Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2169822 C 2006/01/10

(11)(21) 2 169 822

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(22) Date de dépôt/Filing Date: 1996/02/19

(41) Mise à la disp. pub./Open to Public Insp.: 1996/08/23

(45) Date de délivrance/Issue Date: 2006/01/10

(30) Priorité/Priority: 1995/02/22 (08/392,099) US

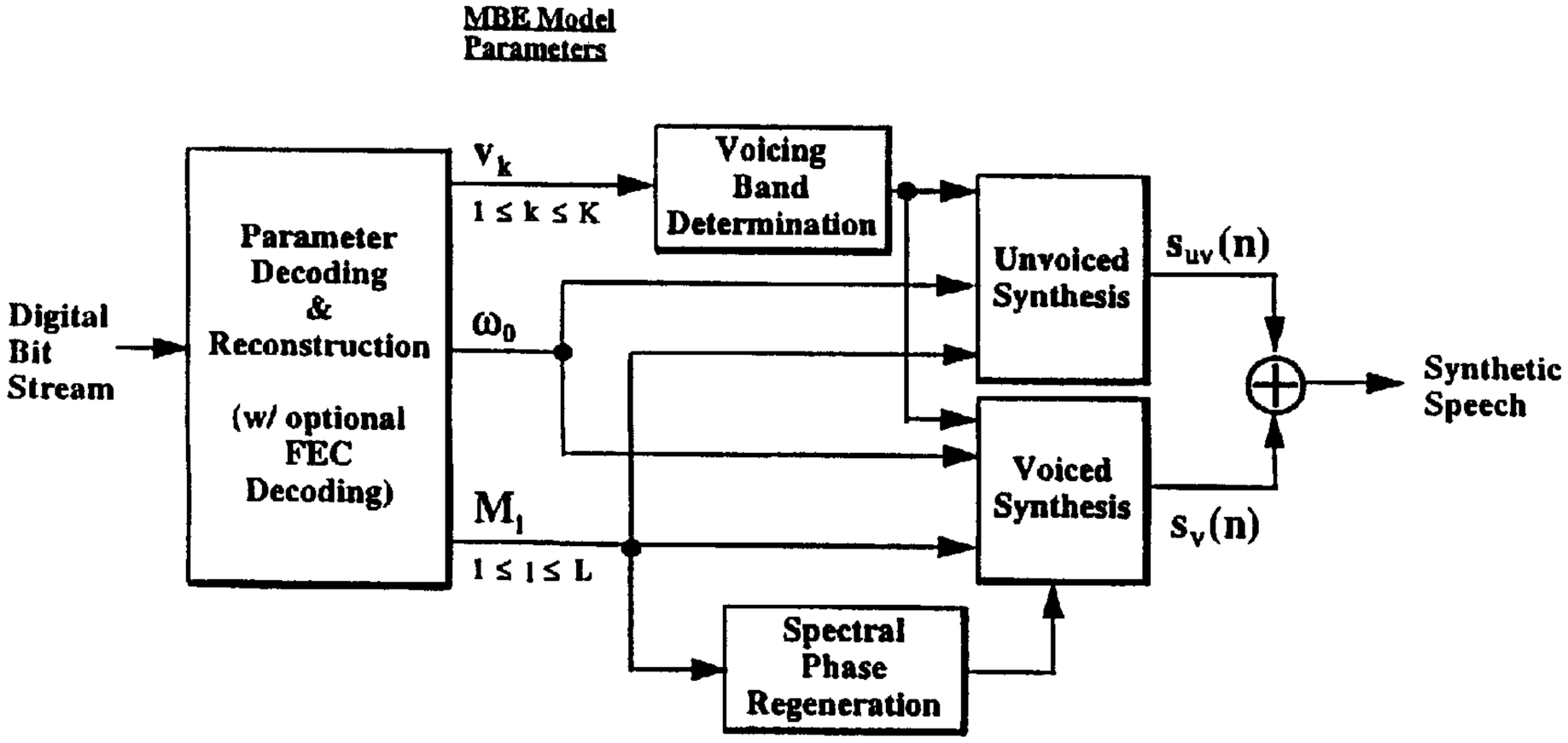
(51) Cl.Int.^{7}/Int.Cl.^{7} G10L 13/04, G10L 19/06, G10L 11/06

(72) Inventeurs/Inventors:
GRIFFIN, DANIEL W., US;
HARDWICK, JOHN C., US

(73) Propriétaire/Owner:
DIGITAL VOICE SYSTEMS, INC., US

(74) Agent: SMART & BIGGAR

(54) Titre : SYNTHESE VOCALE UTILISANT DES INFORMATIONS DE PHASE REGENEREES
(54) Title: SYNTHESIS OF SPEECH USING REGENERATED PHASE INFORMATION



(57) Abrégé/Abstract:
The spectral magnitude and phase representation used in Multi-Band Excitation (MBE) based speech coding systems is improved. At the encoder the digital speech signal is divided into frames, and a fundamental frequency, voicing information, and a set of spectral magnitudes are estimated for each frame. A spectral magnitude is computed at each harmonic frequency (ie. multiples of the estimated fundamental frequency) using a new estimation method which is independent of voicing state and which corrects for any offset between the harmonic and the frequency sampling grid. The result is a fast, FFT compatible method which produces a smooth set of spectral magnitudes without the sharp discontinuities introduced by voicing transitions as found in prior MBE based speech coders. Quantization efficiency is thereby improved, producing higher speech quality at lower bit rates. In addition, smoothing methods, typically used to reduce the effect of bit errors or to enhance formants, are more effective since they are not confused by false edges (i.e. discontinuities) at voicing transitions. Overall speech quality and intelligibility are improved. At the decoder a bit stream is received and then used to reconstruct a fundamental frequency, voicing information, and a set of spectral magnitudes for a sequence of frames. The voicing information is used to label each harmonic as either voiced or unvoiced, and for voiced harmonics an individual phase is regenerated as a function of the spectral magnitudes localized about that harmonic frequency. The decoder then synthesizes the voiced and unvoiced component and adds them to produce the synthesized speech. The regenerated phase more closely approximates actual speech in terms of peak-to-rms value relative to the prior art, thereby yielding improved dynamic range. In addition the synthesized speech is perceived as more natural and exhibits fewer phase related distortions.

Synthesis of Speech Using Regenerated Phase Information

Abstract

The spectral magnitude and phase representation used in Multi-Band Excitation (MBE) based speech coding systems is improved. At the encoder the digital speech signal is divided into frames, and a fundamental frequency, voicing information, and a set of spectral magnitudes are estimated for each frame. A spectral magnitude is computed at each harmonic frequency (ie. multiples of the estimated fundamental frequency) using a new estimation method which is independent of voicing state and which corrects for any offset between the harmonic and the frequency sampling grid. The result is a fast, FFT compatible method which produces a smooth set of spectral magnitudes without the sharp discontinuities introduced by voicing transitions as found in prior MBE based speech coders. Quantization efficiency is thereby improved, producing higher speech quality at lower bit rates. In addition, smoothing methods, typically used to reduce the effect of bit errors or to enhance formants, are more effective since they are not confused by false edges (i.e. discontinuities) at voicing transitions. Overall speech quality and intelligibility are improved. At the decoder a bit stream is received and then used to reconstruct a fundamental frequency, voicing information, and a set of spectral magnitudes for a sequence of frames. The voicing information is used to label each harmonic as either voiced or unvoiced, and for voiced harmonics an individual phase is regenerated as a function of the spectral magnitudes localized about that harmonic frequency. The decoder then synthesizes the voiced and unvoiced component and adds them to produce the synthesized speech. The regenerated phase more closely approximates actual speech in terms of peak-to-rms value relative to the prior art, thereby yielding improved dynamic range. In addition the synthesized speech is perceived as more natural and exhibits fewer phase related distortions.

Synthesis of Speech Using Regenerated Phase Information

Background of the Invention

The present invention relates to methods for representing speech to facilitate efficient low to medium rate encoding and decoding.

