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

特許第6367748号
(P6367748)

(45)発行日　平成30年8月1日(2018.8.1)　(24)登録日　平成30年7月13日(2018.7.13)

(51)Int.Cl.		F I		
H04N 21/472	(2011.01)	H04N	21/472	
H04N 21/4788	(2011.01)	H04N	21/4788	
H04N 21/431	(2011.01)	H04N	21/431	
G10L 15/00	(2013.01)	G10L	15/00	200G
G06F 13/00	(2006.01)	G06F	13/00	560A

請求項の数 9　(全 13 頁)

(21)出願番号　特願2015-80592(P2015-80592)
(22)出願日　平成27年4月10日(2015.4.10)
(65)公開番号　特開2016-201678(P2016-201678A)
(43)公開日　平成28年12月1日(2016.12.1)
審査請求日　平成29年4月25日(2017.4.25)

(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】認識装置、映像コンテンツ提示システム

(57)【特許請求の範囲】
【請求項１】
収音された少なくとも発話以外の音を含む音信号に基づき、上記発話以外の音に対応し、かつ、上記発話以外の音をそのまま表記した文字列ではない、所定の視覚情報を得る認識部を含み、
上記所定の視覚情報は、映像コンテンツに重畳して表示するために得るものであり、
上記認識部は、
映像コンテンツに既に重畳されている１種類以上の視覚情報から上記所定の視覚情報を得る、
認識装置。
【請求項２】
請求項１の認識装置であって、
上記認識部は、
上記音信号の収音時刻に対応する映像コンテンツの時刻に、既に重畳されている視覚情報が複数種類ある場合には、重畳されている数が最も多い種類の視覚情報を、上記所定の視覚情報として得る、
認識装置。
【請求項３】
請求項１の認識装置であって、
上記認識部は、(1-a)上記音信号の収音時刻に対応する映像コンテンツの時刻に、既に

重畳されている視覚情報が1種類である場合、または、(1-b)複数種類あるがそのうち1種類の割合が極めて高い場合には、当該種類の視覚情報を上記所定の視覚情報として得て、(2)上記以外の場合には、映像コンテンツに重畳して表示するために予め用意されている複数種類の視覚情報から上記所定の視覚情報を得る、
認識装置。
【請求項４】
請求項１の認識装置であって、
上記認識部は、(1)上記音信号の収音時刻に対応する映像コンテンツの時刻に、既に重畳されている視覚情報に占める1種類の視覚情報の割合が高い場合には、当該種類の視覚情報を上記所定の視覚情報として得ることを優先し、(2)上記以外の場合には、映像コンテンツに重畳して表示するために予め用意されている複数種類の視覚情報から上記所定の視覚情報を得ることを優先することで、上記所定の視覚情報を得る、
認識装置。
【請求項５】
請求項１の認識装置であって、
上記認識部は、
上記音信号の収音時刻に対応する映像コンテンツの時刻に、既に重畳されている視覚情報が複数種類ある場合には、既に重畳されている複数種類の視覚情報から、重畳されている数の割合の確率に応じてそれぞれの所定の視覚情報が得られるように、ランダムに上記所定の視覚情報を得る、
認識装置。
【請求項６】
収音された少なくとも発話以外の音を含む音信号に基づき、上記発話以外の音に対応し、かつ、上記発話以外の音をそのまま表記した文字列ではない、所定の視覚情報を得る認識部を含み、
上記所定の視覚情報は、映像コンテンツに重畳して表示するために得るものであり、
上記認識部は、
映像コンテンツに既に重畳されている１種類以上の視覚情報、映像コンテンツに重畳して表示するために予め用意されている１種類以上の視覚情報の少なくとも何れかから第一の視覚情報を得、
上記第一の視覚情報の繰り返し回数が得られた場合に、得られた繰り返し回数に応じて、前記第一の視覚情報を繰り返した情報と同じ意味を表す他の視覚情報を前記所定の視覚情報とする、
認識装置。
【請求項７】
請求項１から請求項６の何れかの認識装置であって、
上記発話以外の音の大きさを表す指標に対応する情報を、上記所定の視覚情報が表示部を介して表示される際の大きさの情報として得る表示サイズ取得部を含む、
認識装置。
【請求項８】
映像コンテンツを表示する第一の表示部と、
上記映像コンテンツに対応して発せられた音を収音して上記音信号とする収音部と、
上記請求項１から請求項７の何れかの認識装置と、
上記認識装置が得た所定の視覚情報を映像コンテンツに重畳して表示する第二の表示部とを含む、
映像コンテンツ提示システム。
【請求項９】
メディアコンテンツを提示する提示部と、
上記メディアコンテンツに対応して発せられた音を収音して上記音信号とする収音部と、

