(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number: **0 140 249**
B1

(12) # EUROPEAN PATENT SPECIFICATION

(45) Date of publication of patent specification: **10.08.88**

(51) Int. Cl.[4]: **G 10 L 9/18**

(21) Application number: **84112266.6**

(22) Date of filing: **12.10.84**

(54) **Speech analysis/synthesis with energy normalization.**

(30) Priority: **13.10.83 US 541410**
13.10.83 US 541497

(43) Date of publication of application:
08.05.85 Bulletin 85/19

(45) Publication of the grant of the patent:
10.08.88 Bulletin 88/32

(84) Designated Contracting States:
DE FR GB

(56) References cited:
EP-A-0 027 066
EP-A-0 047 589
FR-A-2 308 248
FR-A-2 380 612
FR-A-2 451 680
US-A-4 071 695
US-A-4 280 192
US-A-4 351 983

(73) Proprietor: **TEXAS INSTRUMENTS INCORPORATED**
13500 North Central Expressway
Dallas Texas 75265 (US)

(72) Inventor: **Doddington, George R.**
910 St. Lukes Drive
Richardson, TX 75080 (US)
Inventor: **Papamichalis, Panos E.**
1704 Blake Drive
Richardson, TX 75081 (US)

(74) Representative: **Leiser, Gottfried, Dipl.-Ing. et al**
Patentanwälte Prinz, Leiser, Bunke & Partner
Manzingerweg 7
D-8000 München 60 (DE)

(56) References cited:
ELECTRONICS LETTERS, vol. 9, no. 14, 12th July 1973, pages 298-300, Stevenage, GB; M.G. CROLL et al.: " 'Nearly instantaneous' digital compandor for transmitting six sound-programme signals in a 2.048Mbit/s multiplex"
NEW ELECTRONICS, vol. 15, no. 2, January 1982, pages 30-32, London, GB; B. DANCE: "A digital speech compressor"

Courier Press, Leamington Spa, England.

EP 0 140 249 B1

(56) References cited:
ICASSP 83, PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, Boston, Massachusetts, 14th-16th April 1983, vol. 2, pages 511-514, IEEE, New York, US; J.A. FELDMAN et al.: "A custom IC for automatic gain control in LPC vocoders"

IBM TECHNICAL DISCLOSURE BULLETIN, vol. 20, no. 12, May 1978, pages 5437-5440, New York, US; S.J. BOIES et al.: "Amplitude-detection method for producing rate-controlled speech"

IBM TECHNICAL DISCLOSURE BULLETIN, vol. 25, no. 7B, December 1982, pages 3678-3680, New York, US; D.R. IRVIN: "Voice activity detector"

ICASSP 79, 1979 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, Washington, D.C., 2nd-4th April 1979, pages 212-215, IEEE, New York, US; R.D. PREUSS: "A frequency domain noise cancelling preprocessor for narrowband speech communications systems"

Description

Background and summary of the invention

The present invention relates to voice coding systems and in particular to a voice mail system and a method of encoding speech as defined in the precharacterizing parts of claims 1, 5, 7 and 8.

A speech encoding system of the type referred to in the preamble of claim 7 is disclosed in EP—A—47 589.

A very large range of applications exists for voice coding systems, including voice mail in microcomputer networks, voice mail sent and received over telephone lines by microcomputers, user-programmed synthetic speech, etc.

In particular, the requirements of many of these applications are quite different from those of simple speech synthesis applications wherein synthetic speech can be carefully encoded and then stored in a ROM or on disk. In such applications, high speed computers with elaborate algorithms, combined with hand tweaking, can be used to optimize encoded speech for good intelligibility and low bit requirements. However, in many other requirements, the speech encoding step does not have such large resources available. This is most obviously true in voice mail microcomputer networks, but it is also important in applications where a user may wish to generate his own reminder messages, diagnostic messages, signals during program operation, etc. For example, a microcomputer system wherein the user could generate synthetic speech messages in his own software would be highly desirable, not only for the individual user but also for the software production houses which do not have trained speech scientists available.

A particular problem in such application is energy variation. That is, not only will a speaker's voice intensity typically contain a large dynamic range related to sentence inflection, but different speakers will have different volume levels, and the same speaker's voice level may vary widely at different times. Untrained speakers are especially likely to use nonuniform uncontrolled variations in volume, which the listener normally ignores. This large dynamic range would mean that the voice coding method used must accommodate a wide dynamic range, and therefore an increased number of bits would be required for coding at reasonable resolution.

However, if energy normalization can be used (i.e. *all* speech adjusted to approximately a constant energy level) these problems are ameliorated.

Energy normalization also improves the intelligibility of the speech received. That is, the dynamic range available from audio amplifiers and loudspeakers is much less than that which can easily be perceived by the human ear. In fact, the dynamic range of loudspeakers is typically much less than that of microphones. This means that a dynamic range which is perfectly intelligible to a human listener may be hard to understand if communicated through a loudspeaker, even if absolutely perfect encoding and decoding is used.

The problems of intelligibility is particularly acute with audio amplifiers and loudspeakers which are not of extremely high fidelity. However, compact low-fidelity loudspeakers must be used in most of the most attractive applications for voice analysis/synthesis, for reasons of compactness, ruggedness, and economy.

A further desideratum is that, in many attractive applications, the person listening to synthesized speech should not be required to twiddle a volume control frequently. Where a volume control is available, dynamic range can be analog-adjusted for each received synthetic speech signal, to shift the narrow window provided by the loudspeaker's narrow dynamic range, but this is obviously undesirable for voice mail systems and many other applications.

In the prior art, analog automatic gain controls have been used to achieve energy normalization of raw signals. However, analog automatic gain controls distort the signal input to the analog to digital converter. That is, where (e.g.) reflection coefficients are used to encode speech data, use of an automatic gain control in the analog signal will introduce error into the calculated reflection coefficients. While it is hard to analyze the nature of this error, error is in fact introduced. Moreover, use of an analog automatic gain control requires an analog part, and every introduction of special analog parts into a digital system greatly increases the cost of the digital system. If an AGC circuit having a fast response is used, the energy levels of consecutive allophones may be inappropriate. For example, in the word "six" the sibilant /s/ will normally show a much lower energy than the vowel /I/. If a fast-response AGC circuit is used, the energy-normalized-word "six" is left with a sound extremely hissy, since the initial /s/ will be raised to the same energy as the /i/, inappropriately. Even if a slower-response AGC circuit is used, substantial problems still may exist, such as raising the noise floor up to signal levels during periods of silence, or inadequate limiting of a loud utterance following a silent period.

Thus, it is an object of the present invention to provide a digital system which can perform energy normalization of voice signals.

It is a further object of the present invention to provide a method for energy normalization of voice signals which will not overemphasize initial constants.

It is a further object of the present invention to provide a method for energy normalization of voice signals which can rapidly respond to energy variations in a speaker's utterance, without excessively distorting the relative energy levels of adjacent allophones with an utterance.

A further general problem with energy normalization is caused by the existence of noise during silent periods. That is, if an energy

normalization system brings the noise floor up towards the expected normal energy level during periods when no speech signal is present, the intelligibility of speech will be degraded and the speech will be unpleasant to listen to. In addition, substantial bandwidth will be wasted encoding noise signals during speech silence periods.

It is a further object of the present invention to provide a voice coding system which will not waste bandwidth on encoding noise during silent periods.

The present invention solves the problems of energy normalization digitally, by using look-ahead energy normalization. That is, an adaptive energy normalization parameter is carried from frame to frame during a speech analysis portion of an analysis-synthesis system. Speech frames are buffered for a fairly long period, e.g. 1/2 second, and then are normalized according to the current energy normalization parameter. That is, energy normalization is "look ahead" normalization in that each frame of speech (e.g. each 20 millisecond interval of speech) is normalized according to the energy normalization value from much later, e.g. from 25 frames later. The energy normalization value is calculated for the frames as received by using a fast-rising slow-falling peak-tracking value.

