(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4963345号
(P4963345)
(45)発行日　平成24年6月27日(2012.6.27)　(24)登録日　平成24年4月6日(2012.4.6)

(51)Int.Cl.　F I
G10L 13/06 (2006.01)　G10L 13/06　240D

請求項の数 3　(全 13 頁)

(21)出願番号　特願2004-270307(P2004-270307)
(22)出願日　平成16年9月16日(2004.9.16)
(65)公開番号　特開2006-84859(P2006-84859A)
(43)公開日　平成18年3月30日(2006.3.30)
審査請求日　平成19年8月10日(2007.8.10)

（出願人による申告）平成１６年度独立行政法人情報通信研究機構、研究テーマ「大規模コーパスベース音声対話翻訳技術の研究開発」に関する委託研究、産業活力再生特別措置法第３０条の適用を受ける特許出願

(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　津崎　実
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　小坂　直敏
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　河井　恒
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】音声合成方法及び音声合成プログラム

(57)【特許請求の範囲】
【請求項１】
末尾に第１の音素を有する第１の音声サンプルと、先頭に、前記第１の音素と同じ音素である第２の音素を有する第２の音声サンプルとを、所定の時間期間内において接続合成する音声合成方法であって、
前記第１の音素の駆動波形と前記第２の音素の駆動波形とを接続して得られる音声波形の継続時間を決定するステップを含み、前記継続時間の先頭は、前記第１の音素の駆動波形の開始時刻であり、前記継続時間の末尾は、前記第２の音素の駆動波形の終了時刻であり、前記所定の時間期間は、前記継続時間内に含まれ、
前記音声合成方法はさらに、
前記所定の時間期間内の第１の時刻、および前記所定の時間期間内で当該第１の時刻より遅い第２の時刻により画定される移行区間を決定するステップと、
前記移行期間内の前記第１の時刻から前記第２の時刻までの間の時点であって、前記第１の音声サンプルと前記第２の音声サンプルとの混合割合が所定の関係を充足する時点を決定するステップと、
前記時点を決定するステップにおいて決定された時点での前記第１の音声サンプルと前記第２の音声サンプルとの駆動波形の位相を整合させるステップと、
前記所定の時間期間の先頭時刻から前記第１の時刻までの区間の合成音声を前記第１の音声サンプルから生成するステップと、
前記第１の音声サンプルから前記第２の音声サンプルへと、前記第１の時刻から前記第

２の時刻までの間の、時間に対する所定の滑らかな関数にしたがって両者の混合割合を変化させて混合することにより、前記移行区間における合成音声を生成するステップとを含み、

前記時点は、前記第１の音素の瞬時音圧値と前記第２の音素の瞬時音圧値との混合割合が実質的に等しくなる時間位置である、音声合成方法。

【請求項２】

前記合成音声を生成するステップは、前記第１の時刻の前記第１の音声サンプルの駆動波形から、前記第２の時刻の前記第２の音声サンプルへと、前記第１の音素の駆動波形の瞬時音圧値と前記第２の音素の駆動波形の瞬時音圧値とを、前記移行期間に対する前記第１の時刻からの経過時刻の重みで加重平均することによって合成音声を生成するステップを含む、請求項１に記載の音声合成方法。

【請求項３】

コンピュータにより実行されると、請求項１又は請求項２に記載の音声合成方法を実行するよう当該コンピュータを制御する、コンピュータで実行可能な音声合成プログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

この発明は、例えばＴＴＳ（Ｔｅｘｔ－Ｔｏ－Ｓｐｅｅｃｈ）システムなどに用いられ、音声コーパスから抽出された音声サンプルを互いに接続して音声合成を行なう音声合成方法及び音声合成プログラムに関する。

【背景技術】

【０００２】

音声合成技術の中に、素片接続型（または波形接続型）音声合成と呼ばれるものがある。この技術では、実際の音声（特定の話者であることが多い。）を収集して音声コーパスを作成する。音声コーパス中の音声データを音素、ダイフォンなどの所定の単位（波形セグメント）に分ける。各波形セグメントには、対応する音声データの音響・韻律に関する特徴量を示す情報が付されている。

【０００３】

音声合成時には、テキストが与えられると、このテキストに対し形態素解析などのテキスト処理を実行し、テキストを音声化した際の各音素などの所定単位ごとに合成目標の音韻・韻律情報および言語情報を生成する。この合成目標にできるだけ合致する音響的な特徴量を有し、かつ互いに接続したときに自然な音声となるような音声サンプルを音声コーパスから抽出する。そして、抽出された音声サンプルを互いに接続することで合成音声波形を生成する。

【０００４】

素片接続型音声合成の基本的な考え方は、次の二つに大別できる。第１は発話単位での接続であり、第２はダイフォン単位での接続である。

【０００５】

発話単位での接続は、発話の基本単位（例えば日本語の場合は、子音と母音との連鎖であるいわゆるＣＶ単位）を互いに接続する方法である。この方法は、自然な発話でも基本単位の間には音響的な不連続性が生じているのであるから、別々の音声サンプルを発話単位でつないで人工的な音響特性の不連続性が生じても、人間の聴覚にとっては許容可能なものであると想定している。