上記請求項１から請求項７の何れかの認識装置と、
上記認識装置が得た所定の視覚情報を上記メディアコンテンツに対応する映像コンテンツに重畳して表示する表示部とを含む、
映像コンテンツ提示システム。
【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。
【先行技術文献】
【非特許文献】
【０００３】
【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】
【０００４】
しかしながら、従来技術では、動画に対してコメントしたいと思ってから、コメントを入力し、コメント投稿ボタンをクリックまたはエンターキーを押下する必要があるため、視聴者がコメントしたいと思ったタイミングから遅れてコメントが表示される場合がある。逆に動画の内容を予め知っている場合には、予めコメントを入力しておき、コメント投稿ボタンをクリックまたはエンターキーを押下するタイミングを視聴者が図ることもできるが、その場合であっても、視聴者がコメントしたいと思ったタイミングよりも早くなったり、または、遅くなったりする場合がある。例えば、ミュージックビデオやライブ映像の楽曲のテンポに合わせて、拍手を意味するテキスト情報「８」をコメントする場合、実際に拍手する場合よりも、ズレてしまう場合が多い、または、ズレ幅が大きくなりやすい。
【０００５】
本発明は、コメント投稿ボタンのクリックまたはエンターキーの押下を行わずに、情報を映像に適切なタイミングで重畳して表示するための認識装置、映像コンテンツ提示システ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】
【０００６】
上記の課題を解決するために、本発明の一態様によれば、認識装置は、収音された少なくとも発話以外の音を含む音信号に基づき、発話以外の音に対応し、かつ、発話以外の音をそのまま表記した文字列ではない、所定の視覚情報を得る認識部を含む。
【発明の効果】
【０００７】
本発明によれば、コメント投稿ボタンのクリックまたはエンターキーの押下を行わずに、情報を映像に適切なタイミングで重畳して表示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】
【０００８】
【図１】第一実施形態に係る映像コンテンツ提示システムの機能ブロック図。
【図２】第一実施形態に係る映像コンテンツ提示システムの処理フローの例を示す図。
【図３】視覚情報付き映像コンテンツの例を示す図。

【図４】視覚情報データベースの例を示す図。
【発明を実施するための形態】
【０００９】
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。
【００１０】
＜第一実施形態に係る映像コンテンツ提示システム１＞
図１は第一実施形態に係る映像コンテンツ提示システム１の機能ブロック図を、図２はその処理フローを示す。
【００１１】
映像コンテンツ提示システム１は、１台以上の視聴者端末９１と、視聴者端末１００と、映像コンテンツを視聴者端末９１及び１００に配信する動画配信サーバ９２とを含む。各視聴者端末９１及び１００と動画配信サーバ９２とは、通信回線を介して通信可能とされている。
【００１２】
＜視聴者端末９１＞
視聴者端末９１は、映像コンテンツ(例えば、動画)を見るもの(例えば、動画の視聴者)によって操作され、入力部（キーボード、マウス、タッチパネル等）と、表示部（ディスプレイ、タッチパネル等）とを含み、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。視聴者は、視聴者端末９１の入力部を介して、動画配信サーバ９２に対して映像コンテンツの再生を要求することができる。また、視聴者端末９１の表示部を介して、映像コンテンツを視聴することができる。さらに、視聴者は、入力部を介して、映像コンテンツに重畳して表示される視覚情報（例えば、コメント）を入力することができる。ここで、「視覚情報」とは、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)「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１参照)、(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２参照）、(5)上述の(1)～(4)に対応するネットスラング。例えば、「笑い」を意味するテキスト情報「ｗｗｗｗｗ…」に対して「草生えた」等のネットスラングがある。
（参考文献１）「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２）「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>
【００１３】
＜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９１及び１００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。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。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響信号と共に提供される映像コンテンツを意味する。動画データベースには、動画と共に動画に付加された視覚情報が記憶される。さらに、視覚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、視覚情報の入力時刻、視覚情報