In a further aspect of the present invention, a novel silence suppression scheme is used. Silence suppression is achieved by tracking 2 additional energy contours. One contour is a slow-rising fast-falling value, which is updated only during unvoiced speech frames, and therefore tracks a lower envelope of the energy contour. (This in effect tracks the ambient noise level). The other parameter is a fast-rising slow-falling parameter, which is updated only during voiced speech frames, and thus tracks an upper envelope of the energy contour. (This in effect tracks the average speech level). A threshold value is calculated as the maximum of respective multiples of these 2 parameters, e.g. the greater of: (5 times the lower envelope parameter), and (one fifth of the upper envelope parameter). Speech is not considered to have begun unless a first frame which *both* has an energy above the threshold level *and* is also voiced in detected. In that case, the system then backtracks among the buffered frames to include as "speech" all immediately preceding frames which also have energy greater than the threshold. That is, after a period during which the frames of parameters received have been identified as silent frames, all succeeding frames are tentively identified as silent frames, until a super-threshold-energy voiced frame is found. At that point, the silence suppression system backtracks among frames immediately preceding this super-threshold energy voiced frame until a broken string subthreshold-energy frames at least to 0.4 seconds long is found. When such a 0.4 second interval of silence is found, backtracking ceases, and only those frames after the 0.4 seconds of silence and before the first voiced super-threshold energy frame are identified as non-silent frames.

At the end of speech, when a voice frame is detected having an energy below the threshold T, a waiting counter is started. If the waiting reaches an upper limit (e.g. 0.4 seconds), without the energy again increasing above T, the utterance is considered to have stopped. The significance of the speech/silence decision is that bits are not wasted on encoding silent frames, energy tracking is not distorted by the presence of silent frames as discussed above, and long utterances can be input from an untrained speakers, who are likely to leave very long silences between consecutive words in a sentence.

In a voice mail system according to the pre-characterising part of claim 1 these objects are achieved by means for normalising the energy parameter of each said speech frame, wherein said energy parameter of each frame is normalised primarily with respect to an energy parameter of a subsequent frame occurring at least 0.1 seconds after said each frame.

According to the invention the method of encoding speech as defined in the pre-characterising part of claim 5 is characterised in that the energy parameters of each of said speech frames is normalised with respect to an energy parameter of a subsequent frame occurring later than each said respective frame by at least 0.1 seconds, with normalisation being done prior to encoding said speech parameters into the data channel.

The speech encoding system as defined in the precharacterising part of claim 7 is characterised by means for normalising the energy parameter of each said speech frame with respect to the energy parameter of a subsequent frame occurring at least 0.1 seconds after said each frame, and wherein said silence suppression means identifies each said frame as silent or non-silent by comparing the energy parameter of each successive one of said frames against a function of first and second adaptively updated threshold values, said first adaptively updated threshold value corresponding to a multiple of an upper envelope of said successive energy parameters of successive ones of said frames and said second threshold value corresponding to a multiple of a lower envelope of said successive values of said frames.

The voice mail system as defined in the pre-characterising part of claim 8 is characterised by means for normalising the energy parameter of each said speech frame with respect to the energy parameter of a subsequent frame occurring at least 0.1 seconds after said each frame, and wherein said silence suppression means identifies each said frame as silent or non-silent by comparing the energy parameter of each successive one of said frames against a function of first and second adaptively updated threshold values, said first adaptively updated threshold value corresponding to a multiple of an upper envelope of said successive energy parameters of

successive ones of said frames and said second threshold value corresponding to a multiple of a lower envelope of said successive values of said frames.

Brief description of the drawings

The present invention will be described with reference to the accompanying drawings, which are hereby incorporated by reference and attested to by the attached Declaration, wherein:

Figure 1 shows one aspect of the present invention, wherein an adaptively normalized energy level ENORM is derived from the successive energy levels of a sequence of speech frames;

Figure 2 shows a further aspect of the present invention, wherein a look-ahead energy normalization curve ENORM* is used for normalization;

Figure 3 shows a further aspect of the present invention, used in silence suppression, wherein high and low envelope curves are continuously maintained for the energy values of a sequence of speech input frames;

Figure 4 shows a further aspect of the invention, wherein the EHIGH and ELOW curves of Figure 3 are used to derive a threshold curve T; and

Figure 5 shows a sample system configuration for practicing the present invention.

Description of the preferred embodiments

The present invention provides a novel speech analysis/synthesis system, which can be configured in a wide variety of embodiments. However, the presently preferred embodiment uses a VAX 11/780 computer, coupled with a Digital Sound Corporation Model 200 A/D and D/A converter to provided high-resolution high-bit-rate digitizing and to provide speech synthesis. Naturally, a conventional microphone and loudspeaker, with an analog amplifier such as a Digital Sound Corporation Model 240, are also used in conjunction with the system.

However, the present invention contains novel teachings which are also particularly applicable to microcomputer-based systems. That is, the high resolution provided by the above digitizer is not necessary, and the computing power available on the VAX is also not necessary. In particular, it is expected that a highly attractive embodiment of the present invention will use a TI Professional Computer (TM), using the built in low-quality speaker and an attached microphone as discussed below.

The system configuration of the presently preferred embodiment is shown schematically in Figure 5. That is, a raw voice input is received by microphone, amplified by microphone amplifier, and digitized by D/A converter. The D/A converter used in the presently presently embodiment, as noted, is an expensive high-resolution instrument, which provides 16 bits of resolution at a sample rate of 8 kHz. The data received at this high sample rate will be transformed to provide speech parameters at a desired frame rate. In the presently preferred embodiment the frame rate is 50 frames per second, but the frame period can easily range between 10 milliseconds and 30 milliseconds, or over an even wider range.

In the presently preferred embodiment, linear predictive coding based analysis is used to encode the speech. That is, the successive samples (at the original high bit rate, of, in this example, 8000 per second) are used as inputs to derive a set of linear predictive coding parameters, for example 10 reflection coefficants k_1—k_{10} plus pitch and energy, as described below.

In practicing the present invention, the audible speech is first translated into a meaningful input for the system. For example, a microphone within range of the audible speech is connected to a microphone preamplifier and to an analog-to-digital converter. In the presently preferred embodiment, the input stream is sampled 8000 times per second, to an accuracy of 16 bits. The stream of input data is then arbitrarily divided up into successive "frames", and, in the presently preferred embodiment, each frame is defined to include 160 samples. That is, the interval between frames is 20 msec, but the LPC parameters of each frame are calculated over a range of 240 samples (30 msec).

In one embodiment, the sequence of samples in each speech input frame is first transformed into a set of inverse filter coefficients a_k, as conventionally defined. See, e.g., Makhoul, "Linear Prediction: A Tutorial Review", proceedings of the IEEE, Volume 63, page 561 (1975). That is, in the linear prediction model, the a_k's are the predictor coefficients with which a signal S_k in a time series can be modeled as the sum of an input u_k and a linear combination of past values S_{k-n} in the series. That is:

$$S_n = -\sum_{k=1}^{p} a_k s_{n-k} + G u_n \qquad (1)$$

Each input frame contains a large number of sampling points, and the sampling points within any one input frame can themselves be considered as a time series. In one embodiment, the actual derivation of the filter coefficients a_k for the sample frame is as follows: First, the time-series autocorrelation values R_i are computed as

$$R(i) = \sum_{n} s_n s_{n+i} \qquad (2)$$

where the summation is taken over the range of samples within the input frame. In this embodiment, 11 autocorrelation values are calculated (R_0—R_{10}). A recursive procedure is now used to derive the inverse filter coefficients as follows:

$$E_o = R(o) \qquad (3)$$

$$k_i = -[R(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)]/E_{i-1} \qquad (4)$$

$$a_i^{(i)}=k_i$$
$$a_j^{(i)}=a_j^{(i-1)}+k_i a_{i-j}^{(i-1)},$$

for

$$1\leq j\leq i-1 \quad (5)$$

$$E_i=(1-k_i^2)E_{i-1} \quad (6)$$

These equations are solved recursively for: i=1, 2, ..., up to the model order p (p=10 in this case). The last iteration gives the final a_k values.

