PITCH-SYNCHRONOUS SPEECH CODING BASED ON TIMBRE VECTORS

ABSTRACT

A pitch-synchronous method and system for speech coding using timbre vectors is disclosed. On the encoder side, speech signal is segmented into pitch-synchronous frames without overlap, then converted into a pitch-synchronous amplitude spectrum using FFT. Using Laguerre functions, the amplitude spectrum is transformed into a timbre vector. Using vector quantization, each timbre vector is converted to a timbre index based on a timbre codebook. The intensity and pitch are also converted into indices respectively using scalar quantization. Those indices are transmitted as encoded speech. On the decoder side, by looking up the same codebooks, pitch, intensity and the timbre vector are recovered. Using Laguerre functions, the amplitude spectrum is recovered. Using Kramers-Kronig relations, the phase spectrum is recovered. Using FFT, the elementary waves are regenerated, and superposed to become the speech signal.
PITCH-SYNCHRONOUS SPEECH CODING BASED ON TIMBRE VECTORS

The present application is a continuation in part of US Patent 8,942,977, entitled "System and Method for Speech Recognition Using Pitch-Synchronous Spectral Parameters", issued 01/27/2015, to inventor Chengjun Julian Chen.

FIELD OF THE INVENTION

[0001] The present invention generally relates to speech coding, in particular to pitch-synchronous speech coding using timbre vectors.

BACKGROUND OF THE INVENTION

[0002] Speech coding is an important field of speech technology. The original speech signal is analog. The transmission of original speech signal takes a huge bandwidth and it is error prone. For several decades, coding methods and systems have been developed, to compress the speech signal to a low-bit-rate digital signal for transmission. The current status of the technology is summarized in a number of monographs, for example, Part C of "Springer Handbook of Speech Processing", Springer Verlag 2007; and "Digital Speech", Second Edition, by A. M. Kondoz, Wiley, 2004. There are several hundreds of patents and patent applications with "speech coding" in the title. The system of speech coding has two components. The encoder converts speech signal to a compressed digital signal. The decoder converts the compressed digital signal back into analog speech signal. The current technology for low bit rate speech coding is based on the following principles:

[0003] For encoding, first, speech signal is segmented into frames with a fixed duration. Second, a program determines whether a frame is voiced or unvoiced. Third, for voiced frames, find the pitch period in the frame. Fourth, extract the linear predictive code (LPC) of each frame. The voicedness index (voice or unvoiced), the pitch period, and LPC coefficients are then quantized to a limited number of bits, to become the encoded speech signal for transmission. In
the decoding process, the voiced segments and the unvoiced segments are treated differently. For voiced segments, a string of pulses are generated according to the pitch period, and then filtered by the LPC based spectrum to generate the voiced sound. For unvoiced segments, a noise signal is generated, and then filtered by the LPC based spectrum to generate an unvoiced consonant. Because pitch period is a property of the frame, each frame must be longer than the maximum pitch period of human voice, which is typically 25 msec. The frame must be multiplied with a window function, typically a Hamming window function, to make the ends approximately matching. To ensure that no information is neglected, each frame must overlap with the previous frame and the following frame, with a typical frame shift of 10 msec.

[0004] The quality of LPC-based speech coding is limited by the intrinsic properties of the LPC coefficients, which is pitch-asynchronous, and has a rather small number of parameters because of non-converging behavior when the number of coefficients is increased. The usual limit is 10 to 16 coefficients. The quality of the LPC-based speech coding is always compared with the 8-kHz sample rate 8 bit voice signal, the so-called legacy telephone standard, toll quality speech signal, or narrow-band speech signal. Coming to the 21th century, all voice recording device and voice production device can provide CD-quality speech signal, with at least 32 kHz sample rate and 16 bit resolution. Toll-quality speech signal is considered poor. Speech coding should be able to generate quality comparable to the CD-quality speech signal.

