

TITLE OF THE INVENTION

Method and Apparatus for Reproducing Speech Signals, Method and Apparatus for Decoding the Speech, Method and Apparatus for Synthesizing the Speech and Portable Radio Terminal Apparatus

BACKGROUND OF THE INVENTION

Field of the Invention

This invention relates to a method and apparatus for reproducing speech signals at a controlled speed, and to a method and apparatus for decoding the speech and a method and apparatus for synthesizing the speech whereby pitch conversion can be realized by a simplified structure. The present invention also relates to a portable radio terminal device for transmitting and receiving pitch-converted speech signals.

Description of the Related Art

There have hitherto been known a variety of encoding methods for encoding an audio signal (inclusive of speech and acoustic signals) for compression by exploiting statistic properties of the signals in the time domain and in the frequency domain and psychoacoustic characteristics of the human ear. The encoding method may roughly be classified into time-domain encoding, frequency domain encoding and analysis/synthesis encoding.

Examples of the high-efficiency encoding of speech signals include sinusoidal analysis encoding, such as harmonic encoding, multi-band excitation (MBE) encoding, sub-band coding (SBC), linear predictive coding (LPC), discrete cosine transform (DCT),

modified DCT (MDCT) and fast Fourier transform (FFT).

Meanwhile, the high-efficiency speech encoding method by the time-axis processing, as typified by code excited linear prediction (CELP) encoding, involves difficulties in expeditious time-axis conversion (modification) because of the necessity of performing voluminous processing operations subsequent to decoder outputting. Moreover, since speed control is performed in the time domain subsequent to decoding, the method cannot be used for bit rate conversion.

On the other hand, if it is attempted to decode speech signals encoded by the above encoding methods, it is frequently desired to vary only the pitch without changing the phoneme of the speech. However, with the usual speech decoding method, the decoded speech has to be pitch-converted using pitch control, thus complicating the structure and raising the cost.

SUMMARY OF THE INVENTION

It is therefore an object of the present invention to provide a method and apparatus for reproducing speech signals whereby speed control to a desired rate over a wide range may be achieved with a high sound quality without changing the phoneme or pitch.

It is another object of the present invention to provide a method and apparatus for decoding the speech and a method and apparatus for synthesizing the speech whereby pitch conversion or pitch control can be achieved by a simplified structure.

It is yet another object of the present invention whereby the

pitch-converted or pitch-controlled speech signals can be transmitted or received by a simplified structure.

With the speech signal reproducing method according to the present invention, the input speech signal is divided on the time axis in terms of pre-set encoding units to produce encoded parameters which are interpolated to produce modified encoded parameters for desired time points, and the speech signal is reproduced based on these modified encoded parameters.

With the speech signal reproducing apparatus according to the present invention, the input speech signal is divided on the time axis in terms of pre-set encoding units to produce encoded parameters which are interpolated to modified encoded parameters for desired time points, and the speech signal is then reproduced based on these modified encoded parameters.

With the speech signal reproducing method, the speech is reproduced with a block length differing from that used for encoding, using encoded parameters obtained on dividing the input speech signal on the time axis in terms of pre-set block as units and encoding the divided speech signal in terms of the encoding blocks.

With the speech decoding method and apparatus according to the present invention, the fundamental frequency and the number in a pre-set band of harmonics of the input encoded speech data are converted and the number of data specifying the amplitude of a spectral component in each input harmonics is interpolated for modifying the pitch.

The pitch frequency is modified at the time of encoding by dimensional conversion in which the number of harmonics is set at a pre-set value.

In this case, the decoder for speech compression may be used simultaneously as a speech synthesizer for text speech synthesis. For routine speech pronunciation, clear playback speech is obtained by compression and expansion, whereas, for special speech synthesis, text synthesis or synthesis under the pre-determined rule is used for constituting an efficient speech output system.

With the speech signal reproducing method and apparatus according to the present invention, an input speech signal is divided in terms of pre-set encoding units on the time axis and encoded in terms of the encoding unit in order to find encoded parameters which are then interpolated to find modified encoded parameters for desired time points. The speech signal is then reproduced based on the modified encoded parameters, so that speed control over a wide range may be realized easily with high quality without changing the phoneme or pitch.

With the speech signal reproducing method and apparatus according to the present invention, the speech is reproduced with a block length differing from that used for encoding, using encoded parameters obtained on dividing the input speech signal on the time axis in terms of pre-set block as units and on encoding the divided speech signal in terms of the encoding blocks. The result is that speed control over a wide range may be realized easily with high

quality without changing the phoneme or pitch.

With the speech decoding method and apparatus according to the present invention, the fundamental frequency and the number in a pre-set band of harmonics of the input encoded speech data are converted and the number of data specifying the amplitude of a spectral component in each input harmonics is interpolated for modifying the pitch. The result is that the pitch may be changed to a desired value by a simplified structure.

In this case, the decoder for speech compression may be used simultaneously as the speech synthesizer for text speech synthesis. For routine speech pronunciation, clear playback speech is obtained by compression and expansion, whereas, for special speech synthesis, text synthesis or synthesis under rule is used for constituting an efficient speech output system.

With the portable radio terminal apparatus, the pitch-converted to pitch-controlled speech signals can be transmitted or received by a simplified structure.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig.1 is a block diagram showing a basic structure of a speech signal reproducing method and a speech signal reproducing apparatus for carrying out the speech signal reproducing method according to the present invention.

Fig.2 is a schematic block diagram showing an encoding unit of the speech signal reproducing apparatus shown in Fig.1.

Fig.3 is a block diagram showing a detailed structure of the

encoding unit.

Fig.4 is a schematic block diagram showing the structure of a decoding unit of the speech signal reproducing apparatus shown in Fig.1.

Fig.5 is a block diagram showing a detailed structure of the decoding unit.

Fig.6 is a flowchart for illustrating the operation of a unit for calculating modified encoding parameters of the decoding unit.

Fig.7 schematically illustrates the modified encoding parameters obtained by the modified encoding parameter calculating unit on the time axis.

Fig.8 is a flowchart for illustrating the detailed interpolation operation performed by the modified encoding parameter calculating unit.

Figs.9A to 9D illustrates the interpolation operation.

Figs.10A to 10C illustrate typical operations performed by the unit for calculating modified encoding parameters.

Figs.11A to 11C illustrate other typical operations performed by the unit for calculating modified encoding parameters.

Fig.12 illustrates an operation in case the frame length is rendered variable to control the speed quickly by the decoding unit.

Fig.13 illustrates an operation in case the frame length is rendered variable to control the speed slowly by the decoding unit.

Fig.14 is a block diagram showing another detailed structure of the decoding unit.

Fig.15 is a block diagram showing an example of application to a speech synthesis device.

Fig.16 is a block diagram showing an example of application to a text speech synthesis device.

Fig.17 is a block diagram showing the structure of a transmitter of a portable terminal employing the encoding unit.

Fig.18 is a block diagram showing the structure of a receiver of a portable terminal employing the decoding unit.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

Referring to the drawings, the speech signal reproducing method and apparatus according to a preferred embodiment of the present invention will be explained. The present embodiment is directed to a speech signal reproducing apparatus 1 for reproducing speech signals based on encoding parameters as found by dividing the input speech signals on the time axis in terms of a pre-set number of frames as encoding units and encoding the divided input speech signals, as shown in Fig.1.

The speech signal reproducing apparatus 1 includes an encoding unit 2 for encoding the speech signals entering an input terminal 101 in terms of frames as units for outputting encoded parameters such as linear prediction encoding (LPC) parameters, line spectrum pair (LSP) parameters, pitch, voiced (V)/unvoiced (UV) or spectral amplitudes A_m , and a period modification unit 3 for modifying an output period of the encoding parameters by time axis compansion. The speech signal reproducing apparatus also includes

a decoding unit 4 for interpolating the encoded parameters outputted at the period modified by the period modification unit 3 for finding the modified encoded parameters for desired time points and for synthesizing the speech signals based on the modified encoded parameters for outputting the synthesized speech signals at an output terminal 201.

The encoding unit 2 is explained by referring to Figs.2 and 3. The encoding unit 2 decides, based on the results of discrimination, whether the input speech signal is voiced or unvoiced, and performs sinusoidal synthetic encoding for a signal portion found to be voiced, while performing vector quantization by a closed-loop search of the optimum vector using an analysis-by-synthesis method for a signal portion found to be unvoiced, for finding the encoded parameters. That is, the encoding unit 2 includes a first encoding unit 110 for finding short-term prediction residuals of the input speech signal, such as linear prediction coding (LPC) residuals, to perform sinusoidal analysis encoding, such as harmonic encoding, and a second encoding unit 120 for performing waveform coding by transmitting phase components of the input speech signal. The first encoding unit 110 and the second encoding unit 120 are used for encoding the voiced (V) portion and the unvoiced (UV) portion, respectively.

In the embodiment of Fig.2, the speech signal supplied to the input terminal 101 is sent to an inverted LPC filter 111 and an LPC analysis quantization unit 113 of the first encoding unit 110. The

LPC coefficient obtained from the LPC analysis/quantization unit 113 or the so-called α -parameter is sent to the inverted LPC filter 111 for taking out the linear prediction residuals (LPC residuals) of the input speech signal by the inverse LPC filter 111. From the LPC analysis/quantization unit 113, a quantized output of the linear spectral pairs (LSP) is taken out as later explained and sent to an output terminal 102. The LPC residuals from the inverted LPC filter 111 are sent to a sinusoidal analysis encoding unit 114. The sinusoidal analysis encoding unit 114 performs pitch detection, spectral envelope amplitude calculations and V/UV discrimination by a voiced (V)/ unvoiced (UV) discrimination unit 115. The spectral envelope amplitude data from the sinusoidal analysis encoding unit 114 are sent to the vector quantization unit 116. The codebook index from the vector quantization unit 116, as a vector-quantized output of the spectral envelope, is sent via a switch 117 to an output terminal 103, while an output of the sinusoidal analysis encoding unit 114 is sent via a switch 118 to an output terminal 104. The V/UV discrimination output from the V/UV discrimination unit 115 is sent to an output terminal 105 and to the switches 117, 118 as switching control signals. For the voiced (V) signal, the index and the pitch are selected so as to be taken out at the output terminals 103, 104. For vector quantization at the vector quantizer 116, a suitable number of dummy data for interpolating amplitude data of an effective band block on the frequency axis from the last amplitude data in the block as far as the first

amplitude data in the block, or dummy data extending the last data and the first data in the block, are appended to the trailing end and to the leading end of the block, for enhancing the number of data to N_F . Then, an O_s -tuple number of amplitude data are found by band-limiting type O_s -tuple oversampling, such as octatuple oversampling. The O_s -tuple number of the amplitude data $((mMx + 1) \times O_s \text{ number of data})$ is further expanded to a larger number of N_M , such as 21048, by linear interpolation. This N_M number data is converted into the pre-set number M (such as 44) by decimation and vector quantization is then performed on the pre-set number of data.

In the present embodiment, the second encoding unit 120 has a code excited linear predictive (CELP) coding configuration and performs vector quantization on the time-domain waveform by a closed-loop search employing an analysis-by-synthesis method. Specifically, an output of a noise codebook 121 is synthesized by a weighted synthesis filter 122 to produce a weighted synthesized speech which is sent to a subtractor 123 where an error between the weighted synthesized speech and the speech supplied to the input terminal 101 and subsequently processed by a perceptually weighting filter 125 is found. A distance calculation circuit 124 calculates the distance and a vector which minimizes the error is searched in the noise codebook 121. This CELP encoding is used for encoding the unvoiced portion as described above. The codebook index as the UV data from the noise codebook 121 is taken out at an output terminal 107 via a switch 127 which is turned on when the results

of V/UV discrimination from the V/UV discrimination unit 115 indicates an unvoiced (UV) sound.

Referring to Fig.3, a more detailed structure of a speech signal encoder shown in Fig.1 is now explained. In Fig.3, the parts or components similar to those shown in Fig.1 are denoted by the same reference numerals.

In the speech signal encoder 2 shown in Fig.3, the speech signals supplied to the input terminal 101 are filtered by a high-pass filter 109 for removing signals of an unneeded range and thence supplied to an LPC analysis circuit 132 of the LPC analysis/quantization unit 113 and to the inverse LPC filter 111.

The LPC analysis circuit 132 of the LPC analysis/ quantization unit 113 applies a Hamming window, with a length of the input signal waveform on the order of 256 samples as a block, and finds linear prediction coefficients, that is so-called α -parameters, by the self-correlation method. The framing interval as a data outputting unit is set to approximately 160 samples. If the sampling frequency f_s is 8 kHz, for example, a one-frame interval is 20 msec or 160 samples.

The α -parameters from the LPC analysis circuit 132 are sent to an α -LSP conversion circuit 133 for conversion into line spectra pair (LSP) parameters. This converts the α -parameters, as found as direct type filter coefficients, into for example, ten, that is five pairs of the LSP parameters. This conversion is carried out by, for example, the Newton-Rhapson method. The reason the α -parameters

are converted into the LSP parameters is that the LSP parameters are superior in interpolation characteristics to the α -parameters.

The LSP parameters from the α -LSP conversion circuit 133 are matrix- or vector-quantized by the LSP quantizer 134. It is possible to take a frame-to-frame difference prior to vector quantization, or to collect plural frames together in order to perform matrix quantization. In the present case, the LSP parameters, calculated every 20 msec, are vector-quantized, with 20 msec as a frame.

