(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6619072号
(P6619072)

(45)発行日　令和1年12月11日(2019.12.11)　(24)登録日　令和1年11月22日(2019.11.22)

(51)Int.Cl.		F I		
G10L 13/08	(2013.01)	G10L	13/08	124
G10L 13/00	(2006.01)	G10L	13/00	100V
G10L 13/02	(2013.01)	G10L	13/02	110Z
H04N 21/431	(2011.01)	H04N	21/431	
H04N 21/435	(2011.01)	H04N	21/435	

請求項の数 7　(全 15 頁)

(21)出願番号　特願2018-191616(P2018-191616)
(22)出願日　平成30年10月10日(2018.10.10)
(62)分割の表示　特願2015-80593(P2015-80593)の分割
原出願日　平成27年4月10日(2015.4.10)
(65)公開番号　特開2019-23747(P2019-23747A)
(43)公開日　平成31年2月14日(2019.2.14)
審査請求日　平成30年10月10日(2018.10.10)

(73)特許権者　000004226
日本電信電話株式会社
東京都千代田区大手町一丁目５番１号
(74)代理人　100121706
弁理士　中尾　直樹
(74)代理人　100128705
弁理士　中村　幸雄
(74)代理人　100147773
弁理士　義村　宗洋
(72)発明者　鎌本　優
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内
(72)発明者　白木　善史
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】音合成装置、音合成方法、及びそのプログラム

(57)【特許請求の範囲】

【請求項１】

時系列の視覚情報に対し、当該視覚情報が表す意味に対応し、かつ、発話以外の音である音刺激を、当該時系列の視覚情報のタイミング及び数の少なくとも何れかに基づき、合成する音刺激合成部を含み、

上記時系列の視覚情報は上記音刺激とは異なる時系列の音信号と対応付けられており、

上記音刺激合成部は上記音刺激を上記時系列の音信号に重畳して出力し、

上記音刺激合成部は、複数の上記視覚情報の中から、同じ意味を表す複数の視覚情報それぞれの入力時刻を抽出し、所定の時間区間毎に、抽出した入力時刻の平均値、最頻値、最小値及び最大値の少なくともいずれかである代表値に基づいて音刺激を重畳するタイミングを求める、

音合成装置。

【請求項２】

請求項１の音合成装置であって、

上記代表値は平均値であり、

上記音刺激合成部は、上記視覚情報の中から、同じ意味を表す複数の視覚情報それぞれの入力時刻を抽出し、抽出した入力時刻の、所定の時間区間毎の平均値と分散とを持つガウス分布に従う乱数に基づいて上記音刺激を重畳するタイミングとする、

音合成装置。

【請求項３】

請求項２の音合成装置であって、
音刺激が記憶される音刺激データベースを含み、
上記音刺激合成部は、上記視覚情報が映像表示部に表示されてから消えるまでの時間繰り返して上記音刺激データベースに記憶された上記音刺激を重畳し、
フレームのインデックスをｉとし、合成後の音刺激をＹ$_i$とし、フレームｉに対して音刺激データベースから取り出した音刺激のテンプレートをＴ$_i$とし、τを音刺激のテンプレートの長さに対応する値とし、前記乱数をσ$_i$とし、音刺激を重畳するタイミングを表すインパルスをδ（ｉ・τ＋σ$_i$）とし、＊を畳み込み演算を表す演算子とし、上記音刺激合成部は、
【数５】

$$\mathrm{Y}_i = \delta(i\cdot\tau+\sigma_i)*\mathrm{T}_i \quad (i=0,1,2,\cdots)$$

により音刺激を合成する、もしくは、
前記乱数をσ$_m$とし、映像表示部に表示される同じ意味を表す視覚情報の個数をＭとし、上記音刺激合成部は、
上記音刺激が拍手の場合、音刺激を重畳するタイミングを表すインパルスを
【数９】

$$\delta\left(\frac{1}{\mathrm{M}}m\cdot\tau+\sigma_m\right)$$

とし、
【数６】

$$\mathrm{Y}_i = \sum_{m=1}^{\mathrm{M}}\delta\left(\frac{1}{\mathrm{M}}m\cdot\tau+\sigma_m\right)*\mathrm{T}_i \quad (i=0,1,2,\cdots)$$

により、音刺激を合成し、
上記音刺激が手拍子の場合、音刺激を重畳するタイミングを表すインパルスをδ（τ＋σ$_m$）とし、
【数７】

$$\mathrm{Y}_i = \sum_{m=1}^{\mathrm{M}}\delta(\tau+\sigma_m)*\mathrm{T}_i \quad (i=0,1,2,\cdots)$$

により、音刺激を合成し、
上記音刺激が笑い声の場合、音刺激を重畳するタイミングを表すインパルスを
【数１０】

$$\delta\left(\frac{1}{\mathrm{M}}m\cdot\tau+\sigma_m\right)$$

とし、
【数８】

$$\mathrm{Y}_i = \sum_{m=1}^{\mathrm{M}}\delta\left(\frac{1}{\mathrm{M}}m\cdot\tau+\sigma_m\right)*\mathrm{T}_i \quad (i=0,1,2,\cdots)$$

