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

特許第4599606号
(P4599606)

(45)発行日　平成22年12月15日(2010.12.15)　(24)登録日　平成22年10月8日(2010.10.8)

(51)Int.Cl.		F I		
G06N 3/00	(2006.01)	G06N	3/00	550G
G10L 13/00	(2006.01)	G10L	13/00	100V
G10L 15/00	(2006.01)	G10L	15/00	200H

請求項の数 5　(全 16 頁)

(21)出願番号　特願2005-218476 (P2005-218476)
(22)出願日　平成17年7月28日 (2005.7.28)
(65)公開番号　特開2007-34788 (P2007-34788A)
(43)公開日　平成19年2月8日 (2007.2.8)
審査請求日　平成19年5月30日 (2007.5.30)

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(74)代理人　100099933
弁理士　清水　敏
(72)発明者　川本　真一
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　四倉　達夫
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　中村　哲
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内

最終頁に続く

(54)【発明の名称】頭部動作自動生成のための頭部動作学習装置及び頭部動作合成装置並びにコンピュータプログラム

(57)【特許請求の範囲】

【請求項１】

発話と、発話に伴う発話主体の頭部の動きとの関係を機械学習により学習するための頭部動作学習装置であって、

発話時の前記発話主体の所定の感情に関する強度を示す感情強度パラメータの入力をユーザより受けるための感情強度入力手段と、

各発話の発話時の前記発話主体の音声から時系列として抽出される所定の音響特徴量と、当該発話に関して、前記感情強度入力手段を介して入力された前記感情強度パラメータと、当該発話時の前記発話主体の頭部の動きを示す情報とから、前記所定の音響特徴量及び前記感情強度パラメータと、前記発話主体の頭部の動きとの間の関係を学習するための学習手段とを含み、

前記学習手段は、

前記発話主体の発話時の音声を受けて、発話開始時から所定時間ごとに当該音声の音響特徴量を抽出するための音響特徴量抽出手段と、

前記音響特徴量抽出手段により抽出される音響特徴量に、当該発話の発話開始からの時間を示す情報を付すための時間情報付与手段と、

前記発話主体の、発話時の頭部の位置又は向きを時刻と対応させて示す情報を取得するための頭部位置情報取得手段と、

ある発話について、前記音響特徴量抽出手段により抽出され、前記時間情報付与手段により発話からの時間情報が付与された音響特徴量と、当該発話に関して前記感情強度入力

手段により入力された感情強度パラメータと、前記頭部位置情報取得手段により取得された前記発話主体の頭部の位置又は向きとを同期させて学習用データを生成するための同期手段と、

前記同期手段によって生成された学習用データを用い、前記ある発話の発話開始からの時刻が付与された前記所定の音響特徴量の時系列及び前記感情強度パラメータと前記発話主体の頭部の位置又は向きとの間の関係を学習するためのニューラルネットワークとを含む、頭部動作学習装置。

【請求項２】

発話から発話に伴う発話主体画像の頭部の動きを合成するための頭部動作合成装置であって、

発話時の発話主体の音声から抽出された所定の音響特徴量の時系列と、発話に関して指定された感情強度パラメータとが与えられると、当該音声の発話時の発話主体の頭部の動きを推定するための頭部位置推定手段と、

音声から前記所定の音響特徴量の時系列を抽出するための音響特徴量抽出手段と、

前記音響特徴量抽出手段により抽出された前記所定の音響特徴量の前記時系列と、指定された感情強度パラメータとを前記頭部位置推定手段に与える事により、前記音声に同期した、前記発話主体の頭部の動作に関する情報を前記頭部位置推定手段からの一連の出力として得るための頭部動作生成手段とを含み、

前記頭部位置推定手段は、発話主体の発話時の音声から抽出された所定の音響特徴量の時系列と、当該発話時の前記発話主体の前記所定の感情に関して指定された感情強度パラメータと、当該発話時の前記発話主体の頭部の動きを示す情報とから、前記所定の音響特徴量の時系列及び前記感情強度パラメータと、発話主体の頭部の位置又は向きとの間の関係を予め学習済のニューラルネットワークを含む、頭部動作合成装置。

【請求項３】

音声を予め格納し、前記音響特徴量抽出手段に与えるための手段をさらに含む、請求項２に記載の頭部動作合成装置。

【請求項４】

ユーザにより入力された感情強度に対応する感情強度パラメータを前記頭部動作生成手段に与えるための手段をさらに含む、請求項２又は請求項３に記載の頭部動作合成装置。

【請求項５】

コンピュータにより実行されると、当該コンピュータを請求項１～４のいずれかに記載の装置として動作させる、コンピュータプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

この発明は、合成された顔画像の頭部動作を音声に合わせて自動的に生成する装置に関し、特に生成される頭部動作によって表現したい感情の強度をユーザが自由にカスタマイズする事ができる頭部動作学習装置及び頭部動作合成装置並びにそれらのためのコンピュータプログラムに関する。

【背景技術】

【０００２】

近年、ＣＧによって作られたキャラクタを使用したアニメーションの製作が盛んになっている。そこで、そのようなアニメーション製作に関する技術の需要は年々高まり、その技術の進歩も大いに期待されている。

【０００３】

アニメーションの製作においては、音声に同期した画像を作成する事が必要である。通常は、先にアニメーション画像を作成し、後に画像に合わせて音声の吹込みを行なう。しかし、アニメーション画像の作成を先に行なうためには、物語の進行に合わせた想像力が必要とされる。また、画像に合わせて音声を吹込むためには、音声を担当するものの技術が必要である。一方で、先にキャラクタの音声の吹込みを台本に従って行ない、その音声

に基づいてキャラクタのアニメーション画像を自動的に生成する手法も存在する。この場合、音声から画像が自動的に生成されるので、音声と画像が自然に同期する事になり、質の高いアニメーション製作を行なう事ができる。また、物語の進行に合わせ、キャラクタの音声を予測しながらそれに合わせて画像を作成する必要がないので、アニメーション製作を効率的に行なう事もできる。

【０００４】

アニメーションで問題となる画像の動きの重要なものにキャラクタの動作がある。このキャラクタの動作としては、表情、ジェスチャ、頭部動作等がある。これらは、画像を見た者に、そのキャラクタによって表現しようとする感情の理解を容易にする事ができる重要な要素である。この様に感情の理解を容易にするために画像の動きを使用するにあたっては、その動きが感情を理解するために十分な助けとなるものである必要がある。だとすれば、画像によって表現しようとしている感情がキャラクタの動作を見た者に自然に理解できるものである事が望ましい。

