PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : G10L 7/02	A1	(11) International Publication Number: WO 92/05539
		(43) International Publication Date: 2 April 1992 (02.04.92)

(21) International Application Number: PCT/US91/06853

(22) International Filing Date: 20 September 1991 (20.09.91)

(30) Priority data:
585,830 20 September 1990 (20.09.90) US

(71) Applicant: DIGITAL VOICE SYSTEMS, INC. [US/US]; One Kendall Square, Building 300, Cambridge, MA 02139 (US).

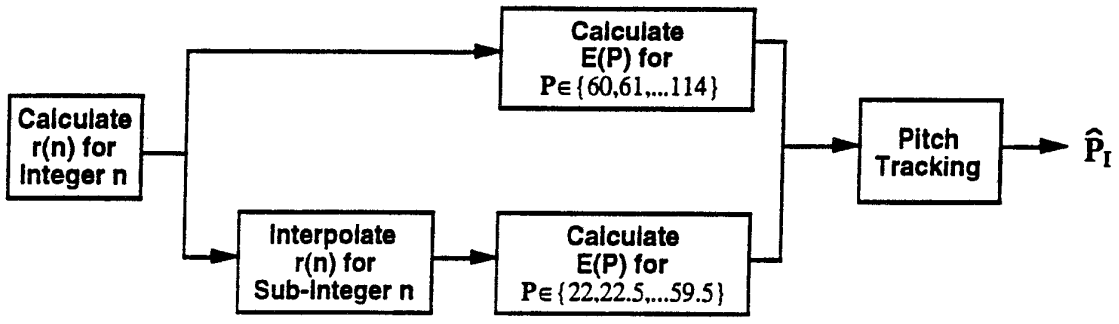
(72) Inventors: HARDWICK, John, C. ; 133 Webster Avenue, Apt. 3, Cambridge, MA 02141 (US). LIM, Jae, S. ; 21 West Chardon Road, Winchester, MA 01890 (US).

(74) Agent: LEE, G., Roger; Fish & Richardson, 225 Franklin Street, Suite 3100, Boston, MA 02110-2804 (US).

(81) Designated States: AT (European patent), AU, BE (European patent), CA, CH (European patent), DE (European patent), DK (European patent), ES (European patent), FI, FR (European patent), GB (European patent), GR (European patent), IT (European patent), JP, KR, LU (European patent), NL (European patent), NO, SE (European patent).

Published
With international search report.

(54) Title: METHODS FOR SPEECH ANALYSIS AND SYNTHESIS



(57) Abstract

Sub-integer resolution pitch values are estimated in making the initial pitch estimate; the sub-integer pitch values are preferably estimated by interpolating intermediate variables between integer values. Pitch regions are used to reduce the amount of computation required in making the initial pitch estimate. Pitch-dependent resolution is used in making the initial pitch estimate, with higher resolution being used for smaller values of pitch.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	**ES**	Spain	**MG**	Madagascar
AU	Australia	**FI**	Finland	**ML**	Mali
BB	Barbados	**FR**	France	**MN**	Mongolia
BE	Belgium	**GA**	Gabon	**MR**	Mauritania
BF	Burkina Faso	**GB**	United Kingdom	**MW**	Malawi
BG	Bulgaria	**GN**	Guinea	**NL**	Netherlands
BJ	Benin	**GR**	Greece	**NO**	Norway
BR	Brazil	**HU**	Hungary	**PL**	Poland
CA	Canada	**IT**	Italy	**RO**	Romania
CF	Central African Republic	**JP**	Japan	**SD**	Sudan
CG	Congo	**KP**	Democratic People's Republic of Korea	**SE**	Sweden
CH	Switzerland			**SN**	Senegal
CI	Côte d'Ivoire	**KR**	Republic of Korea	**SU**+	Soviet Union
CM	Cameroon	**LI**	Liechtenstein	**TD**	Chad
CS	Czechoslovakia	**LK**	Sri Lanka	**TG**	Togo
DE*	Germany	**LU**	Luxembourg	**US**	United States of America
DK	Denmark	**MC**	Monaco		

+ Any designation of "SU" has effect in the Russian Federation. It is not yet known whether any such designation has effect in other States of the former Soviet Union.

Methods for Speech Analysis and Synthesis

Background of the Invention

This invention relates to methods for encoding and synthesizing speech.

Relevant publications include: Flanagan, Speech Analysis, Synthesis and Perception, Springer-Verlag, 1972, pp. 378-386, (discusses phase vocoder – frequency-based speech analysis-synthesis system); Quatieri, et al., "Speech Transformations Based on a Sinusoidal Representation", IEEE TASSP, Vol, ASSP34, No. 6, Dec. 1986, pp. 1449-1986, (discusses analysis-synthesis technique based on a sinsusoidal representation); Griffin, et al., "Multiband Excitation Vocoder", Ph.D. Thesis, M.I.T, 1987, (discusses Multi-Band Excitation analysis-synthesis); Griffin, et al., "A New Pitch Detection Algorithm", Int. Conf. on DSP, Florence, Italy , Sept. 5-8, 1984, (discusses pitch estimation); Griffin, et al., "A New Model-Based Speech Analysis/Synthesis System", Proc ICASSP 85, pp. 513-516, Tampa, FL., March 26-29, 1985, (discusses alternative pitch likelihood functions and voicing measures); Hardwick, "A 4.8 kbps Multi-Band Excitation Speech Coder", S.M. Thesis, M.I.T, May 1988, (discusses a 4.8 kbps speech coder based on the Multi-Band Excitation speech model); McAulay et al., "Mid-Rate Coding Based on a Sinusoidal Representation of Speech", Proc. ICASSP 85, pp. 945-948, Tampa, FL., March 26-29, 1985, (discusses speech coding based on a sinusoidal representation); Almieda et al., "Harmonic Coding with Variable Frequency Synthesis", Proc. 1983 Spain Workshop on Sig. Proc. and its Applications", Sitges, Spain, Sept., 1983, (discusses time domain voiced synthesis); Almieda et al., "Variable Frequency Synthesis: An Improved Harmonic Coding Scheme", Proc ICASSP 84, San Diego, CA., pp. 289-292, 1984, (discusses time domain voiced synthesis); McAulay et al., "Computationally Efficient Sine-Wave Synthesis and its Application to Sinusoidal Transform Coding", Proc. ICASSP 88, New York, NY., pp. 370-373, April 1988, (discusses frequency domain voiced synthesis); Griffin et al., "Signal Estimation From Modified Short-Time Fourier Transform", IEEE TASSP, Vol. 32, No.

2, pp. 236-243, April 1984, (discusses weighted overlap-add synthesis). The contents of these publications are incorporated herein by reference.

The problem of analyzing and synthesizing speech has a large number of applications, and as a result has received considerable attention in the literature. One class of speech analysis/synthesis systems (vocoders) which have been extensively studied and used in practice is based on an underlying model of speech. Examples of vocoders include linear prediction vocoders, homomorphic vocoders, and channel vocoders. In these vocoders, speech is modeled on a short-time basis as the response of a linear system excited by a periodic impulse train for voiced sounds or random noise for unvoiced sounds. For this class of vocoders, speech is analyzed by first segmenting speech using a window such as a Hamming window. Then, for each segment of speech, the excitation parameters and system parameters are determined. The excitation parameters consist of the voiced/unvoiced decision and the pitch period. The system parameters consist of the spectral envelope or the impulse response of the system. In order to synthesize speech, the excitation parameters are used to synthesize an excitation signal consisting of a periodic impulse train in voiced regions or random noise in unvoiced regions. This excitation signal is then filtered using the estimated system parameters.

Even though vocoders based on this underlying speech model have been quite successful in synthesizing intelligible speech, they have not been successful in synthesizing high-quality speech. As a consequence, they have not been widely used in applications such as time-scale modification of speech, speech enhancement, or high-quality speech coding. The poor quality of the synthesized speech is in part, due to the inaccurate estimation of the pitch, which is an important speech model parameter.

To improve the performance of pitch detection, a new method was developed by Griffin and Lim in 1984. This method was further refined by Griffin and Lim in 1988. This method is useful for a variety of different vocoders, and is particularly useful for

a Multi-Band Excitation (MBE) vocoder.

Let $s(n)$ denote a speech signal obtained by sampling an analog speech signal. The sampling rate typically used for voice coding applications ranges between 6khz and 10khz. The method works well for any sampling rate with corresponding change in the various parameters used in the method.