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、視覚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９１及び１００がメタデータとして視覚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。

【００１４】

＜視聴者端末１００＞

視聴者端末１００は、映像コンテンツ(例えば、動画)を見るもの(例えば、動画の視聴者)によって操作される。視聴者端末１００は、表示部（ディスプレイ、タッチパネル等）１１０と収音部（マイクロホン等）１２０と認識部１３０と表示サイズ取得部１４０とを含み、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。視聴者は、視聴者端末１００の入力部（収音部１２０、キーボード、マウス、タッチパネル等）を介して、動画配信サーバ９２に対して映像コンテンツの再生を要求することができる。また、視聴者端末１００の表示部１１０を介して、映像コンテンツを視聴することができる。視聴者端末１００の視聴者は、入力部を介して、映像コンテンツに重畳して表示される視覚情報（例えば、コメント）を入力することができる。

【００１５】

視聴者端末１００は、表示部１１０を介して、映像コンテンツを表示する（Ｓ１１０Ａ）。なお、映像コンテンツは、動画配信サーバ９２から配信される。視聴者端末１００は、収音部１２０を介して、映像コンテンツの視聴者が、映像コンテンツに対応して発する音を音信号x(t)として収音する（Ｓ１２０）。なお、tは時刻を表すインデックスである。

【００１６】

＜認識部１３０＞

認識部１３０は、収音された少なくとも発話以外の音を含む音信号x(t)を受け取り、この音信号x(t)に基づき、発話以外の音に対応し、かつ、発話以外の音をそのまま表記した文字列ではない、視聴者端末９１の表示部または表示部１１０を介して視覚的に認識可能な所定の視覚情報v(t)を得（Ｓ１３０）、通信回線を介して動画配信サーバ９２に送信する。

【００１７】

動画配信サーバ９２は、視覚情報v(t)を受け取り、動画に重畳して配信する。なお、動画データベースに、動画と共に動画に付加された視覚情報v(t)を格納する。視聴者端末９１の表示部または表示部１１０は、視覚情報v(t)が重畳された映像コンテンツを受け取り、表示する（Ｓ１１０Ｂ）。なお、視覚情報v(t)を送信した際の再生時には、視覚情報v(t)を重畳せずに動画のみを配信し、それ以降の再生時に、視覚情報を重畳した動画を配信する構成としてもよい。

【００１８】

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。

【００１９】

「所定の視覚情報」は、前述の発話以外の音（例えば拍手音、笑い声）に対応するものであり、本実施形態では、「所定の視覚情報」として拍手音を表すテキスト情報「８」や笑い声を表すテキスト情報「ｗ」等を用いるものとする。

【００２０】

「発話以外の音をそのまま表記した文字列」とは、要は、発話以外の音を、従来の音声認識装置に入力して得られるテキスト情報である。従来の音声認識装置は、「発話」、つまり、「言語を音声として発し、その結果として発せられた音声」を認識対象としているため、「発話以外の音」を認識対象とした場合、適切な認識結果を得ることができない。例えば、発話以外の音が拍手音の場合、従来の音声認識装置に拍手音を入力しても、「パチパチパチ」といったテキスト情報を得られる可能性は低く、ノイズとして音声認識の対象とされない可能性が高い。従来の音声認識装置を用いて、音声認識の結果、「パチパチ

パチ」というテキスト情報を得たいのであれば、人が「パチパチパチ」と発音する必要がある。同様に、発話以外の音が笑い声の場合、従来の音声認識装置に笑い声を入力しても、「ワハハッ」というテキスト情報を得られる可能性は低く、ノイズとして音声認識の対象とされないか、または、笑い声とは判断できないような音声認識結果が得られる可能性がある。そこで、認識部１３０は、発話以外の音に対して、発話以外の音を従来の音声認識装置に入力して得られるテキスト情報とは異なる所定の視覚情報（例えば、拍手音を表すテキスト情報「８」や笑い声を表すテキスト情報「ｗ」）を得る。