In the two previous patents by the current applicant (US Patent 8,719,030 entitled "System and Method for Speech Synthesis", US Patent 8,942,977 entitled "System and Method for Speech Recognition Using Pitch-Synchronous Spectral Parameters"), a pitch-synchronous segmentation scheme and a new mathematical representation, timbre vectors, are proposed, as an alternative to the fixed-window-size segmentation and LPC coefficients. The new methods enable the parameterization and reproduction of wide-band speech signal with high fidelity, thus provide a new method of speech coding, especially for CD-quality speech signals. The current patent application discloses systems and methods of speech coding using timbre vectors.

**SUMMARY OF THE INVENTION**

The present invention discloses a pitch-synchronous method and system for speech coding using timbre vectors, following US Patent 8,719,030 and US Patent 8,942,977.

According to an exemplary embodiment of the invention, see FIG. 1, a speech signal is first going through a pitch-marks picking program to pick the pitch marks. The pitch marks are extended to unvoiced sections to generate a complete set of segmentation points. The speech signal is segmented into pitch-synchronous frames according to the said segmentation points. An ends-meeting program is executed to make the values at the two ends of every frame equal. Using FFT (fast Fourier transform), the speech signal in each frame is converted into a pitch-synchronous amplitude spectrum, then use Laguerre functions to convert the said pitch-synchronous amplitude spectrum into a unit vector characteristic to the instantaneous timbre, referred to as the timbre vector. Using scalar quantization, the pitch period and the intensity are converted into a pitch index and an intensity index using a pitch codebook and an intensity codebook. Using vector quantization, each timbre vector is converted to a timbre index using a timbre codebook. Together with the type index (silence, unvoiced consonants, voiced consonants, and vowel), those indices are transmitted as encoded speech.

On the decoding side, as shown in FIG. 2, the type index is first fetched. According to the type, indices for pitch, intensity, and timbre are then fetched, and corresponding codebooks are chosen. Then a look-up program picks up the pitch, intensity and timbre vector for the said
pitch period. The rest of the process follows US Patent 8, 719,030, to generate voice signal from
the type, pitch, intensity and timbre of the said frame (pitch period).

[0010] Because the period by period process duplicates the natural process of speech
production, and the timbre vectors catches detailed information about the spectrum of the speech
segment, the decoded voice can have a much higher quality than the speech coding algorithm
based on fixed-duration frames and linear prediction coding (LPC) parameterization, and can
still be transmitted with very low bandwidth.

**BRIEF DESCRIPTION OF DRAWINGS**

[0011] FIG. 1 is a block diagram of an encoding system using pitch-synchronous speech
parameterization through timbre vectors.

[0012] FIG. 2 is a block diagram of a decoding system using pitch-synchronous speech
parameterization through timbre vectors.

[0013] FIG. 3 is an example of the asymmetric window for finding pitch marks.

[0014] FIG. 4 is an example of the profile function for finding the pitch marks.

[0015] FIG. 5 shows an example of the spectrograms of the original speech and the decoded
speech.

[0016] FIG. 6 shows the octal values of a sample of encoded speech.

**DETAILED DESCRIPTION OF THE INVENTION**

[0017] Various exemplary embodiments of the present invention are implemented on a
computer system including one or more processors and one or more memory units. In this
regard, according to exemplary embodiments, steps of the various methods described herein are
performed on one or more computer processors according to instructions encoded on a
computer-readable medium.

[0019] FIG. 1 is a block diagram of speech encoding system according to an exemplary
embodiment of the present invention. The input signal 102, typically in PCM (pulse-code
modulation) format, is first convoluted with an asymmetric window 101, to generate a profile
function 104. The peaks 105 in the profile function, with values greater than a threshold, are assigned as pitch marks 106 of the speech signal, which are the frame endpoints in the voice section of the input speech signal 102. The pitch marks only exist for the voiced sections of the speech signal. Using a procedure 107, those frame endpoints are extended into unvoiced and silence sections of the PCM signal, typically by dividing those sections with a constant time interval, in the exemplary embodiment it is 8 msec. A complete set of frame endpoints 108 is generated. Through a segmenter 109, using the said frame endpoints, the PCM signal 102 is then segmented into raw frames 110. In general, the PCM values of the two ends of a raw frame do not match. By performing Fourier analysis on those raw frames, artifacts would be generated. An ends-matching procedure 111 is applied on each raw frame to convert it into a cyclic frame 112 which can be legitimately treated as a sample of a continuous periodic function. Then, a fast Fourier transform (FFT) unit 113 is applied to each said frame 112 to generate an amplitude spectrum 114. The intensity of the spectrum is calculated as the intensity value 124, and then normalized by unit 115. The normalized amplitude spectrum is then expanded using Laguerre functions 116, to generate a set of expansion coefficients, referred to as a timbre vector 117, similar to the timbre vectors in US Patent 8,719,030 and US Patent 8,942,977.