【０００６】

ただし、有声の子音、半母音などは音響的に前後の音と切離されている度合いが低くなるし、また母音の連鎖ということも音声合成上では生じ得る。すなわち、実際には前後の音と連続している音声であるにもかかわらず、前後と切離して得た音声素片を音声合成時に用いることがある。その結果、自然音声でも発話単位の間に連続的な遷移部が存在すると、素片接続により音声合成した結果が不自然になる可能性が高い。

【０００７】

一方、ダイフォン単位での接続合成方法は、母音を代表とする変化の比較的緩やかな区間の中間で接続する手法である。現在信頼性の高い音響的特徴のほとんどが静的な特性を捉えたものであり、変化が少ない部分では静的な特性でマッチした区間同士を中間でつなげば、物理的に非常に少ない不連続の範囲でつなぎやすいという利点がある。

【非特許文献１】Ｈ．カワハラ他、「加重平均群遅延に対する定点法に基づく正確な音声事象検出法」、ＩＣＳＬＰ－２０００予稿集、北京、ｐｐ．６６４－６６７、２０００年（Hideki Kawahara et al., "Accurate vocal event detection method based on a fixed-point to weighted average group delay", ICSLP-2000, pp. 664-667, Beijin, 2000）

【発明の開示】

【発明が解決しようとする課題】

【０００８】

しかし、このダイフォン単位接続法のように定常部で接続する方法では、パワースペクトル、基本周波数Ｆ０、波形としての連続性、などのすべての物理的側面でまったく等価な素片を常に保守することは不可能であり、通常はごく微小ではあるが物理的不連続の発生が不可避だという問題がある。しかも、人間の聴覚系は定常的状態に生じた僅かな差分に対しては非常に感度が鋭いのが現実であり、ごく微小な物理的不連続であっても、これを敏感に知覚してしまうという問題がある。

【０００９】

したがって、本発明の一つの目的は、接続部における物理的不連続の発生を抑制して知覚的な自然性を高めることができる音声合成方法を提供することである。

【００１０】

本発明の他の目的は、接続部において音響特徴が連続的に変化するように波形セグメントを接続し、知覚的な自然性を高めることができる音声合成方法を提供することである。

【課題を解決するための手段】

【００１１】

本発明の第１の局面に係る音声合成方法は、末尾に第１の音素を有する第１の音声サンプルと、先頭に第２の音素を有する第２の音声サンプルとを、所定の時間期間内において接続合成する音声合成方法であって、所定の時間期間内の第１の時刻、および所定の時間期間内で当該第１の時刻より遅い第２の時刻により画定される移行区間を決定するステップと、所定の時間期間の先頭時刻から第１の時刻までの区間の合成音声を第１の音声サンプルから生成するステップと、第１の音声サンプルから第２の音声サンプルへと、時間に対する所定の滑らかな関数にしたがって両者の混合割合を変化させて混合することにより、移行区間における合成音声を生成するステップとを含む。

【００１２】

移行区間において、第１の音声サンプルから第２の音声サンプルへと、両者の混合割合を滑らかな関数にしたがって変化させて混合させる。音声の不連続が発生することが避けられ、より自然な合成音声を生成できる。合成のための所定の時間期間の先頭から第１の時刻までの間は第１の音声サンプルを合成音声とする。したがってその直前の音素との連続が保たれ、自然な音声が合成できる。

【００１３】

さらに好ましくは、合成音声を生成するステップは、第１の音声サンプルから第２の音声サンプルへと、第１の音素の瞬時音圧値と第２の音素の瞬時音圧値とを、移行期間に対する第１の時刻からの経過時刻の重みで加重平均することによって合成音声を生成するステップを含む。

【００１４】

このような加重平均により第１の音声サンプルと第２の音声サンプルとを混合することで、両者の間の変化が直線的になり、簡単な処理で合成音声の接続部分を自然なものにすることができる。

【００１５】

より好ましくは、合成音声を生成するステップは、移行期間内であって、第１の音声サンプルと第２の音声サンプルとの混合割合が所定の関係を充足する時点を決定するステップと、このステップにおいて決定された時点での第１の音声と第２の音声との駆動波形の位相を整合させて、第１の音素の瞬時音圧値と第２の音素の瞬時音圧値とを重みで加重平均することによって合成音声を生成するステップを含むようにしてもよい。

【００１６】

両者の駆動波形の位相を整合させることにより、合成後の音声波形は合成前の音声の特徴とよく一致する特徴を示し、接続部分がより自然なものとなる。

【００１７】

好ましくは、合成音声を生成するステップは、移行期間内における第１の音声サンプルと第２の音声サンプルとの駆動波形の位相のずれの重み付きの和が最小化するように第１の音声サンプルと第２の音声サンプルとの駆動波形の位相を整合させて、第１の音素の瞬時音圧値と第２の音素の瞬時音圧値とを加重平均することによって合成音声を生成するステップを含む。ずれの和を計算する際の重みは、第１の音素の瞬時音圧値と第２の音素の瞬時音圧値との混合割合が実質的に等しくなる時間位置で最大となり、混合割合が前記時間位置から遠ざかるにつれて減少するように選択される。