We multiply $s(n)$ by a window $w(n)$ to obtain a windowed signal $s_w(n)$. The window used is typically a Hamming window or Kaiser window. The windowing operation picks out a small segment of $s(n)$. A speech segment is also referred to as a speech frame.

The objective in pitch detection is to estimate the pitch corresponding to the segment $s_w(n)$. We will refer to $s_w(n)$ as the current speech segment and the pitch corresponding to the current speech segment will be denoted by P_0, where "0" refers to the "current" speech segment. We will also use P to denote P_0 for convenience. We then slide the window by some amount (typically around 20 msec or so), and obtain a new speech frame and estimate the pitch for the new frame. We will denote the pitch of this new speech segment as P_1. In a similar fashion, P_{-1} refers to the pitch of the past speech segment. The notations useful in this description are P_0 corresponding to the pitch of the current frame, P_{-2} and P_{-1} corresponding to the pitch of the past two consecutive speech frames, and P_1 and P_2 corresponding to the pitch of the future speech frames.

The synthesized speech at the synthesizer, corresponding to $s_w(n)$ will be denoted by $\hat{s}_w(n)$. The Fourier transforms of $s_w(n)$ and $\hat{s}_w(n)$ will be denoted by $S_w(\omega)$ and $\hat{S}_w(\omega)$.

The overall pitch detection method is shown in Figure 1. The pitch P is estimated using a two-step procedure. We first obtain an initial pitch estimate denoted by $\hat{P}_I$. The initial estimate is restricted to integer values. The initial estimate is then refined to obtain the final estimate $\hat{P}$, which can be a non-integer value. The two-step procedure reduces the amount of computation involved.

To obtain the initial pitch estimate, we determine a pitch likelihood function $E(P)$, as a function of pitch. This likelihood function provides a means for the numerical comparison of candidate pitch values. Pitch tracking is used on this pitch likelihood function as shown in Figure 2. In all our discussions in the initial pitch estimation, P is restricted to integer values. The function $E(P)$ is obtained by,

$$E(P) = \frac{\sum_{j=-\infty}^{\infty} s^2(j)w^2(j) - P \cdot \sum_{n=-\infty}^{\infty} r(n \cdot P)}{(\sum_{j=-\infty}^{\infty} s^2(j)w^2(j))(1 - P \cdot \sum_{j=-\infty}^{\infty} w^4(j))} \quad (1)$$

where $r(n)$ is an autcorrelation function given by

$$r(n) = \sum_{j=-\infty}^{\infty} s(j)w^2(j)s(j+n)w^2(j+n) \quad (2)$$

and where,

$$\sum_{j=-\infty}^{\infty} w^2(j) = 1 \quad (3)$$

Equations (1) and (2) can be used to determine $E(P)$ for only integer values of P, since $s(n)$ and $w(n)$ are discrete signals.

The pitch likelihood function $E(P)$ can be viewed as an error function, and typically it is desirable to choose the pitch estimate such that $E(P)$ is small. We will see soon why we do not simply choose the P that minimizes $E(P)$. Note also that $E(P)$ is one example of a pitch likelihood function that can be used in estimating the pitch. Other reasonable functions may be used.

Pitch tracking is used to improve the pitch estimate by attempting to limit the amount the pitch changes between consecutive frames. If the pitch estimate is chosen to strictly minimize $E(P)$, then the pitch estimate may change abruptly between succeeding frames. This abrupt change in the pitch can cause degradation in the synthesized speech. In addition, pitch typically changes slowly; therefore, the pitch estimates from neighboring frames can aid in estimating the pitch of the current frame.

Look-back tracking is used to attempt to preserve some continuity of P from the past frames. Even though an arbitrary number of past frames can be used, we will use two past frames in our discussion.

Let $\hat{P}_{-1}$ and $\hat{P}_{-2}$ denote the initial pitch estimates of P_{-1} and P_{-2}. In the current frame processing, $\hat{P}_{-1}$ and $\hat{P}_{-2}$ are already available from previous analysis. Let $E_{-1}(P)$ and $E_{-2}(P)$ denote the functions of Equation (1) obtained from the previous two frames. Then $E_{-1}(\hat{P}_{-1})$ and $E_{-2}(\hat{P}_{-2})$ will have some specific values.

Since we want continuity of P, we consider P in the range near $\hat{P}_{-1}$. The typical range used is

$$(1-\alpha)\cdot\hat{P}_{-1} \leq P \leq (1+\alpha)\cdot\hat{P}_{-1} \tag{4}$$

where α is some constant.

We now choose the P that has the minimum $E(P)$ within the range of P given by (4). We denote this P as P^*. We now use the following decision rule.

$$\text{If } E_{-2}(\hat{P}_{-2}) + E_{-1}(\hat{P}_{-1}) + E(P^*) \leq \text{Threshold},$$
$$\hat{P}_I = P^* \text{ where } \hat{P}_I \text{ is the initial pitch estimate of } P. \tag{5}$$

If the condition in Equation (5) is satisfied, we now have the initial pitch estimate $\hat{P}_I$. If the condition is not satisfied, then we move to the look-ahead tracking.

Look-ahead tracking attempts to preserve some continuity of P with the future frames. Even though as many frames as desirable can be used, we will use two future frames for our discussion. From the current frame, we have $E(P)$. We can also compute this function for the next two future frames. We will denote these as $E_1(P)$ and $E_2(P)$. This means that there will be a delay in processing by the amount that corresponds to two future frames.

We consider a reasonable range of P that covers essentially all reasonable values of P corresponding to human voice. For speech sampled at 8khz rate, a good range of P to consider (expressed as the number of speech samples in each pitch period) is $22 \leq P < 115$.

For each P within this range, we choose a P_1 and P_2 such that CE(P) as given by (6) is minimized,

$$CE(P) = E(P) + E_1(P_1) + E_2(P_2) \tag{6}$$

subject to the constraint that P_1 is "close" to P and P_2 is "close" to P_1. Typically these "closeness" constraints are expressed as:

$$(1-\alpha)P \leq P_1 \leq (1+\alpha)P \tag{7}$$

and

$$(1-\beta)P_1 \leq P_2 \leq (1+\beta)P_1 \tag{8}$$

This procedure is sketched in Figure 3. Typical values for α and β are $\alpha = \beta = .2$

For each P, we can use the above procedure to obtain $CE(P)$. We then have $CE(P)$ as a function of P. We use the notation CE to denote the "cumulative error".

Very naturally, we wish to choose the P that gives the minimum $CE(P)$. However there is one problem called "pitch doubling problem". The pitch doubling problem arises because $CE(2P)$ is typically small when $CE(P)$ is small. Therefore, the method based strictly on the minimization of the function $CE(\cdot)$ may choose $2P$ as the pitch even though P is the correct choice. When the pitch doubling problem occurs, there is considerable degradation in the quality of synthesized speech. The pitch doubling problem is avoided by using the method described below. Suppose P' is the value of P that gives rise to the minimum $CE(P)$. Then we consider $P = P', \frac{P'}{2}, \frac{P'}{3}, \frac{P'}{4}, \ldots$ in the allowed range of P (typically $22 \leq P < 115$). If $\frac{P'}{2}, \frac{P'}{3}, \frac{P'}{4}, \ldots$ are not integers, we choose the integers closest to them. Let's suppose $P', \frac{P'}{2} and \frac{P'}{3}$, are in the proper range. We begin with the smallest value of P, in this case $\frac{P'}{3}$, and use the following rule in the order presented.

If

$$CE(\frac{P'}{3}) \leq \alpha_1 \text{ and } \frac{CE(\frac{P'}{3})}{CE(P')} \leq \alpha_2, \text{ then } \hat{P}_F = \frac{P'}{3}. \tag{9}$$

where $\hat{P}_F$ is the estimate from forward look-ahead feature.

If

$$CE(\frac{P'}{3}) \leq \beta_1 \text{ and } \frac{CE(\frac{P'}{3})}{CE(P')} \leq \beta_2, \text{ then } \hat{P}_F = \frac{P'}{3}. \tag{10}$$

Some typical values of $\alpha_1, \alpha_2, \beta_1, \beta_2$ are:

$$\alpha_1 = .15 \qquad \alpha_2 = 5.0$$

$$\beta_1 = .75 \qquad \beta_2 = 2.0$$

If $\frac{P'}{3}$ is not chosen by the above rule, then we go to the next lowest, which is $\frac{P'}{2}$ in the above example. Eventually one will be chosen, or we reach $P = P'$. If $P = P'$ is reached without any choice, then the estimate $\hat{P}_F$ is given by P'.

