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

(11) Publication number: **0 440 335 A2**

(12) EUROPEAN PATENT APPLICATION

(21) Application number: **91300106.1**

(51) Int. Cl.[5]: **G10L 9/14**

(22) Date of filing: **08.01.91**

(30) Priority: **01.02.90 GB 9002282**

(43) Date of publication of application:
07.08.91 Bulletin 91/32

(84) Designated Contracting States:
AT BE DE ES FR GB IT NL SE

(71) Applicant: **PSION PLC**
Alexander House, 85 Frampton Street
London, NW8 8NQ(GB)

(72) Inventor: **Wolovitz, Lionel Berel**
66 Strahan Road
London E3 5DB(GB)

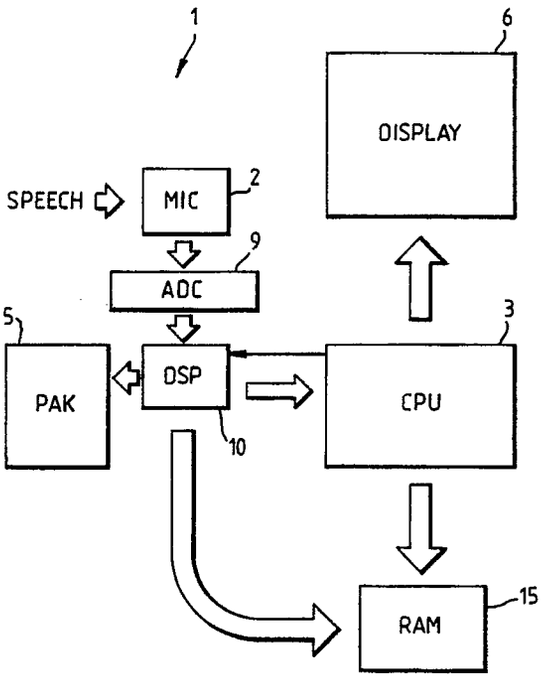
(74) Representative: **Wells, David et al**
Gill Jennings & Every 53/64 Chancery Lane
London WC2A 1HN(GB)

(54) Encoding speech.

(57) An apparatus for encoding speech includes circuits (2,9) which sample a speech signal and generate digital data representative of the signal. An encoder (10,11) connected to the sampling circuits encodes the digital data by linear predictive coding and generates parameters representing the speech signal. The parameters are transformed (12) to obtain sum and difference polynomials. The roots of the sum and difference polynomials are determined (13) producing a series of line spectrum data. A storage device (14,15) stores the line spectrum data.

In one example, the line spectrum data are quantised non-uniformly in the frequency domain and the roots of the polynomials are found by evaluating those polynomials at the quantisation frequencies only.

Fig. 1.



EP 0 440 335 A2

ENCODING SPEECH

The present invention relates to data compression techniques, and in particular to the encoding of speech in a form suitable for storage in the memory of a computer.

It is desirable to provide personal computers, particularly portable hand held or lap-top machines, with the facility to record speech. Recorded speech might be used, for example, to annotate a document in a word processor operating on the machine or might be used to give prompts in association with a diary controlled by the machine.

It is known to provide personal computers with DSPs (digital signal processors) which can be used to process sampled sounds including speech for storage by the computer. However in order to record sound with acceptable quality known systems require a very high data rate, so that in practice no more than a few seconds of speech can be stored in the few megabytes of RAM typically available, and even the tens or hundreds of megabytes available on mass storage devices such as hard disks are quickly exhausted. In order to overcome this problem data compression techniques might be used, however known compression algorithms capable of producing the very high compression ratios desirable in this context are computationally very intensive. Therefore, because of the dual constraints of limited memory and limited processor power in personal computers, it has not previously been possible to record extended periods of speech on such systems.

According to a first aspect of the present invention, a method of encoding speech for storage in a computer comprises, sampling a speech signal, encoding the sampled speech data by linear predictive coding, thereby producing parameters representing the speech data, transforming the parameters to obtain sum and difference polynomials, determining the roots of the polynomials thereby producing line spectrum data, and storing the line spectrum data in the computer.

The present inventor has developed an encoding technique using a combination of features which together provide markedly superior performance to known systems, making possible compression ratios as high as 100:1. At the same time, the algorithm is sufficiently computationally efficient to be implemented on a relatively low-powered hand-held personal computer using a conventional off-the-shelf DSP.