【００１８】

第１の音声サンプルと第２の音声サンプルとは、互いに異なる音声データから得られたものであることが通常である。したがって両者の駆動波形の周期が多少異なる場合があり、両者の駆動波形を移行期間の全体にわたって一致させることはできない。そこで、駆動波形のずれの和を最小化するように両者の位相を整合させるのが合理的である。ただし、この場合に各時点でのずれを対等に扱うのではなく、両者の混合割合が実質的に等しくなる時間位置で最大となり、そこから遠ざかるにつれて減少するような重みを用いると、音声に対する悪影響を小さくとどめることができ、自然な合成音声を生成できる。

【００１９】

本発明の第２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、上記したいずれかの音声合成方法を実行するよう当該コンピュータを制御するものである。

【発明を実施するための最良の形態】

【００２０】

以下に述べる本発明の実施の形態は、コンピュータおよびコンピュータ上で動作するソフトウェアにより実現される。もちろん、以下に述べる機能の一部又は全部を、ソフトウェアでなくハードウェアで実現することも可能である。

【００２１】

図１に、本発明の実施の形態で利用されるコンピュータシステム２０の外観図を、図２にコンピュータシステム２０のブロック図を、それぞれ示す。なおここに示すコンピュータシステム２０はあくまで一例であり、この他にも種々の構成が可能である。

【００２２】

図１を参照して、コンピュータシステム２０は、コンピュータ４０と、いずれもこのコンピュータ４０に接続されたモニタ４２、キーボード４６、およびマウス４８を含む。コンピュータ４０にはさらに、ＣＤ－ＲＯＭ（Ｃｏｍｐａｃｔ Ｄｉｓｋ Ｒｅａｄ－Ｏｎｌｙ Ｍｅｍｏｒｙ）ドライブ５０と、ＦＤ（Ｆｌｅｘｉｂｌｅ Ｄｉｓｋ）ドライブ５２とが内蔵されている。

【００２３】

図２を参照して、コンピュータシステム２０はさらに、コンピュータ４０に接続されるプリンタ４４を含むが、これは図１には示していない。またコンピュータ４０はさらに、ＣＤ－ＲＯＭドライブ５０およびＦＤドライブ５２に接続されたバス６６と、いずれもバス６６に接続された中央演算装置（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ：ＣＰＵ）５６、コンピュータ４０のブートアッププログラムなどを記憶したＲＯＭ（Ｒｅａｄ－Ｏｎｌｙ Ｍｅｍｏｒｙ）５８、ＣＰＵ５６が使用する作業エリアおよびＣＰＵ５

６により実行されるプログラムの格納エリアを提供するＲＡＭ(Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）６０、および後述する音声データベースを格納したハードディスク５４を含む。

【００２４】

以下に述べる実施の形態のシステムを実現するソフトウェアは、たとえば、ＣＤ－ＲＯＭ６２のような記録媒体上に記録されて流通し、ＣＤ－ＲＯＭドライブ５０のような読取装置を介してコンピュータ４０に読込まれ、ハードディスク５４に格納される。ＣＰＵ５６がこのプログラムを実行する際には、ハードディスク５４からこのプログラムを読出してＲＡＭ６０に格納し、図示しないプログラムカウンタによって指定されるアドレスから命令を読出して実行する。ＣＰＵ５６は、処理対象のデータをハードディスク５４から読出し、処理結果を同じくハードディスク５４に格納する。

【００２５】

コンピュータシステム２０の動作自体は周知であるので、ここではその詳細については繰返さない。

【００２６】

なお、ソフトウェアの流通形態は上記したように記憶媒体に固定された形には限定されない。たとえば、ネットワークを通じて接続された他のコンピュータからデータを受取る形で流通することもあり得る。また、ソフトウェアの一部が予めハードディスク５４中に格納されており、ソフトウェアの残りの部分をネットワーク経由でハードディスク５４に取込んで実行時に統合するような形の流通形態もあり得る。

【００２７】

一般的に、現代のプログラムはコンピュータのオペレーティングシステム（ＯＳ）によって提供される汎用の機能を利用し、それらを所望の目的にしたがって組織化した形態で実行することにより前記した所望の目的を達成する。したがって、以下に述べる本実施の形態の各機能のうち、ＯＳまたはサードパーティが提供する汎用的な機能を含まず、それら汎用的な機能の実行順序の組合せだけを指定するプログラム（群）であっても、それらを利用して全体的として所望の目的を達成する制御構造を有するプログラム(群）である限り、それらが本発明の技術的範囲に含まれることは明らかである。

【００２８】

［原理］

本実施の形態のプログラムによってコンピュータ４０が実行する音声合成方法の原理を説明する。なお、以下の説明では、音素Ｐ１、音素Ｖ２、音素Ｐ３、音素Ｖ２ａ、音素Ｖ２ｂ、音素Ｐａ、音素Ｐｂをそれぞれ単にＰ１、Ｖ２、Ｐ３、Ｖ２ａ、Ｖ２ｂ、Ｐａ、Ｐｂという。またＶ２ａとＶ２ｂは、同じ音素Ｖ２に対応する、互いに異なる実サンプル中の音素（波形）であるものとする。

