US 20070083377A1

(19) **United States**

(12) **Patent Application Publication**

Trautmann et al.

(10) **Pub. No.: US 2007/0083377 A1**

(43) **Pub. Date: Apr. 12, 2007**

(54) **TIME SCALE MODIFICATION OF AUDIO USING BARK BANDS**

(76) Inventors: **Steven Trautmann**, Tsukuba (JP); **Atsuhiro Sakurai**, Tsukuba-shi (JP); **Daniel L. Zelazo**, Seattle, WA (US)

Correspondence Address:
TEXAS INSTRUMENTS INCORPORATED
P O BOX 655474, M/S 3999
DALLAS, TX 75265

(21) Appl. No.: **11/248,132**

(22) Filed: **Oct. 12, 2005**

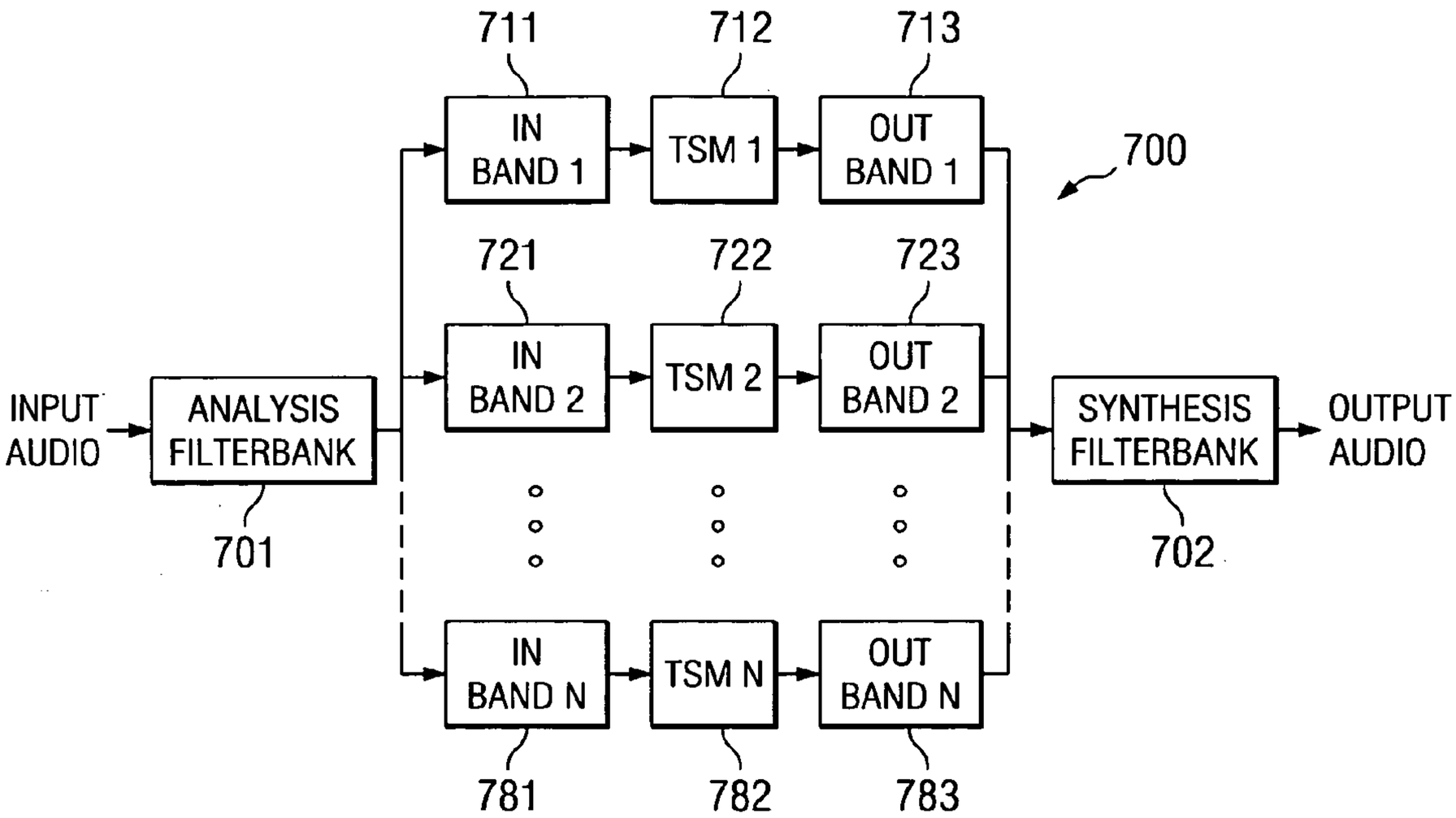
Publication Classification

(51) **Int. Cl.**
G10L 21/04 (2006.01)

(52) **U.S. Cl.** **704/503**

(57) **ABSTRACT**

This invention involves time-scale modification of audio signals. In this invention the input audio signal is separated into a plurality of frequency bands selected according to a Bark scale where each frequency band has an extent dependent upon human frequency perception via a filter bank. Time-scale modification is applied separately to the individual frequency bands. The thus modified signals are recombined for output.



