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

特許第4829477号
(P4829477)

(45)発行日　平成23年12月7日(2011.12.7)　(24)登録日　平成23年9月22日(2011.9.22)

(51)Int.Cl.		F I		
G10L 13/08	(2006.01)	G10L	13/08	13OB
G10L 11/00	(2006.01)	G10L	11/00	101
G10L 13/06	(2006.01)	G10L	13/06	2OO
G10L 21/04	(2006.01)	G10L	21/04	12OA

請求項の数 27　(全 34 頁)

(21)出願番号　特願2004-79079(P2004-79079)
(22)出願日　平成16年3月18日(2004.3.18)
(65)公開番号　特開2005-266349(P2005-266349A)
(43)公開日　平成17年9月29日(2005.9.29)
審査請求日　平成19年1月15日(2007.1.15)
審判番号　不服2010-7652(P2010-7652/J1)
審判請求日　平成22年4月12日(2010.4.12)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100080816
弁理士　加藤　朝道
(72)発明者　三井　康行
東京都港区芝五丁目７番１号　日本電気株式会社内

合議体
審判長　板橋　通孝
審判官　加藤　恵一
審判官　溝本　安展

最終頁に続く

(54)【発明の名称】声質変換装置および声質変換方法ならびに声質変換プログラム

(57)【特許請求の範囲】
【請求項１】
変換目標となる話者の音声（「目標話者音声」という）を入力する目標話者音声入力部と、
音声合成用のデータを記憶する音声合成用データ記憶部と、
前記入力された目標話者音声と同一又は類似の発声内容を記述する発音記号列を入力する発音記号入力部と、
前記入力された発音記号列にしたがって、前記音声合成用データ記憶部に記憶されている音声合成用のデータに基づき、音声を合成して出力する音声合成部と、
前記音声合成部から出力される音声信号（「合成音」という）の特徴パラメータを抽出する合成音特徴パラメータ抽出部と、
前記目標話者音声の特徴パラメータを抽出する目標話者音声特徴パラメータ抽出部と、
前記合成音特徴パラメータ抽出部からの前記合成音の特徴パラメータと、目標話者音声特徴パラメータ抽出部からの前記目標話者音声の特徴パラメータとを入力し、特徴パラメータを表す空間において、前記合成音の第１の部分と、前記目標話者音声の第２の部分との時間軸上の対応関係を求め、前記合成音の前記第１の部分におけるスペクトル形状を、前記目標話者音声の前記第２の部分におけるスペクトル形状に変換する変換関数の同定を行う変換関数生成部と、
を備えている、ことを特徴とする声質変換装置。
【請求項２】

変換対象の音声信号を入力し、前記変換関数生成部で同定された前記変換関数を用いて、前記変換対象となる音声信号を、前記目標話者音声の声質を持つ音声信号に変換して出力する声質変換部をさらに備えている、ことを特徴とする請求項１記載の声質変換装置。

【請求項３】

前記目標話者音声入力部により入力された前記目標話者音声を入力して音声認識し、前記発音記号列を生成して前記音声合成部に出力する音声認識部をさらに備えている、ことを特徴とする請求項１記載の声質変換装置。

【請求項４】

前記音声合成部で前記合成音が生成される際に用いられた、少なくとも時刻情報に対応づいた音素列を含むセグメンテーション情報を記憶し、前記変換関数生成部に出力する第１のセグメンテーション情報記憶部をさらに備えている、ことを特徴とする請求項１～３のいずれか一に記載の声質変換装置。

【請求項５】

前記音声合成部で作成される前記合成音の、少なくとも時刻情報に対応づいた音素列を含むセグメンテーション情報を入力して、前記音声合成部に出力するセグメンテーション情報入力部をさらに備えている、ことを特徴とする請求項１～３のいずれか一に記載の声質変換装置。

【請求項６】

前記音声認識部で認識された目標話者音声の、少なくとも時刻情報に対応づいた音素列を含むセグメンテーション情報を記憶し、前記変換関数生成部および前記音声合成部の少なくとも一方へ出力する第２のセグメンテーション情報記憶部を備えている、ことを特徴とする請求項３又は４記載の声質変換装置。

【請求項７】

前記音声合成用データ記憶部に記憶される前記音声合成用のデータとして、互いに異なる複数の話者の音声データを含むことを特徴とする請求項１～３のいずれか一に記載の声質変換装置。

【請求項８】

請求項２記載の声質変換装置の前記声質変換部から出力される変換後の信号を音声合成用データとして記憶するデータ記憶部と、

発声内容を記述する記号列を入力する記号列入力部と、

前記入力された記号列にしたがって前記音声合成用データ記憶部内の音声合成用データから音声を合成して出力する音声合成出力部と、

を備えている、ことを特徴とする合成音生成装置。

【請求項９】

請求項１記載の声質変換装置における前記合成音の特徴パラメータと前記特徴パラメータに対応して同定された変換関数とを対応付けて記憶する変換データ記憶部と、

被変換音声信号を入力する音声入力部と、

前記入力された被変換音声信号の特徴パラメータを求め、求めた特徴パラメータに対応する前記変換データ記憶部内の前記合成音の特徴パラメータを探索し、前記合成音の特徴パラメータに対応する前記変換関数を読み出し、前記変換関数によって前記被変換音声信号を変換して出力する音声変換出力部と、

を備えている、ことを特徴とする合成音生成装置。

【請求項１０】

声質変換装置により声質を変換する方法であって、

声質の変換目標となる話者の音声（「目標話者音声」という）を入力するステップと、

前記目標話者音声の発声内容を記述する発音記号列を入力し、記憶部に予め記憶されている音声合成用のデータを用いて、前記発音記号列から、合成音を作成するステップと、

前記合成音を分析し、前記合成音の特徴パラメータを抽出するステップと、

前記目標話者音声を分析し、前記目標話者音声の特徴パラメータを抽出するステップと、

前記目標話者音声の特徴パラメータと前記合成音の特徴パラメータとの時間軸上の対応付けを行うステップと、
対応付けがなされた前記目標話者音声及び前記合成音の特徴パラメータに基づき、前記合成音のスペクトル形状を、前記目標話者音声のスペクトル形状に、変換するための変換関数を生成するステップと、
を含む、ことを特徴とする声質変換方法。
【請求項１１】
前記変換関数を用いて、変換対象となる音声信号を前記目標話者音声の声質を持つ音声信号に変換するステップをさらに含む、ことを特徴とする請求項１０記載の声質変換方法。
【請求項１２】
入力される前記目標話者音声を音声認識するステップを含み、
前記音声認識の結果から前記発音記号列を生成する、ことを特徴とする請求項１０記載の声質変換方法。
【請求項１３】
前記合成音の少なくとも時刻情報に対応づいた音素列を含むセグメンテーション情報を入力して、前記セグメンテーション情報に基づき前記合成音を生成する、ことを特徴とする請求項１０～１２のいずれか一に記載の声質変換方法。
【請求項１４】
前記合成音の生成時に用いられる、少なくとも時刻情報に対応づいた音素列を含む第１のセグメンテーション情報によって、前記時間軸上の対応付けを行うステップを含む、ことを特徴とする請求項１０～１２のいずれか一に記載の声質変換方法。
【請求項１５】
前記音声認識された目標話者音声の、少なくとも時刻情報に対応づいた音素列を含む第２のセグメンテーション情報によって、前記時間軸上の対応付けを行うステップを含む、ことを特徴とする請求項１０～１２、１４のいずれか一に記載の声質変換方法。
【請求項１６】
前記音声認識された目標話者音声の、少なくとも時刻情報に対応づいた音素列を含む第２のセグメンテーション情報に基づき、前記合成音を生成するステップを含む、ことを特徴とする請求項１０～１２、１４のいずれか一に記載の声質変換方法。
【請求項１７】
前記音声認識された目標話者音声の、少なくとも時刻情報に対応づいた音素列を含む第２のセグメンテーション情報によって、前記時間軸上の対応付けを行い、前記第２のセグメンテーション情報に基づき、前記合成音を生成するステップを含む、ことを特徴とする請求項１０～１２、１４のいずれか一に記載の声質変換方法。
【請求項１８】
声質変換装置により声質を変換する方法であって、
声質の変換目標となる話者の音声（「目標話者音声」という）を入力するステップと、
入力される前記目標話者音声を音声認識するステップと、
前記音声認識の結果から前記目標話者音声の発声内容を記述する発音記号列を生成するステップと、
を含み、さらに、
（ａ）記憶部に予め記憶されている音声合成用のデータを用いて、前記発音記号列から、合成音を作成するステップと、
（ｂ）前記合成音と前記目標話者音声とを分析し、それぞれの特徴パラメータを抽出するステップと、
（ｃ）前記目標話者音声の特徴パラメータと前記合成音の特徴パラメータとの時間軸上の対応付けを行うステップと、
（ｄ）対応付けされた２つの前記特徴パラメータに基づいて、前記合成音のスペクトル形状を前記目標話者音声のスペクトル形状に変換する変換関数を生成するステップと、

（ｅ）前記生成された変換関数を用いて変換対象となる音声の声質を変換するステップと、
（ｆ）前記声質の変換結果を、前記音声合成用のデータとして前記記憶部に格納するステップと、
を所定の収束条件に至るまで繰り返すことを特徴とする声質変換方法。
【請求項１９】
声質変換装置を構成するコンピュータに、
声質の変換目標となる話者の音声（「目標話者音声」という）を入力する処理と、
前記目標話者音声の発声内容を記述する発音記号列を入力し、記憶部に予め記憶されている音声合成用のデータを用いて、前記発音記号列から、合成音を作成する処理と、
前記合成音と前記目標話者音声とを分析し、それぞれの特徴パラメータを抽出する処理と、
前記目標話者音声の特徴パラメータと前記合成音の特徴パラメータとの時間軸上の対応付けを行う処理と、
対応付けがなされた前記目標話者音声及び前記合成音の特徴パラメータに基づき、前記合成音のスペクトル形状を、前記目標話者音声のスペクトル形状に、変換するための変換関数を生成する処理と、
を実行させるプログラム。
【請求項２０】
請求項１９記載のプログラムにおいて、
前記変換関数を用いて、変換対象となる音声信号を前記目標話者音声の声質を持つ音声信号に変換する処理を、さらに前記コンピュータに実行させるプログラム。
【請求項２１】
請求項１９記載のプログラムにおいて、
入力される前記目標話者音声を音声認識する処理と、
前記音声認識の結果から前記発音記号列を生成する処理とを、前記コンピュータに実行させるプログラム。
【請求項２２】
請求項１９～２１のいずれか一に記載のプログラムにおいて、
前記合成音の少なくとも時刻情報に対応づいた音素列を含むセグメンテーション情報を入力して、前記セグメンテーション情報に基づき前記合成音を生成する処理を、前記コンピュータに実行させるプログラム。
【請求項２３】
請求項１９～２１のいずれか一に記載のプログラムにおいて、
前記合成音の生成時に用いられる、少なくとも時刻情報に対応づいた音素列を含む第１のセグメンテーション情報によって、前記時間軸上の対応付けを行う処理を、前記コンピュータに実行させるプログラム。
【請求項２４】
請求項１９～２１、２３のいずれか一に記載のプログラムにおいて、
前記音声認識された目標話者音声の、少なくとも時刻情報に対応づいた音素列を含む第２のセグメンテーション情報によって、前記時間軸上の対応付けを行う処理を、前記コンピュータに実行させるプログラム。
【請求項２５】
請求項１９～２１、２３のいずれか一に記載のプログラムにおいて、
前記音声認識された目標話者音声の、少なくとも時刻情報に対応づいた音素列を含む第２のセグメンテーション情報に基づき、前記合成音を生成する処理を、前記コンピュータに実行させるプログラム。
【請求項２６】
請求項１９～２１、２３のいずれか一に記載のプログラムにおいて、
前記音声認識された目標話者音声の、少なくとも時刻情報に対応づいた音素列を含む第

２のセグメンテーション情報によって、前記時間軸上の対応付けを行い、前記第２のセグメンテーション情報に基づき前記合成音を生成する処理を、前記コンピュータに実行させるプログラム。

【請求項２７】

声質変換装置を構成するコンピュータに、

声質の変換目標となる話者の音声（「目標話者音声」という）を入力する処理と、

入力される前記目標話者音声を音声認識する処理と、

前記音声認識の結果から前記目標話者音声の発声内容を記述する発音記号列を生成する処理と、を実行させ、さらに、

（ａ）記憶部に予め記憶されている音声合成用のデータを用いて、前記発音記号列から合成音を作成する処理と、

（ｂ）前記合成音と前記目標話者音声とを分析し、それぞれの特徴パラメータを抽出する処理と、

（ｃ）前記目標話者音声の特徴パラメータと前記合成音の特徴パラメータとの時間軸上の対応付けを行う処理と、

（ｄ）対応付けされた２つの前記特徴パラメータに基づいて、前記合成音のスペクトル形状を前記目標話者音声のスペクトル形状に変換する変換関数を生成する処理と、

（ｅ）前記生成された変換関数を用いて変換対象となる音声の声質を変換する処理と、

（ｆ）前記声質の変換結果を、前記音声合成用のデータとして前記記憶部に格納する処理と、

を所定の収束条件に至るまで繰り返す、

処理を実行させるプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、声質変換装置および声質変換方法ならびに声質変換プログラムに関する。

