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

(11) Publication number: **0 689 189 A1**

(12) EUROPEAN PATENT APPLICATION

(21) Application number: **95108870.7**

(22) Date of filing: **08.06.95**

(51) Int. Cl.[6]: **G10L 3/00**, G10L 9/14, G10L 3/02, G10L 9/18

(30) Priority: **20.06.94 IT MI941283**

(43) Date of publication of application:
27.12.95 Bulletin 95/52

(84) Designated Contracting States:
DE FR GB IT

(71) Applicant: **ALCATEL ITALIA S.p.A.**
Via L. Bodio, 33/39
I-20158 Milano (IT)

(72) Inventor: **Cucchi, Silvio**
Via S. Ibenzio 9
I-20090 Gaggiano (MI) (IT)
Inventor: **Fratti, Marco**
Via della Birona 9
I-20052 Monza (MI) (IT)

(74) Representative: **Pohl, Herbert, Dipl.-Ing. et al**
c/o Alcatel SEL AG,
Zentralbereich Patente und Lizenzen,
Postfach 30 09 29
D-70449 Stuttgart (DE)

(54) Voice coders

(57) The invention relates to a method of improving the features of voice encoders based on linear prediction and analysis-by-synthesis techniques making use of an objective function to minimize.

This objective function comprises jointly or alternately the free evolution of the objective signal and of the synthetic signal and a weighing with respect to the error between the prediction residue and the synthetic excitation.

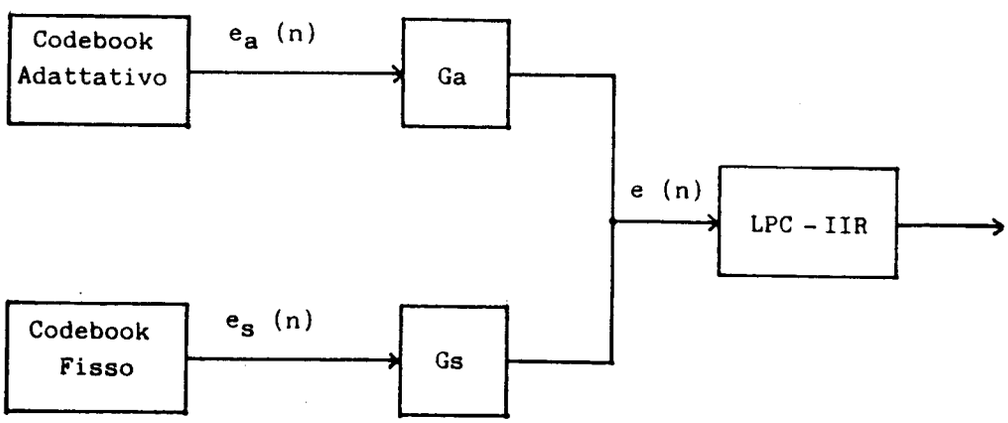

FIG. 1

EP 0 689 189 A1

1. DESCRIPTION OF THE PRIOR ART

Speech coding is of application in several communication fields: from transmission via satellite to radiomobile, store-forward systems, automatic responders, etc.

In particular there is a strong need of techniques effective for voice signal coding where there are remarkable band limitations (consider the "limited" availability of band in the ether); therefore, it is important to be able to reduce drastically the bit-rate to be transmitted, still maintaining a high quality of the received signal.

Various voice signal coding techniques are used for this purpose; the most usual (assuring a high quality of the received signal at various bit-rates) are based upon the LP (Linear Prediction) and A-b-S (Analysis-by-Synthesis) principles (P. Kroon, E.F. Deprettere "A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbits/s", IEEE Journal on Selected Areas in Communications, Vol. 6, No. 2, pages 353-363, Feb. 1988).

The present specification discloses some techniques for improving the features of speech coders based on the above-mentioned techniques.

The voice coders based on the Linear Prediction (LP) are parametric coders; typically Analysis-by-Synthesis (A-b-S) techniques are used for the correct determination of the parameters of the system. Such coders synthesize the voice through the use of a suitable input excitation to a synthesis LP filter.

In particular, the excitation should have the characteristics of the "physical" excitation wave which, coming from the glottis, is then spectrally modified in function of the characteristics of the system that simulates the voice segment (LP filter).

The most recent A-b-S coders make use of an excitation structure which is composed of an Adaptive Codebook and of a Fixed Codebook (eventually structured). Without prejudicing the generality, it can be assumed that the Fixed Codebook is composed of independent vectors of random numbers, as in the case of CELP coders (M.R. Schroeder, B.S. Atal, "Code Excited Linear Prediction (CELP): high-quality speech at very low bit rates", Proc. ICASSP '85, pages 26-29).