【００２１】

以下、「発話以外の音」を認識し、認識結果として視覚情報を取得する方法について説明する。なお、本実施形態において、所定の視覚情報は、映像コンテンツに重畳して表示するために得るものである。

【００２２】

認識部１３０は、受け取った音信号x(t)から、その大きさを表す指標（例えば、音量、パワー、エネルギー）を求め、大きさを表す指標と所定の閾値との大小関係に基づき、無音か否かを判定し、無音ではなく、何らかの音を収音できたと判定したときに、以下の方法により、所定の視覚情報を取得する。例えば、大きさを表す指標が音が大きいほど大きくなる値（例えば、音量、パワー、エネルギー）の場合には、大きさを表す指標が所定の閾値未満のときに無音と判定し、閾値以上のときに何らかの音を収音できたと判定する。また、大きさを表す指標が音信号が大きいほど小さくなる値の場合には、大きさを表す指標が所定の閾値より大きいときに無音と判定し、閾値以下のときに何らかの音を収音できたと判定する。本実施形態では、音信号x(t)の音量が所定の閾値以上となったときに何らかの音を収音できたと判定する。

【００２３】

（取得方法１）

認識部１３０は、所定の視覚情報を、映像コンテンツに既に重畳されている1種類以上の視覚情報から得る。例えば、音信号x(t)の音量が所定の閾値以上となる時刻において、映像コンテンツに既に重畳されている1種類以上の視覚情報の中から1種類の視覚情報を選択し、所定の視覚情報とする。例えば、(1-1)重畳されている数が最も多い種類の視覚情報を、所定の視覚情報を選択する。また、(1-2)重畳されている数の割合に応じて、ランダムに所定の視覚情報を選択する。(1-3)重畳されている1種類以上の視覚情報の中からランダムに所定の視覚情報を選択する。

【００２４】

例えば、音信号x(t)の音量が所定の閾値以上になった時刻において、図３のような視覚情報付き映像コンテンツを受け取った場合、拍手音を表す「８」というテキスト情報と、笑い声を表す「ｗ」というテキスト情報との、２種類の視覚情報が、映像コンテンツに既に重畳されており、それぞれの視覚情報の重畳されている個数は4個と2個である。なお、本実施形態では、ある視覚情報（例えば「８」というテキスト情報）とその視覚情報の繰り返し（例えば「８８８…」というテキスト情報）とは同じ種類の視覚情報として取り扱う。ただし、異なる種類の視覚情報として取り扱ってもよい。この２種類の視覚情報から何れか一方の視覚情報を選択して、所定の視覚情報を得る。(1-1)の場合、重畳されている数が最も多い種類の視覚情報は、拍手音を表す「８」というテキスト情報なので、これを所定の視覚情報として得る。(1-2)の場合、拍手音を表す「８」というテキスト情報が重畳されている数の割合は4/6であり、笑い声を表す「ｗ」というテキスト情報が重畳されている数の割合は2/6であり、この割合に応じて、ランダムに所定の視覚情報を選択する。例えば、4/6の確率で拍手音を表す「８」というテキスト情報を所定の視覚情報として選択し、2/6の確率で笑い声を表す「ｗ」というテキスト情報を所定の視覚情報として選択する。(1-3)の場合、1/2の確率で拍手音を表す「８」というテキスト情報を所定の視覚情報として選択し、1/2の確率で笑い声を表す「ｗ」というテキスト情報を所定の視覚情報として選択する。

【００２５】

また、例えば、「音信号x(t)の音量が所定の閾値以上となる時刻において、」ではなく、「音信号x(t)の音量が所定の閾値以上となる時刻までに、」映像コンテンツに既に重畳されている1種類以上の視覚情報の中から1種類の視覚情報を選択し、所定の視覚情報としてもよい。例えば、音信号x(t)の音量が所定の閾値以上となったときに、それ以前に得ていた1種以上の視覚情報の中から1種類の視覚情報を選択し、所定の視覚情報としてもよい。選択の方法としては、(1-1)～(1-3)の方法を用いればよい。