[0020] During the above process, the type of the said frame (pitch period) is determined, see 118. If the amplitude is smaller than a silence threshold, the frame is silence, type 0. If the intensity is higher than the silence threshold but there is no pitch marks, the frame is unvoiced, type 1. For frames bounded by pitch marks, if the amplitude spectrum is concentrated in the low-frequency range (0 to 5 kHz), than the period is voiced, type 3. If the amplitude spectrum in the higher-frequency range (5 to 16 kHz) is substantial, for example, has 30% or more power, then the period is transitional which is voices fricative or a transition frame between voiced and unvoiced, type 2. The type information is encoded in a 2-bit type index, 119. For voiced periods, the pitch value, 120, is conveniently expressed in MIDI unit. Using a pitch codebook 121, the said pitch is scalar-quantized by unit 122. The said intensity 124 is conveniently expressed in decibel (dB) unit. Using an intensity codebook 125, through scalar quantization 126, the intensity index 127 of the frame is generated. Furthermore, using a timbre codebook 128, using vector quantization 129, the timbre index 130 of the frame is generated. Notice that for each type of
Comparing with LPC, the timbre vector is a better subject for vector quantization, because the distance measure (or distortion measure) of the timbre vectors is very simple. According to US Patent 8,719,030 and US Patent 8,942,977, it is

\[ \delta = \sum_{n=0}^{N} \left[ c^{(1)}_n - c^{(2)}_n \right]^2. \]

[0021] FIG. 2 shows the decoding process. From the signals transmitted to the decoder, the 2-bit type index is first fetched. If the frame is silence, a silence PCM, 8 msec of zeros, is sent to the output. If the frame is voiced, type 3, or transitional, type 2, the pitch index 203, the intensity index 204, and the timbre index 205, are fetched. Using the pitch codebook for voiced frames or the pitch codebook for transitional frames, 206, through a look-up procedure 207, the pitch period 208 is identified. Using the intensity codebook for voiced frames or the intensity codebook for transitional frames, 209, through a look-up procedure 210, the intensity of the frame 211 is identified. The intensity 211 and the timbre vector 214 are sent to the waveform recovery unit, 215 through 221, to generate the elementary wave for that frame. The procedure is described in detail in US Patent 8,719,030, especially page 3, lines 42-50. Briefly, using Laguerre transform 215, the timbre vector 214 is converted back to amplitude spectrum 216. Using a phase generator 217 based on Kramers-Kronig relations, the phase spectrum 218 is generated from the amplitude spectrum 216. Using fast Fourier transform (FFT) 219, an elementary waveform 221 is generated. Those elementary waves are lineally superposed using superposition unit 223 according to the time delay 222 defined by the pitch period 208, to generate PCM output 224. For unvoiced frames, type 1, the procedure is identical, except the pitch period, or the frame duration, is a fixed value, which is 8 msec in the current exemplary embodiment. And the phase is random over the entire frequency scale.

