

(19)



Europäisches Patentamt
European Patent Office
Office européen des brevets



(11) Publication number:

0 527 529 A2

(12)

EUROPEAN PATENT APPLICATION

(21) Application number: **92202374.2**

(51) Int. Cl.⁵: **G10L 3/02**

(22) Date of filing: **31.07.92**

(30) Priority: **09.08.91 EP 91202044**
24.02.92 EP 92200521

(43) Date of publication of application:
17.02.93 Bulletin 93/07

(84) Designated Contracting States:
DE FR GB IT

(71) Applicant: **N.V. Philips' Gloeilampenfabrieken**
Groenewoudseweg 1
NL-5621 BA Eindhoven(NL)

(72) Inventor: **Vogten, Leonardus Lambertus Maria**
c/o INT. OCTROOIBUREAU B.V. Prof.
Holstlaan 6
NL-5656 AA Eindhoven(NL)
Inventor: **Ma, Chang Xue**
c/o INT. OCTROOIBUREAU B.V. Prof.
Holstlaan 6
NL-5656 AA Eindhoven(NL)
Inventor: **Verhelst, Werner Desiré Elisabeth**
c/o INT. OCTROOIBUREAU B.V. Prof.
Holstlaan 6
NL-5656 AA Eindhoven(NL)
Inventor: **Eggen, Josephus Hubertus**
c/o INT. OCTROOIBUREAU B.V. Prof.
Holstlaan 6
NL-5656 AA Eindhoven(NL)

(74) Representative: **Strijland, Wilfred et al**
INTERNATIONAAL OCTROOIBUREAU B.V.
Prof. Holstlaan 6
NL-5656 AA Eindhoven (NL)

(54) **Method and apparatus for manipulating duration of a physical audio signal, and a storage medium containing a representation of such physical audio signal.**

(57) A method is described for manipulating an audio equivalent signal through positioning of mutually overlapping time windows. These are displaced with respect to each other over an actual pitch period length, and multiplied by a window function. Systematic repeating, suppressing, and maintaining of segment signals will accommodate the signal to an intended overall length.

EP 0 527 529 A2

FIELD OF THE INVENTION

The invention relates to a method for manipulating an audio equivalent signal, comprising positioning of a chain of mutually overlapping time windows with respect to the audio equivalent signal, as based on periodicity measurements on said audio equivalent signal, and wherein a positional displacement between adjacent windows substantially corresponds to a principal period of said periodicity, and synthesizing an audio output signal by chained superposition of segment signals, each deriving from the audio equivalent signal through weighting with the associated window function. Such method has been described in earlier non pre-published European Application 91202044.3 (PHN 13.801) co-authored by the present inventors and assigned to the same assignee, herein incorporated by reference, insofar as not actually included. As an alternative of the art recited, a still earlier, prepublished technology can be used, as per European Patent Application EP-A-363 233. The object of the references is to change prosody for synthesized speech, the pitch of the output signal, and the duration of stretches of speech. In combination with the junior reference, the method should be performed automatically, be robust against noise, and retain a high audio quality for the output signal.

The inventors of the present invention have realized that the manipulation of the duration can be used in various situations where there are external constraints to the total length of a self-contained unit of speech, which constraints may specify both the maximum and the minimum duration of such unit.

SUMMARY TO THE INVENTION

Accordingly, amongst other things, it is an object of the present invention to position the manipulated audio equivalent signal in a predetermined time length that differs from the original length, while on the one hand filling the interval more or less completely, and on the other hand keeping the impression of the eventual representation as natural as possible.

Now, according to one of its aspects, the object is realized in that the invention is characterized by manipulating a duration of said output signal through systematically repeating, maintaining, and/or suppressing said segment signals, to a resulting predetermined overall length that differs from a corresponding duration of said audio equivalent signal.