【００２６】

（取得方法２）

認識部１３０は、所定の視覚情報を、映像コンテンツに重畳して表示するために予め用意されている1種類以上の視覚情報から得る。例えば、図４に示すような視覚情報データベースを予め用意しておき、(2)1種類以上の視覚情報の中からランダムに所定の視覚情報を選択する。なお、所定の視覚情報として、所定の行為、例えば、「笑い」を意味する情報のみを選択したい場合には、視覚情報データベースに「笑い」を意味する情報のみ、例えば、「w」「（笑）」「:-)」「(^o^)」等を用意しておけばよい。

【００２７】

（取得方法３）

（取得方法２）と、音信号x(t)がどのような音なのかを認識する処理を組合せてもよい。

【００２８】

(3-1)例えば、音信号x(t)に対して、VAD(voice activity detection:音声区間検出)を行い、音声区間と判定した場合には、「発話以外の音」が「笑い声」に対応すると判定し、笑い声に対応する視覚情報を、所定の視覚情報として選択する。また、非音声区間と判定した場合には、「発話以外の音」が「拍手音」に対応すると判定し、拍手に対応する視覚情報を、所定の視覚情報として選択する。この場合、所定の視覚情報が意味する行為が2種類以上ある場合には、視覚情報データベースには、視覚情報が音声区間に対応するものか、非音声区間に対応するものかを記憶しておく。なお、音声区間や非音声区間に対応する視覚情報が複数種類ある場合には、その中からランダムに所定の視覚情報を選択する。

【００２９】

(3-2)例えば、予め「発話以外の音」から音声特徴量を抽出しておき、音信号x(t)から抽出した音声特徴量との類似度を求め、類似度が所定の閾値以上となる場合に、その「発話以外の音」に対応する視覚情報を、所定の視覚情報として選択する。なお、笑い声や拍手等に対応する視覚情報が複数種類ある場合には、その中からランダムに所定の視覚情報を選択してもよい。なお、従来の音声認識装置では、発話から音声特徴量を抽出していたのに対し、本実施形態では「発話以外の音」から音声特徴量を抽出する。また、この場合、「発話以外の音」は視聴者の所定の行為(笑いや拍手)を意味し（所定の行為に対応し）、背景雑音等を含まない。

【００３０】

（取得方法４）

（取得方法１）と、（取得方法２）または（取得方法３）とを組合せてもよい。

【００３１】

認識部１３０は、(4-a)音信号の収音時刻に対応する映像コンテンツの時刻に、既に重畳されている視覚情報が1種類である場合、または、(4-b)複数種類あるがそのうち1種類の割合が極めて高い場合には、当該種類の視覚情報を所定の視覚情報として得る。

【００３２】

一方、(4-a)及び(4-b)以外の場合には、映像コンテンツに重畳して表示するために予め用意されている複数種類の視覚情報から所定の視覚情報を得る。

【００３３】

例えば、音信号x(t)の音量が所定の閾値以上となる時刻において、（または「音信号x(t)の音量が所定の閾値以上となる時刻までに、」）映像コンテンツに既に重畳されている

視覚情報が1種類か、2種類以上かを判定する。1種類の場合には、その視覚情報を所定の視覚情報として得る。2種類以上の場合には、重畳されている数が最も多い種類の視覚情報の割合を求め、その割合が所定の閾値(例えば0.5)より大きいときに、その視覚情報を所定の視覚情報として選択する。重畳されている数が最も多い種類の視覚情報の割合が所定の閾値以下のときに、（取得方法２）または（取得方法３）の方法により、所定の視覚情報を選択する。