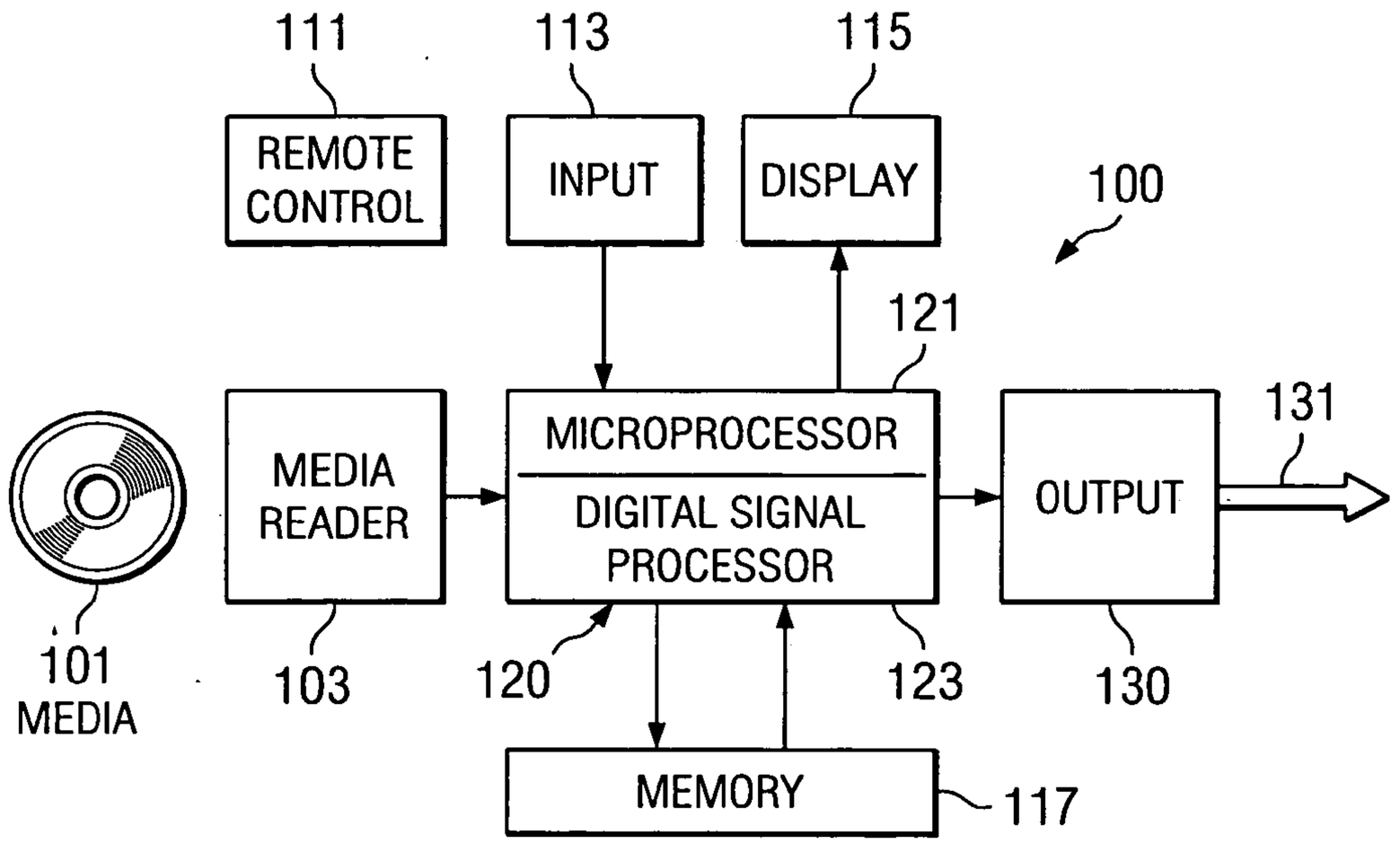


FIG. 1

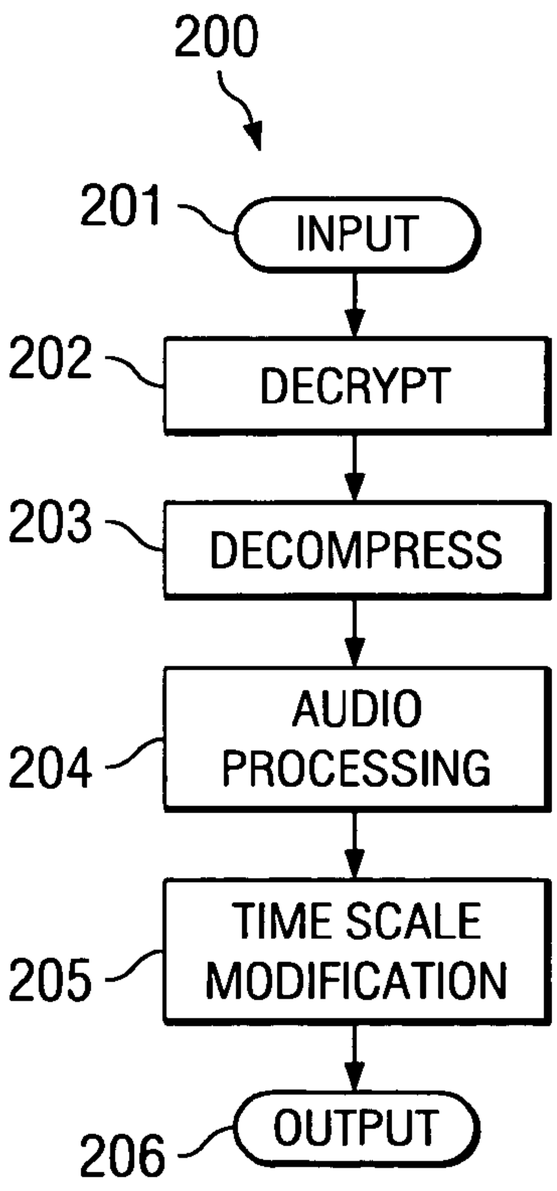


FIG. 2

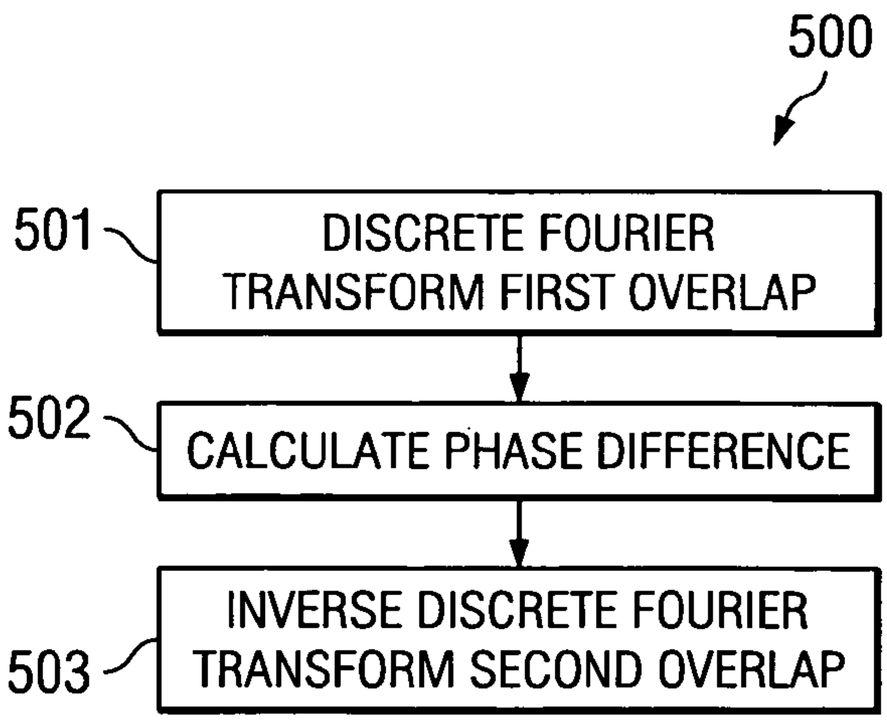