により、音刺激を合成する、
音合成装置。
【請求項４】
請求項１から請求項３の何れかの音合成装置であって、
音刺激が記憶される音刺激データベースを含み、

上記音刺激合成部は、前記代表値を、一番初めに音刺激を音信号に重畳するタイミングとして利用し、上記視覚情報が映像表示部に表示されてから消えるまでの時間繰り返して上記音刺激を重畳する、
音合成装置。
【請求項５】
請求項１から請求項４の何れかの音合成装置であって、
視覚情報と、個数と、視覚情報が表す意味に対応する、個数に応じた音量の音刺激とが対応付けて記憶される音刺激データベースを含み、
上記音刺激合成部は、上記時系列の視覚情報の中から抽出される同じ意味を表す複数の視覚情報とその視覚情報が映像表示部に表示される個数とに対応する音刺激を上記音刺激データベースから選択し、重畳する、
音合成装置。
【請求項６】
音合成装置による音合成方法であって、
音刺激合成部が、時系列の視覚情報に対し、当該視覚情報が表す意味に対応し、かつ、発話以外の音である音刺激を、当該時系列の視覚情報のタイミング及び数の少なくとも何れかに基づき、合成する音刺激合成ステップを含み、
上記時系列の視覚情報は上記音刺激とは異なる時系列の音信号と対応付けられており、
上記音刺激合成ステップは上記音刺激を上記時系列の音信号に重畳して出力し、
上記音刺激合成ステップは、複数の上記視覚情報の中から、同じ意味を表す複数の視覚情報それぞれの入力時刻を抽出し、所定の時間区間毎に、抽出した入力時刻の平均値、最頻値、最小値及び最大値の少なくともいずれかである代表値に基づいて音刺激を重畳するタイミングを求める、
音合成方法。
【請求項７】
請求項１から請求項５の何れかの音合成装置として、コンピュータを機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、映像を見るものによって入力されるテキスト情報を、その映像に重畳して表示する技術に関する。
【背景技術】
【０００２】
映像を見るものによって入力されるテキスト情報を、その映像に重畳して表示する従来技術として非特許文献１が知られている。非特許文献１では、視聴者は、動画を視聴しながら、コメントを投稿することができる。
【０００３】
また、伝送元において収音された拍手や手拍子音、声援・掛け声などの環境音を効率よく伝送し、伝送先で伝送元の場の雰囲気を再現する従来技術として特許文献１が知られている。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】特開２０１４－６３１４５号公報
【非特許文献】
【０００５】
【非特許文献１】「動画の視聴　コメントの投稿」、[online]、NIWANGO.INC、[平成27年2月2日検索]、インターネット<URL : http://info.nicovideo.jp/help/player/howto/>
【発明の概要】

【発明が解決しようとする課題】
【０００６】
しかしながら、従来技術では、テキスト情報が重畳された映像の雰囲気を再現することは難しい。
【０００７】
本発明は、テキスト情報が重畳された映像の雰囲気を再現する音合成装置、音合成方法及びそのプログラムを提供することを目的とする。
【課題を解決するための手段】
【０００８】
上記の課題を解決するために、本発明の一態様によれば、音合成装置は、時系列の視覚情報に対し、当該視覚情報が表す意味に対応し、かつ、発話以外の音である音刺激を、当該時系列の視覚情報のタイミング及び数の少なくとも何れかに基づき、合成する音刺激合成部を含み、時系列の視覚情報は音刺激とは異なる時系列の音信号と対応付けられており、音刺激合成部は音刺激を時系列の音信号に重畳して出力し、音刺激合成部は、複数の視覚情報の中から、同じ意味を表す複数の視覚情報それぞれの入力時刻を抽出し、所定の時間区間毎に、抽出した入力時刻の平均値、最頻値、最小値及び最大値の少なくともいずれかである代表値に基づいて音刺激を重畳するタイミングを求める。
【０００９】
上記の課題を解決するために、本発明の他の態様によれば、音合成方法は、音刺激合成部が、時系列の視覚情報に対し、当該視覚情報が表す意味に対応し、かつ、発話以外の音である音刺激を、当該時系列の視覚情報のタイミング及び数の少なくとも何れかに基づき、合成する音刺激合成ステップを含み、時系列の視覚情報は音刺激とは異なる時系列の音信号と対応付けられており、音刺激合成ステップは音刺激を時系列の音信号に重畳して出力し、音刺激合成ステップは、複数の視覚情報の中から、同じ意味を表す複数の視覚情報それぞれの入力時刻を抽出し、所定の時間区間毎に、抽出した入力時刻の平均値、最頻値、最小値及び最大値の少なくともいずれかである代表値に基づいて音刺激を重畳するタイミングを求める。
【発明の効果】
【００１０】
本発明によれば、テキスト情報が重畳された映像の雰囲気を再現することができるという効果を奏する。
【図面の簡単な説明】
【００１１】
【図１】第一実施形態に係る音合成装置の機能ブロック図。
【図２】第一実施形態に係る音合成装置の処理フローの例を示す図。
【図３】テキスト情報付き映像信号の例を示す図。
【図４】合成した音刺激を重畳した音信号の例を示す図。
【図５】音刺激データベースのデータ例を示す図。
【図６】テキスト情報が表示されるタイミングで、音刺激を音信号に重畳する例を示す図。
【図７】テキスト情報の個数が多いタイミングに合わせて、音刺激を音信号に重畳する例を示す図。
【図８】音刺激合成部の音刺激合成手順を例示する図。
【図９】テキスト情報の個数に応じて、音刺激の音量を変更し、変更後の音刺激を音信号に重畳する例を示す図。
【図１０】音刺激データベースのデータ例を示す図。
【発明を実施するための形態】
【００１２】
以下、本発明の実施形態について、説明する。なお、以下の説明に用いる図面では、同じ機能を持つ構成部や同じ処理を行うステップには同一の符号を記し、重複説明を省略す