【背景技術】

【０００２】

従来から、ある話者が発声した音声を、別の話者の声質を持つ音声へと変換する、声質変換技術についての研究がなされてきた。例えば、特許文献１において、話者Ａの声質を話者Ｂの声質へと変換する技術が開示されている。特許文献１における声質変換法を図３８に示す。この声質変換法では、話者Ａの音声をＬＰＣ分析１０１によって分析し、話者Ａのコードブック１０３を用い、ベクトル量子化１０５によって量子化する。また、話者Ｂの音声をＬＰＣ分析１０２によって分析し、話者Ｂのコードブック１０４を用い、ベクトル量子化１０６によって量子化する。それぞれ量子化されたデータは、時間軸の整合をＤＰマッチングによる対応付け１０７によって対応付け（写像）がなされ、写像を元に、ヒストグラムを用いた変換コードブック作成１０８によってスペクトル変換コードブック１０９を作成する。話者Ａの声質がスペクトル変換コードブック１０９を用いて声質変換される。

【０００３】

さらに近年では、一人ないし複数の話者の音声データから生成ならびに蓄積された音声合成用データベースを用いて、入力されたテキストの内容を音声として出力する音声合成装置において、所望の声質を持つ音声を合成するための声質変換技術も研究されている。この方法の利点は、ある声質を持つ合成音を生成する際に、その声質を持つ話者の音声データをその都度録音してデータベースを生成ならびに蓄積する必要がなく、１人ないし複数の標準話者データベースを予め蓄積しておけば、そのデータベースを目的の声質を持つように声質変換することで、所望の声質を持つ合成音が生成できるという点にある。

【０００４】

例えば、特許文献２において、予め記憶してある音声データベースと目標話者の音声との間の写像コードブックを作成し、これを変換関数として変換された音声データベースを

テキスト音声合成用データベースとして用いる技術が開示されている。特許文献２における声質変換法を図３９に示す。複数の登録話者の音響特徴パラメータを含む音声データベースを予め記憶する。この声質変換装置は、複数の登録話者の音響特徴パラメータを含む音声データベース２０３とそのコードブック２０４を予め記憶しておく。入力された目標話者の少なくとも１単語の音声信号に基づいて、声質変換をすべき目標話者に最も近い話者を、複数の登録話者の中から選択する選択手段２０１と、選択手段２０１によって選択された話者の音響空間と目標話者の音響空間との間の差分を計算することにより、選択された話者から目標話者への写像コードブック２０５を計算する生成手段２０２と、入力された音声合成すべき文字列に基づいて、音声データベース２０３に記憶された選択された話者の音声の音響特徴パラメータを上記選択された話者のコードブックを用いて量子化し、選択された話者のコードブックと写像コードブックの対応関係に基づいて文字列に対応する目標話者の音声信号の音響特徴パラメータを生成する写像処理手段２０６と、写像処理手段２０６によって生成された目標話者の音声信号の音響特徴パラメータに基づいて、文字列に対応する目標話者の音声信号を発生して出力する音声合成手段２０７とを備えている。また、生成手段２０２は、移動ベクトル場平滑化法を用いて、選択された話者から目標話者への写像コードブックを計算するようにしている。

【０００５】

なお、音声信号処理、音声認識の一般的な技術については、非特許文献１、非特許文献２において解説されている。

【０００６】

【特許文献１】特開平１－９７９９７号公報　（図４）
【特許文献２】特開平８－２４８９９４号公報　（図１）
【非特許文献１】古井貞熙著、「ディジタル音声処理」、東海大学出版会、１９８５年
【非特許文献２】中川聖一著、「確率モデルによる音声認識」、電子情報通信学会、１９８８年

【発明の開示】

【発明が解決しようとする課題】

【０００７】

特許文献１に開示されている声質変換方法では、変換元話者Ａから変換目標話者Ｂへと声質変換する場合に、ＡとＢが全く同一内容の発声をしている原音の音声データ（あるいは特徴量データ）を用いて変換関数を作成する必要があった。このため、この方法で声質変換を実施するためには、話者Ａと話者Ｂの全く同一内容の音声データがその都度必要となってしまい、自由な声質変換ができなかった。

【０００８】

また、特許文献２に開示されている声質変換および音声合成方法では、目標話者の音声の音響空間と音声データベースの音響空間との間の差分を取って移動ベクトル場平滑化法を用いて写像コードブックを求めるため、データベースの情報量が目標話者の情報量に比べ圧倒的に多い場合、非常に疎な対応付けを平滑化して全体の対応付けとする必要があった。このため、対応付けの精度が低く、変換後の音声の音質が劣化する、あるいは目標音声の声質との類似度が低いという問題点が生じた。

【０００９】

したがって、本発明の主たる目的は、変換目標話者の発声内容に依存しない自由な声質変換を実現する装置および方法ならびにプログラムを提供することにある。

【００１０】

本発明の他の目的は、所望の声質へ高精度に変換できる変換関数の同定を可能とする装置および方法ならびにプログラムを提供することにある。

【００１１】

本発明のさらに他の目的は、変換目標話者音声と同一あるいは類似の発声内容を持つ合成音の自動生成を可能とする装置および方法ならびにプログラムを提供することにある。

【課題を解決するための手段】

【００１２】
前記目的を達成するために、本発明に係る声質変換装置は、第１のアスペクトによれば、変換目標となる話者の音声（「目標話者音声」という）を入力する目標話者音声入力部と、音声合成用のデータを記憶する音声合成用データ記憶部と、入力された目標話者音声と同一又は類似の発声内容を記述する発音記号列を入力する発音記号入力部と、を備える。また、入力された発音記号列にしたがって、音声合成用データ記憶部に記憶されている音声合成用のデータに基づき、音声を合成して出力する音声合成部と、音声合成部から出力される音声信号（「合成音」という）の特徴パラメータを抽出する合成音特徴パラメータ抽出部と、目標話者音声の特徴パラメータを抽出する目標話者音声特徴パラメータ抽出部と、を備える。さらに、合成音特徴パラメータ抽出部からの合成音の特徴パラメータと、目標話者音声特徴パラメータ抽出部からの目標話者音声の特徴パラメータとを入力し、特徴パラメータを表す空間において、合成音の第１の部分と、目標話者音声の第２の部分との時間軸上の対応関係を求め、合成音の第１の部分におけるスペクトル形状を、目標話者音声の第２の部分におけるスペクトル形状に変換する変換関数の同定を行う変換関数生成部と、を備える構成とされる。
【００１３】
また、本発明に係る声質変換方法は、第２のアスペクトによれば、声質変換装置により声質を変換する方法である。声質変換方法は、声質の変換目標となる話者の音声（「目標話者音声」という）を入力するステップと、目標話者音声の発声内容を記述する発音記号列を入力し、記憶部に予め記憶されている音声合成用のデータを用いて、発音記号列から、合成音を作成するステップと、を含む。また、合成音と目標話者音声とを分析し、それぞれの特徴パラメータを抽出するステップと、目標話者音声の特徴パラメータと合成音の特徴パラメータとの時間軸上の対応付けを行うステップと、対応付けがなされた目標話者音声及び合成音の特徴パラメータに基づき、合成音のスペクトル形状を、目標話者音声のスペクトル形状に、変換するための変換関数を生成するステップと、を含む。
【００１４】
さらに、本発明に係る声質変換方法は、第３のアスペクトによれば、声質変換装置により声質を変換する方法である。声質変換方法は、声質の変換目標となる話者の音声（「目標話者音声」という）を入力するステップと、入力される目標話者音声を音声認識するステップと、音声認識の結果から目標話者音声の発声内容を記述する発音記号列を生成するステップと、を含む。さらに、（ａ）記憶部に予め記憶されている音声合成用のデータを用いて、発音記号列から、合成音を作成するステップと、（ｂ）合成音と目標話者音声とを分析し、それぞれの特徴パラメータを抽出するステップと、（ｃ）目標話者音声の特徴パラメータと合成音の特徴パラメータとの時間軸上の対応付けを行うステップと、（ｄ）対応付けされた２つの特徴パラメータに基づいて、合成音のスペクトル形状を目標話者音声のスペクトル形状に変換する変換関数を生成するステップと、（ｅ）生成された変換関数を用いて変換対象となる音声の声質を変換するステップと、（ｆ）声質の変換結果を、音声合成用のデータとして記憶部に格納するステップと、を収束条件に至るまで繰り返す。
【００１５】
また、本発明に係る声質変換プログラムは、第４のアスペクトによれば、声質変換装置を構成するコンピュータに実行させるプログラムである。このプログラムは、声質の変換目標となる話者の音声（「目標話者音声」という）を入力する処理と、目標話者音声の発声内容を記述する発音記号列を入力し、記憶部に予め記憶されている音声合成用のデータを用いて、発音記号列から、合成音を作成する処理と、を実行させる。また、合成音と目標話者音声とを分析し、それぞれの特徴パラメータを抽出する処理と、目標話者音声の特徴パラメータと合成音の特徴パラメータとの時間軸上の対応付けを行う処理と、対応付けがなされた目標話者音声及び合成音の特徴パラメータに基づき、合成音のスペクトル形状を、目標話者音声のスペクトル形状に、変換するための変換関数を生成する処理と、を実行させる。

【００１６】
さらに、本発明に係る声質変換プログラムは、第５のアスペクトによれば、声質変換装置を構成するコンピュータに実行させるプログラムである。このプログラムは、声質の変換目標となる話者の音声（「目標話者音声」という）を入力する処理と、入力される目標話者音声を音声認識する処理と、音声認識の結果から目標話者音声の発声内容を記述する発音記号列を生成する処理と、を実行させる。さらに、（ａ）記憶部に予め記憶されている音声合成用のデータを用いて、発音記号列から合成音を作成する処理と、（ｂ）合成音と目標話者音声とを分析し、それぞれの特徴パラメータを抽出する処理と、（ｃ）目標話者音声の特徴パラメータと合成音の特徴パラメータとの時間軸上の対応付けを行う処理と、（ｄ）対応付けされた２つの特徴パラメータに基づいて、合成音のスペクトル形状を目標話者音声のスペクトル形状に変換する変換関数を生成する処理と、（ｅ）生成された変換関数を用いて変換対象となる音声の声質を変換する処理と、（ｆ）声質の変換結果を、音声合成用のデータとして記憶部に格納する処理と、を収束条件に至るまで繰り返す処理を実行させる。
【発明の効果】
【００１７】
本発明によれば、目標話者音声と同一あるいは類似の発声内容を持つ変換元音声を予め用意することなく、目標話者音声の声質への変換を行うことができる。したがって、使用者の負担が軽減する。この理由は、目標話者音声と同一あるいは類似の発声内容を表す発音記号列を入力することで、目標話者音声と同一あるいは類似の発声内容を持つ合成音を生成することができるためである。
【００１８】
また、学習用音声の対応付けおよび声質変換を行うための変換関数の同定が高精度にできる。したがって、従来に比べて所望の声質に近い声質を持つ高音質な音声が得られる。この理由の１つは、合成音作成に使用されるデータを目標話者音声の情報から推定することによって、目標話者音声への変換関数を同定しやすいように合成音を作成することができるためである。もう１つの理由は、合成音作成時に使用されたデータを記憶しておくことによって、このデータを音声から抽出する特徴パラメータに加えて変換関数生成に用いることができるためである。
【００１９】
さらに、目標話者音声と同一あるいは類似の発声内容を表す発音記号列を予め用意することなく、目標話者音声と同一あるいは類似の発声内容の合成音が作成できる。したがって、処理が自動化し、処理速度が向上する。この理由は、目標話者音声を音声認識することによって、目標話者音声と同一あるいは類似の発声内容を表す発音記号列を自動的に生成することができるためである。
【発明を実施するための最良の形態】
【００２０】
［第１の実施形態］
図１に、本発明の第１の実施形態に係る声質変換装置のブロック図を示す。声質変換装置は、目標話者音声入力部１１と、発音記号列入力部１２と、音声合成用データ記憶部１３と、音声合成部１４と、目標話者音声特徴パラメータ抽出部１５と、合成音特徴パラメータ抽出部１６と、変換関数生成部１７と、声質変換部１８とを備えている。
【００２１】
目標話者音声入力部１１は、目標とする声質を持つ音声データを入力する。発音記号列入力部１２は、作成したい音声の発声内容が記述されている発音記号列を入力する。この時、入力される発音記号列は、目標話者音声入力部１１から入力される目標話者音声と同一あるいは類似の発声内容の音声を合成するように記述してあるものとする。音声合成用データ記憶部１３は、音声合成部１４で用いる、音声や音節等の単位、時間長情報等のデータを記憶している。音声合成部１４は、発音記号列に対応するデータにしたがって、音声合成用データ記憶部１３に蓄えられたデータから合成用のデータを算出し、これを用い