The quantized output of the quantizer 134, that is the index data of the LSP quantization, are taken out to the decoding unit 103 at a terminal 102, while the quantized LSP vector is sent to an LSP interpolation circuit 136.

The LSP interpolation circuit 136 interpolates the LSP vectors, quantized every 20 msec or 40 msec, in order to provide an octatuple rate. That is, the LSP vector is updated every 2.5 msec. The reason is that, if the residual waveform is processed with the analysis/synthesis by the harmonic encoding/decoding method, the envelope of the synthetic waveform presents an extremely smooth waveform, so that, if the LPC coefficients are changed abruptly every 20 msec, a foreign noise is likely to be produced. That is, if the LPC coefficient is changed gradually every 2.5 msec, such foreign noise may be prevented from being produced.

For inverted filtering of the input speech using the interpolated LSP vectors, produced every 2.5 msec, the LSP

parameters are converted by an LSP to α conversion circuit 137 into α -parameters as coefficients of, for example, ten-order direct type filter. An output of the LSP to α conversion circuit 137 is sent to the LPC inverted filter circuit 111 which then performs inverted filtering for producing a smooth output using α -parameters updated every 2.5 msec. An output of the inverted LPC filter 111 is sent to an orthogonal transform circuit 145, such as a DCT circuit, of the sinusoidal analysis encoding unit 114, such as a harmonic encoding circuit.

The α -parameters from the LPC analysis circuit 132 of the LPC analysis/quantization unit 113 are sent to a perceptual weighting filter calculating circuit 139 where data for perceptual weighting is found. These weighting data are sent to the perceptual weighting vector quantizer 116, perceptual weighting filter 125 of the second encoding unit 120 and to the perceptual weighted synthesis filter 122.

The sinusoidal analysis encoding unit 114 of the harmonic encoding circuit analyzes the output of the inverted LPC filter 111 by a method of harmonic encoding. That is, pitch detection, calculations of the amplitudes A_m of the respective harmonics and voiced (V)/ unvoiced (UV) discrimination, are carried out, and the numbers of the amplitudes A_m or the envelopes of the respective harmonics, varied with the pitch, are made constant by dimensional conversion.

In an illustrative example of the sinusoidal analysis encoding

unit 114 shown in Fig.3, commonplace harmonic encoding is used. In particular, in multi-band excitation (MBE) encoding, it is assumed in modelling that voiced portions and unvoiced portions are present in the frequency area or band at the same time point (in the same block or frame). In other harmonic encoding techniques, it is uniquely judged whether the speech in one block or in one frame is voiced or unvoiced. In the following description, a given frame is judged to be UV if the totality of the band is UV, insofar as the MBE encoding is concerned.

The open-loop pitch search unit 141 and the zero-crossing counter 142 of the sinusoidal analysis encoding unit 114 of Fig.3 is fed with the input speech signal from the input terminal 101 and with the signal from the high-pass filter (HPF) 109, respectively. The orthogonal transform circuit 145 of the sinusoidal analysis encoding unit 114 is supplied with LPC residuals or linear prediction residuals from the inverted LPC filter 111. The open loop pitch search unit 141 takes the LPC residuals of the input signals to perform relatively rough pitch search by open loop. The extracted rough pitch data is sent to a fine pitch search unit 146 by closed loop search as later explained. From the open loop pitch search unit 141, the maximum value of the normalized autocorrelation $r(p)$, obtained by normalizing the maximum value of the self-correlation of the LPC residuals along with the rough pitch data, are taken out along with the rough pitch data so as to be sent to the V/UV discrimination unit 115.

The orthogonal transform circuit 145 performs orthogonal transform, such as discrete Fourier transform (DFT), for converting the LPC residuals on the time axis into spectral amplitude data on the frequency axis. An output of the orthogonal transform circuit 145 is sent to the fine pitch search unit 146 and a spectral evaluation unit 148 for evaluating the spectral amplitude or envelope.

The fine pitch search unit 146 is fed with relatively rough pitch data extracted by the open loop pitch search unit 141 and with frequency-domain data obtained by DFT by the orthogonal transform unit 145. The fine pitch search unit 146 swings the pitch data by $\pm$ several samples, at a rate of 0.2 to 0.5, centered about the rough pitch value data, in order to arrive ultimately at the value of the fine pitch data having an optimum decimal point (floating point). The analysis by synthesis method is used as the fine search technique for selecting a pitch so that the power spectrum will be closest to the power spectrum of the original sound. Pitch data from the closed-loop fine pitch search unit 146 is sent to an output terminal 104 via a switch 118.

In the spectral evaluation unit 148, the amplitude of each harmonics and the spectral envelope as the sum of the harmonics are evaluated based on the spectral amplitude and the pitch as the orthogonal transform output of the LPC residuals and sent to the fine pitch search unit 146, V/UV discrimination unit 115 and to the perceptually weighted vector quantization unit 116.

The V/UV discrimination unit 115 discriminates V/UV of a frame based on an output of the orthogonal transform circuit 145, an optimum pitch from the fine pitch search unit 146, spectral amplitude data from the spectral evaluation unit 148, maximum value of the normalized self-correlation $r(p)$ from the open loop pitch search unit 141 and the zero-crossing count value from the zero-crossing counter 142. In addition, the boundary position of the band-based V/UV discrimination for MBE may also be used as a condition for V/UV discrimination. A discrimination output of the V/UV discrimination unit 115 is taken out at the output terminal 105.

An output unit of the spectrum evaluation unit 148 or an input unit of the vector quantization unit 116 is provided with a number of data conversion unit (a unit performing a sort of sampling rate conversion). The data number conversion unit is used for setting the amplitude data $|Am|$ of an envelope taking into account the fact that the number of bands split on the frequency axis and the number of data differ with the pitch. That is, if the effective band is up to 3400 kHz, the effective band can be split into 8 to 63 bands depending on the pitch. The number of $mMX + 1$ of the amplitude data $|Am|$, obtained from band to band, is changed in a range from 8 to 63. Thus the data number conversion unit 119 converts the amplitude data of the variable number $mMx + 1$ to a pre-set number M of data, such as 44 data.

The amplitude data or envelope data of the pre-set number M ,

such as 44, from the data number conversion unit, provided at an output unit of the spectral evaluation unit 148 or at an input unit of the vector quantization unit 116, are gathered in terms of a pre-set number of data, such as 44 data, as units, and vector-quantized by the vector quantization unit 116. This weight is supplied by an output of the perceptual weighting filter calculation circuit 139. The index of the envelope from the vector quantizer 116 is taken out by a switch 117 at an output terminal 103. Prior to weighted vector quantization, it is advisable to take inter-frame difference using a suitable leakage coefficient for a vector made up of a pre-set number of data.

The second encoding unit 120 is explained. The second encoding unit 120 is of the code excited linear prediction (CELP) coding structure and is used in particular for encoding the unvoiced portion of the input speech signal. In the CELP encoding configuration for the unvoiced speech portion, a noise output corresponding to LPC residuals of an unvoiced speech portion as a representative output of the noise codebook, that is the so-called stochastic codebook 121, is sent via gain circuit 126 to the perceptually weighted synthesis filter 122. The speech signal supplied from the input terminal 101 via high-pass filter (HPF) 109 and perceptually weighted by the perceptually weighting filter 125 is fed to the subtractor 123 where a difference or error of the perceptually weighted speech signal from the signal from the synthesis filter 122 is found. This error is fed to a distance

calculation circuit 124 for finding the distance and a representative value vector which will minimize the error is searched by the noise codebook 121. The above is the summary of the vector quantization of the time-domain waveform employing the closed-loop search in turn employing the analysis by synthesis method.

As data for the unvoiced (UV) portion from the second encoder 120 employing the CELP coding structure, the shape index of the codebook from the noise codebook 121 and the gain index of the codebook from the gain circuit 126 are taken out. The shape index, which is the UV data from the noise codebook 121, is sent via a switch 127s to an output terminal 107s, while the gain index, which is the UV data of the gain circuit 126, is sent via a switch 127g to an output terminal 107g.

These switches 127s, 127g and the switches 117, 118 are turned on and off depending on the results of V/UV decision from the V/UV discrimination unit 115. Specifically, the switches 117, 118 are turned on, if the results of V/UV discrimination of the speech signal of the frame about to be transmitted indicates voiced (V), while the switches 127s, 127g are turned on if the speech signal of the frame about to be transmitted is unvoiced (UV).

The encoded parameters, outputted by the encoding unit 2, are supplied to the period modification unit 3. The period modification unit 3 modifies an output period of the encoded parameters by time axis compression/expansion. The encoded parameters, outputted at a period modified by the period modification unit 3, are sent to the

decoding unit 4.

The decoding unit 4 includes a parameter modification unit 5 for interpolating the encoded parameters, compressed along time axis by the period modification unit 3, by way of an example, for generating modified encoded parameters associated with time points of pre-set frames, and a speech synthesis unit 6 for synthesizing the voiced speech signal portion and the unvoiced speech signal portion based on the modified encoded parameters.

Referring to Figs.4 and 5, the decoding unit 4 is explained. In Fig.4, the codebook index data, as quantized output data of the linear spectrum pairs (LSPs) from the period modification unit 3, are supplied to an input terminal 202. Outputs of the period modification unit 3, that is index data, as quantized envelope data, pitch data and V/UV discrimination output data, are supplied to input terminals 203, 204 and 205, respectively. Index data from the period modification unit 3, as data for an unvoiced speech portion, is also supplied to an input terminal 207.

The index data from the input terminal 203, as the quantized envelope output, is sent to an inverse vector quantizer 212 for vector quantization to find a spectral envelope of the LPC residuals. Before being sent to a voiced speech synthesis unit 211, the spectral envelope of the LPC residuals is transiently taken out at near a point indicated by arrow P₁ in Fig.4 by the parameter processor 5 for parameter modification as will be explained subsequently. The index data is then sent to the voiced speech

synthesis unit 211.

The voiced speech synthesis unit 211 synthesizes the LPC residuals of the voiced speech signal portion by sinusoidal synthesis. The pitch and the V/UV discrimination data, entering the input terminals 204, 205, respectively and transiently taken out at points P_2 and P_3 in Fig.4 by the parameter modification unit 5 for parameter modification, are similarly supplied to the synthesis speech synthesis unit 211. The LPC residuals of the voiced speech from the voiced speech synthesis unit 211 are sent to an LPC synthesis filter 214.

The index data of the UV data from the input terminal 207 is sent to an unvoiced speech synthesis unit 220. The index data of the UV data is turned into LPC residuals of the unvoiced speech portion by the unvoiced speech synthesis unit 220 by having reference to the noise codebook. The index data of the UV data are transiently taken out from the unvoiced speech synthesis unit 220 by the parameter modification unit 5 as indicated at P_4 in Fig.4 for parameter modification. The LPC residuals, thus processed with parameter modification, are also sent to the LPC synthesis filter 214.

The LPC synthesis filter 214 performs independent LPC synthesis on the LPC residuals of the voiced speech signal portion and on the LPC residuals of the unvoiced speech signal portion. Alternatively, the LPC synthesis may be performed on the LPC residuals of the voiced speech signal portion and the LPC residuals

of the unvoiced speech signal portion summed together.

The LSP index data from the input terminal 202 are sent to an LPC parameter regenerating unit 213. Although the α -parameters of the LPC are ultimately produced by the LPC parameter regenerating unit 213, the inverse vector quantized data of the LSP are taken out partway by the parameter modification unit 5 as indicated by arrow P, for parameter modification.

The dequantized data, thus processed with parameter modification, is returned to this LPC parameter regenerating unit 213 for LPC interpolation. The dequantized data is then turned into α -parameters of the LPC which are supplied to the LPC synthesis filter 14. The speech signals, obtained by LPC synthesis by the LPC synthesis filter 214, are taken out at the output terminal 201. The speech synthesis unit 6, shown in Fig.4, receives the modified encoded parameters, calculated by the parameter modification unit 5 as described above, and outputs the synthesized speech. The actual configuration of the speech synthesis unit is as shown in Fig.5, in which parts or components corresponding to those shown in Fig.4 are depicted by the same numerals.

Referring to Fig.5, the LSP index data, entering the input terminal 202, is sent to an inverse vector quantizer 231 for LSPs in the LPC parameter regenerating unit 213 so as to be inverse vector quantized into LSPs (line spectrum pairs) which are supplied to the parameter modification unit 5.

The vector-quantized index data of the spectral envelope A_m

from the input terminal is sent to the inverse vector quantizer 212 for inverse vector quantization and turned into data of the spectral envelope which is sent to the parameter modification unit 5.

The pitch data and the V/UV discrimination data from the input terminals 204, 205 are also sent to the parameter modification unit 5.

To input terminals 207s and 207g of Fig.5 are supplied shape index data and gain index data as UV data from output terminals 107s and 107g of Fig.3 via period modification unit 3. The shape index data and the gain index data are thence supplied to the unvoiced speech synthesis unit 220. The shape index data from the terminal 207s and the gain index data from the terminal 207g are supplied to a noise codebook 221 and to a gain circuit 222 of the unvoiced speech synthesis unit 220, respectively. A representative value output read out from the noise codebook 221 is the noise signal component corresponding to the LPC residuals of the unvoiced speech and becomes an amplitude of a pre-set gain in the gain circuit 22. The resulting signal is supplied to the parameter modification unit 5.