FIG. 5
(PRIOR ART)

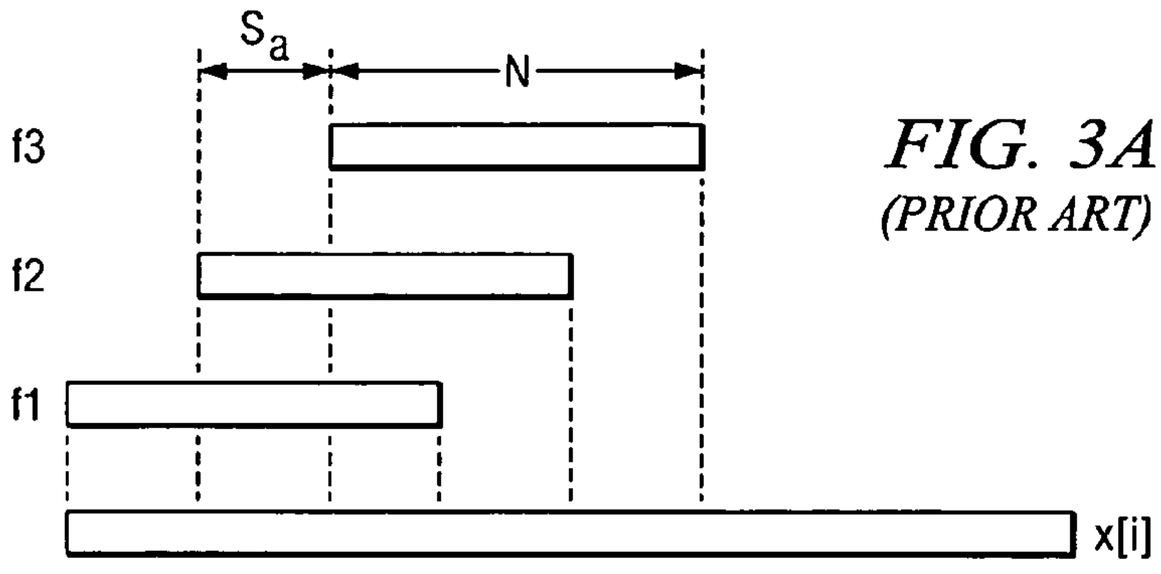


FIG. 3A
(PRIOR ART)

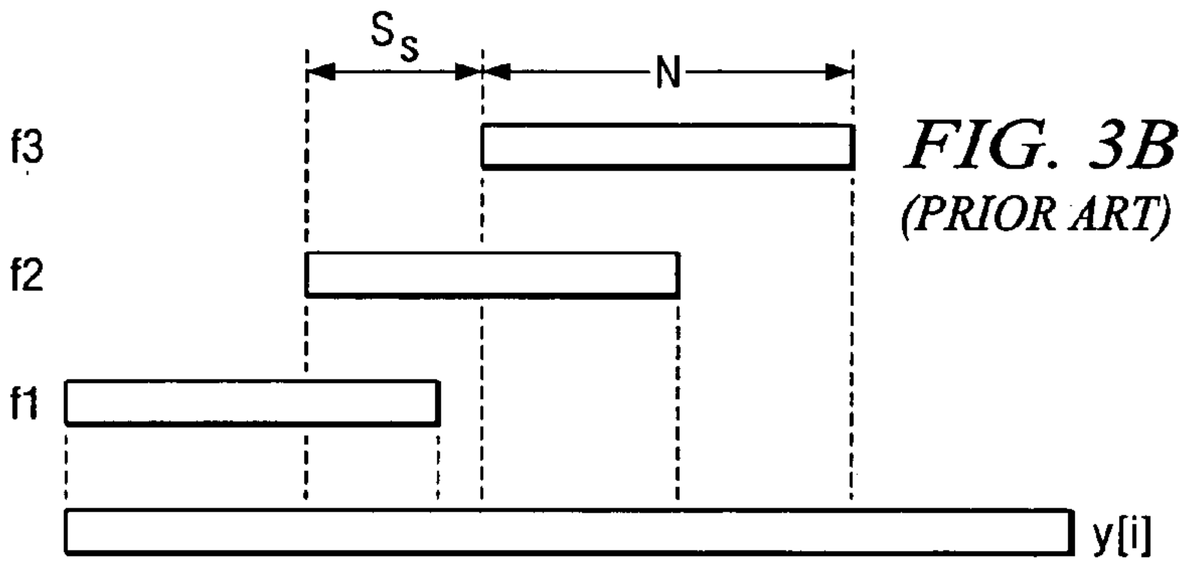


FIG. 3B
(PRIOR ART)