て音声を合成して出力する。

【００２２】

また、目標話者音声特徴パラメータ抽出部１５は、目標話者音声入力部１１から入力された音声データに対しスペクトル分析を施して、特徴パラメータを抽出する。合成音特徴パラメータ抽出部１６は、音声合成部１４から入力された音声データに対しスペクトル分析を施して、特徴パラメータを抽出する。ここで、目標話者音声および合成音から抽出する特徴パラメータとしては、例えば非特許文献１に記載してあるようなＬＰＣ（linear predictive Coding）係数、フォルマント周波数ならびにバンド帯域幅、韻律データ等の内、少なくとも１つを含む。変換関数生成部１７は、目標話者音声から抽出された特徴パラメータおよび合成音から抽出された特徴パラメータによって、合成音の声質から目標話者音声の声質へと変換する関数を同定する。声質変換部１８は、同定された変換関数を用いて、変換対象である被変換入力信号を変換して変換後出力信号を出力する。

【００２３】

次に、本発明の第１の実施形態に係る声質変換装置の動作を図１および図１３を参照して説明する。図１３は、本発明の第１の実施形態に係る声質変換装置の動作を示すフローチャートである。まず、ユーザが望む声質を持つ話者の音声データを目標音声として用意し、目標話者音声入力部１１に入力する（ステップＳ１１）。次に、目標話者音声入力部１１に入力した目標話者音声と同一あるいは類似の発声内容を持つように発音記号列を作成し、発音記号列入力部１２に入力する。入力された発音記号列に対応するデータにしたがって音声合成用データ記憶部１３から合成用のデータを算出し、音声合成部１４において合成音を生成する。この時、目標話者音声と合成音の発声の時間長は、ずれていても構わない。音声合成用データと発音記号との対応付けの規則は、発音記号が指定されると音声合成用データ記憶部１３に記憶されている音声素片データや時間長情報等の音声合成用データから当該発音記号に対応した音声合成用データを算出するように予め作成されている（ステップＳ１２）。

【００２４】

また、目標話者音声特徴パラメータ抽出部１５において、目標話者音声に対し分析を行い、目標話者音声の特徴パラメータを抽出する。また、合成音特徴パラメータ抽出部１６において、合成音に対し分析を行い、合成音の特徴パラメータを抽出する（ステップＳ１３）。合成音と目標音声の時間軸のずれを修正するために、目標話者音声から抽出したパラメータと合成音から抽出した特徴パラメータを用いて、合成音と目標音声の間で時間軸の対応付けを行う。対応付けの方法としては、ＤＰマッチングによる時間軸伸縮等が考えられる（ステップＳ１４）。全学習データ（目標話者音声と合成音の組）の時間軸の対応付けが行われた後、この時間軸対応付け済みの特徴パラメータを用いて、変換関数生成部１７で合成音の声質から目標話者音声の声質へと変換する関数を作成する（ステップＳ１５）。

【００２５】

さらに、被変換入力信号のどの部分にどのような変換関数が適用されるかを決定するために、被変換入力信号（素片データ）と作成された変換関数との対応付けを行う（ステップＳ１６）。被変換入力信号に対応付けられた変換関数を用いて、被変換入力信号を変換する（ステップＳ１７）。変換されたデータを変換後出力信号として出力する（ステップＳ１８）。

【００２６】

本発明の第１の実施形態によれば、目標話者音声と同一あるいは類似の発声内容の合成音を音声合成で生成してから分析するため、目標話者音声の発声内容にかかわらず、高精度の対応付けを行ったうえで変換関数を生成することが可能となる。このため、変換目標話者の発声内容に依存しない自由な声質変換を実現することができる。したがって、目標話者音声の発声内容に合わせて、その都度変換元音声を入力する必要がなくなるので、使用者の負担を軽減し、処理を迅速化することができる。

【００２７】

［第２の実施形態］
図２に、本発明の第２の実施形態に係る声質変換装置のブロック図を示す。図２において、図１と同一の符号は、同一部あるいは相当部を示し、その説明を省略する。第２の実施形態に係る声質変換装置は、第１の実施形態に係る声質変換装置に、さらに音声認識部１９を備えている。音声認識部１９は、目標話者音声入力部１１ａで入力された目標話者音声を音声認識し、目標話者音声の発声内容を発音記号とともに発声記号列入力部１２ａへ出力する。
【００２８】
次に、本発明の第２の実施形態に係る声質変換装置の動作を図２および図１４を参照して説明する。ただし、図１４において、ステップＳ２１およびＳ２３～Ｓ２８は、それぞれ図１３のＳ１１およびＳ１３～Ｓ１８と同様の動作をするため、新たに追加したステップの動作のみを説明する。ステップＳ２１で目標話者音声入力部１１ａから入力された目標話者音声を音声認識部１９で音声認識して（ステップＳ２２１）発音記号列を生成し（ステップＳ２２２）、生成された発音記号列を用いて合成音を作成する（ステップＳ２２３）。
【００２９】
本発明の第２の実施形態によれば、目標話者音声の音声認識結果を合成音生成のための発音記号列として入力するため、ユーザがテキスト等で発音記号列を入力する必要が無く、処理の自動化を図ることが可能となる。本発明は、目標話者音声として多くの文を用いて対応付けの高精度化を図る際に有効である。
【００３０】
［第３の実施形態］
本発明の第３の実施形態に係る声質変換装置は、第１の実施形態に係る声質変換装置の構成に加え、音声合成部で使用したセグメンテーション情報を記憶し、変換関数生成部に出力するセグメンテーション情報記憶部により構成される。このような構成を採用し、変換関数生成部でセグメンテーション情報を用いることにより、所望の声質へ高精度に変換できる変換関数の同定が可能である。なお、以下の説明において、セグメンテーション情報とは、時刻情報に対応づいた音素（音素ラベル）を主とし、ピッチパターンあるいはパワー等の韻律情報を付加したものでもよい。
【００３１】
図３に、本発明の第３の実施形態に係る声質変換装置のブロック図を示す。図３において、目標話者音声入力部１１と、発音記号列入力部１２と、音声合成用データ記憶部１３と、目標話者音声特徴パラメータ抽出部１５と、合成音特徴パラメータ抽出部１６と、声質変換部１８は、図１と同等であり、その説明を省略する。本発明の第３の実施形態に係る声質変換装置は、さらに合成音セグメンテーション情報記憶部２０を備えている。合成音セグメンテーション情報記憶部２０は、音声合成部１４ａで発音記号列に従って音声を合成する際に用いられるセグメンテーション情報を記憶して、変換関数生成部１７ａへと出力する。
【００３２】
次に、本発明の第３の実施形態に係る声質変換装置の動作を図３および図１５を参照して説明する。ただし、図１５において、図１３の同一の符号は、同一あるいは同等の処理を行い、その説明を省略し、新たに追加したステップの動作のみを説明する。ステップＳ１２ａでは、音声合成部１４で音声を合成する際に用いられたセグメンテーション情報２３ａを合成音セグメンテーション情報記憶部２０に記憶し、変換関数生成部１７ａへ出力する。変換関数生成部１７ａに入力されたセグメンテーション情報２３ａは、時間軸の対応付け（ステップＳ１４ａ）、および変換関数の作成（ステップＳ１５ａ）で用いられる。
【００３３】
本発明の第３の実施形態によれば、合成音を生成する際に利用するセグメンテーション情報を変換関数生成時に活用するため、合成音特徴パラメータと目標話者音声特徴パラメ

ータとを対応付ける際の処理の簡素化、高精度化を図ることが可能となる。

【００３４】

［第４の実施形態］

図４に、本発明の第４の実施形態に係る声質変換装置のブロック図を示す。図４において、図２と同一の符号は、同一部あるいは相当部を示し、その説明を省略する。本発明の第４の実施形態に係る声質変換装置は、図２に対し、さらに合成音セグメンテーション情報記憶部２０を備えている。なお、合成音セグメンテーション情報記憶部２０は、図３で説明したものと同様である。

【００３５】

次に、本発明の第４の実施形態に係る声質変換装置の動作を図４および図１６を参照して説明する。ただし、図１６において、図１３の同一の符号は、同一あるいは同等の処理を行い、その説明を省略し、新たに追加したステップの動作のみを説明する。ステップＳ２２３ａでは、音声合成部１４ａで音声を合成する際に用いられたセグメンテーション情報２３ｂを出力し、変換関数生成部１７ａへ入力する。入力されたセグメンテーション情報２３ｂは、時間軸の対応付け（ステップＳ２４ａ）、および変換関数の作成（ステップＳ２５ｂ）で用いられる。

【００３６】

本発明の第４の実施形態によれば、目標話者音声を音声合成して発音記号列を自動生成し、これを用いて合成音を生成する際に利用するセグメンテーション情報を変換関数生成時に活用するため、処理の自動化を図りつつ、合成音特徴パラメータと目標話者音声特徴パラメータとを対応付ける際の処理の簡素化、高精度化を図ることが可能となる。

【００３７】

［第５の実施形態］

図５に、本発明の第５の実施形態に係る声質変換装置のブロック図を示す。図５において、図１と同一の符号は、同一部あるいは相当部を示し、その説明を省略する。本発明の第５の実施形態に係る声質変換装置は、さらにセグメンテーション情報入力部２２を備えている。セグメンテーション情報入力部２２は、目標話者音声入力部１１で入力された目標話者音声を不図示の手段で分析してセグメンテーション情報２３ｃを抽出し、音声合成部１４ｂに出力する。

【００３８】

次に、本発明の第５の実施形態に係る声質変換装置の動作を図５および図１７を参照して説明する。ただし、図１７において、図１３の同一の符号は、同一あるいは同等の処理を行い、その説明を省略し、新たに追加したステップの動作のみを説明する。ステップＳ１２ｂでは、目標話者音声入力部１１で入力された目標話者音声に基づいたセグメンテーション情報２３ｃを入力する。入力されたセグメンテーション情報２３ｃは、音声合成部１４ｂに出力され、発音記号列入力部１２で入力された発音記号列の発声内容を持ち、発声の時間長が目標話者音声と同一である音声が音声合成部１４ｂにおいて合成される。

【００３９】

本発明の第５の実施形態によれば、目標話者音声に基づいたセグメンテーション情報を入力して合成音を生成するため、生成された合成音の時間長情報は当該目標話者音声の時間長情報と等しくなり、変換関数生成部での時間軸の対応付け処理を不必要とする、あるいは簡素化することが可能となる。

【００４０】

［第６の実施形態］

図６に、本発明の第６の実施形態に係る声質変換装置のブロック図を示す。図６において、図２と同一の符号は、同一部あるいは相当部を示し、その説明を省略する。本発明の第６の実施形態に係る声質変換装置は、さらにセグメンテーション情報入力部２２を備えている。セグメンテーション情報入力部２２は、目標話者音声入力部１１ａで入力された目標話者音声を不図示の手段で分析してセグメンテーション情報２３ｃを抽出し、音声合成部１４ｂに出力する。