The foregoing has described an embodiment using Durbin's recursive procedure to calculate the a_k's for the sample frame. However, the presently preferred embodiment uses a procedure due to Leroux-Gueguen. In this procedure, the normalized error energy E (i.e. the self-residual energy of the input frame) is produced as a direct byproduct of the algorithm. The Leroux-Gueguen algorithm also produces the reflection coefficients (also referred to as partial correlation coefficients) k_i. The reflection coefficients k_r are very stable parameters, and are insensitive to coding errors (quantization noise).

The Leroux-Gueguen procedure is set forth, for example, in IEEE Transactions on Acoustic Speech and Signal Processing, page 257 (June 1977). This algorithm is a recursive procedure, defined as follows:

$$k_h=-e_{h+1}^{(h)}/e_o^{(h)} \quad (7)$$

$$e_o^{(h+1)}=e_o^{(h)}(1-k_h^2) \quad (8)$$

$$e_i^{(h+i)}=e_i^{(h)}+k_h e_{h+1-i}^{(h)} \quad (9)$$

This algorithm computres the reflection coefficients k_i using as intermediaries impulse response estimates e_k rather then the filter coefficients a_k.

Linear predictive coding models generally are well known in the art, and can be found extensively discussed in such references as Rabiner and Schafer, *Digital Processing of Speech Signals* (1978), Markel and Gray, *Linear Predictive Coding of Speech* (1976). It should be noted that the excitation coding transmitted need not be merely energy and pitch, but may also contain some additional information regarding a residual signal. For example, it would be possible to encode a bandwidth of the residual signal which was an integral multiple of the pitch, and approximately equal to 1000 Hz, as an excitation signal. Many other well-known variations of encoding the excitation information can also be used alternatively. Similarly, the LPC parameters can be encoded in various ways. For example, as is also well known in the art, there are numerous equivalent formulations of linear predictive coefficients. These can be expressed as the LPC filter coefficients a_k, or as the reflection coefficients k_i, or as the autocorrelations R_i, or as other parameter sets such as the impulse response estimates parameters E(i) which are provided by the LeRoux-Guegen procedure. Moreover, the LPC model order is not necessarily 10, but can be 8, 12, 14, or other.

Moreover, it should be noted that the present invention does not necessarily have to be used in combination with an LPC speech encoding model at all. That is, the present invention provides an energy normalization method which digitally modifies only the energy of each of a sequence of speech frames, with regard to only the energy and voicing of each of a sequence of speech frames. Thus, the present invention is applicable to energy normalization of the systems using any one of a great variety of speech encoding methods, including transform techniques, formant encoding techniques, etc.

Thus, after the input samples have been converted to a sequence of speech frames each having a data vector including an energy value, the present invention operates on the energy value of the data vectors. In the presently preferred embodiment, the encoded parameters are the reflection coefficients k_1—k_{10}, the energy, and pitch. (The pitch parameter includes the voicing decision, since an unvoiced frame is encoded as pitch=zero).

The novel operations in the system of the present invention begin at this point. That is, a sequence of encoded frames, each including an energy parameter and modeling parameters, is provided as the raw output of the speech analysis section. Note that, at this stage, the resolution of the energy parameter coding is much higher than it will be in the encoded information which is actually transmitted over the communications or storage channel 40. The way in which the present invention performs energy normalization on successive frames, and suppresses coding of silient frames, may be seen with regard to the energy diagrams of Figures 1—4. These show examples of the energy values E(i) seen in successive frames i within a sequence of frames, as received as raw output in the speech analysis section.

An adaptive parameter ENORM(i) is then generated, approximately as shown in Figure 1. That is, ENORM(0) is an initial choice for that parameter, e.g. ENORM(0)=100. ENORM is subsequently updated, for each successive frame, as follows:

If E(i) is greater than ENORM(i−1), then ENORM (i) is set equal to alpha times E(i)+(1-alpha) times ENORM(i−1);

Otherwise, ENORM(i) is set equal to beta times E(i)+(1-beta) times ENORM(i−1), where alpha is given a value close to 1 to provide a fast rising time constant (preferably about 0.1 seconds), and Beta has given a value close to 0, to provide a slow falling time constant (preferably in the neighborhood of 4 seconds).

It should be noted that in the software attached as appendix A, which is hereby incorporated by reference, the parameter alpha is stated as "alpha-up", and the parameter beta is stated as "alpha-down". Thus, the adaptive parameter ENORM provides an envelope tracking measure

which tracks the peak energy of the sequence of frames I.

This adaptive peak-tracking parameter ENORM(i) is used to normalize the energy of the frames, but this not done directly. The energy of each frame I is normalized by dividing it by a look ahead normalized energy ENORM*(i), where ENORM*(i) is defined to be equal to ENORM(i+d), where d represents a number of frames of delay which is typically chosen to be equivalent to 1/2 second (but must be at least 0.1 seconds. Thus, the energy E(i) of each frame is normalized by dividing by the normalized energy ENORM*(i):

E*(i) is set equal to E(i)/ENORM*(i). This is accomplished by buffering a number of speech frames equal to the delay d, so that the value of ENORM for the last frame loaded into the buffer provides the value of ENORM* for the oldest frame in the buffer, i.e. for the frame currently being taken out of the buffer.

The introduction of this delay in the energy normalization means that the energy of initial low-energy periods will be normalized with respect to the energy of immediately following high-energy periods, so that the relative energy of initial consonants will not be distorted. That is, unvoiced frames of speech will typically have an energy value which is much lower than that of voiced frames of speech. Thus, in the word "six" the initial allophone /s/ should be normalized with respect to the energy level of the vowel allophone /i/. If the allophone /s/ is normalized with respect to its own energy, then it will be raised to an improperly high energy, and the initial consonant /s/ will be greatly overemphasized.

Since the falling time constant (corresponding to the parameter beta) is so long, energy normalization at the end of a word will not be distorted by the approximately zero-energy value of the following frames of silence. (In addition, when silence suppression is used, as is preferable, the silence suppression will prevent ENORM from falling very far in this situation). That is, for a final unvoiced consonant, the long time constant corresponding to beta will mean that the energy normalization value ENORM of the silient frames 1/2 second after the end of a word will be still be dominated by the voiced phonemes immediately preceding the final unvoiced consonant. Thus, the final unvoiced constant will be normalized with respect to preceding voiced frames, and its energy also will not be unduly raised.

Thus, the foregoing steps provide a normalized energy E*(i) for each speech frame i. In the presently preferred embodiment, a further novel step is used to suppress silent periods. As shown in the diagram of Figure 5, silence detection is used to selectively prevent certain frames from being encoded. Those frames which are encoded are encoded with a normalized energy E*(i), together with the remaining speech parameters in the chosen model (which in the presently preferred embodiment are the pitch P and the reflection coefficients k_1—k_{10}).

Silence suppression is accomplished in a further novel aspect of the present invention, by carrying 2 envelope parameters: ELOW and EHIGH. Both of these parameters are started from some initial value (e.g. 100) and then are updated depending on the energy E(i) of each frame i and on the voiced or unvoiced status of that frame. If the frame is unvoiced, then only the lower parameter ELOW is updated as follows:

If E(i) is greater than ELOW, then ELOW is set equal to gamma times E(i)+(1-gamma) times ELOW;

otherwise, ELOW is set equal to delta times E(i)+(1-delta) times ELOW,

where gamme corresponds to a slow rising time constant (typically 1 second), and delta corresponds to a fast falling time constant (typically 0.1 second). Thus, ELOW in effect tracks a lower envelope of the energy contour of EI. The parameters gamma and delta are referred to in the accompanying software as ALOWUP and ALOWDN.

If the frame I is voiced, then only EHIGH is updated, as follows:

If E(i) is greater then EHIGH, the EHIGH is set equal to epsilon times E(i)+(1-epsilon) times EHIGH;

otherwise, EHIGH is set equal to zeta times E(i)+(1-zeta) times EHIGH,

where epsilon corresponds to a fast rising time constant (typically 0.1 seconds), and zeta corresponds to a fast falling time constant (typically 1 second). Thus, EHIGH tracts an upper envelope of the energy contour. The parameters ELOW and EHIGH are shown in Figure 3. Note that the parameter EHIGH is not updated during the initial unvoiced series of frames, and the parameter ELOW is not disturbed during the following voiced series of frames.