[0022] FIG. 3 shows an example of the asymmetric window function (item 101 of FIG. 1) for pitch mark identification. On an interval (-N<n<N), the formula is

\[ w(n) = \pm \sin \left( \frac{\pi n}{N} \right) \left[ 1 + \cos \left( \frac{\pi n}{N} \right) \right]. \]
The ± sign is used to accommodate the polarity of the PCM signals. If a positive sign is taken, the value is positive for 0<n<N, but becomes zero at n=N; and it is negative for -N<n<0, again becomes zero at n=-N. Denoting the PCM signal as p(n), a profile function is generated

\[ f(m) = \sum_{n=-N}^{n<N} w(n)[p(m+n) - p(m+n-1)] \]

Typical result is shown in FIG. 4. Here, 401 is the voice signal. Item 402 indicates the starting point of each pitch period, where the variation of signal is the greatest. 403 is the profile function generated using the asymmetric window function w(n). As shown, the peak positions 404 of the profile function 403 are pointing to the locations with weak variation 405. Its mechanism is also shown in FIG. 4: Each pitch period starts with a large variation of PCM signal at 402. The variation decreases gradually and becomes weak near the end of each pitch period. However, it depends on the relative polarity of the signal and the asymmetric window. If the polarity of the asymmetric window if reversed, then the peak 406 points to the middle of a pitch period, 407. The polarity of the speech signal depends on the microphone and the amplifier circuit, and it should be identified before the encoding process.

[0023] FIG. 5 shows a particular design of the bit allocation for the indices. The design is a proof-of-the-concept coding scheme, not optimized for quality and minimizing the bandwidth. In the said design, only integer number of bytes is used. Therefore, it can be viewed by displaying the octal values of each byte. In the said design, the number of frame repetition is encoded, represented by a repetition index, see below.

[0024] As shown in FIG. 5, the decoder first fetches a byte 501. The highest two bits indicate the type of the frame. If the highest bits are 00, see 502, the frame is silence. The rest 6 bits 503 represent the repetition index, from 0 to 63. In a proof-of-concept prototype, each silence frame is 8 msec. The maximum silence time which can be represented by a single byte is 512 msec, or one half of a second. Such a designation will not cause coding delay. When the speaker is silent, the encoder is waiting for the end of the silence, than output a silence byte. On the decoder side, no signal is transmitted, the output is naturally silence, and until the silence byte arrives.
[0025] If the first byte of a group of bytes has highest bits of 01, see 504 and 505, the frame is unvoiced. The frame duration is also 8 msec. Pitch index is not required. The rest 6 bits are the intensity index, 506. By looking up from an unvoiced intensity codebook 507, the intensity of the said unvoiced frame is determined. Each unvoiced frame is represented by two bytes. The first two bits of the second byte represent number of repetition. If two consecutive frames have the identical timbre vector, the repetition index is 1. If three consecutive frames have the identical timbre vector, the repetition index is 2. The maximum repetition is set to 3. This upper bound is designed for two purposes. First, the intensity of the repeated frames has to be interpolated from the end-point frames. To ensure quality, a limit of four frames is needed. Second, the encoding of four repeated unvoiced frames takes 32 msec. Because the tolerable encoding delay is 70 to 80 msec, as 32 msec is acceptable, too many frames would cause too much encoding delay.

[0026] If the first two bits of the leading byte 512 or 513 are 10 or 11, see 513 and 523, the frame is voiced or transitional, and two following bytes should be fetched from the transmission stream, ch1 and ch2. Similar to the case of unvoiced frames, the rest 6 bits of the leading byte represent intensity index, 514 or 524. By looking up from an intensity codebook, 515 or 525, the intensity is determined. The second byte, 516 or 526, carries a repetition index, 516 or 526, and a pitch index, 518 or 528. The repetition index is limited to 4, and both intensity and pitch have to be linearly interpolated from the two ending-point frames. By looking up from a pitch codebook, 519 or 529, the pitch value is determined. The third byte 520 or 530 is timbre index. By looking up from a timbre codebook, 521 or 531, the timbre vector is determined. Because the type of frame is separated, a codebook size of 256 for each type seems adequate.

[0027] During encoding, the determination of type 2 (transitional) and type 3 (voiced) is based on the spectral distribution, as presented above: If the speech power in a frame with a well-defined pitch period is concentrated in the low-frequency range (0 to 5 kHz), the frame is voiced. If the power in the high frequency range (5 kHz and up) is substantial, then it is a transitional frame. During encoding, different types of frames are treated differently. For voiced frames, below 5 kHz, the phase is generated by the Kramers-Kronig relations; and above 5 kHz, the phase is random. For transitional frames, below 2.5 kHz, the phase is generated by the Kramers-
Knonig relations; and above 2.5 kHz, the phase is random. For unvoiced frames, the phase is random on the entire frequency scale. For details, see US Patent 8,719,030.