An advantage of the method of positioning windows according to the junior reference is that it can be machine-executed without any window-to-win-

dow human control being necessary. Furthermore, it has been found that the duration can be changed by a factor between 2 and $\frac{1}{2}$ without seriously impairing understandability of speech. For lesser degrees of manipulating the duration, such as by + or - 30%, not only remains the understandability very good, but also the natural quality of speech is maintained, and a listener would hardly feel the change of duration as unnatural. A prerequisite to applying the method is that the pitch can indeed be measured, which for human speech is a problem knowing various solutions. Situations where the duration of speech should be manipulated are various, such as in post-synchronizing of movies or other video representative material, adapting a speech explanation or other matter to physical motion of objects, such as the closing instant of a door, and many other instances. In movies, actor utterances should preferably coincide with their facial motions, or at least with their moving around in general. Typical time scales of the total duration of the utterance are 0.3 to several seconds. In this short time frame, prior art had not succeeded in duration manipulation with also preserving naturalness. On a much longer time scale, the length of a pause can be manipulated, such as is often done by human interpreters. If the available time is known beforehand, sometimes a different verbalization can be used, but all these methods require specialized human skills. The present method is easily applicable and just requires the setting of a speed-up or slow-down percentage. Of course, the use of the present invention is also for amending longer durations than in the seconds range.

The invention relates also to an apparatus for executing the method and to a storage medium containing a representation of audio signal equivalent. The invention allows to fill the available space for a unit of speech (sentence, partial sentence, exclamation, or other) well high completely.

A particular application is Compact Disc Interactive, especially so in a multi-language environment. Editing CD-I is by itself a complicated task. Sizing the duration of speech utterances may now be performed by the machine for relieving the program editor from this tedium. By itself, CD-I is a well-published storage medium with associated development platform, the storage itself being an extension from Compact Disc Audio.

Various advantageous aspects of the invention are recited in dependent Claims.

BRIEF DESCRIPTION OF THE FIGURE

These and other advantages will be described with reference to a preferred embodiment which is shown in a number of Figures, of which

Figure 1 shows editing of CDI-program for storage on a CD-I-disc;

The following Figures especially show the technology of the junior reference:

Figures 2a,b,c show speech signals with windows placed according to the invention;

Figure 3 shows an apparatus for changing the pitch and/or duration of a signal;

Figure 4 shows multiplication means and window function value selection means for use in an apparatus for changing the pitch and/or duration of a signal;

Figure 5 shows window position selection means for implementing the invention;

Figure 6 shows a subsystem for combining several segment signals.

DESCRIPTION OF A PREFERRED EMBODIMENT

As is commonly understood, the audio or speech equivalent signal may be direct analog speech, or it may be speech that is stored as a sequence of codes for on the basis thereon generating synthetic speech. The length of the various windows may be non-uniform, and in a particular embodiment, the length of each window may be substantially equal to a local actual pitch period length. Within the window, the window function is uniform, which means that the window function scales linearly with the width of the window, which means that generally, there may be an appreciable variation between the widths of successive windows. The systematical character of the repeating, maintaining, or suppression implies that there is a certain prescription for the sequence of window positions that, first, restricts to either repeating or suppressing, either possibly in combination with maintaining, and furthermore, that the repeating or maintaining is done under control of an actual or emulated recurrent cycle. Examples are:

each third window is repeated once, the others are maintained;

of each five successive windows, #2 and #4 are suppressed;

at each next window, a count is incremented by a particular amount and overflow controls actual suppression or repetition.

It is commented that the systematical character would not need to be completely uniform. For example, in post-synchronization of a movie, it could be advantageous to amend the time durations of various parts of a sentence somewhat differently from each other, as long as the natural character of the resulting speech would remain. In particular the movement of a face while speaking speech could to a certain extent be followed by the dynamics of the audio speech. Also, different sentences in various places of the post-synchronizing

now may have uniform pitch among each other.

The different representations in parallel may be different languages; it has been found that the same sentence, translated to another language, would have different length, counted for example, as a number of syllables: in particular, the German language caused a longer duration as compared with English and French.

Other, in particular exotic, languages may lead to even more extreme situations. Similar situations may distinguish child voices from adult voices.

In the Figure for a three language CD-I track the pictorial material 200 is shown with accompanying speech representations in French (202), German (204) and English (206) before editing. It is intended to lend each language representation (among which a user may choose) exactly the same duration as the pictorial material (movie, animation, etcetera). As shown, on line 202, a single window is suppressed, on line 204, five windows are suppressed. On line 206, six windows are repeated one (crosses). The result after editing is not shown. It has been found that analysis of the results can prove infringement. Especially the occurrence of the repeated windows is well traceable. Moreover, the substantially equal lengths of the various representations is, together with the high subjective quality of the rendering is a clear indication for the use of the present technology.