【０００５】

キャラクタの動作には、前述した様に、表情、ジェスチャ、頭部動作等がある。このうち、表情、ジェスチャ等はある感情に対応する動作に見られる個人差が比較的大きい上に発話意図にも大きく依存する。それゆえ、ある音声からその音声に対応する自然でかつ、誰が見ても個々のキャラクタの感情を容易に推測する事のできる表情、ジェスチャ等を自動生成する事は難しい。つまり、ある表情やジェスチャ等が必ずしもすべてのキャラクタに適用できるとは限らない。一方で、頭部動作については、他の動作に比べ、同じ動作が他のキャラクタにもある程度、違和感なく適用できる。そこで、音声からキャラクタの動作を自動生成するにあたっては、頭部動作を生成する方法を採用して、画像を見た者がキャラクタの感情を適切に理解する事ができる様にする事が望ましい。

【０００６】

この様に、音声から、自然な頭部動作を自動的に生成する際には、予め学習されたパラメータを使用するという方法が考えられる。この学習にあたっては、音声と動作とを同時に収集してそのデータを集積し、いかなる音声からいかなる動作が生じるのが妥当であるかという関係を学習する方法が採用される。このような学習から得られた関係を使用して音声から頭部動作を生成する方法には、非特許文献１に示される様にニューラルネットワークを用いたものがある。ニューラルネットワークは非線型的な手法であり、実際の人間の頭部動作に良く似た動きを生成する事ができる。

【０００７】

図１に、非特許文献１に開示の従来技術による頭部動作の自動生成システムについて示す。図１を参照して、この頭部動作自動生成システムは、学習のための発話者の頭部動作３０を撮影して、学習のための頭部動作データを収集するためのカメラ３６と、発話者の発話音声３２を録音するためのマイクロフォン３４と、マイクロフォン３４によって録音された学習のための音声を格納するための音声格納部３８と、カメラ３６によって収集された頭部動作に関するデータを格納するための頭部動作格納部４０とを含む。録音音声と頭部動作に関するデータとの双方には、時刻情報が記録されている。この時刻情報は、録音機器と録画機器とで共有されており、従って音声と頭部画像との対応関係をとる事が可能である。

【０００８】

頭部動作自動生成システムはさらに、時刻を共有する音声から音響特徴量を算出し、この特徴量データと頭部動作に関するデータとから、ニューラルネットワーク学習のための学習データ（これを「音声－頭部動作データ」と呼ぶ。）を作成するための音声－頭部動作同期部４２と、音声－頭部動作同期部４２の作成した音声－頭部動作データを格納するための音声－頭部動作データ格納部４４と、格納された音声－頭部動作データを使用して、所定のニューラルネットワークに音声とそれに同期する頭部動作との関係について学習を行なわせるための学習部４６と、その学習によって得られたニューラルネットワークのパラメータを格納するためのニューラルネットワークパラメータ格納部４８とを含む。

【０００９】
頭部動作自動生成システムはさらに、予め台本に従い録音される、キャラクタの音声５４に使用される音声を格納するための音声格納部５２と、ニューラルネットワークパラメータ格納部４８に格納されたニューラルネットワークパラメータを使用して、音声格納部５２に格納された音声に基づいて、キャラクタの頭部動作５６を生成するための頭部動作生成部５０とを含む。
【００１０】
この頭部動作自動生成システムにおいては、発話者の発話音声３２がマイクロフォン３４によって録音され、音声格納部３８に格納される。一方、カメラ３６で記録された発話者の頭部動作データは、頭部動作格納部４０に格納される。両者には共通した時刻情報が含まれている。
【００１１】
音声－頭部動作同期部４２が、音声格納部３８に格納された音声と頭部動作格納部４０に格納された頭部動作に関するデータとのうち、時刻情報を共有するデータを用いて、所定のニューラルネットワークの学習のための音声－頭部動作データを生成する。その生成された音声－頭部動作データは音声－頭部動作データ格納部４４に格納される。この音声－頭部動作データを用いて、学習部４６で、音声とそれに対応する頭部動作との間の関係をニューラルネットワークに学習させる。その学習によって得られた、ニューラルネットワークの動作を規定するパラメータ（ニューラルネットワークパラメータ）がニューラルネットワークパラメータ格納部４８に格納される。
【００１２】
このニューラルネットワークパラメータを用いて、頭部動作生成部５０で音声から頭部動作の自動生成が行なわれる。具体的には、まず、台本に基づいてキャラクタの音声が録音され、音声格納部５２に格納される。一方、ニューラルネットワークパラメータ格納部４８に格納されたパラメータにより、ニューラルネットワークを予め設定しておく。音声格納部５２に格納された音声から、学習時に音声－頭部動作同期部４２が算出したものと同じ種類の音響特徴量を算出する。この音響特徴量を入力としてニューラルネットワークに与える事により、その出力として入力音声に対応する頭部の動き（頭部特徴点の座標）がニューラルネットワークから出力される。この値を各フレームで算出する事により、キャラクタの頭部動作５６が自動生成される。
【００１３】
この様に、最初に台本に基づいてキャラクタの音声を録音し、その後に音声に同期した頭部動作が生成される。この頭部動作は、学習に基づいてニューラルネットワークにより生成されるので、自然なものとなる。それゆえ、画像を見た者にとっては頭部動作で表わされたキャラクタの感情が理解しやすくなるし、ユーザにとってはアニメーション製作を効率的に行なう事ができるという利点がある。
【非特許文献１】川本真一、松下義則、中井満、下平博、嵯峨山茂樹、「擬人化音声対話エージェントのための発話時の頭部挙動の自動生成」、日本音響学会誌、２００２年秋。
【発明の開示】
【発明が解決しようとする課題】
【００１４】
非特許文献１に開示の技術における様に、音声から算出される音響特徴量から全自動で頭部動作を生成すると、画像生成の効率性の点及び音声に同期した自然な頭部動作を生成できるという点についての問題はない。しかし、音声から算出される音響特徴量から頭部動作の全自動生成を行なうと、ユーザであるアニメーションのクリエータの感性を反映する余地のない頭部動作が生成される。すなわち、キャラクタの頭部動作は学習時に採取された音声と台本に従って録音された音声との音響特徴量によって制限されてしまう事になり、多種多様なキャラクタの個性を頭部動作によって表現する事が難しくなる。
【００１５】
そこで、本発明の目的は、ユーザの感性を反映する事のできる頭部動作自動生成のため

の頭部動作学習装置及び頭部動作合成装置を提供する事である。

【課題を解決するための手段】

【００１６】

本発明の第１の局面に係る頭部動作学習装置は、発話時の発話主体の所定の感情に関する強度を示す感情強度パラメータの入力をユーザより受けるための感情強度入力手段と、各発話の発話時の発話主体の音声から時系列として抽出される所定の音響特徴量と、当該発話に関して、感情強度入力手段を介して入力された感情強度パラメータと、当該発話時の発話主体の頭部の動きを示す情報とから、所定の音響特徴量及び感情強度パラメータと、発話主体の頭部の動きとの間の関係を学習するための学習手段とを含む。