る。
【００１３】
＜第一実施形態＞
図１は第一実施形態に係る音合成装置１００の機能ブロック図を、図２はその処理フローを示す。
【００１４】
音合成装置１００は、例えば、動画の視聴者によって操作される視聴者端末内に組み込まれる。なお、本実施形態において、動画とは、時間軸に同期させた音信号と共に提供される映像信号を意味する。視聴者端末は、入力部（キーボード、マウス、タッチパネル等）と、映像表示部（ディスプレイ、タッチパネル等）と音再生部（スピーカ等）を含み、例えば、パーソナルコンピュータ、スマートホン、タブレット等からなる。視聴者端末及び動画配信サーバ９２は、通信回線を介して通信可能とされている。視聴者は、視聴者端末の入力部を介して、動画配信サーバ９２に対して動画の再生を要求する（Ｓ１）ことができる。
【００１５】
＜動画配信サーバ９２＞
動画配信サーバ９２は、動画データベース及びビデオカメラから動画を受け取り、視聴者端末の要求に応じて、動画データベース内に格納されている動画、または、ビデオカメラで収録した動画をリアルタイムで配信する（Ｓ２）。また、ビデオカメラで収録された動画に限らず、リアルタイムで合成・編集されたＣＧやモーションキャプチャ等から合成されたＣＧをリアルタイム配信することもある。動画データベースには、動画と共に動画に付加されたテキスト情報が記憶され、動画と共に配信される。さらに、テキスト情報にはメタデータが付加されている。メタデータとしては、テキスト情報の入力時刻、テキスト情報の大きさ、その色、その出現方法、その移動速度や、移動位置等がある。例えば、大きさ、色、出現方法、移動速度、移動位置等は、テキスト情報の入力者が選択できるものとしてもよく、視聴者端末がメタデータとしてテキスト情報と一緒に送信し、動画データベースに動画と共に記憶される。なお、テキスト情報付き動画に含まれるテキスト情報、音信号及び映像信号は、時間軸において同期しており、それぞれ時系列において対応付けられている。
【００１６】
＜音合成装置１００＞
音合成装置１００には、テキスト情報付きの動画、より詳しく言うと、時間軸において同期しているテキスト情報、音信号及び映像信号が入力され、テキスト情報付き映像信号（図３参照）に合わせて、テキスト情報に対応する音刺激を合成し（Ｓ１２０）、合成した音刺激を重畳した音信号（図４参照）を出力する。時間軸において同期しているテキスト情報、映像信号、音刺激及び音信号を併せて音刺激及びテキスト情報付き動画ともいう。
【００１７】
視聴者端末の映像表示部及び音再生部は、音刺激及びコメント情報付きの動画を再生し（Ｓ４）、視聴者は、視聴者端末の映像表示部及び音再生部を介して、音刺激及びコメント情報付きの動画を視聴することができる。
【００１８】
例えば、非特許文献１のニコニコ動画（登録商標）では視聴者側から拍手や手拍子を表現するために「８」という文字を入力して手を打ったことを表現するテキスト情報が使われている。また、笑いを表すために「ｗ」という文字が使われている。仮に、これらのテキスト情報に対して、従来の音声合成技術を適用した場合、これらのテキスト情報に対してそれぞれ、「ハチ」「ダブリュ」という音声が合成される。一方、本実施形態では、「８」は拍手音及び手拍子音の何れか、「ｗ」は笑い声というように変換し合成音を出力する。
【００１９】

音合成装置１００は、音刺激データベース１１０と、音刺激合成部１２０とを含む。

【００２０】

＜音刺激データベース１１０＞

音刺激データベース１１０には、テキスト情報と、そのテキスト情報が表す意味に対応する音刺激のテンプレートとが対応付けられて記憶されている（図５参照）。なお、本実施形態ではテキスト情報は所定の行為を意味するものとする。また、音刺激は、発話以外の音である。ここで、「発話以外の音」とは、「言語を音声として発し、その結果として発せられた音声」以外の音を意味し、例えば、笑い声、拍手音である。例えば、笑いを意味するテキスト情報「w」に対して、「ダブリュ」という音声波形のテンプレートではなく、「笑い声」の音の波形のテンプレート（笑い声の場合、例えば数秒分のテンプレート）が記憶されている。また、拍手及び手拍子の何れかを意味するテキスト情報「８」に対して、「ハチ」という音声波形のテンプレートではなく、「拍手音及び手拍子音の何れか」の音の波形のテンプレート（拍手音及び手拍子音の場合、例えば、数百ミリ秒分のテンプレート）が記憶されている。なお、図５の例では、テキスト情報と音刺激のテンプレートとが１対１で対応しているが、１対多、多対１、多対多で対応してもよい。つまり、(1)所定の意味（例えば笑い）を表す１つのテキスト情報（例えば「w」）とその意味に対応する複数の音刺激のテンプレート（複数の笑い声のバリエーションを用意する）とが１対多で対応してもよいし、(2)所定の意味（例えば笑い）を表す複数のテキスト情報（例えば「w」「（笑）」「(^o^)」）とその意味に対応する１つの音刺激のテンプレート（１つの笑い声を用意する）とが多対１で対応してもよいし、(3)所定の意味（例えば笑い）を表す複数のテキスト情報（例えば「w」「（笑）」「(^o^)」）とその意味に対応する複数の音刺激のテンプレート（複数の笑い声のバリエーションを用意する）とが多対多で対応してもよい。なお、拍手音、手拍子音及び笑い声等は、その時々により、異なるほうがより自然に聞こえるため、テキスト情報が重畳された映像の雰囲気をより自然に再現しようとするならば、複数の音刺激のテンプレートを用意するほうがよい。