The 2 envelope parameters ELOW and EHIGH are then used to generate 2 threshold parameters TLOW and THIGH, defined as:

TLOW=PL times ELOW

THIGH=PH times EHIGH,

where PL and PH are scaling factors (e.g. PL=5 and PH=0.2). A threshold T is then set as the maximum of TLOW and THIGH.

Based on this threshold T, a decision is made whether a frame is nonsilent or silent, as follows:

If the current frame is a silient frame, all following frames will be tentatively assumed to be silent unless a voiced super-threshold-energy (and therefore nonsilent) frame is deteched. The frames tentatively assumed to be silent will be stored in a buffer (preferable containing at least one second of data), since they may be identified later as *not* silent. A speech frame is detected only when some frame is found which has a frame energy E(i) greater than the threshold T *and which is voiced*. That is, an unvoiced super-threshold-energy frame is not by itself enough to cause a decision that speech has begun. However, once a voiced high energy frame is found, the prior frames in the buffer are reexamined, and all immediately preceding unvoiced frames which

have an energy greater than T are then identified as nonsilent frames. Thus, in the sample word "six", the unvoiced super-threshold-energy frames in the constant /s/ would not immediately trigger a decision that a speech signal had begun, but, when the voiced super-threshold-energy frames in the /i/ are detected, the immediately preceding frames are reexamined, and the frames corresponding to t /s/ which have energy greater than T are then also designated as "speech" frames.

If the current frame is a "speech" (nonsilent) frame, the end of the word (i.e. the beginning of "silent" frames which need not be encoded) is detected as follows. When a voiced frame is found which has its energy E(i) less than T, a waiting counter is started. If the waiting reaches an upper limit (e.g. 0.4 seconds) without the energy ever rising above T, then speech is determined to have stopped, and frames after the last frame which had energy E(i) greater than T are considered to be silent frames. These frames are therefore not encoded.

It should be noted that the energy normalization and silence suppression features of the system of the present invention are both dependent in important ways on the voicing decision. It is preferable, although not strictly necessary, that the voicing decision be made by means of a dynamic programming procedure which makes pitch and voicing decision simultaneously, using an interrelated distance measure.

The actual encoding can now be performed with a minimum bit rate. In the presently preferred embodiment, 5 bits are used to encode the energy of each frame, 3 bits are used for each of the ten reflection coefficients, and 5 bits are used for the pitch. However, this bit rate can be further compressed by one of the many variations of delta coding, e.g. by fitting a polynomial to the sequence of parameter values across successive frames and then encoding merely the coefficients of that polynomial, by simple linear delta coding, or by any of the various well known methods.

In a further attractive contemplated embodiment of the invention, an analysis system as described above is combined with speech synthesis capability, to provide a voice mail station, or a station capable of generating used-generated spoken reminder messages. This combination is easily accomplished with minimal required hardware addition. The encoded output of the analysis section, as described above, is connected to a data channel of some sort. This may be a wire to which an RS 232 UART chip is connected, or may be a telephone line accessed by a modem, or may be simply a local data buss which is also connected to a memory board or memory chips, or may of course be any of a tremendous variety of other data channels. Naturally, connection to any of these normal data channels is easily and conveniently made two way, so that data may be received from a communications channel or recalled from memory. Such data received from the channel will thus contain a plurality of speech parameters, including an energy value.

In the presently preferred embodiment, where LPC speech modeling is used, the encoded data received from the data channel will contain LPC filter parameters for each speech frame, as well as some excitation information. In the presently preferred embodiment, the data vector for each speech frame contains 10 reflection coefficients as well as pitch and energy. The reflection co-efficients configure a tense-order lattice filter, and an excitation signal is generated from the excitation parameters and provided as input to this lattice filter. For example, where the excitation parameters are pitch and energy, a pulse, at intervals equal to the pitch period, is provided as the excitation function during voiced frames (i.e. during frames when the encoded value of pitch is non zero), and pseudo-random noise is provided as the excitation function when pitch has been encoded as equal to zero (unvoiced frames). In either case, the energy parameter can be used to define the power provided in the excitation function. The output of the lattice filter provides the LPC-modeled synthetic signal, which will typically be of good intelligible quality, although not absolutely transparent. This output is then digital-to-analog converted, and the analogue output of the d-a converter is provided to an audio amplifier, which drives a loudspeaker or headphones.

In a further attractive alternative embodiment of the present invention, such a voice mail system is configured in a microcomputer-based system. This configuration uses a 8088-based system, together with a special board having a TMS 320 numeric processure chip mounted thereon. The fast multiple provided by the TMS 320 is very convenient in performing signal processing functions. A pair of audio amplifiers for input and output is also provided on the speech board, as is an 8 bit mu-law codec. The function of this embodiment is essentially identical to that of the VAX embodiment described above, except for a slight difference regarding the converters. The 8 bit codec performs mu-law conversion, which is non linear but provides enhanced dynamic range. A lookup table is used to transform the 8 bit mu-law output provided from the codec chip into a 13 bit linear output. Similarly, in a speech synthesis operation, the linear output of the lattice filter operation is pre-converted, using the same lookup table, to an 8-bit word which will give an appropriate analog output signal from the codec. This microcomputer embodiment also includes an internal speaker, and a microphone jack.

A further preferred realization is the use of multiple micro-computer based voice mail stations, as described above, to configure a microcomputer-based voice mail system. In such a system, microcomputers are conventionally connected in a local area network, using one of the many conventional LAN protoacalls, or are connected using PBX tilids. The only slightly

distinctive feature of this voice mail system embodiment is that the transfer mechanism used must be able to pass binary data, and not merely ASCII data. As between microcomputer stations which have the voice mail analysis/synthesis capabilities discussed above, the voice mail operation is simply a straight forward file transfer, wherein a file representing encoded speech data is generated by an analysis operation at one station, is transferrd as a file to another station, and then is converted to analog speech data by a synthesis operation at the second station.

Thus, the crucial changes taught by the present invention are changes in the analysis portion of an analysis/synthesis system, but these changes affect the system as a whole. That is, the system as a whole will achieve higher throughput of intelligible speech information per transmitted bit, better perceptual quality of synthesized sound at the synthesis section, and other system-level advantages. In particular, microcomputer network voice mail systems perform better with minimized channel loading according to the present invention.

Thus, the present invention provides the objects described above, of energy normalization and of silent suppression, as well as other objects, advantageously.

Claims

1. A voice mail system, comprising an analyzer connected to receive a digital speech signal and to generate therefrom a sequence of frames of speech parameters, said parameters at each frame including an energy parameter, excitation parameters, and linear predictive coding parameters, output means for loading said parameters for each speech frame into a data channel, input means for receiving a sequence of frames of speech parameters, means for configuring a lattice filter in accordance with said linear predictive coding parameters, means for generating an excitation signal in accordance with said excitation parameters, said excitation being provided as input to said lattice filter, and means for modulating the output of said lattice filter in accordance with said energy parameter to provide a speech signal output characterized in that:

means are provided for normalising the energy parameter of each said speech frame, wherein said energy parameter of each frame is normalised primarily with respect to an energy parameter of a subsequent frame occurring at least 0.1 seconds after said each frame.

2. The system of claim 1, wherein said energy parameter of each said speech frame is normalised with respect to a peak-tracking parameter of said subsequent frames, said peak-tracking parameter corresponding generally to an upper envelope of the sequence of said energy parameters of said frames.

3. The system of claim 1, wherein said speech parameters of each of said frame also indicate the voiced/unvoiced status of said respective frame.

4. The system of claim 3, wherein said parameters also include pitch information for each of said speech frames, and wherein said analyzer jointly determines pitch and voicing of each frame, so that the said pitch and voicing decisions vary as smoothly as possible across adjacent frames.

5. A method of encoding speech, comprising the steps of analyzing a speech signal to provide a sequence of frames as speech parameters, each said frame of said sequence of parameters including an energy parameter, and encoding said speech parameters into a data channel, characterized in that:

the energy parameters of each of said speech frames is normalised with respect to an energy parameter of a subsequent frame occurring later than each said respective frame by at least 0.1 seconds, with normalisation being done prior to encoding said speech parameters into the data channel.