【００２９】

例として３つの母音からなる音素の連鎖を接続合成によって実現する場合を考える。図３（ａ）に示すように、実現したい音素の連鎖がＰ１－Ｖ２－Ｐ３であるとする。また、音声コーパス中に存在する、Ｐ１とＶ２の連鎖Ｐ－Ｖ２ａからなる実サンプル（第１の音声サンプル）１０１（図３（ｂ）に示す）と、音素Ｖ２とＰ３の連鎖Ｖ２ｂ－Ｐ３からなる実サンプル（第２の音声サンプル）１０２（図３（ｃ）に示す）を接続する場合を想定する。なお、図３（ｂ）のＰａは、音声コーパス中で第１の音声サンプル１０１のＶ２ａの直後に存在している音素であり、図３（ｃ）に示すＰｂは、音声コーパス中で第２の音声サンプル１０２の音素Ｖ２ｂの直前に存在している音素であるものとする。

【００３０】

従来法によるダイフォン接続の場合は、図３（ｄ）に示すように、Ｖ２ａとＶ２ｂの中間点１１０で接続をする。Ｖ２ａもＶ２ｂも合成目標であるＶ２に合致する音響特徴量を有しており、実質的にその継続長もほぼ同じである。したがってこのように互いの中間点でＶ２ａとＶ２ｂとを接続することで、Ｖ２と一致する継続長を持つ音素Ｖ２ａ／Ｖ２ｂが得られる。

【００３１】
しかしこの場合、接続点である中間点１１０を境に音響的特徴が微妙に変化し、聴覚的な不自然さを感じる。両者の差が偶然小さければ聴覚的にも滑らかな接続が達成できることになるが、それが必ず保証されているわけではない。
【００３２】
そこで、この実施の形態では、Ｖ２ａとＶ２ｂを中間点１１０でいきなり繋ぐのではなく、以下のようにする。図３（ｅ）を参照して、Ｖ２の継続時間１４０内で第１の時点１４２と、Ｖ２の継続時間１４０内でかつ第１の時点１４２より後の第２の時点１４４とを決定する。この二つの時点１４２および１４４により移行期間１０３が画定される。Ｖ２の継続時間１４０内で移行期間１０３の前の期間１０４をＶ２ａの保存期間、後の期間１０５をＶ２ｂの保存期間１０５とする。すなわち、Ｖ２の継続時間１４０をＶ２ａの保存期間１０４、Ｖ２ｂの保存期間１０５、およびＶ２ａからＶ２ｂへの移行期間１０３に分ける。そして、この移行期間１０３でＶ２ａ的な音からＶ２ｂ的な音へと音響的特徴を徐々に移行させることによって知覚的に有害な急激な変化を排除する。なお、図３（ｅ）における記号「Ｖ２ａｂ」は、Ｖ２の継続期間においてＶ２ａとＶ２ｂが混合されて接続されていることを表している。
【００３３】
本実施の形態では、第１の時点１４２および第２の時点１４４については、継続時間１４０の全体に対しどのような割合の時点とするかを予め定めておき、合成目標が与えられ継続時間１４０が定まったところで決定する。
【００３４】
ＶＶ２ａからＶ２ｂへの音声の変化の様子を時間の関数として示したのが図３（ｆ）である。図３を参照して、実線１３０で示したように、本実施の形態ではＶ２ａの保存期間１０４ではＶ２ａが１００％、Ｖ２ｂの保存期間１０５ではＶ２ｂが１００％、移行期間１０３ではＶ２ａからＶ２ｂまで直線的に変化する割合で両者を混合する。
【００３５】
これに対し図３（ｄ）に示した従来のダイフォン接続に対応して時間の関数で表せば、そのグラフは図３（ｆ）において破線１２０で示すようになる。破線１２０で示すように、接続点である中間点１１０を境に音素がＶ２ａからＶ２ｂに完全に変化し、その結果、音響的特徴もこの時点で変化する。そのため生じる音響的特徴の不連続により、聴覚的な不自然さを感じることがある。
【００３６】
本実施の形態において目標とする効果を上げるためには、図３（ｅ）及び（ｆ）に示す移行期間１０３を、実現したいＶ２の継続時間１４０を上限として、その中で十分に長く取る必要がある。最大ではＶ２の継続時間１４０の全体を移行期間とし、Ｖ２ａ、Ｖ２ｂの保存期間１０４、１０５の長さをゼロとすることもできる。
【００３７】
しかし、Ｖ２ａの開始時点はＰ１からの調音結合による影響が残っている可能性が高い。この部分でＶ２ｂを混入させ始めると、その部分にはＶ２ｂの抽出前の環境であるＰｂからの影響も混入することとなり、Ｐ１からＶ２ａへの連続性に悪影響を及ぼすとともに、Ｖ２としての音韻性が低下する可能性が考えられる。したがって一般的には、移行期間１０３の直前にＶ２ａの保存期間１０４をある程度の長さ設けた方が望ましい。
【００３８】
一方、Ｖ２ｂの後半はＰ３に対する遷移部と見なすことができ、自然音声であっても遷移部として音韻性が曖昧になる部分である。この部分に多少他の環境からの混入が生じたとしても、Ｖ２としての知覚は既に確立されており、Ｐ３への滑らかな移行が生じている限り、知覚的にはそれほど害とならないと考えられる。したがって、本実施の形態では、Ｖ２ａの保存期間１０４をＶ２の継続時間１４０の半分の長さとし、残りの半分を移行期間１０３とする。この場合、Ｖ２ｂの保存期間１０５の長さはゼロとなる。
【００３９】