The parameter modification unit 5 interpolates the encoded parameters, outputted by the encoding unit 2 and having an output period modified by the period modification unit 3, for generating modified encoded parameters, which are supplied to the speech synthesis unit 6. The parameter modification unit 3 speed-modifies the encoded parameters. This eliminates the operation of speed

modification after decoder outputting and enables the speech signals reproducing apparatus 1 to deal with fixed rates different with similar algorithms.

Referring to the flowcharts of Figs.6 and 8, the operation of the period modification unit 3 and the parameter modification unit 5 is explained.

At step S1 of Fig.6, the period modification unit 3 receives encoded parameters, such as LSPs, pitch, voiced/unvoiced (V/UV), spectral envelope Am or LPC residuals. The LSPs, pitch, V/UV, Am and the LPC residuals are represented as $l_{sp}[n][p]$, $P_{ch}[n]$, $vu_v[n]$, $a_m[n][k]$ and $r_{ch}[n][i][j]$, respectively.

The modified encoded parameters, ultimately calculated by the parameter modification unit 5, are represented as $mod_l_{sp}[m][p]$, $mod_p_{ch}[m]$, $mod_vu_v[m]$, $mod_a_m[m][k]$ and $mod_r_{ch}[m][i][j]$, where k and p denote the number of harmonics and the number of LSP orders, respectively. Each of n and m denotes frame numbers corresponding to time-domain index data prior and subsequent to time axis conversion, respectively. Meanwhile, each of n and m denotes an index of a frame having an interval of 20 msec, while i and j denote a sub-frame number and a sample number, respectively.

The period modification unit 3 then sets the number of frames representing the original time duration to and the number of frames representing the time duration after modification to N_1 , N_2 , respectively, as shown at step S2. The period modification unit then proceeds to time-axis compression of the speech N_1 to the speech

N_2 as shown at step S3. That is, the time-axis compression ratio spd at the period modification unit 3 is found as $spd = N_2/N_1$, on the proviso that $0 \leq n < N_1$ and $0 \leq m < N_2$.

The parameter modification unit 5 then sets m , corresponding to the frame number corresponding in turn to the index of the time axis after time axis modification, to 2.

The parameter modification unit 5 then finds two frames f_{r0} and f_{r1} and the differences *left* and *right* between the two frames f_{r0} and f_{r1} and the ratio m/spd .

If the parameters l_{sp} , p_{ch} , vu_w , a_m and r_{cs} are denoted as $*$, mod $*[m]$ may be represented by the general formula

$$mod_*[m] = *[m/spd]$$

where $0 \leq m < N_2$. However, since m/spd is not an integer, the modified encoded parameter at m/spd is produced by interpolation from two frames of

$$f_{r0} = \lfloor m/spd \rfloor$$

and

$$f_{r1} = f_{r0} + 1.$$

Between the frame f_{r0} , m/spd and the frame f_{r1} , the relation shown in Fig.7, namely

$$left = m/spd - f_{r0}$$

$$right = f_{r1} - m/spd$$

holds.

The encoded parameters for m/spd in Fig.7, namely the modified encoded parameters, may be found by interpolation as

shown at step S6.

The modified encoded parameter is simply found by linear interpolation by:

$$\text{mod_}[m] = *[f_{r0}] \times \text{right} + *[f_{r1}] \times \text{left}$$

However, with interpolation between the two frames f_{r0} and f_{r1} , the above general formula cannot be used if the two frames are different as to V/UV, that is if one of the two frames is V and the other is UV. Therefore, the parameter modification unit 5 changes the method for finding the encoded parameters depending on the voiced (V) or unvoiced (UV) character of the two frames f_{r0} and f_{r1} as indicated by steps S11 ff. of Fig.8.

First, the voiced (V) or unvoiced (UV) character of the two frames f_{r0} and f_{r1} is determined, as shown at step S11. If the two frames f_{r0} and f_{r1} are both found to be voiced (V), processing transfers to step s12 where all parameters are linearly interpolated and represented by:

$$\text{mod_} p_{ch}[m] = p_{ch}[f_{r0}] \times \text{right} + p_{ch}[f_{r1}] \times \text{left}$$

$$\text{mod_} a_m[m][k] = a_m[f_{r0}][k] \times \text{right} + a_m[f_{r1}][k] \times \text{left}$$

where $0 \leq k < L$, where L is the maximum possible number of harmonics. For $a_m[n][k]$, 0 is inserted at such positions where there are no harmonics. If the number of harmonics differs between the frames f_{r0} and f_{r1} , 0s are inserted in vacant positions. Alternatively, a fixed number such as $0 \leq k < L$, where $L = 43$, may be used if prior to passage through a number of data converter on the decoder side.

$$\text{mod_}l_{,p}[m][p] = l_{,p}[f_{,d}][p] \times \text{right} + l_{,p}[f_{,u}][p] \times \text{left}$$

where $0 \leq p < P$, where P denotes the number of orders of the LSPs and is usually 10.

$$\text{mod_}vu_v[m] = 1$$

In V/UV discrimination, 1 and 0 denote voiced (V) and unvoiced (UV), respectively.

If, at step S11, none of the two frames $f_{,0}$ and $f_{,1}$ is judged to be voiced (V), it is judged at step S13 whether both the two frames $f_{,0}$ and $f_{,1}$ are unvoiced (UV). If the result of judgment at step S13 is Yes, that is if the two frames are both unvoiced, the interpolation unit 5 slices 80 samples ahead and at back of $r_{,s}$, with m/spd as center and with $p_{,ch}$ as a maximum value, as indicated at step S14.

In effect, if $\text{left} < \text{right}$ at step S14, 80 samples ahead and at back of $r_{,s}$, centered about m/spd , are sliced, and inserted into $\text{mod } r_{,s}$, as shown in Fig.9A. That is,

for ($j = 0; j < \text{FRM} \times (1/2 - m/\text{spd} + f_{,0}); j^{++}$) {
 $\text{mod } r_{,s}[m][0][j] = r_{,s}[f_{,d}][0][j + (m/\text{spd} - f_{,d}) \times \text{FRM}];$ };

for ($j = \text{FRM} \times (1/2 - m/\text{spd} + f_{,0}); j < \text{FRM}/2; j^{++}$) {
 $\text{mod } r_{,s}[m][0][j] = r_{,s}^{**}[m][0][j] = r_{,s}[f_{,d}][1][j - \text{FRM} \times (1/2 - m/\text{spd} + f_{,0})];$ };

for ($j = 0; j < \text{FRM} \times (1/2 - m/\text{spd} + f_{,0}); j^{++}$) {
 $\text{mod } r_{,s}[m][1][j] = r_{,s}[f_{,d}][1][j + (m/\text{spd} - f_{,d}) \times \text{FRM}];$ };

for ($j = \text{FRM} \times (1/2 - m/\text{spd} + f_{,0}); j = \text{FRM}/2; j^{++}$) {
 $\text{mod } r_{,s}[m][1][j] = r_{,s}[f_{,d}][0][j + \text{FRM} \times (1/2 - m/\text{spd} + f_{,d})];$ };

where FRM is e.g., 160.

On the other hand, if $\text{left} \geq \text{right}$ at this step S14, the interpolation unit 5 slices 80 samples ahead and at back of $r_{e,}$, centered about m/spd , to produce $\text{mod_}r_{e,}$, as shown in Fig.9B.

If the condition of step S13 is not met, processing transfers to step S15 where it is judged whether the frame $f_{r,0}$ is voiced (V) and the frame $f_{r,1}$ is unvoiced (UV). If the result of judgment is YES, that is if the frame $f_{r,0}$ is voiced (V) and the frame $f_{r,1}$ is unvoiced (UV), processing transfers to step S16. If the result of judgment is NO, that is if the frame $f_{r,0}$ is unvoiced (UV) and the frame $f_{r,1}$ is voiced (V), processing transfers to step S17.

In the processing downstream of the step S15 ff., the two frames $f_{r,0}$ and $f_{r,1}$ are different as to V/UV, that is voiced (V) to unvoiced (UV). This takes into account the fact that if parameters are interpolated between two frames $f_{r,0}$ and $f_{r,1}$ which are different as to V/UV, the result of interpolation becomes meaningless.

At step S16, the size of left ($=m/\text{spd} - f_{r,0}$) and that of right ($= f_{r,1} - m/\text{spd}$) are compared to each other, in order judge if the frame $f_{r,0}$ is closer to m/spd .

If the frame $f_{r,0}$ is closer to m/spd , the modified encoded parameters are set, using the parameters of the frame $f_{r,0}$, so that

$$\text{mod_}p_{eh}[m] = p_{eh}[f_{r,0}]$$

$$\text{mod_}a_m[m][k] = a_m[f_{r,0}][k], \text{ where } 0 \leq k < L;$$

$$\text{mod_}l_{sp}[m][p] = l_{sp}[f_{r,0}][p], \text{ where } 0 \leq p < I; \text{ and}$$

$$\text{mod_}vu_v[m] = 1$$

as shown at step S18.

If the result of judgment at step S16 is NO, left $\geq$ right, so that the frame f_{r1} is closer, so that processing transfers to step S19 to maximize the pitch. Also, r_{e1} of the frame f_{r1} is directly used as shown in Fig.9C and set as $\text{mod_}r_{e1}$. That is, $\text{mod_}r_{e1}[m][i][j] = r_{e1}[f_{r1}][i][j]$. The reason is that, for voiced frame f_{r0} the LPC residuals r_{e1} are not transmitted.

At step S17, judgment similar to that at step S16 is given on the basis of judgment given at step S15 that the two frames f_{r0} and f_{r1} are unvoiced (UV) and voiced (V), respectively. That is, the sizes of left ($=m/\text{spd} - f_{r0}$) and right ($=f_{r1} - m/\text{spd}$) are compared to each other in order to judge whether or not the frame f_{r0} is closer to m/spd .

If the frame f_{r0} is closer, processing transfers to step S18 to maximize the pitch. Also, r_{e0} of the frame f_{r0} is directly used and set as $\text{mod_}r_{e0}$. That is, $\text{mod_}r_{e0}[m][i][j] = r_{e0}[f_{r0}][i][j]$. The reason is that, for voiced frame f_{r1} , the LPC residuals r_{e0} are not transmitted.

If the result of judgment at step S17 is NO, left $\geq$ right and hence the frame f_{r0} is closer to m/spd , so that processing proceeds to step S21 and the modified encoded parameters are set, using the parameters of the frame f_{r1} , so that

$$\text{mod_}p_{ch}[m] = p_{ch}[f_{r1}]$$

$$\text{mod_}a_m[m][k] = a_m[f_{r1}][k], \text{ where } 0 \leq k < L;$$

$$\text{mod_}l_{sp}[m][p] = l_{sp}[f_{r1}][p], \text{ where } 0 \leq p < I; \text{ and}$$

$$\text{mod_}vu_u[m] = 1$$

In this manner, the interpolation unit 5 provides different operations for the interpolation of step S6 of Fig.6 shown in detail in Fig.8, depending on the V/UV character of the two frames $f_{r,0}$ and $f_{r,1}$. After the end of the interpolation at step S6, processing transfers to step S6 for incrementing the value of m . The operations of steps S5 and S6 are repeated until the value of m becomes equal to N_2 .

The operations of the period modification unit 3 and the parameter modification unit 5 are explained collectively by referring to Fig.10. Referring to Fig.10A, the period of the encoding parameters, extracted every 20 msec of a period by the encoding unit 2, is modified by the period modification unit 5 by time axis compression to 15 msec, as shown in Fig.10A. By the interpolation operation, responsive to the state of V/UV of the two frames $f_{r,0}$ and $f_{r,1}$, the parameter modification unit 5 calculates the modified encoded parameters every 20 msec, as shown in Fig.10C.

The operations by the period modification unit 3 and the parameter modification unit 5 may be reversed in sequence, that is, the encoded parameters shown in Fig.11A are first interpolated as shown in Fig.11B and subsequently compressed as shown in Fig.11C for calculating the modified encoded parameters.

Returning to Fig.5, the modified encoded parameters $mod\ l_p[m][p]$ on the LSP data, calculated by the parameter calculation unit 5, are sent to LSP interpolation circuits 232_v, 232_u for LSP interpolation. The resulting data is converted by LSP to α

converting circuits 234_v , 234_{uv} for conversion into an α -parameter for linear predictive coding (LPC) which is sent to the LPC synthesis filter 214. The LSP interpolation circuit 232_v and the LSP to α converting circuit 234_v are used for the voiced (V) signal portion, while the LSP interpolation circuit 232_u and the LSP to α converting circuit 234_u are used for the unvoiced (UV) signal portion. The LPC synthesis filter 214 is made up of an LPC synthesis filter 236 for the voiced portion and an LPC synthesis filter 237 for the unvoiced portion. That is, the LPC coefficient interpolation is performed independently for the voiced portion and the unvoiced portion for preventing ill effects otherwise produced by interpolation of LSPs of totally different character at a transient region from the voiced portion to the unvoiced portion or at a transient region from the voiced portion to the unvoiced portion.