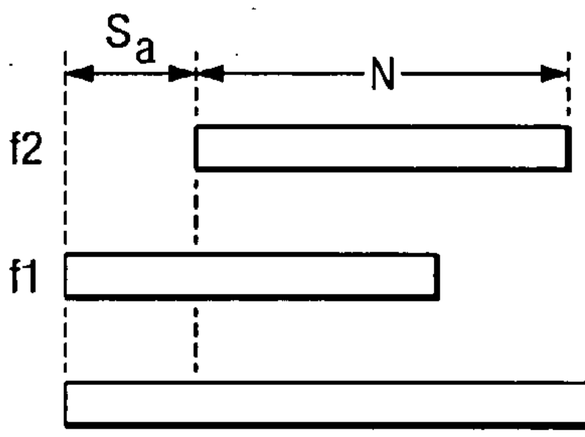


FIG. 4A
(PRIOR ART)

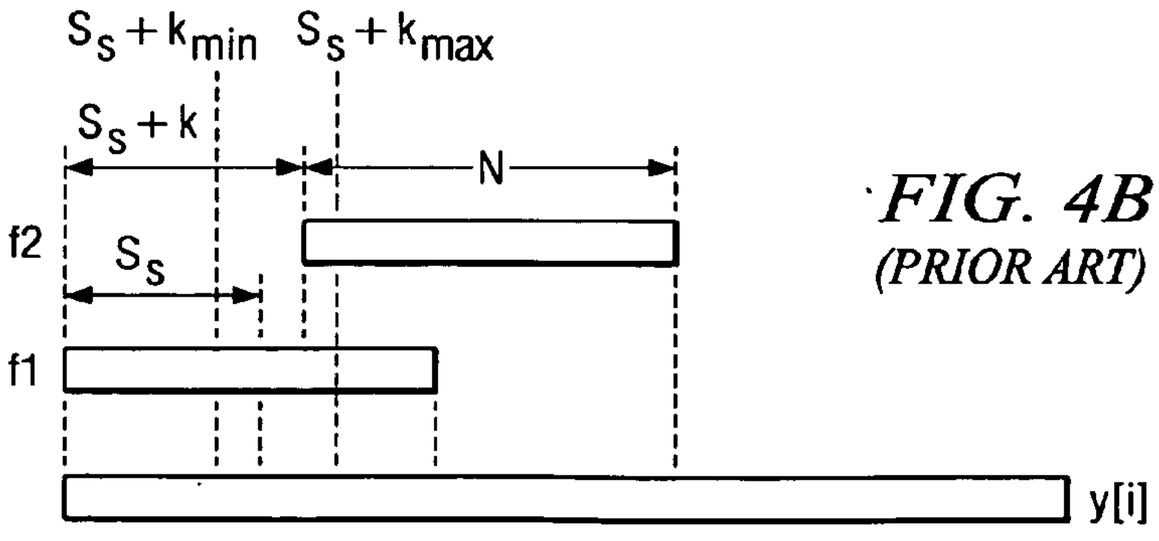


FIG. 4B
(PRIOR ART)