【００１７】

この頭部動作学習装置によると、上記した関係の学習にあたって、ユーザが任意の感情強度パラメータを入力する事ができる。ユーザが必要だと考える感情強度パラメータと音響特徴量と、頭部動作との間の関係を学習させる事ができる。学習後は、任意の値の感情強度パラメータに対し、上記した関係にもとづいて、妥当な頭部の動作を生成できる。従って、ユーザの感性を頭部動作学習結果に反映するための適切な頭部動作学習装置を提供する事ができる。

【００１８】

好ましくは、学習手段は、発話主体の発話時の音声を受けて、発話開始時から所定時間ごとに当該音声の音響特徴量を抽出するための音響特徴量抽出手段と、音響特徴量抽出手段により抽出される音響特徴量に、当該発話の発話開始からの時間を示す情報を付すための時間情報付与手段と、発話主体の、発話時の頭部の位置又は向きを時刻と対応させて示す情報を取得するための頭部位置情報取得手段と、ある発話について、音響特徴量抽出手段により抽出され、時間情報付与手段により発話からの時間情報が付与された音響特徴量と、当該発話に関して感情強度入力手段により入力された感情強度パラメータと、頭部位置情報取得手段により取得された発話主体の頭部の位置又は向きとを同期させて学習用データを生成するための同期手段と、同期手段によって生成された学習用データを用いて、音響特徴量の時系列及び感情強度パラメータと、発話主体の頭部の位置又は向きの変化との間の関係を学習するための手段とを含む。

【００１９】

この頭部動作学習装置によると、音声から抽出された音響特徴量に時間情報を付与し、それと発話開始からの時刻に対応させた頭部の位置又は向きと感情強度入力部によって入力された感情強度とを同期させて学習用データを生成し、そのデータを元に音響特徴量の時系列及び感情強度パラメータと発話主体の頭部の位置又は向きの変化との間の関係を学習する。その結果、音声と同期した頭部動作の関係を学習するための適切な頭部動作学習装置を提供する事ができる。

【００２０】

さらに好ましくは、学習するための手段は、同期手段によって生成された学習用データを用い、ある発話の発話開始からの時刻が付与された所定の音響特徴量の時系列及び感情強度パラメータと発話主体の頭部の位置又は向きとの間の関係を、所定の非線型関数近似により学習するための関数近似学習手段を含む。

【００２１】

この頭部動作学習装置によると、非線型関数近似により音響特徴量の時系列及び感情強度パラメータと発話主体の頭部の位置又は向きとの間の関係を学習する。従って、線型関数近似では表わせない頭部の位置又は向きの非線型な変化を学習する事ができる。その結果、より自然な頭部動作に近い頭部動作の学習をするための適切な頭部動作学習装置を提供する事ができる。

【００２２】

さらに好ましくは、関数近似学習手段は、同期手段によって生成された学習用データを学習用データとして、音響特徴量及び感情強度パラメータと発話主体の頭部の位置又は向きとの間の関係を学習するためのニューラルネットワークを含む。

【００２３】
この頭部動作学習装置によると、ニューラルネットワークを用いた非線型関数近似により音響特徴量の時系列及び感情強度パラメータと発話主体の頭部の位置又は向きとの間の関係を学習する。従って、ニューラルネットワークを用いない線型関数近似では学習する事のできない頭部の位置又は向きの非線型な変化を学習するための適切な頭部動作を学習する事ができる。その結果、より自然な頭部動作に近い頭部動作の学習をするための適切な頭部動作学習装置を提供する事ができる。
【００２４】
本発明の第２の局面に係る頭部動作合成装置は、発話時の発話主体の音声から抽出された所定の音響特徴量の時系列と、発話に関して指定された感情強度パラメータとが与えられると、当該音声の発話時の発話主体の頭部の動きを推定するための頭部位置推定手段と、音声から所定の音響特徴量の時系列を抽出するための音響特徴量抽出手段と、音響特徴量抽出手段により抽出された所定の音響特徴量の時系列と、指定された感情強度パラメータとを頭部位置推定手段に与える事により、音声に同期した、発話主体の頭部の動作に関する情報を頭部位置推定手段からの一連の出力として得るための頭部動作生成手段とを含む。
【００２５】
この頭部動作合成装置によると、音声を入力すると、その音声に含まれた情報から自動的に画像の頭部動作が合成される。従って、音声に同期した自然な頭部動作を効率的に合成するための適切な頭部動作合成装置を提供する事ができる。
【００２６】
好ましくは、頭部動作合成装置は、音声を予め格納し、音響特徴量抽出手段に与えるための手段をさらに含む。
【００２７】
この頭部動作合成装置によると、音声を予め録音し、それを格納する事ができる。従って、以前に録音された音声を元に画像の頭部動作を合成するための適切な頭部動作合成装置を提供する事ができる。
【００２８】
さらに好ましくは、この頭部動作合成装置は、ユーザにより入力された感情強度に対応する感情強度パラメータを頭部動作生成手段に与えるための手段をさらに含む。
【００２９】
この頭部動作合成装置によると、ユーザが任意に感情強度を入力する事ができる。従って、ユーザの感性を、合成される頭部動作に適切に反映できる頭部動作合成装置を提供する事ができる。
【００３０】
さらに好ましくは、頭部位置推定手段は、予め、発話主体の発話時の音声から抽出された所定の音響特徴量の時系列と、当該発話時の発話主体の所定の感情に関して指定された感情強度パラメータと、当該発話時の発話主体の頭部の動きを示す情報とから、所定の音響特徴量の時系列及び感情強度パラメータと、発話主体の頭部の位置又は向きとの間の関係を予め学習済の機械学習手段を含む。
【００３１】
この頭部動作合成装置によると、機械学習手段が、音響特徴量の時系列と、感情強度パラメータと、発話主体の頭部の動きを示す情報とから、これらの間の関係を予め学習する。従って、ユーザが音響特徴量と感情強度パラメータとを機械学習手段に入力して、入力に対応する頭部動作を得る事ができる。
【００３２】
さらに好ましくは、機械学習手段は、発話主体の発話時の音声から抽出された所定の音響特徴量の時系列と、当該発話時の発話主体の所定の感情に関して指定された感情強度パラメータと、当該発話時の発話主体の頭部の動きを示す情報とから、所定の音響特徴量の時系列及び感情強度パラメータと、発話主体の頭部の位置又は向きとの間の関係を予め非

線型関数近似により学習済の関数近似学習手段を含む。

【００３３】

この頭部動作合成装置によると、頭部動作合成の際に非線型関数近似による頭部動作の合成が可能になる。従って、線型関数近似によると表現できない様な非線型な頭部動作の合成をするための適切な頭部動作合成装置を提供する事ができる。