【００４１】
次に、本発明の第６の実施形態に係る声質変換装置の動作を図６および図１８を参照して説明する。ただし、図１８において、図１３の同一の符号は、同一あるいは同等の処理を行い、その説明を省略し、新たに追加したステップの動作のみを説明する。ステップＳ２２２ａでは、目標話者音声入力部１１ａで入力された目標話者音声に基づいたセグメンテーション情報２３ｃを入力する。入力されたセグメンテーション情報２３ｃは音声合成部１４ｂに出力され、発音記号列入力部１２ａで入力された発音記号列の発声内容を持ち、発声の時間長が目標話者音声と同一である音声が合成される。
【００４２】
本発明の第６の実施形態によれば、目標話者音声に基づいたセグメンテーション情報を入力して合成音を生成するため、生成された合成音の時間長情報は当該目標話者音声の時間長情報と等しくなり、変換関数生成部での時間軸の対応付け処理を不必要とする、あるいは簡素化することが可能となる。さらに、本発明では、音声認識によって目標話者音声から発音記号列を生成しているため、処理の自動化を図りつつ、発音記号列と目標話者音声から抽出される時間長情報との整合性が取りやすくすることが可能となる。
【００４３】
［第７の実施形態］
図７に、本発明の第７の実施形態に係る声質変換装置のブロック図を示す。図７において、図２と同一の符号は、同一部あるいは相当部を示し、その説明を省略する。本発明の第７の実施形態に係る声質変換装置は、さらに目標話者音声セグメンテーション情報記憶部２１を備えている。目標話者音声セグメンテーション情報記憶部２１は、目標話者音声入力部１１ａで入力された目標話者音声のセグメンテーション情報を記憶して、変換関数生成部１７ｂに出力する。
【００４４】
次に、本発明の第７の実施形態に係る声質変換装置の動作を図７および図１９を参照して説明する。ただし、図１９において、図１４の同一の符号は、同一あるいは同等の処理を行い、その説明を省略し、新たに追加したステップの動作のみを説明する。ステップＳ２２１ａでは、目標話者音声信号を音声認識部１９で音声認識した結果、抽出されるセグメンテーション情報２３ｄを出力する。セグメンテーション情報２３ｄは、変換関数生成部１７ｂへ出力され、時間軸の対応付け（ステップＳ２４ａ）、および変換関数の作成（ステップＳ２５ａ）で用いられる。
【００４５】
本発明の第７の実施形態によれば、目標話者音声を音声認識する際に同時に目標話者音声のセグメンテーション情報を抽出して変換関数生成部に入力するため、目標話者音声の特徴パラメータに加えてセグメンテーション情報を変換関数生成に用いることができ、合成音特徴パラメータと目標話者音声特徴パラメータとを対応付ける際の処理の簡素化、高精度化を図ることが可能となる。
【００４６】
［第８の実施形態］
図８に、本発明の第８の実施形態に係る声質変換装置のブロック図を示す。図８において、図２と同一の符号は、同一部あるいは相当部を示し、その説明を省略する。本発明の第８の実施形態に係る声質変換装置は、さらに目標話者音声セグメンテーション情報記憶部２１ａを備えている。目標話者音声セグメンテーション情報記憶部２１ａは、目標話者音声入力部１１ａで入力された目標話者音声のセグメンテーション情報を記憶して、音声合成部１４ｂに出力する。
【００４７】
次に、本発明の第８の実施形態に係る声質変換装置の動作を図８および図２０を参照して説明する。ただし、図２０において、図１４の同一の符号は、同一あるいは同等の処理を行い、その説明を省略し、新たに追加したステップの動作のみを説明する。ステップＳ２２１ａでは、目標話者を音声認識部１９で音声認識した結果抽出されるセグメンテーシ

ョン情報２３ｅを出力する。ステップＳ２２３ｂにおいて、音声合成部１４ｂは、セグメンテーション情報２３ｅを入力し、発音記号列入力部１２ａで入力された発音記号列の発声内容を持ち、目標話者音声のセグメンテーション情報２３ｅに基づいた音声を合成する。

【００４８】

本発明の第８の実施形態によれば、目標話者音声を音声認識する際に同時に同音声のセグメンテーション情報を抽出して音声合成部に入力するため、目標話者音声に基づいたセグメンテーション情報を用いて合成音を生成することができ、合成音特徴パラメータと目標話者音声特徴パラメータとを対応付ける際の処理の簡素化、高精度化を図ることが可能となる。

【００４９】

［第９の実施形態］

図９に、本発明の第９の実施形態に係る声質変換装置のブロック図を示す。図９において、図２と同一の符号は、同一部あるいは相当部を示し、その説明を省略する。本発明の第９の実施形態に係る声質変換装置は、さらに目標話者音声セグメンテーション情報記憶部２１ｂを備えている。目標話者音声セグメンテーション情報記憶部２１ｂは、目標話者音声入力部１１ａで入力された目標話者音声のセグメンテーション情報を記憶して、音声合成部１４ｂおよび変換関数生成部１７ｂに出力する。

【００５０】

次に、本発明の第９の実施形態に係る声質変換装置の動作を図９および図２１を参照して説明する。ただし、図２１において、図１４の同一の符号は、同一あるいは同等の処理を行い、その説明を省略し、新たに追加したステップの動作のみを説明する。ステップＳ２２１ａでは、目標話者を音声認識部１９で音声認識した結果抽出されるセグメンテーション情報２３ｆを出力する。セグメンテーション情報２３ｆは、音声合成部１４ｂへ出力され、発音記号列入力部１２ａで入力された発音記号列の発声内容を持ち、目標話者音声のセグメンテーション情報２３ｆに基づいた音声を合成する。さらに、セグメンテーション情報２３ｆは、変換関数生成部１７ｂにも出力され、時間軸の対応付け（ステップＳ２４ａ）、および変換関数の作成（ステップＳ２５ａ）で用いられる。

【００５１】

本発明の第９の実施形態によれば、目標話者音声を音声認識する際に同時に同音声のセグメンテーション情報を抽出して音声合成部および変換関数生成部に入力するため、目標話者音声に基づいたセグメンテーション情報を用いて合成音を生成することができ、かつ、目標話者音声の特徴パラメータに加えてセグメンテーション情報を変換関数生成に用いることができる。したがって、合成音特徴パラメータと目標話者音声特徴パラメータとを対応付ける際の処理の簡素化、高精度化を図ることが可能となる。

【００５２】

［第１０の実施形態］

図１０に、本発明の第１０の実施形態に係る声質変換装置のブロック図を示す。図１０において、図２と同一の符号は、同一部あるいは相当部を示し、その説明を省略する。本発明の第１０の実施形態に係る声質変換装置は、さらに合成音セグメンテーション情報記憶部２０と目標話者音声セグメンテーション情報記憶部２１を備えている。合成音セグメンテーション情報記憶部２０は、図３で説明した合成音セグメンテーション情報記憶部２０と同じである。また目標話者音声セグメンテーション情報記憶部２１は、図７で説明した目標話者音声セグメンテーション情報記憶部２１と同様のものである。

【００５３】

次に、本発明の第１０の実施形態に係る声質変換装置の動作を図１０および図２２を参照して説明する。ただし、図２２において、図１３の同一の符号は、同一あるいは同等の処理を行い、その説明を省略し、新たに追加したステップの動作のみを説明する。ステップＳ２２１ａでは、目標話者を音声認識部１９で音声認識した結果抽出されるセグメンテーション情報２３ｄを出力する。セグメンテーション情報２３ｄは、変換関数生成部１７

ｃに出力される。また、ステップＳ２２３ａでは、音声合成部１４ａで音声を合成する際に用いられたセグメンテーション情報２３ｇを記憶し、変換関数生成部１７ｃへ出力される。変換関数生成部１７ｃに入力されたセグメンテーション情報２３ｆおよび２３ｇは時間軸の対応付け（ステップＳ２４ｂ）、および変換関数の作成（ステップＳ２５ｂ）で用いられる。

【００５４】

本発明の第１０の実施形態によれば、変換関数生成部において合成音と目標話者音声両方のセグメンテーション情報を利用できるため、合成音および目標話者音声の特徴パラメータに加えてセグメンテーション情報を変換関数生成に用いることができ、合成音特徴パラメータと目標話者音声特徴パラメータとを対応付ける際の処理の簡素化、高精度化を図ることが可能となる。

【００５５】

［第１１の実施形態］

図１１に、本発明の第１１の実施形態に係る声質変換装置のブロック図を示す。図１１において、図２と同一の符号は、同一部あるいは相当部を示し、その説明を省略する。本発明の第１１の実施形態に係る声質変換装置は、さらに合成音セグメンテーション情報記憶部２０と目標話者音声セグメンテーション情報記憶部２１ａを備えている。合成音セグメンテーション情報記憶部２０および目標話者音声セグメンテーション情報記憶部２１ａは、図３の合成音セグメンテーション情報記憶部２０および図８の目標話者音声セグメンテーション情報記憶部２１ａと同様のものである。

【００５６】

次に、本発明の第１１の実施形態に係る声質変換装置の動作を図１１および図２３を参照して説明する。ただし、図２３において、図２０の同一の符号は、同一あるいは同等の処理を行い、その説明を省略し、新たに追加したステップの動作のみを説明する。ステップＳ２２３ｃでは、音声合成部１４ｃで音声を合成する際に用いられたセグメンテーション情報２３ｇを記憶し、変換関数生成部１７ａへ入力する。入力されたセグメンテーション情報２３ｇは時間軸の対応付け（ステップＳ２４ａ）、および変換関数の作成（ステップＳ２５ａ）で用いられる。

【００５７】

本発明の第１１の実施形態によれば、目標話者音声を音声認識する際に同時に同音声のセグメンテーション情報を抽出して音声合成部に入力するため、目標話者音声に基づいたセグメンテーション情報を用いて合成音を生成することができる。また、合成音を生成する際に利用するセグメンテーション情報を変換関数生成時に活用するため、合成音の特徴パラメータに加えてセグメンテーション情報を変換関数生成に用いることができる。したがって、合成音特徴パラメータと目標話者音声特徴パラメータとを対応付ける際の処理の簡素化、高精度化を図ることが可能となる。

【００５８】

［第１２の実施形態］

図１２に、本発明の第１２の実施形態に係る声質変換装置のブロック図を示す。図１２において、図２と同一の符号は、同一部あるいは相当部を示し、その説明を省略する。本発明の第１２の実施形態に係る声質変換装置は、さらに合成音セグメンテーション情報記憶部２０と目標話者音声セグメンテーション情報記憶部２１ｂを備えている。合成音セグメンテーション情報記憶部２０および目標話者音声セグメンテーション情報記憶部２１ｂは、それぞれ図３の合成音セグメンテーション情報記憶部２０および図９の目標話者音声セグメンテーション情報記憶部２１ｂと同様のものである。

【００５９】

次に、本発明の第１２の実施形態に係る声質変換装置の動作を図１２および図２４を参照して説明する。ただし、図２４において、図２１の同一の符号は、同一あるいは同等の処理を行い、その説明を省略し、新たに追加したステップの動作のみを説明する。ステップＳ２２３ｃでは、音声合成部１４ｃで音声を合成する際に用いられたセグメンテーショ

ン情報２３ｇを記憶し、変換関数生成部１７ｃへ入力する。入力されたセグメンテーション情報２３ｇは時間軸の対応付け（ステップＳ２４ｂ）、および変換関数の作成（ステップＳ２５ｂ）で用いられる。

【００６０】

本発明の第１２の実施形態によれば、目標話者音声を音声認識する際に同時に同音声のセグメンテーション情報を抽出して音声合成部および変換関数生成部に入力するため、目標話者音声に基づいたセグメンテーション情報を用いて合成音を生成することができる。また、目標話者音声の特徴パラメータに加えてセグメンテーション情報を変換関数生成に用いることができる。また、変換関数生成部において合成音と目標話者音声両方のセグメンテーション情報を利用できるため、合成音および目標話者音声の特徴パラメータに加えてセグメンテーション情報を変換関数生成に用いることができる。したがって、合成音特徴パラメータと目標話者音声特徴パラメータとを対応付ける際の処理の簡素化、高精度化を図ることが可能となる。

【実施例１】

【００６１】

図２５は、本発明の第１の実施例に係る声質変換装置のブロック図である。声質変換装置は、目標話者音声入力部１１、発音記号列入力部１２、音声合成用データ記憶部１３、音声合成部１４、ＬＰＣ分析部２５、２６、３０、ＤＰマッチング部２７、スペクトル形状変換部２８、変換関数導出部２９、対応フレーム探索部３１、スペクトル変換部３２を備える。なお、図２５において、図１と同一の符号は、同一物あるいは相当物を示し、特に記載無き場合、その説明を省略する。

【００６２】

本実施例では、あらかじめ少なくとも１人以上の話者から音声合成用のデータベースが作成されており、音声合成用データ記憶部１３に保存されているものとする。データベースは、音声素片、継続時間長、ピッチパターン等のデータを含んでいる。ただし、必ずしもこれらの全てのデータをデータベース内に記憶しておく必要はなく、これらのうち１つないし２つのデータだけでもよい。

【００６３】

今、目標話者音声の発声内容Ｃが明らかになっており、予めこの発声内容と同一あるいは類似の発声内容を持つ合成音を生成できるような発音記号列が発音記号列入力部１２から入力され用意されているとする。この発音記号列に従って、音声合成部１４は、音声合成用データ内の素片の声質を持った発声の合成音Ｂを合成する。目標話者音声入力部１１から入力される目標話者音声としては、操作開始時点で録音装置に発声する場合、予め保存してあった音声データを用いる場合等が考えられる。また、目標話者音声が複数あっても構わない。