[0028] To improve naturalness, jitter may be added to the pitch values. To do this, a few percentages (usually 1% to 3%) of random number is added to the pitch value. Furthermore, shimmer may also be added to the intensity value. To do this, a few percentages (usually 1% to 3%) of random number is added to the intensity value.

[0029] Fast Fourier transform (FFT) is an efficient method for Fourier analysis. However, FFT is much more efficient if the period is an integer power of 2, such as 64, 128, 256, etc. For voiced frames, the pitch period is a variable. In order to utilize FFT, the PCM values in each pitch period is first linearly interpolated into $2^n$ points, in the exemplary embodiment presented here, it is $8 \times 32 = 256$ points. After FFT, the amplitude spectrum is reversely interpolated to the true values of the pitch period.

[0030] The art of building of codebooks is well known in the literature, see for example, A. Gersho and R. M. Gray, "Vector Quantization and Signal Compression", Kluwer Academic Publishers, Boston, 1991. The basic method of building codebooks is the K-means clustering algorithm. A brief summary of the said algorithm can be found in F. Jelinek, "Statistical Methods for Speech Recognition", The MIT Press, Cambridge MA, 1997, page 10-11. Briefly, the K-means clustering process for timbre vectors is as follows: A large database of timbre vectors of a category (voiced, unvoiced or transitional) is collected; choose randomly a fixed number of timbre vectors as seeds; divide the entire vector space to find clusters of timbre vectors closest to each seed; find the center of each cluster. Use the cluster centers as the new seeds, repeat the said process until the centers of clusters converge. The number of seeds, and consequently the number of cluster centers, is called the size of the codebook.

[0031] An example of the encoded speech is shown in FIG. 6, encoded from sentence a0008, spoken by a U.S. English speaker bdl, in ARCTIC databases, published by CMU Language Technologies Institute, 2003. The duration of the speech is 2.5 seconds. The encoded speech has 543 bytes, or 4344 bits. Therefore, the bit rate is $4.344/2.5 = 1.737$ kb/s, in the very-low-bit-rate range. Nevertheless, nearly CD-quality voice is regenerated. The advantages of the current
method are predicable from its principle. First, the maximum bandwidth according to the current invention can be 16 kHz or greater using a PCM speech signal of 32 kHz sampling rate or higher and 16 bit resolution. The legacy speech coding is based on 4 kHz bandwidth (8 kHz PCM sampling rate, 8 bit), the fricatives such as [f] and [s] are not distinguishable. Using the algorithm disclosed in the current invention, the fricatives [f] and [s] are clearly distinguishable. Furthermore, while the legacy low-bit-rate speech coding is based on an all-pole model of speech signal which fails to represent the nasal sounds, the technology disclosed in the current invention reproduces the entire spectrum, and the nasal sounds are reproduced faithfully.

[0032] While this invention has been described in conjunction with the exemplary embodiments outlined above, it is evident that many alternatives, modifications and variations will be apparent to those skilled in the art. Accordingly, the exemplary embodiments of the invention, as set forth above, are intended to be illustrative, not limiting. Various changes may be made without departing from the spirit and scope of the invention.
CLAIMS

I claim:

1. A method of speech communication from a transmitter to a receiver using a plurality of processors comprising an encoder to compress the speech signal into a digital form and a decoder to recover speech signal from the said compressed digital form comprising:

   (A) an encoder in the transmitter comprising the following elements:
   
   segment the voice-signal into non-overlapping frames, wherein for voiced sections the frames are pitch periods and for unvoiced sections the frame duration is a constant;
   
   identify the type of a said frame to generate a type index;
   
   identify the pitch period of a said frame from the segmentation process;
   
   generate amplitude spectra of a said frame using Fourier analysis;
   
   generate an intensity parameter of a said frame from the amplitude spectrum;
   
   transform the said amplitude spectrum into timbre vectors using Laguerre functions;
   
   apply vector quantization to the said timbre vector using a timbre-vector codebook to generate a timbre index;
   
   apply scalar quantization to said intensity parameter using an intensity codebook to generate an intensity index;
   
   apply scalar quantization to said pitch period with a pitch codebook to generate a pitch index;
   
   transmit the type index, intensity index, pitch index and timbre index to the receiver;