但し、本方法はＴＴＳシステムの中で使われる可能性が高いので、Ｐ１、Ｖ２、Ｐ３、Ｐａ、Ｐｂに関する情報を事前に持っていると想定できる。Ｖ２ａ、Ｖ２ｂの保存期間、移行期間の最適な割合についてはそれらの環境の組合せに依存する可能性が高い。その場合は、抽出環境と接続の組合せ毎に適切な形で混合することも可能である。

【００４０】

以下、接続合成方法の具体的な内容について説明する。

【００４１】

［第１の実施の形態（駆動周期同期型のクロスフェード法）］

この実施の形態では、移行期間１０３において、Ｖ２ａとＶ２ｂとの瞬時音圧値を移行期間１０３の先頭からの経過時間に応じた重みで加重平均することによってクロスフェードする。

【００４２】

この方法においてさらに品質向上が必要な場合は、次のような方策が可能となる。一般的にクロスフェードの難点として途中で二つの音が混ざる場合に、二つの音の中間的なひとつの音が聞こえて欲しいにもかかわらず、単純に二つの音が混ざって同時に聞こえてしまうということがある。この欠点を最小化するために、カワハラ（非特許文献１）によって提案された事象検出アルゴリズムにより駆動時点のマーキングを取り、２音間で位相のずれを最小化した形でクロスフェードをかける。または、音声コーパスの各サンプルに、駆動波形のピークを示すピッチマークを付しておき、そうしたマークを利用して位相のずれを最小化するようにしてもよい。

【００４３】

図４は、位相のずれをまったく考慮しない場合に生じうる問題を示したものである。図４（ａ）の波形２０１が母音Ｖ２ａの波形であり、図４（ｂ）の波形２０２が母音Ｖ２ｂの波形であるものとする。符号２１１、符号２１２で示す矢印が、それぞれＶ２ａ、Ｖ２ｂのピッチマークである。Ｖ２ａ、Ｖ２ｂ間の位相がずれたままで混合した結果が図４（ｃ）に示されている。波形２０３上から駆動周期を観察することが困難になることが分かる。このような場合、聴覚系は二つの母音が平行して存在しているという知覚像を持ちやすい。

【００４４】

図５はこれに対してＶ２ａ、Ｖ２ｂ間で位相を整合した場合を示す。この処理では、ピッチマーク２１１、２１２を参照して二つの波形の位相を整合させる。図５（ａ）（ｂ）はそれぞれ混合前のＶ２ａ、Ｖ２ｂの波形２０１、２０２であり、図５（ｃ）は混合後の波形２０４を示す。混合前のそれぞれの母音の波形とよく似た周期構造が保存されていることが分かる。したがって、このように二つの波形の位相を整合させることにより、二つの母音が同時に聞こえる不都合を回避できる。

【００４５】

現実的には２母音間で位相（駆動周期）が完全に一致することは期待できないため、ある時点でのずれをなくすと別の時点ではずれが生じざるを得ない。但し、本方法では混合の期間は短時間であるので、混合の割合が釣合う時点でのマーカーのずれが生じた場合のペナルティーを重く評価することが合理的な位相整合の取り方となる。したがって本実施の形態では、混合の割合がＶ２ａ、Ｖ２ｂそれぞれ５０％となる時点で両者の位相を整合させるようにする。

【００４６】

第１の実施の形態の方法を図６にまとめる。これはコンピュータにより実現するときのプログラムの制御構造を示すものでもある。まず、移行期間１０３の開始時点および終了時点を計算により設定し、第１の音声サンプル１０１と第２の音声サンプル１０２の接続対象音素Ｖ２ａとＶ２ｂの駆動波形の位相を整合させる（ステップ３００）。次に、Ｖ２ａとＶ２ｂの瞬時音圧値を移行期間１０３に対する移行期間の先頭からの経過時間を重みとして加重平均して移行期間１０３における混合波形を生成し（ステップ３０１）、第１の音声サンプル１０１と第２の音声サンプル１０２を接続する（ステップ３０２）。