【００３４】

さらに好ましくは、関数近似学習手段は、発話主体の発話時の音声から抽出された所定の音響特徴量の時系列と、当該発話時の発話主体の所定の感情に関して指定された感情強度パラメータと、当該発話時の発話主体の頭部の動きを示す情報とから、所定の音響特徴量の時系列及び感情強度パラメータと、発話主体の頭部の位置又は向きとの間の関係を予め学習済のニューラルネットワークを含む。

【００３５】

この頭部動作合成装置によると、頭部動作合成の際にニューラルネットワークによる非線型関数近似による頭部動作の合成が可能になる。従って、ニューラルネットワークを用いない線型関数近似によると表現できない様な非線型な頭部動作の合成をするための適切な頭部動作合成装置を提供する事ができる。

【００３６】

本発明の第３の局面に係るコンピュータプログラムは、コンピュータにより実行されると、当該コンピュータを上記のいずれかに記載の装置として動作させるものである。従って上述した頭部動作学習装置及び頭部動作合成装置のいずれかと同様の効果を得る事ができる。

【発明を実施するための最良の形態】

【００３７】

以下、図面を参照し発明の実施の一形態を説明する。本実施の形態は、音声の音響特徴量だけでなく、ユーザによる指示を反映した形で頭部動作を生成する装置に関するものである。

【００３８】

＜構成＞

図２に、本発明の実施の一形態に係る頭部動作の自動生成システムのブロック図を示す。図２を参照して、この頭部動作自動生成システムは、学習用データの発話時の発話者の頭部動作６０を撮影して、頭部動作データを収集するためのカメラ６８と、学習用データの発話者の発話音声６２を録音するためのマイクロフォン６６と、感情の強さを明示するためにユーザにより指定される感情強度の入力を受けるための感情強度入力部６４と、マイクロフォン６６によって録音された音声と感情強度入力部６４を介して入力された感情強度のパラメータとを格納するための音声格納部７０と、カメラ６８によって収集された頭部動作に関するデータを格納するための頭部動作格納部７２とを含む。音声格納部７０の格納する音声データと、頭部動作格納部７２の格納する頭部動作に関するデータとはいずれもフレーム化されており、共通した時刻情報を含んでいる。従って、両者を同期させる事が可能である。

【００３９】

なお、発話者は、学習用の発話を行なう際には、感情強度入力部６４で入力された感情強度に対応した形で、感情を込めて行なう。感情強度入力部６４からは、各発話について指定された感情強度を示すパラメータが入力される。なお、本実施の形態では感情強度としては「感情なし（通常）」、「感情を含む」、及び「非常に強い感情を含む」という三つの段階を使用する。これらは本実施の形態ではそれぞれ「０」、「０．５」及び「１」という値で指定され音声格納部７０に格納される。

【００４０】

このシステムはさらに、音声格納部７０に格納された音声から所定の音響特徴量を算出し、頭部動作格納部７２に格納された頭部動作に関するデータ及び音声格納部７０に格納された、当該発話に対して指定された感情強度とともにニューラルネットワークの学習の

ためのデータ（音声－頭部動作データ）を生成するための音声－頭部動作同期部７４と、その学習のための音声－頭部動作データを格納するための音声－頭部動作データ格納部７６と、格納された音声－頭部動作データを使用してニューラルネットワークに、音声と、それに対応する感情強度と、対応する頭部動作との間の関係についての学習を行なわせるための学習部７８と、その学習によって得られたニューラルネットワークパラメータを格納するためのニューラルネットワークパラメータ格納部８０とを含む。

【００４１】

このシステムはさらに、アニメーションにあるキャラクタについて、担当の声優により予め台本に基づいて録音された音声を格納するための音声格納部８６と、台本に基づき、かつユーザの判断に従って、キャラクタの発話時の感情強度パラメータをユーザが入力するために使用する感情強度設定部８４とを含む。感情強度設定部８４により設定される感情強度パラメータは、学習時と同様、「０」、「０．５」、及び「１」であるものとする。

【００４２】

このシステムはさらに、ニューラルネットワークパラメータ格納部８０に格納されたニューラルネットワークパラメータを使用して、音声格納部８６に格納された音声から得られる音響特徴量と、感情強度設定部８４により設定された感情強度パラメータとに基づいてキャラクタの頭部動作９０を生成するための頭部動作生成部８２とを含む。

【００４３】

ここで、感情強度とは、感情の強さによって頭部動作を制御するための要素である。本実施の形態では、「怒り感情強度」を扱うものとする。怒りという感情に関しては、通常状態（まったく怒っていない状態）～怒り～激怒の様な感情強度を考える事ができる。前述した様に「通常状態」を感情強度値（ＥＦ値）０で、「怒り」をＥＦ値０．５で、「激怒」をＥＦ値１．０で表わす。ユーザはこの感情強度（ＥＦ）の値を任意に選択する事により、キャラクタの多種多様な感情の強さを状況に応じて自由に表現する事ができる。例えば、ＥＦ値＝０．２を選択して「弱い怒り」を表現する事ができる。また、ＥＦ値は１．０以上を選択する事も可能であるので、例えばＥＦ値１．５等を選択して「あり得ないほどの激怒」等を表現する事もできる。

【００４４】

また、この感情強度は怒り感情強度に限られない。例えば、通常状態～喜び～歓喜といった喜び感情強度を設定する事も可能である。

【００４５】

図３に頭部動作生成部８２の詳細なブロック図を示す。図３を参照して、この頭部動作生成部８２は、入力された音声の音声信号を音素に分解し、入力音声に対応する音響特徴量及び音素列を出力するための音声認識部１１０と、音声認識部１１０より出力された音響特徴量及び音素列を格納するための音声情報格納部１１２と、ニューラルネットワークパラメータ格納部１１６に格納されたパラメータにより予め設定されたニューラルネットワークと、音声情報格納部１１２に格納された音響特徴量と、ユーザによって設定された感情強度とから頭部動作パラメータを生成するための頭部動作自動生成部１１４と、頭部動作自動生成部１１４により生成された頭部動作パラメータを格納するための頭部動作パラメータ格納部１１８と、頭部動作パラメータ格納部１１８に格納された頭部動作パラメータを使用して画像の頭部動作を合成するための頭部動作合成部１２０とを含む。

【００４６】

図４に、図３に示す頭部動作自動生成部１１４のブロック図を示す。図４を参照して、この頭部動作自動生成部１１４は、音声情報格納部１１２に格納された各フレームの音響特徴量から、声の高さに相当する基本周波数（Ｆ０）１７０、声の大きさに相当するパワー１７２、及び発話時間情報１７４を抽出するとともに、算出したフレームを含む直前１１フレームの音響特徴量を記憶するための特徴量抽出部１７６と、特徴量抽出部１７６に記憶された直前の１１フレーム分の音響特徴量と、感情強度設定部８４を介してユーザによって設定された感情強度１７８とを入力として受け、キャラクタの頭部動作に関する情