【００２１】

＜音刺激合成部１２０＞

音刺激合成部１２０は、テキスト情報付き動画（テキスト情報+音信号+映像信号）を受け取り、時系列のテキスト情報に対し、テキスト情報が表す意味に対応する音刺激のテンプレートを音刺激データベース１１０から取り出し、時系列のテキスト情報のタイミングと数の少なくとも何れかに基づき、音刺激を合成し（Ｓ１２０）、時系列の音信号に対応付けて出力する。なお、あるテキスト情報が表す意味に対応する音刺激のテンプレートが複数存在する場合には、その中から１つをランダムに選択すればよい。音刺激合成部１２０は、選択した音刺激のテンプレートを、必要に応じて前のフレームと補間をして、所定の時間長のフレーム単位（例えば映像の１フレームに対応する時間長）で、１フレームごとに励起される音刺激を合成する。音刺激合成部１２０は、合成した音刺激を受け取った音信号に重畳して（時系列の音信号に対応付けて）、出力する。

【００２２】

例えば、拍手及び手拍子の何れかを意味するテキスト情報「８」に対し、音刺激データベース１１０から対応する拍手音及び手拍子音の何れかの音の波形のテンプレート（例えば数百ミリ秒分のテンプレート）を取り出し、必要に応じて前のフレームと補間をして、１フレームごとに励起される拍手音及び手拍子音の何れかの音刺激を合成する。そして、所定の時間分の拍手音及び手拍子音の何れかの音刺激を合成し、音信号に重畳する。同様の方法により、笑いを意味するテキスト情報「w」に対し、所定の時間分の笑い声の音刺激を合成し、音信号に重畳してもよい。

【００２３】

なお、発話を意味するテキスト情報に合わせて従来の音声合成装置を用いて音声を合成してもよい。この場合、テキスト情報に対して、まず、本実施形態の音刺激合成を行い、音刺激合成の対象とならないテキスト情報に対して従来の音声合成装置を用いて音声を合成すればよい。例えば「素晴らしい８８」というテキスト情報が入力された場合、「すば

らしいハチハチ」という音声を合成するのではなく、「すばらしい（音声）＋拍手音（音刺激）」という音を合成し、音信号に重畳する。このような構成とすることで、音信号に対して、従来の音声合成により合成された音声（話し声）と共に、音刺激（拍手音、手拍子音及び笑い声等）が重畳され、テキスト情報が重畳された映像の雰囲気をより自然に再現することができる。

【００２４】

なお、上述の通り、テキスト情報「８」は拍手及び手拍子の何れかを意味する。何れも手を叩く行為であるが、「手拍子」は一定のテンポに合わせて手を叩く行為であり、「拍手」は一定のテンポを持たずに手を叩く行為である。ここで、「手拍子」と「拍手」とは、手を叩く時間的間隔や音量的差異が異なるため（参考文献１）、例えば、音信号等に基づいて、テキスト情報が何れの行為を意味するのかを判別することができる。

（参考文献１）鎌本優，河原一彦，尾本章，守谷健弘，「音楽鑑賞時に励起される拍手音・手拍子音の低遅延伝送に向けた基礎的検討」、日本音響学会 2014年秋季研究発表会，1-Q-17、2014年.

【００２５】

例えば、音信号が曲を表し、一定のテンポがある場合には、テキスト情報「８」は手拍子を意味する可能性が高い。また、曲が終了後のテキスト情報「８」は拍手を意味する可能性が高い。また、テキスト情報「８」が一度の入力で連続している場合、つまりテキスト情報「８８…」が入力された場合には、「パチパチ…」を意味し、拍手を意味する可能性が高い。また、テキスト情報「８」が周期的に入力される場合には、手拍子を意味する可能性が高い。

【００２６】

(音刺激を重畳するタイミング)

(1)テキスト情報が表示されるタイミング(例えばテキスト情報の入力時刻)で、音刺激を音信号に重畳する（図６参照）。

【００２７】

(2)テキスト情報の個数が多いタイミングに合わせて、音刺激を音信号に重畳する（図７参照）。

【００２８】

複数のテキスト情報の中から、同じ意味を表すテキスト情報を抽出する。例えば、音刺激データベース１１０を参照して、音刺激合成部１２０は、同じ意味を表すテキスト情報毎に分類し、テキスト情報の入力時刻を抽出する。

【００２９】

音刺激合成部１２０は、抽出した入力時刻の統計量に基づいて音刺激を音信号に重畳するタイミングを求める。例えば、抽出した入力時刻を用いて、所定の時間区間毎に、時間区間毎の代表値(平均値、最頻値、最小値及び最大値等の複数の入力時刻を代表する何らかの値)を求め、重畳するタイミングとして検出する。例えば、抽出した入力時刻を用いて、ヒストグラムを作成し、多数決により重畳するタイミングを求める。つまり、最頻値を重畳するタイミングとする。

【００３０】

例えば、テキスト情報が手拍子を意味し、動画がミュージックビデオであり、曲のテンポが148BPM(Beats per Minutes)の場合、一拍の間隔は405ms程度なので、所定の時間区間を405msとする。また、例えば、テキスト情報が「拍手」または「笑い」を意味する場合、所定の時間区間を一連の行為「拍手」または「笑い」が、継続しうる最大の時間に設定する。例えば、何らかの事象に対して、「拍手」を送るのは、長くとも30秒程度であろうと想定される場合、最初に「拍手」を意味するテキスト情報が表示されてから１分以内に表示される「拍手」を意味する他のテキスト情報から代表値を求め、重畳するタイミングとして検出する。