Linear predictive coding is in itself a well-known technique for speech analysis and synthesis and is described, for example, in the paper by B.S. Atal et al "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave" published at pp637-655 of The Journal of the Acoustical Society of America vol. 50 no. 2 (part 2) 1971. The use of Line Spectrum Pairs is described in "The Computation of Line spectral Frequencies using Chebyshev Polynomials" P. Kabal et al IEEE Trans. on Acoustics Speech and Signal Processing, vol. ASSP-34, no. 6, December 1986 and in the paper by George S. Kang et al. published at pp 568-571 of IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-35 no. 4, April 1987. However it has not previously been thought possible in practice to implement such techniques in relatively low-powered computers because of the computationally intensive nature of the root-finding algorithms used in deriving the line spectrum pairs, while without the use of line spectral pairs the compression ratios achieved are insufficiently high.

Preferably the stored line spectrum pairs are quantised in the frequency domain with a non-uniform quantisation and the roots of the polynomials are found by evaluating the polynomials at the quantisation frequencies only.

It is found that both the compression ratio and the computational efficiency of the algorithm can be improved by quantising the line spectral pairs in the frequency domain with an appropriate quantisation of the frequency range, with the line spectrum pairs being found by evaluating the relevant functions at the quantisation frequencies only.

Preferably the sampled voice data are divided into frames having a predetermined period, successive frames being compared and when a new frame differs by less than a predetermined degree from a preceding frame the new frame being coded by an instruction to repeat the preceding frame.

The compression ratio is further improved by only storing a new frame when it differs by more than a predetermined degree from the preceding frame. When the frames are the same then the new frame is encoded by a bit in a certain field which is recognised, on reconstruction of the data, as an instruction to repeat the preceding frame. It is found that with appropriate criteria for comparison between the frames, as described in further detail below, the number of frames to be stored can be reduced by as much as 50%, without significant loss in quality.

Preferably the method includes displaying graphically a representation of the voice signal, and editing the stored parameterised voice signal in response to operations on the graphical representation of the signal carried out by the user.

Using compression techniques such as those provided by the first aspect of the present invention it becomes feasible to store and edit speech using

the computer in much the same way as text is stored and edited using a word processor. However the problem arises of providing an appropriate user interface. It is known in personal computers incorporating simple digital samplers to provide a display of sampled data in the form of a plot of amplitude against time. While such a display may be of some use in handling relatively uniform signals such as music it is found to be more difficult to handle a signal as complex as speech. The present inventor has found that, when the signal is stored in compressed and parameterised form a display can be generated from the stored parameters which greatly facilitates the editing of complex signals such as speech. The user can then carry out most of the operations normally provided in a text processor, including cutting and pasting, copying and erasing.

Preferably the method further comprises selecting a point in the voice signal for editing by moving a cursor with respect to the graphical representation of the signal while at the same time reproducing the portion of the signal marked by the cursor.

According to a second aspect of the present invention there is provided an apparatus for encoding speech comprising:

means for sampling a speech signal thereby generating digital data representative of the speech signal,

an encoder connected to the means for sampling and arranged to encode the digital data by linear predictive coding thereby generating parameters representative of the speech signal,

means for transforming the said parameters and calculating corresponding sum and difference polynomials,

means for determining the roots of the sum and difference polynomials thereby generating a series of line spectrum data, and

storage means for storing the line spectrum data.

According to a third aspect of the present invention there is provided a computer including means for encoding and decoding speech, said means for encoding and decoding comprising:

means for sampling a speech signal thereby generating digital data representative of the speech signal,

an encoder connected to the means for sampling and arranged to encode the digital data by linear predictive coding thereby generating parameters representative of the speech signal,

means for transforming the parameters and calculating corresponding sum and difference polynomials,

means for determining the roots of the sum and difference polynomials thereby generating a series of line spectrum data,

means for storing the line spectrum data,

means for retrieving the line spectrum data from the means for storing,

means for determining sum and difference polynomials corresponding to the retrieved line spectrum data,

means for transforming the sum and difference polynomials and generating the parameters representative of said speech signal,

a decoder arranged to decode the parameters thereby generating digital data representative of the speech signal, and

means for generating and outputting an analogue signal corresponding to the digital data.

A method and apparatus in accordance with the present invention will now be described in detail with reference to the accompanying drawings in which:

Figure 1 is a block diagram of a computer for use in the present invention;

Figure 2 is a graph showing the root of a polynomial function;

Figures 3A and 3B are block diagrams showing schematically an encoder and decoder respectfully; and

Figure 4 is a diagram showing a display for use in editing a voice signal.

A laptop personal computer 1 includes a microphone 2 connected via an analogue to digital converter ADC, to a digital signal processor (DSP). Where appropriate, data output by the DSP is written under control of the main CPU 3 to a mass storage device 4 which in the present example is a battery-packed RAM cartridge or to the main RAM 5 of the computer 1.