In certain situations, apart from changing the duration per se, the slowing down or speeding may lend the speech a character, such as nervous (fast) or majestic (slow). Also such use is sometimes advantageous. Changing the duration of the audio equivalent signal may be combined with changing the pitch. The two types of manipulation may be both in the same direction, for example in that both effectively shorten the duration. In other circumstances, they could to some degree compensate the effects, so that the change in duration would be less or even be zero. The change of duration may be according to a time-varying pattern, whereby the overall change of duration is the integral or sum of the elementary changes-of-duration.

DESCRIPTION OF A PREFERRED TECHNOLOGY

Hereinafter, a description of the preferred technology according to the junior reference is given.

Figures 2a, 2b and 2c show speech signals with marks 52 placed apart by distances determined with a pitch meter (that may be conventional), that is, without a fixed phase reference. In Figure 2a, two successive periods were marked as voiceless by placing their pitch period length indication outside the scale. The pitch marks (lower scale) were obtained by interpolating the period length. Although the pitch period lengths were de-

terminated without smoothing other than that inherent in determining spectra of the speech signal extending over several pitch periods, a very regular curve was obtained automatically.

The incremental placement of windows also solves another problem. For unvoiced stretches, that contain fricatives like the sound "ssss", in which the vocal cords are not excited, the windows are placed incrementally just like for voiced stretches. The pitch period length is interpolated between the lengths measured for unvoiced stretches adjacent to the voiced stretch. This provides regularly spaced windows without audible artefacts.

The placement of windows is easy if the input audio equivalent signal is monotonous. In this case, the windows may be placed simply at fixed distances from each other. This may be effected by preprocessing the signal, so as to change its pitch to a single monotonous value. The final manipulation to obtain a desired pitch and/or duration starting can then be performed with windows at uniform spacing.

An exemplary apparatus.

Figure 3 shows an exemplary embodiment of an apparatus for changing the pitch and/or duration of an audible signal. The input audio equivalent signal arrives at an input 60, and the output signal leaves at an output 63. The input signal is multiplied by the window function in multiplication means 61, and stored segment signal by segment signal in segment slots in storage means 62. To synthesize the output signal on output 63, speech samples from various segment signals are summed in summing means 64. The manipulation of speech signals, in terms of pitch change and/or duration manipulation, is effected by addressing the storage means 62 and selecting window function values. Accordingly, selection of storage addresses for storing the segments is controlled by window position selection means 65, which also control window function value selection means 69; selection of readout addresses is controlled by combination means 66.

In order to explain the operation of the components of the apparatus shown in Figure 3 it will be briefly recalled that signal segments S are to be derived from the input signal X (at 60), the segments being defined by

$$S_i(t) = W(t/L_i) X(t + t_i) \quad (-L_i < t < 0)$$

$$S_i(t) = W(t/L_{i+1}) X(t + t_i) \quad (0 < t < L_{i+1})$$

and these segments are to be superposed to produce the output signal Y (at 63):

$$Y(t) = \sum_i S_i(t - T_i)$$

(The sum being limited to indices i for which $-L_i < t - T_i < L_{i+1}$).

At any point in time t' a signal $X(t')$ is supplied at the input 60, which contributes to two segments i , $i + 1$ at respective t values $t_a = t' - t_i$ and $t_b = t' - t_{i+1}$ - (these being the only possibilities that $-L_i < t < L_{i+1}$).

Figure 4 shows the multiplication means 61 and the window function value selection means 69. The respective t values t_a , t_b described above are multiplied by the inverse of the period length L_i - (determined from the period length in an inverter 74) in scaling multipliers 70a, 70b to determine the corresponding arguments of the window function W . These arguments are supplied to window function evaluators 71a, 71b (implemented for example in case of discrete arguments as a lookup table) which outputs the corresponding values of the window function, which are multiplied with the input signal in two multipliers 72a, 72b. This produces the segment signal values S_i , S_{i+1} at two inputs 73a, 73b to the storage means 62.