報である頭部回転角Ｒｘ１８２、Ｒｙ１８４、Ｒｚ１８６を出力するためのニューラルネットワーク１８０とを含む。頭部回転角と発話時間情報の詳細については後述する。ニューラルネットワーク１８０は、ニューラルネットワークパラメータ格納部８０に記憶されたニューラルネットワークパラメータを用いて予め設定される。従ってニューラルネットワーク１８０は、図２に示す学習部７８による学習に従い、入力される音響特徴量及び感情強度パラメータに従った頭部動作に関する情報を出力する事が可能である。

【００４７】

図５に、図２に示す学習部７８が行なう処理であるニューラルネットワークの学習に関係する処理の詳細について示す。図５を参照して、まず、処理１３４では学習用のテキスト１３２に基づいて学習用の音声の収録が行なわれる。収録された音声は記憶装置１３６に格納される。

【００４８】

処理１４２は記憶装置１３６に格納された音声を音素に分割する。分割された音素は処理１５２に送られる。この音素は、頭部動作データの作成ではなく、キャラクタの口の動きの合成に使用される。

【００４９】

処理１３８では、記憶装置１３６に格納された音声を所定のフレーム長でかつ所定のシフト長のフレームにフレーム化する。フレーム化された音声は処理１４０と処理１４４とに送られる。

【００５０】

処理１４４では、処理１３８によって得られたフレームから発話区間を検出する。発話区間の検出には、種々の手法を用いる事ができる。学習用のテキスト１３２の録音は通常はスタジオで行なわれるので、発話区間と無音区間との識別は容易である。検出された発話区間を特定する情報は処理１５２に送られる。

【００５１】

処理１４０では、フレーム化された音声の各フレームから、基本周波数及びパワーを含む音響特徴量を算出する。算出された音響特徴量は処理１５２に送られる。

【００５２】

次に、処理１３０では、ユーザによって任意の感情強度が設定される。設定された感情強度パラメータは処理１５２に送られる。

【００５３】

一方、処理１４８では、学習用のテキスト１３２を発話する際の発話者の頭部動作データを、カメラ１４６を用いたモーションキャプチャによって収集する。モーションキャプチャによって得られた頭部動作データは、記憶装置１５０に格納される。記憶装置１５０に格納された頭部動作データは、処理１５２に与えられる。

【００５４】

処理１５２では感情強度、音響特徴量、発話区間、音素、及び、頭部動作データを参照して、音声と頭部動作との同期を行なった上で学習用のデータを作成する。すなわち、各フレームの音響特徴量と、指定された感情強度パラメータと、当該フレームに対応する頭部動作データとから学習用の音声－頭部動作データを作成する。この学習用データは記憶装置１５４に与えられる。

【００５５】

この記憶装置１５４に格納された音声－頭部動作データに基づいてニューラルネットワークの学習を行なう事により、音声から頭部動作を自動生成する際に使用されるニューラルネットワークパラメータが得られる。

【００５６】

図６に発話時間情報の詳細について示す。発話時間情報とは、話し始めに頭部を上げて話し終わりに頭部を下げるというような、発話時に一般的に見られる発話経過時間に関連すると思われる頭部動作を表現するために音響特徴量の一つとして取り入れたものである。

【００５７】
図６を参照して、縦軸に発話時間情報をとり、横軸に発話開始から終了までの時間をとる。発話時間情報が０であるとは、発話の開始時又は発話がなされていない状態を示す。発話時間情報が１であるとは、発話の終了時を示す。
【００５８】
図７に頭部回転角の詳細を示す。図７を参照して、顔画像中に示す様に、頭部回転角Ｒｘ１６０、Ｒｙ１６２、Ｒｚ１６４はそれぞれ予め定められた３次元座標の３軸（ｘ軸、ｙ軸、及びｚ軸）周りの回転角度を表わす。Ｒｘは頭部の上下方向の動き、すなわちうなずいたり上を向いたりするような動きに用いるための角度である。Ｒｙは左右方向に首をかしげるような動きに用いるための角度である。Ｒｚは左右方向に顔を向ける動きに用いるための角度である。この３軸の回転角を組み合わせる事によって３次元的な回転により、頭部動作の表現が可能になる。
【００５９】
なお、本発明の実施の形態では頭部動作として頭部角度による頭部の向き（回転）を例にとって説明しているが、頭部の位置、つまり、頭部が並進するような動きに基づく頭部動作を合成する事も可能であり、さらには並進運動と回転運動とが組み合わされた頭部動作を合成する事もできる。
【００６０】
＜動作＞
上記した本実施の形態に係る頭部動作自動生成システムは以下の様に動作する。このシステムの動作には三つのフェーズがある。第１のフェーズは学習フェーズであり、第２のフェーズは台本に基づく音声の録音フェーズであり、第３のフェーズは録音された音声に基づき、キャラクタの頭部動作データを作成するフェーズである。アニメーション作成システム全体としては、これ以外にキャラクタをデザインし、アニメーションの表情を作成したりする処理があるが、それらについては本願発明とは関係がないのでここでは説明は省略する。
【００６１】
図２を参照して、学習時には、この頭部動作自動生成システムにおいては、学習時の発話者の発話音声６２がカメラ６８と同期したマイクロフォン６６によって録音され、音声格納部７０に格納される。一方、カメラ６８で記録された発話者の頭部動作データは、頭部動作格納部７２に格納される。この際、両方のデータには共通の時間情報が付される。発話者に対しては所定の感情強度（本実施の形態の場合には、怒りに関して、「通常」、「怒り」、及び「激怒」という３種類）のいずれかに合わせて発話する様に指示が出される。そして、感情強度入力部６４によって、その感情強度を示す感情強度パラメータ（「０」、「０．５」及び「１」のいずれか）がユーザにより手入力される。この感情強度パラメータは対応する発話音声とともに音声格納部７０に格納される。
【００６２】
音声格納部７０に格納された音声及び感情強度と、頭部動作格納部７２に格納された頭部動作に関するデータとはともにフレーム化され、同じ時刻のフレームに対応する音声データ及び感情強度パラメータと頭部動作パラメータとが、音声－頭部動作同期部７４によって一つのデータにまとめられる。こうして学習用の音声－頭部動作データが作成され、音声－頭部動作データ格納部７６に格納される。
【００６３】
続いて、音声－頭部動作データ格納部７６に格納された音声－頭部動作データを用いて、学習部７８で、音声及び感情強度とそれに対応する頭部動作との関係をニューラルネットワークに学習させる。その学習によって得られたニューラルネットワークパラメータがニューラルネットワークパラメータ格納部８０に格納される。
【００６４】
台本の録音時には、声優が、台本を見ながら所定のキャラクタの台詞を発話する。この音声は録音され、図２に示す音声格納部８６に格納される。