The ADC samples the speech waveform received by the microphone, producing a data stream at 64 kbits/s. The resulting data stream in divided into frames having, in the present example, a duration of 25 ms. For each frame a number of parameters are calculated. These parameters represent the speech within the frame as an exciting frequency and a number of predictor coefficients which define a discreet time-varying linear filter having a transfer function which models the effect of the vocal tract on the excitation produced by the vocal cords. If the frame comprises p samples, a_k is a general predictor coefficient, and s_n is a general sample value then the prediction error E_n is the difference between the speech sample s_n and its predicted value $\hat{s}_n$ given by

$$\hat{S}_n = \sum_{k=1}^{p} a_k s_{n-k}$$

E_n is then given by

$$E_n = S_n - \hat{S}_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}.$$

The values for the predictor coefficients are chosen to minimise the mean-squared prediction error $\langle E_n^2 \rangle_{av}$. As described in greater detail in the above cited paper by B.S. Atal et al, applying this constraint to the set of sampled data results in a set of simultaneous linear equations which are solved to give the predictor coefficients.

In order to achieve further data compression the predictor coefficients are not stored directly but are used to define a polynomial function which is decomposed into sum and difference polynomials. The roots of these sum and difference polynomials constitute line spectrum pairs and it is these line spectrum pairs which are output by the DSP for storage by the computer and which may subsequently be used for a complementary synthesis process. The root finding algorithm and the complementary algorithm for deriving the predictor coefficient from the stored line-spectrum pairs is described in detail in the accompanying appendix. The roots are found by evaluating the polynomials at different frequencies. Rather than using a continuous range of frequencies the functions are evaluated at predetermined intervals with a non-uniform spacing. In the present example 42 logarithmically spaced frequencies are used for quantisation. These frequencies, listed in Table 1, are chosen to provide more quantisation levels in the regions to which the ear is most sensitive. Then, as shown in Figure 2 the frequencies f_1, f_2 between which a change in sign occurs define the position of the root. Of the two frequencies, f_1, f_2 the one at which the polynomial p(f) is closer to zero is chosen to represent the root. For clarity, Figure 2 shows a much simplified function as representing the polynomial.

As a result of the above described process, each frame is represented by a series of first, second,... nth roots, each root comprising an index from 1 to 42 corresponding to a particular frequency quantisation level. It is found that a further marked increase in the compression ratio can be achieved by comparing successive frames and when a new frame differs by less than a predetermined degree encoding that frame by an instruction to repeat the previous frame. The comparison is made between corresponding roots of the different frames, that is the first root in one frame is compared with the first root in the other frame and the second root in one frame with the second root in the other frame and so on. The difference in terms of the quantisation units of the corresponding roots is determined. It is found that a frame can be replaced by a "repeat" instruction when none of the roots compared in this manner differ by more than 3, without significant loss of quality in reproduction. If roots are allowed to differ by as much as 5 then audible degradation occurs. Using a criterion of 3 around 50% of frames are repeated. In order further to improve the quality of reproduction tighter constraints can be used. If, for example, a frame is repeated only when no root differs by more than 1 from the corresponding root in the preceding frame then while the percentage of frames which can be encoded as repeats is reduced there is a corresponding increase in the quality of reproduction. It is found that the higher frequency roots can be allowed to vary by more than the lower frequency roots and so, in the preferred embodiment, the criterion for a repeat frame is that the first four roots should differ by no more than 1 quantisation unit and the remaining roots should differ by no more than 2.

In total each frame output by the DSP comprises a first parameter indicating whether the frame is voiced or unvoiced, a second parameter defining the pitch period of the exciting tone together with the line-spectrum pairs calculated from the predictor coefficients $a_1 \ldots a_p$ The pitch period may be determined using any one of a number of conventional pitch determining techniques, such as that described in the paper by B. Gold et al, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain" published at pp442-448, J. Acoust. Soc. Amer., vol. 46 August 1969.

Once the line spectrum pairs have been stored then the speech can be reproduced at any time by recalling the stored pairs from memory, and applying a complementary algorithm to calculate the predictor coefficient from the line spectrum pairs. An appropriate algorithm is listed in appendix 2. The resulting data are output by a DAC, audio amplifier and loudspeaker (not shown).

The voice data stored in the computer may be edited by the user. A window 7 is opened in the display 6 of the computer and a selected portion of the parameterised voice data displayed. Using a cursor 8 under the control of an input device such as a mouse or digitising pad the user can manipulate the displayed data, carrying out functions such as cutting and pasting, copying or erasing. In accordance with the input from the input device, corresponding changes are made in the stored parameters.