The modified encoded parameter on the spectral envelope data $\text{mod_}a_m[m][k]$, as found by the parameter modification unit 5, is sent to a sinusoidal synthesis circuit 215 of the voiced speech synthesis unit 211. This voiced speech synthesis unit 211 is also fed with the modified encoded parameter on the pitch $\text{mod_}p_{ch}[m]$ and the modified encoded parameter $\text{mod_}vu_v[m]$ on the V/UV decision data, as calculated by the parameter modification unit 5. From the sinusoidal synthesis circuit 215, the LPC residual data corresponding to the output of the LPC inverted filter 111 of Fig.3 are taken out and sent to an adder 218.

The modified encoded parameter on the spectral envelope data

$\text{mod_a}_m[m][k]$, modified encoded parameter on the pitch $\text{mod_p}_{ch}[m]$ and the modified encoded parameter on the V/UV decision data $\text{mod_vu}_u[m]$, as found by the parameter modification unit 5, are sent to a noise synthesis circuit 216 for noise addition for the voiced (V) portion. An output of the noise synthesis circuit 216 is sent to an adder 218 via a weighted overlap-and-add circuit 217. Specifically, the noise taking into account the parameters derived from the encoded speech data, such as pitch spectral envelope amplitudes, maximum amplitude in the frame or residual signal level, is added to the voiced portion of the LPC residual signal of the LPC synthesis filter input, that is excitation, in consideration that, if the input to the LPC synthesis filter of the voiced speech, that is excitation, is produced by sinusoidal synthesis, "stuffed" feeling is produced in the low-pitch sound, such as male speech, while the sound quality is rapidly changed between the V and UV speech portions, thus producing a non-spontaneous feeling.

A sum output of the adder 218 is sent to the synthesis filter 236 for the voiced speech where the time waveform data is produced by LPC synthesis. In addition, resulting time waveform data is filtered by a post-filter 238v and thence supplied to an adder 239.

It is noted that the LPC synthesis filter 214 is separated into the synthesis filter for V 236 and the synthesis filter for UV 237, as explained previously. If the synthesis filter is not separated in this manner, that is if the LSPs are interpolated continuously every

20 samples or every 2.5 msec without making distinction between the V and UV signal portions, the LSPs of totally different character are interpolated at the U to UV to UV to V transient portions, thus producing foreign sound. For preventing such ill effects, the LPC synthesis filter is separated into the filter for V and the filter for UV for interpolating the LPC coefficients independently for V and UV.

The modified encoded parameters on the LPC residuals $mod_{r_{a,m}[i][j]}$, as calculated by the parameter modification unit 5, are sent to the windowing circuit 223 for windowing for smoothing the junction portions with the voiced speech portion.

An output of the windowing circuit 223 is sent to the synthesis filter 237 for UV of the LPC synthesis filter 214 as an output of the unvoiced speech synthesis unit 220. The synthesis filter 237 performs LPC synthesis on the data to provide time waveform data for the unvoiced portion which is filtered by a post-filter for unvoiced speech 238u and thence supplied to an adder 239.

The adder 239 adds the time waveform signal of the voiced portion from the post-filter 238v for voiced speech to the time waveform data for the unvoiced speech portion from the post-filter for the unvoiced speech portion 238u and outputs the resulting data at an output terminal 201.

With the present speech signal reproducing apparatus 1, an array of modified encoded parameters $mod_*[m]$, where $0 \leq m < N_2$

is decoded in this manner instead of the inherent array $*[n]$, where $0 \leq n < N_1$. The frame interval during decoding may be fixed such as at 20 msec as conventionally. In such case, time axis compression and resulting speed-up of the reproducing rate may be realized for $N_2 < N_1$, while time axis expansion and resulting speed-down of the reproducing rate may be realized for $N_2 > N_1$.

With the present system, the ultimately obtained parameter string is arrayed in an inherent spacing of 20 msec for decoding, so that optional speed-up may be realized easily. Moreover, speed-up and speed-down may be realized without any distinction by the same processing operation.

Consequently, the contents of solid-state recording can be reproduced at a speed twice the real-time speed. Since the pitch and the phoneme remain unchanged despite increased playback speed, the recording contents can be discerned despite reproduction at a significantly increased playback speed.

If $N_2 < N_1$, that is if the playback speed is lowered, the playback sound tends to become non-spontaneous since plural parameters $\text{mod_}r_{e,i}$ are produced from the same LPC residuals $r_{e,i}$ in the case of the unvoiced frame. In such case, an appropriate amount of noise may be added to the parameters $\text{mod_}r_{e,i}$ for eliminating such non-spontaneousness to some extent. Instead of adding the noise, the parameters $\text{mod_}r_{e,i}$ may be replaced by suitably generated Gaussian noise, or the excitation vector, randomly selected from the codebook, may also be employed.

With the above-described speech signal reproducing apparatus 1, the time axis of the output period of the encoded parameters from the encoding unit 2 is compressed by the period modification unit 3 for speed-up of the reproducing speed. However, the frame length may be rendered variable by the decoding unit 4 for controlling the reproducing speed.

In such case, since the frame length is rendered variable, the frame number n is not changed before and after parameter generation by the parameter modification unit 5 of the decoding unit 4.

Also, the parameter modification unit 5 modifies the parameters $l_p[n][p]$ and $vu_v[n]$ to $mod_l_p[n][p]$ and to $mod_vu_v[n]$, respectively, regardless of whether the frame in subject is voiced or unvoiced.

If $mod_vu_v[n]$ is 1, that is if the frame in subject is voiced (V), the parameters $p_{ch}[n]$ and $a_m[n][k]$ are modified to $mod_p_{ch}[n]$ and to $mod_a_m[n][k]$, respectively.

If $mod_vu_v[n]$ is 0, that is if the frame in subject is unvoiced (UV), the parameter $r_{ss}[n][i][j]$ is modified to $mod_r_{ss}[n][i][j]$.

The parameter modification unit 5 directly modifies $l_p[n][p]$, $p_{ch}[n]$, $vu_v[n]$ and $a_m[n][k]$ directly to $mod_l_p[n][p]$, $p_{ch}[n]$, $mod_vu_v[n]$ and to $mod_a_m[n][k]$. However, the parameter modification unit varies the residual signal $mod_r_{ss}[n][i][j]$ depending on the speed spd .

If the speed $spd < 1.0$, that is if the speed is faster, the

residual signals of the original signal are sliced at a mid portion, as shown in Fig.12. If the original frame length is orgFrmL , $(\text{orgFrmL} - \text{frmL})/2 \leq j \leq (\text{orgFrmL} + \text{frmL})/2$ is sliced from the original frame $r_{\text{e}}[n][i]$ to give $\text{mod_}r_{\text{e}}[n][i]$. It is also possible that slicing be made at the leading end of the original frame.

If the speed $\text{spd} > 1.0$, that is if the speed is slower, the original frame is used and an original frame added to with noise components is used for any deficit portion. A decoded excitation vector added to with a suitably generated noise may also be used. The Gaussian noise may be generated and used as the excitation vector for reducing the alien feeling produced by continuation of frames of the same waveform. The above noise components may also be added to both ends of the original frame.

Thus, in the case of the speech signal reproducing apparatus 1 configured for changing the speed control by varying the frame length, the speech synthesis unit 6 is constructed and designed so that the LSP interpolation unit 232v and 232u, sinusoidal synthesis unit 215 and the windowing unit 223 will perform different operations for controlling the speed by time axis compansion.

The LSP interpolation unit 232_v finds the smallest integer p satisfying the relation $\text{frmL}/p \leq 20$ if the frame in subject is voiced (V). The LSP interpolation unit 232_u finds the smallest integer p satisfying the relation $\text{frmL}/p \leq 80$ if the frame in subject is unvoiced (UV). The range of the sub-frame $\text{subl}[i][j]$ for LSP interpolation is determined by the following equation:

$\text{nint}(\text{frmL}/p \times i) \leq j \leq \text{nint}(\text{frmL}/p \times (i+1))$, where $0 \leq i \leq p-1$.

In the above equation, $\text{nint}(x)$ is a function which returns an integer closest to x by rounding the first sub-decimal order. For both the voiced and unvoiced sounds, $p = 1$ if frmL is less than 20 or 80.

For example, for the i 'th sub-frame, since the center of the sub-frame is $\text{frmL} \times (2i + 1)/2p$, LSPs are interpolated at a rate of $\text{frmL} \times (2p - 2i - 1)/(20:\text{frmL} \times (2i + 1)/2p)$, as disclosed in our co-pending JP Patent Application No.6-198451.

Alternatively, the number of the sub-frames may be fixed and the LSPs of each sub-frame may be interpolated at all times at the same ratio. The sinusoidal synthesis unit 223 modifies the window length for matching to the frame length frmL .

With the above-described speech signal reproducing apparatus 1, the encoded parameters, the output period of which has been companded on the time axis, are modified using the period modification unit 3 and the parameter modification unit 5 for varying the reproducing speed without changing the pitch or phoneme. However, it is also possible to omit the period modification unit 3 and to process the encoded data from the encoding unit 2 by a number of data conversion unit 270 of the decoding unit 8 shown in Fig.14 for varying the pitch without varying the phoneme. In Fig.14, the parts and components corresponding to those shown in Fig.4 are indicated by the same reference numerals.

The basic concept underlying the decoding unit 8 is to convert the basic frequency of the harmonics of the encoded speech data entered from the encoding unit 2 and the number of amplitude data in a pre-set band by the number of data conversion unit 270 operating as data conversion means to perform only the pitch without changing the phoneme. The number of data conversion unit 270 varies the pitch by modifying the number of data specifying the size of spectral components in each input harmonics.

Referring to Fig.14, a vector quantized output of LSPs, corresponding to an output of the output terminal 102 of Figs.2 and 3, or codebook indices, are supplied to the input terminal 202.

The LSP index data is sent to an inverse vector quantizer 231 of the LPC parameter reproducing unit 213 for inverse vector quantization to line spectrum pairs (LSPs). The LSPs are sent to LSP interpolation circuits 232, 233 for interpolation and thence supplied to LSP to α conversion circuits 234, 235 for conversion to α -parameters of the linear prediction codes. These α -parameters are sent to the LPC synthesis filter 214. The LSP interpolation circuit 232 and the LSP to α converting circuit 234 are used for the voiced (V) signal portion, while the LSP interpolation circuit 233 and the LSP to α converting circuit 235 are used for the unvoiced (UV) signal portion. The LPC synthesis filter 214 is made up of an LPC synthesis filter 236 for the voiced portion and an LPC synthesis filter 237 for the unvoiced portion. That is, the LPC coefficient interpolation is performed independently for the voiced portion and

the unvoiced portion for preventing ill effects otherwise produced by interpolation of LSPs of totally different character at a transient region from the voiced portion to the unvoiced portion or at a transient region from the voiced portion to the unvoiced portion.

To an input terminal 203 of Fig.14 is supplied weighted vector quantized code index data of the spectral envelope A_m corresponding to an output of the terminal 103 of the encoder shown in Figs.2 and 3. To an input terminal 205 is supplied V/UV decision data from the terminal 105 of Figs.2 and 3.

The vector quantized index data of the spectral envelope A_m from the input terminal 203 is sent to the inverse vector quantizer 212 for inverse vector quantization. The number of amplitude data of the inverse vector quantized envelope is fixed at a pre-set value of, for example, 44. Basically, the number of data is converted to give the number of harmonics corresponding to the pitch data. If it is desired to change the pitch, as in the present embodiment, the envelope data from the inverse vector quantizer 212 is sent to the number of data conversion unit 270 for varying the number of amplitude data by, for example, interpolation, depending on the desired pitch value.

The number of data conversion unit 270 is also fed with pitch data from the input terminal 204 such that the pitch at the encoding time is changed to a desired pitch which is outputted. The amplitude data and the modified pitch data are sent to the sinusoidal synthesis circuit 215 of the voiced speech synthesis unit

211. The number of the amplitude data supplied to the synthesis circuit 215 corresponds the modified pitch of the spectral envelope of the LPC residuals from the number of data conversion unit 270.

There are a variety of interpolation methods for converting the number of amplitude data of the spectral envelope of the LPC residuals by the number of data conversion unit 270. For example, a suitable number of dummy data for interpolating amplitude data of an effective band block on the frequency axis from the last amplitude data in the block as far as the first amplitude data in the block or dummy data extending the left-hand end (first data) and the right-hand end (last data) in the block, are appended to the amplitude data in the block, for enhancing the number of data to N_F . Then, an O_s -tuple number of amplitude data are found by band-limiting type O_s -tuple oversampling, such as octatuple oversampling. The O_s -tuple number of the amplitude data $((mMx + 1) \times O_s \text{ number of data})$ is further expanded to a larger number of N_M , such as 2048, by linear interpolation. This N_M number data is converted into the pre-set number M (such as 44) by decimation and vector quantization is then carried out on the pre-set number of data.

As an illustrative operation in the number of data conversion unit 270, the case in which the frequency $F_0 = f/L$ for a pitch lag L to F_x , where f_s is a sampling frequency such that $f_s = 8 \text{ kHz} = 8000 \text{ Hz}$, is explained.

In this case, the pitch frequency $F_0 = 8000/L$, while there are

$n = L/2$ harmonics set up to 4000 Hz. In the usual speech range of 3400 Hz, the number of harmonics is $(L/2) \times (3400/4000)$. This is converted by the above data number conversion or dimensional conversion to, for example, 44, before proceeding to vector quantization. There is no necessity of performing quantization if simply the pitch is to be varied.