【００３４】

このような取得方法により、発話以外の音を、そのとき表示部に出ている視覚情報の中で多数を占める視覚情報に変換して画面に表示することができる。

（取得方法５）

（取得方法１）と、（取得方法２）または（取得方法３）との組合せとしては以下のような方法も考えられる。

【００３５】

認識部１３０は、(5)音信号の収音時刻に対応する映像コンテンツの時刻に、既に重畳されている視覚情報に占める1種類の視覚情報の割合が高い場合には、当該種類の視覚情報を所定の視覚情報として得ることを優先し、(5)以外の場合には、映像コンテンツに重畳して表示するために予め用意されている複数種類の視覚情報から所定の視覚情報を得ることを優先することで、所定の視覚情報を得る。

【００３６】

例えば、音信号x(t)の音量が所定の閾値以上となる時刻において、（または「音信号x(t)の音量が所定の閾値以上となる時刻までに、」）映像コンテンツに既に重畳されている視覚情報の種類毎にそれぞれの割合を求め、その割合が所定の閾値a(例えばa>0.5)より大きいときに、所定の確率b(0.5<b<1)でその割合に対応する視覚情報を所定の視覚情報として選択し、(1-b)の確率で、（取得方法２）または（取得方法３）の方法により、所定の視覚情報を選択する。一方、その割合が所定の閾値a以下のときに、所定の確率c(0.5<c<1)で、（取得方法２）または（取得方法３）の方法により、所定の視覚情報を選択し、(1-c)の確率でその割合に対応する視覚情報を所定の視覚情報として選択する。

【００３７】

＜表示サイズ取得部１４０＞

表示サイズ取得部１４０は、収音された少なくとも発話以外の音を含む音信号x(t)を受け取り、その大きさを表す指標（例えば、音量、パワー、エネルギー）を求め、大きさを表す指標と所定の閾値との大小関係に基づき、無音か否かを判定する。例えば、大きさを表す指標が音が大きいほど大きくなる値（例えば、音量、パワー、エネルギー）の場合には、大きさを表す指標が所定の閾値未満のときに無音と判定し、閾値以上のときに何らかの音を収音できたと判定する。また、大きさを表す指標が音信号が大きいほど小さくなる値の場合には、大きさを表す指標が所定の閾値より大きいときに無音と判定し、閾値以下のときに何らかの音を収音できたと判定する。さらに、無音ではなく、何らかの音を収音できたと判定したときに、表示サイズ取得部１４０は、大きさを表す指標に対応する情報を、所定の視覚情報が視聴者端末９１の表示部または表示部１１０を介して表示される際の大きさの情報s(t)として得（Ｓ１４０）、通信回線を介して動画配信サーバ９２に送信する。例えば、音信号x(t)の大きさが大きいほど、表示される際の大きさが大きくなるように情報s(t)を取得する。