These segment signal values are stored in the storage means 62 in segment slots at addresses in the slots corresponding to their respective time point values t_a , t_b and to respective slot numbers. These addresses are controlled by window position selection means 65. Window position selection means suitable for implementing the invention are shown in Figure 5. The time point values t_a , t_b are addressed by counters 81, 82, the segment slots numbers are addressed by indexing means 84, (which output the segment indices i , $i + 1$). The counters 81, 82 and the indexing means 84 output addresses with a width as appropriate to distinguish the various positions within the slots and the various slot respectively, but are shown symbolically only as single lines in Figure 5.

The two counters 81, 82 are clocked at a fixed clock rate and count from an initial value loaded from a load input (L), upon a trigger signal at trigger input (T). The indexing means 84 increment the index values upon reception of this trigger signal. According to one embodiment, pitch measuring means 86 determine a pitch value from input 60, and control the scale factor for the scaling multipliers 70a, 70b, and provide the initial value of the first counter 81 (the initial count being minus the pitch value), whereas the trigger signal is generated internally in the window position selection means, once the counter reaches zero, as detected by a comparator 88. This means that successive windows are placed by incrementing the location of a previous window by the time needed by the first counter 81 to reach zero.

In another embodiment, a monotonized signal is applied to the input 60 (this monotonized signal

being obtained by prior processing in which the pitch is adjusted to a time independent value). In this monotonized case, a constant value, corresponding to the monotonized pitch is fed as initial value to the first counter 81. In this case the scaling multipliers 70a, 70b can be omitted since the windows have a fixed size.

The combination means 66 of Figure 3 are shown in Figure 10. The purpose of the output side is to superpose segments from the storage means 62 according to

$$Y(t) = \sum_i S_i(t-T_i)$$

The sum being limited to index values i for which $-L_i < t - T_i < L_{i+1}$;

in principle, any number of index values may contribute to the sum at one time point t. But when the pitch is not changed by more than a factor of 3/2, at most 3 index values will contribute at a time. By way of example, therefore, Figures 3 and 7 show an apparatus which provides for only three active indices at a time; extension to more than three segments is straightforward.

For addressing the segments, the combination means 66 are quite similar to the input side: they comprise three counters 101, 102, 103 (clocked at a fixed rate), outputting the time point values $t-T_i$ for the three segment signals. The three counters receive the same trigger signal, which triggers loading of minus the desired output pitch interval in the first of the three counters 101. Upon the trigger signal the last position of the first counter 101 is loaded into the second counter 102, and in the third counter 103 the last position of the second counter 102 is loaded. The trigger signal is generated by a comparator 104, which detects zero crossing of the first counter 101. The trigger signal also updates indexing means 106.

The indexing means address the segment slot numbers which must be read out and the counters address the position within the slots. The counters and indexing means address three segments, which are output from the storage means 62 to the summing means 64 in order to produce the output signal.

By applying desired pitch interval values at the pitch control input 68a, one can thus control the pitch value. The duration of the speech signal is controlled by a duration control input 68b to the indexing means. Without duration manipulation, the indexing means simply produce three successive segment slot numbers. At the trigger signal, the values of the first and second output are copied to the second and third output respectively, and the first output is increased by one. When the duration is increased, the first output is kept constant once every so many cycles, as determined by the dura-

tion control input 68b. To decrease the duration, the first output is increased by two every so many cycles. The change in duration is determined by the net number of skipped or repeated indices. When the apparatus is used to change the pitch and duration of a signal independently (for example changing the pitch and keeping the duration constant), the duration input 68b should be controlled to give a net frequency F at which indices should be skipped or repeated according to

$$F = (Dt/T) - 1$$

(D being the factor by which the duration is changed, t being the pitch period length of the input signal and T being the period length of the output signal; a negative value of F corresponds to skipping of indices, a positive value corresponds to repetition).

Figure 3 only provides one embodiment by way of example. The principal point is the incremental placement of windows at the input side with a phase determined from the phase of a previous window. There are many ways of generating the addresses for the storage means 62, of which Figure 5 is but one. For example, the addresses may be generated using a computer program, and the starting addresses need not have the values given in the example.

Figure 3 can be implemented in various ways, for example using digital samples at input 60, where the sampling rate has at any convenient value, for example 10000 samples per second; conversely, it may use continuous signal techniques, where the clocks 81, 82, 101, 102, 103 provide continuous ramp signals, and the storage means provide for continuously controlled access like a magnetic disk. Furthermore, in Figure 3 in practice segment slots may be reused after some time, as they are not needed permanently. Not all components of Figure 4 need to be implemented by discrete function blocks: often it may be implemented in whole or part by a computer.

