, pp. 323-329(1980).

[17] Carr, P. B. and Trill, D., ”Long-term larynx excitation spectra.”, J. Acoust. Soc. Am., 36, 11, pp. 2033-2040(1964).

[18] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-24, No. 5, pp. 399–418 (1976).

[19] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average Magnitude Difference Function Pitch Extractor.", IEEE Trans. Acoust. Speech and Signal Proc. ASSP-22, pp. 353-362(1974).

[20] M. M. Sondhi, "New Methods of Pitch Extraction.", IEEE Trans. Audio and Electroacoustics, AU-16, No. 2, pp. 262-266(1968).

[21] B. Gold and L. R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain.", J. Acoust. Soc. Am., Vol. 46, No. 2, Pt. 2, pp. 442-448, August(1969).

[22] A. M. Noll, "Cepstrum Pitch Determination.", J. Acoust. Soc. Am., 41, pp. 293-309(1967).

[23] Kitawaki, Itakura, Saito, "Optimal code structure in PARCOR-type speech analysis and synthesis system", IECE, J61-A, 2, pp. 119-126(1978).

[24] Kitawaki, Itakura, "Efficient coding of speech using nonlinear quantization and nonuniform sampling of PARCOR coefficients.", IECE, J61-A, 6, pp. 543-550(1978).

[25] Sugamura, Itakura, "Speech information compression using line spectrum pair (LSP) speech analysis and synthesis method.", IECE, J64-A, 8, pp. 599-606(1981).

[26] B. Bogert, M. Healy, and J. Tukey, "The Quefrency Analysis of Time Series for Echoes.", Proc. Symp. Time Series Analysis, Chap. 15, pp. 209-243(1963).

[27] Imai, Abe, "Extraction of spectral envelope by improved cepstral method.", IECE, J62-A, 4, pp. 217-223(1979).

[28] Imai, Fuichi, "Impulse train equivalent sound source for high quality speech synthesis.", IECE, J68-A, 11, pp. 1242-1252(1985).

[29] Atal, B. S. et al, "A new method of LPC excitation for producing natural-sounding speech at low bit rates.", Proc. IEEE, ICASSP82, pp. 614-617(1982).

[30] Atal, B. S. and Schroeder, M. R., "Stochastic coding of speech at very low bit rate.", Proc. ICASSP 84, pp. 1610-1613(1984).

[31] Un, C. K., Magill, D. T., "The residual-excited linear prediction vocoder with transmission rate below 9. 6 k bit/s.", IEEE Trans., COM-23, 12, pp. 1466-1474(1975).

[32] Fumitada Itakura, "A New Speech Analysis-Synthesis Method: PARCOR," *Nikkei Electronics*, February 12, pp. 58–75 (1973), in Japanese.

[33] Atsushi Mano and Shinji Ozawa, "Excitation-Source Information Extraction from the Amplitude-Envelope Characteristics of Residual Waveforms Using LPC Analysis," *Transactions of the IECE of Japan*, Vol. J67-A, No. 1, pp. 72–73 (1984), in Japanese.

[34] Rabiner, L. R., Schafer, R. W., Translated by Hisayoshi Suzuki, "Audio digital signal processing.", Corona Corp., p. 172, p. 199(1983).

[35] Brigham, E. Oran, Translated by Miyagawa and Imai, "Fast fourier transform.", Japan science and technology(1981).

[36] Agui, Nakajima, "Computer audio processing.", Sanpo publications, pp. 132-134(1981).

[37] Itakura, Saito, "Estimation of speech spectral density and formant frequency using statistical methods.", IECE, J53-A, 1, pp. 35-42(1970).

[38] Saito, Nakada, "Fundamentals of speech information processing.", Ohmsha, P. 120(1981).

[39] Atsushi Mano and Shinji Ozawa, "Excitation-Pulse Model with Zeros Based on Pitch-Synchronous Analysis of LPC Voiced-Speech Residuals," *Transactions of the IEICE of Japan, Series A*, Vol. J70-A, No. 6, pp. 960–968 (1987), in Japanese.

[40] M. V. Mathews, J. E. Miller and E. E. David, Jr., "Pitch Synchronous Analysis of Voiced Sounds.", J. Acoust. Soc. Am., 33, pp. 179-186(1961).

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

[42] Miller, R. L., "Nature of the vocal cord wave.", J. Acoust. Soc. Am., 31, p. 667 (1959).

[43] Kashiwagi, Nakamura, Takanashi, "Speaker identification using spectral envelope of linear prediction residuals.", IECE, J68-A, 7, pp. 702-703(1985).

[44] Atsushi Mano and Shinji Ozawa, "Pitch-Synchronous Mel-Inverse-LSP Analysis-Synthesis of LPC Voiced-Speech Residuals," *Transactions of the IEICE of Japan, Series A*, Vol. J71-A, No. 3, pp. 634–641 (1988), in Japanese.

[45] S. S. Stevens and J. Volkmann, "A Scale for the Measurement of the Psychological Magnitude Pitch," *Journal of the Acoustical Society of America*, Vol. 8, pp. 185–190 (1937).

[46] H. W. Strube, "Linear Prediction on a Warped Frequency Scale," *Journal of the Acoustical Society of America*, Vol. 68, No. 4, pp. 1071–1076 (October 1980).

[47] Imai, Sumita, and Furuichi, "Mel Log Spectrum Approximation (MLSA) Filter for Speech Synthesis," *Transactions of the IECE of Japan*, Vol. J66-A, No. 2, pp. 122–129 (1983), in Japanese.

[48] Sadaoki Furui, *Digital Speech Processing*, Tokai University Press, p. 130 (1985), in Japanese.