【００４７】
この実施の形態によれば、音声波形の瞬時音圧値の加重平均という比較的簡単な処理で二つの音素をその中間で結合し、一つの音素（特にダイフォン接続における母音部）を生成できる。その結合個所では、合成波形は第１の実サンプルの波形から第２の実サンプルの波形に滑らかに変化する。したがって人間が音響的な特徴の不連続を知覚する可能性が低くなる。さらに、二つの波形の位相を整合させることにより、接続後の音声がより自然なものとなるという効果がある。
【００４８】
上記した実施の形態では、移行期間では、単純に加重平均をとることで第１の波形から第２の波形に滑らかに、かつ直線的に波形をクロスフェードしている。しかし本発明はそのような実施の形態には限定されない。例えば、時間に関して２次以上の関数で、かつ移行期間の両端でそれぞれＶ２ａおよびＶ２ｂとなるという境界条件を満足するような関数によって、移行期間における両者の混合割合を決定するようにしてもよい。この場合、この関数の値がある時間におけるＶ２ａの混合割合を表すものとすれば、関数の値が時間に対して単調減少となることが好ましい。
【００４９】
［第２の実施の形態（正弦波モデルによるモルフィング法）］
第２の実施の形態にかかる音声合成方法は、二つの音素Ｖ２ａとＶ２ｂの混合を単純な時間波形レベルではなく、正弦波成分に分解した後に行なう方法である。母音音声に代表されるような音響信号は、振幅・周波数の異なる複数の正弦振動の加算として表現可能である。この方法では、混合する基となる二つの音声信号Ｖ２ａとＶ２ｂをフーリエ変換によりそれぞれ複数の正弦波成分に分解し、その間の対応付けを取った後、Ｖ２ａとＶ２ｂの間で各成分の周波数、振幅項が連続的な変化を生じるように変化させることによって、中間的な音を実現する。
【００５０】
例えばＶ２ａ、Ｖ２ｂの問に基本周波数のずれがあるような場合、前述の第１の実施の形態では、中間部に２種類の基本周波数成分が出現することが避けられない。聴覚系はそのような場合には二つの音を知覚しがちである。しかしこの第２の実施の形態に係る方法では、基本周波数成分は基本周波数成分として連続的に変化を起こす。したがって混合部が２音に分離して聞こえる印象を回避することが可能となる。
【００５１】
第２の実施の形態の方法を図７で説明する。この方法をコンピュータプログラムで実現する場合、そのためのプログラムの制御構造もこの様な形となる。第１の音声サンプル１０１の接続対象音素Ｖ２ａを、ｎ個の正弦波の成分に分解する（ステップ４００）。同様に、第２の音声サンプル１０２の接続対象音素Ｖ２ｂを、ｎ個の正弦波の成分に分解する（ステップ４０１）。なおｎは予め定められた整数であるものとする。
【００５２】
次に、分解した正弦波成分の対応するもの同士を混合するに際して、両波形がデジタルデータであるため、一方の波形のある成分と他方の波形のどの成分とを組合せて加重平均するかを決定する（ステップ４０２）。
【００５３】
次に、移行期間を設定し、対応するｎ個の正弦波同士の、対応を決定した点により表される振幅を、移行期間１０３の重みで加重平均することにより混合した後（ステップ４０３）、ｎ個の波形をフーリエ逆変換により合成して移行期間１０３における混合波形を生成し（ステップ４０４）、次いで第１の音声サンプル１０１と第２の音声サンプル１０２を接続する（ステップ４０５）。
【００５４】
なおこの実施の形態では、処理を簡単にするためにｎを固定するものとしている。しかし本発明はそのような実施の形態には限定されず、フレームごとにｎが変化するような方

法も可能である。

【００５５】

［第３の実施の形態（音響ボコーダ・モデルによるモルフィング法）］

第２の実施の形態が音声信号を正弦波成分に分解するのに対して、この実施例はソースフィルター原理に基づいて音声信号を駆動源情報成分と共振特性成分とに分離し、それぞれの次元での連続的な変化を実現した後、それらからボコーダにより音声を合成する。

【００５６】

母音信号に代表されるような音は、駆動音源情報の成分とその伝達系の共振特性の成分へと分解可能である。前者は声帯振動の周期性によってフーリエスペクトル上には基本周波数とその高調波の位置として主に反映される。後者はそれらの高調波成分の包絡を決定するもので、主に声道の形状の変化によって人間の発声の場合は実現される。したがって、声道形状を変化させずに声の高さだけを変えたり、反対に声の高さを変えずに声道形状だけを変えたりといった独立の制御が原理的には可能である。

【００５７】

第３の実施の形態では、このように分解した音源情報成分と共振特性成分という独立の次元の変数上でＶ２ａ、Ｖ２ｂの混合を実施する。第２の実施の形態に比べてより人間の音声生成器官で生じていることに近いため、不自然な混合を起こす可能性の低減が見込まれる。

【００５８】

第３の実施の形態の方法を図８で説明する。コンピュータにより本実施の形態にかかる方法を実現する場合、そのためのプログラムの制御構造は図８に示されるようなものとなる。図８を参照して、第１の音声サンプル１０１の接続対象音素Ｖ２ａを、音源情報成分と共振特性成分とに分解する（ステップ５００）。同様に、第２の音声サンプル１０２の接続対象音素Ｖ２ｂを、音源情報成分と共振特性成分とに分解する（ステップ５０１）。

【００５９】

次に、移行期間を設定し、音源情報成分同士の振幅を、移行期間１０３（図３（ｅ）および（ｆ）参照）に対する移行期間１０３の先頭からの経過時間の重みで加重平均することにより混合し、１個の音源情報成分を生成する（ステップ５０４）。具体的には、標本周波数値のパラメータ上での加重平均をとる。同様に、共振特性成分同士の振幅を、移行期間１０３内での経過時間の重みで加重平均することにより混合し、１個の共振特性成分を生成する（ステップ５０５）。具体的には、共振特性のスペクトルのパラメータ上での加重平均をとる。ただしこの共振特性の加重平均の場合、フォルマント間の対応をとって周波数軸の不均等圧縮・伸長を行なう。

【００６０】

次に、生成した音源情報成分と共振特性成分とを音響ボコーダにより合成して、移行期間１０３における混合波形を生成し（ステップ５０６）、次いで第１の音声サンプル１０１と第２の音声サンプル１０２とを接続する（ステップ５０７）。

【００６１】

［変形例］

上記した実施の形態では、移行期間では、単純に加重平均をとることで第１の波形から第２の波形に滑らかに直線的に波形をクロスフェードしている。しかし本発明はそのような実施の形態には限定されない。例えば、時間に関して２次以上の関数で、かつ移行期間の両端でそれぞれＶ２ａおよびＶ２ｂとなるという境界条件を満足するような関数によって、移行期間における両者の混合割合を決定するようにしてもよい。この場合、この関数の値がある時間におけるＶ２ａの混合割合を表すものとすれば、関数の値が時間に対して単調減少となることが好ましい。