Relevant publications include: J. L. Flanagan, Speech Analysis, Synthesis and Perception, Springer-Verlag, 1972, pp. 378-386, (discusses phase vocoder – frequency-based speech analysis-synthesis system); Jayant et al., Digital Coding of Waveforms, Prentice-Hall, 1984, (discusses speech coding in general); U.S. Patent No. 4,885,790 (discloses sinusoidal processing method); U.S. Patent No. 5,054,072 (discloses sinusoidal coding method); Almeida et al., "Nonstationary Modelling of Voiced Speech", IEEE TASSP, Vol. ASSP-31, No. 3, June 1983, pp 664-677, (discloses harmonic modelling and coder); Almeida et al., "Variable-Frequency Synthesis: An Improved Harmonic Coding Scheme", IEEE Proc. ICASSP 84, pp 27.5.1-27.5.4, (discloses polynomial voiced synthesis method); 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 sinusoidal representation); 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 the sinusoidal transform speech coder); Griffin, "Multiband Excitation Vocoder", Ph.D. Thesis, M.I.T, 1987, (discusses Multi-Band Excitation (MBE) speech model and an 8000 bps MBE speech coder); Hardwick, "A 4.8 kbps

Multi-Band Excitation Speech Coder", SM. Thesis, M.I.T, May 1988, (discusses a 4800 bps Multi-Band Excitation speech coder); Telecommunications Industry Association (TIA), "APCO Project 25 Vocoder Description", Version 1.3, July 15, 1993, IS102BABA (discusses 7.2 kbps $IMBE^{TM}$ speech coder for APCO Project 25 standard); US patent No. 5,081,681 (discloses MBE random phase synthesis); US patent No. 5,247,579 (discloses MBE channel error mitigation method and formant enhancement method); US patent No. 5,226,084 (discloses MBE quantization and error mitigation methods).

(IMBE is a trademark of Digital Voice Systems, Inc.)

The problem of encoding and decoding speech has a large number of applications and hence it has been studied extensively. In many cases it is desirable to reduce the data rate needed to represent a speech signal without substantially reducing the quality or intelligibility of the speech. This problem, commonly referred to as "speech compression", is performed by a speech coder or vocoder.

A speech coder is generally viewed as a two part process. The first part, commonly referred to as the encoder, starts with a digital representation of speech, such as that generated by passing the output of a microphone through an A-to-D converter, and outputs a compressed stream of bits. The second part, commonly referred to as the decoder, converts the compressed bit stream back into a digital representation of speech which is suitable for playback through a D-to-A converter and a speaker. In many applications the encoder and decoder are physically separated and the bit steam is transmitted between them via some communication channel.

A key parameter of a speech coder is the amount of compression it achieves, which is measured via its bit rate. The actual compressed bit rate achieved is generally a function of the desired fidelity (i.e., speech quality) and the type of speech. Different types of speech coders have been designed to operate at high rates (greater than 8 kbps), mid-rates (3 - 8 kbps) and low rates (less than 3 kbps). Recently,

mid-rate speech coders have been the subject of strong interest in a wide range of mobile communication applications (cellular, satellite telephony, land mobile radio, in-flight phones, etc...). These applications typically require high quality speech and robustness to artifacts caused by acoustic noise and channel noise (bit errors).

One class of speech coders, which have been shown to be highly applicable to mobile communications, is based upon an underlying model of speech. Examples from this class include linear prediction vocoders, homomorphic vocoders, sinusoidal transform coders, multi-band excitation speech coders and channel vocoders. In these vocoders, speech is divided into short segments (typically 10-40 ms) and each segment is characterized by a set of model parameters. These parameters typically represent a few basic elements, including the pitch, the voicing state and spectral envelope, of each speech segment. A model-based speech coder can use one of a number of known representations for each of these parameters. For example the pitch may be represented as a pitch period, a fundamental frequency, or a long-term prediction delay as in CELP coders. Similarly the voicing state can be represented through one or more voiced/unvoiced decisions, a voicing probability measure, or by the ratio of periodic to stochastic energy. The spectral envelope is often represented by an all-pole filter response (LPC) but may equally be characterized by a set of harmonic amplitudes or other spectral measurements. Since usually only a small number of parameters are needed to represent a speech segment, model based speech coders are typically able to operate at medium to low data rates. However, the quality of a model-based system is dependent on the accuracy of the underlying model. Therefore a high fidelity model must be used if these speech coders are to achieve high speech quality.

One speech model which has been shown to provide good quality speech and to work well at medium to low bit rates is the Multi-Band Excitation (MBE) speech model developed by Griffin and Lim. This model uses a flexible voicing structure which allows it to produce more natural sounding speech, and which makes it more

robust to the presence of acoustic background noise. These properties have caused the MBE speech model to be employed in a number of commercial mobile communication applications.

The MBE speech model represents segments of speech using a fundamental frequency, a set of binary voiced or unvoiced (V/UV) decisions and a set of harmonic amplitudes. The primary advantage of the MBE model over more traditional models is in the voicing representation. The MBE model generalizes the traditional single V/UV decision per segment into a set of decisions, each representing the voicing state within a particular frequency band. This added flexibility in the voicing model allows the MBE model to better accommodate mixed voicing sounds, such as some voiced fricatives. In addition this added flexibility allows a more accurate representation of speech corrupted by acoustic background noise. Extensive testing has shown that this generalization results in improved voice quality and intelligibility.

The encoder of an MBE based speech coder estimates the set of model parameters for each speech segment. The MBE model parameters consist of a fundamental frequency, which is the reciprocal of the pitch period; a set of V/UV decisions which characterize the voicing state; and a set of spectral amplitudes which characterize the spectral envelope. Once the MBE model parameters have been estimated for each segment, they are quantized at the encoder to produce a frame of bits. These bits are then optionally protected with error correction/detection codes (ECC) and the resulting bit stream is then transmitted to a corresponding decoder. The decoder converts the received bit stream back into individual frames, and performs optional error control decoding to correct and/or detect bit errors. The resulting bits are then used to reconstruct the MBE model parameters from which the decoder synthesizes a speech signal which is perceptually close to the original. In practice the decoder synthesizes separate voiced and unvoiced components and adds the two components to produce the final output.

In MBE based systems a spectral amplitude is used to represent the spectral

envelope at each harmonic of the estimated fundamental frequency. Typically each harmonic is labeled as either voiced or unvoiced depending upon whether the frequency band containing the corresponding harmonic has been declared voiced or unvoiced. The encoder then estimates a spectral amplitude for each harmonic frequency, and in prior art MBE systems a different amplitude estimator is used depending upon whether it has been labeled voiced or unvoiced. At the decoder the voiced and unvoiced harmonics are again identified and separate voiced and unvoiced components are synthesized using different procedures. The unvoiced component is synthesized using a weighted overlap-add method to filter a white noise signal. The filter is set to zero all frequency regions declared voiced while otherwise matching the spectral amplitudes labeled unvoiced. The voiced component is synthesized using a tuned oscillator bank, with one oscillator assigned to each harmonic labeled voiced. The instantaneous amplitude, frequency and phase is interpolated to match the corresponding parameters at neighboring segments. Although MBE based speech coders have been shown to offer good performance, a number of problems have been identified which lead to some degradation in speech quality. Listening tests have established that in the frequency domain both the magnitude and phase of the synthesized signal must be carefully controlled in order to obtain high speech quality and intelligibility. Artifacts in the spectral magnitude can have a wide range of effects, but one common problem at mid-to-low bit rates is the introduction of a muffled quality and/or an increase in the perceived nasality of the speech. These problems are usually the result of significant quantization errors (caused by too few bits) in the reconstructed magnitudes. Speech formant enhancements methods, which amplify the spectral magnitudes corresponding to the speech formants, while attenuating the remaining spectral magnitudes, have been employed to try to correct these problems. These methods improve perceived quality up to a point, but eventually the distortion they introduce becomes too great and quality begins to deteriorate.

Performance is often further reduced by the introduction of phase artifacts, which are caused by the fact that the decoder must regenerate the phase of the voiced speech component. At low to medium data rates there are not sufficient bits to transmit any phase information between the encoder and the decoder. Consequently, the encoder ignores the actual signal phase, and the decoder must artificially regenerate the voiced phase in a manner which produces natural sounding speech.