Claims

1. A method for manipulating an audio equivalent signal, comprising positioning of a chain of mutually overlapping time windows with respect to the audio equivalent signal, as based on periodicity measurements on said audio equivalent signal, and wherein a positional displacement between adjacent windows substantially corresponds to a principal period of said periodicity, and synthesizing an audio output signal by chained superposition of segment signals, each deriving from the audio equivalent signal through weighting with the asso-

ciated window function, characterized by manipulating a duration of said output signal through systematically repeating, maintaining, and/or suppressing said segment signals, to a resulting predetermined overall length that differs from a corresponding duration of said audio equivalent signal. 5

2. A method as claimed in Claim 1, wherein said predetermined length applies to a plurality of speech equivalent signals in parallel that correspond in content but have differences in representation. 10
3. A method as claimed in Claim 2 wherein said differences originate from said plurality of audio equivalent signals being in as many different languages. 15
4. A method as claimed in Claim 1, 2 or 3, wherein said predetermined length pertains to an intermission between non-manipulated audio equivalent signals. 20
5. A method as claimed in any of Claims 1 to 4 for post-synchronizing human speech as featured by a video representable item. 25
6. A method as claimed in any of Claims 1 to 5, wherein said windows are shifted with respect to a previous window in said sequence over an actual pitch period length of said audio equivalent signal. 30
7. A unitary storage medium comprising audio equivalent signals generated according to the method as claimed in any of Claims 1 to 6. 35
8. An apparatus for manipulating an audio equivalent signal according to a method as claimed in any of Claims 1 to 6. 40