【００３１】

なお、この方法を用いる場合、所定の時間区間分のテキスト情報付き動画（テキスト情

報+音信号+映像信号）をバッファリングしておき、音刺激を重畳して、音刺激及びテキスト情報付き動画を出力すればよい。

【００３２】

(繰り返し重畳する場合)

なお、一人の人間による一拍分の音刺激（拍手音、手拍子音、笑い声等）を音刺激データベース１１０に記憶しておき、音刺激を音信号に繰り返し重畳する構成としてもよい。その場合、(音刺激を重畳するタイミング)の(1)及び(2)で求めたタイミングを、一番初めに音刺激を重畳するタイミングとして利用する。所定の時間区間分（例えば、テキスト情報が映像表示部に表示されてから消えるまで）繰り返し重畳すればよい。

【００３３】

例えば、同じ意味を表すテキスト情報が映像表示部に表示される個数が１個の場合は、図８Ａのように、所定の間隔毎（例えば、拍手の場合約３００ｍｓごと）に、音刺激を重畳する。なお、音刺激を重畳するタイミングに揺らぎを持たせてもよい。例えば、音刺激が拍手の場合、所定の間隔は約３００ｍｓでよいが、より好ましくは３００ｍｓを中心として時間間隔に揺らぎを持たせる。時間間隔に揺らぎを持たせることによってさらに自然な拍手音を合成することができる（参考文献１参照）。たとえば３００ｍｓを中心としてガウス分布にしたがう乱数により、±数１０ｍｓの揺らぎを持たせればよい。例えば、フレームのインデックスをｉとし、合成後の音刺激（拍手音）をY_iとし、フレームｉに対して音刺激データベース１１０から取り出した音刺激のテンプレートをT_iとし、テンプレートT_iの長さ（テンプレートT_iに含まれる全フレームに含まれる、音刺激のデータのサンプル数）をＰとし、音刺激合成部１２０は、テンプレートT_i＝（t_i［１］　t_i［２］　…　t_i［Ｐ］）と拍手タイミングを表すインパルスδ（ｉ・τ＋σ_i）の畳み込み演算でY_iを出力とする。テンプレートの長さは所定の間隔（拍手の場合では約３００ｍｓ程度）よりも短いほうが、音が重ならないため好ましい。

【００３４】

【数１】

$$Y_i = \delta(i \cdot \tau + \sigma_i) * T_i \quad (i = 0,1,2,\cdots)$$

【００３５】

ここで＊は畳み込み演算を表す。ここで、τ＝３００ｍｓであり、σ_iは－１０ｍｓ≦σ_i≦＋１０ｍｓの範囲で生成した乱数である。音刺激を重畳するタイミングはフレーム間隔で特定し、１フレームごとに励起される音刺激を合成し、その結果として、音の波形のテンプレート分（例えば、笑い声の場合、数秒分、拍手音及び手拍子音の場合、数百ミリ秒分）の音刺激を合成し重畳する。１人分の手拍子を合成する場合、音刺激を重畳する間隔は、手拍子の対象に応じて変化し、曲のテンポが１４８ＢＰＭ（Beats Per Minute）の場合、４０５ｍｓ前後とする。さらに、時間間隔の揺らぎは、拍手の場合よりも手拍子の場合のほうが小さく設定したほうがよく、例えば、手拍子の場合のσ_iの範囲が拍手の場合のσ_iの範囲よりも小さくなるように設定する。

【００３６】

同じ意味を表すテキスト情報が映像表示部に表示される個数に応じて、音刺激を重畳する時間間隔を変更してもよい。例えば、拍手を表すＭ個のテキスト情報が映像表示部に表示されている場合、図８Ｂのように、時間間隔を約３００／Ｍ（ｍｓ）ごとに音刺激を重畳する。個数Ｍの逆数を使って、時間間隔を約３００／Ｍ（ｍｓ）と設定することで、拍手を表すテキスト情報の個数Ｍが増えるに従って時間間隔が小さくなるように設定することができる。この場合もガウス分布やラプラス分布に従う乱数によって、揺らぎを持たせることができる。例えば音刺激合成部１２０は、

【００３７】

【数２】

$$\mathrm{Y}_i = \sum_{m=1}^{\mathrm{M}} \delta\left(\frac{1}{\mathrm{M}} m \cdot \tau + \sigma_m\right) * \mathrm{T}_i \quad (i = 0,1,2,\cdots)$$

【００３８】

によりテンプレートを変換した音刺激Ｙｉ（ｉ＝０，１，２，・・・）を重畳する。テキスト情報の個数Ｍの手拍子を合成する場合、Ｍ個の手拍子のタイミングはほぼ同じなので、例えば、

【００３９】

【数３】

$$\mathrm{Y}_i = \sum_{m=1}^{\mathrm{M}} \delta(\tau + \sigma_m) * \mathrm{T}_i \quad (i = 0,1,2,\cdots)$$

【００４０】

によりテンプレートを変換した音刺激Ｙｉ（ｉ＝０，１，２，・・・）を重畳する。なお、この場合も時間間隔の揺らぎは、拍手の場合よりも手拍子の場合のほうが小さく設定したほうがよい（参考文献１参照）。

【００４１】

音刺激の例として拍手音及び手拍子音の何れかを対象として説明したが、これに限らず拍手音及び手拍子音以外の音刺激（たとえば、一人の人間による笑い声）を対象としても良い。

【００４２】

なお、音刺激合成部１２０において、音刺激のテンプレートＴｉ＝（ｔｉ［１］　ｔｉ［２］　…　ｔｉ［Ｐ］）と笑い声を表すインパルスδ（ｍ・τ＋σｍ）の畳み込み演算でＹｉを出力としても良い。