【００６５】
頭部動作作成時には、学習時に得られたニューラルネットワークパラメータを用いて、頭部動作生成部８２で音声から頭部動作の自動生成が行なわれる。具体的には、まず、ニューラルネットワークをニューラルネットワークパラメータ格納部８０に格納されたパラメータを用いて設定する。さらに、各発話に対し感情強度設定部８４を用いてユーザが０、０．５又は１のうちのいずれかの感情強度パラメータを設定する。頭部動作生成部８２が、音声格納部８６から読出された音声から音響特徴量（Ｆ０、パワー、及び発話時間情報）を抽出し、その音響特徴量とユーザによって当該発話に対し設定された感情強度パラメータとを用いて、ニューラルネットワークに対する入力を作成し与える。この入力に応答して、ニューラルネットワークにより、キャラクタの頭部動作９０が自動生成される。

【００６６】
この頭部動作生成部８２の動作の詳細について、図３を参照して説明する。まず、音声格納部８６に格納された音声データが音声認識部１１０に与えられ、音声認識部１１０で音素に分解され音響特徴量が付された音素列として出力される。その音響特徴量及び音素列は音声情報格納部１１２に格納される。

【００６７】
音声情報格納部１１２に格納された音声情報の中で、音響特徴量に関するものは頭部動作を生成するために頭部動作自動生成部１１４に出力される。音素列はキャラクタの口の動きを合成するために使用される。

【００６８】
一方、ユーザが感情強度設定部８４で設定した感情強度パラメータも頭部動作自動生成部１１４に与えられる。

【００６９】
頭部動作自動生成部１１４においては、音声情報格納部１１２から与えられた音響特徴量と、感情強度設定部８４で設定された感情強度とから、ニューラルネットワークへの入力データが作成される。この入力データに関しては、音声データの最後の１１フレーム分が記憶される。これら１１フレーム分の入力データがニューラルネットワークへの入力として与えられる。これに応答して、対応する頭部動作パラメータがニューラルネットワークから出力される。この頭部動作自動生成部１１４での具体的な変換処理については後述する。

【００７０】
頭部動作自動生成部１１４から出力された頭部動作パラメータは図３に示す頭部動作パラメータ格納部１１８に格納される。頭部動作パラメータ格納部１１８に格納された頭部動作パラメータを使って頭部動作合成部１２０で画像の頭部動作が合成される。

【００７１】
頭部動作自動生成部１１４は以下の様に動作する。

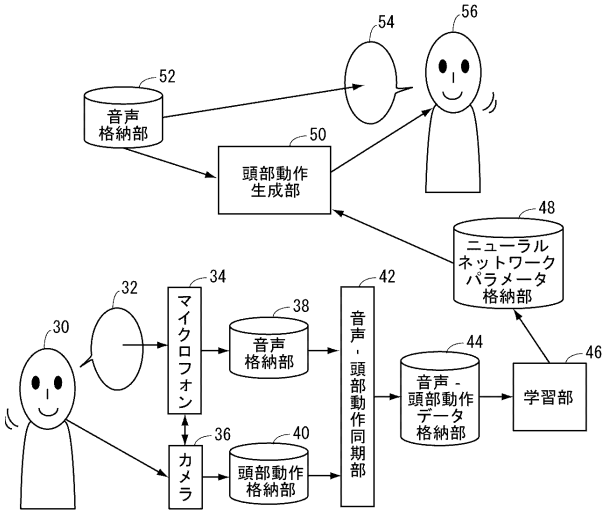
【００７２】
図４を参照してまず、音声情報格納部１１２から出力された各フレームの音響特徴量に関する情報のうち、声の高さ１７０と、声の大きさ１７２と、発話時間情報１７４とが特徴量抽出部１７６により抽出される。この情報は、直前の１１フレーム分にわたり特徴量抽出部１７６内に保持される。これら直前の１１フレーム分の特徴量と感情強度設定部８４でユーザによって任意に設定された感情強度とが、ニューラルネットワーク１８０に与えられる。これに応答して、ニューラルネットワーク１８０が、頭部動作を合成するためのパラメータである頭部動作パラメータ（頭部回転角Ｒｘ１８２，Ｒｙ１８４及びＲｚ１８６）を出力する。ニューラルネットワーク１８０はニューラルネットワークパラメータ格納部８０に記憶されたニューラルネットワークパラメータにより予め設定される。このパラメータは、学習結果に従い、ニューラルネットワーク１８０における入力（音声）と出力（頭部動作パラメータ）との関係を示すものである。出力された頭部動作パラメータは頭部動作パラメータ格納部１１８に格納される。