After inverse vector quantization, the number of 44 of the harmonics can be changed to a desired number, that is to a desired pitch frequency F_x , by dimensional conversion by the number of data conversion unit 270. The pitch lag L_x corresponding to the pitch frequency $F_x(\text{Hz})$ is $L_x = 8000/F_x$, such that the number of harmonics set up to 3400 Hz is $(L_x/2) \times (3400/4000) = (4000/F_x) \times (3400/4000) = 3400/F_x$

that is $3400/F_x$. That is, it suffices to perform conversion from 44 to $3400/F_x$ by dimensional conversion or number of data conversion in the number of data conversion unit 270.

If the frame-to-frame difference is found at the time of encoding prior to vector quantization of spectral data, the frame-to-frame difference is decoded after the inverse vector quantization. The number of data conversion is then performed for producing spectral envelope data.

The sinusoidal synthesis circuit 215 is supplied not only with pitch data and spectral envelope amplitude data of LPC residuals from the number of data conversion unit 270, but also with the V/UV decision data from the input terminal 205. From the

sinusoidal synthesis circuit 215, the LPC residual data are taken out and sent to the adder 218.

The envelope data from the inverse vector quantizer 212, the pitch data from the input terminal 204 and the V/UV decision data from the input terminal 205 are sent to the noise addition circuit 216 for noise addition for the voiced (V) portion. Specifically, the noise taking into account the parameters derived from the encoded speech data, such as pitch spectral envelope amplitudes, maximum amplitude in the frame or residual signal level, is added to the voiced portion of the LPC residual signal for the LPC synthesis filter input, that is excitation, in consideration that, if the input to the LPC synthesis filter of the voiced speech, that is excitation, is produced by sinusoidal synthesis, "stuffed" feeling is produced in the low-pitch sound, such as male speech, while the sound quality is rapidly changed between the V and UV speech portions thus producing an non-spontaneous feeling.

A sum output of the adder 218 is sent to the synthesis filter 236 for the voiced speech where the time waveform data is produced by LPC synthesis. In addition, resulting time waveform data is filtered by a post-filter 238v for voiced data and thence supplied to an adder 239.

To input terminals 207s and 207g of Fig.14 are supplied shape index data and gain index data as UV data from output terminals 107s and 107g of Fig.3 via period modification unit 3. The shape index data and the gain index data are thence supplied to the

unvoiced speech synthesis unit 220. The shape index data from the terminal 207s and the gain index data from the terminal 207g are supplied to a noise codebook 221 and a gain circuit 222 of the unvoiced speech synthesis unit 220, respectively. A representative value output read out from the noise codebook 221 is the noise signal component corresponding to the LPC residuals of the unvoiced speech and becomes an amplitude of a pre-set gain in the gain circuit 222. The representative value output of the pre-set gain amplitude is sent to a windowing circuit 223 for windowing for smoothing a junction portion to the voiced signal portion.

An output of the windowing circuit 223 is sent as an output of the unvoiced speech synthesis unit 220 to a synthesis filter 237 for unvoiced (UV) portion of the LPC synthesis filter 214. The output of the windowing circuit 223 is processed by the synthesis filter 237 by LPV synthesis to give time-domain waveform signals of the unvoiced speech signal portion which is then filtered by a post-filter for unvoiced speech portion 238u and thence supplied to the adder 239.

The adder 239 sums the time-domain waveform signal for the voiced speech signal portion from the post-filter 238v for the voiced speech to the time-domain waveform data for the unvoiced speech signal portion, from the post-filter for the unvoiced speech signal portion 238u. The resulting sum signal is outputted at the output terminal 201.

It is seen from above that the pitch can be varied without

changing the phoneme of the speech by changing the number of harmonics without changing the shape of the spectral envelope. Thus, if encoded data of a speech pattern, that is an encoded bitstream, is available, its pitch may be optionally varied for synthesis.

Referring to Fig.15, an encoded bitstream or encoded data, obtained on encoding by the encoder of Figs.2 and 3, are outputted by an encoded data outputting unit 301. Of these data, at least the pitch data and spectral envelope data are sent via a data conversion unit 302 to a waveform synthesis unit 303. The data irrelevant to pitch conversion, such as voiced/unvoiced (V/UV) decision data, are directly sent to the waveform synthesis unit 303.

The waveform synthesis unit 303 synthesizes the speech waveform based on the spectral envelope data or pitch data. Of course, in the case of the synthesis device shown in Figs.4 or 5, LSP data or CELP data are also taken out from the outputting unit 301 and supplied as described above.

In the configuration of Fig.15, at least pitch data or spectral envelope data are converted by the data conversion unit 302 depending on the desired pitch as described above and thence supplied to the waveform synthesis unit 303 where the speech waveform is synthesized from the converted data. Thus the speech signals changed in pitch without changing the phoneme can be taken out at an output terminal 304.

The above-described technique can be used for synthesis of

speech by rule or text.

Fig.16 shows an example of application of the present invention to speech text synthesis. In the present embodiment, the above-described decoder for speech encoding for compression may be used simultaneously as a text speech synthesizer. In the example of Fig.16, regeneration of speech data is used in combination.

In Fig.6, the speech rule synthesizer and the speech synthesizer with data conversion for pitch modification as described above are comprised in a speech-by-rule synthesis unit 300. Data from a text analysis unit 310 is supplied to the speech-by-rule synthesis unit 300 from which the synthesized speech having the desired pitch is outputted and sent to a fixed contact *a* of a changeover switch 330. A speech reproducing unit 320 reads out speech data occasionally compressed and stored in a memory such as ROM and decodes the data for expansion. The decoded data is sent to the other fixed contact *b* of the changeover switch 330. One of the synthesized speech signals and the reproduced speech signals is selected by the changeover switch 330 and outputted at an output terminal 340.

The device shown in Fig.16 may be used in, for example, a navigation system for a vehicle. In such case, the reproduced speech of high quality and high clarity from the speech regenerator 320 may be used for routine speech, such as "Please turn to right" for bearing indication, while the synthesized speech from the speech-by-rule generator 300 may be used for speech of special designations

for e.g. a building or territory, which is voluminous and cannot be stored as speech information in a ROM.

The present invention has an additional merit that the same hardware may be used for the computer speech synthesizer 300 and the speech regenerator 320.

The present invention is not limited to the above-described embodiments. For example, the construction of the speech analysis side (encoder) of Figs.1 and 3 or the speech synthesis side (decoder) of Fig.14, described above as hardware, may be realized by a software program using, for example, a digital signal processor (DSP). The data of plural frames may be handled together and quantized by matrix quantization in place of vector quantization. The present invention may also be applied to a variety of speech analysis/synthesis methods. The present invention is also not limited to transmission or recording/reproduction and may be applied to a variety of usages such as pitch conversion, speed or rate conversion, synthesis of the speech-by-rule or noise suppression.

The above-described signal encoding and signal decoding apparatus may be used as a speech codec employed in, for example, a portable communication terminal or a portable telephone set shown in Fig.14.

Fig.17 shows a transmitting side of a portable terminal employing a speech encoding unit 160 configured as shown in Figs.2 and 3. The speech signals collected by a microphone 161 are amplified by an amplifier 162 and converted by an analog/digital

(A/D) converter 163 into digital signals which are sent to the speech encoding unit 160 configured as shown in Figs.1 and 3. The digital signals from the A/D converter 163 are supplied to the input terminal 101. The speech encoding unit 160 performs encoding as explained in connection with Figs.1 and 3. Output signals of output terminals of Figs.1 and 2 are sent as output signals of the speech encoding unit 160 to a transmission channel encoding unit 164 which then performs channel coding on the supplied signals. Output signals of the transmission channel encoding unit 164 are sent to a modulation circuit 165 for modulation and thence supplied to an antenna 168 via a digital/analog (D/A) converter 166 and an RF amplifier 167.

Fig.18 shows a reception side of a portable terminal employing a speech decoding unit 260 configured as shown in Figs.5 and 14. The speech signals received by the antenna 261 of Fig.14 are amplified in an RF amplifier 262 and sent via an analog/digital (A/D) converter 263 to a demodulation circuit 264, from which demodulated signals are sent to a transmission channel decoding unit 265. An output signal of the decoding unit 265 is supplied to a speech decoding unit 260 configured as shown in Figs.5 and 14. The speech decoding unit 260 decodes the signals as explained in connection with Figs.5 and 14. An output signal at an output terminal 201 of Figs.2 and 4 is sent as a signal of the speech decoding unit 260 to a digital/analog (D/A) converter 266. An analog speech signal from the D/A converter 266 is sent to a

CLAIMS

Claim 1. A method for reproducing speech signals based on encoded parameters as found by dividing an input speech signal in terms of pre-set encoding units on the time axis and by encoding the divided input speech signals,

wherein the improvement comprises interpolating said encoded parameters for finding modified encoded parameters for desired time points and reproducing the speech signals based on the modified encoded parameters.

Claim 2. The method as claimed in claim 1 wherein said encoded parameters are those obtained on harmonic coding.

Claim 3. The method as claimed in claim 1 wherein it is judged whether said input speech signal is voiced or unvoiced and, based on the result of judgment, the portion of the input speech signal judged to be voiced is encoded by sinusoidal synthetic encoding, and the portion of the input speech signal judged to be unvoiced is quantized by vector quantization by closed loop search of an optimum vector using an analysis by synthesis method.

Claim 4. The method as claimed in claim 1 further comprising:

a period modifying step companding the time axis of the encoded parameters obtained from one encoding unit to another for modifying an output period of said encoded parameters;

an interpolation step for interpolating the modified parameters to find modified encoded parameters associated with time points corresponding to said encoding units; and



a speech synthesis step of synthesizing said voiced and unvoiced portions based on said modified encoded parameters.

Claim 5. The method as claimed in claim 1 wherein a noise component is added to an excitation signal at the time of synthesis of said unvoiced portion, said noise component is substituted for said excitation signal or an excitation vector randomly selected from a codebook is used.

Claim 6. An apparatus for reproducing speech signals based on encoded parameters as found by dividing an input speech signal in terms of pre-set encoding units on the time axis and by encoding the divided input speech signals,

wherein the improvement consists in that said encoded parameters are interpolated for finding modified encoded parameters for desired time points and the speech signal is reproduced based on the modified encoded parameters.

Claim 7. The apparatus as claimed in claim 6 wherein said encoded parameters are those obtained on harmonic coding.

Claim 8. The apparatus as claimed in claim 1 wherein it is judged whether said input speech signal is voiced or unvoiced and, based on the result of judgment, the portion of the input speech signal judged to be voiced is encoded by sinusoidal synthetic encoding, and the portion of the input speech signal judged to be unvoiced is quantized by vector quantization by closed loop search for an optimum vector using an analysis by synthesis method.

Claim 9. The apparatus as claimed in claim 6 further comprising:



a period modifying step companding the time axis of the encoded parameters obtained from one encoding unit to another for modifying an output period of said encoded parameters;

an interpolation step for interpolating the modified parameters to find modified encoded parameters associated with time points corresponding to said encoding units; and

a speech synthesis step of synthesizing said voiced and unvoiced portions based on said modified encoded parameters.

Claim 10. The apparatus as claimed in claim 6 wherein a noise component is added to an excitation signal at the time of synthesis of said unvoiced portion, said noise component is substituted for said excitation signal or an excitation vector randomly selected from a codebook is used.

Claim 11. A method for reproducing speech signals based on encoded parameters as found by dividing an input speech signal in terms of pre-set blocks on the time axis and by encoding the divided input speech signals,

wherein the speech is reproduced using said encoded parameters and with blocks of lengths different from those at the time.

Claim 12. The method as claimed in claim 11 wherein said encoded parameters are found by a first step of harmonic coding or harmonic coding of LPC residuals and a second step of waveform encoding.

Claim 13. The method as claimed in claim 11 wherein it is judged whether the input speech signal is voiced or unvoiced and, based on



the result of judgment, the portion of the input speech signal judged to be voiced is encoded by harmonic coding or harmonic coding of LPC residuals and the portion of the input speech signal judged to be unvoiced is quantized by vector quantization of the time waveform of LPC residuals employing the analysis by synthesis method.

Claim 14. The method as claimed in claim 12 wherein a sub-frame length of interpolation of LSPs representing a spectral envelope is modified responsive to the reproducing speed designated by a decoder.

Claim 15. The method as claimed in claim 11 wherein the length of an excitation signal for an unvoiced signal portion is modified in such a manner that, if the length is decreased, the inherent excitation signal is used and, if the length is in deficit, the inherent excitation signal added to with a noise component, a noise component or an excitation vector randomly selected from an excitation signal codebook is used.

Claim 16. A speech decoding method comprising:

a step of converting the fundamental frequency and the number in a pre-set band of harmonics of an input encoded speech data, and

a step of interpolating and modifying the number of data specifying the size of spectral component in each input harmonics for modifying the pitch of a synthesized speech.

Claim 17. The method as claimed in claim 16 wherein said interpolation is executed using a band-limited type oversampling



filter.

Claim 18. A speech decoding apparatus comprising:

means for converting the fundamental frequency and the number in a pre-set band of harmonics of an input encoded speech data, and

means for interpolating and modifying the number of data specifying the size of spectral component in each input harmonics for modifying the pitch of a synthesized speech.

Claim 19. The apparatus as claimed in claim 18 wherein said interpolation is executed using a band-limited type oversampling filter.

