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

(11) Publication number: **0 119 033**
A1

(12) EUROPEAN PATENT APPLICATION

(21) Application number: 84301302.0

(22) Date of filing: 28.02.84

(51) Int. Cl.³: G 10 L 1/08

(30) Priority: 11.03.83 GB 8306685
10.12.83 GB 8333037

(43) Date of publication of application:
19.09.84 Bulletin 84/38

(84) Designated Contracting States:
BE CH DE FR GB LI NL

(71) Applicant: PRUTEC LIMITED
142 Holborn Bars
London EC1N 2NH(GB)

(72) Inventor: Senensieb, Gideon Abraham
11, Lawrance Lea
Harston Cambridge CB2 5QR(GB)

(72) Inventor: Milbourn, Anthony John
65, High Street Hail Weston
St. Neots Cambridge(GB)

(74) Representative: Messulam, Alec Moses et al,
A. Messulam & Co. 24 Broadway
Leigh on Sea Essex SS9 1BN(GB)

(54) Speech encoder.

(57) The invention relates to a speech encoder using linear predictive coding and proposes a code comprising the parameters of a linear predictor and an excitation signal consisting of a plurality of pulses of which the timing and the amplitude is selected for each frame of speech.

To enable the excitation signal pulses for the recursive filter to be evaluated in real time, the speech signal is passed through a pole-zero filter 38 to suppress the effects of reverberations and the output of the filter 38 is correlated with the time weighted impulse response of the recursive filter with the encoded parameters.

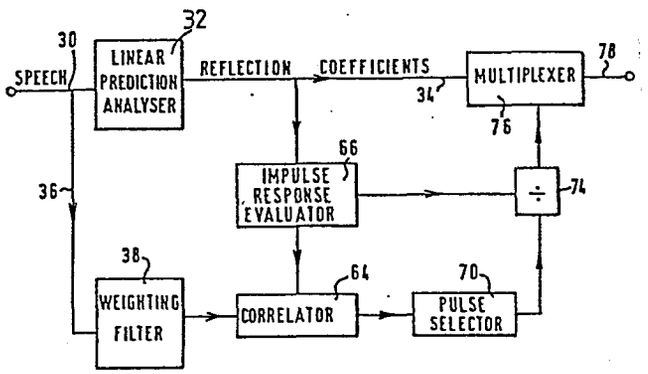

FIG. 2.

EP 0 119 033 A1

SPEECH ENCODER

This invention relates to a speech encoder, this being a circuit for converting a speech signal into a pulse train. The pulse train may either be transmitted, encrypted, or stored and from it the original speech can be reproduced.

The design of a vocoder for commercial application requires a careful compromise between three main parameters: perceived voice quality, data rate, and complexity (roughly equivalent to cost) of the hardware implementation. Other important performance parameters that must be considered are voice quality in the presence of acoustic noise at the input, robustness against errors in the low bit-rate digital stream and performance in tandem with other voice coding equipment.

There is already known in the art a system of speech encoding which makes use of the technique of linear predictive coding (LPC). In order to explain the principles employed in this method of encoding, reference will first be made to Figure 1 which shows a linear predictor.

The linear predictor in Figure 1 is a recursive digital filter comprising a summation circuit 10 which has an input line 12 and an output line 14. The output line 14 is connected to a shift register or to a tapped delay line 16 each tapping of which is fed back to the summation circuit by way of a respective multiplication circuit 18_1 to 18_n.

Assume that it is desired to produce a particular sequence of output signals corresponding to a sampled speech signal. At any given instant, the output signal has a first component determined by the weighted summed outputs from the tappings of the delay line and a second

component determined by the value of the input signal at that instant. The first of these two components may be regarded as the predicted value based on previous values of the output signal and the second as the residual error. If the weighting parameters p_1 to p_n of the circuits 18 are optimised then the residual error will be minimised. To enable the reproduction by a linear predictor of an original speech signal it is only necessary to transmit or store in each frame the weighting parameters and an excitation signal. The residual error, if used as the excitation, yields perfect reproduction of the original speech.