Extensive experimentation has shown that the regenerated phase has a significant effect on perceived quality. Early methods of regenerating the phase involved simple integration of the harmonic frequencies from some set of initial phases. This procedure ensured the voiced component was continuous at segment boundaries; however, choosing a set of initial phases which resulted in high quality speech was found to be problematic. If the initial phases were set to zero, the resulting speech was judged to be "buzzy", while if the initial phase was randomized the speech was judged "reverberant". This result led to a better approach described in US patent No. 5,081,681, where depending on the V/UV decisions, a controlled amount of randomness was added to the phase in order to adjust the balance between "buzziness" and "reverberance". Listening tests showed that less randomness was preferred when the voiced component dominated the speech, while more phase randomness was preferred when the unvoiced component dominated. Consequently, a simple voicing ratio was computed to control the amount of phase randomness in this manner. Although voicing dependent random phase was shown to be adequate for many applications, listening experiments still traced a number of quality problems to the voiced component phase. Tests confirmed that the voice quality could be significantly improved by removing the use of random phase, and instead individually controlling the phase at each harmonic frequency in a manner which more closely matched actual speech. This discovery has led to the present invention, described here in the context of the preferred embodiment.

Summary of the Invention

This invention provides a method for decoding and synthesizing a synthetic digital speech signal from a plurality of digital bits of the type produced by dividing a speech signal into a plurality of frames, determining voicing information representing whether each of a plurality of frequency bands of each frame should be synthesized as voiced or unvoiced bands; processing the speech frames to determine spectral envelope information representative of the magnitudes of the spectrum in the frequency bands, and quantizing and encoding the spectral envelope and voicing information, wherein the method for decoding and synthesizing the synthetic digital speech signal comprises the steps of:

decoding the plurality of bits to provide spectral envelope and voicing information for each of a plurality of frames;

processing the spectral envelope information to determine regenerated spectral phase information for each of the plurality of frames;

determining from the voicing information whether frequency bands for a particular frame are voiced or unvoiced;

synthesizing speech components for voiced frequency bands using the regenerated spectral phase information;

synthesizing a speech component representing the speech signal in at least one unvoiced frequency band; and

synthesizing the speech signal by combining the synthesized speech components for voiced and unvoiced frequency bands.

This invention also provides an apparatus for decoding and synthesizing a synthetic digital speech signal from a plurality of digital bits of the type produced by dividing a speech signal into a plurality of frames, determining voicing information representing whether each of a plurality of frequency bands of each frame should be synthesized as voiced or unvoiced bands; processing the speech frames to determine spectral envelope information representative of the magnitudes of the spectrum in the frequency bands, and quantizing and encoding the spectral envelope and voicing information, wherein the apparatus for decoding and synthesizing the synthetic digital speech comprises:

means for decoding the plurality of bits to provide spectral envelope and voicing information for each of a plurality of frames;

means for processing the spectral envelope information to determine regenerated spectral phase information for each of the plurality of frames;

means for determining from the voicing information whether frequency bands for a particular frame are voiced or unvoiced;

means for synthesizing speech components for voiced frequency bands using the regenerated spectral phase information;

means for synthesizing a speech component representing the speech signal in at least one unvoiced frequency band; and

means for synthesizing the speech signal by combining the synthesized speech components for voiced and unvoiced frequency bands.

In a first aspect, the invention features an improved method of regenerating the voiced component phase in speech synthesis. The phase is estimated from the spectral envelope of the voiced component (e.g., from the shape of the spectral envelope in the vicinity of the voiced component). The decoder reconstructs the spectral envelope and voicing information for each of a plurality of frames, and the voicing information is used to determine whether frequency bands for a particular frame are voiced or unvoiced. Speech components are synthesized for voiced frequency bands using the regenerated spectral phase information. Components for unvoiced frequency bands are generated using other techniques, e.g., from a filter response to a random noise signal, wherein the filter has approximately the spectral envelope in the unvoiced bands and approximately zero magnitude in the voiced bands.

Preferably, the digital bits from which the synthetic speech signal is synthesized include bits representing fundamental frequency information, and the spectral envelope information comprises spectral magnitudes at harmonic multiples of the fundamental frequency. The voicing information is used to label each frequency band (and each of the harmonics within a band) as either voiced or unvoiced, and for harmonies within a voiced band an individual phase is regenerated as a function of the spectral envelope (the spectral shape represented by the spectral magnitudes) localized about that harmonic frequency.

Preferably, the spectral magnitudes represent the spectral envelope independently of whether a frequency band is voiced or unvoiced. The regenerated spectral phase information is determined by applying an edge detection kernel to a representation of the spectral envelope, and the representation of the spectral envelope to which the edge detection kernel is applied has been compressed. The voice speech components are determined at least in part using a bank of sinusoidal oscillators, with the oscillator characteristics being determined from the fundamental frequency and regenerated spectral phase information.

The invention produces synthesized speech that more closely approximates ac-

tual speech in terms of peak-to-rms value relative to the prior art, thereby yielding improved dynamic range. In addition to synthesized speech is perceived as more natural and exhibits fewer phase related distortions.

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

Figure 1 is a drawing of the invention, embodied in the new MBE based speech encoder. A digital speech signal $s(n)$ is first segmented with a sliding window function $w(n - iS)$ where the frame shift S is typically equal to 20 ms. The resulting segment of speech, denoted $s_w(n)$ is then processed to estimate the fundamental frequency ω_0, a set of Voiced/Unvoiced decisions, v_k, and a set of spectral magnitudes, M_l. The spectral magnitudes are computed, independent of the voicing information, after transforming the speech segment into the spectral domain with a Fast Fourier Transform (FFT). The frame of MBE model parameters are then quantized and encoded into a digital bit stream. Optional FEC redundancy is added to protect the bit stream against bit errors during transmission.

Figure 2 is a drawing of the invention embodied in the new MBE based speech decoder. The digital bit stream, generated by the corresponding encoder as shown in Figure 1, is first decoded and used to reconstruct each frame of MBE model parameters. The reconstructed voicing information, v_k, is used to reconstruct K voicing bands and to label each harmonic frequency as either voiced or unvoiced, depending upon the voicing state of the band in which it is contained. Spectral phases, ϕ_l are regenerated from the spectral magnitudes, M_l, and then used to synthesize the voiced component $s_v(n)$, representing all harmonic frequencies labelled voiced. The voiced component is then added to the unvoiced component (representing unvoiced bands) to create the synthetic speech signal.

Preferred Embodiment of the Invention

The preferred embodiment of the invention is described in the context of a new MBE based speech coder. This system is applicable to a wide range of environments, including mobile communication applications such as mobile satellite, cellular telephony, land mobile radio (SMR, PMR), etc.... This new speech coder combines the standard MBE speech model with a novel analysis/synthesis procedure for computing the model parameters and synthesizing speech from these parameters. The new method allows speech quality to be improved while lowering the bit rate needed to encode and transmit the speech signal. Although the invention is described in the context of this particular MBE based speech coder, the techniques and methods disclosed herein can readily be applied to other systems and techniques by someone skilled in the art without departing from the spirit and scope of this invention.

In the new MBE based speech coder a digital speech signal sampled at 8 kHz is first divided into overlapping segments by multiplying the digital speech signal by a short (20-40 ms) window function such as a Hamming window. Frames are typically computed in this manner every 20 ms, and for each frame the fundamental frequency and voicing decisions are computed. In the new MBE based speech coder these parameters are computed according to the method described in Canadian patent applications 2144823 and 2167025, both entitled "ESTIMATION OF EXCITATION PARAMETERS". Alternatively,

the fundamental frequency and voicing decisions could be computed as described in TIA Interim Standard IS102BABA, entitled "APCO Project 25 Vocoder". In either case a small number of voicing decisions (typically twelve or less) is used to model the voicing state of different frequency bands within each frame. For example, in a 3.6 kbps speech coder eight V/UV decisions are typically used to represent the voicing state over eight different frequency bands spaced between 0 and 4 kHz.