Claim 20. A speech synthesis method comprising:

a regular speech synthesis step of synthesizing a regular speech in accordance with a pre-set rule for outputting amplitude data of harmonics;

a number of data conversion step of converting the fundamental frequency of harmonics of input data and the number of amplitudes in a pre-set band; and

a step of interpolating the data specifying the size of spectral component in each input harmonics for modifying the pitch of a synthesized speech.

Claim 21. The speech synthesis method as claimed in claim 20 wherein said interpolation is executed using a band-limited type oversampling filter.

Claim 22. A speech synthesis apparatus comprising:



regular speech synthesis means of synthesizing a regular speech in accordance with a text for outputting amplitude data of harmonics;

number of data conversion means for converting the fundamental frequency of harmonics of input data and the number of amplitudes in a pre-set band; and

means for interpolating the data specifying the size of spectral component in each input harmonics for modifying the pitch of a synthesized speech.

Claim 23. The speech synthesis apparatus as claimed in claim 22 wherein said interpolation is executed using a band-limited type oversampling filter.

Claim 24. A portable radio terminal apparatus comprising:

amplifier means for amplifying a reception signal; demodulation means for A/D converting and subsequently demodulating said amplified signal;

transmission path decoding means for channel-coding said demodulated signal;

speech-decoding means for speech-encoding an output of said transmission path decoding means; and

D/A conversion means for D/A conversion of decoded speech signals from said speech decoding means for producing analog speech signal.

Claim 25. The portable radio terminal apparatus as claimed in claim 24 wherein said speech decoding means includes number of data



conversion means for converting the fundamental frequency of harmonics of input data and the number of amplitudes in a pre-set band, and

means for interpolating the data specifying the size of spectral component in each input harmonics for modifying the pitch of a synthesized speech.



* C L M S *

METHOD AND APPARATUS FOR REPRODUCING SPEECH SIGNALS,
METHOD AND APPARATUS FOR DECODING THE SPEECH, METHOD AND
APPARATUS FOR SYNTHESIZING THE SPEECH AND PORTABLE
RADIO TERMINAL APPARATUS

ABSTRACT

A method for reproducing speech signals at a controlled speed whereby rate conversion of the time axis may be facilitated, and a method for synthesizing the speech whereby pitch conversion can be realized by a simplified structure based on the encoded speech data without changing the phoneme. With the speech reproducing method, an encoding unit 2 discriminates whether an input speech signal is voiced or unvoiced. Based on the results of discrimination, the encoding unit 2 performs sinusoidal synthesis and encoding for a signal portion found to be voiced, while performing vector quantization by closed-loop search for an optimum vector for a portion found to be unvoiced using an analysis-by-synthesis method, in order to find encoded parameters. The decoding unit 3 compands the time axis of the encoded parameters obtained every pre-set frames at a period modification unit 4 for modifying the output period of the parameters for creating modified encoded parameters associated with different time points corresponding to the pre-set frames. A speech synthesis unit 6 synthesizes the voiced speech portion and the unvoiced speech portion based on the modified encoded parameters. With the speech synthesizing unit, an encoded bit stream or encoded data is outputted by an encoded data outputting unit 301. Of these data, at least pitch data and amplitude data of the spectral envelope are sent via a data conversion unit 302 to a waveform synthesis unit 302, where the number of amplitude data of the spectral envelope is changed without changing the shape of the spectral envelope depending on

a pitch desired pitch value. A waveform synthesis unit 303 synthesizes the speech waveform based on the converted spectral envelope data and pitch data.