【００６４】

次に、ＬＰＣ分析部２５は、目標話者音声入力部１１から入力された、話者Ａの発声内容Ｃの音声（以降では、目標Ａとする）を例えば２０ｍｓｅｃといった長さの分析フレームごとにＬＰＣ分析し、目標ＡのＬＰＣ係数３５ａを抽出する。また、ＬＰＣ分析部２６は、音声合成部１４で作成した合成音Ｂの発声内容Ｃの音声（以降では、合成音Ｂとする）を同様にＬＰＣ分析し、合成音ＢのＬＰＣ係数３５ｂを抽出する。本実施例では分析方法（特徴パラメータ抽出）として、ＬＰＣ分析を説明するが、他に、ＬＳＰ（line spectrum pair）分析、ＰＡＲＣＯＲ（partial auto-correlation）分析等のスペクトル分析や零交叉計数法、自己相関法等のピッチ抽出法、およびこれら複数の組み合わせによる分析等が考えられる。また、抽出する特徴パラメータとしてＬＰＣ係数のほかに、自己相関係数、ケプストラム係数等や、ピッチ周波数等の韻律パラメータを用いることもできる。

【００６５】

次に、声質変換の主要部となる、ＤＰマッチング部２７、スペクトル形状変換部２８、変換関数導出部２９、対応フレーム探索部３１、スペクトル変換部３２における処理の流れを説明する。図２６は、声質変換の主要部における音声データの処理の流れを説明する

図である。ＬＰＣ分析された目標話者音声（目標Ａ）と合成音Ｂとは、ＤＰマッチング部２７によってＤＰマッチングが行われて対応関係が求められる。求められた対応関係を用いて、合成音Ｂから目標Ａへの変換関数が、スペクトル形状変換部２８、および変換関数導出部２９により生成される。

【００６６】

変換対象となる音声素片は、対応フレーム探索部３１によって合成音素片との対応付けがなされる。この場合、対応付けは、例えば変換対象となる音声素片と合成音素片とのそれぞれのＬＰＣ係数での特徴空間で距離が小さくなるものを選択することでなされる。変換対象となる音声素片は、対応付けがされた合成音素片に対応する先に求めた変換関数によってスペクトル変換がなされ、目標の声質を持つ音声素片が生成されることとなる。以下に、各部についてより詳しく説明する。

【００６７】

ＤＰ（dynamic programming）マッチング部２７は、目標ＡのＬＰＣ係数３５ａと合成音ＢのＬＰＣ係数３５ｂとを用いて、目標Ａと合成音Ｂの時間軸を合わせるために、ＤＰ（dynamic programming）マッチングによる時間軸伸縮を行う。これにより、目標Ａと合成音Ｂとの分析フレームごとの対応が作成される。この時、ＤＰマッチングで用いる特徴量空間内距離を表す尺度としては、差分ベクトルの二乗和、ＷＬＲ（weighted likelihood ratio）尺度、最尤スペクトル距離、ＬＰＣケプストラム距離等を用いることができる。

【００６８】

図２７は、分析フレームごとのＤＰマッチングを模式的に表した図である。目標話者音声（目標Ａ）のフレームをＡ１、Ａ２、Ａ３、Ａ４、Ａ５、・・Ａｍ、・・とし、合成音Ｂのフレームを Ｂ１、Ｂ２、Ｂ３、Ｂ４、Ｂ５、・・Ｂｎ、・・とする。目標話者音声（目標Ａ）と合成音Ｂとをフレーム毎にＬＰＣ分析し、その上で、例えば、フレームＡ１とフレームＢ１との特徴量空間内距離、フレームＡ１とフレームＢ２との特徴量空間内距離、フレームＡ２とフレームＢ１との特徴量空間内距離、の中で最も距離の短い対応関係を選択する。ここでは、フレームＡ２とフレームＢ１との特徴量空間内距離が選択されたとすると、次には、例えば、フレームＡ３とフレームＢ１との特徴量空間内距離、フレームＡ２とフレームＢ２との特徴量空間内距離、フレームＡ３とフレームＢ２との特徴量空間内距離、の中で最も距離の短い対応を選択する。このようにして、特徴量空間内距離の最小の距離の対応を順次求めていき、目標話者音声（目標Ａ）と合成音ＢとのＤＰマッチングが行われる。

【００６９】

次に、スペクトル形状変換部２８は、分析フレームごとに対応付けられた目標ＡのＬＰＣ係数３５ａと合成音ＢのＬＰＣ係数３５ｂとを用いて、合成音Ｂの声質が目標Ａの声質へと変換されるような変換関数を同定する。合成音Ｂから目標Ａへと声質変換されるような変換を実現するには、合成音Ｂの分析フレームごとの周波数特性が目標Ａの周波数特性とできるだけ等しくなるように、合成音Ｂの周波数領域のデータを変換関数で変換すればよい。これを分析フレームごとに考えると、合成音Ｂの１つの分析フレームＢｎ（ｎ番目の分析フレームＢｎとする）の周波数特性が、この分析フレームに時間軸上で対応付けられた目標Ａの分析フレームＡｍ（ｍ番目の分析フレームＡｍとする）の周波数特性に変換されるような変換関数を同定すればよいということになる。そこで、分析フレームをフーリエ変換して波形の周波数領域の形状（スペクトル包絡）を求め、周波数軸上でＤＰマッチング等による伸縮を行い、合成音Ｂから目標Ａへの変換関数を求める。すなわち、合成音Ｂの分析フレームＢｎのスペクトル形状を、目標Ａの分析フレームＡｍのスペクトル形状に合致させるような変換を行う。

【００７０】

さらに、スペクトル形状変換について説明する。図２８は、周波数軸上でのＤＰマッチングを模式的に表した図である。目標話者音声（目標Ａ）と合成音Ｂのスペクトル包絡をそれぞれＳＡ、ＳＢとする。周波数軸上でスペクトル包絡ＳＡとスペクトル包絡ＳＢとの

対応関係をＤＰマッチングによって求める。対応関係を表すパスＰ０、・・Ｐｉ、・・Ｐｎが変換関数に相当し、合成音Ｂのスペクトルをパス Ｐ０、・・Ｐｉ、・・Ｐｎによって非線型に写像することで目標話者音声（目標Ａ）のスペクトルが得られることとなる。なお、予めスペクトル包絡ＳＡ、ＳＢに対し、高域強調あるいは低域強調等の前処理を行い、前処理がなされたスペクトル包絡に対してＤＰマッチングを行うようにしてもよい。

【００７１】

以上で説明したように、変換関数同定においては、ＬＰＣ分析部２６で合成音Ｂのｎ番目の分析フレームＢｎをＬＰＣ分析した結果から、スペクトル包絡ＳＢｎを導出する。同様に、Ｂｎに時間軸上で対応付けられた目標Ａの分析フレームＡｍのスペクトル包絡ＳＡｍを同定する。その上で、ＳＢｎがＳＡｍに変換されるような変換関数を作成する。このようにして、合成音Ｂのｎ番目の分析フレームＢｎから、時間軸上で対応付けられた目標Ａの分析フレームＡｍへの変換関数が求まるので、これを当該フレーム全てに適用し、合成音Ａ全体に対する変換関数を求めることができる。

【００７２】

なお、以上の説明では、目標Ａと合成音Ｂといった１文だけを学習データとして用いて変換関数をフレームごとに求める例を示したが、学習データをさらに増やした場合、目標Ａと合成音ＢのＬＰＣ係数を特徴量空間内で、例えば音素ごとにクラスタリングし、クラスタの代表点ごとに変換関数を同定することも可能である。また、ＧＭＭ（Guassian Mixture model）等の確率密度分布で特徴量空間を表現して、密度分布状態で対応付ける方法も考えられる。

【００７３】

図２９は、特徴量空間内でクラスタリングされた音素間の変換関数を模式的に示す図である。合成音Ｂの特徴パラメータ空間内と目標Ａの特徴パラメータ空間内とにおいて、例えば音素（ａ、ｉ、ｕ、ｅ、ｏ、ｓ）がそれぞれクラスタに分類され、合成音Ｂの特徴量空間内のクラスタ中の代表点が変換関数によって目標Ａの特徴量空間内のクラスタ中の代表点に変換される。

【００７４】

次に、変換対象である音声合成用素片データと変換関数との対応付けを行う。そのために、ＬＰＣ分析部３０は、あらかじめ音声合成用の素片信号をＬＰＣ分析して、素片データの分析フレームごとのＬＰＣ係数を求めておく。

【００７５】

対応フレーム探索部３１は、ＬＰＣ分析部３０が出力する素片データの分析フレーム毎のＬＰＣ係数に対して合成音Ｂのフレーム毎のＬＰＣ係数３５ｂ中の特徴量空間内距離が最も近いものを対応フレームとして求める。

【００７６】

変換関数導出部２９は、スペクトル形状変換部２８によって求めた合成音Ｂのフレームごとの変換関数を素片データの変換関数とする。この時、学習データを増やした場合は、図２９で説明した変換関数の同定時と同様に、フレームごとではなくクラスタごとに変換関数を設定するといったことも可能である。

【００７７】

スペクトル変換部３２は、変換関数導出部２９において求めた素片に対する変換関数の中から、先の対応フレームに対する変換関数を選択し、選択された変換関数を用いて素片信号（素片データ）のスペクトルを変換する。これにより、目標Ａの声質を持つ変換後素片信号（合成音を作成できる素片データ）を得ることができる。

【００７８】

さらに、この素片データを、音声合成用データ記憶部１３に保存されている変換前の音声合成用データベース内の素片データと差し替えることで、任意のテキストに対して目標Ａの声質を持つ合成音を作成できるデータベースが完成する。

【００７９】

なお、第１の実施例では、合成音を作成するための発音記号列が予め分かっているもの

としたが、目標話者音声の発声内容を漢字かな混じり文等のテキストデータで表して形態素解析等を用いて分析し、その結果によって発音記号列を生成することも可能である。

【００８０】

また、被変換対象である入力信号は、上記の素片データ以外にも、上記以外の素片データを使うこともできる。例えば、音声合成用データ記憶部１３に保存されている素片データが旅行用の会話のデータであり、入力信号となる素片データが一般の会話用のデータである場合などがある。

【００８１】

さらに、上記の素片データで作成された合成音、上記以外の素片データで作成された合成音、ユーザの発声による音声等を被変換対象である入力信号とすることもできる。

【００８２】

さらに、第１の実施例の変形として、変換関数生成部内の合成音と目標話者音声の時間軸対応付け処理を行わないで変換関数の同定を行うという方法も考えられる。この方法を実現できる理由を以下に示す。本発明では、合成音と目標話者音声の発声内容を同一あるいは類似のものにして処理を行っているため、合成音および目標話者音声の音素情報は同一の部分が多く、特徴量空間内の特徴パラメータの分布が互いに類似している。このため、例えば、音素ごとにクラスタリングして変換関数を求める場合、合成音と目標話者音声の時間軸対応を取らなくても、変換関数を生成することができる。この方法を用いれば、時間軸対応付け処理を行わないので、処理速度の大幅な向上を図ることが可能である。

【実施例２】

【００８３】

図３０は、本発明の第２の実施例に係る声質変換装置のブロック図である。声質変換装置は、目標話者音声入力部１１、発音記号列入力部１２、音声合成用データ記憶部１３、音声合成部１４ａ、音声認識部１９、合成音セグメンテーション情報記憶部２０、目標話者音声セグメンテーション情報記憶部２１、ＬＰＣ分析部２５、２６、３０、ＤＰマッチング部２７ａ、スペクトル形状変換部２８、変換関数導出部２９、対応フレーム探索部３１、スペクトル変換部３２を備える。なお、図３０において、図２５と同一の符号は、同一物あるいは相当物を示し、特に記載無き場合、その説明を省略する。

【００８４】

音声認識部１９は、入力された目標話者音声信号に対し音声認識を行うことによって、目標話者音声信号と同一あるいは類似の発声内容を持つ合成音を生成するための発音記号列を生成し、発音記号列入力部１２に出力する。今、発声内容Ｃを持つ音声が目標話者音声であるとすると、これを音声認識することによって、発声内容Ｃを持つ発音記号列を自動的に生成することができる。このため、発音記号列が予め明らかでない場合でも、目標話者音声と同一あるいは類似の発生内容を持つ合成音を生成することができ、目標話者音声のみを入力してしまえば、第１の実施例と同様の処理を自動的に行うことができる。この時の音声認識の方法としては、例えば非特許文献２にあるようなＨＭＭ（hidden Markov model）による音声認識法等がある。