The final step is to compare $\hat{P}_F$ with the estimate obtained from look-back tracking, P^*. Either $\hat{P}_F$ or P^* is chosen as the initial pitch estimate, $\hat{P}_I$, depending upon the outcome of this decision. One common set of decision rules which is used to compare the two pitch estimates is:

If

$$CE(\hat{P}_F) < E_{-2}(\hat{P}_{-2}) + E_{-1}(\hat{P}_{-1}) + E(P^*) \text{ then } \hat{P}_I = \hat{P}_F \tag{11}$$

Else if

$$CE(\hat{P}_F) \geq E_{-2}(\hat{P}_{-2}) + E_{-1}(\hat{P}_{-1}) + E(P^*) \text{ then } \hat{P}_I = P^* \tag{12}$$

Other decision rules could be used to compare the two candidate pitch values.

The initial pitch estimation method discussed above generates an integer value of pitch. A block diagram of this method is shown in Figure 4. Pitch refinement increases the resolution of the pitch estimate to a higher sub-integer resolution. Typically the refined pitch has a resolution of $\frac{1}{4}$ integer or $\frac{1}{8}$ integer.

We consider a small number (typically 4 to 8) of high resolution values of P near $\hat{P}_I$. We evaluate $E_r(P)$ given by

$$E_r(P) = \int_{\omega=-\pi}^{\pi} G(\omega)|S_w(\omega) - \hat{S}_w(\omega)|^2 d\omega \tag{13}$$

where $G(\omega)$ is an arbitrary weighting function and where

$$S_w(\omega) = \sum_{n=-\infty}^{\infty} s_w(n) e^{-j\omega n} \tag{14}$$

and

$$\hat{S}_w(\omega) = \sum_{m=-\infty}^{\infty} A_M W_r(\omega - m\omega_0) \tag{15}$$

The parameter $\omega_0 = \frac{2\pi}{P}$ is the fundamental frequency and $W_r(\omega)$ is the Fourier Transform of the pitch refinement window, $w_r(n)$ (see Figure 1). The complex coefficients, A_M, in (16), represent the complex amplitudes at the harmonics of ω_0. These coefficients are given by

$$A_M = \frac{\int_{a_M}^{b_M} S_w(\omega) W_r(\omega - m\omega_0) dw}{\int_{a_M}^{b_M} |W_r(\omega - M\omega_0)|^2 dw} \tag{16}$$

where

$$a_M = (m - .5)\omega_0 \text{ and } b_M = (m + .5)\omega_0 \tag{17}$$

The form of $\hat{S}_w(\omega)$ given in (15) corresponds to a voiced or periodic spectrum.

Note that other reasonable error functions can be used in place of (13), for example

$$\hat{E}_r(P) = \int_{-\pi}^{\pi} G(\omega) |S_w(\omega) - \hat{S}_w(\omega)|^2 dw \tag{18}$$

Typically the window function $w_r(n)$ is different from the window function used in the initial pitch estimation step.

An important speech model parameter is the voicing/unvoicing information. This information determines whether the speech is primarily composed of the harmonics of a single fundamental frequency (voiced), or whether it is composed of wideband "noise like" energy (unvoiced). In many previous vocoders, such as Linear Predictive Vocoders or Homomorphic Vocoders, each speech frame is classified as either entirely voiced or entirely unvoiced. In the MBE vocoder the speech spectrum, $S_w(\omega)$, is divided into a number of disjoint frequency bands, and a single voiced/unvoiced (V/UV) decision is made for each band.

The voiced/unvoiced decisions in the MBE vocoder are determined by dividing the frequency range $0 \leq \omega \leq \pi$ into L bands as shown in Figure 5. The constants $\Omega_0 = 0$, Ω_1, ... Ω_{L-1}, $\Omega_L = \pi$, are the boundaries between the L frequency bands.

Within each band a V/UV decision is made by comparing some voicing measure with a known threshold. One common voicing measure is given by

$$D_l = \frac{\int_{\Omega_l}^{\Omega_{l+1}} |S_w(\omega) - \hat{S}_w(\omega)|^2 dw}{\int_{\Omega_l}^{\Omega_{l+1}} |S_w(\omega)|^2 dw} \quad (19)$$

where $\hat{S}_w(\omega)$ is given by Equations (15) through (17). Other voicing measures could be used in place (19). One example of an alternative voicing measure is given by

$$D_l = \frac{\int_{\Omega_l}^{\Omega_{l+1}} ||S_w(\omega)| - |\hat{S}_w(\omega)||^2 dw}{\int_{\Omega_l}^{\Omega_{l+1}} |S_w(\omega)|^2 dw} \quad (20)$$

The voicing measure D_l defined by (19) is the difference between $S_w(\omega)$ and $\hat{S}_w(\omega)$ over the l'th frequency band, which corresponds to $\Omega_l < \omega < \Omega_{l+1}$. D_l is compared against a threshold function. If D_l is less than the threshold function then the l'th frequency band is determined to be voiced. Otherwise the l'th frequency band is determined to be unvoiced. The threshold function typically depends on the pitch, and the center frequency of each band.

In a number of vocoders, including the MBE Vocoder, the Sinusoidal Transform Coder, and the Harmonic Coder the synthesized speech is generated all or in part by the sum of harmonics of a single fundamental frequency. In the MBE vocoder this comprises the voiced portion of the synthesized speech, $v(n)$. The unvoiced portion of the synthesized speech is generated separately and then added to the voiced portion to produce the complete synthesized speech signal.

There are two different techniques which have been used in the past to synthesize a voiced speech signal. The first technique synthesizes each harmonic separately in the time domain using a bank of sinusiodal oscillators. The phase of each oscillator is generated from a low-order piecewise phase polynomial which smoothly interpolates between the estimated parameters. The advantage of this technique is that the resulting speech quality is very high. The disadvantage is that a large number of computations are needed to generate each sinusiodal oscillator. This computational cost of this technique may be prohibitive if a large number of harmonics must be synthesized.

The second technique which has been used in the past to synthesize a voiced speech signal is to synthesize all of the harmonics in the frequency domain, and then to use a Fast Fourier Transform (FFT) to simultaneously convert all of the synthesized harmonics into the time domain. A weighted overlap add method is then used to smoothly interpolate the output of the FFT between speech frames. Since this technique does not require the computations involved with the generation of the sinusoidal oscillators, it is computationally much more efficient than the time-domain technique discussed above. The disadvantage of this technique is that for typical frame rates used in speech coding (20-30 ms.), the voiced speech quality is reduced in comparison with the time-domain technique.

Summary of the Invention

In a first aspect, the invention features an improved pitch estimation method in which sub-integer resolution pitch values are estimated in making the initial pitch estimate. In preferred embodiments, the non-integer values of an intermediate autocorrelation function used for sub-integer resolution pitch values are estimated by interpolating between integer values of the autocorrelation function.

In a second aspect, the invention features the use of pitch regions to reduce the amount of computation required in making the initial pitch estimate. The allowed range of pitch is divided into a plurality of pitch values and a plurality of regions. All regions contain at least one pitch value and at least one region contains a plurality of pitch values. For each region a pitch likelihood function (or error function) is minimized over all pitch values within that region, and the pitch value corresponding to the minimum and the associated value of the error function are stored. The pitch of a current segment is then chosen using look-back tracking, in which the pitch chosen for a current segment is the value that minimizes the error function and is within a first predetermined range of regions above or below the region of a prior segment. Look-ahead tracking can also be used by itself or in conjunction with look-back tracking; the pitch chosen for the current segment is the value that minimizes

a cumulative error function. The cumulative error function provides an estimate of the cumulative error of the current segment and future segments, with the pitches of future segments being constrained to be within a second predetermined range of regions above or below the region of the current segment. The regions can have nonuniform pitch width (i.e., the range of pitches within the regions is not the same size for all regions).

In a third aspect, the invention features an improved pitch estimation method in which pitch-dependent resolution is used in making the initial pitch estimate, with higher resolution being used for some values of pitch (typically smaller values of pitch) than for other values of pitch (typically larger values of pitch).