6. The method of claim 5, wherein said energy value of each said speech frame is normalized with respect to a peak-tracking parameter of said subsequent frames, said peak-tracking parameter corresponding generally to an upper envelope of the sequence of said energy values of said frame.

7. A speech coding system, comprising an analyzer connected to receive speech input data and to generate therefrom a sequence of frames of speech parameters, said frames being provided at a predetermined frame rate, said frames comprising plural parameters including an energy parameter, an encoder for encoding successive speech frames as digital values, and silence suppression means connected to said encoding means, said silence suppression means preventing said encoder from encoding the ones of said sequence of frames which do not correspond to an actual speech signal, and output means for loading said encoded digital values into a data channel, characterized by: means for normalizing the energy parameter of each said speech frame with respect to the energy parameter of a subsequent frame occurring at least 0.1 seconds after said each frame, and wherein said silence suppression means identifies each said frame as silent or non-silent by comparing the energy parameter of each successive one of said frames against a function of first and second adaptively updated threshold values, said first adaptively updated threshold value corresponding to a multipole of an upper envelope of said successive energy parameters of successive ones of said frames and said second threshold value corresponding to a multiple of a lower envelope of said successive values of said frames.

8. A voice mail system, comprising an analyzer connected to receive speech input data and to generate therefrom a sequence of frames of speech parameters, said frames being provided at a predetermined frame rate, said frames

comprising plural parameters including an energy parameter, an encoder for encoding successive speech frames as digital values, and silence suppression means connected to said encoding means, said silence suppression means preventing said encoder from encoding the ones of said sequence of frames which do not correspond to an actual speech signal, output means for loading said encoded digital values into a data channel, input means for receiving a sequence of frames of speech parameters, means for configuring a lattice filter in accordance with said linear predictive coding parameters, means for generating an excitation signal in accordance with said excitation parameters, said excitation being provided as input to said lattice filter, and means for modulating the output of said lattice filter in accordance with said energy parameter to provide a speech signal output, characterized by:

means for normalizing the energy parameter of each said speech frame with respect to the energy parameter of a subsequent frame occurring at least 0.1 seconds after said each frame, and wherein said silence suppression means identifies each said frame as silent or non-silent by comparing the energy parameter of each successive one of said frames against a function of first and second adaptively updated threshold values, said first adaptively updated threshold value corresponding to a multiple of an upper envelope of said successive energy parameters of successive ones of said frames and said second threshold value corresponding to a multiple of a lower envelope of said successive values of said frames.

9. The system of claim 8, wherein said analyzer provides a voicing decision for each of said speech frames, and wherein said silence suppression means updates said first threshold only during voiced ones of said frames and updates said second threshold only during unvoiced ones of said frames.

10. The system of claim 8, wherein said silence suppression means, once a silent frame has been identified, does not identify a nonsilent frame thereafter until a voiced super-threshold-energy frame is detected, in which case said voiced super-threshold-energy frame and all preceding unvoiced super-threshold-energy speech frames which are not separated from said super-threshold-energy voiced frame by at least a predetermined number of successive frames each having an energy level below said threshold level, are identified as nonsilent.

11. The system of claim 8, wherein said silence suppression means, once a nonsilent frame has been identified, identifies a silent frame only when a continuous succession of subthreshold-energy frames has been identified over a predetermined time interval.

12. The system of either of claims 10 or 11, wherein said predetermined time interval is between 0.2 and 0.8 seconds.

13. The system of claim 8, wherein said energy value of each said speech frame is normalized with respect to said energy values primarily of those of said frames which are later than said respective frame by at least 0.1 seconds.

14. The system of any of claims 8 and 13, wherein said energy value of each said speech frame is normalized with respect to a peak-tracking parameter of said subsequent frames, said peak-tracking parameter corresponding generally to an upper envelope of the sequence of said energy values of said frames.

15. The system of claim 11, wherein said silence suppression means, once a nonsilent frame has been identified, identifies a silent frame only if said continuous succession of subthreshold energy frames over said predetermined time interval is found after a voiced subthreshold energy frame.

Patentansprüche

1. Sprachpostsystem mit einem Analysator, der so angeschlossen ist, daß er ein digitales Sprachsignal empfängt und daraus eine Folge von Rahmen aus Sprachparametern erzeugt, wobei die Parameter jedes Rahmens einen Energieparameter, Anregungsparameter und Parameter für die lineare Vorausssagecodierung enthalten, Ausgangsmitteln zum Laden der Parameter für jeden Sprachrahmen in einen Datenkanal, Eingabemitteln zum Empfangen einer Folge von Rahmen aus Sprachparametern, Mitteln zum Konfigurieren eines Gitterfilters entsprechend den Parametern für die lineare Voraussagecodierung, Mitteln zum Erzeugen eines Anregungssignals entsprechend den Anregungsparametern, wobei das Anregungssignal als Eingangssignal für das Gitterfilter vorgesehen ist, und Mitteln zum Modulieren des Ausgangssignals des Gitterfilters entsprechend dem Energieparameter zur Lieferung eines Ausgangssprachsignals, dadurch gekennzeichnet, daß Mittel zum Normieren des Energieparameters jedes Sprachrahmens vorgesehen sind, wobei der Energieparameter jedes Sprachrahmens in erster Linie bezüglich eines Energieparameters eines nachfolgenden Rahmens normiert wird, der wenigstens 0,1 Sekunden nach dem betreffenden Rahmen erscheint.

2. System nach Anspruch 1, bei welchem der Energieparameter jedes Sprachrahmens bezüglich eines Spitzenwert - Nachführungsparameters der nachfolgenden Rahmen normiert wird, wobei der Spitzenwert - Nachführungsparameter allgemein einer oberen Hüllkurve der Folge der Energieparameter der Rahmen entspricht.

3. System nach Anspruch 1, bei welchem die Sprachparameter jedes Rahmens auch den Stimmhaft/Stimmlos-Zustand des jeweiligen Rahmens angeben.

4. System nach Anspruch 3, bei welchem die Parameter auch eine Tohöheninformation für jeden der Sprachrahmen enthalten und der Analysator gemeinsam die Tonhöhe und den Stimmtyp jedes Rahmens bestimmt, so daß die

Tonhöhe und die Stimmtypentscheidungen sich so glatt wie möglich über benachbarte Rahmen ändern.

5. Verfahren zum Codieren von Sprache, enthaltend die Schritte des Analysierens eines Sprachsignals zur Erzeugung einer Folge von Rahmen aus Sprachparametern, wobie jeder Rahmen der Folge von Sprachparametern einen Energieparameter enthält, sowie des Codierens der Sprachparameter in einen Datenkanal, dadurch gekennzeichnet, daß die Energieparameter jedes der Sprachrahmen bezüglich eines Energieparameters eines nachfolgenden Rahmens normiert werden, der um wenigstens als 0,1 Sekunden später als der betroffene Rahmen auftritt, wobei die Normierung vor der Codierung der Sprachparameter in den Datenkanal durchgeführt wird.

6. Verfahren nach Anspruch 5, bei welchem der Energiewert jedes Sprachrahmens bezüglich eines Spitzen - Nachführungsparameters der nachfolgenden Rahmen normiert wird, wobei der Spitzenwert - Nachführungsparameter allgemein einer oberen Hüllkurve der Folge der Energiewerte des Rahmens entspricht.