The technique described above works well for speech signals because the operation simulates the acoustic properties of the human vocal tract. When a sound is uttered a vibration is transmitted down the vocal tract which is configured to produce the desired sound.

The configuration of the vocal tract, being due to physical movement of articulatory organs, can only change quite slowly. The analogy between the configuration of the vocal tract and the weighting parameters allows much of the information in the speech signal to be transmitted at a low data rate. While this ensures good intelligibility, the quality and naturalness of the reproduced speech is largely dependent on the excitation signal used.

In a system which has been proposed in the past, the parameters of the predictor are transmitted or stored and the excitation signal is selected either as white noise or as a regular series of pulses depending on the type of sound to be produced. Even using such crude simulation of the residual signal it was possible to produce recognisable speech. However, though the quality was acceptable for certain applications, for example military applications where maximum signal compression

was of most importance, it fell below acceptable commercial standards.

In order to improve the quality of the reproduced speech, it is necessary to put more information into the excitation signal so that it should resemble the residual signal more closely. With this aim in mind, it has been proposed that in each frame the predictor should be excited by a train of pulses, in which the timing and the magnitude of each pulse in the train should be selected in order to minimise the difference between the re-synthesised speech and the original speech signal. In this last case, the excitation signal does not depend on the type of sound to be produced but for each frame the ideal excitation pulse train is computed.

The multi-pulse-exicted linear-reductive model of speech generation was presented by Atal and Remde in their paper "A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates" Proc. ICASSP 1982 pp 614-617. This model bypasses neatly the problem of inflexible classification of speech segments into voiced and unvoiced sounds encountered in previous approaches to vocoding. It has been used to demonstrate very good reproduction of speech from parameters encoded at bit-rates estimated to be in the region of 10 kbits/s.

There now follows the derivation of multi-phase excitation parameters based on a model comprising a multi-pulse excitation generator coupled to a linear predictor. The multi-pulse excitation signal, u(n), is a sequence of samples whose values are zero in all but a few positions. The amplitudes and positions of the non-zero samples are chosen so as to minimise a perceptually meaningful error. The details of a possible formulation are summarized below.

A sequence $\hat{s}(n)$ is to be synthesised over the interval n = 1 .. N by exciting a linear predictor with the multi-pulse sequence u(n). The linear predictor has a transfer function H(z), corresponding to an impulse response h(n). The sequence u(n) contains at most K non-zero samples $u(n_k)$, k=1..K, where K < N. The positions n_k and the values $u(n_k)$ are to be chosen so as to minimise the energy in the error sequence.

$$e(n) = \left[s(n) - \hat{S}(n) \right] * w(n) \qquad (1)$$

where s(n) is the sequence of samples of original speech, w(n) is the impulse response corresponding to a spectral weighting function W(z), and * denotes convolution.

The problem is therefore to determine

n_k, $u(n_k)$ for k = 1..K

so as to minimise

$$E = \sum_{n=1}^{N} e^2(n) \qquad (2)$$

In order to avoid the complexity of determining all 2K unknowns simultaneously, an iterative procedure can be adopted in which position and amplitude are evaluated for one non-zero sample at a time.

The jth iteration establishes the values n_j, $u(n_j)$ once n_k , $u(n_k)$ for k=1..j-1 have been determined by the previous j-1 iterations and with $u(n_k)$ set to zero for k > j. At the jth iteraion we minimise

$$E_j = \sum_{n=1}^{N} e_j^2(n) = \sum_{n=1}^{N} \left(\left[s(n) - \hat{S}(n) \right] * w(n) \right)^2 \qquad (3)$$

Setting

$$h1_{(n)} = h(n)*w(n) \tag{4}$$

and noting that

$$e_j(n) = e_{j-1}(n) - u(n_j).h^1(n-n_j) \tag{5}$$

we have

$$E_j = \sum_{n=}^{N} [e_{j-1}(n) - u(n_j).h'(n-n_j)]^2 \tag{6}$$

The optimum value of $u(n_j)$ is found by differentiating E_j partially with respect ot $u(n_j)$ and setting the derivative to 0. This yields

$$u(n_j) = \frac{\sum_{n=1}^{N} e_{j-1}(n).h'(n-n_j)}{\sum_{n=1}^{N} [h'(n-n_j)]^2} \tag{7}$$

The minimum value of E_j for a given n_j can then be obtained by combining (6) and (7) into

$$E_{j,min} = \sum_{n=1}^{N} e_{j-1}^2(n) - u^2 n_j \sum_{n=1}^{N} [h'(n-n_j)]^2 \tag{8}$$

From (6), E_{jmin} cannot be negative. Therefore E_j is minimised in (8) if n_j is chosen such that $|u(n_j)|$ is a maximum.