In a fourth aspect, the invention features improving the accuracy of the voiced/unvoiced decision by making the decision dependent on the energy of the current segment relative to the energy of recent prior segments. If the relative energy is low, the current segment favors an unvoiced decision; if high, the current segment favors a voiced decision.

In a fifth aspect, the invention features an improved method for generating the harmonics used in synthesizing the voiced portion of synthesized speech. Some voiced harmonics (typically low-frequency harmonics) are generated in the time domain, whereas the remaining voiced harmonics are generated in the frequency domain. This preserves much of the computational savings of the frequency domain approach, while it preserves the speech quality of the time domain approach.

In a sixth aspect, the invention features an improved method for generating the voiced harmonics in the frequency domain. Linear frequency scaling is used to shift the frequency of the voiced harmonics, and then an Inverse Discrete Fourier Transform (DFT) is used to convert the frequency scaled harmonics into the time domain. Interpolation and time scaling are then used to correct for the effect of the linear frequency scaling. This technique has the advantage of improved frequency accuracy.

Other features and advantages of the invention will be apparent from the following

description of preferred embodiments and from the claims.

Brief Description of the Drawings

FIGS. 1-5 are diagrams showing prior art pitch estimation methods.

FIG. 6 is a flow chart showing a preferred embodiment of the invention in which sub-integer resolution pitch values are estimated

FIG. 7 is a flow chart showing a preferred embodiment of the invention in which pitch regions are used in making the pitch estimate.

FIG. 8 is a flow chart showing a preferred embodiment of the invention in which pitch-dependent resolution is used in making the pitch estimate.

FIG. 9 is a flow chart showing a preferred embodiment of the invention in which the voiced/unvoiced decision is made dependent on the relative energy of the current segment and recent prior segments.

FIG 10 is a block diagram showing a preferred embodiment of the invention in which a hybrid time and frequency domain synthesis method is used.

FIG 11 is a block diagram showing a preferred embodiment of the invention in which a modified frequency domain synthesis is used.

Description of Preferred Embodiments of the Invention

In the prior art, the initial pitch estimate is estimated with integer resolution. The performance of the method can be improved significantly by using sub-integer resolution (e.g. the resolution of $\frac{1}{2}$ integer). This requires modification of the method. If $E(P)$ in Equation (1) is used as an error criterion, for example, evaluation of $E(P)$ for non-integer P requires evaluation of $r(n)$ in (2) for non-integer values of n. This can be accomplished by

$$r(n+d) = (1-d) \cdot r(n) + d \cdot r(n+1) \quad \text{for} \quad 0 \leq d \leq 1 \tag{21}$$

Equation (21) is a simple linear interpolation equation; however, other forms of interpolation could be used instead of linear interpolation. The intention is to require the

initial pitch estimate to have sub-integer resolution, and to use (21) for the calculation of $E(P)$ in (1). This procedure is sketched in Figure 6.

In the initial pitch estimate, prior techniques typically consider approximately 100 different values ($22 \leq P < 115$) of P. If we allow sub-integer resolution, say $\frac{1}{2}$ integer, then we have to consider 186 different values of P. This requires a great deal of computation, particularly in the look-ahead tracking. To reduce computations, we can divide the allowed range of P into a small number of non-uniform regions. A reasonable number is 20. An example of twenty non-uniform regions is as follows:

Region 1:	$22 \leq P < 24$
Region 2:	$24 \leq P < 26$
Region 3:	$26 \leq P < 28$
Region 4:	$28 \leq P < 31$
Region 5:	$31 \leq P < 34$
⋮	⋮
Region 19:	$99 \leq P < 107$
Region 20:	$107 \leq P < 115$

Within each region, we keep the value of P for which $E(P)$ is minimum and the corresponding value of $E(P)$. All other information concerning $E(P)$ is discarded. The pitch tracking method (look-back and look-ahead) uses these values to determine the initial pitch estimate, $\hat{P}_I$. The pitch continuity constraints are modified such that the pitch can only change by a fixed number of regions in either the look-back tracking or look-ahead tracking.

For example if $P_{-1} = 26$, which is in pitch region 3, then P may be constrained to lie in pitch region 2, 3 or 4. This would correspond to an allowable pitch difference of 1 region in the "look-back" pitch tracking.

Similarly, if $P = 26$, which is in pitch region 3, then P_1 may be constrained to lie in pitch region 1, 2, 3, 4 or 5. This would correspond to an allowable pitch difference of 2 regions in the "look-ahead" pitch tracking. Note how the allowable pitch difference may be different for the "look-ahead" tracking than it is for the "look-back" tracking.

The reduction of from approximately 200 values of P to approximately 20 regions reduces the computational requirements for the look-ahead pitch tracking by orders of magnitude with little difference in performance. In addition the storage requirements are reduced, since $E(P)$ only needs to be stored at 20 different values of P_1 rather than 100-200.

Further substantial reduction in the number of regions will reduce computations but will also degrade the performance. If two candidate pitches fall in the same region, for example, the choice between the two will be strictly a function of which results in a lower $E(P)$. In this case the benefits of pitch tracking will be lost. Figure 7 shows a flow chart of the pitch estimation method which uses pitch regions to estimate the initial pitch.

In various vocoders such as MBE and LPC, the pitch estimated has a fixed resolution, for example integer sample resolution or $\frac{1}{2}$-sample resolution. The fundamental frequency, ω_0, is inversely related to the pitch P, and therefore a fixed pitch resolution corresponds to much less fundamental frequency resolution for small P than it does for large P. Varying the resolution of P as a function of P can improve the system performance, by removing some of the pitch dependency of the fundamental frequency resolution. Typically this is accomplished by using higher pitch resolution for small values of P than for larger values of P. For example the function, $E(P)$, can be evaluated with half-sample resolution for pitch values in the range $22 \leq P < 60$, and with integer sample resolution for pitch values in the range $60 \leq P < 115$. Another example would be to evaluate $E(P)$ with half sample resolution in the range $22 \leq P < 40$, to evaluate $E(P)$ with integer sample resolution for the range $42 \leq P < 80$, and to evaluate $E(P)$ with resolution 2 (i.e. only for even values of P) for the range $80 \leq P < 115$. The invention has the advantage that $E(P)$ is evaluated with more resolution only for the values of P which are most sensitive to the pitch doubling problem, thereby saving computation. Figure 8 shows a flow chart of the pitch estimation method which uses pitch dependent resolution.

The method of pitch-dependent resolution can be combined with the pitch estimation method using pitch regions. The pitch tracking method based on pitch regions is modified to evaluate $E(P)$ at the correct resolution (i.e. pitch dependent), when finding the minimum value of $E(P)$ within each region.

In prior vocoder implementations, the V/UV decision for each frequency band is made by comparing some measure of the difference between $S_w(\omega)$ and $\hat{S}_w(\omega)$ with some threshold. The threshold is typically a function of the pitch P and the frequencies in the band. The performance can be improved considerably by using a threshold which is a function of not only the pitch P and the frequencies in the band but also the energy of the signal (as shown in Figure 9). By tracking the signal energy, we can estimate the signal energy in the current frame relative to the recent past history. If the relative energy is low, then the signal is more likely to be unvoiced, and therefore the threshold is adjusted to give a biased decision favoring unvoicing. If the relative energy is high, the signal is likely to be voiced, and therefore the threshold is adjusted to give a biased decision favoring voicing. The energy dependent voicing threshold is implemented as follows. Let ξ_0 be an energy measure which is calculated as follows,

$$\xi_0 = \int_{-\pi}^{\pi} H(\omega)|S_w(\omega)|^2 dw \tag{22}$$

where $S_w(\omega)$ is defined in (14), and $H(\omega)$ is a frequency dependent weighting function. Various other energy measures could be used in place of (22), for example,

$$\xi_0 = \int_{-\pi}^{\pi} H(\omega)|S_w(\omega)|dw \tag{23}$$

The intention is to use a measure which registers the relative intensity of each speech segment.