7. Sprachcodierungssystem mit einem Analysator, der so angeschlossen ist, daß er Spracheingangsdaten empfängt und daraus eine Folge von Rahmen aus Sprachparametern erzeugt, wobei die Rahmen mit einer vorbestimmten Rahmenfolgefrequenz geliefert werden und mehrere Parameter einschließlich einem Energieparameter enthalten, einem Codierer zum Codieren aufeinanderfolgender Sprachrahmen als digitale Werte und Stummunterdrückungsmitteln, die an den Codierer angeschlossen sind, wobei die Stummunterdrückungsmittel den Codierer daran hindern, die 1-Werte der Folge von Rahmen zu codieren, die nicht einem tatsächlichen Sprachsignal entsprechen, und Ausgabemitteln zum Laden der codierten digitalen Werte in einen Datenkanal, gekennzeichnet durch Mittel zum Normierendes Energieparameters bezüglich des Energieparameters eines nachfolgenden Rahmens, der wenigstens 0,1 Sekunden nach jedem Rahmen auftritt, wobei die Stummunterdrückungsmittel jeden der Rahmen als stumm oder nicht stumm identifizieren, indem der Energieparameter jedes nachfolgenden Rahmens mit einer Funktion aus ersten und zweiten adaptiv aktualisierten Schwellenwerten verglichen werden, wobei der erste adaptiv aktualisierte Schwellenwert einem Vielfachen einer oberen Hüllkurve der aufeinanderfolgenden Energieparameter aufeinanderfolgender Rahmen entspricht, während der zweite Schwellenwert einer Vielfachen einer unteren Hüllkurve der aufeinanderfolgenden Werte der Rahmen entspricht.

8. Sprachpostsystem mit einem Analysator, der so angeschlossen ist, daß er Spracheingangsdaten empfängt und aus diesen eine Folge von Rahmen aus Sprachparametern erzeugt, wobei die Rahmen mit einer vorbestimmten Rahmenfolgefrequenz erzeugt werden und mehrere Parameter einschließlich eines Energieparameters enthalten, einem Codierer zum Codieren aufeinanderfolgender Sprachrahmen als digitale Werte und Stummunterdrückungsmitteln, die an den Codierer angeschlossen sind, wobei die Stummunterdrückungsmittel den Codierer daran hindern, die 1-Werte der Folge von Rahmen zu codieren, die nicht einem tatsächlichen Sprachsignal entsprechen, Ausgabemitteln zum Laden der codierten digitalen Werte in einen Datenkanal, Eingangsmitteln zum Empfangen einer Folge von Rahmen aus Sprachparametern, Mitteln zum Konfigurieren eines Gitterfilters entsprechend den Parametern für die lineare Voraussagecodierung, Mitteln zum Erzeugen eines Anregungssignals gemäß den Anregungsparametern, wobei die Anregung als Eingangssignal für das Gitterfilter geliefert wird, und Mitteln zum Modulieren des Ausgangssignals des Gitterfilters entsprechend dem Energieparameter zur Abgabe eines Sprachausgangssignals, gekennzeichnet durch Mittel zum Normieren des Energieparameters jedes Sprachrahmens bezüglich des Energieparameters eines nachfolgenden Rahmens, der wenigstens 0,1 Sekunden nach dem Rahmen auftritt, wobei die Stummunterdrückungsmittel jeden Rahmen als stumm oder nicht stumm identifizieren, indem der Energieparameter jedes nachfolgenden Rahmens mit einer Funktion aus ersten und zweiten adaptiv aktualisierten Schwellenwerten verglichen wird, wobei der erste adaptive aktualisierte Schwellenwert einer Vielfachen einer oberen Hüllkurve der aufeinanderfolgenden Energieparameter nachfolgender Rahmen entspricht, während der zweite Schwellenwert einer Vielfachen einer unteren Hüllkurve der aufeinanderfolgenden Werte der Rahmen entspricht.

9. System nach Anspruch 8 bei welchem der Analysator eine Stimmtypscheidung für jeden der Sprachrahmen trifft und die Stummunterdrückungsmittel den ersten Schwellenwert nur während stimmhafter Rahmen und den zweiten Schwellenwert nur während stimmloser Rahmen aktualisieren.

10. System nach Anspruch 8, bei welchem die Stummunterdrückungsmittel nach der Identifizierung eines stummen Rahmens im Anschluß daran keinen nicht stummen Rahmen identifizieren, bis ein stimmhafter Rahmen mit einer über dem Schwellenwert liegenden Energie festgestellt wird, wobei in diesem Fall der stimmhafte Rahmen mit einer über dem Schwellenwert liegenden Energie und alle vorangehenden stimmlosen Sprachrahmen mit einer über dem Schwellenwert liegenden Energie, die nicht um wenigstens eine vorbestimmte Anzahl aufeinanderfolgender Rahmen mit einem jeweils unter dem Schwellenwert liegenden Energiepegel getrennt sind, als nicht stumm identifiziert werden.

11. System nach Anspruch 8, bei welchem die Stummunterdrückungsmittel nach Identifizierung eines nicht stummen Rahmens einen stummen Rahmen nur dann identifizieren, wenn eine kontinuierliche Folge von Rahmen mit einem

unterhalb des Schwellenwerts liegenden Energiepegel über ein vorbestimmtes Zeitintervall identifiziert worden sind.

12. System nach Anspruch 10 oder 11, bei welchem das vorbestimmte Zeitintervall zwischen 0,2 und 0,8 Sekunden liegt.

13. System nach Anspruch 8, bei welchem der Energiewert jedes Sprachrahmens hauptsächlich in bezug auf die Energiewerte von solchen Rahmen normiert wird, die um wenigstens 0,1 Sekunden später als der betreffende Rahmen liegen.

14. System nach einem der Ansprüche 8 und 13, bei welchem der Energiewert jedes Sprachrahmens bezüglich auf einen Spitzenwert - Nachführungsparameter der nachfolgenden Rahmen normiert wird, wobei der Spitzenwert - Nachführungsparameter allgemein einer oberen Hüllkurve der Folge der Energiewerte der Rahmen entspricht.

15. System nach Anspruch 11, bei welchem die Stummunterdrückungsmittel nach der Identifizierung eines nicht stummen Rahmens einen stummen Rahmen nur dann identifizieren, wenn die kontinuierliche Folge der Rahmen mit einem unterhalb des Schwellenwertes liegenden Energiepegel für die Dauer des vorbestimmte Zeitintervalls nach einem stimmhaften Rahmen mit einem unterhalb des Schwellenwertes liegenden Energiepegel gefunden worden ist.

Revendications

1. Système de communication vocale, comportant un analyseur connecté pour recevoir un signal de parole numérique et pour produire à partir de ce dernier une séquence de trames de paramètres de parole, lesdits paramètres de chaque trame contenant un paramètre d'énergie, des paramètres d'excitation et des paramètres de codage linéaire par prévision, un dispositif de sortie pour charge desdits paramètres de chaque trame de parole dans un canal de données, un dispositif d'entrée pour recevoir une séquence de trames de paramètres de parole, un dispositif de configuration d'un filtre en réseau en fonction desdits paramètres de codage linéaire par prévision, un dispositif générateur d'un signal d'excitation en fonction desdits paramètres d'excitation, ladite excitation étant produite comme entrée audit filtre en réseau, et un dispositif de modulation de la sortie dudit filtre en réseau en fonction dudit paramétre d'énergie pour produire une sortie de signal de parole, caractérisé en ce que:

un dispositif est prévu pour normalizer le paramètre d'énergie de chacune desdites trames de parole, ledit paramètre d'énergie de chaque trame étant normalisé principalement par rapport à un paramètre d'énergie d'une trame ultérieure apparaissant au moins 0,1 seconde après chacune desdites trames.

2. Système selon la revendication 1, dans lequel ledit paramètre d'énergie de chacune desdites trames de parole est normalisé par rapport à un paramètre de poursuite de crête desdites trames suivantes, ledit paramètre de poursuite de crête correspondant généralement à une enveloppe supérieure de la séquence desdits paramètres d'énergie desdites trames.

3. Système selon la revendication 1, dans lequel lesdits paramètres de parole de chacune desdites trames indiquent également l'état sonore/sourd de chacune desdites trames respectives.

4. Système selon la revendication 3, dans lequel lesdits paramètres comprennent également des informations de hauteur pour chacune desdites trames de parole, et dans lequel ledit analyseur détermine conjointement la hauteur et la sonorité de chaque trame de manière que lesdites décisions de hauteur de de sonorité varient aussi régulièrement que possible entre des trames voisines.