【００４３】

【数４】

$$\mathrm{Y}_i = \sum_{m=1}^{\mathrm{M}} \delta\left(\frac{1}{\mathrm{M}} m \cdot \tau + \sigma_m\right) * \mathrm{T}_i \quad (i = 0,1,2,\cdots)$$

【００４４】

この場合、τは笑い声のテンプレートの長さ（数秒）に対応する値とする。

【００４５】

(揺らぎのバリエーション)

揺らぎを持たせる際のバリエーションについて説明する。

【００４６】

例えば、抽出した入力時刻を用いて、所定の時間区間毎に、入力時刻の平均値と分散とを求め、その平均値と分散とを持つガウス分布に従う乱数を重畳するタイミングとしてもよい。この方法により、音刺激を重畳するタイミングにゆらぎを与えることができ、より自然なタイミングで音刺激を再生することができる。

【００４７】

テキスト情報が手拍子を意味する場合、予め手拍子を行う際に一般的に生じる分散の値を求めておき、その分散に基づき、重畳するタイミングを求めてもよい。例えば、上述の方法で重畳するタイミングを求め、その重畳するタイミングを中心として、手拍子を行う際に一般的に生じる分散を持つガウス分布に従う乱数を新たな(最終的に用いる)重畳するタイミングとする。

【００４８】

(音刺激の音量を調整する方法)
音刺激の音量を調整する方法を説明する。
【００４９】
(1)テキスト情報の個数に応じて、音刺激の音量を変更し、変更後の音刺激を音信号に重畳する（図９参照）。例えば、映像表示部に表示されるテキスト情報の個数が多くなるほど音量が大きくなるように変更する。
【００５０】
(2)テキスト情報の大きさに応じて、音刺激の音量を変更し、変更後の音刺激を音信号に重畳する。例えば、映像表示部に表示されるテキスト情報の大きさが大きくなるほど音量が大きくなるように変更する。
【００５１】
(3)音刺激データベース１１０に個数に応じて異なる音量の音刺激を収音し記憶しておき、音刺激合成部１２０は、テキスト情報とその個数に応じて音刺激を選択してもよい。この場合、音刺激データベース１１０には、テキスト情報と、個数と、そのテキスト情報が表す意味に対応する音刺激とが対応付けられて記憶されている。例えば、テキスト情報「w」と、個数１と、１人分の笑い声からなる音刺激とが対応付けられて記憶されており、テキスト情報「w」と、個数２と、２人分の笑い声からなる音刺激とが対応付けられて記憶されている。なお、テキスト情報と個数と音刺激とが１対１対１で対応してもよいし、１対１対多、多対１対１、多対１対多で対応してもよい。
【００５２】
＜効果＞
以上の構成により、テキスト情報が重畳された映像の雰囲気を再現することができる。
【００５３】
＜変形例＞
本実施形態では、音合成装置１００は、音刺激及びテキスト情報付き動画を出力しているが、本実施形態のポイントは、音刺激を合成することであり、少なくとも音刺激を出力すればよい。例えば、本実施例のように視聴者端末（パーソナルコンピュータ、スマートホン、タブレット等）内に本実施形態の音合成装置１００が組み込まれてもよいし、動画配信サーバ内に音合成装置１００が組み込まれてもよい。また、音合成装置１００を独立した装置として構成してもよい。少なくとも時間軸において音信号または映像信号に同期しているテキスト情報を受け取り、音刺激を合成し、出力することができればよい。音刺激を音信号に同期させる処理等は別装置において行ってもよい。
【００５４】
本実施形態では、視聴者によって入力され、映像信号に重畳して表示される情報としてテキスト情報の例を示したが、他の視覚情報であってもよい。ここで、「視覚情報」とは、映像表示部を介して視覚的に認識可能な情報であって、例えば、文字、図形若しくは記号若しくはこれらの結合又はこれらと色彩との結合である。また、静止画に限らず、動く画像であってもよい。例えば、(1)本実施形態のように、「笑い」や「拍手」等の所定の行為を意味するテキスト情報（例えば「w」や「８」等）、(2)テキスト情報以外の「笑い」や「拍手」等の所定の行為を意味し、識別するためのコンピュータ上のビット情報、(3)顔文字、絵文字等、通常のテキスト情報で無いもの。例えば、キャリアの異なる携帯電話間で共通絵文字(参考文献２参照)、(4)アスキーアート等，全体としてはテキスト情報とテキスト情報の配置情報を用いた絵のようになっているもの（参考文献３参照）
（参考文献２）「docomo／au共通絵文字」、株式会社NTTドコモ、[online]、[平成27年2月9日検索]、インターネット<URL: https://www.nttdocomo.co.jp/service/developer/smart_phone/make_contents/pictograph/>
（参考文献３）「アスキーアート」、[online]、2015年2月2日、ウィキペディア、[平成27年2月9日検索]、インターネット<URL: http://ja.wikipedia.org/wiki/%E3%82%A2%E3%82%B9%E3%82%AD%E3%83%BC%E3%82%A2%E3%83%BC%E3%83%88>
このテキスト情報以外の視覚情報を含む場合の音刺激データベース１１０に格納されるデ

ータの例を図１０に示す。

【００５５】

＜その他の変形例＞

本発明は上記の実施形態及び変形例に限定されるものではない。例えば、上述の各種の処理は、記載に従って時系列に実行されるのみならず、処理を実行する装置の処理能力あるいは必要に応じて並列的にあるいは個別に実行されてもよい。その他、本発明の趣旨を逸脱しない範囲で適宜変更が可能である。