The sequence e(n) and values of $u(n_j)$ for all possible n_j must be recomputed at each iteration over the interval of interest. The procedure can be refined by re-adjusting the amplitudes of all selected samples simultaneously, once their positions are all known.

The procedure described above is ill-suited to implementation in real-time at low-cost with current hardware technology because of the large computation

rate and because of the inherent block-processed structure of the algorithm.

The present invention is intended to encode and decode speech using linear predictive coding in which the LPC filter is excited by a series of pulses whose positions and amplitude are capable of being computed in real time.

According to the present invention, there is provided an encoder for encoding speech signals, comprising means for sampling frames of the speech signal to be encoded, a linear prediction analyser for determining for each frame the weighting parameters of a linear predictor to minimise the residual signal for the sampled frame, and means for producing an excitation signal for transmission or storage in conjunction with the parameters to enable each frame of the speech signal to be resynthesised, in which the means for producing an excitation signal comprises means for correlating a signal derived from the speech signal in that frame with the time weighted impulse response of a linear predictor having the weighting parameters determined by the analyser by the analyser.

The expression "time weighted" is intended to signify that the response has the same shape but decays more rapidly, this being achieved by multiplying the parameter p_n by a factor k^n, where $k<1$.

A linear recursive filter if excited by a single pulse may have an impulse response of very long time duration and provided that it is not unstable will eventually decay rather than oscillate. The effect of a long time response is that responses from consecutive excitation pulses tend to run into each other and it is difficult when performing a correlation to separate the pulse response of one excitation from another.

In the preferred embodiment of the invention, the speech signal is passed through a weighting filter, preferably a pole-zero filter, which has the effect of damping reverberations. The weighting filter has a non-recursive part the weighting parameters of which are of the same magnitude as, but of opposite sign to, those of the linear predictor in the decoder. In the analogy mentioned above one may regard the purpose of the non-recursive side of the weighting filter as negating the effect of the vocal tract on the pulses originally generated within the throat of the speaker. The other side of the filter, on the other hand, the recursive part, has weighting coefficients which are related to those of the linear predictor but are weighted by a factor which follows a power law of k^n, $(k<1)$, so that time-weighting of the impulse response is achieved.

If one correlates the speech signal after passing it through such a weighting filter with the impulse response of a filter which consists only of the recursive side of the weighting filter when excited by a single excitation pulse, then the correlator will produce a high correlation output at the times when impulses should be applied to the linear prediction filter in order to simulate the speech signal.

Thus, in the preferred embodiment, the weighting filter is followed by a correlator of which the output is fed to an impulse selector. The purpose of the impulse selector is to select from amongst the peaks of the output of the correlator a number of peaks having the highest magnitude. These peaks determine the time at which the residual signal should be applied to the linear predictor in the decoder in order to resynthesise the speech signal.

Also in the preferred embodiment, the peaks are selected such that they are all of the same polarity. This

polarity can be set so as to match the polarity of the microphone being used. If the polarities of the peak selection and microphone are correctly matched, then this improves the quality of the resynthesised speech by helping to preserve its harmonic content.

It is also preferred that the excitation pulses should have an amplitude related to the amplitude of the peak produced by the correlator. Because the auto-correlation functions of the pulse responses of the LPC filter are not constant but vary with the weighting parameters, it is preferred that the excitation pulse amplitude should be derived by dividing the correlator output by the value of the auto-correlation function of the impulse response of the filter with the prevailing time weighted parameters.

The invention will now be described further, by way of example, with reference to the accompanying drawings, in which:

Figure 1 is, as earlier described, a diagram of a linear predictive filter;

Figure 2 is a block circuit diagram of an encoder in accordance with the present invention; and

Figure 3 is a diagram showing a weighting filter.