In Fig. 1 there is represented a block diagram of a typical CELP voice synthesizer; block LPC-IIR denotes the synthesis filter for reconstructing the voice waveform; $e_a(n)$ is the adaptive codebook vector (and Ga is the corresponding scaling factor) and $e_s(n)$ is the fixed codebook vector (and Gs is the corresponding scaling factor); e(n) is the composite excitation vector. For a detailed description of the synthesizer, reference can be made to W.B. Kleijn, D.J. Krasinski, R.H. Ketchum "Improved Speech Quality and Efficient Vector Quantization in SELP", Proc. ICASSP '88, pages 155-158.

In general, $e_a(n)$ and $e_s(n)$ are selected from a suitable set of vectors and are determined simultaneously with respective Ga and Gs. The determination occurs in a time interval of about 5 to 10 ms (analysis frame) and is based on the minimization of the objective function according to the well-known criterion of the perceptively weighted minimum-squared error (see M.R. Schroeder, B.S. Atal, "Code Excited Linear Prediction (CELP): high-quality speech at very low bit-rates", Proc. ICASSP '85, pages 26 to 29), according to the following expression :

$$E = \sum_{n=0}^{N-1} [r_s(n) - Gu_i(n)]^2 \qquad (1)$$

where N is the length of the time interval for minimization; $u_i(n)$ is the zero-state synthesis filter response at the the i-th input of the Codebook (either adaptive or fixed) and G is the corresponding gain; lastly, $r_s(n)$ is the reference signal or "objective" signal (i.e. the original voice segment from which the contribution of the reconstruction filter memory deriving from previous synthesis has been subtracted).

The objective function described at (1), even if usually used, cannot be optimal for the choice of the parameters. In particular, it must be kept in mind that the system is random: this entails that the contribution to the synthetic signal made by the excitation samples in the vicinity of n = 0, in general is greater than the contribution made by the excitation samples in the vicinity of n = N - 1. This fact may cause a poor approximation of the ideal excitation during segments of voiced signal. In this circumstance, the ideal excitation exhibits the characteristic quasi-periodic "pitch pulses". The synthetic excitation, in this case, shall contain the pitch pulses with the correct time alignment and the correct amplitude. In the case in which the impulses of ideal excitation (commonly called "prediction residue") are located at the end of the minimization interval (i.e. for n comprised in the vicinity of N - 1), its reconstruction becomes more problematic, since their contribution "weighs" less within the minimization interval.

This phenomenon becomes more apparent during signal transients, i.e. in the passages from unvoiced segments to voiced segments and within the voice portions in the segments in which the ideal excitation changes its shape (still maintaining the "quasi-period ic" characteristic) because of prediction filter variations.

In the following, two possible approaches are described for overcoming the problems described above; these approaches can be used both separately and jointly and allow the characteristics of the A-b-S coders operating at various bit-rates to be improved.

2. FREE-EVOLUTION BASED APPROACH

A first approach consists in using a signal $r_s^{el}(n)$ longer than N samples as a reference signal of the objective function (i.e. signal $r_s(n)$ of eq. (1)). Such a signal is obtained from the time linkage of the signal $r_s(n)$ (for n = o...N - 1) and from the free evolution of such a signal, said free evolution el(n) being obtained by charging the last p samples of $r_s(n)$ in the synthesis filter memory LPC-IIR (p being the order of the filter) and letting the filter "discharge" (i.e. calculating its output corresponding to a null input).

Therefore, it is obtained that:

$$r_s^{el}(n) = r_s(n), n = 0..N-1 \tag{2}$$

$$r_s^{el}(n) = el(n), n = N..N-1+M \tag{3}$$

M being the free-evolution length.

Such approach can be justified in the following manner: the voice can always be considered as obtained from an ideal excitation that constitutes the input of an all-pole synthesis filter (the filter denoted by LPC-IIR in Fig. 1). Such ideal excitation is nothing else than the prediction residue, obtained by filtering the voice through the "inverse filter", i.e. the all-zero filter derived from LPC-IIR.

Assume to carry out a dashed stationary analysis of the voice signal: then, within the analysis interval, the ideal excitation constitutes the forcing term of the synthesis filter. But, if at the end of the analysis interval, the input of the filter is "turned off" (i.e. the ideal excitation is set to zero), the synthesis filter is discharged according to a waveform depending on its poles and on the samples of the ideal excitation (especially those right preceding the time instant n = N - 1).