Figure 1 & 15

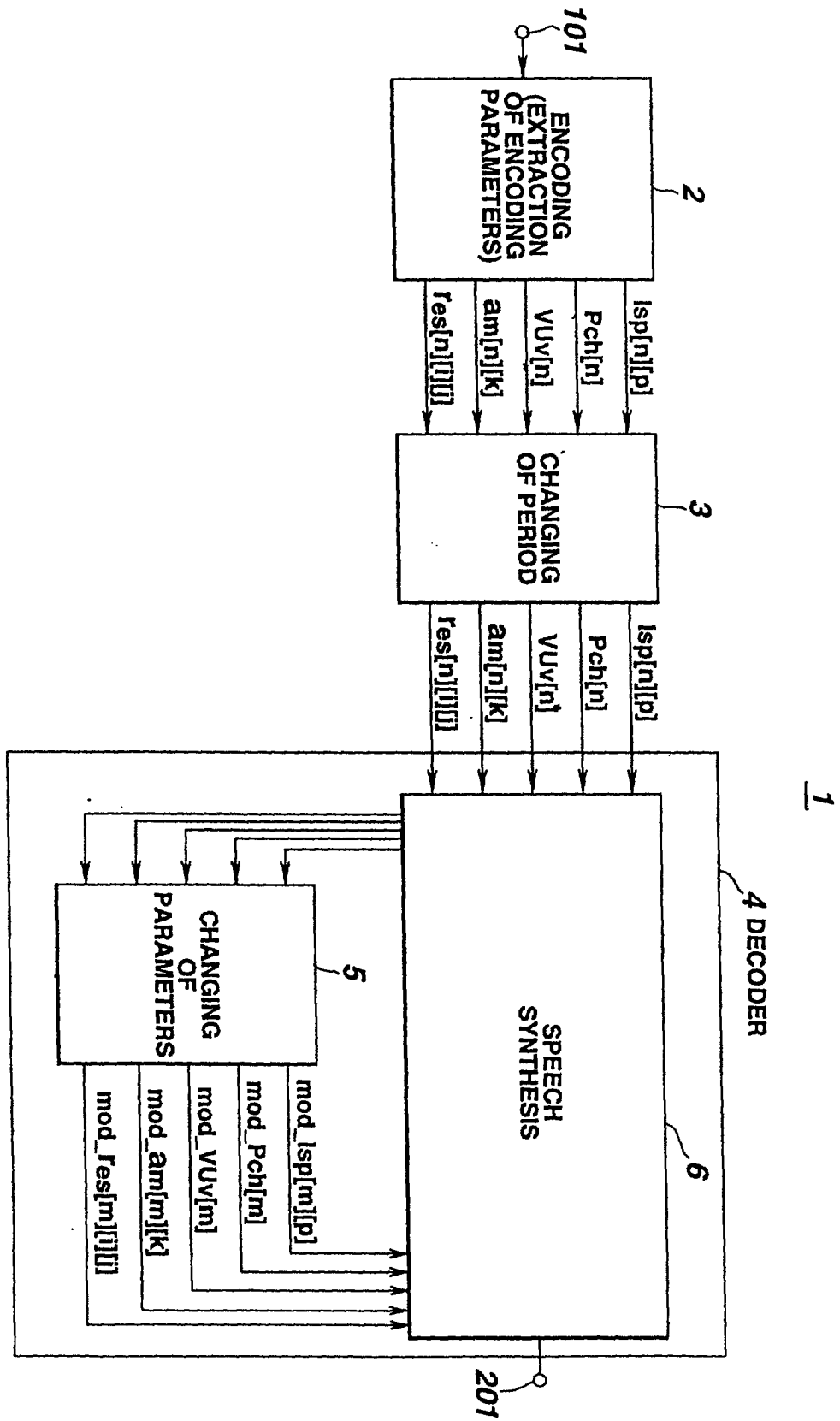


FIG.1



FIG. 2

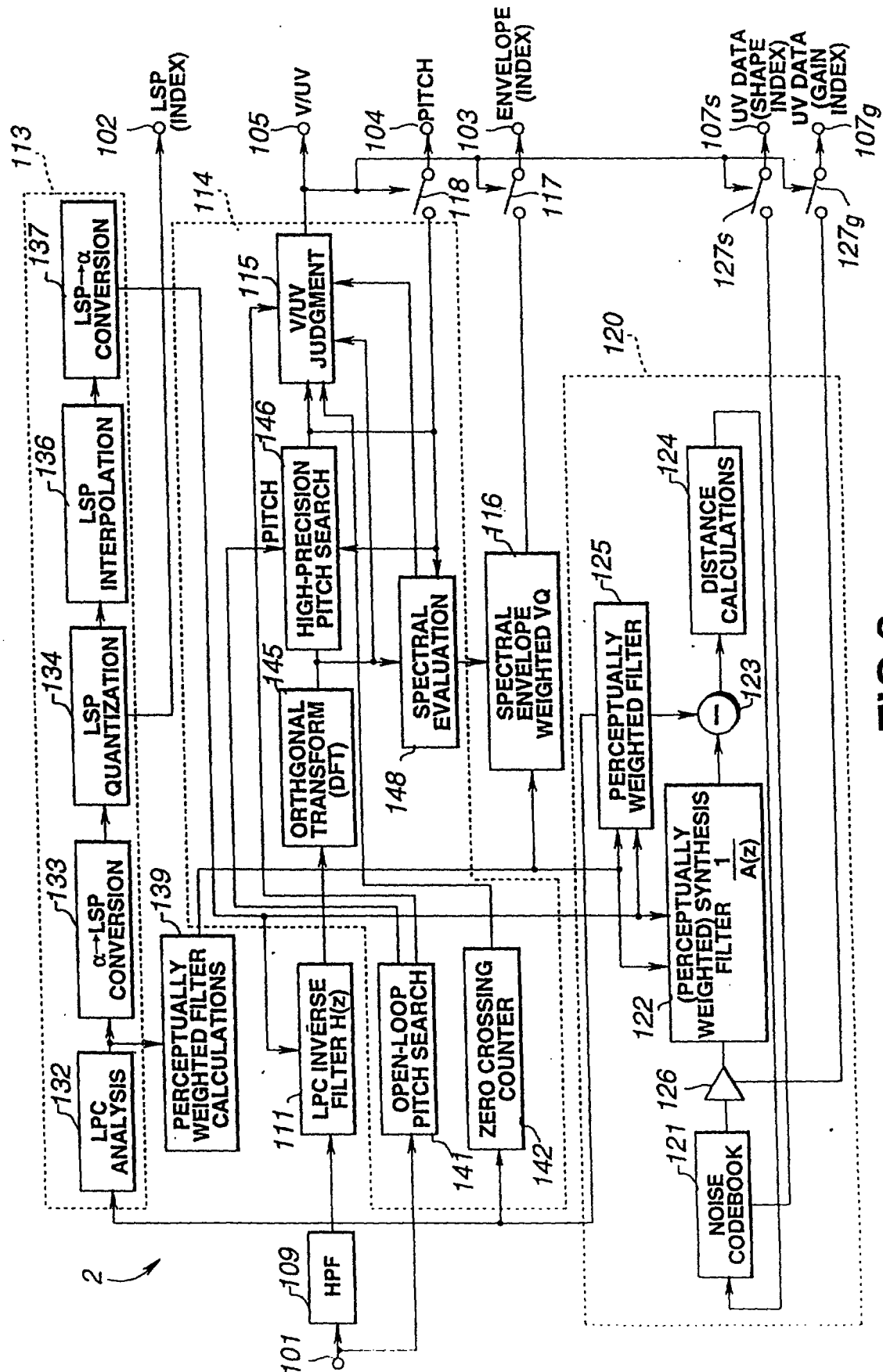


FIG. 3

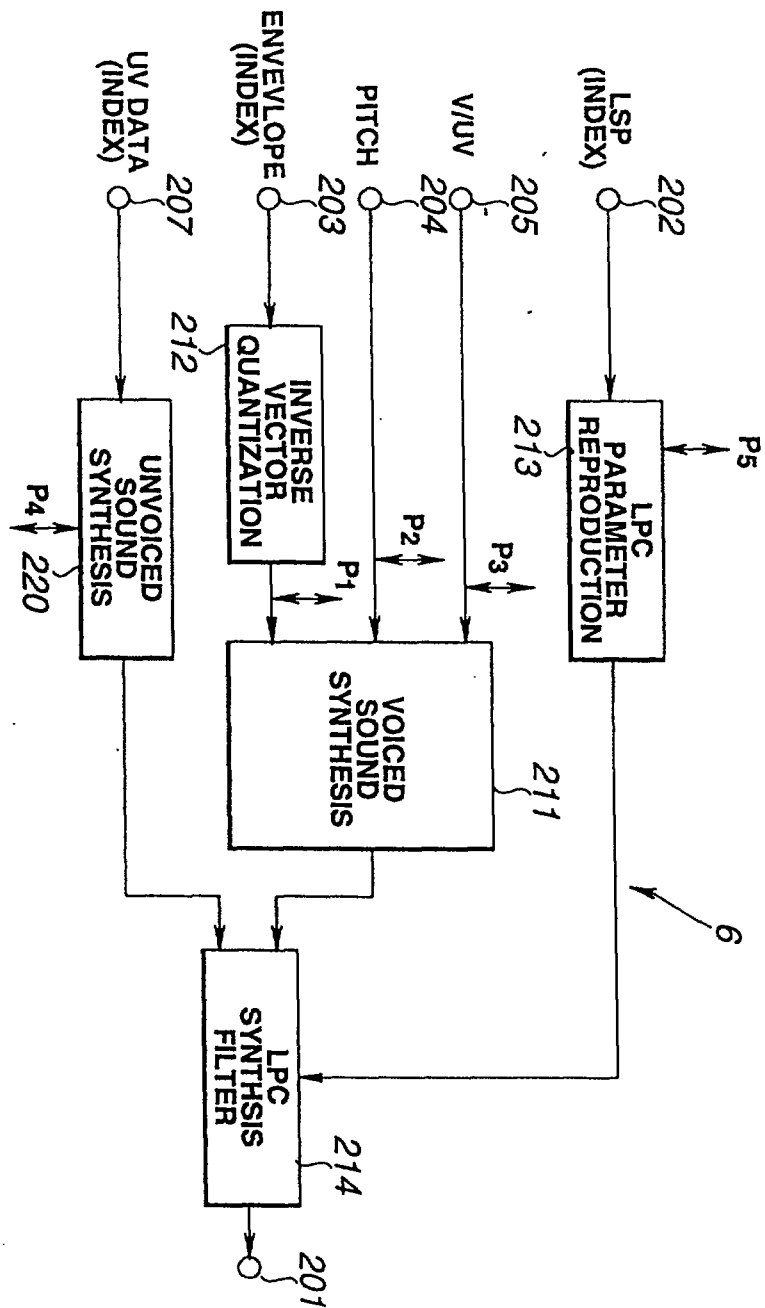


FIG. 4

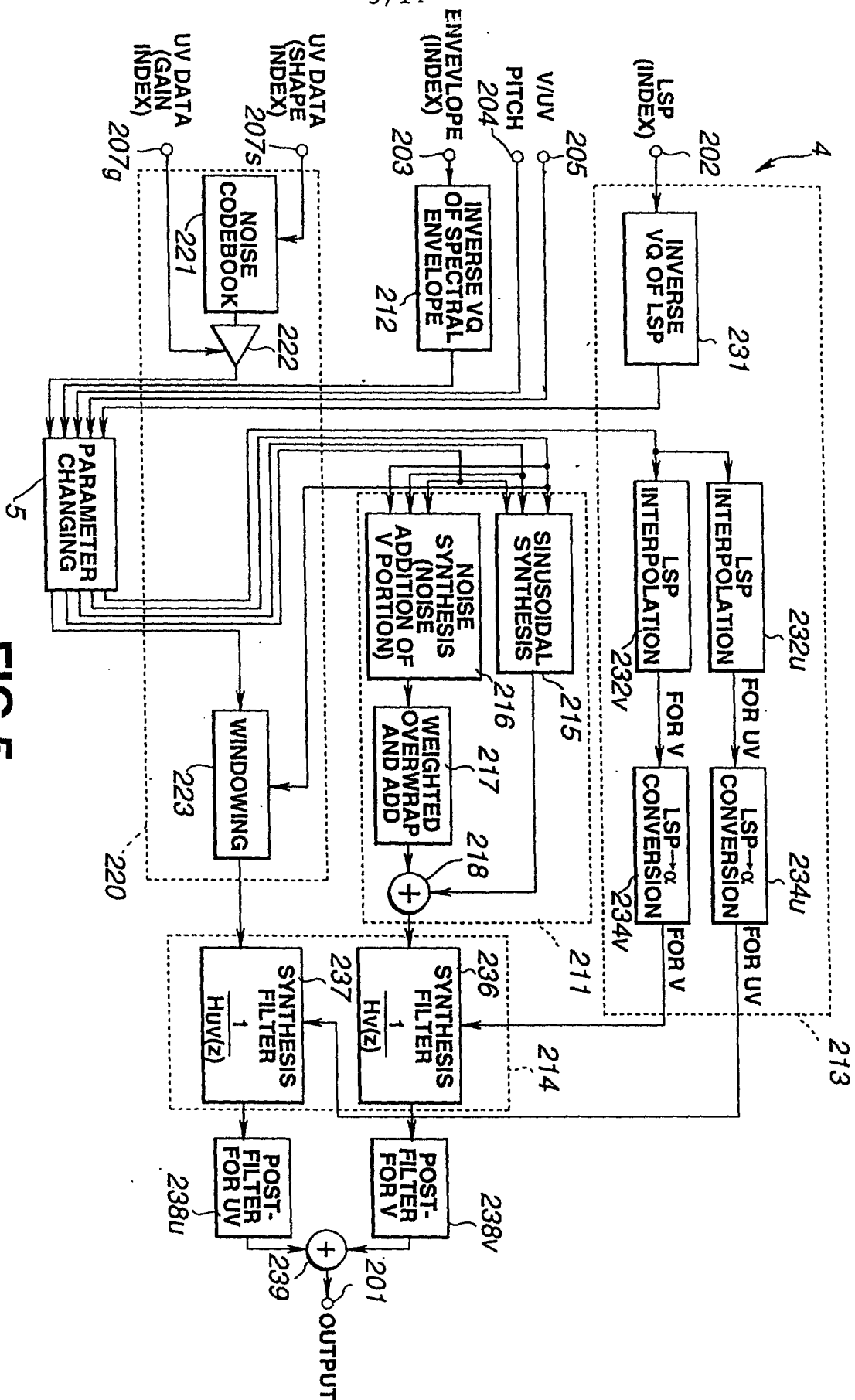
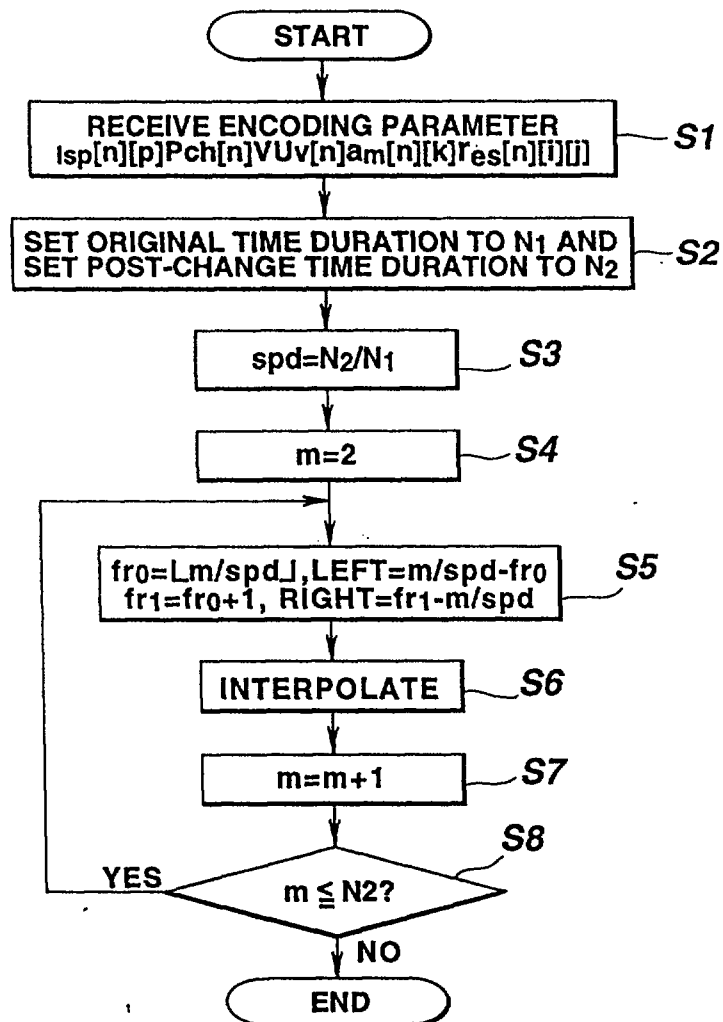
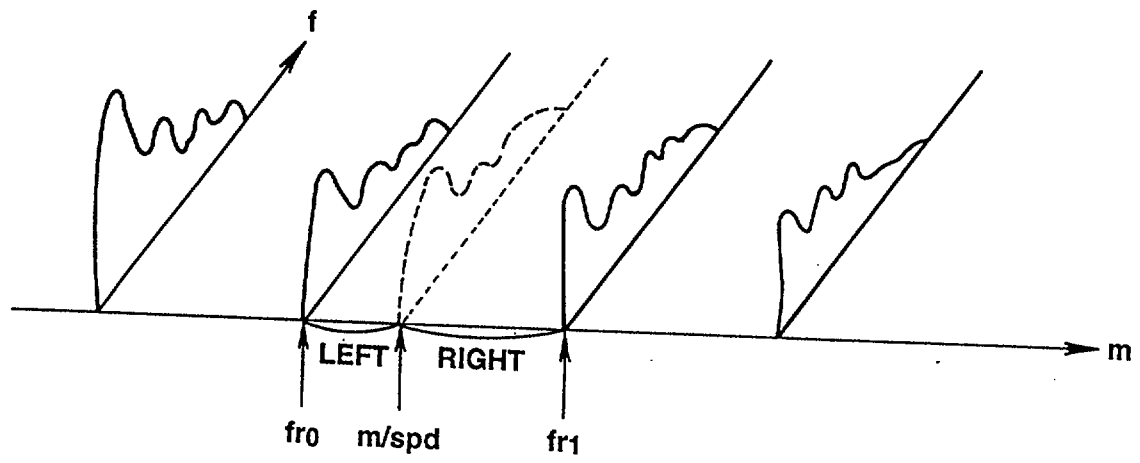


FIG. 5

**FIG.6**

**FIG.7**

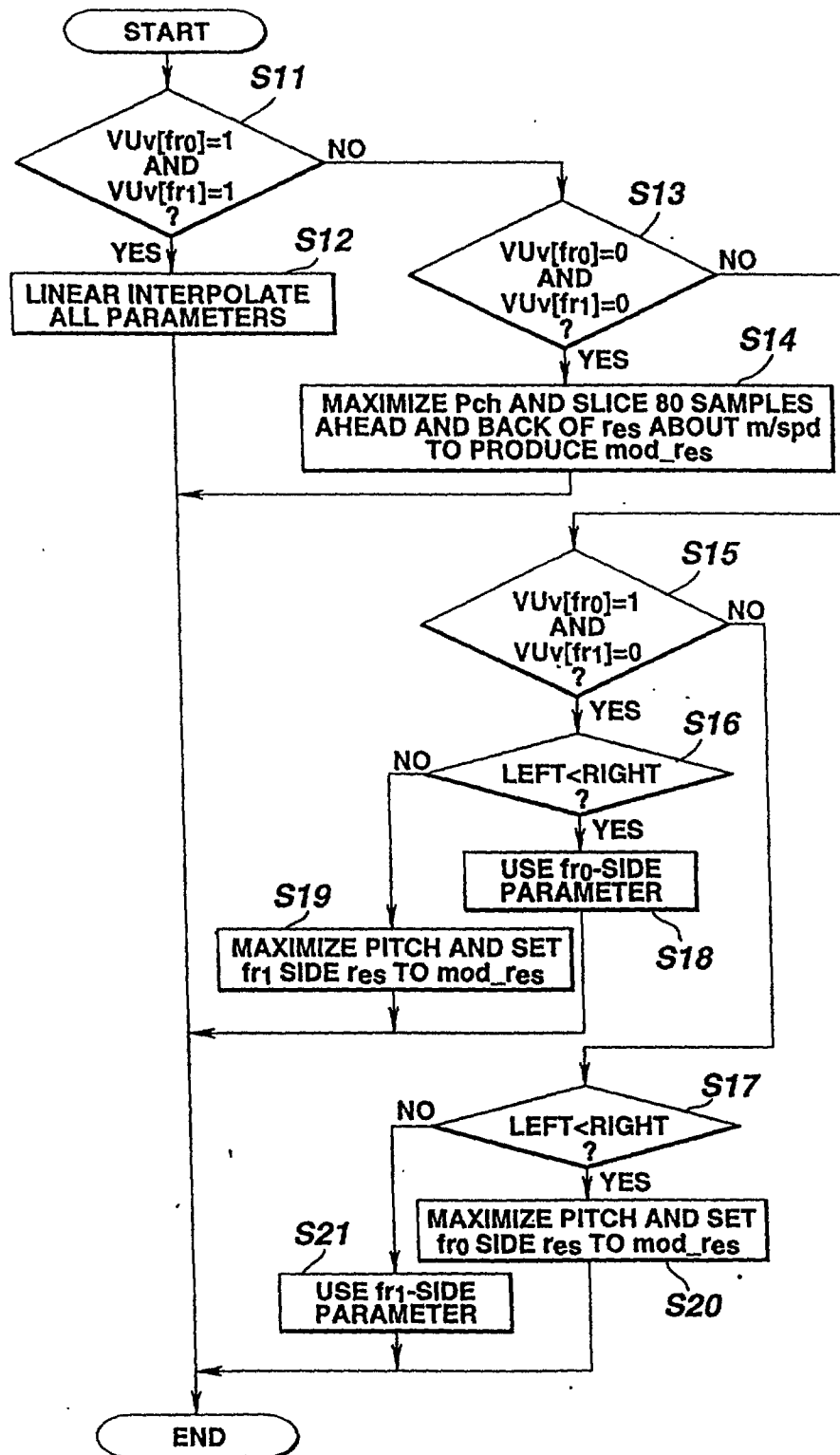
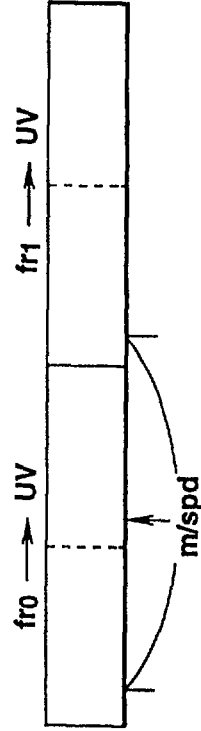
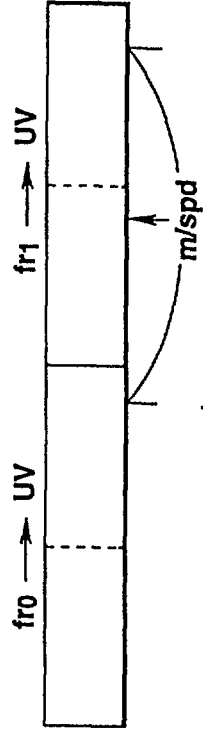