【００８５】

合成音セグメンテーション情報記憶部２０は、音声合成部１４ａにおいて音声を合成する際に音声合成用データ内から算出されて用いられるセグメンテーション情報を、変換関数の生成時に活用するためにＤＰマッチング部２７ａに出力する。今、発声内容Ｃを持つ音声が目標話者音声であるとすると、これを音声認識部１９において音声認識することによって、発声内容Ｃを持つ発音記号列を自動的に生成し発音記号列入力部１２に出力される。発音記号列が音声合成部１４ａに入力されたとすると、音声合成部１４ａは、この発音記号列にしたがって音声合成用データ記憶部１３から発声内容Ｃに対応したセグメンテーション情報を算出する。合成音セグメンテーション情報記憶部２０は、このセグメンテーション情報を記憶しておき、セグメンテーション情報２３ｉとしてＤＰマッチング部２７ａに出力し、変換関数生成の際に利用する。例えば、目標話者音声を分析して抽出された特徴パラメータの時間変化に対して、セグメンテーション情報と合成音を分析して抽出

された特徴パラメータの時間変化を対応付けることで、目標話者音声と合成音の時間軸対応付けを簡素化、高精度化することができる。セグメンテーション情報としては、図３１に示すような、フレーム番号と音素が対応付けられた表（例えば、フレーム番号１、２、・・１０、１１、・・、５０、５１、・・に対しそれぞれ音素「ｉ」、「ｉ」、・・「ｉ」、「ｅ」、・・「ｕ」、「ｓ」、・・が対応している）を用いる。その他の利用法としては、目標話者音声と合成音の組が複数文であった場合、音素セグメンテーション情報を用いてクラスタリングするといった方法も考えられる。

【００８６】

目標話者音声セグメンテーション情報記憶部２１は、音声認識部１９が目標話者音声信号を音声認識する際に、出力される目標話者音声のセグメンテーション情報を入力する。今、発声内容Ｃを持つ目標話者音声を目標話者音声入力部１１に入力したとする。音声認識部１９は、この目標話者音声を音声認識して、発声内容Ｃを記述する発音記号列を生成して発音記号列入力部１２に出力する一方で、音声認識の結果であるセグメンテーション情報を目標話者音声セグメンテーション情報記憶部２１に出力し、目標話者音声セグメンテーション情報記憶部２１は、セグメンテーション情報を記憶しておく。このセグメンテーション情報２３ｈを、ＤＰマッチング部２７ａに出力して合成音と目標話者音声との対応付け処理に利用する。利用法としては、合成音を分析して抽出された特徴パラメータの時間変化に対して、セグメンテーション情報と目標話者音声を分析して抽出された特徴パラメータの時間変化を対応付けることで、目標話者音声と合成音の時間軸対応付けを簡素化、高精度化することができる。なお、セグメンテーション情報としては、図３１で示した表と同様のものを用いる。

【００８７】

ＤＰマッチング部２７ａは、実施例１で説明したＤＰマッチング部２７と同様に、目標ＡのＬＰＣ係数３５ａと合成音ＢのＬＰＣ係数３５ｂとを用いて、目標Ａと合成音Ｂの時間軸を合わせるために、ＤＰマッチングによる時間軸伸縮を行う。さらに、ＤＰマッチング部２７ａは、合成音セグメンテーション情報記憶部２０が出力するセグメンテーション情報２３ｈと、目標話者音声セグメンテーション情報記憶部２１が出力するセグメンテーション情報２３ｉとを用いてＤＰマッチングによる時間軸対応付けの簡素化、高精度化を図っている。

【００８８】

次にセグメンテーション情報を用いたＤＰマッチングによる時間軸対応付けの一例について説明する。図３２は、セグメンテーション情報を用いたＤＰマッチングによる時間軸対応付けの第１の例を示す図である。縦軸方向に分析された目標Ａ（音素「ａ」「ｓ」「ｕ」）、横軸方向に分析された合成音Ｂ（音素「ａ」「ｓ」「ｕ」）が配置されている。セグメンテーション情報２３ｈとして、目標Ａの音素「ａ」と「ｓ」の間に音素境界Ｐ１が、目標Ａの音素「ｓ」と「ｕ」の間に音素境界Ｑ１が音声合成部１４ａにおいて設定されている。また、セグメンテーション情報２３ｉとして、合成音Ｂの音素「ａ」と「ｓ」の間に音素境界Ｐ２が、目標Ａの音素「ｓ」と「ｕ」の間に音素境界Ｑ２が音声認識部１９において設定されている。この時、音素境界Ｐ１と音素境界Ｐ２との交点を拘束点Ｐとし、音素境界Ｑ１と音素境界Ｑ２との交点を拘束点Ｑとする。ＤＰマッチングを音素「ａ」、「ｓ」、「ｕ」に対して順次行い、ＤＰパスを求めていく。この時、ＤＰパスが拘束点Ｐ、拘束点Ｑを通るように制約を課してＤＰパスを求めるようにする。すなわち、拘束点を通るようにＤＰパスを決定することにより、制約条件付のＤＰマッチングを行い、音素境界同士が対応付けられるようにする。なお、拘束点は、必ずしも通らなくてもよく、拘束点の近傍を通るような緩やかな制約を課してもよい。

【００８９】

セグメンテーション情報は、ある発声内容に対して、各音素の開始時刻と終了時刻、および音素のラベル等が記述された情報である。セグメンテーション情報によって、音素のラベルとその音素の開始時刻、終了時刻が示されるために、目標Ａあるいは合成音Ｂにおける音素の境界が明確に示されることとなる。したがって、目標Ａと合成音Ｂとの各フレ

ーム間の時間軸上の対応付け（ＤＰパス）を求める際に、セグメンテーション情報を用いて制約を付けることで、ＤＰマッチングによる対応付けが音素境界付近であいまいになった場合であっても、精度の高い対応付けを実現することができることとなる。

【００９０】

次にセグメンテーション情報を用いたＤＰマッチングによる時間軸対応付けの他の例について説明する。図３３は、セグメンテーション情報を用いたＤＰマッチングによる時間軸対応付けの第２の例を示す図である。目標Ａと合成音Ｂは、図３２で説明したと同様に配置されている。ただし、図３３において、図３２と異なる点は、セグメンテーション情報２３ｈが入力されていない。すなわち、目標話者音声セグメンテーション情報記憶部２１が存在せず、目標Ａに音素境界がないことである。この場合、通常のＤＰマッチングによってＤＰパスを求めた上で、ＤＰパスが合成音Ｂの音素境界Ｐ１を通る点に対応する目標Ａの推定音素境界Ｐ３を求める。また、ＤＰパスが合成音Ｂの音素境界Ｑ１を通る点に対応する目標Ａの推定音素境界Ｑ３を求める。すなわち、制約条件なしでＤＰマッチングを行った後に、セグメンテーション情報がある音声の音素境界と対応付いた箇所を、セグメンテーション情報がない音声の音素境界として推定することができる。通常、ＤＰマッチングを行っただけでは、音素境界を判定することができないため、音素境界が推定できるこの方法は、変換関数を選択する際などにより有効な手段となる。

【００９１】

以上の説明では、セグメンテーション情報をＤＰマッチングに適用する例について説明したが、他にセグメンテーション情報を、スペクトル変換部３２において被変換入力信号（素片信号）を変換する際に適用することも可能である。すなわち、被変換入力信号の各フレームがどの変換関数で変換されるかを判定する際に、そのフレームのセグメンテーション情報に付するラベル情報を用いれば、どの集合（例えば、音素毎のクラスタ等）に属するかを容易に判別することができる。この様子を図３４に示す。図３４において、ラベル情報（「ｉ」「ｉ」「ｅ」「ｕ」）は、特徴パラメータ空間中の各クラスタに対応付けがなされ、ラベル情報から直接的に特徴量空間内のクラスタ中の代表点に変換することができる。

【実施例３】

【００９２】

次に声質変換方法を繰り返し行い精度を高めて行く例について説明する。図３５は、本発明の第３の実施例に係る声質変換方法を表すフローチャート図である。図３５において、ステップＳ３１～Ｓ３９は、それぞれ図１４のステップＳ２１、Ｓ２２１、Ｓ２２２、Ｓ２２３、Ｓ２３、Ｓ２４、Ｓ２５、Ｓ２６、Ｓ２７と同等の処理を行うステップであり、その説明を省略する。ステップＳ４０において、変換後の信号を音声合成用のデータとして音声合成用データ記憶部１３に登録する。

【００９３】

ステップＳ４０において、繰り返しが所定の収束条件に達したか否かを判定する。達していなければ、ステップＳ３４に戻り、変換後の音声合成用データを用いて合成音を生成する。達していればステップＳ４２で一連の処理が終了する。収束条件としては、例えば、目標話者音声と合成音とのＬＰＣ係数空間内距離が一定値以下になった場合、もしくはこの一定値以下の状態が一定回数継続した場合がある。また、スペクトルやパワー等の前回との差分値の合計値が一定値以下になった場合、もしくはこの一定値以下の状態が一定回数継続した場合を収束条件としてもよい。さらに繰り返し回数が一定回数繰り返した場合等を収束条件としてもよい。

【００９４】

第３の実施例では、第１の実施例のように、素片を被変換入力信号として音声合成用データの声質を変換する場合には、変換後の音声合成用データを用いて再度目標話者音声と同一あるいは類似の発声内容の合成音を生成し、実施例１と同様の方法で変換関数の生成を行い、素片を変換するという処理を複数回繰り返す。さらに、第２の実施例を組み合わせて用いることで、最初に目標話者音声を入力してしまえば、繰り返しの処理も全て自動

で行うことができ、変換精度（声質の類似度）を高めていくことが可能となる。なお、図３５では、セグメンテーション情報を用いない例を示したが、勿論セグメンテーション情報を用いてもよい。

【００９５】

なお、以上説明した実施例１～３において、音声合成用データ記憶部１３に複数人数分の音声データからなる音声合成用データを記憶しておき、入力された目標話者の声質によって使用する音声合成用データを選択することができるようにすることも可能である。例えば、目標話者音声が男声であった場合は男声の素片データを用い、女声であった場合は女声の素片データを用いるといった方法が考えられる。この方法を用いれば、極端な変換を避けて変換処理による音質の劣化を少なくすることが可能となる。

【実施例４】

【００９６】

次に、本発明の声質変換装置、声質変換方法あるいは声質変換プログラムによって生成された合成音データを用いて合成音を生成する装置について説明する。図３６は、本発明の第４の実施例に係る合成音生成装置を表すブロック図である。図３６において、記号列入力部４１、音声合成出力部４２、データ記憶部４３は、それぞれ図１の発音記号列入力部１２、音声合成部１４、音声合成用データ記憶部１３と同等の機能を有するものである。ただし、音声合成出力部４２は、生成された合成音を出力する機能を持つ。また、データ記憶部４３は、先に説明した声質変換装置、声質変換方法あるいは声質変換プログラムによって生成される変換後出力信号を合成音データとして蓄えるものである。音声合成出力部４２は、記号列入力部４１に入力される記号列に基づいてデータ記憶部４３から読み出した合成音データを用いて合成音を生成して出力する。出力される合成音は、変換後出力信号、すなわち目標話者の声質に変換されたデータに基づいて生成されるので、目標話者の声質を備えた合成音となる。

【実施例５】

【００９７】

次に、本発明の声質変換装置、声質変換方法あるいは声質変換プログラムによって生成された合成音データを用いて合成音を生成する他の装置について説明する。図３７は、本発明の第５の実施例に係る合成音生成装置を表すブロック図である。図３７において、音声入力部５１は、ユーザの発する音声信号を入力する。変換データ記憶部５３は、先に説明した声質変換装置、声質変換方法あるいは声質変換プログラムによって生成される、分析フレーム毎のＬＰＣ係数と対応する変換関数とを記憶している。音声変換出力部５２は、音声入力部５１から出力される音声信号をフレーム毎に、例えばＬＰＣ分析し、分析されたフレームにパラメータ空間距離の最も近いフレームを変換データ記憶部５３内において探索して対応する変換関数を読出し、この変換関数によって音声信号のスペクトルを変換して出力する。出力される合成音は、目標話者の声質に変換されたデータに基づいて生成されるので、目標話者の声質を備えた合成音となる。

【産業上の利用可能性】

【００９８】

本発明によれば、電話機やトランシーバー等の通信機器で、自由に声質を変えるといった用途に適用できる。また、パーソナルコンピュータや携帯電話等で、電子メールやチャットのテキストを読み上げる際の合成音の声質をユーザの望む声質にするといった用途にも適用できる。さらに、アニメーションの音声の録音や外国映画の吹き替え等をテキスト音声合成で行う場合に、登場人物に合ったキャラクタ音声の声質を生成するといった用途にも適用できる。