   (B) a decoder in the receiver comprising the following elements:
   
   take the transmitted intensity index, look-up into the intensity codebook to identify the intensity;
   
   take the transmitted pitch index, look-up into the pitch codebook to identify the pitch;
take the transmitted timbre index, look-up into the timbre-vector codebook to identify the timber vector;

inverse transform the said timbre vector into amplitude spectra using Laguerre functions;

generate phase spectrum from the amplitude spectrum using Kramers-Kronig relations;

use fast Fourier transform to generate an elementary waveform from the said amplitude spectrum, phase spectrum, and intensity;

superpose the said elementary waves according to the timing provided by the pitch period to generate an output speech signal.

2. The method of claim 1, wherein the speech signal is segmenting by steps comprising:

convolute the speech signal with an asymmetric window to generate a profile function;

take the peaks of the said profile function that is greater than a threshold as the segmentation points in the voiced section of the said speech signal;

extend the segmentation points to unvoiced sections where no peaks in the said profile function above a threshold with a fixed time interval.

3. The method of claim 1, wherein the pitch period is defined as the time difference of two consecutive peaks above a threshold value in the said profile function.

4. The method of claim 1, wherein the types of a frame is defined as:

type 0, silence, when the intensity is smaller than a silence threshold;

type 1, unvoiced, when there is no pitch marks detected;

type 2, transitional, when a pitch mark is found and the speech power in the upper frequency range is greater than a percentage, as an example, greater than 30% above 5 kHz;

type 3, voiced, when a pitch mark is found and the speech power in the upper frequency range is smaller than a percentage, as an example, smaller than 30% above 5 kHz.

5. The method of claim 1, wherein the timbre vector codebooks are constructed using the K-means clustering algorithm comprising:

collect a large number of timbre vectors of a given type (voiced, unvoiced, or transitional) from a database of speech;
according to the desired size N of codebook, randomly select N timber vectors as seeds;
for each seed, find the timber vectors closest to the said seed to form a cluster;
find the center of the said cluster;
use the said cluster centers as the new seeds, repeat the process until the values converge.

6. The method of claim 1, wherein the intensity codebooks and the pitch codebooks are
constructed using scalar quantization from large databases;

7. The method of claim 1, wherein the bit rate of encoded speech is further reduced by using a
repetition index to represent repeated indices;

8. The method of claim 1, wherein the naturalness of output speech is improved by adding
shimmer to the intensity values.

9. The method of claim 1, wherein the naturalness of output speech is improved by adding
jitter to the pitch values.

10. The method of claim 1, wherein the said Fourier analysis in the encoding stage is executed
using a scaled fast Fourier transform (FFT) comprising:
    interpolate the PCM values in a pitch period into an integer power of 2, for example 256;
    perform FFT on the said interpolated signals to generate an amplitude spectrum;
    linearly interpolate the said amplitude spectrum to the correct frequency scale.

11. An apparatus of speech communication from a transmitter to a receiver using a plurality
of processors comprising an encoder to compress the speech signal into a digital form and a
decoder to recover speech signal from the said compressed digital form comprising:
    (A) an encoder in the transmitter comprising the following elements:
        segment the voice-signal into non-overlapping frames, wherein for voiced sections the frames
        are pitch periods and for unvoiced sections the frame duration is a constant;
        identify the type of a said frame to generate a type index;
        identify the pitch period of a said frame from the segmentation process;
        generate amplitude spectra of a said frame using Fourier analysis;
generate an intensity parameter of a said frame from the amplitude spectrum;

transform the said amplitude spectrum into timbre vectors using Laguerre functions;

apply vector quantization to the said timbre vector using a timbre-vector codebook to generate a timbre index;

apply scalar quantization to said intensity parameter using an intensity codebook to generate an intensity index;

apply scalar quantization to said pitch period with a pitch codebook to generate a pitch index;

transmit the type index, intensity index, pitch index and timbre index to the receiver;