【００５６】

＜プログラム及び記録媒体＞

また、上記の実施形態及び変形例で説明した各装置における各種の処理機能をコンピュータによって実現してもよい。その場合、各装置が有すべき機能の処理内容はプログラムによって記述される。そして、このプログラムをコンピュータで実行することにより、上記各装置における各種の処理機能がコンピュータ上で実現される。

【００５７】

この処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体としては、例えば、磁気記録装置、光ディスク、光磁気記録媒体、半導体メモリ等どのようなものでもよい。

【００５８】

また、このプログラムの流通は、例えば、そのプログラムを記録したＤＶＤ、ＣＤ－ＲＯＭ等の可搬型記録媒体を販売、譲渡、貸与等することによって行う。さらに、このプログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することにより、このプログラムを流通させてもよい。

【００５９】

このようなプログラムを実行するコンピュータは、例えば、まず、可搬型記録媒体に記録されたプログラムもしくはサーバコンピュータから転送されたプログラムを、一旦、自己の記憶部に格納する。そして、処理の実行時、このコンピュータは、自己の記憶部に格納されたプログラムを読み取り、読み取ったプログラムに従った処理を実行する。また、このプログラムの別の実施形態として、コンピュータが可搬型記録媒体から直接プログラムを読み取り、そのプログラムに従った処理を実行することとしてもよい。さらに、このコンピュータにサーバコンピュータからプログラムが転送されるたびに、逐次、受け取ったプログラムに従った処理を実行することとしてもよい。また、サーバコンピュータから、このコンピュータへのプログラムの転送は行わず、その実行指示と結果取得のみによって処理機能を実現する、いわゆるＡＳＰ（Application Service Provider）型のサービスによって、上述の処理を実行する構成としてもよい。なお、プログラムには、電子計算機による処理の用に供する情報であってプログラムに準ずるもの（コンピュータに対する直接の指令ではないがコンピュータの処理を規定する性質を有するデータ等）を含むものとする。

【００６０】

また、コンピュータ上で所定のプログラムを実行させることにより、各装置を構成することとしたが、これらの処理内容の少なくとも一部をハードウェア的に実現することとしてもよい。