5. Procédé de codage de parole, consistant à analyser un signal de parole pour produire une séquence de trames comme des paramètres de parole, chacune desdites trames de ladite séquence de paramètres contenant un paramètre d'énergie, et à coder lesdits paramètres de parole dans un canal de données, caractérisé en ce que les paramètres d'énergie de chacune desdites trames de parole sont normalisés par rapport à un paramètre d'énergie d'une trame suivante apparaissant plus tard que chacune desdites trames respectives, d'au moins 0,1 seconde, la normalisation étant faite avant le codage desdits paramètres de parole dans le canal de données.

6. Procédé selon la revendication 5, dans lequel ladite valeur d'énergie de chacune desdites trames de parole est normalisée par rapport à un paramètre de poursuite de crête desdites trames suivantes, ledit paramètre de poursuite de crête correspondant généralement à une enveloppe supérieure de la séquence desdites valeurs d'énergie de ladite trame.

7. Système de codage de la parole, comportant un analyseur connecté pour recevoir des données d'entrée de parole et pour produite à partir de ces données une séquence de trames de paramètres de parole, lesdites trames étant produites à une fréquence de trames prédéterminée, lesdites trames contenant plusieurs paramètres comprenant un paramètre d'énergie, un codeur pour coder des trames de parole successives comme des valeurs numériques et un dispositif de suppression de silence connecté audit dispositif de codage, ledit dispositif de suppression de silence évitant que ledit codeur code celles desdites séquences de trames qui ne correspondnt pas à un signal de parole réelle, et un dispositif de sortie pour charger lesdites valeurs numériques codées dans un canal de données, caractérisé par: un dispositif de normalisation du paramètre d'énergie de chacune desdites trames de parole par rapport au paramètre d'énergie d'une trame ultérieure apparaissant au moins 0,1 seconde après chacune desdites trames, et dans lequel ledit dispositif de suppression de silence identifie chacune desdites trames comme silencieuses ou

non silencieuses en comparant le paramètre d'énergie de chacune successive desdites trames à une fonction d'une première et d'une seconde valeurs seuil corrigées de façon adaptative, ladite première valeur seuil corrigée de façon adaptative correspondant à un multiple d'une enveloppe supérieure desdits paramères d'énergie successifs de certaines successives desdites trames et ladite seconde valeur seuil correspondant à un multiple d'une enveloppe inférieure desdites valeurs successives desdites trames.

8. Système de communication vocale, comportant un analyseur connecté pour recevoir des données d'entrée de parole et pour produire à partir de ces données une séquence de trames de paramètres de parole, lesdites trames étant produites avec un débit de trames prédéterminé, lesdites trames contenant plusieurs paramètres comprenant un paramètre d'énergie, un codeur pour coder des trames de parole successives comme des valeurs numériques et un dispositf de suppression de silence connecté audit dispositif de codage, ledit dispositif de suppression de silence évitant que ledit codeur ne code celles de ladite séquence de trames ne correspondant pas à un signal de parole réelle, un dispositif de sortir pour charger lesdites valeurs numériques codées dans un canal de données, un dispositif d'entrée pour recevoir une séquence de trames de paramètres de parole, un dispositif de configuration d'un filtre en réseau en fonction desdits paramètres de codage linéaires à prévision, un dispositif générateur d'un signal d'excitation en fonction desdits paramètres d'excitation, ladite excitation étant produite comme une entrée dudit filtre en réseau et un dispositif de modulation de la sortie dudit filtre en réseau en fonction dudit paramètre d'énergie pour produire une sortie de signal de parole, caractérisé par:

un dispositif de normalisation du paramètre d'énergie de chacune desdites trames de parole par rapport au paramètre d'énergie d'une trame ultérieure apparaissant au moins 0,1 seconde après chacune desdits trames et dans lequel dedit dispositif de suppression de silence identifie chacune desdites trames comme silencieuse ou non silencieuse en comparant le paramètre d'énergie de chacune successive desdites trames à une fonction d'une première et une seconde valeurs seuil corrigées de façon adaptative, ladite première valeur seuil corrigée de façon adaptative correspondant à un multiple d'une enveloppe supérieure desdits paramètres successifs d'énergie de certaines successives desdites trames et la seconde valeur seuil correspondant à un multiple d'une enveloppe inférieure desdites valeurs successives desdites trames.

9. Système selon la revendication 8, dans lequel l'analyseur produite une décision de vocalisation pour chacune desdites trames de parole, et dans lequel ledit dispositif de suppression de silence corrige ledit premier seuil seulement pendant celles sonores desdites trames et corrige seulement ledit second seuil pendant celles sourdes desdites trames.

10. Système selon la revendication 8, dans lequel ledit dispositif de suppression de silence, une fois qu'une trame silencieuse a été identifiée, n'identifie pas une trame non silencieuse ensuite jusqu'à ce qu'une trame sonore d'énergie supérieure au seuil soit détectée, auquel cas ladite trame sonore d'énergie supérieure au seuil et toutes les trames de parole sourdes d'énergie supérieure au seuil qui ne sont pas séparées de ladite trame sonore d'énergie supérieure au seuil par au moins un nombre prédéterminé des trames successives ayant chacune un niveau d'énergie au-dessous dudit niveau seuil, sont identifiées comme non silencieuses.

11. Systèm selon la revendication 8, dans lequel le dispositif de suppression de silence, une fois qu'une trame non silencieuse a été identifiée, identifie une trame silencieuse seulement lorsqu'une succession continue de trames d'énergie au-dessous du seuil a été identifiée pendant un intervalle de temps prédéterminé.

12. Système selon la revendication 10 ou 11, dans lequel ledit intervalle de temps prédéterminé est compris entre 0,2 et 0,8 secondes.

13. Système selon la revendication 8, dans lequel ladite valeur d'énergie de chacune desdites trames de parole est normalisée par rapport auxdites valeurs d'énergie, principalement celles desdites trames qui sont ultérieures à ladite trame respective d'au moins 0,1 seconde.

14. Système selon l'une quelconque des revendications 8 et 13, dans lequel ladite valeur d'énergie de chacune desdites trames de parole est normalisée par rapport à un paramètre de poursuite de crête desdites trames suivantes, ledit paramètre de poursuite de crête correspondant généralement à une enveloppe supérieure de la séquence desdites valeurs d'énergie desdites trames.

15. Système selon la revendication 11, dans lequel ledit dispositif de suppression de silence, une fois qu'une trame non silencieuse a été identifiée, identifie une trame silencieuse seulement si ladite succession continue de trames d'énergie au-dessous du seuil pendant un intervalle de temps prédéterminée est trouvée après une trame sonore d'énergie au-dessous du seuil.