【００３８】

このような構成により、音量に合わせても文字の大きさを変えて表示部に表示することができ、より視聴者の雰囲気を詳細に伝えることができる。

【００３９】

＜効果＞

以上の構成により、コメント投稿ボタンのクリックまたはエンターキーの押下を行わずに、視覚情報を映像コンテンツに適切なタイミングで重畳して表示することができる。

【００４０】

＜変形例＞
本実施形態では、表示部１１０は、映像コンテンツと共にそれに重畳される視覚情報を表示しているが、映像コンテンツのみを表示する表示部を別途設けてもよい。
【００４１】
本実施形態の視聴者端末１００内に従来の音声認識装置を組み込んでもよい。例えば、認識部１３０の前段に従来の音声認識装置を組み込み、適切な音声認識ができなかった場合にのみ認識部１３０で認識処理を行う構成としてもよい。
【００４２】
本実施形態では、表示サイズ取得部１４０を備えるが、必ずしも備えなくともよい。なお、表示サイズ取得部１４０を備えない場合、視覚情報の大きさを表す情報として予めデフォルト値を設定しておけばよい。また、視覚情報の大きさは視聴者の操作により図示しない入力部から変更可能としてもよい。
【００４３】
本実施形態では、認識部１３０が、視聴者端末１００に組み込まれる構成としたが、独立した認識装置として構成してもよい。また、認識部１３０が、動画配信サーバ９２、または、視聴者端末１００以外の動画を再生する側の視聴者端末９１に組み込まれる構成としてもよい。その場合には、認識部１３０が組み込まれた装置に、音信号x(t)を送信する必要がある。データの伝送量を考慮すると、本実施形態のように、視聴者端末１００に認識部１３０が組み込まれ、視覚情報v(t)を送信する構成が望ましい。
【００４４】
なお、本実施形態では、表示部１１０において、視聴者に対して映像コンテンツを提示しているが、他のコンテンツを提示してもよい。端末は、対象者に対して何らかの刺激によってコンテンツを提示することができればよく、本実施形態のように音刺激及び光刺激による映像コンテンツを提示してもよいし、音刺激のみによる音響コンテンツ(ラジオ放送等)を提示してもよいし、対象者が持つ他の感覚器(触覚器、嗅覚器、味覚器)で受け取ることができる他の刺激(化学物質、温度、圧力)、または、各刺激の組合せによってコンテンツを提示してもよい。その場合であっても、表示部１１０は所定の視覚情報を表示するために用いる。なお、対象者が持つ感覚器で受け取ることができる刺激(光、音、化学物質、温度、圧力等)、または、それらの組合せによって提示されるコンテンツを纏めて「メディアコンテンツ」ともいい、メディアコンテンツを対象者に提示するための構成を提示部という。収音部１２０では、メディアコンテンツから刺激を感じ取った対象者がメディアコンテンツに対応して発する音を収音する。
【００４５】
視聴者端末１００は、図示しない表示長取得部を含んでもよい。表示長取得部は、収音された少なくとも発話以外の音を含む音信号x(t)を受け取り、その大きさを表す指標（例えば、音量の移動平均）を求め、大きさを表す指標と所定の閾値との大小関係に基づき、無音か否かを判定する。例えば、大きさを表す指標が音が大きいほど大きくなる値（例えば、音量、パワー、エネルギー）の場合には、大きさを表す指標が所定の閾値未満のときに無音と判定し、閾値以上のときに何らかの音を収音できたと判定する。また、大きさを表す指標が音信号が大きいほど小さくなる値の場合には、大きさを表す指標が所定の閾値より大きいときに無音と判定し、閾値以下のときに何らかの音を収音できたと判定する。表示長取得部は、無音ではなく、何らかの音を収音できたと判定したときに、その音の継続時間に対応する情報を、所定の視覚情報が視聴者端末９１の表示部または表示部１１０を介して表示される際の繰り返し回数として得、認識部１３０に出力する。例えば、音信号x(t)の継続時間が大きいほど、繰り返し回数が大きくなるような構成とする。認識部１３０は、求めた所定の視覚情報v(t)を繰り返し回数に応じて、繰り返した情報を、改めて所定の視覚情報として動画配信サーバ９２に出力する。例えば、所定の視覚情報v(t)が「ｗ」であり、繰り返し回数が5回のとき、改めて所定の視覚情報v(t)として「ｗｗｗｗｗ」、または、これと同じ意味を表すネットスラングである「草生えた」等を通信回線を介して動画配信サーバに送信する。このような構成により、より視聴者の雰囲気を詳細に伝

えることができる。なお、認識部１３０では、繰り返し回数を得てから、所定の視覚情報v(t)を出力するため、視覚情報v(t)を送信した際の再生時において、視覚情報v(t)を動画に重畳する場合、発話以外の音が発生してから、所定の視覚情報v(t)を出力するまでに、ズレが生じる。このズレをなくすために、送信した際の再生時においては、繰り返さずに所定の視覚情報v(t)を動画配信サーバ９２に送信してもよい。そして、動画配信サーバ９２は、視覚情報v(t)を受け取り、繰り返さずに動画に重畳して配信する。認識部１３０では、繰り返し回数を得てから、求めた所定の視覚情報v(t)を繰り返し回数に応じて、繰り返した情報を、改めて所定の視覚情報として動画配信サーバ９２に出力する。動画配信サーバ９２では、動画データベースに、動画と共に繰り返した視覚情報v(t)を格納する。このような構成により、動画配信サーバ９２では、繰り返した視覚情報v(t)を格納した後になされた再生要求に対して、繰り返した視覚情報を重畳した動画を配信することができる。