【００７３】

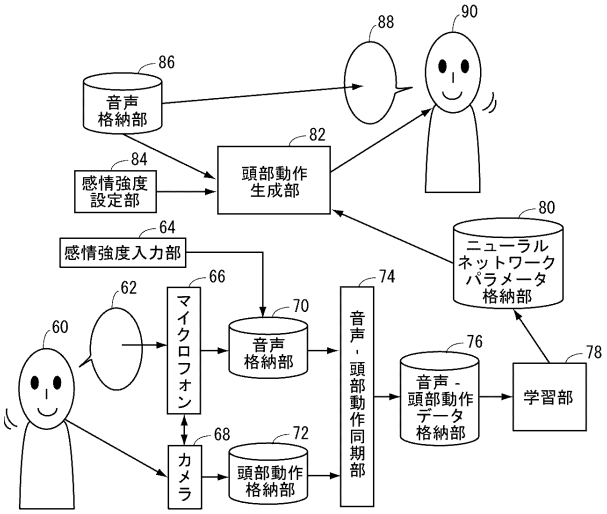
＜頭部動作生成の具体例＞
図８に、様々な感情強度を入力する事によって変化する頭部動作の具体例を示す。
【００７４】
図８を参照して、縦軸には頭部回転角の一つで、頭部の上下方向の動きを表わす角Ｒｘの値を、横軸には発話開始時からの時間（発話時間情報）を単位１／１０００秒で示す。グラフ中に書かれた英語の文章は、この具体例で使用された発話文である。この具体例では感情強度は怒り感情強度を使用した。また、その怒り感情強度は０．０（通常状態）から１．０（激怒）までの間で０．２刻みで推移させたものを使用した。
【００７５】
図８に示される様に、感情強度を変化させると、波形で示される頭部の上下動が変化する。そして、感情強度が強くなる、すなわち、より「激怒」に近づくにつれて、０．７５秒から１．３秒にかけての頭部の上下動が大きくなる。つまり頭部画像の動きが激しくなる。
【００７６】
なお、図８において、特に１秒付近の角Ｒｘの値は、感情強度の値に対し、非線型に変化している。これは、ニューラルネットワークの様に非線型の変換を行なう場合に特徴的な事であり、実際に発話者の頭部動作はこのような非線型性を示す。ニューラルネットワークとは異なり、線型的な手法を用いて頭部動作を生成すると、このような上下動の非線型性が失われ、感情強度に対して動きの大きさが単に線型にしか変化しないような、単調な変形結果しか得られない。
【００７７】
＜性能評価のための実験＞
本実施の形態に係る頭部動作生成のための装置の有効性を評価するために、学習に用いたデータの一部である、実際の頭部の頭部回転角Ｒｘと、頭部動作を本実施の形態に係るシステムで生成した場合の頭部回転角Ｒｘとの比較を行なった結果を図９に示す。図９においては、縦軸に頭部回転角度Ｒｘをとり、横軸に発話開始からの時間を単位１／１０００秒でとっている。この性能評価のための実験で用いた感情強度は怒り感情強度である。
【００７８】
図９に示される様に、通常の状態（図９の上段：感情強度ＥＦ＝０．０）、怒り（図９の中段：ＥＦ＝０．５）、激怒（図９の下段：ＥＦ＝１．０）のいずれにおいても、実際の頭部動作の回転角Ｒｘと、本実施の形態に係るシステムにより生成された頭部動作の回転角Ｒｘとの波形は互いに非常に類似している。
【００７９】
上述した波形の類似から、予め用意された学習用の音声から得られる音響特徴量と、ユーザによって設定された感情強度とを用いて生成された頭部回転角は、いずれの感情強度の場合にも、学習に用いられた実際の頭部動作の頭部回転角と類似する事が分かる。この結果から、ユーザが任意に感情強度を設定しても、音声に同期した自然な頭部動作が生成される事が期待できると言える。従って、音声からニューラルネットワークを使用して頭部動作を生成する際に任意の感情強度をユーザが入力する事によって、ユーザの感性に応じた、かつ、自然な頭部動作の生成が可能になる。
【００８０】
以上より、音声から頭部動作を自動生成する際に、ユーザによって任意に設定できる感情強度を使用する事で、感情強度に応じ、音声から自然な頭部動作を生成する事ができる。このような手法で頭部動作を生成する事により、ユーザの感性に応じた、かつ、効率的な頭部動作の生成が可能となる。
【００８１】
今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味及び範囲内でのすべての変更を含む。

【図面の簡単な説明】
【００８２】
【図１】従来技術による頭部動作自動生成システムを示す図である。
【図２】本発明に係る頭部動作自動生成システムを示す図である。
【図３】頭部動作生成部の詳細を示す図である。
【図４】頭部動作自動生成部の詳細を示す図である。
【図５】ニューラルネットワークの学習方法に係る処理の詳細について示す図である。
【図６】発話時間情報の詳細について示す図である。
【図７】頭部回転角の詳細を示す図である。
【図８】様々な感情強度を入力する事によって変化する頭部動作の具体例を示す図である。
【図９】本発明の性能評価の結果を示す図である。
【符号の説明】
【００８３】
６４　感情強度入力部
７４　音声－頭部動作同期部
７８　学習部
８４　感情強度設定部
８６　音声格納部
１１４　頭部動作自動生成部
１４８　モーションキャプチャ
１７６　特徴量抽出部
１８０　ニューラルネットワーク

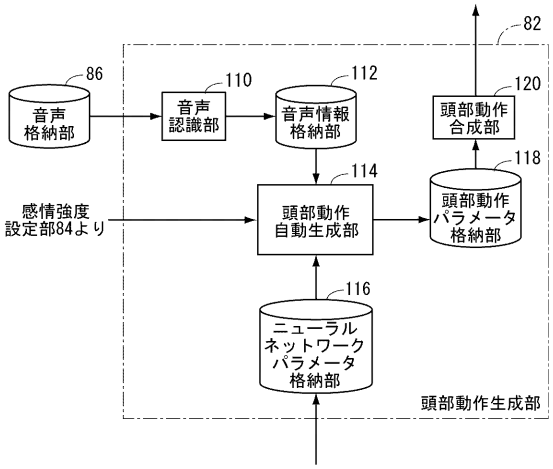
【図１】



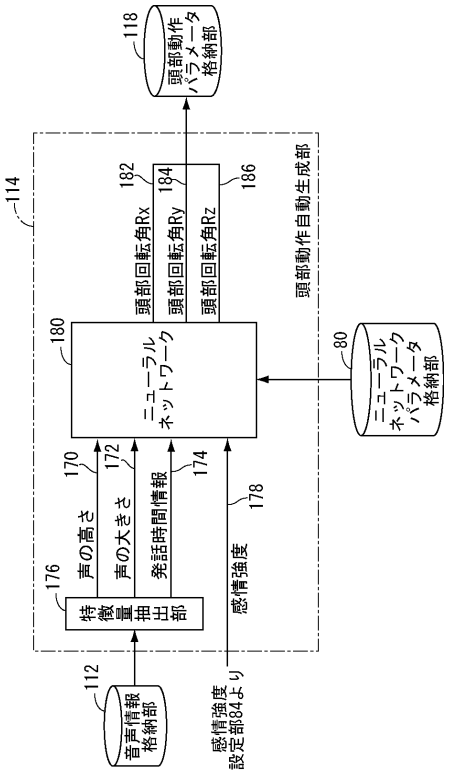
【図２】



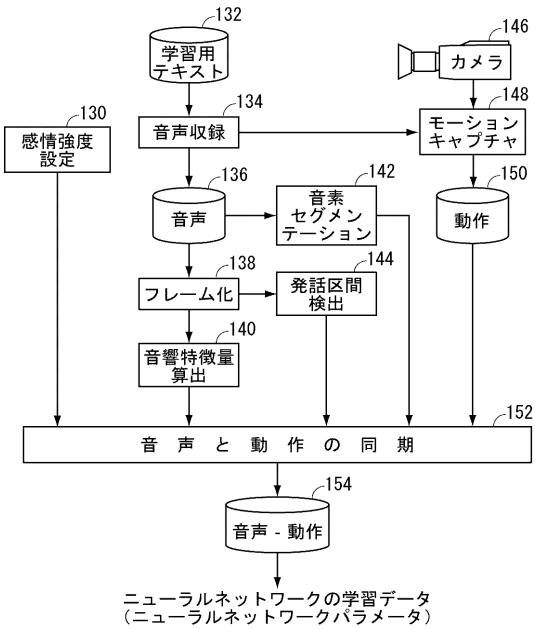
【図３】


82
86
音声
格納部
110
音声
認識部
112
音声情報
格納部
120
頭部動作
合成部
118
頭部動作
パラメータ
格納部
114
感情強度
設定部84より
頭部動作
自動生成部
116
ニューラル
ネットワーク
パラメータ
格納部
頭部動作生成部