FIG. 6

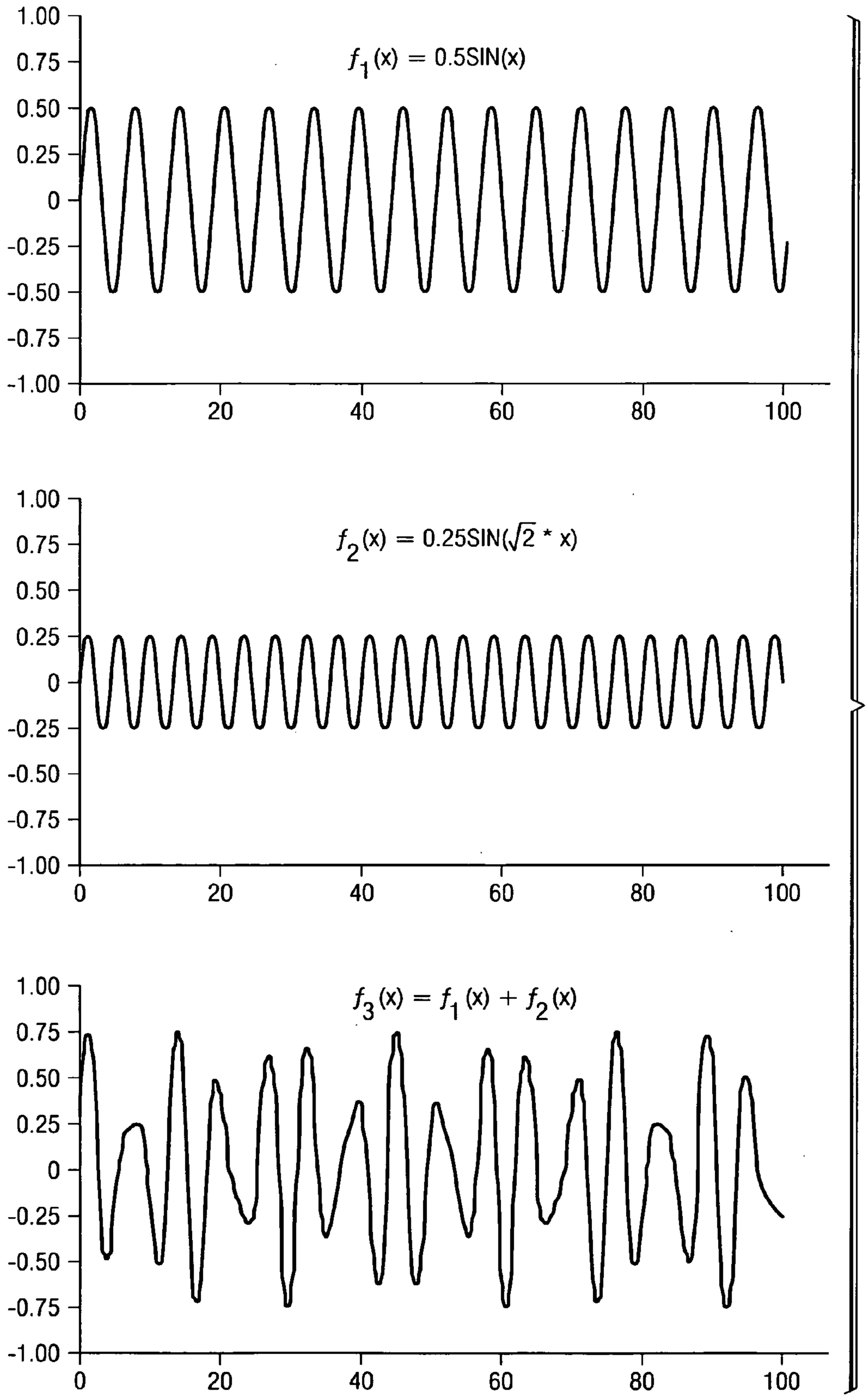
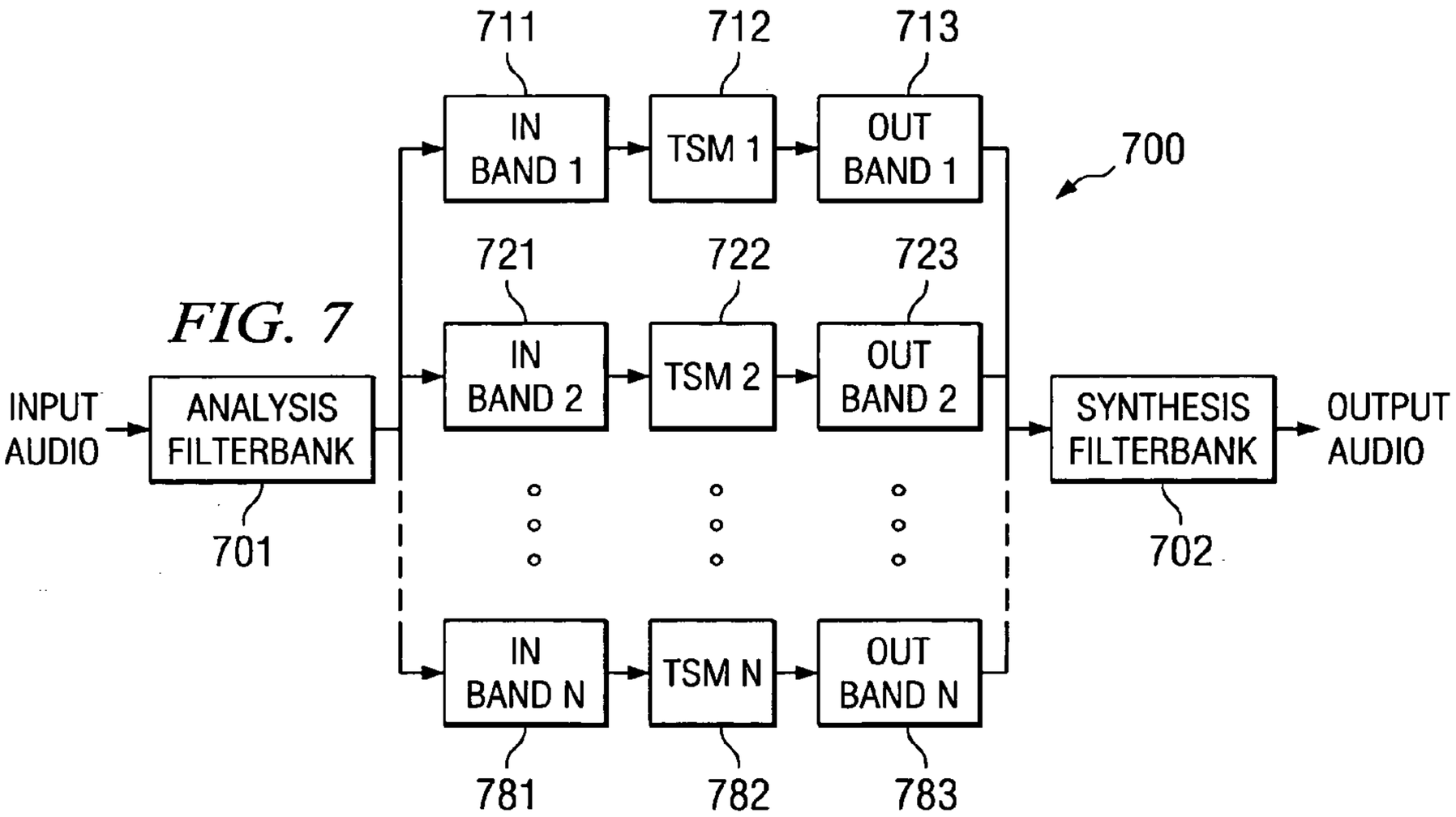

$f_1(x) = 0.5\text{SIN}(x)$
$f_2(x) = 0.25\text{SIN}(\sqrt{2} * x)$
$f_3(x) = f_1(x) + f_2(x)$
1.00
0.75
0.50
0.25
0
-0.25
-0.50
-0.75
-1.00
0
20
40
60
80
100

FIG. 7


711
712
713
IN BAND 1
TSM 1
OUT BAND 1
700
721
722
723
IN BAND 2
TSM 2
OUT BAND 2
INPUT AUDIO
ANALYSIS FILTERBANK
701
SYNTHESIS FILTERBANK
OUTPUT AUDIO
702
IN BAND N
TSM N
OUT BAND N
781
782
783

TIME SCALE MODIFICATION OF AUDIO USING BARK BANDS

TECHNICAL FIELD OF THE INVENTION

[0001] The technical field of this invention is digital audio time scale modification.

BACKGROUND OF THE INVENTION

[0002] Time-scale modification (TSM) is an emerging topic in audio digital signal processing due to the advance of low-cost, high-speed hardware that enables real-time processing by portable devices. Possible applications include intelligible sound in fast-forward play, real-time music manipulation, foreign language training, etc. Most time scale modification algorithms can be classified as either frequency-domain time scale modification or time-domain time scale modification. Frequency-domain time scale modification provides higher quality for polyphonic sounds, while time-domain time scale modification is more suitable for narrow-band signals such as voice. Time-domain time scale modification is the natural choice in resource-limited applications due to its lower computational cost.