Three quantities, roughly corresponding to the average local energy, maximum local energy, and minimum local energy, are updated each speech frame according to the following rules:

$$\xi_{avg} = (1-\gamma_0)\xi_{avg} + \gamma_0 \cdot \xi_0 \tag{24}$$

$$\xi_{max} = \begin{cases} (1-\gamma_1)\xi_{max} + \gamma_1 \cdot \xi_0 & \text{if } \xi_0 > \xi_{max} \\ (1-\gamma_2)\xi_{max} + \gamma_2 \cdot \xi_0 & \text{if } \xi_0 \leq \xi_{max} \end{cases} \quad (25)$$

$$\xi_{min} = \begin{cases} (1-\gamma_3)\xi_{min} + \gamma_3 \cdot \xi_0 & \text{if } \xi_0 \leq \xi_{min} \\ (1-\gamma_4)\xi_{min} + \gamma_4 \cdot \xi_0 & \text{if } \xi_{min} \leq \xi_0 < \mu \cdot \xi_{min} \\ (1+\gamma_4)\xi_{min} & \text{if } \xi_0 > \mu \cdot \xi_{min} \end{cases} \quad (26)$$

For the first speech frame, the values of ξ_{avg}, ξ_{max}, and ξ_{min} are initialized to some arbitrary positive number. The constants $\gamma_0, \gamma_1, \ldots \gamma_4$, and μ control the adaptivity of the method. Typical values would be:

$$\gamma_0 = .067$$
$$\gamma_1 = .5$$
$$\gamma_2 = .01$$
$$\gamma_3 = .5$$
$$\gamma_4 = .025$$
$$\mu = 2.0$$

The functions in (24) (25) and (26) are only examples, and other functions may also be possible. The values of $\xi_0, \xi_{avg}, \xi_{min}$ and ξ_{max} affect the V/UV threshold function as follows. Let $T(P,\omega)$ be a pitch and frequency dependent threshold. We define the new energy dependent threshold, $T_\xi(P,W)$, by

$$T_\xi(P,\omega) = T(P,\omega) \cdot M(\xi_o, \xi_{avg}, \xi_{min}, \xi_{max}) \quad (27)$$

where $M(\xi_0, \xi_{avg}, \xi_{min}, \xi_{max})$ is given by

$$M(\xi_0, \xi_{avg}, \xi_{min}, \xi_{max}) = \begin{cases} \lambda_0 \text{ , if } \xi_{avg} \leq \xi_{silence} & \\ \frac{(\xi_{min}+\xi_0)(\xi_{max}+\lambda_1\xi_0)}{(\xi_0+\lambda_2\xi_{max})(\xi_0+\xi_{max})} \text{ , } & \text{if } \xi_{avg} > \xi_{silence} \\ & \text{and } \xi_{min} \geq \lambda_2\xi_{max} \\ 1 \text{ , } & \text{if } \xi_{avg} > \xi_{silence} \\ & \text{and } \xi_{min} \leq \lambda_2\xi_{max} \end{cases} \quad (28)$$

Typical values of the constants $\lambda_0, \lambda_1, \lambda_2$ and $\xi_{silence}$ are:

$$\lambda_0 = .5$$

$$\lambda_1 = 2.0$$

$$\lambda_2 = .0075$$

$$\xi_{silence} = 200.0$$

The V/UV information is determined by comparing D_l, defined in (19), with the energy dependent threshold, $T_\xi(\hat{P}, \frac{\Omega_l+\Omega_{l+1}}{2})$. If D_l is less than the threshold then the l'th frequency band is determined to be voiced. Otherwise the l'th frequency band is determined to be unvoiced.

$T(P,\omega)$ in Equation (27) can be modified to include dependence on variables other than just pitch and frequency without effecting this aspect of the invention. In addition, the pitch dependence and/or the frequency dependence of $T(P,\omega)$ can be eliminated (in its simplist form $T(P,\omega)$ can equal a constant) without effecting this aspect of the invention.

In another aspect of the invention, a new hybrid voiced speech synthesis method combines the advantages of both the time domain and frequency domain methods used previously. We have discovered that if the time domain method is used for a small number of low-frequency harmonics, and the frequency domain method is used for the remaining harmonics there is little loss in speech quality. Since only a small number of harmonics are generated with the time domain method, our new method preserves much of the computational savings of the total frequency domain approach. The hybrid voiced speech synthesis method is shown in Figure 10

Our new hybrid voiced speech synthesis method operates in the following manner. The voiced speech signal, $v(n)$, is synthesized according to

$$v(n) = v_1(n) + v_2(n) \tag{29}$$

where $v_1(n)$ is a low frequency component generated with a time domain voiced synthesis method, and $v_2(n)$ is a high frequency component generated with a frequency

domain synthesis method.

Typically the low frequency component, $v_1(n)$, is synthesized by,

$$\sum_{k=1}^{K} a_k(n) \cos \Theta_k(n) \tag{30}$$

where $a_k(n)$ is a piecewise linear polynomial, and $\Theta_k(n)$ is a low-order piecewise phase polynomial. The value of K in Equation (30) controls the maximum number of harmonics which are synthesized in the time domain. We typically use a value of K in the range $4 \leq K \leq 12$. Any remaining high frequency voiced harmonics are synthesized using a frequency domain voiced synthesis method.

In another aspect of the invention, we have developed a new frequency domain sythesis method which is more efficient and has better frequency accuracy than the frequency domain method of McAulay and Quatieri. In our new method the voiced harmonics are linearly frequency scaled according to the mapping $\omega_0 \rightarrow \frac{2\pi}{L}$, where L is a small integer (typically $L < 1000$). This linear frequency scaling shifts the frequency of the k'th harmonic from a frequency $\omega_k = k \cdot \omega_0$, where ω_0 is the fundamental frequency, to a new frequency $\frac{2\pi k}{L}$. Since the frequencies $\frac{2\pi k}{L}$ correspond to the sample frequencies of an L-point Discrete Fourier Transform (DFT), an L-point Inverse DFT can be used to simultaneously transform all of the mapped harmonics into the time domain signal, $\hat{v}_2(n)$. A number of efficient algorithms exist for computing the Inverse DFT. Some examples include the Fast Fourier Transform (FFT), the Winograd Fourier Transform and the Prime Factor Algorithm. Each of these algorithms places different constraints on the allowable values of L. For example the FFT requires L to be a highly composite number such as 2^7, 3^5, $2^4 \cdot 3^2$, etc... .

Because of the linear frequency scaling, $\hat{v}_2(n)$ is a time scaled version of the desired signal, $v_2(n)$. Therefore $v_2(n)$ can be recovered from $\hat{v}_2(n)$ through equations (31)-(33) which correspond to linear interpolation and time scaling of $\hat{v}_2(n)$

$$v_2(n) = (1 - \delta_n)\hat{v}(m_n) + \delta_n \cdot \hat{v}(m_n + 1) \tag{31}$$

$$m_n = \lfloor \frac{\omega_0 L n}{2\pi} \rfloor \quad \text{where } \lfloor x \rfloor = \text{the smallest integer} \leq x \tag{32}$$

$$\delta_n = \frac{\omega_0 L n}{2\pi} - m_n \tag{33}$$

Other forms of interpolation could be used in place of linear interpolation. This procedure is sketched in Figure 11.

Other embodiments of the invention are within the following claims. Error function as used in the claims has a broad meaning and includes pitch likelihood functions.

Claims

1. A method for estimating the pitch of individual segments of speech, said pitch estimation method comprising the steps of:

dividing the allowable range of pitch into a plurality of pitch values with sub-integer resolution;

evaluating an error function for each of said pitch values, said error function providing a numerical means for comparing the said pitch values for the current segment; and

using look-back tracking to choose for the current segment a pitch value that reduces said error function within a first predetermined range above or below the pitch of a prior segment.

2. A method for estimating the pitch of individual segments of speech, said pitch estimation method comprising the steps of:

dividing the allowable range of pitch into a plurality of pitch values with sub-integer resolution;

evaluating an error function for each of said pitch values, said error function providing a numerical means for comparing the said pitch values for the current segment; and

using look-ahead tracking to choose for the current speech segment a value of pitch that reduces a cumulative error function, said cumulative error function providing an estimate of the cumulative error of the current segment and future segments as a function of the current pitch, the pitch of future segments being constrained to be within a second predetermined range of the pitch of the preceding segment.

3. The method of claim 1 further comprising the steps of:

using look-ahead tracking to choose for the current speech segment a value of pitch that reduces a cumulative error function, said cumulative error function providing an estimate of the cumulative error of the current segment and future segments as a function of the current pitch, the pitch of future segments being constrained to be

within a second predetermined range of the pitch of the preceding segment; and

deciding to use as the pitch of the current segment either the pitch chosen with look-back tracking or the pitch chosen with look-ahead tracking.