Therefore, it is evident that in the case in which the last samples of the ideal excitation are significant (e.g. a pitch pulse is present) and the filter is near to instability (e.g. during segments of voiced signal), the free evolution of the filter due to the ideal excitation, typically will exhibit sinusoidal oscillations which will damp rather slowly and therefore the term el(n) of equation (3) will contribute significantly.

For a high-quality of the reconstructed signal it is very important that the synthetic excitation has spectral and time location (e.g. the pitch pulse) characteristics similar to those of the ideal excitation. Therefore, it is evident that by including, in the objective function, the contributions of the free evolutions due to both the ideal excitation and to the synthetic excitation, it is possible to carry out a more correct choice of the latter. In fact, depending on the spectral/time characteristics of the signal, the difference between the ideal free evolution and the synthetic one may have a preponderant weight in the modified objective function.

In formulas, the above-mentioned concepts may be expressed according to the revised objective function:

$$E1 = \sum_{n=0}^{N-1+M} \left[r_s^{el}(n) - G u_i^{el}(n) \right]^2 \tag{4}$$

in which

$$u_i^{el}(n) = u_i(n), n = 0..N-1 \quad (5)$$

$$u_i^{el}(n) = el_i(n), n = N..N-1+M \quad (6)$$

where $u_i(n)$ is the (zero state) synthesis filter response at the i-th input and $el_i(n)$ is the corresponding "synthetic" free evolution.

The excitation parameters (i.e. the i-th index and the corresponding gain G) are then choosen in such a way as to minimize the modified objective function (4).

For instance, to obtain the "original" free evolution e1(n) one could proceed in the following way:

- Inverse filtering (through an all-zero filter) of the voice signal along the interval 0...N - 1, thus obtaining the ideal excitation (prediction residue), limited to the time interval 0...N - 1.
- Providing at the input of the synthesis filter LPC-IIR the ideal excitation thus attained, obtaining again at the output the original voice signal within the time interval 0... N - 1.
- Starting from the final status of the synthesis filter thus attained, provide at the input of the synthesis filter a null input and let the filter "discharge" for a number M of samples equal to the length of the free evolution to be obtained.

From the procedure described above it can be noted at once that there is no need of computing the prediction residue. In order to obtain the desired free evolution it is sufficient to force into the state of the synthesis filter the last p samples (p being the order of the filter) of the original voice signal (i.e. the samples N - 1, N - 2, ..., N - p) and letting the null-input filter discharge. Evidently one can proceed in a similar fashion for computing the synthetic free evolution.

To be noted, lastly, that this approach does not entail an increase in the coding delay since, in the objective function, the voice samples beyond the time interval 0... N - 1 are not used.

3. THE WEIGHT-BASED APPROACH

In the previous paragraph it has been pointed out that to obtain a high-quality of the reconstructed signal it is very important that the synthetic excitation has spectral and time location (e.g. pitch pulse) characteristics, similar to the ones present in the ideal excitation. From this it derives that it may be important to obtain not only a good similarity between the original voice and the synthetic voice, but also a good similarity between ideal excitation and synthetic excitation.

In fact, by using an approach to the minimum squares in the classical objective function, the parameters of the reconstructed excitation allow the achievement of a synthetic voice which "averagely" is similar to the original voice.

Actually, from the perceptive point of view it is often more important that the synthetic voice is similar to the original voice only locally (for instance it is very important to reconstruct the connection from an unvoiced segment to a voiced segment with the correct time alignment and with the correct dynamics. It is not rare to find connection transients whose time duration is much shorter than the duration of the synthesis frame).

For a fair local reconstruction it is then important to maintain a certain similarity degree, also with the ideal excitation.

The objective function may than be composed of two contributes, in function of the original voice and of the ideal excitation, respectively, and it assumes the following expression:

$$E2 = \alpha E + (1 - \alpha)E3 \quad (7)$$

where:

$$E = \sum_{n=0}^{N-1} [r_s(n) - Gu_i(n)]^2 \quad (8)$$

$$E3 = \sum_{n=0}^{N-1} [e_s(n) - Ge_i(n)]^2 \qquad (9)$$

In equation (9) $e_s(n)$ is the prediction residue obtained from the reference signal $r_s(n)$ and $e_i(n)$ is the codebook excitation generating the synthetic signal $u_i(n)$. To be noted that the prediction residue $e_s(n)$ must be calculated starting from $r_s(n)$ through inverse filtering (with all-zero filter) with null initial state. In fact, as it is known, the reference has been obtained from the voice signal by subtracting its reconstruction filter memory deriving from the previous synthesis. The reference signal is then "free" from every contribution due to the filter memory and can be considered as obtained from a suitable ideal excitation $e_s(n)$ coming into the synthesis filter with null initial state.