In Figure 2, the speech signal to be encoded is received over an input line 30. The input signal is applied to a known circuit 32 which is a linear prediction analyser. This circuit computes the values of the weighting parameters of a digital recursive filter which would minimise the residual signal and outputs these parameters. As is known, a linear prediction analyser more readily computes so called reflection co-efficients which are not the same as the weighting parameters but

from which these parameters can be computed. The reflection co-efficients are applied to a line 34.

The speech signal is also applied via a line 36 to a weighting filter 38 which will now be described by reference to Figure 3. The weighting filter comprises an input line 40 connected to a summation circuit 42 having an output line 44. A multi-tapped delay line (or shift register) 46 is connected to the input line 40 and a similar multi-tapped delay line 48 is connected to the output line 44. The tappings of the delay line 46 are connected by way of a first set of weighting circuits 50 to the circuit 42 which also receives signals from the tappings of the delay line 48 through weighting circuits 52. The values of the parameters used in the multiplication circuits of the weighting filter 38 in Figure 3 are derived from the linear prediction analyser 32.

In a block 60, the weighting parameters p_1 to p_n, equivalent to the reflection coefficients are computed. In the coefficient weighting circuits 32, two sets of parameters are derived from the parameters p_1 to p_n for setting the parameters of the weighting filter 38. The first set of parameters is applied to the weighting circuits 50 and are equal to $-p_1$ to $-p_n$. Thus the combination of the summation circuit with the delay line 46 and the weighting circuits 50 results in a digital non-recursive filter having parameters which are the opposite of those used in the receiving circuit to re-synthesize the speech signal. As previously stated, the effect of the non-recursive part of the weighting filter is to negate the effect of the vocal tract.

The second set of parameters evaluated by the co-efficient weighting circuit 62 is equal to $k.p_1$ to $k^n.p_n$, where k is less than 1. Thus, the delay line 48 and the weighting circuits 52 produce in conjunction with the summation circuit 42 a recursive digital filter

whose pulse response is similar to that of the filter used to resynthesize the speech but with more rapid decay. The effect of combination of the non-recursive and recursive filters which constitute the weighting filter 38, which is also termed a pole-zero filter, is to produce from the speech signal one in which reverberations are more severely damped to reduce the interaction between the effects of consecutive excitation pulses.

The output of the digital weighting filter 38 is applied to a correlator 64 connected to a circuit 66 which evaluates the impulse response of a digital recursive filter of the same construction as that shown in Fig. 1 but with weighting parameters $k.p_1$ to $k^n.p_n$.

The correlator 64 may consist of a shift register whose tapping are connected to multiplication circuits the multiplication factors of which are determined by the impulse response evaluating circuit 66. When there is a high level of correlation between the output of the weighting filter 38 and the impulse response evaluated by the circuit 66, a high output is produced by the correlator. The output of the correlator 64 thus contains peaks which coincide with impulses in the excitation signal which, if applied to the linear predictor at the decoder, will cause a good approximation to the original speech signal to be produced. However, in order to reduce the bit rate, it is necessary to select from amongst the correlator output only a small number of pulses and these should coincide with the impulses of maximum energy in the excitation signal.

The purpose of the pulse selector circuit 70 in Figure 2 is to select the timing of the pulses which are to be encoded. One could merely store the output values from the correlator and select the highest peaks but this

could result in consecutive high values being used to produce excitation pulses when they are truly the flanks of the same pulse. Therefore, it is preferable that the impulse circuit locate local maxima and minima and disregard the values adjacent to these peaks. One possible algorithm would be to disregard high values adjacent a local maximum or minimum if they are not separated from the local maximum or minimum by a zero crossing or a turning point. Another possible algorithm is to select a fixed number of the greatest peaks in each time frame and to ensure that they are separated by at least some minimum time.