4. The method of claim 3 wherein the pitch of the current segment is equal to the pitch chosen with look-back tracking if the sum of the errors (derived from the error function used for look-back tracking) for the current segment and selected prior segments is less than a predetermined threshold; otherwise the pitch of the current segment is equal to the pitch chosen with look-back tracking if the sum of the errors (derived from the error function used for look-back tracking) for the current segment and selected prior segments is less than the cumulative error (derived from the cumulative error function used for look-ahead tracking); otherwise the pitch of the current segment is equal to the pitch chosen with look-ahead tracking.

5. The method of claim 1, 2 or 3 wherein the pitch is chosen to minimize said error function or cumulative error function.

6. The method of claim 1, 2 or 3 wherein the said error function or cumulative error function is dependent on an autocorrelation function.

7. The method of claim 1, 2 or 3 wherein the error function is that shown in equations (1), (2) and (3).

8. The method of claim 6 wherein said autocorrelation function for non-integer values is estimated by interpolating between integer values of said autocorrelation function.

9. The method of claim 7 wherein r(n) for non-integer values is estimated by interpolating between integer values of r(n).

10. The method of claim 9 wherein the interpolation is performed using the expression of equation (21).

11. The method of claim 1, 2 or 3 comprising the further step of refining the pitch estimate.

12. A method for estimating the pitch of individual segments of speech, said pitch

estimation method comprising the steps of:

dividing the allowed range of pitch into a plurality of pitch values;

dividing the allowed range of pitch into a plurality of regions, all regions containing at least one of said pitch values and at least one region containing a plurality of said pitch values;

evaluating an error function for each of said pitch values, said error function providing a numerical means for comparing the said pitch values for the current segment;

finding for each region the pitch that generally minimizes said error function over all pitch values within that region and storing the associated value of said error function within that region; and

using look-back tracking to choose for the current segment a pitch that generally minimizes said error function and is within a first predetermined range of regions above or below the region containing the pitch of the prior segment.

13. A method for estimating the pitch of individual segments of speech, said pitch estimation method comprising the steps of:

dividing the allowed range of pitch into a plurality of pitch values;

dividing the allowed range of pitch into a plurality of regions, all regions containing at least one of said pitch values and at least one region containing a plurality of said pitch values;

evaluating an error function for each of said pitch values, said error function providing a numerical means for comparing the said pitch values for the current segment;

finding for each region the pitch that generally minimizes said error function over all pitch values within that region and storing the associated value of said error function within that region; and

using look-ahead tracking to choose for the current segment a pitch that generally minimizes a cumulative error function, said cumulative error function providing an

estimate of the cumulative error of the current segment and future segments as a function of the current pitch, the pitch of future segments being constrained to be within a second predetermined range of regions above or below the region containing the pitch of the preceding segment.

14. The method of claim 12 further comprising the steps of:

using look-ahead tracking to choose for the current segment a pitch that generally minimizes a cumulative error function, said cumulative error function providing an estimate of the cumulative error of the current segment and future segments as a function of the current pitch, the pitch of future segments being constrained to be within a second predetermined range of regions above or below the region containing the pitch of the preceding segment; and

deciding to use as the pitch of the current segment either the pitch chosen with look-back tracking or the pitch chosen with look-ahead tracking.

15. The method of claim 14 wherein the pitch of the current segment is equal to the pitch chosen with look-back tracking if the sum of the errors (derived from the error function used for look-back tracking) for the current segment and selected prior segments is less than a predetermined threshold; otherwise the pitch of the current segment is equal to the pitch chosen with look-back tracking if the sum of the errors (derived from the error function used for look-back tracking) for the current segment and selected prior segments is less than the cumulative error (derived from the cumulative error function used for look-ahead tracking); otherwise the pitch of the current segment is equal to the pitch chosen with look-ahead tracking.

16. The method of claim 14 or 15 wherein the first and second ranges extend across different numbers of regions.

17. The method of claim 12, 13 or 14 wherein the number of pitch values within each region varies between regions.

18. The method of claim 12, 13 or 14 comprising the further step of refining the pitch estimate.

19. The method of claim 12, 13 or 14 wherein the allowable range of pitch is divided into a plurality of pitch values with sub-integer resolution.

20. The method of claim 19 wherein the said error function or cumulative error function is dependent on an autocorrelation function; said autocorrelation function being estimated for non-integer values by interpolating between integer values of said autocorrelation function.

21. The method of claim 12, 13 or 14 wherein the allowed range of pitch is divided into a plurality of pitch values using pitch dependent resolution.

22. The method of claim 21 wherein smaller values of said pitch values have higher resolution.

23. The method of claim 22 wherein smaller values of said pitch values have sub-integer resolution

24. The method of claim 22 wherein larger values of said pitch values have greater than integer resolution.

25. A method for estimating the pitch of individual segments of speech, said pitch estimation method comprising the steps of:

dividing the allowable range of pitch into a plurality of pitch values using pitch dependent resolution;

evaluating an error function for each of said pitch values, said error function providing a numerical means for comparing the said pitch values for the current segment; and

choosing for the pitch of the current segment a pitch value that reduces said error function.

26. A method for estimating the pitch of individual segments of speech, said pitch estimation method comprising the steps of:

dividing the allowable range of pitch into a plurality of pitch values using pitch dependent resolution;

evaluating an error function for each of said pitch values, said error function

providing a numerical means for comparing the said pitch values for the current segment; and

using look-back tracking to choose for the current segment a pitch value that reduces said error function within a first predetermined range above or below the pitch of a prior segment.

27. A method for estimating the pitch of individual segments of speech, said pitch estimation method comprising the steps of:

dividing the allowable range of pitch into a plurality of pitch values using pitch dependent resolution;

evaluating an error function for each of said pitch values, said error function providing a numerical means for comparing the said pitch values for the current segment; and

using look-ahead tracking to choose for the current speech segment a value of pitch that reduces a cumulative error function, said cumulative error function providing an estimate of the cumulative error of the current segment and future segments as a function of the current pitch, the pitch of future segments being constrained to be within a second predetermined range of the pitch of the preceding segment;

28. The method of claim 26 further comprising the steps of:

using look-ahead tracking to choose for the current speech segment a value of pitch that reduces a cumulative error function, said cumulative error function providing an estimate of the cumulative error of the current segment and future segments as a function of the current pitch, the pitch of future segments being constrained to be within a second predetermined range of the pitch of the preceding segment;

deciding to use as the pitch of the current segment either the pitch chosen with look-back tracking or the pitch chosen with look-ahead tracking.

29. The method of claim 28 wherein the pitch of the current segment is equal to the pitch chosen with look-back tracking if the sum of the errors (derived from the error function used for look-back tracking) for the current segment and selected

prior segments is less than a predetermined threshold; otherwise the pitch of the current segment is equal to the pitch chosen with look-back tracking if the sum of the errors (derived from the error function used for look-back tracking) for the current segment and selected prior segments is less than the cumulative error (derived from the cumulative error function used for look-ahead tracking); otherwise the pitch of the current segment is equal to the pitch chosen with look-ahead tracking.

30. The method of claim 25, 26, 27 or 28 wherein a pitch is chosen to minimize said error function or cumulative error function.

31. The method of claim 25, 26, 27 or 28 wherein higher resolution is used for smaller values of pitch.

32. The method of claim 31 wherein smaller values of said pitch values have sub-integer resolution

33. The method of claim 31 wherein larger values of said pitch values have greater than integer resolution.

34. A method for making the voiced/unvoiced decision for a particular frequency band, the method comprising the steps of:

evaluating a voicing measure for said frequency band;

making the voiced/unvoiced decision for said frequency band based upon a comparison between the voicing measure and a threshold;

determining an energy measure of the current segment, and comparing it to the signal energy of one or more recent prior segments; and

adjusting the threshold to make a voiced decision more likely when the energy of the current segment is relatively high compared to the energy of the recent prior segments.

35. A method for making the voiced/unvoiced decision for a particular frequency band, the method comprising the steps of:

evaluating a voicing measure for said frequency band;

making the voiced/unvoiced decision for said frequency band based upon a com-

parison between the voicing measure and a threshold;

determining an energy measure of the current segment, and comparing it to the signal energy of one or more recent prior segments; and

adjusting the threshold to make an unvoiced decision more likely when the energy of the current segment is relatively low compared to the energy of the recent prior segments.