As the cursor marking the edit point is moved relative to the display the frames corresponding to the point under the cursor at any instant are repro-

duced in a sequence and at a speed determined by the speed and direction of movement of the cursor. The user is therefore able to select an edit point by ear without having to proceed by trial and error, selecting an edit point, replaying the selected portion in a separate subsequent operation and so on. since the data is parameterised with the pitch coded independently of the other parameters the variation in speed does not change the pitch of the reproduced voice, and so the cursor can be moved at different speeds without loss of intelligibility.

In an alternative mode of editing the cursor is used to mark an anchor point. The voice data from the anchor point onwards is reproduced. If the cursor marking the anchor point is moved then immediately the reproduction of data starts again from the new anchor point.

The way in which the display is derived from the stored data can be varied according to the requirements of any given field of use and the limitations of a particular display device. In the present example each point plotted in the display represents the mean amplitude of four successive frames.

Claims

1. A method of encoding speech for storage in a computer comprising sampling a speech signal, encoding the sampled speech data by linear predictive coding, thereby producing parameters representing the speech data, transforming the parameters to obtain sum and difference polynomials, determining the roots of the polynomials thereby producing a series of line spectrum data, and storing the line spectrum data in the computer.

2. A method according to claim 1, in which the stored line spectrum data are quantised in the frequency domain with a non-uniform quantisation and the roots of the polynomials are found by evaluating the polynomials at the quantisation frequencies only.

3. A method according to claims 1 or 2, in which the sampled voice data are divided into frames having a predetermined period, successive frames being compared and when a new frame differs by less than a predetermined degree from a preceding frame the new frame being coded by an instruction to repeat the preceding frame.

4. A method according to claim 3, in which the frames are compared by determining the difference in quantisation units between corresponding roots in the respective series of line spectrum data, and a frame is repeated when none of the differences exceed a predetermined limit.

5. A method according to claim 4, in which different predetermined limits are provided for different parts of the respective series, with a lower predetermined limit being provided for lower frequency terms.

6. A method of storing and editing voice data comprising encoding and storing a voice signal by a method according to any one of the preceding claims, displaying graphically a representation of the voice signal, and editing the stored parameterised voice signal in response to operations on the graphical representation of the signal carried out by the user.

7. A method according to claim 6, further comprising selecting a point in the voice signal for editing by moving a cursor with respect to the graphical representation of the signal while at the same time reproducing the portion of the signal marked by the cursor.

8. A method according to claim 6 or 7, in which the step of displaying graphically the voice signal includes plotting as a function of time a parameter representative of the signal averaged over a plurality of frames.

9. An apparatus for encoding speech comprising:

means (2,9) for sampling a speech signal thereby generating digital data representative of the speech signal,

an encoder (10,11) connected to the means for sampling and arranged to encode the digital data by linear predictive coding thereby generating parameters representative of the speech signal,

means (12) for transforming the said parameters and calculating corresponding sum and difference polynomials,

means (13) for determining the roots of the sum and difference polynomials thereby generating a series of line spectrum data, and

storage means (14) for storing the line spectrum data.

10. An apparatus according to claim 9, further comprising means (19) for comparing successive frames of voice data having predetermined periods, and

means responsive to the means for comparing for generating and storing an instruction to repeat a preceding frame when the means for comparing determine that a current frame

differs by less than a predetermined degree from the preceding frame.

11. An apparatus according to claim 10, wherein the means (19) for comparing comprise means for reading series of line spectrum data corresponding to the respective frames, and means for determining the differences between corresponding terms in the series of line spectrum data and comparing the differences with a predetermined limit (δ), the frames being determined to differ by less than the said predetermined degree when none of the differences exceed the predetermined limit.

12. An apparatus according to claim 3, wherein the means (19) for determining and comparing differences include means for storing first and second different predetermined limits (δ_1,δ_2) for respective higher and lower frequency terms of the series, the magnitude of the predetermined limit for the lower frequency terms being less than the magnitude of the predetermined limit for the higher frequency terms.

13. An apparatus according to any one of claims 9 to 12 in which the means (13) for determining the roots of the sum and difference polynomials include means for evaluating the said polynomials at non-uniform quantisation frequencies only and thereby determining roots of said sum and difference polynomials, and the means (14) for storing the line spectrum store data quantised in the frequency domain at the said non-uniform quantisation frequencies.

14. An apparatus according to any one of claims 9 to 13, further comprising means (6) for displaying graphically a representation of the voice signal, and editing means for editing the stored parameterised voice signal in response to operations on the graphical representation carried out by the user.