The amplitude of the selected pulses will be related to the amplitude of an optimal excitation signal. In order to normalise these pulses to take into account the different values of the auto-correlation function of the impulse responses, the impulse reponse circuit 66 additionally evaluates the auto-correlation function of each pulse response and applies a signal over a line 72 to a divider circuit 74. In the divider circuit 74, the selected pulses are divided by the auto-correlation value and the output signal from the divider is fed to a multiplexer 76 which encodes the reflection coefficient received over the line 34 and the signals from the divider 74 to produce the encoded signal on output line 78 for transmission or storage.

The mathematical considerations underlying the invention are now considered for completeness but the successful operation of the apparatus of the invention is not dependent upon the accuracy of the analysis.

The preferred embodiment of the invention proposes making some simplifying assumptions in order to derive a modified algorithm which permits implementation in real-time of a 7.2 kbits/s vocoder using standard components on a double Eurosize circuit board.

Defining the linear predictor of order M as

$$H(z) = \frac{1}{1 - \sum_{m=1}^{M} a_m z^{-m}} \qquad (9)$$

We now define the weighting function, W(z) by

$$W(z) = \frac{1 - \sum_{m=1}^{M} a_m z^{-m}}{1 - \sum_{m=1}^{M} a_m \gamma^m z^{-m}} \qquad (10)$$

where γ is a real number between 0 and 1. The filter W(z) serves to de-emphasize the error signal e(n) in the formant regions, reflecting the fact that distortion in these regions is masked by relatively large concentrations of energy in the speech signal. Broadly speaking, the de-emphasis effect is enhanced by reducing γ.

If the linear-redictive analysis method employed leads to an unconditionally-stable linear predictor, then the envelope of its impulse response, h(n), decays with time.

The impluse response $h^1(n)$, defined in (4), corresponds to the cascade of the transfer functions H(z) and W(z). Some thought shows that

$$h'(n) = h(n).\gamma^n \qquad (11)$$

Since $\gamma < 1$, the envelope of $h^1(n)$ decays more rapidly than that of h(n). Combining this with the causality of the linear predictor we can write

$$h^1(n) = 0 \text{ for } n < 0$$
$$h^1(n) \approx 0 \text{ for } n \Rightarrow n_g \qquad (12)$$

In (12) n_g is an arbitary positive integer. The approximation in (12) can be improved by increasing n_g and/or by reducing τ.

Furthermore, (12) can be applied to (5) to yield

$$e_j(n) = e_{j-1}(n) \quad \text{for} \quad (n-n_j) < 0 \qquad (13)$$
$$e_j(n) \approx e_{j-1}(n) \quad \text{for} \quad (n-n_j) \Rightarrow ng$$

We now apply the restriction

$$|n_j - n_i| \Rightarrow ng \text{ for } i = 1..j-1 \qquad (14)$$

requiring a minimum separation of n_g between non-zero samples of the excitation sequence. This restriction can extend (13) to

$$e_{j-1}(n) \approx e_0(n) \quad \text{for} \quad |n-n_i| \Rightarrow n_g, i=1..j-1 \qquad (15)$$

The sequence $e_0(n)$ is defined as

$$e_0(n) = [s(n) - \hat{s}_0(n)]*w(n) \qquad (16)$$

where $\hat{s}_0(n)$ is the output of the linear predictor driven with zero input.

Using equations (12) and (15), we can rewrite (7) in approximate form as

$$u(n_j) = \frac{\sum_{n=n_j}^{n_j+n_g-1} e_0(n) \cdot h'(n-n_j)}{\sum_{n=0}^{n_g-1} [h'(n)]^2} \qquad (17)$$

subject to the restriction of (14).

An approximate solution to the problem of determing the positions and amplitudes of the non-zero excitation samples can therefore be found by computing the following equation :

$$\phi(n) = \frac{\sum_{m=0}^{n_g-1} e_o(n+m) \cdot h(m)}{\sum_{m=0}^{n_g-1} [h'(m)]^2} \qquad (18)$$

and selecting values of $n=n_k$, for which the corresponding values of $|\phi(n)|$ are local maxima, subject to the restriction of (14).

The modified computation exploits an alternative interpretation of the role of the weighting function W(z). The effect of the weighting function can be viewed as an attempt to separate the response of the system to successive non-zero excitation samples. If these samples are far enough apart, their values can be optimized independently.