[0003] The basic operation of time domain time-scale modification is successively overlapping and adding audio frames, where time scaling is achieved by changing the spacing between them. It is known in the art to calculate the exact overlap point based on a measure of similarity between the signals to be overlapped. This measure of similarity is generally based on cross-correlation.

[0004] Most time-domain time-scale modification algorithms are derived from the synchronous overlap-and-add method (SOLA). The synchronous overlap-and-add algorithm and its variations are based on successive overlap and addition of audio frames. For the overlap, the overlap point is adjusted by computing a measure of signal similarity between the overlapping regions for each possible overlap position, which is limited by a minimum and maximum overlap points. The position of maximum similarity is selected. The signal similarity measure can be represented as a full cross-correlation function or simplified versions. This similarity calculation represents about 80% or more of the total computation required by the algorithm.

[0005] Even though SOLA based methods represent an attractive low-cost solution to the time-scale modification problem, their limitation stands out in the case of polyphonic music signals. Their intrinsic problem is that the audio signal is treated as a whole without consideration for its individual frequency components, so that the overlap point adjustment based on signal similarity cannot simultaneously generate smooth transitions for the multiple frequency components of the signal.

[0006] A family of methods known as phase vocoder does time-scale modification in the frequency domain. The input signal is analyzed at equally spaced overlapping windowed frames using a short-time discrete Fourier transform. Next the phase difference for spectral peaks is calculated. This phase difference is the difference in phase between an input phase and a time scale modified signal phase. An intrinsic sinusoidal model is generally used. The frequency is represented by the sum $\Omega_k+\omega_{ik}$: where carrier $\mathbf{106}_k$ is $2\pi k/N$; and ω_{ik} is an instantaneous frequency modulator. This produces an estimate ω_{ik} for each spectral line by obtaining the phase difference between two consecutive analysis frames. Here, k is the spectral line and N is the size of the short-time discrete Fourier transform. The process reconstructs an output signal from the analyzed frames using a short-time inverse discrete Fourier transform. The frames are overlapped by a different overlap factor to achieve the desired time scaling. The instantaneous frequency ω_{ik} is used to calculate the phase corresponding to each spectral line in the time shifted instant.

[0007] Even though phase vocoders can potentially achieve higher quality than time-domain methods, a severe limitation is the large amount of computation required in the forward and inverse discrete Fourier transforms and also in the spectrum manipulation process. Practical implementations on fixed-point processors result in a computational cost up to 10 times higher than time-domain time-scale modification methods. In addition, maintaining phase coherence between frames is not an easy task and can be the source of artifacts.

SUMMARY OF THE INVENTION

[0008] This invention involves time-scale modification of audio signals. In this invention the input audio signal is separated into a plurality of frequency bands selected according to a Bark scale where each frequency band has an extent dependent upon human frequency perception via a filter bank. The human auditory system has a limited, frequency-dependent resolution. The perceptually uniform measure of frequency can be expressed in terms of the width of the critical or Bark bands. It is less than 100 Hz at the lowest audible frequencies, and more than 4 kHz at the high end. The complete audio frequency range can be partitioned into 25 critical, or Bark bands. Time-scale modification is applied separately to these individual frequency bands. The thus modified signals are then recombined for output.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] These and other aspects of this invention are illustrated in the drawings, in which:

[0010] FIG. **1** is a block diagram of a digital audio system to which this invention is applicable;

[0011] FIG. **2** is a flow chart illustrating the data processing operations involved in time-scale modification employing the digital audio system of FIG. **1**;

[0012] FIG. **3***a* illustrates the analysis step in the overlap and add method of time scale modification according to the prior art;

[0013] FIG. **3***b* illustrates the synthesis step in the overlap and add method of time-scale modification according to the prior art;

[0014] FIG. **4***a* illustrates the analysis step in synchronous overlap and add method of time scale modification according to the prior art;

[0015] FIG. **4***b* illustrates the synthesis step in the synchronous overlap and add method of time-scale modification according to the prior art;

[0016] FIG. **5** is a flow chart illustrating the steps in the prior art phase vocoder time scale modification technique;

[0017] FIG. **6** is a view of several waveforms used in explanation of this invention; and

[0018] FIG. **7** is a process diagram illustrating the processes of this invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0019] FIG. **1** is a block diagram illustrating a system to which this invention is applicable. The preferred embodiment is a DVD player or DVD player/recorder in which the time scale modification of this invention is employed with fast forward or slow motion video to provide audio synchronized with the video in these modes.

[0020] System **100** received digital audio data on media **101** via media reader **103**. In the preferred embodiment media **101** is a DVD optical disk and media reader **103** is the corresponding disk reader. It is feasible to apply this technique to other media and corresponding reader such as audio CDs, removable magnetic disks (i.e. floppy disk), memory cards or similar devices. Media reader **103** delivers digital data corresponding to the desired audio to processor **120**.

[0021] Processor **120** performs data processing operations required of system **100** including the time scale modification of this invention. Processor **120** may include two different processors, microprocessor **121** and digital signal processor **123**. Microprocessor **121** is preferably employed for control functions such as data movement, responding to user input and generating user output. Digital signal processor **123** is preferably employed in data filtering and manipulation functions such as the time scale modification of this invention. A Texas Instruments digital signal processor from the TMS320C5000 family is suitable for this invention.

[0022] Processor **120** is connected to several peripheral devices. Processor **12**Q receives user inputs via input device **113**. Input device **113** can be a keypad device, a set of push buttons or a receiver for input signals from remote control **111**. Input device **113** receives user inputs which control the operation of system **100**. Processor **120** produces outputs via display **115**. Display **115** may be a set of LCD (liquid crystal display) or LED (light emitting diode) indicators or an LCD display screen. Display **115** provides user feedback regarding the current operating condition of system **100** and may also be used to produce prompts for operator inputs. As an alternative for the case where system **100** is a DVD player or player/recorder connectable to a video display, system **100** may generate a display output using the attached video display. Memory **117** preferably stores programs for control of microprocessor **121** and digital signal processor **123**, constants needed during operation and intermediate data being manipulated. Memory **117** can take many forms such as read only memory, volatile read/write memory, nonvolatile read/write memory or magnetic memory such as fixed or removable disks. Output **130** produces an output **131** of system **100**. In the case of a DVD player or player/recorder, this output would be in the form of an audio/video signal such as a composite video signal, separate audio signals and video component signals and the like.