【図面の簡単な説明】

【００９９】

【図１】本発明の第１の実施形態に係る声質変換装置の構成を示すブロック図である。

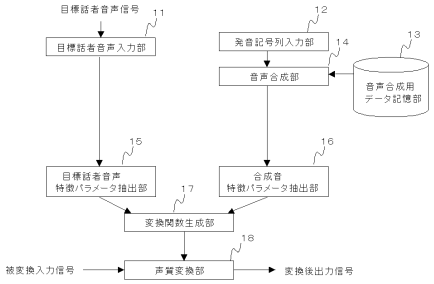
【図２】本発明の第２の実施形態に係る声質変換装置の構成を示すブロック図である。

【図３】本発明の第３の実施形態に係る声質変換装置の構成を示すブロック図である。

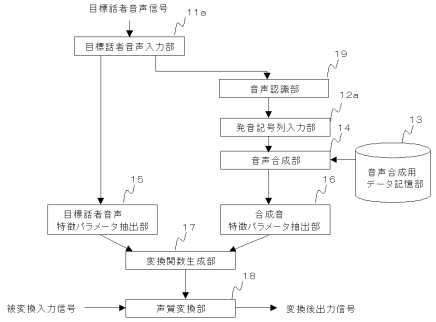
【図４】本発明の第４の実施形態に係る声質変換装置の構成を示すブロック図である。
【図５】本発明の第５の実施形態に係る声質変換装置の構成を示すブロック図である。
【図６】本発明の第６の実施形態に係る声質変換装置の構成を示すブロック図である。
【図７】本発明の第７の実施形態に係る声質変換装置の構成を示すブロック図である。
【図８】本発明の第８の実施形態に係る声質変換装置の構成を示すブロック図である。
【図９】本発明の第９の実施形態に係る声質変換装置の構成を示すブロック図である。
【図１０】本発明の第１０の実施形態に係る声質変換装置の構成を示すブロック図である。
【図１１】本発明の第１１の実施形態に係る声質変換装置の構成を示すブロック図である。
【図１２】本発明の第１２の実施形態に係る声質変換装置の構成を示すブロック図である。
【図１３】本発明の第１の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図１４】本発明の第２の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図１５】本発明の第３の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図１６】本発明の第４の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図１７】本発明の第５の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図１８】本発明の第６の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図１９】本発明の第７の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図２０】本発明の第８の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図２１】本発明の第９の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図２２】本発明の第１０の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図２３】本発明の第１１の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図２４】本発明の第１２の実施形態に係る声質変換装置の動作を示すフローチャートである。
【図２５】本発明の第１の実施例に係る声質変換装置の構成を示すブロック図である。
【図２６】声質変換の主要部における音声データの処理の流れを説明する図である。
【図２７】分析フレームごとのＤＰマッチングを模式的に表した図である。
【図２８】周波数軸上でのＤＰマッチングを模式的に表した図である。
【図２９】特徴量空間内でクラスタリングされた音素間の変換関数を模式的に示す図である。
【図３０】本発明の第２の実施例に係る声質変換装置のブロック図である。
【図３１】フレーム番号と音素との対応付けを示す図である。
【図３２】セグメンテーション情報を用いたＤＰマッチングによる時間軸対応付けの第１の例を示す図である。
【図３３】セグメンテーション情報を用いたＤＰマッチングによる時間軸対応付けの第２の例を示す図である。
【図３４】ラベル情報とクラスタとの対応付けを示す図である。
【図３５】本発明の第３の実施例に係る声質変換方法を表すフローチャート図である。

【図３６】本発明の第４の実施例に係る合成音生成装置を表すブロック図である。
【図３７】本発明の第５の実施例に係る合成音生成装置を表すブロック図である。
【図３８】第１の従来例の声質変換方法を表すブロック図である。
【図３９】第２の従来例の声質変換方法を表すブロック図である。
【符号の説明】
【０１００】
１１、１１ａ　目標話者音声入力部
１２、１２ａ　発音記号列入力部
１３　音声合成用データ記憶部
１４、１４ａ、１４ｂ、１４ｃ　音声合成部
１５　目標話者音声特徴パラメータ抽出部
１６　合成音特徴パラメータ抽出部
１７、１７ａ、１７ｂ、１７ｃ　変換関数生成部
１８　声質変換部
１９　音声認識部
２０　合成音セグメンテーション情報記憶部
２１、２１ａ、２１ｂ　目標話者音声セグメンテーション情報記憶部
２２　セグメンテーション情報入力部
２３ａ、２３ｂ、２３ｃ、２３ｄ、２３ｅ、２３ｆ、２３ｇ、２３ｈ、２３ｉ　セグメンテーション情報
２５、２６、３０　ＬＰＣ分析部
２７、２７ａ　ＤＰマッチング部
２８　スペクトル形状変換部
２９　変換関数導出部
３１　対応フレーム探索部
３２　スペクトル変換部
３５ａ、３５ｂ　ＬＰＣ係数
４１　記号列入力部
４２　音声合成出力部
４３　データ記憶部
５１　音声入力部
５２　音声変換出力部
５３　変換データ記憶部

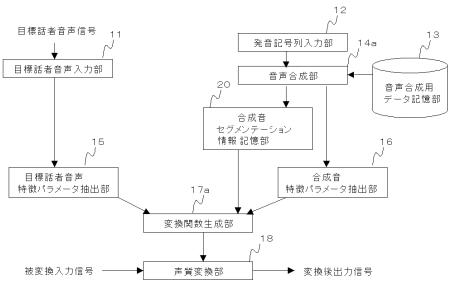
【図１】


目標話者音声信号
11
目標話者音声入力部
12
発音記号列入力部
14
音声合成部
13
音声合成用
データ記憶部
15
目標話者音声
特徴パラメータ抽出部
16
合成音
特徴パラメータ抽出部
17
変換関数生成部
18
被変換入力信号
声質変換部
変換後出力信号

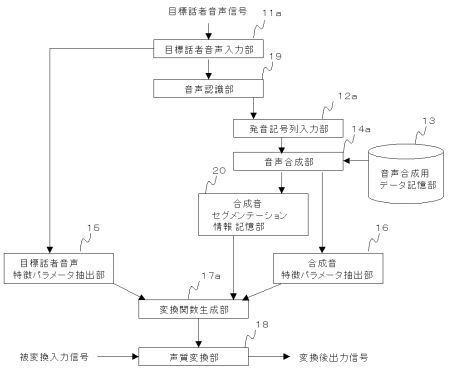
【図２】


目標話者音声信号
11a
目標話者音声入力部
19
音声認識部
12a
発音記号列入力部
14
音声合成部
13
音声合成用
データ記憶部
15
目標話者音声
特徴パラメータ抽出部
16
合成音
特徴パラメータ抽出部
17
変換関数生成部
18
被変換入力信号
声質変換部
変換後出力信号

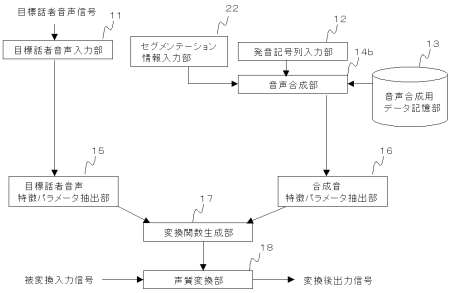
【図３】


目標話者音声信号
11
目標話者音声入力部
12
発音記号列入力部
14a
音声合成部
13
音声合成用
データ記憶部
20
合成音
セグメンテーション
情報 記憶部
15
目標話者音声
特徴パラメータ抽出部
16
合成音
特徴パラメータ抽出部
17a
変換関数生成部
18
被変換入力信号
声質変換部
変換後出力信号

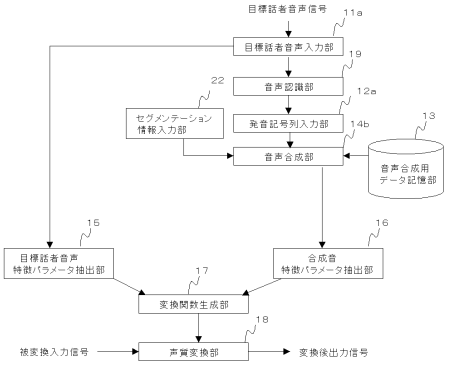
【図４】


目標話者音声信号
11a
目標話者音声入力部
19
音声認識部
12a
発音記号列入力部
14a
音声合成部
13
音声合成用
データ記憶部
20
合成音
セグメンテーション
情報 記憶部
15
目標話者音声
特徴パラメータ抽出部
16
合成音
特徴パラメータ抽出部
17a
変換関数生成部
18
被変換入力信号
声質変換部
変換後出力信号

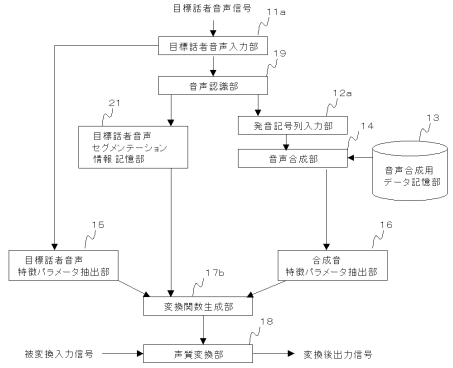
【図５】


目標話者音声信号
11
目標話者音声入力部
22
セグメンテーション情報入力部
12
発音記号列入力部
14b
音声合成部
13
音声合成用データ記憶部
15
目標話者音声特徴パラメータ抽出部
16
合成音特徴パラメータ抽出部
17
変換関数生成部
18
被変換入力信号
声質変換部
変換後出力信号

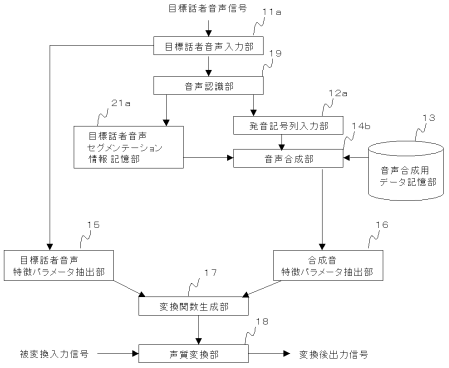
【図６】


目標話者音声信号
11a
目標話者音声入力部
19
音声認識部
22
セグメンテーション情報入力部
12a
発音記号列入力部
14b
音声合成部
13
音声合成用データ記憶部
15
目標話者音声特徴パラメータ抽出部
16
合成音特徴パラメータ抽出部
17
変換関数生成部
18
被変換入力信号
声質変換部
変換後出力信号

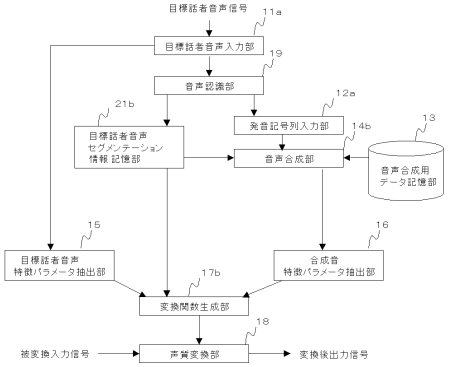
【図７】


目標話者音声信号
11a
目標話者音声入力部
19
音声認識部
21
目標話者音声セグメンテーション情報記憶部
12a
発音記号列入力部
14
音声合成部
13
音声合成用データ記憶部
15
目標話者音声特徴パラメータ抽出部
16
合成音特徴パラメータ抽出部
17b
変換関数生成部
18
被変換入力信号
声質変換部
変換後出力信号

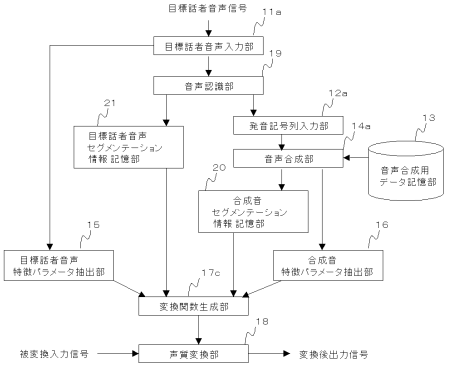
【図８】


目標話者音声信号
11a
目標話者音声入力部
19
音声認識部
21a
目標話者音声セグメンテーション情報記憶部
12a
発音記号列入力部
14b
音声合成部
13
音声合成用データ記憶部
15
目標話者音声特徴パラメータ抽出部
16
合成音特徴パラメータ抽出部
17
変換関数生成部
18
被変換入力信号
声質変換部
変換後出力信号