In low bit-rate applications it is desirable to place the non-zero samples far apart, so as to distribute them roughly evenly across the interval of synthesis.

CLAIMS

1. An encoder for encoding speech signals, comprising means for sampling frames of the speech signal to be encoded, a linear prediction analyser for determining for each frame the weighting parameters of a linear predictor to minimise the residual signal for the sampled frame, and means for producing an excitation signal for transmission or storage in conjunction with the parameters to enable each frame of the speech signal to be resynthesised, characterised in that the means for producing an excitation signal comprises means (64) for correlating a signal derived from the speech signal in that frame with the time weighted impulse response of a linear predictor having the weighting parameters determined by the analyser.

2. A signal encoder as claimed in Claim 1, in which the signal derived from the speech signal is obtained by means of a weighting filter (38) which is operative to damp reverberations within the speech signal caused by resonances in the vocal tract and precedes the correlating means (64).

3. A signal encoder as claimed in Claim 2, in which the weighting filter (38) comprises a pole-zero filter.

4. A signal encoder as claimed in any preceding claim, in which the correlating means (64) comprises a tapped delay line, means for multiplying the tapped signals by the said time weighted impulse response, and means for summing the outputs of the multiplication circuits.

5. A signal encoder as claimed in any preceding claim in which the output of the correlating means is connected to a pulse selector (70) which is operative to select a number of pulses from the correlator output.

6. A signal encoder as claimed in Claim 5, in which the pulse selector (70) comprises means for detecting local peaks and means for selecting amongst the local peaks, those having the most positive or the most negative amplitudes.

7. A signal encoder as claimed in any preceding claim, comprising means (66,74) for determining the magnitude of the transmitted pulses by dividing the output of the correlating means (64) by the auto-correlation function of the said time weighted impulse response.

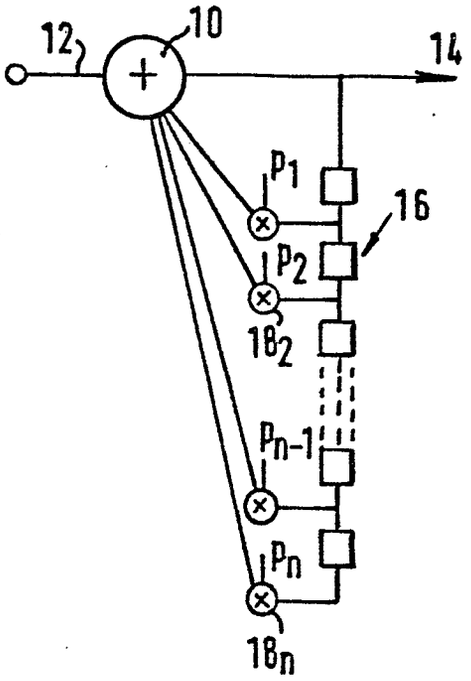

FIG.1.