Letting $s(n)$ represent the discrete speech signal, the speech spectrum for the i'th frame, $S_w(\omega, i \cdot S)$ is computed according to the following equation:

$$S_w(\omega, i) = \sum_n s(n) w(n - i \cdot S) e^{-j\omega n} \quad (1)$$

where $w(n)$ is the window function and S is the frame size which is typically 20 ms (160 samples at 8 kHz). The estimated fundamental frequency and voicing decisions for the i'th frame are then represented as $\omega_0(i \cdot S)$ and $v_k(i \cdot S)$ for $1 \leq k \leq K$, respectively, where K is the total number of V/UV decision (typically $K = 8$). For notational simplicity the frame index $i \cdot S$ can be dropped when referring to the current frame, thereby denoting the current spectrum, fundamental, and voicing decisions as: $S_w(\omega)$, ω_0 and v_k, respectively.

In MBE systems the spectral envelope is typically represented as a set of spectral amplitudes which are estimated from the speech spectrum $S_w(\omega)$. Spectral amplitudes are typically computed at each harmonic frequency (i.e. at $\omega = \omega_0 l$, for $l = 0, 1, \ldots$). Unlike the prior art MBE systems, the invention features a new method for estimating these spectral amplitudes which is independent of the voicing state. This results in a smoother set of spectral amplitudes since the discontinuities are eliminated, which are normally present in prior art MBE systems whenever a voicing transition occurs. The invention features the additional advantage of providing an exact representation of the local spectral energy, thereby preserving perceived loudness. Furthermore, the invention preserves local spectral energy while compensating for the effects of the frequency sampling grid normally employed by a highly efficient Fast Fourier Transform (FFT). This also contributes to achieving a smooth set of spectral amplitudes. Smoothness is important for overall performance since it increases quantization efficiency and it allows better formant enhancement (i.e. postfiltering) as well as channel error mitigation.

In order to compute a smooth set of the spectral magnitudes, it is necessary to consider the properties of both voiced and unvoiced speech. For voiced speech, the spectral energy (i.e. $|S_w(\omega)|^2$) is concentrated around the harmonic frequencies, while for unvoiced speech, the spectral energy is more evenly distributed. In prior art MBE systems, unvoiced spectral magnitudes are computed as the average spectral energy over a frequency interval (typically equal to the estimated fundamental)

centered about each corresponding harmonic frequency. In contrast, the voiced spectral magnitudes in prior art MBE systems are set equal to some fraction (often one) of the total spectral energy in the same frequency interval. Since the average energy and the total energy can be very different, especially when the frequency interval is wide (i.e. a large fundamental), a discontinuity is often introduced in the spectral magnitudes, whenever consecutive harmonics transition between voicing states (i.e. voiced to unvoiced, or unvoiced to voiced).

One spectral magnitude representation which can solve the aforementioned problem found in prior art MBE systems is to represent each spectral magnitude as either the average spectral energy or the total spectral energy within a corresponding interval. While both of these solutions would remove the discontinuties at voicing transistions, both would introduce other fluctuations when combined with a spectral transformation such as a Fast Fourier Transform (FFT) or equivalently a Discrete Fourier Transform (DFT). In practice an FFT is normally used to evaluate $S_w(\omega)$ on a uniform sampling grid determined by the FFT length, N, which is typically a power of two. For example an N point FFT would produce N frequency samples between 0 and 2π as shown in the following equation:

$$S_w(m) = \sum_{n=0}^{N-1} s(n)w(n - i \cdot S)e^{\frac{-j2\pi mn}{N}} \qquad \text{for } 0 \leq m < N \qquad (2)$$

In the preferred embodiment the spectrum is computed using an FFT with $N = 256$, and $w(n)$ is typically set equal to the 255 point symmetric window function presented in Table 1.

It is desirable to use an FFT to compute the spectrum due to it's low complexity. However, the resulting sampling interval, $2\pi/N$, is not generally an inverse multiple of the fundamental frequency. Consequently, the number of FFT samples between any two consecutive harmonic frequencies is not constant between harmonics. The result is that if average spectral energy is used to represent the harmonic magnitudes, then voiced harmonics, which have a concentrated spectral distribution, will experience fluctuations between harmonics due to the varying number of FFT sam-

ples used to compute each average. Similarly, if total spectral energy is used to represent the harmonic magnitudes, then unvoiced harmonics, which have a more uniform spectral distribution, will experience fluctuations between harmonics due to the varying number of FFT samples over which the total energy is computed. In either case the small number of frequency samples available from the FFT can introduce sharp fluctuations into the spectral magnitudes, particularly when the fundamental frequency is small.

The invention uses a compensated total energy method for all spectral magnitudes to remove discontinuities at voicing transitions. The invention's compensation method also prevents FFT related fluctuations from distorting either the voiced or unvoiced magnitudes. In particular, the invention computes the set of spectral magnitudes for the current frame, denoted by M_l for $0 \leq l \leq L$ according to the following equation:

$$M_l = \left[\frac{\sum_{m=0}^{N-1} |S_w(m)|^2 G(\frac{2\pi m}{N} - l\omega_0)}{N \sum_{n=0}^{N-1} w^2(n)} \right]^{\frac{1}{2}} \tag{3}$$

It can be seen from this equation, that each spectral magnitude is computed as a weighted sum of the spectral energy $|S_w(m)|^2$, where the weighting function is offset by the harmonic frequency for each particular spectral magnitude. The weighting function $G(\omega)$ is designed to compensate for the offset between the harmonic frequency $l\omega_0$ and the FFT frequency samples which occur at $2\pi m/N$. This function is changed each frame to reflect the estimated fundamental frequency as follows:

$$G(\omega) = \begin{cases} 1 & \text{for } |\omega| < \frac{\omega_0}{2} - \frac{\pi}{N} \\ \frac{1}{2} - \frac{N}{2\pi}(\omega - \frac{\omega_0}{2}) & \text{for } \frac{\omega_0}{2} - \frac{\pi}{N} \leq |\omega| < \frac{\omega_0}{2} + \frac{\pi}{N} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

One valuable property of this spectral magnitude representation is that it is based on the local spectral energy (i.e $|S_w(m)|^2$) for both voiced and unvoiced harmonics. Spectral energy is generally considered to be a close approximation of the way humans perceive speech, since it conveys both the relative frequency content and the loudness information without being effected by the phase of the speech signal.

Since the new magnitude representation is independent of the voicing state, there are no fluctuations or discontinuities in the representation due to transitions between voiced and unvoiced regions or due to a mixture of voiced and unvoiced energy. The weighting function $G(\omega)$ further removes any fluctuations due to the FFT sampling grid. This is achieved by interpolating the energy measured between harmonics of the estimated fundamental in a smooth manner. An additional advantage of the weighting functions disclosed in Equation (4) is that the total energy in the speech is preserved in the spectral magnitudes. This can be seen more clearly by examining the following equation for the total energy in the set of spectral magnitudes.

$$\sum_{l=0}^{L} |M_l|^2 = \frac{1}{N \sum_{n=0}^{N-1} w^2(n)} \sum_{m=0}^{N-1} |S_w(m)|^2 \sum_{l=0}^{L} G(\frac{2\pi m}{N} - l\omega_0) \qquad (5)$$

This equation can be simplified by recognizing that the sum over $G(\frac{2\pi m}{N} - l\omega_0)$ is equal to one over the interval $0 \leq m \leq \lfloor \frac{L\omega_0 N}{2\pi} \rfloor$. This means that the total energy in the speech is preserved over this interval, since the energy in the spectral magnitudes is equal to the energy in the speech spectrum. Note that the denominator in Equation (5) simply compensates for the window function $w(n)$ used in computing $S_w(m)$ according to Equation (1) . Another important point is that the bandwidth of the representation is dependent on the product $L\omega_0$. In practice the desired bandwidth is usually some fraction of the Nyquist frequency which is represented by π. Consequently the total number of spectral magnitudes, L, is inversely related to the estimated fundamental frequency for the current frame and is typically computed as follows:

$$L = \lfloor \frac{\alpha\pi}{\omega_0} \rfloor \qquad (6)$$

where $0 \leq \alpha < 1$. A 3.6 kbps system which uses an 8 kHz sampling rate has been designed with $\alpha = .925$ giving a bandwidth of 3700 Hz.

Weighting functions other than that described above can also be used in Equation (3). In fact, total power is maintained if the sum over $G(\omega)$ in Equation (5) is approximately equal to a constant (typically one) over some effective bandwidth.