【００６２】

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および範囲内

でのすべての変更を含む。

【図面の簡単な説明】

【００６３】

【図１】この発明の一実施の形態の音声合成プログラムを実行するコンピュータシステムの外観図である。

【図２】図１のコンピュータシステムのブロック図である。

【図３】この発明の一実施の形態の音声合成方法の原理を説明するための模式図である。

【図４】位相を整合させることなく音素を混合する場合の波形図である。

【図５】位相を整合させて音素を混合する場合の波形図である。

【図６】第１の実施の形態の音声合成方法を説明するためのフローチャートである。

【図７】第２の実施の形態の音声合成方法を説明するためのフローチャートである。

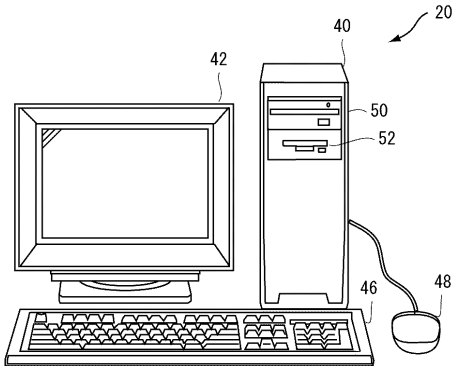
【図８】第３の実施の形態の音声合成方法を説明するためのフローチャートである。

【符号の説明】

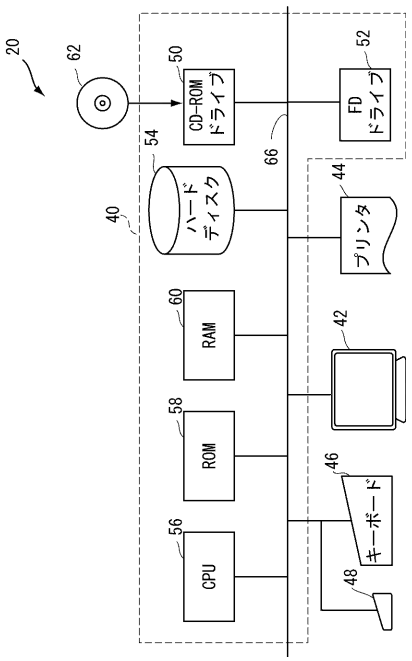
【００６４】

４０　コンピュータ、１０１　第１の音声サンプル、１０２　第２の音声サンプル、１０３　移行期間、１０４　第１の音素の保存期間、１０５　第２の音素の保存期間、Ｖ２ａ　第１の音素、Ｖ２ｂ　第２の音素