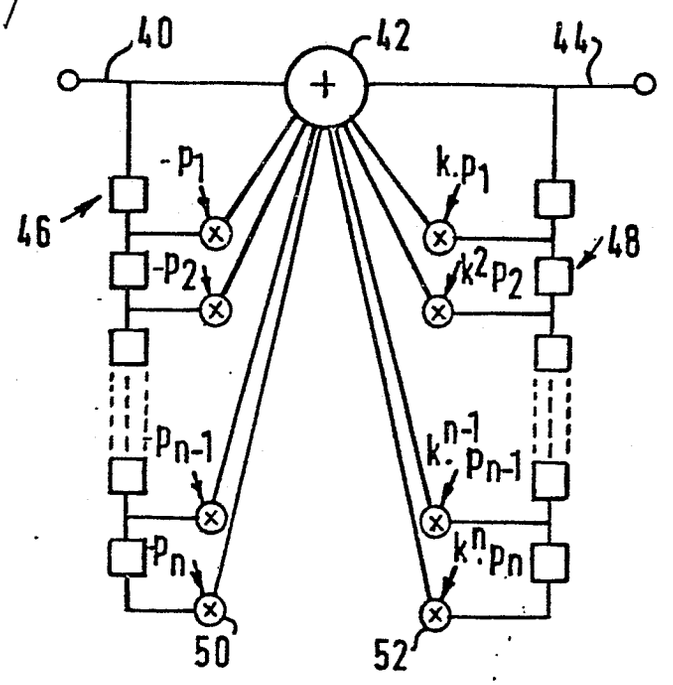

FIG. 3

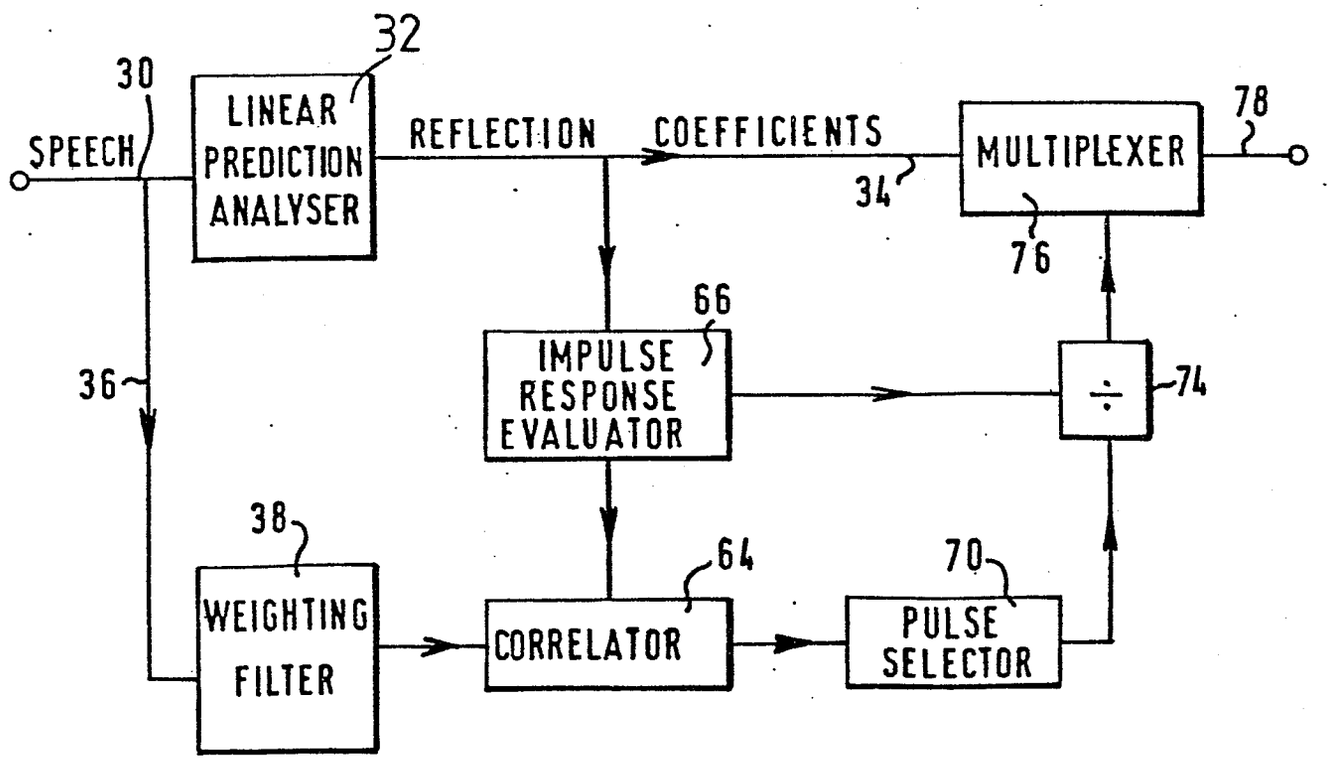

FIG. 2.

European Patent Office

EUROPEAN SEARCH REPORT

0119033
Application number

EP 84 30 1302

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. ³)
P,X	GB-A-2 110 906 (WESTERN ELECTRIC CO.) * Figure 1 *	1	G 10 L 1/08

TECHNICAL FIELDS SEARCHED (Int. Cl. ³)

G 10 L 1/08

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	22-05-1984	ARMSPACH J.F.A.M.

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