36. The method of claim 34 comprising the further step of adjusting the threshold to make a voiced decision more likely when the energy of the current segment is relatively high compared to the energy of the recent prior segments.

37. The method of claim 34, 35 or 36 wherein the energy measure is that shown in Equation (21).

38. The method of claim 34, 35 or 36 wherein the voicing measure is that shown in Equation (19).

39. The method of claim 34, 35 or 36 wherein the energy dependence of the said threshold is that shown in Equations (24), (25), (26), (27) and (28).

40. A method for generating the harmonics for use in synthesizing the voiced portion of synthesized speech, the method comprising the steps of:

generating some voiced harmonics using a time domain synthesis method; and

generating other harmonics with a frequency domain synthesis method.

41. The method of claim 40 wherein low-frequency harmonics are generated using a time domain synthesis method.

42. The method of claim 40 or 41 wherein high-frequency harmonics are generated using a frequency domain synthesis method.

43. The method of claim 40 wherein said time domain synthesis is performed by generating a low-order piecewise phase polynomial.

44. The method of claim 42 wherein said time domain synthesis is performed by generating a low-order piecewise phase polynomial.

45. The method of claim 42 wherein said harmonics generated in the frequency

domain are generated using the method comprising the steps of:

linearly frequency scaling the voiced harmonics according to the mapping $\hat{\omega}_0 \rightarrow \frac{2\pi}{L}$, where L is some small integer;

performing an L-point Inverse Discrete Fourier Transform (DFT) to simultaneously transform the frequency scaled harmonics into the time domain; and

performing interpolation and time scaling to generate the output.

46. A method for generating the harmonics for use in synthesizing the voiced portion of synthesized speech, the method comprising the steps of:

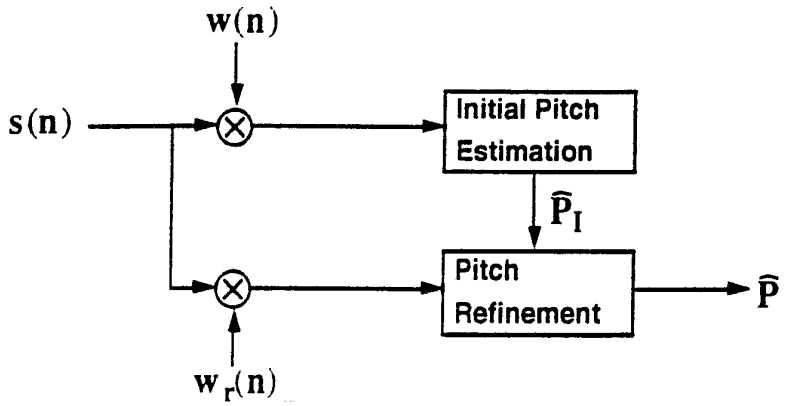
linearly frequency scaling the voiced harmonics according to the mapping $\hat{\omega}_0 \rightarrow \frac{2\pi}{L}$, where L is some small integer;

performing an L-point Inverse Discrete Fourier Transform (DFT) to simultaneously transform the frequency scaled harmonics into the time domain; and

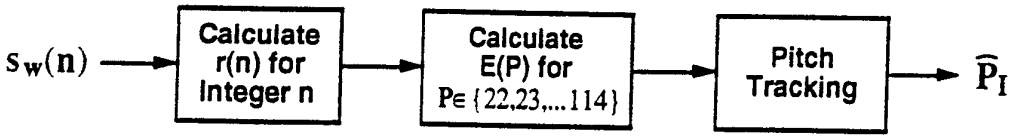
performing interpolation and time scaling to generate the output.

47. The method of claim 45 or 46 wherein said DFT is computed with a Fast Fourier Transform, and L is a highly composite number.

48. The method of claim 45 or 46 wherein said interpolation is performed with linear interpolation.

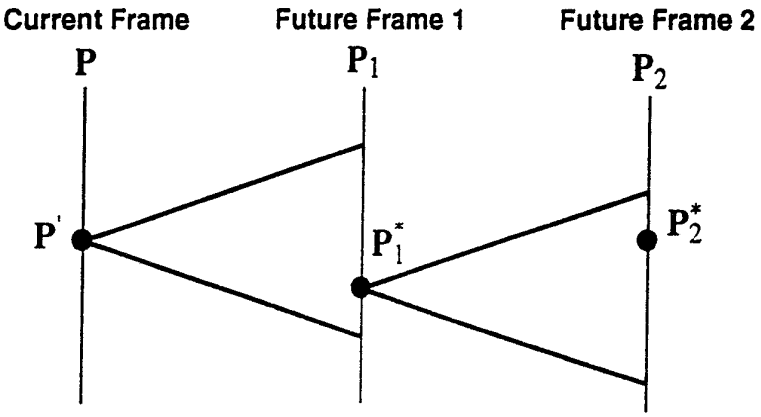


PRIOR ART
FIG. 1

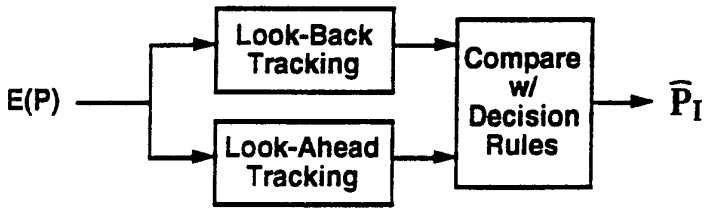


PRIOR ART

FIG. 2

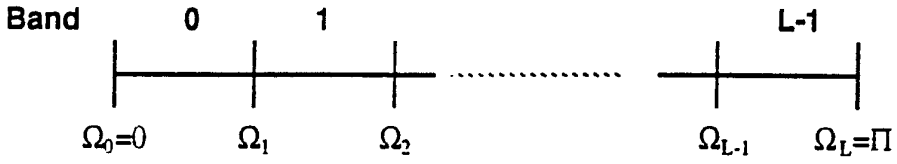


PRIOR ART
FIG. 3

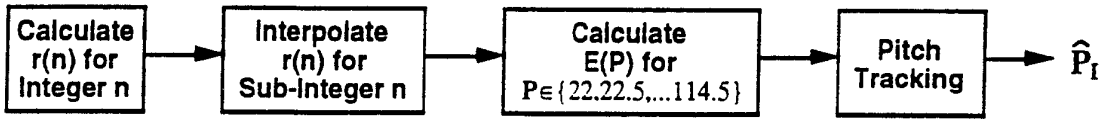


PRIOR ART

FIG. 4



PRIOR ART
FIG. 5



PRIOR ART

FIG. 6

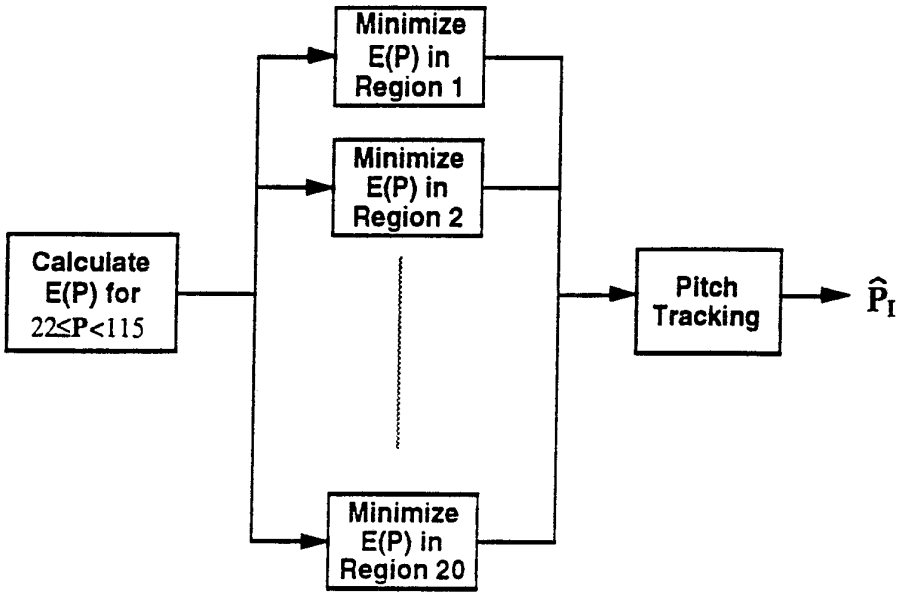

Minimize E(P) in Region 1
Minimize E(P) in Region 2
Calculate E(P) for 22≤P<115
Pitch Tracking
$\hat{P}_I$
Minimize E(P) in Region 20