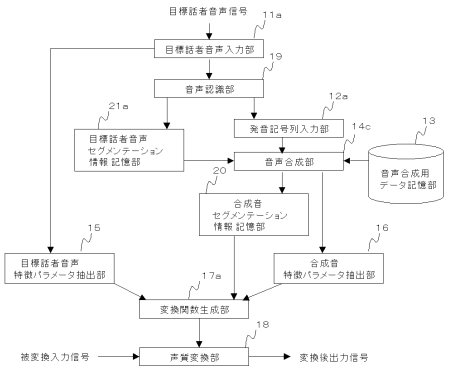
【図９】



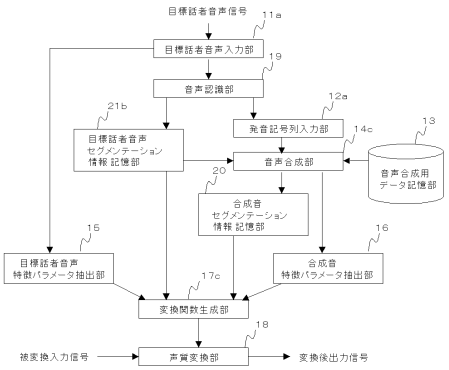
【図１０】



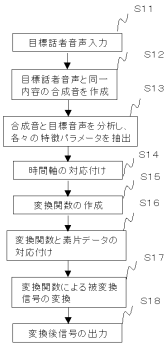
【図１１】



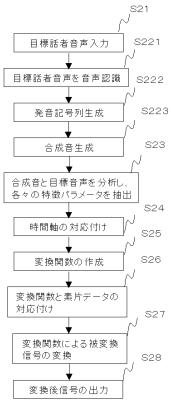
【図１２】



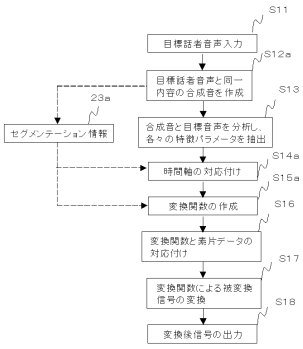
【図１３】



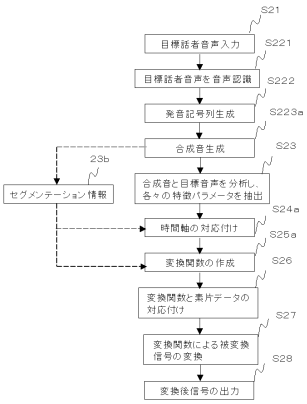
【図１４】



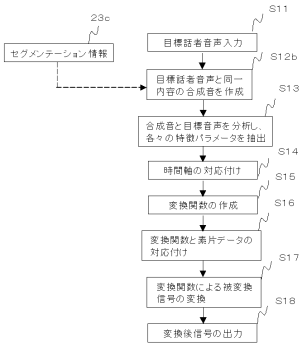
【図１５】



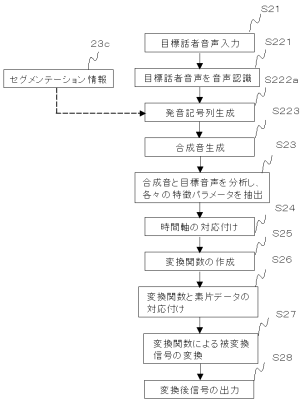
【図１６】



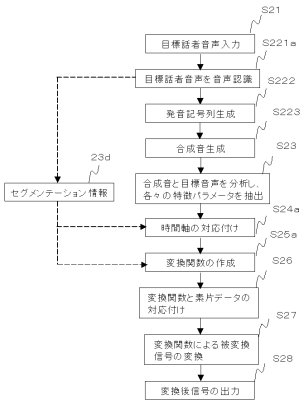
【図１７】



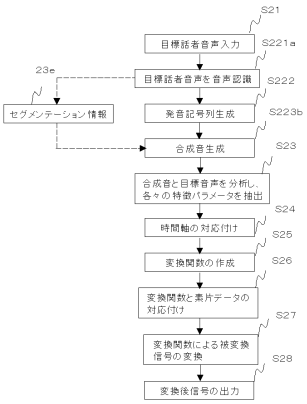
【図１８】



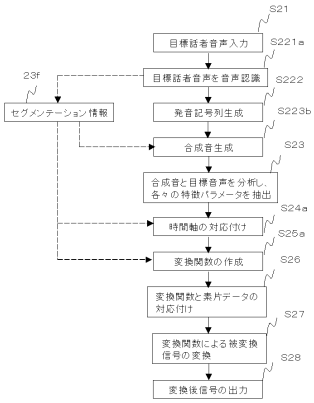
【図１９】



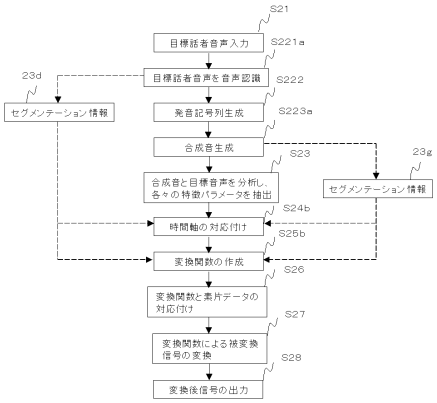
【図２０】



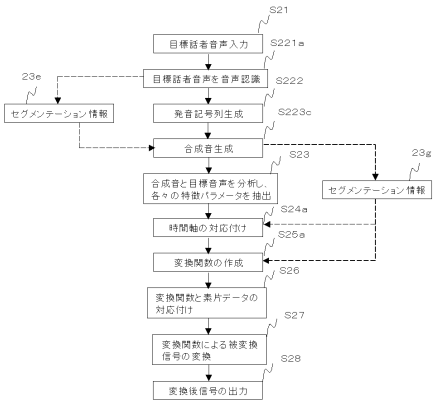
【図２１】



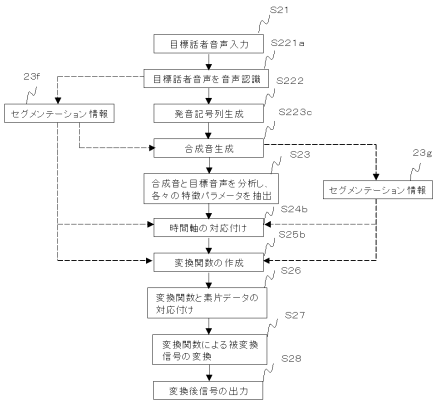
【図２２】



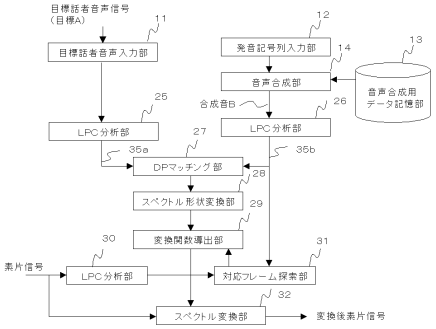
【図２３】



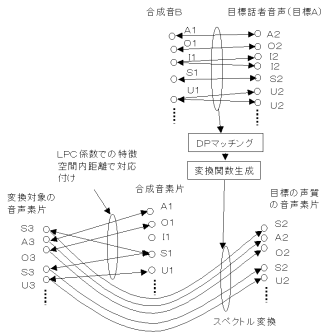
【図２４】



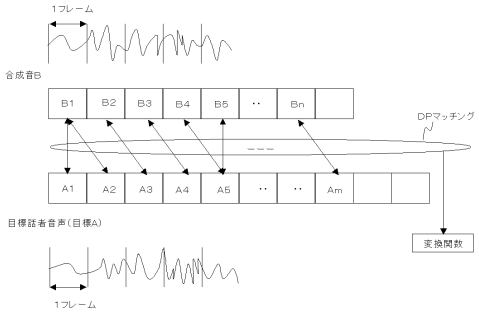
【図２５】



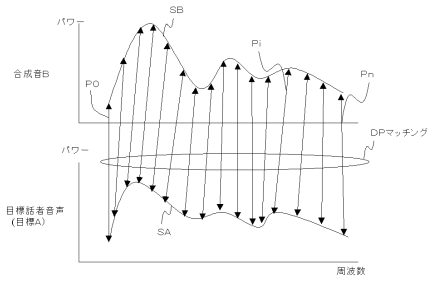
【図２６】



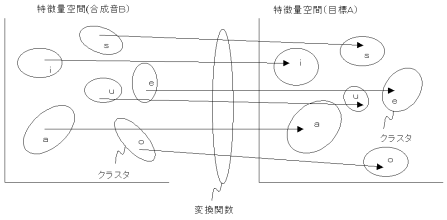
【図２７】



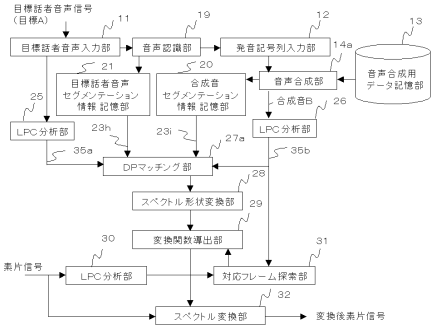
【図２８】



【図２９】



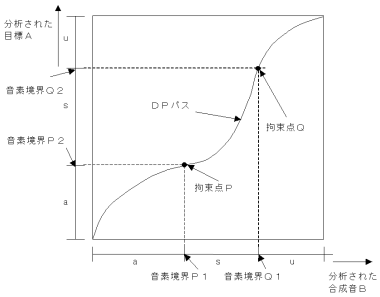
【図３０】



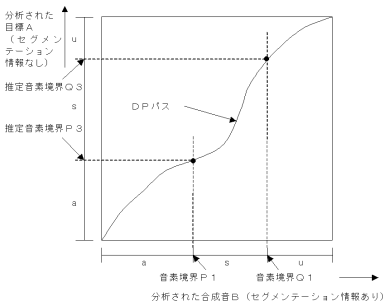
【図３１】

フレーム番号	1	2	..	..	10	11	..	50	51	..
音素	i	i	..	..	j	e	..	u	s	..

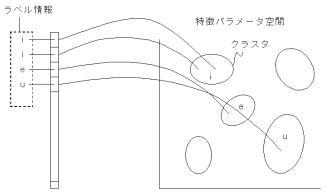
【図３２】



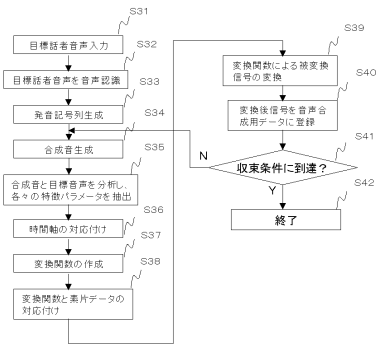
【図３３】



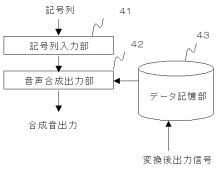
【図３４】



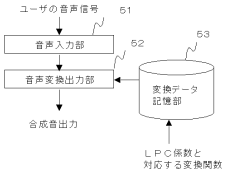
【図３５】



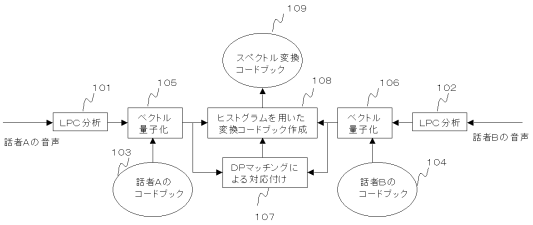
【図３６】



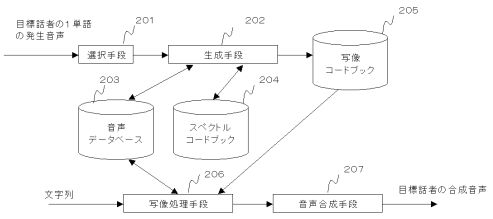
【図３７】


ユーザの音声信号
51
音声入力部
52
音声変換出力部
53
変換データ
記憶部
合成音出力
ＬＰＣ係数と
対応する変換関数

【図３８】


109
スペクトル変換
コードブック
101
105
108
106
102
LPC分析
ベクトル
量子化
ヒストグラムを用いた
変換コードブック作成
ベクトル
量子化
LPC分析
話者Aの音声
話者Bの音声
103
DPマッチングに
よる対応付け
104
話者Aの
コードブック
話者Bの
コードブック
107

【図３９】


目標話者の1単語
の発生音声
201
選択手段
202
生成手段
205
写像
コードブック
203
204
音声
データベース
スペクトル
コードブック
206
207
文字列
写像処理手段
音声合成手段
目標話者の合成音声

フロントページの続き

(56)参考文献 特開２００２－２１５１９９（ＪＰ，Ａ）
特開２０００－２９４８７（ＪＰ，Ａ）
特開平１０－３０１５９９（ＪＰ，Ａ）
特開平０９－１７９５７６（ＪＰ，Ａ）
特開平０９－２４４６９４（ＪＰ，Ａ）
特開平０９－３０５１９７（ＪＰ，Ａ）
特許第２８８０５０８（ＪＰ，Ｂ２）
特開２００１－０３４２８０（ＪＰ，Ａ）