[0023] FIG. **2** is a flow chart illustrating process **200** including the major processing functions of system **100**. Flow chart **200** begins with data input at input block **201**. Data processing begins with an optional decryption function (block **202**) to decode encrypted data delivered from media **101**. Data encryption would typically be used for control of copying for theatrical movies delivered on DVD, for example. System **100** in conjunction with the data on media **101** determines if this is an authorized use and permits decryption if the use is authorized.

[0024] The next step is optional decompression (block **203**). Data is often delivered in a compressed format to save memory space and transmit bandwidth. There are several motion picture data compression techniques proposed by the Motion Picture Experts Group (MPEG). These video compression standards typically include audio compression standards such as MPEG Layer 3 commonly known as MP3. There are other audio compression standards. The result of decompression for the purposes of this invention is a sampled data signal corresponding to the desired audio. Audio CDs typically directly store the sampled audio data and thus require no decompression.

[0025] The next step is audio processing (block **204**). System **100** will typically include audio data processing other than the time scale modification of this invention. This might include band equalization filtering, conversion between the various surround sound formats and the like. This other audio processing is not relevant to this invention and will not be discussed further.

[0026] The next step is time scale modification (block **205**). This time scale modification is the subject of this invention and various techniques of the prior art and of this invention will be described below in conjunction with FIGS. **3** to **6**. Flow chart **200** ends with data output (block **206**).

[0027] FIG. **3** illustrates this process. In FIG. **3**(*a*), x(i) is the analysis signals represented as a sequence with index i. Similarly, FIG. **3**(*b*) illustrates synthesis signal y(i) having a sequence index i. The quantity N is the frame size. S_a is the analysis frame interval between consecutive frames f_j (where j=1, 2 . . .). Ss is the similar synthesis frame interval. The relationship between the analysis frame interval S_a and the synthesis frame interval S_s sets the time scale modification. The overlap-and-add time scale modification algorithm is simple and provides acceptable results for small time-scale factors. In general this method yields poor quality compared to other methods described below.

[0028] The synchronous overlap-and-add time scale modification algorithm is an improvement over the previous overlap-and-add approach. Instead of using a fixed overlap interval for synthesis, the overlap point is adjusted by computing the normalized cross-correlation between the overlapping regions for each possible overlap position within minimum and maximum deviation values. The overlap position of maximum cross-correlation is selected. The cross-correlation is calculated using the following formula, where L_k is the length of the overlapping window:

$$R[k] = \frac{\sum_{i=0}^{L_k-1} y[mS_s + k + i]x[mS_a + i]}{\left[\sum_{i=0}^{L_k-1} y^2[mS_s + k + i] \sum_{i=0}^{L_k-1} x^2[mS_a + i]\right]^{1/2}} \quad (1)$$

FIG. **4** illustrates the synchronous overlap-and-add time scale modification algorithm. The same variables are used in FIG. **4**(*a*) for analysis as FIG. **3**(*a*) and used in. FIG. **4**(*b*) for synthesis as in **3**(*b*). In FIG. **4**, k is the deviation of the

overlap position, with k limited to the range between k_{min} and k_{max}. Note that k=0 is equivalent to the overlap-and-add time scale modification algorithm illustrated in FIGS. **3** (*a*) and **3** (*b*). The synchronous overlap-and-add time scale modification algorithm requires a large amount of computation to calculate the normalized cross-correlation used in equation 1. The similarity computation can be reduced using a more efficient normalized cross-correlation formula or another measure of signal similarity instead of equation 1. Even such a reduced computation will still be the most computation-expensive part of the algorithm. The following discussion applies to whatever normalized cross-correlation formula or measure of signal similarity is used. This computation enables better phase matching for each overlapping frame, thus improving the resulting sound quality.

[**0029**] FIG. **5** is a flow chart illustrating process **500** including the basic phase vocoder as known in the art. At block **501** the input signal is analyzed at equally spaced overlapping windowed frames using a short-time discrete Fourier transform. The resulting data describes short time intervals of the audio data in the frequency domain. Next the phase difference for spectral peaks is calculated (block **502**). This phase difference is the difference in phase between an input phase and a time scale modified signal phase. Block **502** uses an intrinsic sinusoidal model where the frequency is represented by the sum $\Omega_k+\omega_{ik}$: where carrier Ω_k is $2\pi k/N$; and ω_{ik} is an instantaneous frequency modulator. Block **502** estimates ω_{ik} for each spectral line by obtaining the phase difference between two consecutive analysis frames. Here, k is the spectral line and N is the size of the short-time discrete Fourier transform.

[**0030**] Process **500** reconstructs an output signal from the analyzed frames using a short-time inverse discrete Fourier transform (block **503**). The frames are overlapped by a different overlap factor to achieve the desired time scaling. The instantaneous frequency ω_{ik} is used to calculate the phase corresponding to each spectral line in the time shifted instant.

[**0031**] Consider a simple signal consisting of non-harmonically related frequencies, such as $f_1=0.5\ \sin(x)$ and $f_2=0.25\ \sin(\sqrt{2}x)$ and their sum f_3 illustrated in FIG. **6**. Because the signals f_1 and f_2 are not harmonically related, any instantaneous relationship between their respective phases will never be repeated exactly because a perfect match would require an integer number of periods of both signals. Thus a time-domain time-scale modification technique would try to find a close match within signal f_3 but there will always be some phase disruption when jumping to a different location. This phase match problem causes artifacts for many time-domain time-scale modification techniques. Now consider separating these components and performing a similar operation on each signal individually. In this case, there is little problem finding a perfect phase match for each signal, though it will be at different locations. Combining the resulting time-scaled signals produces an artifact-free time-scaled whole. Unfortunately in the real world, even narrow band signals do not repeat perfectly due to changes in pitch and amplitude, and to interference among close frequencies. However analysis in separate frequency bands gives each band great flexibility in finding the best overlap point. This improves overall quality.