FIG. 7

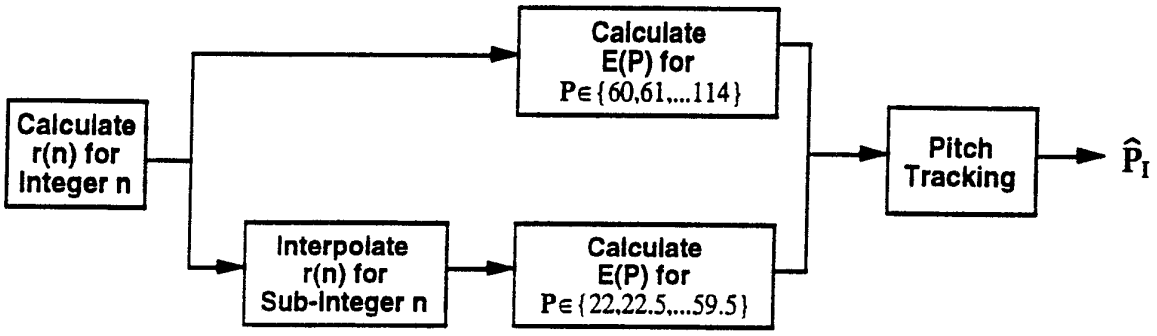

Calculate r(n) for Integer n
Calculate E(P) for P∈{60,61,....114}
Interpolate r(n) for Sub-Integer n
Calculate E(P) for P∈{22.22.5,....59.5}
Pitch Tracking
$\hat{P}_I$

FIG. 8

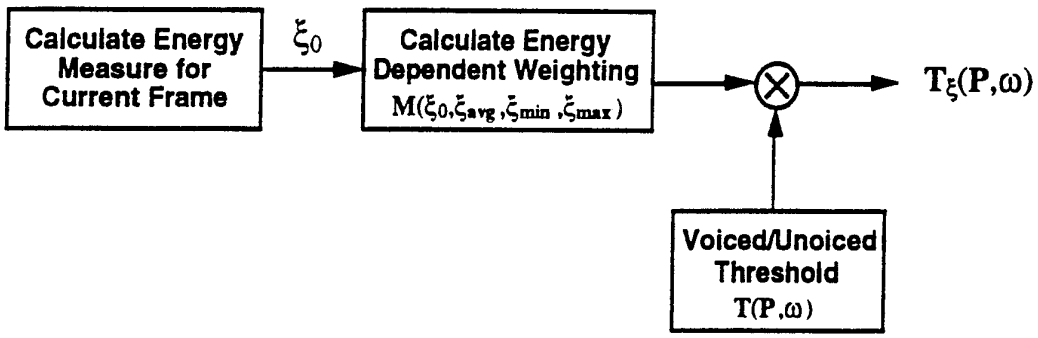

Calculate Energy Measure for Current Frame
ξ_0
Calculate Energy Dependent Weighting
$M(\xi_0,\xi_{avg},\xi_{min},\xi_{max})$
$T_\xi(P,\omega)$
Voiced/Unoiced Threshold
$T(P,\omega)$

FIG. 9

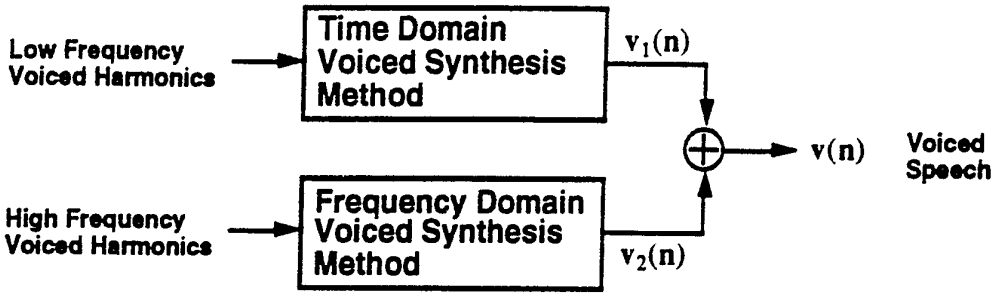

Low Frequency
Voiced Harmonics
Time Domain
Voiced Synthesis
Method
$v_1(n)$
$v(n)$
Voiced
Speech
High Frequency
Voiced Harmonics
Frequency Domain
Voiced Synthesis
Method
$v_2(n)$

FIG. 10

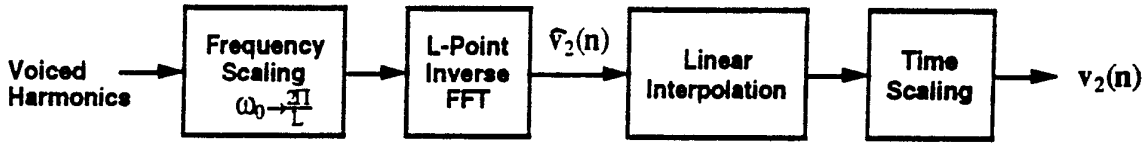

Voiced Harmonics
Frequency Scaling
$\omega_0 \rightarrow \frac{2\pi}{L}$
L-Point Inverse FFT
$\hat{v}_2(n)$
Linear Interpolation
Time Scaling
$v_2(n)$

FIG. 11

INTERNATIONAL SEARCH REPORT

International Application No. PCT/US91/06853

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all) [6]

According to International Patent Classification (IPC) or to both National Classification and IPC

IPC(5): G10L 7/02
US CL : 381/49

II. FIELDS SEARCHED

Minimum Documentation Searched [7]

Classification System	Classification Symbols
US	381/36-41,47,49

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched [8]

III. DOCUMENTS CONSIDERED TO BE RELEVANT [9]

Category *	Citation of Document, [11] with indication, where appropriate, of the relevant passages [12]	Relevant to Claim No. [13]
A,P	US, A, 4,989,247 (VAN HEMERT) 29 JANUARY 1991 See abstract.	1-33
A	US, A, 4,282,405 (TAGUCHI) 04 AUGUST 1981 See abstract.	1-33
A	US, A, 4,696,038 (DODDINGTON ET AL.) 22 SEPTEMBER 1987 See abstract.	1-33
A	GRIFFIN ET AL., "MULTIBAND EXCITATION VOCODER" vol. 36, no. 8., IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, AUGUST 1988, PGS. 1223-1235.	1-33

* Special categories of cited documents: [10]

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search

02 DECEMBER 1991

Date of Mailing of this International Search Report

10 JAN 1992

International Searching Authority

ISA/US

Signature of Authorized Officer

MICHELLE DOERRLER

FURTHER INFORMATION CONTINUED FROM THE SECOND SHEET

V ☐ OBSERVATIONS WHERE CERTAIN CLAIMS WERE FOUND UNSEARCHABLE [1]

This international search report has not been established in respect of certain claims under Article 17(2) (a) for the following reasons:

1. ☐ Claim numbers because they relate to subject matter [2] not required to be searched by this Authority, namely:

2. ☐ Claim numbers , because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out [3], specifically:

3. ☐ Claim numbers_____, because they are dependent claims not drafted in accordance with the second and third sentences of PCT Rule 6.4(a).

VI. ☒ OBSERVATIONS WHERE UNITY OF INVENTION IS LACKING [2]

This International Searching Authority found multiple inventions in this international application as follows:

I. Claims 1-33 are drawn to a method for estimating pitch of individual segments of speech in class 381 subclass 49.

II. ClAims 34-39 are drawn to a method for making voiced/unvoiced decisions in class 381 subclass 46.

III. Claims 40-48 are drawn to amethod for generating harmonies for use

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims of the international application.

2. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims of the international application for which fees were paid, specifically claims:

3. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claim numbers:

1-33

4. ☐ As all searchable claims could be searched without effort justifying an additional fee, the International Searching Authority did not invite payment of any additional fee.

Remark on Protest

☐ The additional search fees were accompanied by applicant's protest.

☐ No protest accompanied the payment of additional search fees.