15. A computer including means for encoding and decoding speech, said means for encoding and decoding comprising:

means (2,9) for sampling a speech signal thereby generating digital data representative of the speech signal,

an encoder (10,11) connected to the means for sampling and arranged to encode the digital data by linear predictive coding thereby generating parameters representative of the speech signal,

means (12) for transforming the parameters and calculating corresponding sum and difference polynomials,

means (13) for determining the roots of the sum and difference polynomials thereby generating a series of line spectrum data,

means (14) for storing the line spectrum data,

means for retrieving the line spectrum data from the means for storing,

means (16) for determining sum and difference polynomials corresponding to the retrieved line spectrum data,

means (17) for transforming the sum and difference polynomials and generating the parameters representative of said speech signal,

a decoder (18) arranged to decode the parameters thereby generating digital data representative of the speech signal, and

means (20) for generating and outputting an analogue signal corresponding to the digital data.

16. The computer of claim 15, wherein the computer is a hand-held or lap-top computer.

17. A hand-held or lap-top computer, including an apparatus according to any one of claims 9 to 14.

Fig. 1.

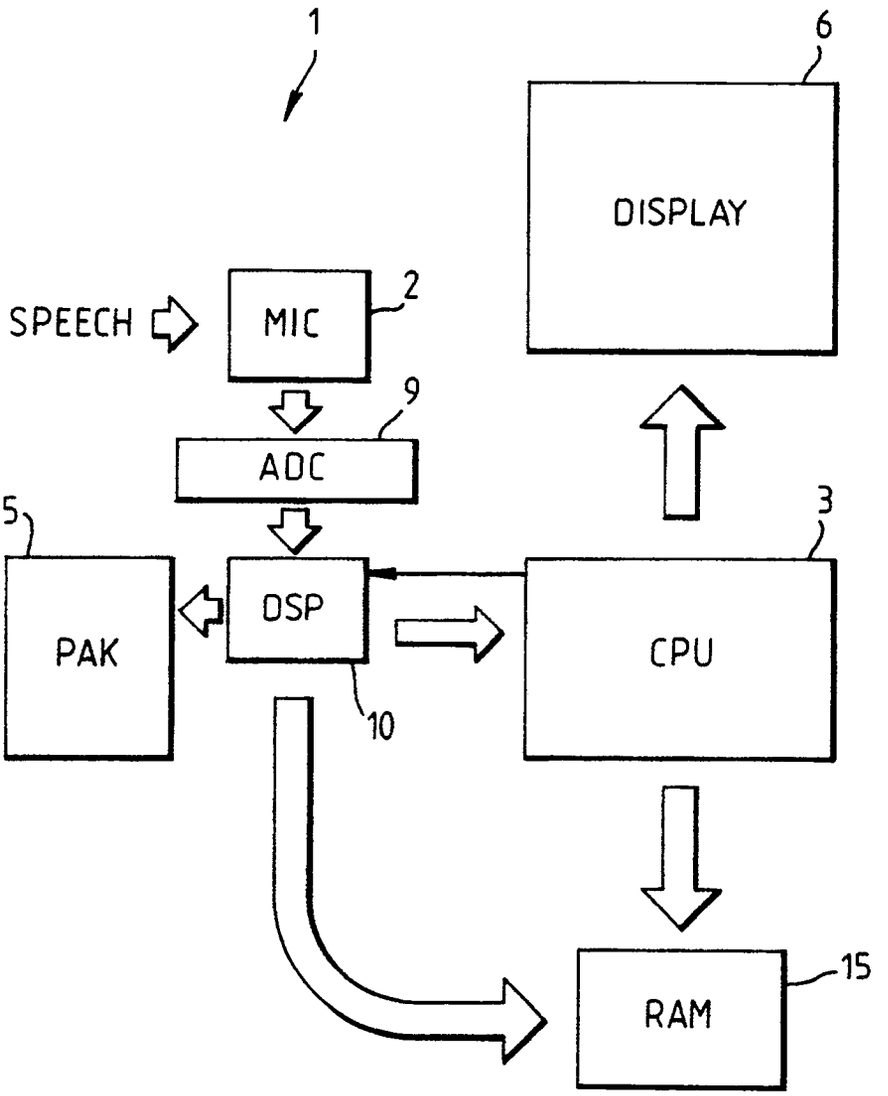
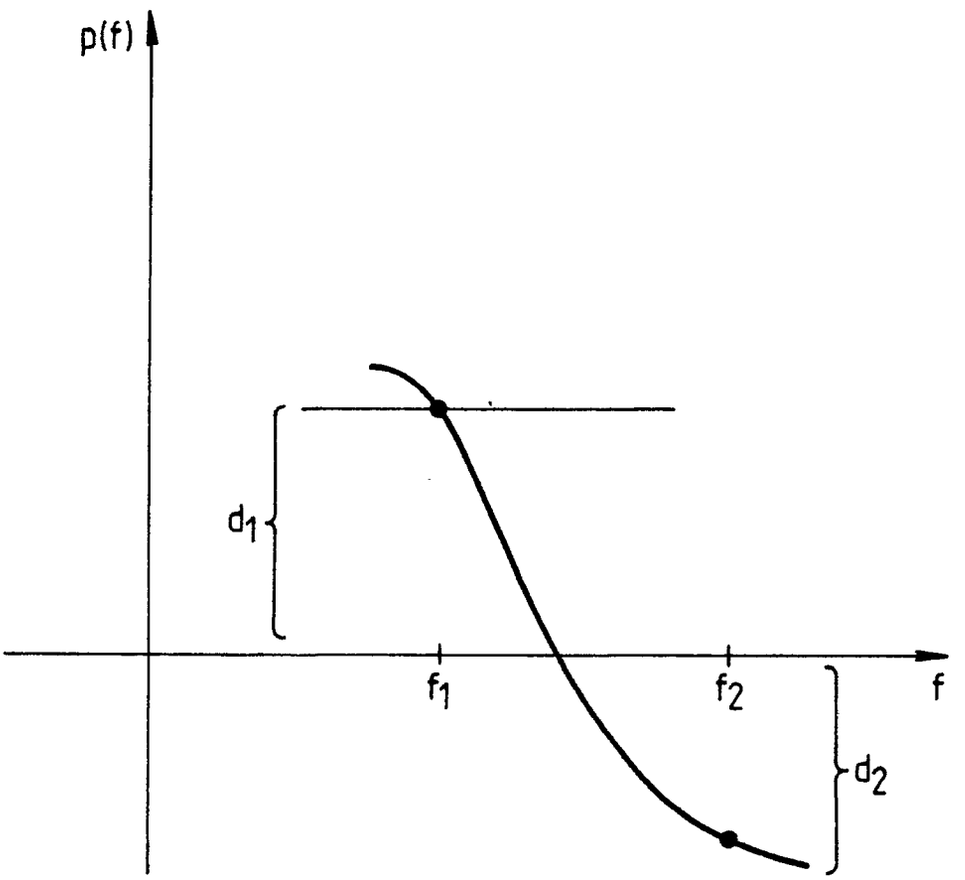