The weighting function given in Equation (4) uses linear interpolation over the FFT sampling interval ($2\pi/N$) to smooth out any fluctuations introduced by the sampling grid. Alternatively, quadratic or other interpolation methods could be incorporated into $G(\omega)$ without departing from the scope of the invention.

Although the invention is described in terms of the MBE speech model's binary V/UV decisions, the invention is also applicable to systems using alternative representations for the voicing information. For example, one alternative popularized in sinsoidal coders is to represent the voicing information in terms of a cut-off frequency, where the spectrum is considered voiced below this cut-off frequency and unvoiced above it. Other extensions such as non-binary voicing information would also benefit from the invention.

The invention improves the smoothness of the magnitude representations since discontinuities at voicing transitions and fluctuations caused by the FFT sampling grid are prevented. A well known result from information theory is that increased smoothness facilitates accurate quantization of the spectral magnitudes with a small number of bits. In the 3.6 kbps system 72 bits are used to quantize the model parameters for each 20 ms frame. Seven (7) bits are used to quantize the fundamental frequency, and 8 bits are used to code the V/UV decisions in 8 different frequency bands (approximately 500 Hz each). The remaining 57 bits per frame are used to quantize the spectral magnitudes for each frame. A differential block Discrete Cosine Transform (DCT) method is applied to the log spectral magnitudes. The invention's increased smoothness compacts more of the signal power into the slowly changing DCT components. The bit allocation and quantizer step sizes are adjusted to account for this effect giving lower spectral distortion for the available number of bits per frame. In mobile communications applications it is often desirable to include additional redundancy to the bit stream prior to transmission across the mobile channel. This redundancy is typically generated by error correction and/or detection codes which add additional redundancy to the bit stream in such a man-

ner that bit errors introduced during transmission can be corrected and/or detected. For example, in a 4.8 kbps mobile satellite application, 1.2 kbps of redundant data is added to the 3.6 kbps of speech data. A combination of one [24,12] Golay code and three [15,11] Hamming Codes is used to generate the additional 24 redundant bits added to each frame. Many other types of error correction codes, such as convolutional, BCH, Reed-Solomon, etc..., could also be employed to change the error robustness to meet virtually any channel condition.

At the receiver the decoder receives the transmitted bit stream and reconstructs the model parameters (fundamental frequency, V/UV decisions and spectral magnitudes) for each frame. In practice the received bit stream may contain bit errors due to noise in the channel. As a consequence the V/UV bits may be decoded in error, causing a voiced magnitude to be interpreted as unvoiced or vice versa. The invention reduces the perceived distortion from these voicing errors since the magnitude itself, is independent of the voicing state. Another advantage of the invention occurs during formant enhancement at the receiver. Experimentation has shown perceived quality is enhanced if the spectral magnitudes at the formant peaks are increased relative to the spectral magnitudes at the formant valleys. This process tends to reverse some of the formant broadening which is introduced during quantization. The speech then sounds crisper and less reverberant. In practice the spectral magnitudes are increased where they are greater than the local average and decreased where they are less than the local average. Unfortunately, discontinuities in the spectral magnitudes can appear as formants, leading to spurious increases or decreases. The invention's improved smoothness helps solve this problem leading to improved formant enhancement while reducing spurious changes.

As in previous MBE systems, the new MBE based encoder does not estimate or transmit any spectral phase information. Consequently, the new MBE based decoder must regenerate a synthetic phase for all voiced harmonics during voiced speech synthesis. The invention features a new magnitude dependent phase generation

method which more closely approximates actual speech and improves overall voice quality. The prior art technique of using random phase in the voiced components is replaced with a measurement of the local smoothness of the spectral envelope. This is justified by linear system theory, where spectral phase is dependent on the pole and zero locations. This can be modeled by linking the phase to the level of smoothness in the spectral magnitudes. In practice an edge detection computation of the following form is applied to the decoded spectral magnitudes for the current frame:

$$\phi_l = \sum_{m=-D}^{D} h(m) B_{l+m} \qquad \text{for } 1 \leq l \leq L \tag{7}$$

where the parameters B_l represent the compressed spectral magnitudes and $h(m)$ is an appropriately scaled edge detection kernel. The output of this equation is a set of regenerated phase values, ϕ_l, which determine the phase relationship between the voiced harmonics. One should note that these values are defined for all harmonics, regardless of the voicing state. However, in MBE based systems only the voiced synthesis procedure uses these phase values, while the unvoiced synthesis procedure ignores them. In practice the regenerated phase values are computed for all harmonics and then stored, since they may be used during the synthesis of the next frame as explained in more detail below (see Equation (20)).

The compressed magnitude parameters B_l are generally computed by passing the spectral magnitudes M_l through a companding function to reduce their dynamic range. In addition extrapolation is performed to generate additional spectral values beyond the edges of the magnitude representation (i.e. $l \leq 0$ and $l > L$). One particularly suitable compression function is the logarithm, since it converts any overall scaling of the spectral magnitudes M_l (i.e. its loudness or volume) into an additive offset in B_l. Assuming that $h(m)$ in Equation (7) is zero mean, then this offset is ignored and the regenerated phase values ϕ_l are independent of scaling. In practice $\log_2$ has been used since it is easily computable on a digital computer. This

leads to the following expression for B_l:

$$B_l = \begin{cases} 0 & \text{for } l = 0 \\ \log_2(M_l) & \text{for } 1 \leq |l| \leq L \\ \log_2(M_L) - \gamma * (l - L) & \text{for } L < |l| \leq L + D \end{cases} \quad (8)$$

The extrapolated values of B_l for $l > L$ are designed to emphasize smoothness at harmonic frequencies above the represented bandwidth. A value of $\gamma = .72$ has been used in the 3.6 kbps system, but this value is not considered critical, since the high frequency components generally contribute less to the overall speech than the low frequency components. Listening tests have shown that the values of B_l for $l \leq 0$ can have a significant effect on perceived quality. The value at $l = 0$ was set to a small value since in many applications such as telephony there is no DC response. In addition listening experiments showed that $B_0 = 0$ was preferable to either positive or negative extremes. The use of a symmetric response $B_{-l} = B_l$ was based on system theory as well as on listening experiments.

The selection of an appropriate edge detection kernel $h(m)$ is important for overall quality. Both the shape and scaling influence the phase variables ϕ_l which are used in voiced synthesis, however a wide range of possible kernels could be successfully employed. Several constraints have been found which generally lead to well designed kernels. Specifically, if $h(m) \geq 0$ for $m > 0$ and if $h(m) = -h(-m)$ then the function is typically better suited to localize discontinuities. In addition it is useful to constrain $h(0) = 0$ to obtain a zero mean kernel for scaling independence. Another desirable property is that the absolute value of $h(m)$ should decay as $|m|$ increases in order to focus on local changes in the spectral magnitudes. This can be achieved by making $h(m)$ inversely proportional to m. One equation (of many) which satisfies all of these constraints is shown in Equation (9).

$$h(m) = \begin{cases} \frac{\lambda}{m} & \text{for } m \text{ odd and } -D \leq m \leq D \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

The preferred embodiment of the invention uses Equation (9) with $\lambda = .44$. This value was found to produce good sounding speech with modest complexity, and the synthesized speech was found to possess a peak-to-rms energy ratio close to that of the original speech. Tests performed with alternate values of λ showed that small variations from the preferred value resulted in nearly equivalent performance. The kernel length D can be adjusted to tradeoff complexity versus the amount of smoothing. Longer values of D are generally preferred by listeners, however a value of $D = 19$ has been found to be essentially equivalent to longer lengths and hence $D = 19$ is used in the new 3.6 kbps system.

One should note that the form of Equation (7) is such that all of the regenerated phase variables for each frame can be computed via a forward and inverse FFT operation. Depending on the processor, an FFT implementation can lead to greater computational efficiency for large D and L than direct computation.