In equation (7), α is a parameter whose value is comprised between 0 and 1 and controls the importance to be attached to the minimization with respect to the reference signal. Letting $\alpha = 1$ the original objective function is found again.

The excitation parameters (i.e. the i-th index and the corresponding gain G) are then chosen in such a way as to minimize the objective function described by equations (7), (8), (9). Parameter α can be either fixed or even made adaptive (i.e. varying with time), for instance in function of certain characteristics of the signal that can be estimated a priori (e.g.: estimate of voiced/unvoiced, estimate of transients, estimate of the pitch period or of the synthesis filter, etc.).

Finally, notice that the contribution due to the free evolution described in the preceding paragraph can be included in the objective function described by equations (7), (8), (9). In this case, term (8) of the objective function is modified as described in the preceding paragraph.

Claims

1. Method of computing the excitation parameters in voice coders based on linear prediction and analysis-by-synthesis techniques using an objective function to be minimized, characterized in that said objective function comprises, jointly or alternately the free evolution of the objective signal and of the synthetic signal and a weighing with respect to the error between prediction residue and synthetic excitation.

2. Method according to claim 1, characterized by using the objective function:

$$Ex = \alpha E1 + (1 - \alpha)E3 \qquad (10)$$

where: function E1 takes into account, besides the error between objective and synthetic signals also the error between the relative free evolutions, function E3 takes into account the error between ideal excitation on and synthetic excitation, and α is a weight factor comprised between 0 and 1.

3. Method according to claim 2, characterized in that said function E1 is given by:

$$E1 = \sum_{n=0}^{N-1+M} \left[r_s^{el}(n) - Gu_i^{el}(n)\right]^2 \qquad (11)$$

4. Method according to claim 2, characterized in that said function E3 is given by:

$$E3 = \sum_{n=0}^{N-1} [e_s(n) - Ge_i(n)]^2 \qquad (12)$$

5. Method according to claim 2, characterized in that said weight factor is allowed to vary with time.

6. Audio coder characterized by using the method according to one of the preceding claims.

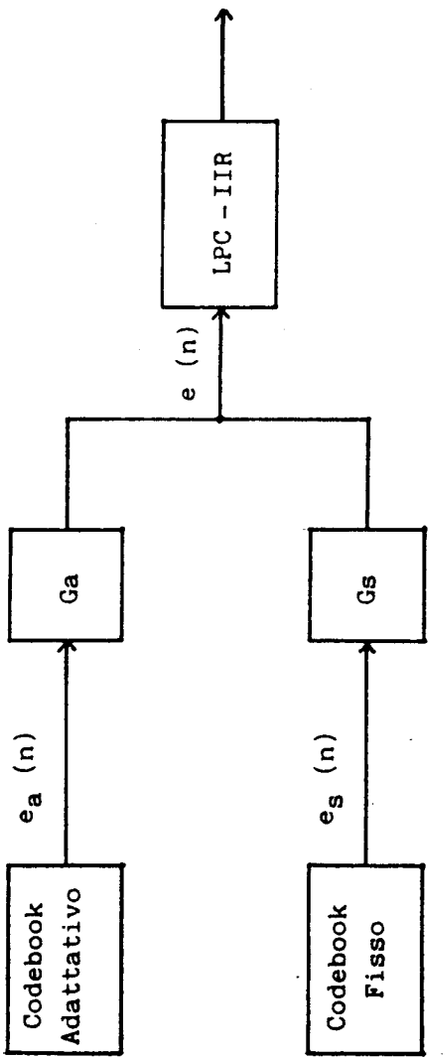

FIG. 1

European Patent Office

EUROPEAN SEARCH REPORT

Application Number EP 95108870.7

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl. 6)
A	EP - A - 0 516 439 (MOTOROLA INC.) * Abstract; Fig. 2A; claims 1,2 * --	1	G 10 L 3/00 G 10 L 9/14 G 10 L 3/02 G 10 L 9/18
A	EP - A - 0 515 138 (NOKIA MOBILE PHONES LTD.) * Fig. 2; abstract; claim 1 * --	1	
A	EP - A - 0 465 057 (AMERICAN TELEPHONE AND TELEGRAPH COMP.) * Fig. 1; abstract; claim 1 * ----	1	

TECHNICAL FIELDS SEARCHED (Int. Cl. 6)

G 10 L 3/00
G 10 L 5/00
G 10 L 7/00
G 10 L 9/00

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
VIENNA	29-08-1995	BERGER

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P0401)