45

50

55

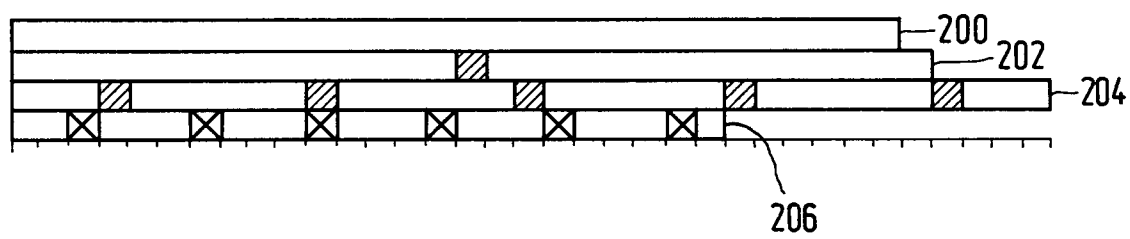


FIG. 1

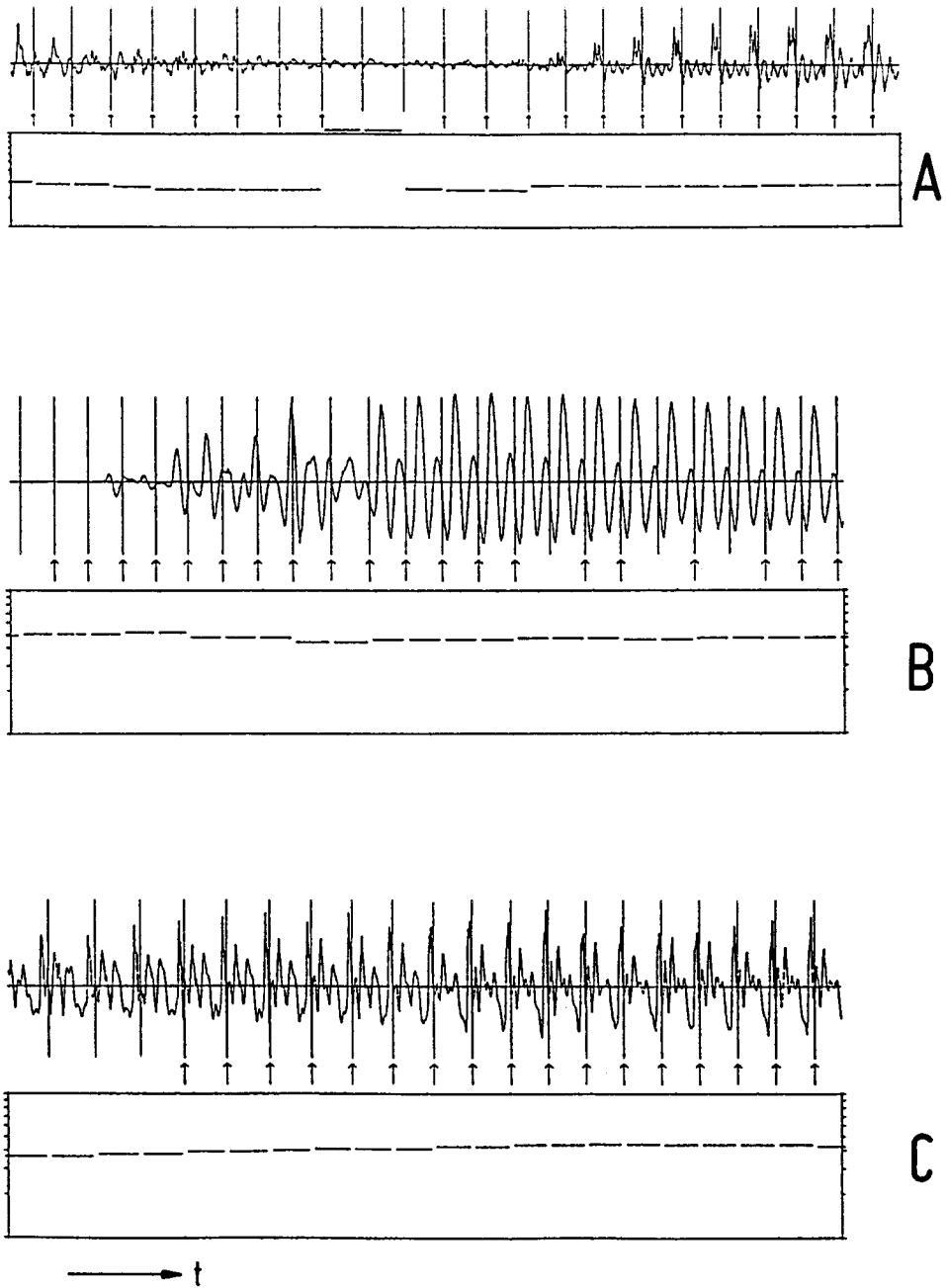


FIG.2

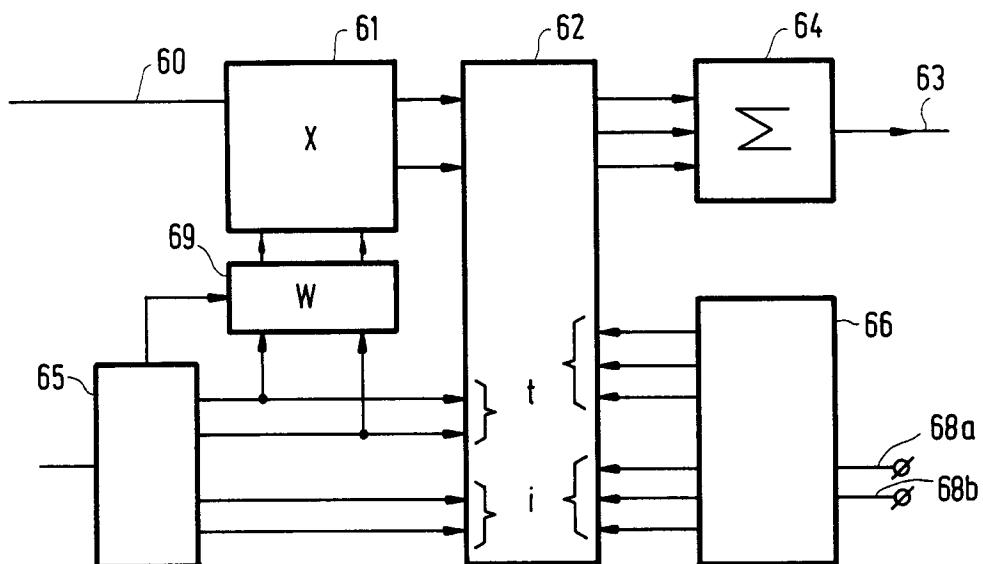