The calculation of the regenerated phase variables is greatly facilitated by the invention's new spectral magnitude representation which is independent of voicing state. As discussed above the kernel applied via Equation (7) accentuates edges or other fluctuations in the spectral envelope. This is done to approximate the phase relationship of a linear system in which the spectral phase is linked to changes in the spectral magnitude via the pole and zero locations. In order to take advantage of this property, the phase regeneration procedure must assume that the spectral magnitudes accurately represent the spectral envelope of the speech. This is facilitated by the invention's new spectral magnitude representation, since it produces a smoother set of spectral magnitudes than the prior art. Removal of discontinuities and fluctuations caused by voicing transitions and the FFT sampling grid allows more accurate assessment of the true changes in the spectral envelope. Consequently phase regeneration is enhanced, and overall speech quality is improved.

Once the regenerated phase variables, ϕ_l, have been computed according to the above procedure, the voiced synthesis process synthesizes the voiced speech $s_v(n)$ as

the sum of individual sinusoidal components as shown in Equation (10). The voiced synthesis method is based on a simple ordered assignment of harmonics to pair the l'th spectral amplitude of the current frame with the l'th spectral amplitude of the previous frame. In this process the number of harmonics, fundamental frequency, V/UV decisions and spectral amplitudes of the current frame are denoted as $L(0)$, $\omega_0(0)$, $v_k(0)$ and $M_l(0)$, respectively, while the same parameters for the previous frame are denoted as $L(-S)$, $\omega_0(-S)$, $v_k(-S)$ and $M_l(-S)$. The value of S is equal to the frame length which is 20 ms (160 samples) in the new 3.6 kbps system.

$$s_v(n) = \sum_{l=1}^{max[L(-S),L(0)]} 2 \cdot s_{v,l}(n) \qquad \text{for } -S < n \leq 0 \tag{10}$$

The voiced component $s_{v,l}(n)$ represents the contribution to the voiced speech from the $l'th$ harmonic pair. In practice the voiced components are designed as slowly varying sinusoids, where the amplitude and phase of each component is adjusted to approximate the model parameters from the previous and current frames at the endpoints of the current synthesis interval (i.e. at $n = -S$ and $n = 0$), while smoothly interpolating between these parameters over the duration of the interval $-S < n < 0$.

In order to accommodate the fact that the number of parameters may be different between successive frames, the synthesis method assumes that all harmonics beyond the allowed bandwidth are equal to zero as shown in the following equations.

$$M_l(0) = 0 \qquad \text{for } l > L(0) \tag{11}$$

$$M_l(-S) = 0 \qquad \text{for } l > L(-S) \tag{12}$$

In addition it assumes that these spectral amplitudes outside the normal bandwidth are labeled as unvoiced. These assumptions are needed for the case where the number of spectral amplitudes in the current frame is not equal to the number of spectral amplitudes in the previous frame (i.e. $L(0) \neq L(-S)$).

The amplitude and phase functions are computed differently for each harmonic pair. In particular the voicing state and the relative change in the fundamental

frequency determine which of four possible functions are used for each harmonic for the current synthesis interval. The first possible case arises if the l'th harmonic is labeled as unvoiced for both the previous and current speech frame, in which event the voiced component is set equal to zero over the interval as shown in the following equation.

$$s_{v,l}(n) = 0 \qquad \text{for } -S < n \leq 0 \tag{13}$$

In this case the speech energy around the l'th harmonic is entirely unvoiced and the unvoiced synthesis procedure is responsible for synthesizing the entire contribution.

Alternatively, if the l'th harmonic is labeled as unvoiced for the current frame and voiced for the previous frame, then $s_{v,l}(n)$ is given by the following equation,

$$s_{v,l}(n) = w_s(n+S)\, M_l(-S)\, \cos[\omega_0(-S)\,(n+S)\,l + \theta_l(-S)] \qquad \text{for } -S < n \leq 0 \tag{14}$$

In this case the energy in this region of the spectrum transitions from the voiced synthesis method to the unvoiced synthesis method over the duration of the synthesis interval.

Similarly, if the l'th harmonic is labeled as voiced for the current frame and unvoiced for the previous frame then $s_{v,l}(n)$ is given by the following equation.

$$s_{v,l}(n) = w_s(n)\, M_l(0)\, \cos[\omega_0(0)\, n\, l + \theta_l(0)] \qquad \text{for } -S < n \leq 0 \tag{15}$$

In this case the energy in this region of the spectrum transitions from the unvoiced synthesis method to the voiced synthesis method.

Otherwise, if the l'th harmonic is labeled as voiced for both the current and the previous frame, and if either $l >= 8$ or $|\omega_0(0) - \omega_0(-S)| \geq .1\ \omega_0(0)$, then $s_{v,l}(n)$ is given by the following equation, where the variable n is restricted to the range $-S < n \leq 0$.

$$\begin{aligned} s_{v,l}(n) &= w_s(n+S)\, M_l(-S)\, \cos[\omega_0(-S)\,(n+S)\,l + \theta_l(-S)] \\ &+ w_s(n)\, M_l(0)\, \cos[\omega_0(0)\, n\, l + \theta_l(0)] \end{aligned} \tag{16}$$

The fact that the harmonic is labeled voiced in both frames, corresponds to the situation where the local spectral energy remains voiced and is completely synthesized within the voiced component. Since this case corresponds to relatively large changes in harmonic frequency, an overlap-add approach is used to combine the contribution from the previous and current frame. The phase variables $\theta_l(-S)$ and $\theta_l(0)$ which are used in Equations (14), (15) and (16) are determined by evaluating the continuous phase function $\theta_l(n)$ described in Equation (20) at $n = -S$ and $n = 0$.

A final synthesis rule is used if the l'th spectral amplitude is voiced for both the current and the previous frame, and if both $l < 8$ and $|\omega_0(0) - \omega_0(-S)| < .1\ \omega_0(0)$. As in the prior case, this event only occurs when the local spectral energy is entirely voiced. However, in this case the frequency difference between the previous and current frames is small enough to allow a continuous transition in the sinusoidal phase over the synthesis interval. In this case the voiced component is computed according to the following equation,

$$s_{v,l}(n) = a_l(n)\ \cos[\theta_l(n)] \qquad \text{for } -S < n \leq 0 \tag{17}$$

where the amplitude function, $a_l(n)$, is computed according to Equation (18), and the phase function, $\theta_l(n)$, is a low order polynomial of the type described in Equations (19) and (20).

$$a_l(n) = w_s(n+S)\, M_l(-S) + w_s(n)\, M_l(0) \tag{18}$$

$$\theta_l(n) = \theta_l(-S) + [\omega_0(-S) \cdot l + \Delta\omega_l](n+S) + [\omega_0(0) - \omega_0(-S)] \cdot \frac{l(n+S)^2}{2S} \tag{19}$$

$$\Delta\omega_l = \frac{1}{S}\left[\phi_l(0) - \phi_l(-S) - 2\pi\left\lfloor \frac{\phi_l(0) - \phi_l(-S) + \pi}{2\pi} \right\rfloor\right] \tag{20}$$

The phase update process described above uses the invention's regenerated phase values for both the previous and current frame (i.e. $\phi_l(0)$ and $\phi_l(-S)$) to control the phase function for the l'th harmonic. This is performed via the second order phase polynomial expressed in Equation (19) which ensures continuity of phase at the ends of the synthesis boundary via a linear phase term and which otherwise meets the

desired regenerated phase. In addition the rate of change of this phase polynomial is approximately equal to the appropriate harmonic frequency at the endpoints of the interval.

The synthesis window $w_s(n)$ used in Equations (14), (15), (16) and (18) is typically designed to interpolate between the model parameters in the current and previous frames. This is facilitated if the following overlap-add equation is satisfied over the entire current synthesis interval.

$$w_s(n) + w_s(n+S) = 1 \qquad \text{for } -S < n \leq 0 \tag{21}$$

One synthesis window which has been found useful in the new 3.6 kbps system and which meets the above constraint is defined as follows:

$$w_s(n) = \begin{cases} 1 & \text{for } |n| \leq (S-\beta)/2 \\ 1 + \frac{(S-\beta)-2n}{2\beta} & \text{for } (S-\beta)/2 < n < (S+\beta)/2 \\ 1 + \frac{(S-\beta)+2n}{2\beta} & \text{for } -(S-\beta)/2 > n > -(S+\beta)/2 \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

For a 20 ms frame size ($S = 160$) a value of $\beta = 50$ is typically used. The synthesis window presented in Equation (22) is essentially equivalent to using linear interpolation.