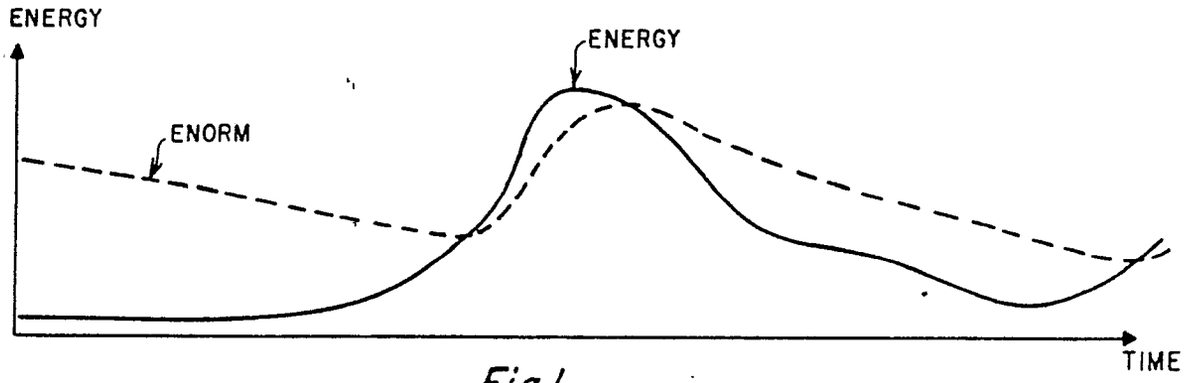

Fig.1

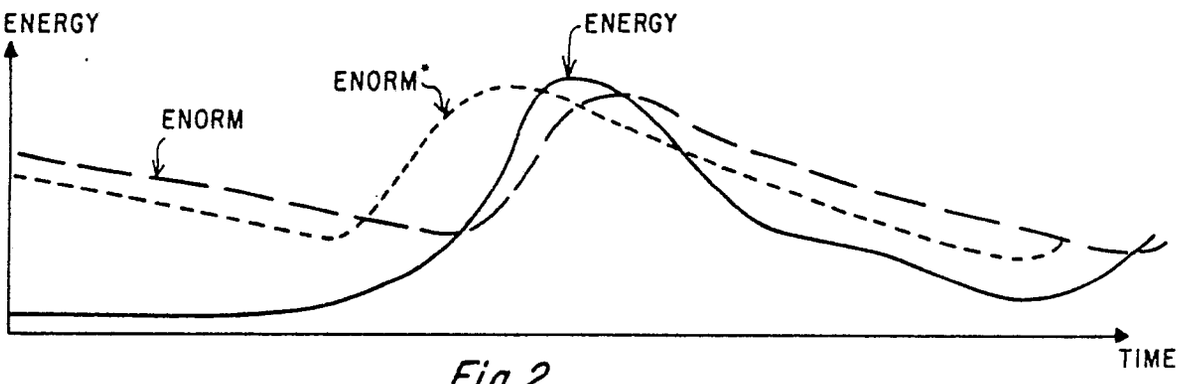

Fig.2

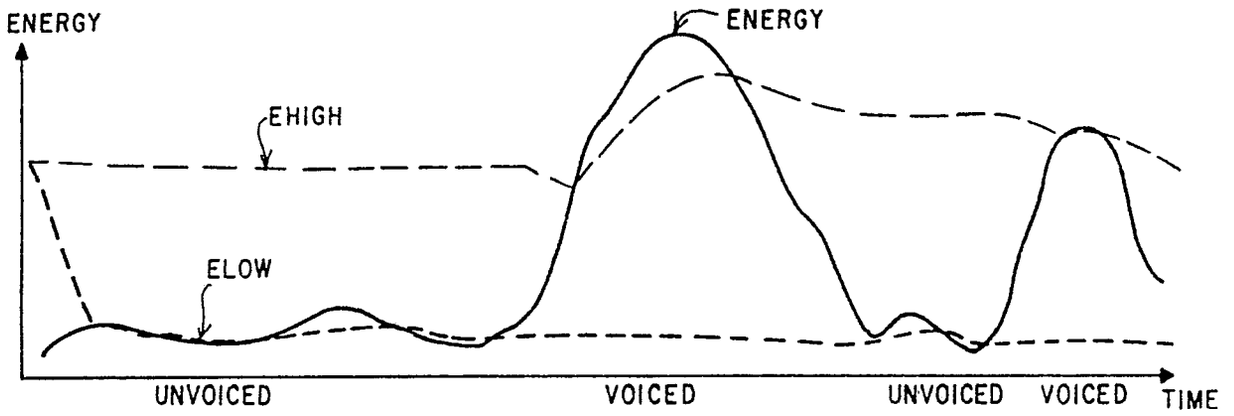
ENERGY
ENERGY
EHIGH
ELOW
UNVOICED
VOICED
UNVOICED
VOICED
TIME

Fig. 3

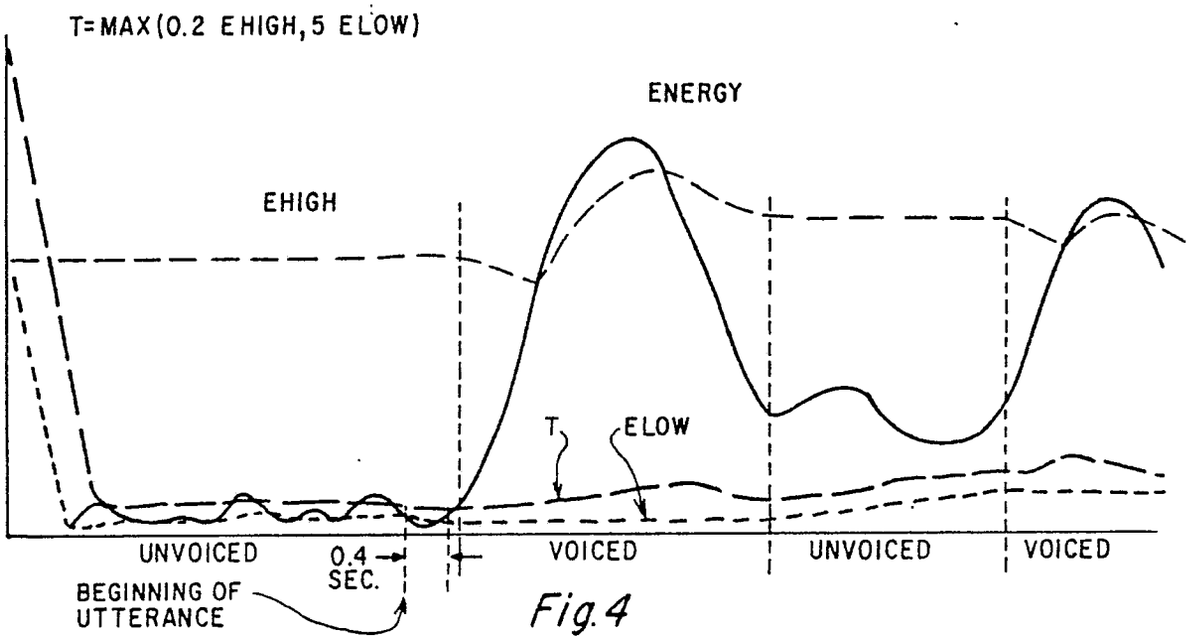
T=MAX(0.2 EHIGH,5 ELOW)
ENERGY
EHIGH
T
ELOW
UNVOICED
0.4 SEC.
VOICED
UNVOICED
VOICED
BEGINNING OF UTTERANCE

Fig. 4

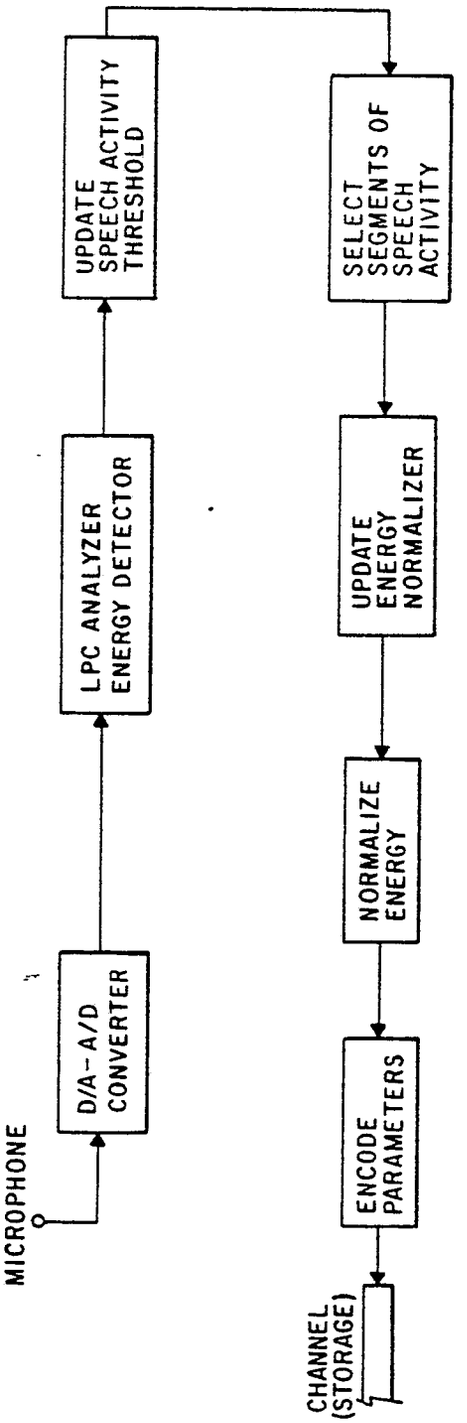

Fig.5 BLOCK DIAGRAM FOR SILENCE SUPPRESSION AND ENERGY NORMALIZATION