【図１】

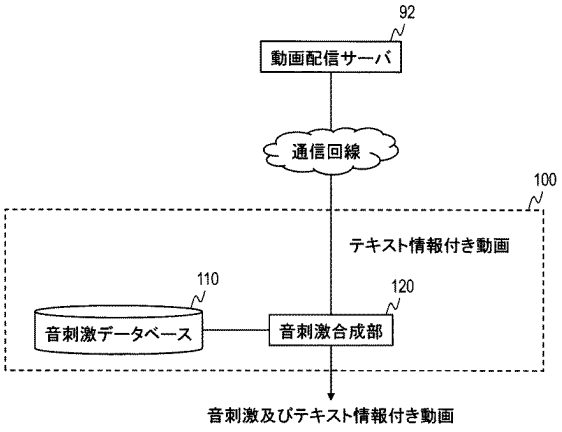

図1

【図２】

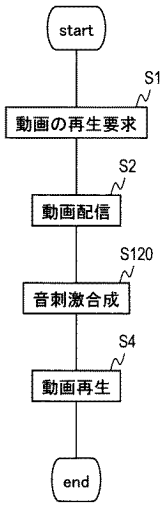

図2

【図３】

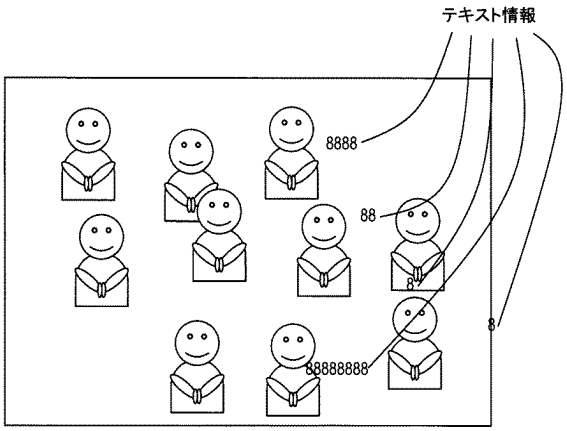

図3

【図４】

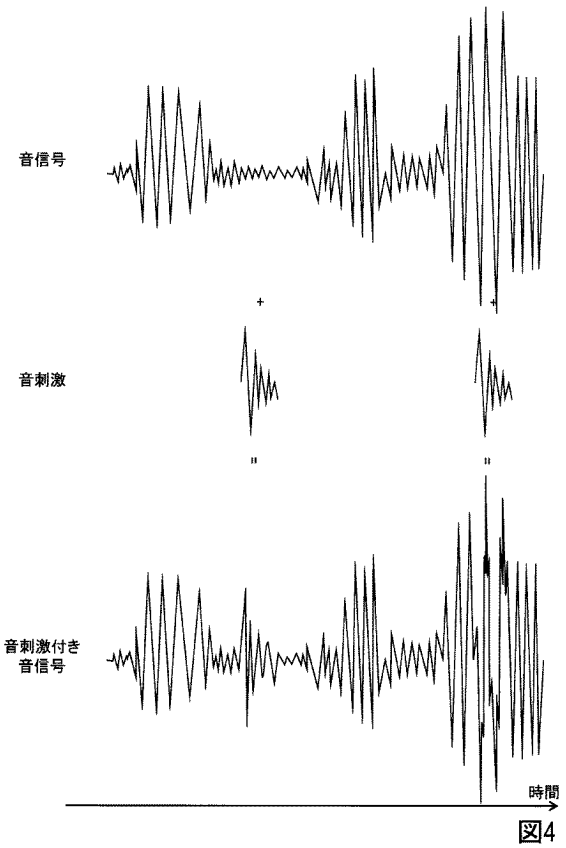

図4

【図５】

	テキスト情報	音刺激
1	w	
2	8	
…	…	…

図5

【図６】

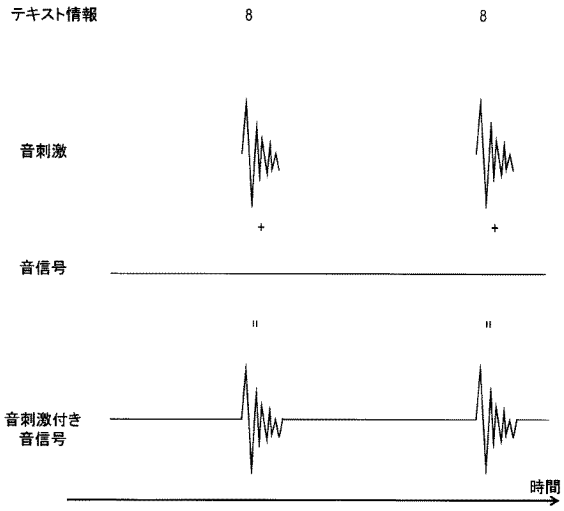

図6

【図７】

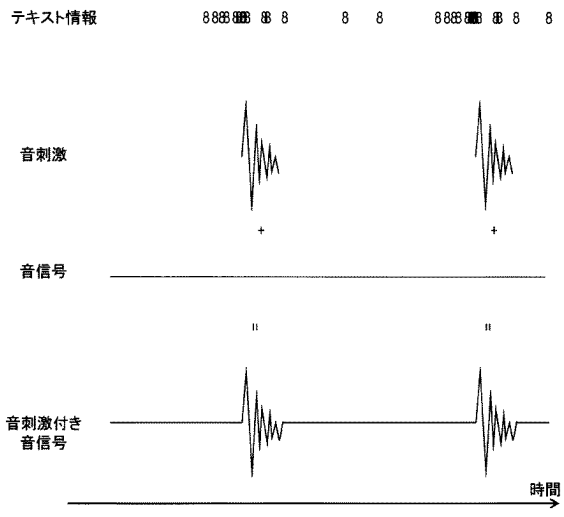

図7

【図８】

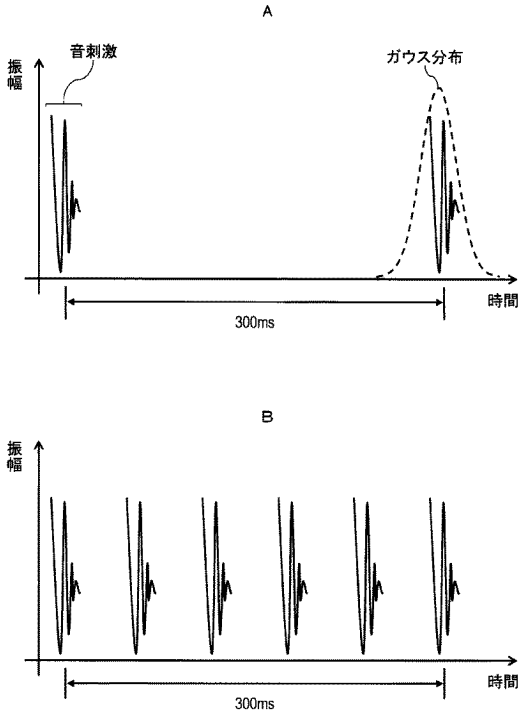

図8

【図９】

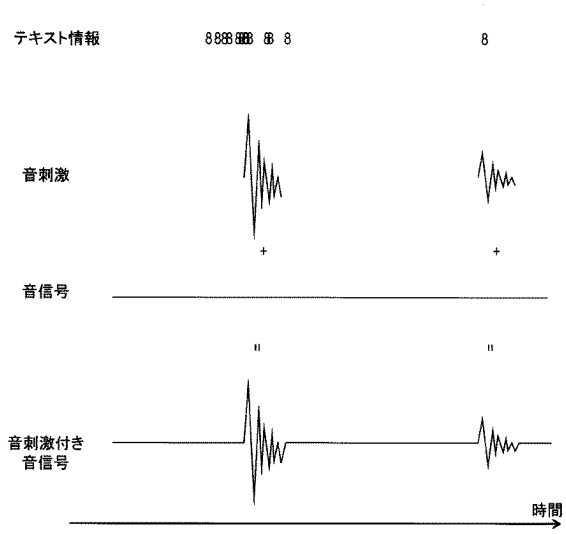

図9

【図１０】

	視覚情報	(意味)	音刺激
1	「w」	(笑い)	
2	「(笑)」		
3			
4	「8」	(拍手)	
5			
..	…		

図10

フロントページの続き

(72)発明者　佐藤　尚
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内
(72)発明者　ガブリエル　パブロ　ナバ
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内
(72)発明者　守谷　健弘
東京都千代田区大手町一丁目５番１号　日本電信電話株式会社内

審査官　岩田　淳

(56)参考文献　特開２０１４－２１２４９０（ＪＰ，Ａ）
特開２０１４－０６３１４５（ＪＰ，Ａ）
特開２０１２－１３３６６２（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ１０Ｌ　　１３／００－１３／１０
Ｈ０４Ｎ　　２１／００－２１／８５８