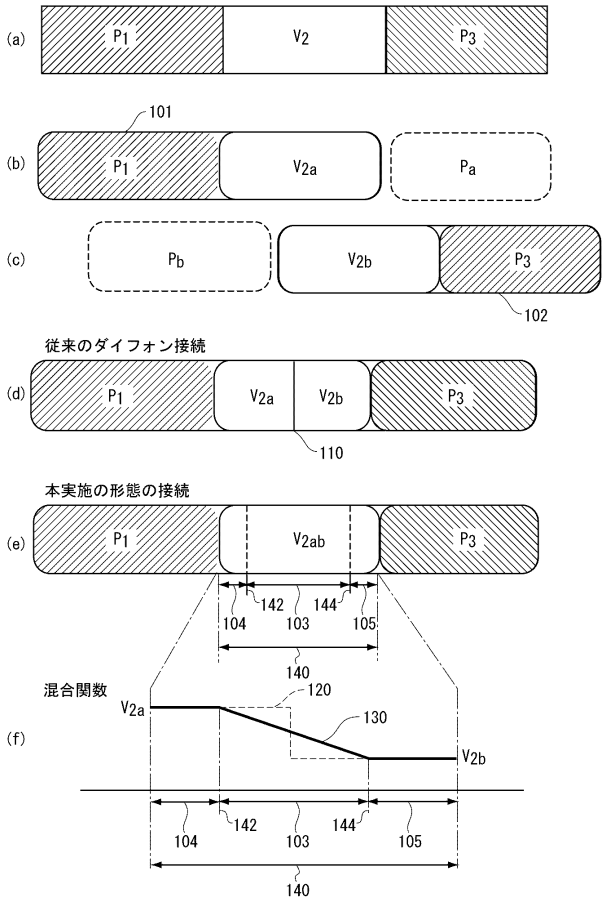
【図１】



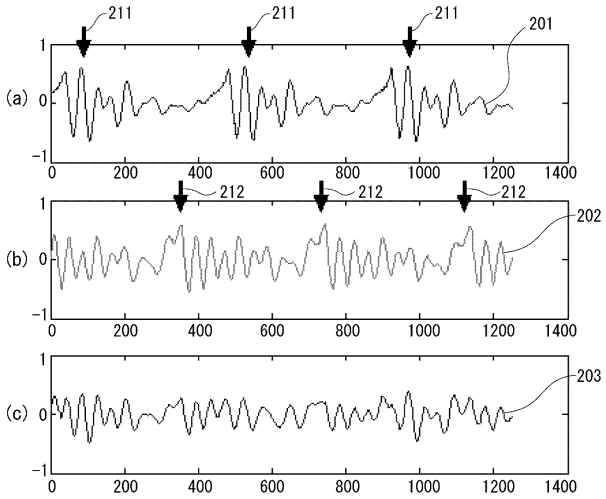
【図２】



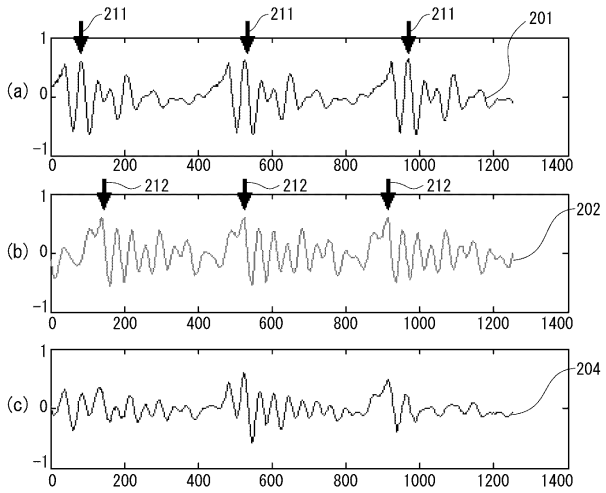
【図３】


(a)
P1
V2
P3
101
(b)
P1
V2a
Pa
(c)
Pb
V2b
P3
102
従来のダイフォン接続
(d)
P1
V2a
V2b
P3
110
本実施の形態の接続
(e)
P1
V2ab
P3
142
144
104
103
105
140
混合関数
120
V2a
130
(f)
V2b
142
144
104
103
105
140

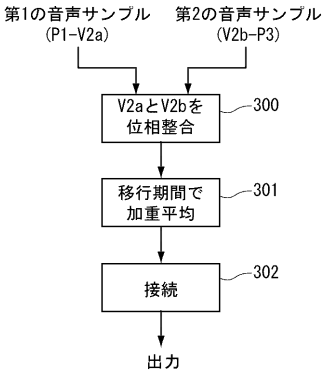
【図４】


211
211
211
201
(a)
212
212
212
202
(b)
203
(c)

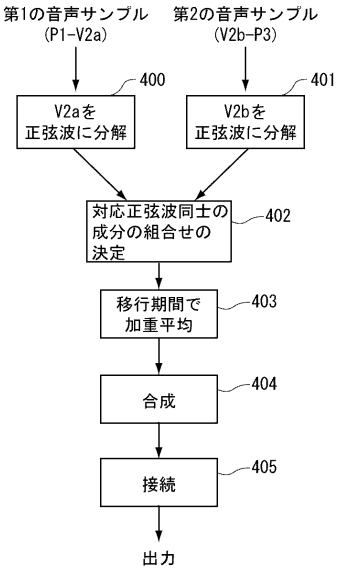
【図５】


211
211
211
201
(a)
212
212
212
202
(b)
204
(c)

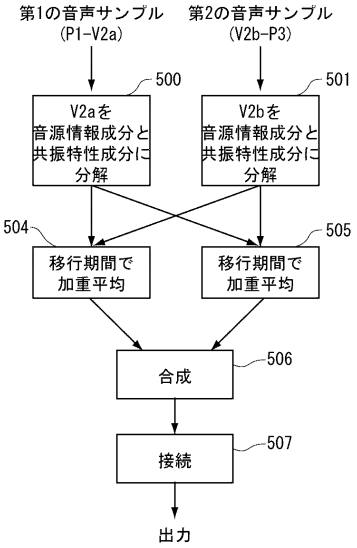
【図６】


第1の音声サンプル
(P1-V2a)
第2の音声サンプル
(V2b-P3)
V2aとV2bを
位相整合
300
移行期間で
加重平均
301
接続
302
出力

【図７】


第1の音声サンプル
(P1-V2a)
第2の音声サンプル
(V2b-P3)
400
401
V2aを
正弦波に分解
V2bを
正弦波に分解
対応正弦波同士の
成分の組合せの
決定
402
移行期間で
加重平均
403
合成
404
接続
405
出力

【図８】


第1の音声サンプル
(P1-V2a)
第2の音声サンプル
(V2b-P3)
500
501
V2aを
音源情報成分と
共振特性成分に
分解
V2bを
音源情報成分と
共振特性成分に
分解
504
505
移行期間で
加重平均
移行期間で
加重平均
506
合成
507
接続
出力

フロントページの続き

審査官　毛利　太郎

(56)参考文献　特開平１１－０３８９８９（ＪＰ，Ａ）
特開平１０－１２４０８２（ＪＰ，Ａ）
特開平１１－１９４７９６（ＪＰ，Ａ）
特開平０５－２９７８９１（ＪＰ，Ａ）
特開２００４－１０２１１８（ＪＰ，Ａ）
特開平１１－２２４０９６（ＪＰ，Ａ）
特表２００５－５２３４７８（ＪＰ，Ａ）
特開２００４－２５８６６０（ＪＰ，Ａ）
Eric MOULINES and Francis CHARPENTIER, Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones, Speech Communication, EP, ISCA, １９９０年, Vol.9, p.453-467

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ１０Ｌ　１３／００－１３／０８