(B) a decoder in the receiver comprising the following elements:

  take the transmitted intensity index, look-up into the intensity codebook to identify the intensity;

  take the transmitted pitch index, look-up into the pitch codebook to identify the pitch;

  take the transmitted timbre index, look-up into the timbre-vector codebook to identify the timber vector;

  inverse transform the said timbre vector into amplitude spectra using Laguerre functions;

  generate phase spectrum from the amplitude spectrum using Kramers-Knonig relations;

  use fast Fourier transform to generate an elementary waveform from the said amplitude spectrum, phase spectrum, and intensity;

  superpose the said elementary waves according to the timing provided by the pitch period to generate an output speech signal.

12. The apparatus of claim 11, wherein the speech signal is segmenting by steps comprising:

  convolute the speech signal with an asymmetric window to generate a profile function;

  take the peaks of the said profile function that is greater than a threshold as the segmentation points in the voiced section of the said speech signal;

  extend the segmentation points to unvoiced sections where no peaks in the said profile function above a threshold with a fixed time interval.
13. The apparatus of claim 11, wherein the pitch period is defined as the time difference of two consecutive peaks above a threshold value in the said profile function.

14. The apparatus of claim 11, wherein the types of a frame is defined as:

    type 0, silence, when the intensity is smaller than a silence threshold;
    type 1, unvoiced, when there is no pitch marks detected;
    type 2, transitional, when a pitch mark is found and the speech power in the upper frequency range is greater than a percentage, as an example, greater than 30% above 5 kHz;
    type 3, voiced, when a pitch mark is found and the speech power in the upper frequency range is smaller than a percentage, as an example, smaller than 30% above 5 kHz.

15. The apparatus of claim 11, wherein the timbre vector codebooks are constructed using the K-means clustering algorithm comprising:

    collect a large number of timbre vectors of a given type (voiced, unvoiced, or transitional) from a database of speech;
    according to the desired size N of codebook, randomly select N timber vectors as seeds;
    for each seed, find the timber vectors closest to the said seed to form a cluster;
    find the center of the said cluster;
    use the said cluster centers as the new seeds, repeat the process until the values converge.

16. The apparatus of claim 11, wherein the intensity codebooks and the pitch codebooks are constructed using scalar quantization from large databases;

17. The apparatus of claim 11, wherein the bit rate of encoded speech is further reduced by using a repetition index to represent repeated indices;

18. The apparatus of claim 11, wherein the naturalness of output speech is improved by adding shimmer to the intensity values.

19. The apparatus of claim 11, wherein the naturalness of output speech is improved by adding jitter to the pitch values.

20. The apparatus of claim 11, wherein the said Fourier analysis in the encoding stage is executed using a scaled fast Fourier transform (FFT) comprising:
interpolate the PCM values in a pitch period into an integer power of 2, for example 256; perform FFT on the said interpolated signals to generate an amplitude spectrum; linearly interpolate the said amplitude spectrum to the correct frequency scale.
ASymmetric Window → PCM SIGNAL

CONVOLUTION → SEGMENTER

PROFILE FUNCTION → RAW FRAMES

PICKING PEAKS → ENDS MATCHING

PITCH MARKS AS FRAME ENDPOINTS → CYCLIC FRAMES

EXTENSION TO UNVOICED AND SILENCE PCMS → FFT

COMPLETE SET OF FRAME ENDPOINTS → AMPLITUDE SPECTRA

NORMALIZATION → LAGUERRE EXPANSION

SEGMENTER → TYPE (0, 1, 2, 3)

2 BIT TYPE INDEX → ENCODED SPEECH

PITCH CODEBOOKS → PITCH INDEX

SCALAR QUANTIZATION → INTENSITY CODEBOOKS

INTENSITY INDEX → SCALAR QUANTIZATION

TIMBRE CODEBOOKS → VECTOR QUANTIZATION

TIMBRE INDEX
FIG. 5