【００４６】

本実施形態では、認識部１３０において、受け取った音信号x(t)から、その大きさを表す指標を求め、大きさを表す指標と所定の閾値との大小関係に基づき、無音か否かを判定しているが、認識部１３０の前段に既存のVAD(voice activity detection)を設け、VADで音信号x(t)が無音か否かを判定してもよい。さらに、何らかの音を収音できたと判定したときに、音信号x(t)が認識部１３０に入力される構成とし、認識部１３０では、音信号x(t)を用いて、所定の視覚情報v(t)を得ればよい。

【００４７】

＜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。

【００４８】

＜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。

【００４９】

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。

【００５０】

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。

【００５１】

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。

【００５２】

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】

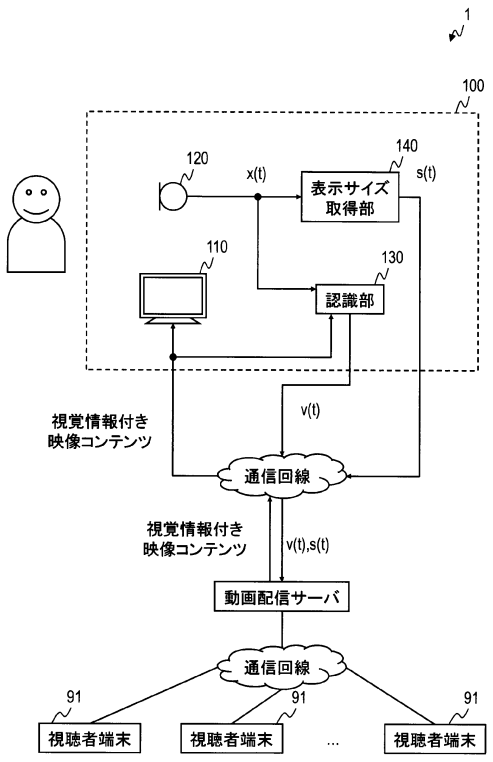

図1

【図２】

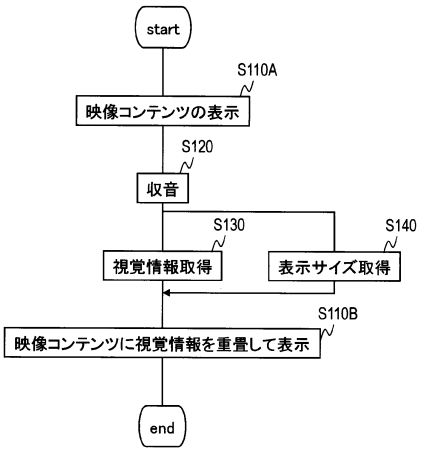

図2

【図３】

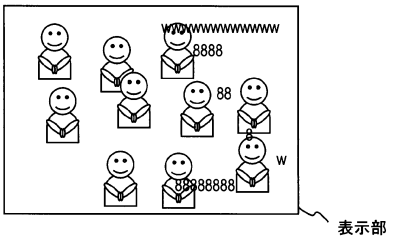

図3

【図４】

	視覚情報
1	「w」
2	「(笑)」
3	☺
4	「8」
5	
…	…

図4

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　富樫　明

(56)参考文献　特開２０１４－１４２９３４（ＪＰ，Ａ）
特開２００５－０６５２５２（ＪＰ，Ａ）
国際公開第２０１４／１９２４５７（WO，Ａ１）
特開平０８－００２０１５（ＪＰ，Ａ）
特開２００４－２５９３７５（ＪＰ，Ａ）
特開２０１２－０９００９１（ＪＰ，Ａ）
特開２００８－２７８１９９（ＪＰ，Ａ）
特開２０１３－０３７６７０（ＪＰ，Ａ）
“ニコニコ実況で新しいＴＶの楽しみ方が増えました。”，２０１２年　１月　８日，[平成30年3月7日検索]，インターネット<URL: https://zigsow.jp/review/108/167745>

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｈ０４Ｎ　２１／００－２１／８５８
Ｇ０６Ｆ　１３／００
Ｇ１０Ｌ　１５／００