[**0032**] FIG. **7** illustrates the filter bank time-scale modification method of this invention. Analysis filter bank **701** receives the input audio and generates N band limited signal in N respective frequency bands. The exact number and nature of these bands depends on the implementation and can be varied to meet various requirements including quality and computational complexity. The frequency bands are selected based on a Bark scale partition of the spectrum where each have about the same relevance in human perception. Bark scale frequency bands are more complex computationally but are better psychoacoustically. Analysis filter bank **701** can be a set of band pass finite impulse response (FIR) filters. These are preferably designed so that the bands could be simply summed in synthesis filter bank **702** to perfectly reconstruct the original signal. Each frequency band may undergo some input processing (In band blocks **711**, **721** . . . **781**). Next each frequency band is subject to time-domain time-scale modification via the corresponding TSM unit **712**, **722** . . . **782**. Following optional output processing (Out band blocks **713**, **723** . . . **783**), synthesis filter bank **702** recombines the outputs.

[**0033**] Filters of this type were tested with a signal that included both music and spoken English. The analysis filter bank separated the signal into 25 frequency bands. To simplify processing, these bands were not decimated. This resulted in 25 times the amount of original data and 25 times the time-scale modification computation. Two different time-domain time-scale modification techniques were applied to these original bands individually before summing. Listening tests showed that both time-domain time-scale modification techniques produced excellent results. These results were much better than employing the same time-domain time-scale modification techniques applied to the signal as a whole.

[**0034**] This filter bank time-scale modification method may be particularly useful for MPEG audio subbands compressed according to the MPEG Layer 3 standard commonly known as MP3. These MP3 files are already divided into Bark sub bands and decimated. This results in no data increase so that the time-scale modification computation is on the same order as needed for operating upon the original signal as a whole.

[**0035**] The filter bank time-scale modification technique of this invention is a fundamental approach that can be applied in many ways. Some but not all of these ways produce excellent results. There are no pre-defined constraints on the filter bank used nor on the time-scale modification method used within each frequency band. There is no requirement that only time-domain time-scale modification techniques be applied to individual bands. Frequency domain time-scale modification or other techniques could also be applied. There can be some relationship between the time-scale modification methods between bands. There may be communication between some or all of the bands when determining the optimum overlap point, and this point may be signal dependent. Different time-scale modification techniques may be applied to different bands. To apply filter bank time-scale modification in a useful way, various design issues must be considered such as the computational resource available and desired level of quality. Psychoacoustic principles will control which implementations are successful and which are not.

What is claimed is:

1. A method of time-scale modification of a digital audio signal comprising the steps of:

separating the digital audio signal into a plurality of frequency bands selected according to a Bark scale where each frequency band has an extent dependent upon human frequency perception;

separately time-scale modifying each of the plurality of frequency bands producing corresponding time-scale modified frequency band signals; and

combining the separate time-scale modified frequency band signals.

2. The method of claim 1, wherein:

said step of separately time-scale modifying each of the plurality of frequency bands includes time-domain time-scale modification.

3. The method of claim 2, wherein:

said step of time-domain time-scale modification of each frequency band includes

analyzing each frequency band in a set of first equally spaced, overlapping time windows having a first overlap amount S_a,

selecting a base overlap S_S for output synthesis corresponding to a desired time scale modification,

calculating a measure of similarity between overlapping frames of each frequency band for a range of overlaps between S_S+k_{min} to S_S+k_{max} of the single audio signal, where k_{min} is a minimum overlap deviation and k_{max} is a maximum overlap deviation,

determining an overlap deviation k yielding the largest measure of similarity for each frequency band,

synthesizing an output signal for each frequency band in a set of second equally spaced, overlapping time windows having a second overlap amount equal to S_S+k.

4. The method of claim 1, wherein:

the digital audio signal consists of an MPEG Layer 3 compressed audio signal; and

said step of separating the digital audio signal into a plurality of frequency bands includes

decoding the MPEG Layer 3 compressed audio signal into a plurality of decimated Bark subbands, and

employing the decimated subbands as the plurality of frequency bands.

5. A digital audio apparatus comprising:

a source of a digital audio signal;

a digital signal processor connected to said source of a digital audio signal programmed to perform time scale modification on the digital audio signal by

separating the digital audio signal into a plurality of frequency bands selected according to a Bark scale where each frequency band has an extent dependent upon human frequency perception,

separately time-scale modify each of the plurality of frequency bands producing corresponding time-scale modified frequency band signals,

combining the separate time-scale modified frequency band signals; and

an output device connected to the digital signal processor for outputting the time scale modified digital audio signal.

6. The digital audio apparatus of claim 5, wherein:

said digital signal processor is programmed to separately time-scale modify each of the plurality of frequency bands by time-domain time-scale modification.

7. The digital audio apparatus of claim 6, wherein:

said digital signal processor is programmed to time-domain time-scale modify each frequency band by

analyzing each frequency band in a set of first equally spaced, overlapping time windows having a first overlap amount S_a,

selecting a base overlap Ss for output synthesis corresponding to a desired time scale modification,

calculating a measure of similarity between overlapping frames of each frequency band for a range of overlaps between S_S+k_{min} to S_S+k_{max} of the single audio signal, where k_{min} is a minimum overlap deviation and k_{max} is a maximum overlap deviation,

determining an overlap deviation k yielding the largest measure of similarity for each frequency band,

synthesizing an output signal for each frequency band in a set of second equally spaced, overlapping time windows having a second overlap amount equal to S_S+k.

8. The digital audio apparatus of claim 5, wherein:

said source of a digital audio signal produces an MPEG Layer 3 compressed audio signal; and

said digital signal processor is programmed to

decode said MPEG Layer 3 compressed audio signal into a plurality of decimated Bark subbands, and

employ the decimated subbands as the plurality of frequency bands.

* * * * *