The voiced speech component synthesized via Equation (10) and the described procedure must still be added to the unvoiced component to complete the synthesis process. The unvoiced speech component, $s_{uv}(n)$, is normally synthesized by filtering a white noise signal with a filter response of zero in voiced frequency bands and with a filter response determined by the spectral magnitudes in frequency bands declared unvoiced. In practice this is performed via a weighted overlap-add procedure which uses a forward and inverse FFT to perform the filtering. Since this procedure is well known, the references should be consulted for complete details.

Various alternatives and extensions to the specific techniques taught here could be used without departing from the spirit and scope of the invention. For example a

third order phase polynomial could be used by replacing the $\Delta\omega_l$ term in Equation (19) with a cubic term having the correct boundary conditions. In addition the prior art describes alternative windows functions and interpolation methods as well as other variations. Other embodiments of the invention are within the following claims.

Claims

1. A method for decoding and synthesizing a synthetic digital speech signal from a plurality of digital bits of the type produced by dividing a speech signal into a plurality of frames, determining voicing information representing whether each of a plurality of frequency bands of each frame should be synthesized as voiced or unvoiced bands; processing the speech frames to determine spectral envelope information representative of the magnitudes of the spectrum in the frequency bands, and quantizing and encoding the spectral envelope and voicing information, wherein the method for decoding and synthesizing the synthetic digital speech signal comprises the steps of:

decoding the plurality of bits to provide spectral envelope and voicing information for each of a plurality of frames;

processing the spectral envelope information to determine regenerated spectral phase information for each of the plurality of frames,

determining from the voicing information whether frequency bands for a particular frame are voiced or unvoiced;

synthesizing speech components for voiced frequency bands using the regenerated spectral phase information,

synthesizing a speech component representing the speech signal in at least one unvoiced frequency band, and

synthesizing the speech signal by combining the synthesized speech components for voiced and unvoiced frequency bands.

2. Apparatus for decoding and synthesizing a synthetic digital speech signal from a plurality of digital bits of the type produced by dividing a speech signal into a plurality of frames, determining voicing information representing whether each of a plurality of frequency bands of each frame should be synthesized as voiced or unvoiced bands; processing the speech frames to determine spectral envelope information representative of the magnitudes of the spectrum in the frequency bands,

and quantizing and encoding the spectral envelope and voicing information, wherein the apparatus for decoding and synthesizing the synthetic digital speech comprises:

means for decoding the plurality of bits to provide spectral envelope and voicing information for each of a plurality of frames;

means for processing the spectral envelope information to determine regenerated spectral phase information for each of the plurality of frames,

means for determining from the voicing information whether frequency bands for a particular frame are voiced or unvoiced;

means for synthesizing speech components for voiced frequency bands using the regenerated spectral phase information,

means for synthesizing a speech component representing the speech signal in at least one unvoiced frequency band, and

means for synthesizing the speech signal by combining the synthesized speech components for voiced and unvoiced frequency bands.

3. The subject matter of claim 1 or 2, wherein the digital bits from which the synthetic speech signal is synthesized include bits representing spectral envelope and voicing information and bits representing fundamental frequency information.

4. The subject matter of claim 3, wherein the spectral envelope information comprises information representing spectral magnitudes at harmonic multiples of the fundamental frequency of the speech signal.

5. The subject matter of claim 4, wherein the spectral magnitudes represent the spectral envelope independently of whether a frequency band is voiced or unvoiced.

6. The subject matter of claim 4 or 5, wherein the regenerated spectral phase information is determined from the shape of the spectral envelope in the vicinity of the harmonic multiple with which the regenerated spectral phase information is associated.

7. The subject matter of claim 4 or 5, wherein the regenerated spectral phase information is determined by applying an edge detection kernel to a representation of the spectral envelope.

8. The subject matter of claim 7, wherein the representation of the spectral envelope to which the edge detection kernel is applied has been compressed.

9. The subject matter of any one of claims 4 to 8, wherein the unvoiced speech component of the synthetic speech signal is determined from a filter response to a random noise signal, wherein the filter has approximately the spectral magnitudes in the unvoiced bands and approximately zero magnitude in the voiced bands.

10. The subject matter of any one of claims 4 to 9, wherein the voiced speech components are determined at least in part using a bank of sinusoidal oscillators, with the oscillator characteristics being determined from the fundamental frequency and regenerated spectral phase information.

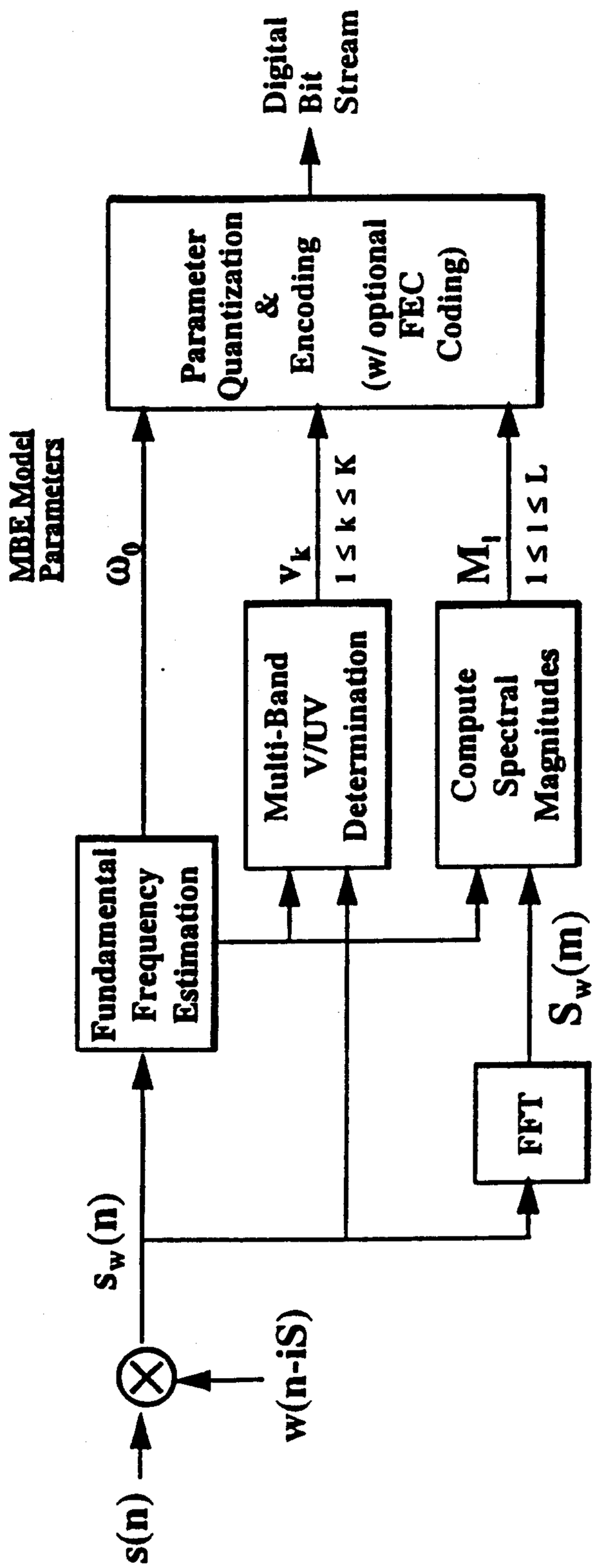

Digital
Bit
Stream
Parameter
Quantization
&
Encoding
(w/ optional
FEC
Coding)
MBE Model
Parameters
ω_0
v_k
$1 \le k \le K$
M_l
$1 \le l \le L$
Multi-Band
V/UV
Determination
Compute
Spectral
Magnitudes
Fundamental
Frequency
Estimation
FFT
$S_w(m)$
$S_w(m)$
w(n-iS)
s(n)