FIG.8

FOR UV-UV
WITH
LEFT < RIGHT



FOR UV-UV
LEFT ≥ RIGHT



FOR V-UV



FOR UV-V



FIG. 9(A)

FIG. 9(B)

FIG. 9(C)

FIG. 9(D)

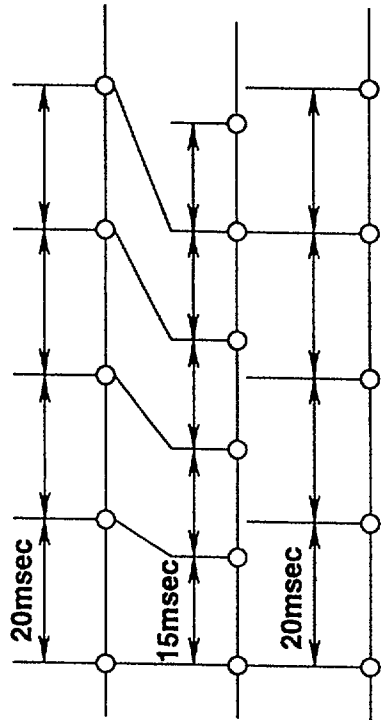


FIG. 10(A)

FIG. 10(B) COMPRESSION

FIG. 10(C) INTERPOLATION

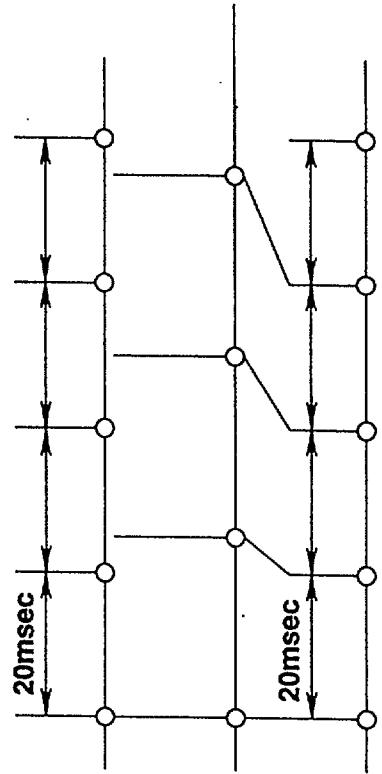


FIG. 11(A)

FIG. 11(B) INTERPOLATION

FIG. 11(C) COMPRESSION

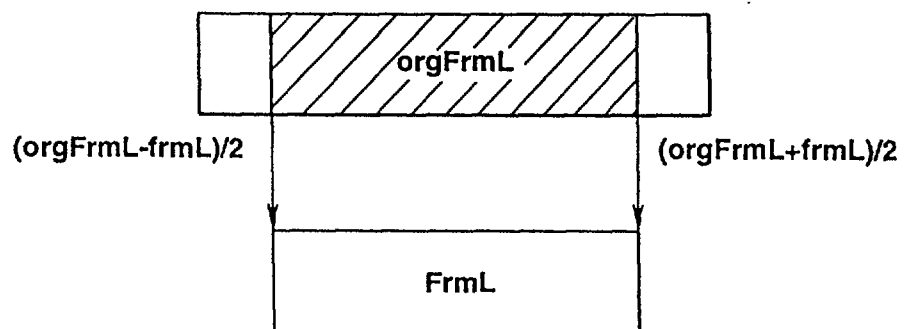


FIG.12

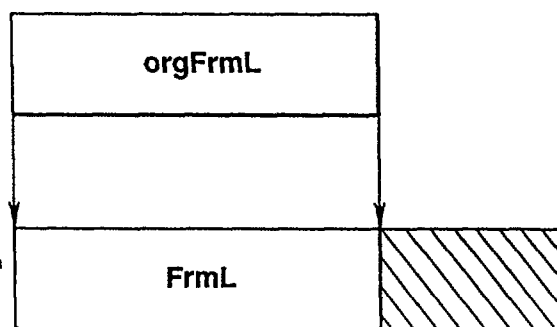


FIG.13

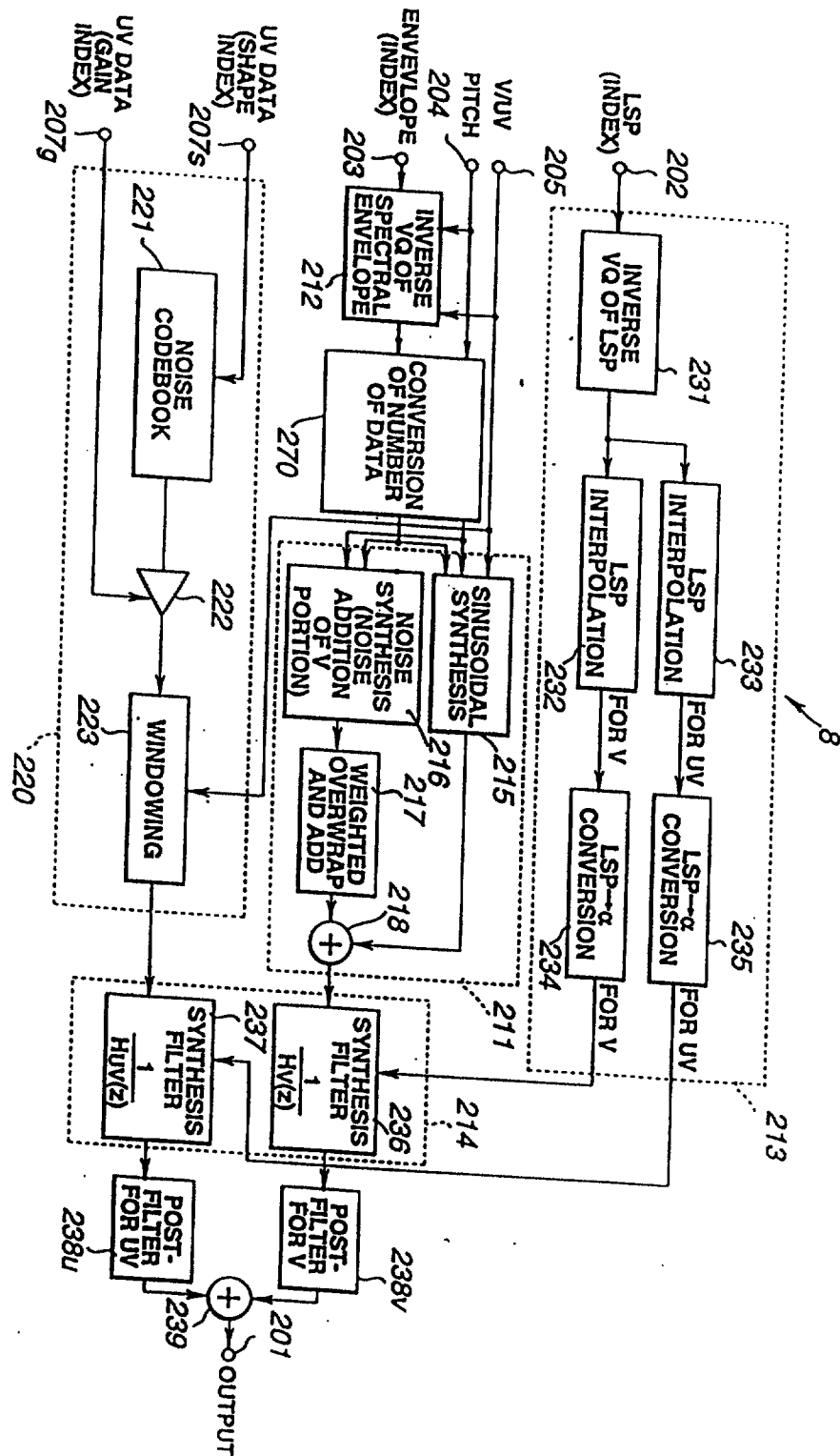


FIG. 14

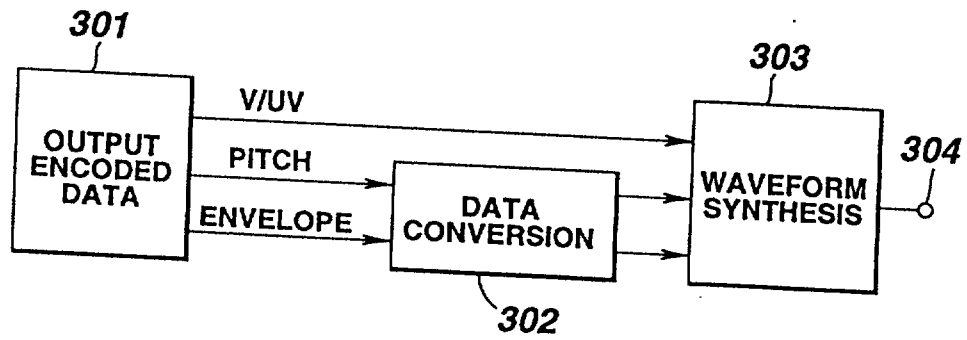


FIG. 15

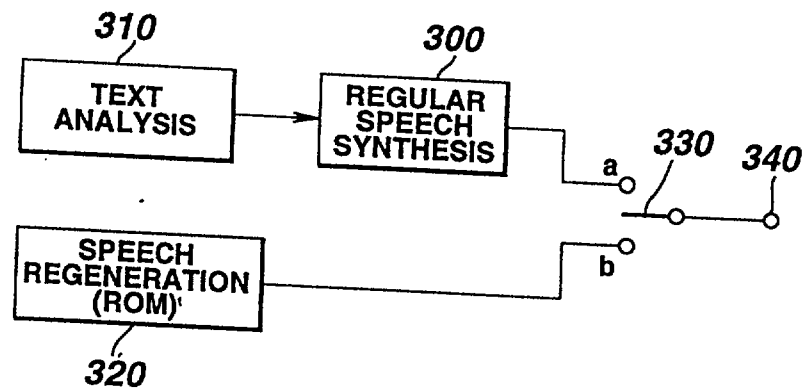


FIG. 16

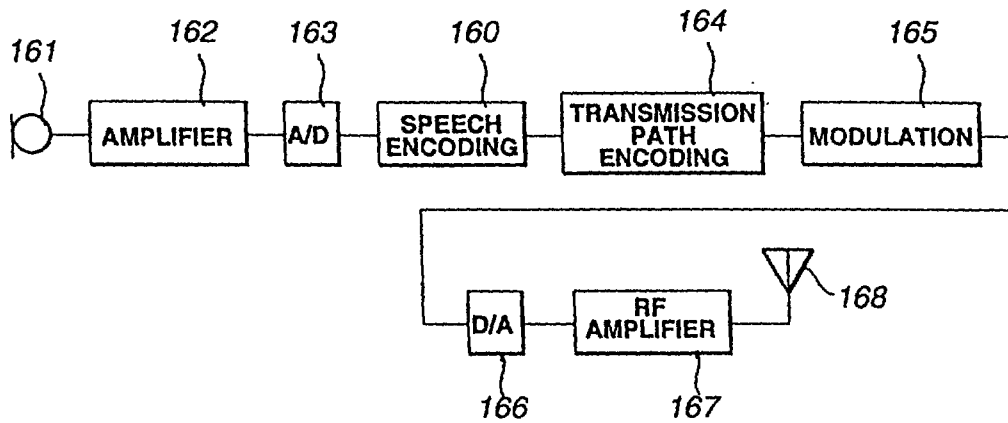


FIG.17

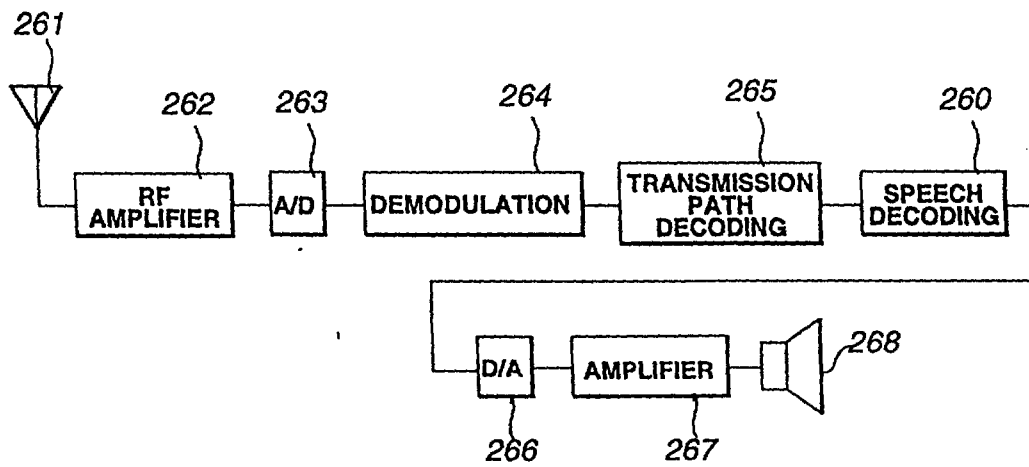


FIG.18