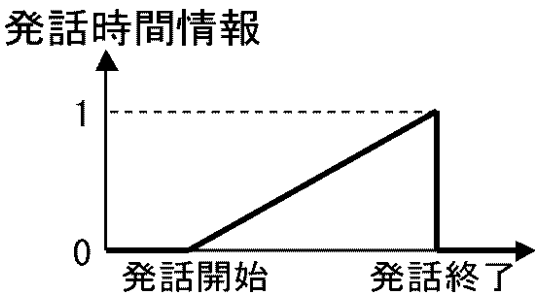
【図４】


118
頭部動作
パラメータ
格納部
114
頭部動作自動生成部
182
頭部回転角Rx
184
頭部回転角Ry
186
頭部回転角Rz
180
ニューラル
ネットワーク
80
ニューラル
ネットワーク
パラメータ
格納部
170
声の高さ
172
声の大きさ
174
発話時間情報
178
感情強度
176
特徴量抽出部
112
音声情報
格納部
感情強度
設定部84より

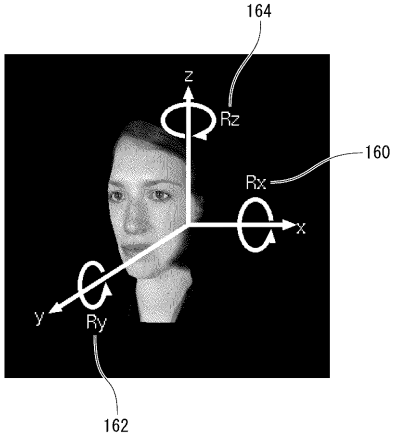
【図５】


132
学習用
テキスト
146
カメラ
130
感情強度
設定
134
音声収録
148
モーション
キャプチャ
136
音声
142
音素
セグメン
テーション
150
動作
138
フレーム化
144
発話区間
検出
140
音響特徴量
算出
152
音 声 と 動 作 の 同 期
154
音声 - 動作
ニューラルネットワークの学習データ
（ニューラルネットワークパラメータ）

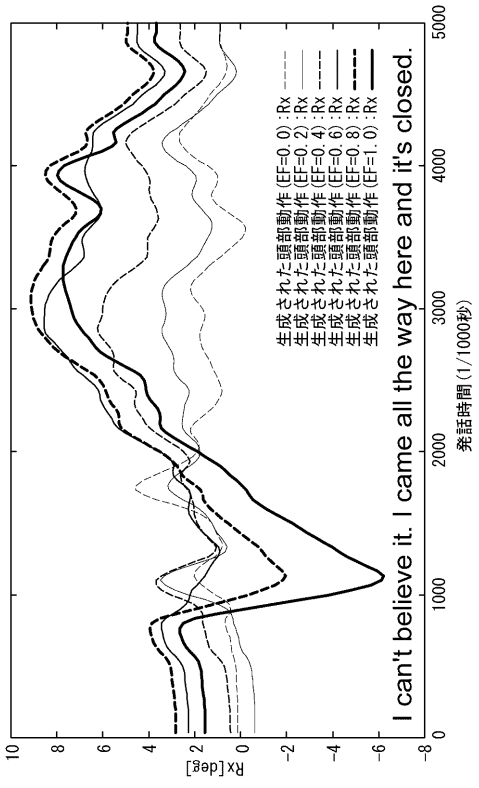
【図６】


発話時間情報
1
0
発話開始
発話終了

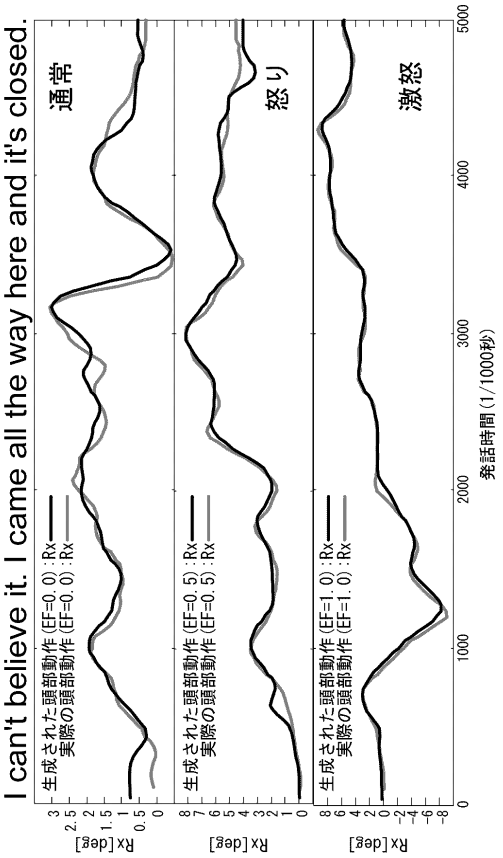
【図７】


164
z
Rz
160
Rx
x
y
Ry
162

【図８】


I can't believe it. I came all the way here and it's closed.
生成された頭部動作 (EF=0. 0) : Rx
生成された頭部動作 (EF=0. 2) : Rx
生成された頭部動作 (EF=0. 4) : Rx
生成された頭部動作 (EF=0. 6) : Rx
生成された頭部動作 (EF=0. 8) : Rx
生成された頭部動作 (EF=1. 0) : Rx
Rx[deg]
発話時間 (1/1000秒)

【図９】


I can't believe it. I came all the way here and it's closed.
通常
生成された頭部動作 (EF=0. 0) : Rx
実際の頭部動作 (EF=0. 0) : Rx
怒り
生成された頭部動作 (EF=0. 5) : Rx
実際の頭部動作 (EF=0. 5) : Rx
激怒
生成された頭部動作 (EF=1. 0) : Rx
実際の頭部動作 (EF=1. 0) : Rx
Rx[deg]
発話時間 (1/1000秒)

フロントページの続き

審査官　北川　純次

(56)参考文献　特開平１０－１５４２３８（ＪＰ，Ａ）
松下　善則，擬人化音声対話エージェントにおける発話時の頭部挙動モデル　A Head-Behavior Synchronization Model with Utterance for Anthropomorphic Spoken-Dialog Agent，電子情報通信学会技術研究報告　Ｖｏｌ．１０１　Ｎｏ．６９９　IEICE Technical Report，日本，社団法人電子情報通信学会　The Institute of Electronics,Information and Communication Engineers，２００２年　3月　1日，第101巻　第699号，pages:9～16

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　　３／００
Ｇ１０Ｌ　１３／００
Ｇ１０Ｌ　１５／００
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍＩＩ）
ＪＳＴ７５８０（ＪＤｒｅａｍＩＩ）