FIG. 1

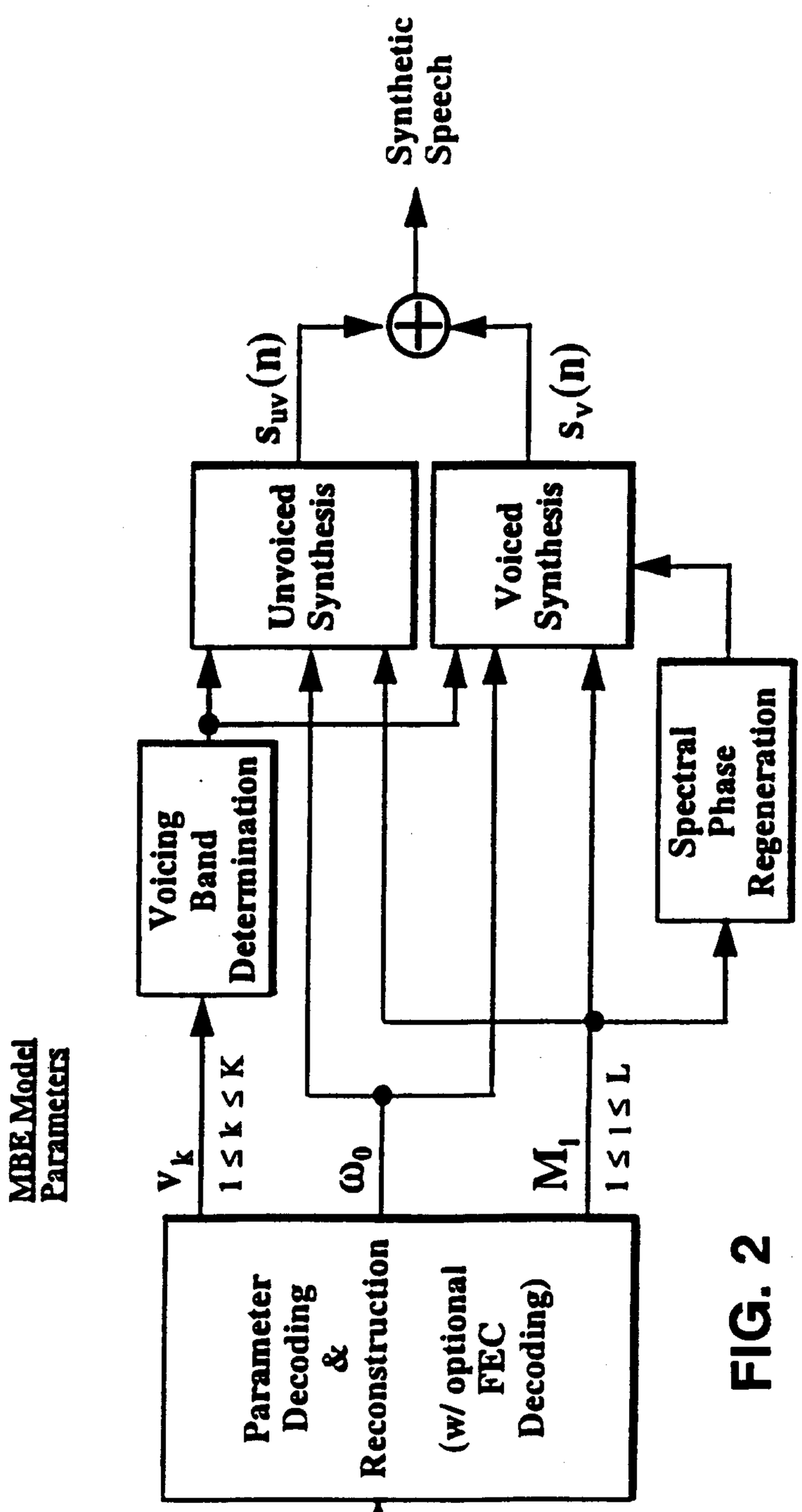
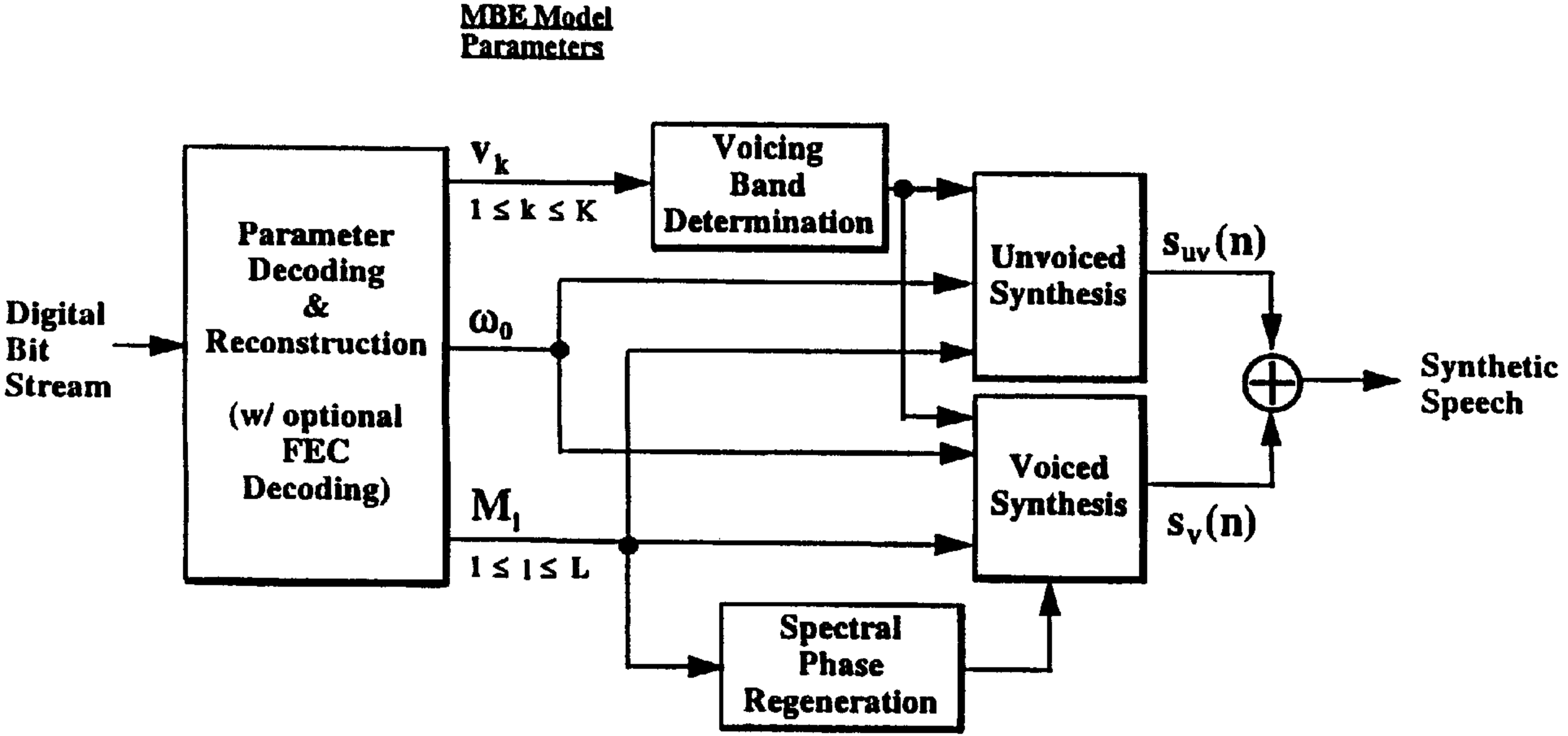

Synthetic Speech
$s_{uv}(n)$
$s_v(n)$
Unvoiced Synthesis
Voiced Synthesis
Voicing Band Determination
Spectral Phase Regeneration
MBE Model Parameters
v_k
$1 \le k \le K$
ω_0
M_l
$1 \le l \le L$
Parameter Decoding & Reconstruction (w/ optional FEC Decoding)
Digital Bit Stream

FIG. 2


MBE Model Parameters
Digital Bit Stream
Parameter Decoding & Reconstruction (w/ optional FEC Decoding)
v_k
$1 \le k \le K$
ω_0
M_l
$1 \le l \le L$
Voicing Band Determination
Unvoiced Synthesis
Voiced Synthesis
Spectral Phase Regeneration
$s_{uv}(n)$
$s_v(n)$
Synthetic Speech