FIG.3

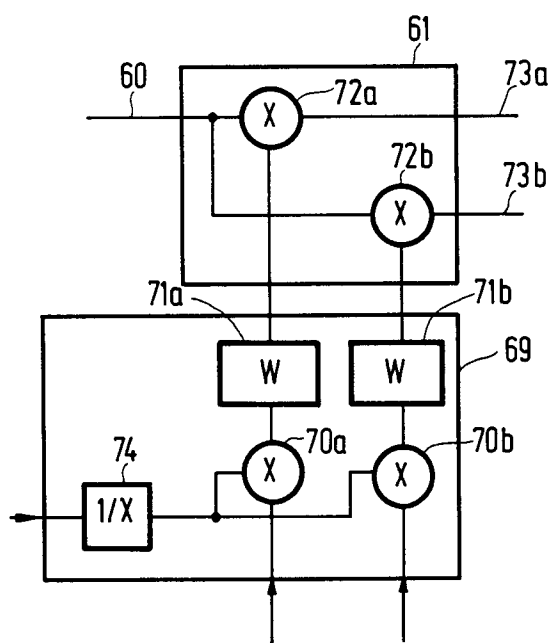


FIG.4

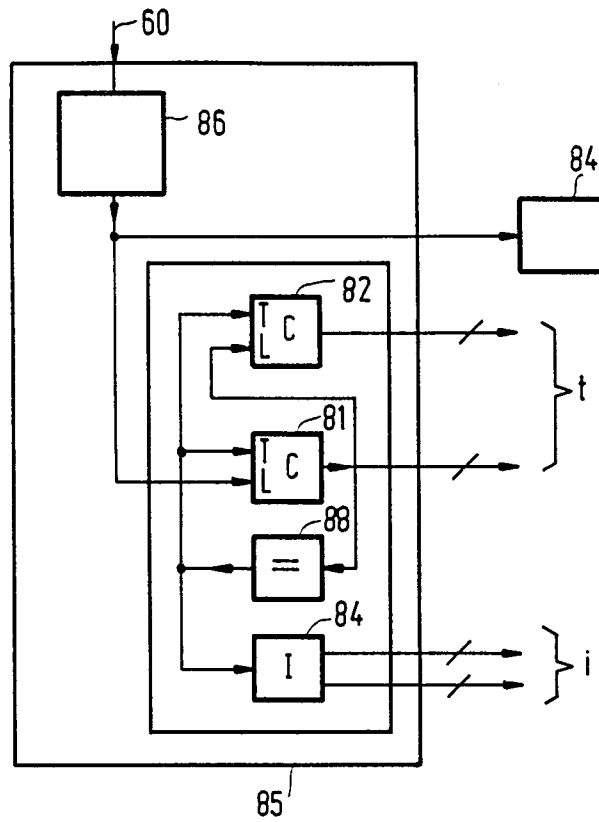


FIG. 5

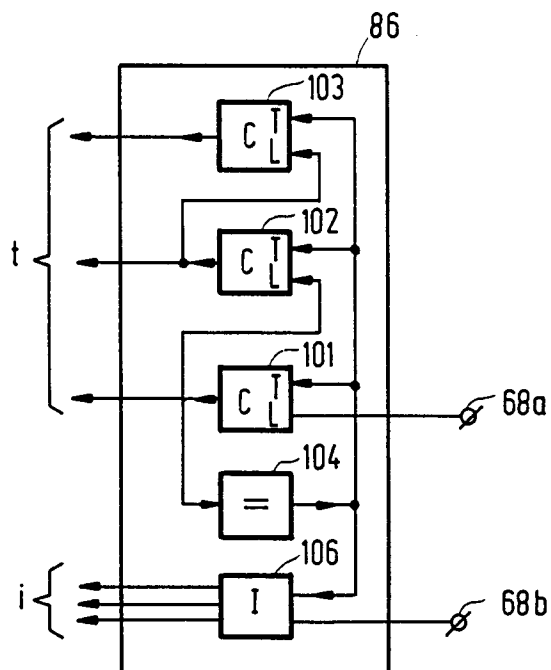


FIG. 6