Fig. 2.



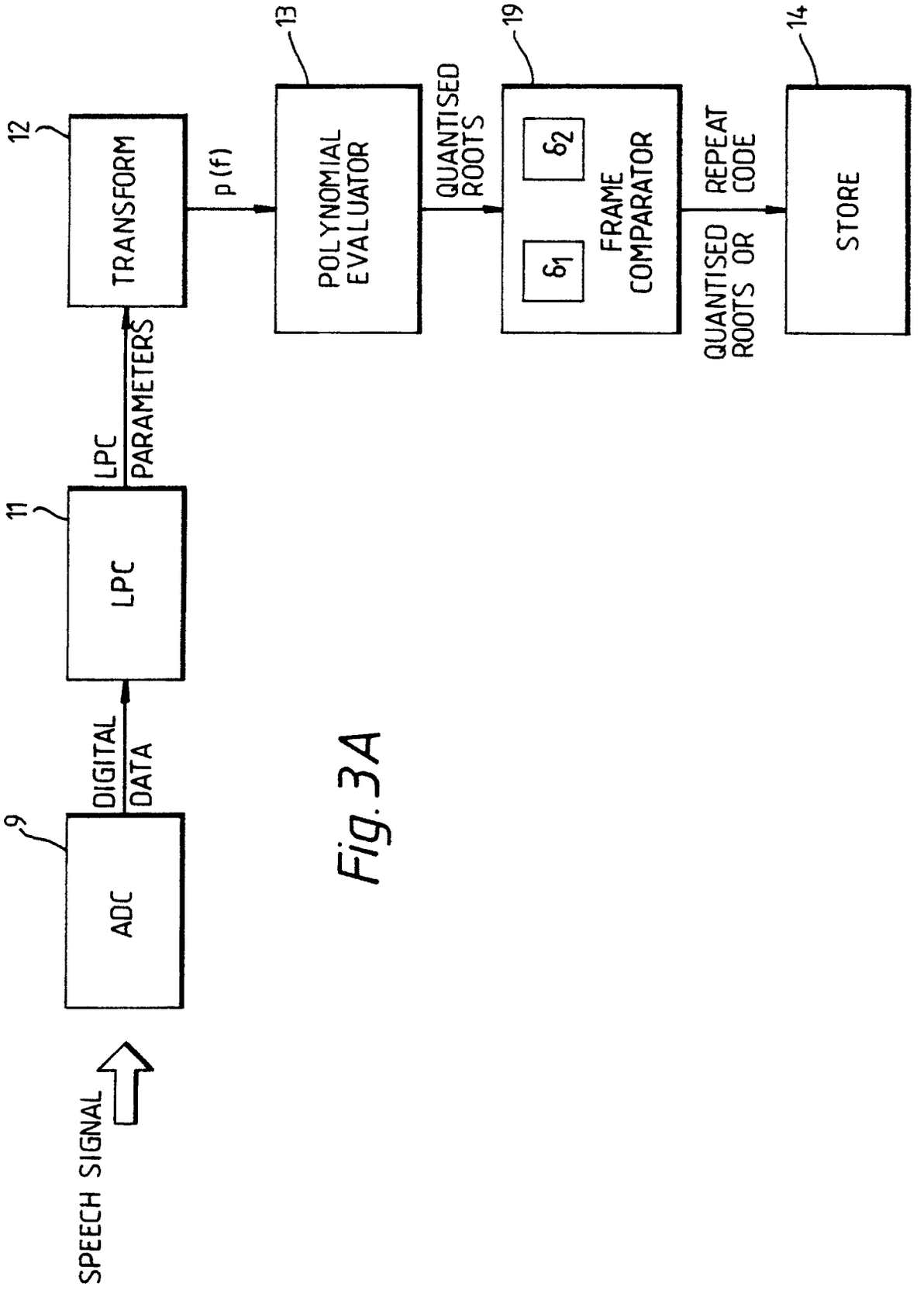
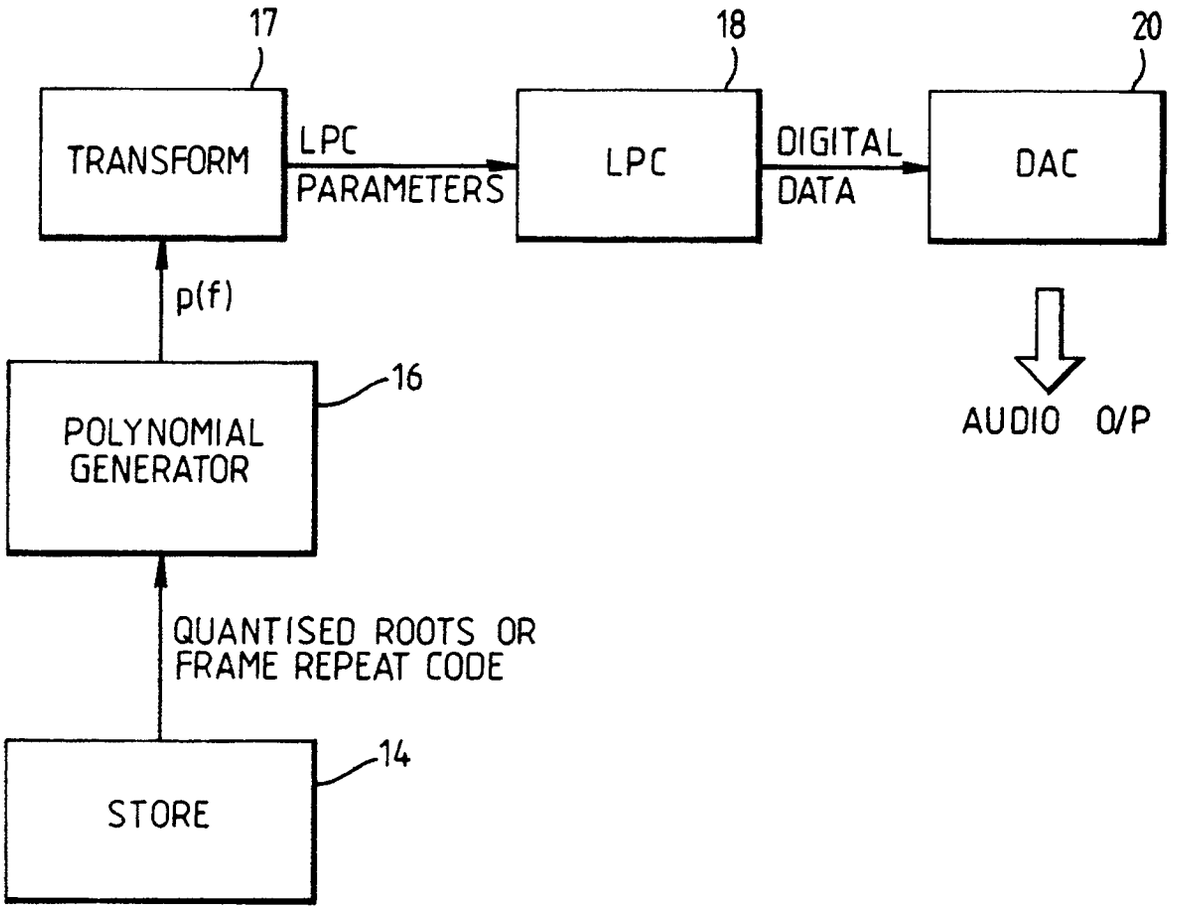
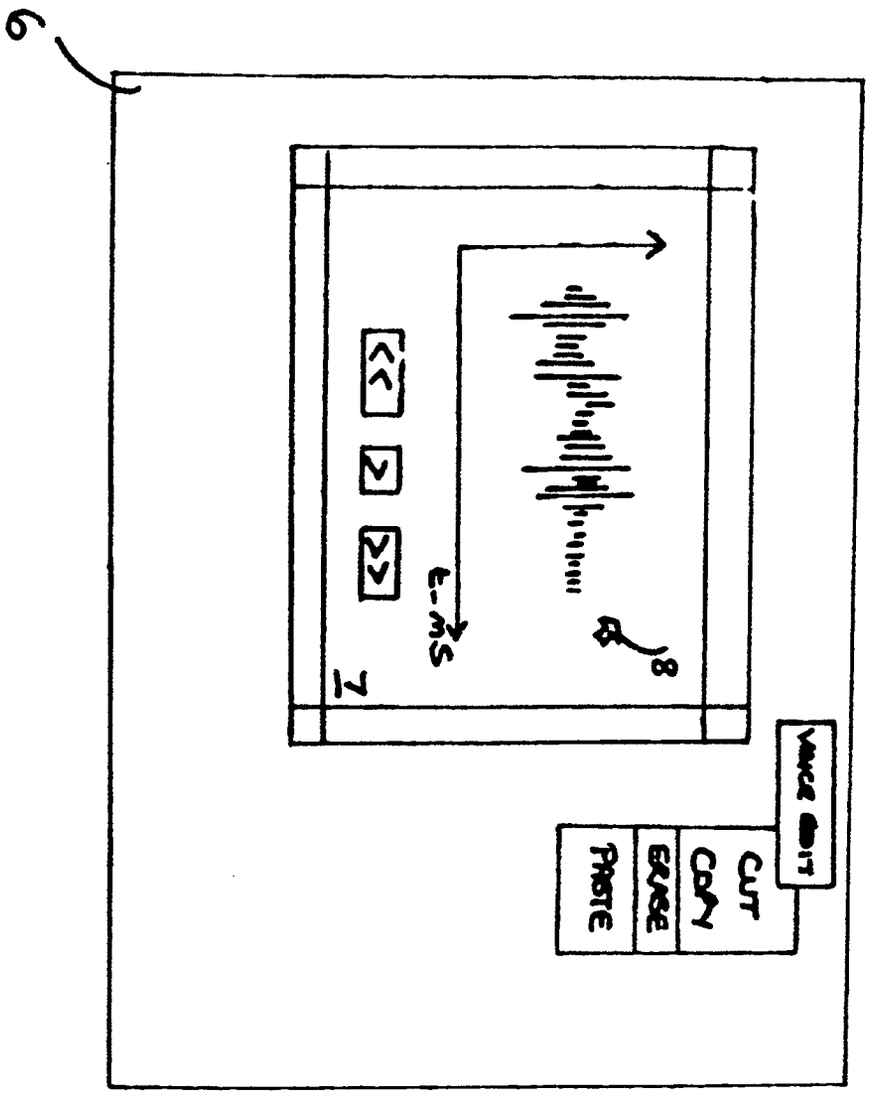

SPEECH SIGNAL
9
ADC
DIGITAL DATA
11
LPC
LPC PARAMETERS
12
TRANSFORM
p (f)
13
POLYNOMIAL EVALUATOR
QUANTISED ROOTS
19
δ1
δ2
FRAME COMPARATOR
QUANTISED ROOTS OR REPEAT CODE
14
STORE

Fig. 3A

Fig. 3B


17
18
20
TRANSFORM
LPC
PARAMETERS
LPC
DIGITAL
DATA
DAC
p(f)
AUDIO O/P
16
POLYNOMIAL
GENERATOR
QUANTISED ROOTS OR
FRAME REPEAT CODE
14
STORE


6
7
8
t-3s
CUT
COPY
ERASE
PASTE