(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2010-8922
(P2010-8922A)
(43)公開日　平成22年1月14日(2010.1.14)

(51)Int.Cl.		F I		テーマコード（参考）
G10L 21/04	(2006.01)	G10L 21/04	120D	
G10L 13/06	(2006.01)	G10L 13/06	240D	

審査請求 未請求　請求項の数 15　OL　（全 28 頁）

(21)出願番号　特願2008-170973 (P2008-170973)
(22)出願日　平成20年6月30日 (2008.6.30)

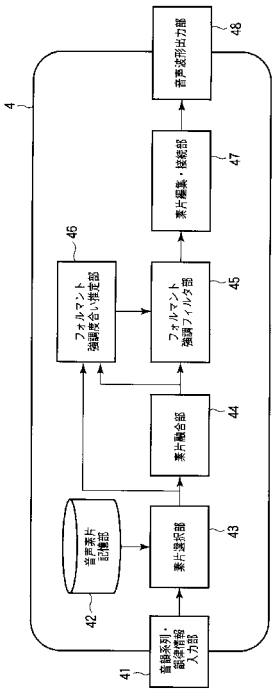
(71)出願人　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(74)代理人　100058479
弁理士　鈴江　武彦
(74)代理人　100108855
弁理士　蔵田　昌俊
(74)代理人　100091351
弁理士　河野　哲
(74)代理人　100088683
弁理士　中村　誠
(74)代理人　100109830
弁理士　福原　淑弘
(74)代理人　100075672
弁理士　峰　隆司

最終頁に続く

(54)【発明の名称】音声処理装置、音声処理方法及びプログラム

(57)【要約】
【課題】こもり感やブザー感が少なく、かつ人工的でない高音質な合成音声を生成できる音声処理装置を提供すること。
【解決手段】素片選択部４３は、目標音声に係る音韻系列のセグメントごとに複数個ずつの音声素片を選択する。素片融合部４４は、セグメントごとに該複数個の音声素片を融合して融合素片を生成する。フォルマント強調度合い推定部４６は、セグメントごとに該複数個の音声素片に関する特徴量と融合素片に関する特徴量との少なくとも一方を用いてフォルマント強調の度合いを推定する。フォルマント強調フィルタ部４５は、セグメントごとに融合素片に対して推定された強調度合い基づきフォルマント強調する。
【選択図】　図２

図２



【特許請求の範囲】
【請求項１】
目標音声に対応する音韻系列を合成単位で区切って得られる複数のセグメントを取得する第１の取得部と、
前記目標音声に対応する各々の前記セグメントの韻律情報を取得する第２の取得部と、
各々の前記セグメントごとに、当該セグメントに対し、当該セグメントの前記韻律情報に基づいて、予め用意された複数の音声素片のうちから、複数個の音声素片を選択する選択部と、
各々の前記セグメントごとに、当該セグメントに対して選択された複数個の前記音声素片を融合することによって、融合素片を生成する融合部と、
各々の前記セグメントごとに、前記選択部により選択された複数個の前記音声素片に関する特徴量と、前記融合部により生成された前記融合素片に関する特徴量との少なくとも一方を用いて、当該セグメントに係る前記融合素片に対して行うべきフォルマント強調における強調度合いを推定する推定部と、
各々の前記セグメントごとに、当該セグメントに係る前記融合素片に対して、前記推定部が推定した前記強調度合い基づくフォルマント強調を行うフォルマント強調フィルタ部とを備えたことを特徴とする音声処理装置。
【請求項２】
各々の前記セグメントについて前記フォルマント強調フィルタ部によりそれぞれ得られたフォルマント強調された前記融合素片に係る音声波形をもとにして、合成音声を生成する生成部を更に備えたことを特徴とする請求項１に記載の音声処理装置。
【請求項３】
各々の前記セグメントについて前記フォルマント強調フィルタ部によりそれぞれ得られたフォルマント強調された前記融合素片をそのまま出力する出力部を更に備えたことを特徴とする請求項１に記載の音声処理装置。
【請求項４】
前記出力部は、前記融合素片を、テキスト音声合成に供するための音声素片を記憶する記憶部に出力することを特徴とする請求項３に記載の音声処理装置。
【請求項５】
前記予め用意された複数の音声素片を記憶する音声素片記憶部を更に備えたことを特徴とする請求項１ないし４のいずれか１項に記載の音声処理装置。
【請求項６】
前記推定部は、各々の前記セグメントごとに、前記融合部により生成された前記融合素片のスペクトル包絡が、前記選択部により選択された前記音声素片のスペクトル包絡から、どの程度鈍ったかを推定し、推定されたスペクトル包絡の鈍り具合が大きいセグメントほど、強めのフォルマント強調度合いを推定することを特徴とする請求項１ないし５のいずれか１項に記載の音声処理装置。
【請求項７】
前記推定部は、各々の前記セグメントごとに、前記融合部により生成された前記融合素片のスペクトル包絡と、前記選択部により選択された前記音声素片のスペクトル包絡の形状との差を推定し、推定されたスペクトル包絡の形状の差が大きいセグメントほど、強めのフォルマント強調度合いを推定することを特徴とする請求項１ないし５のいずれか１項に記載の音声処理装置。
【請求項８】
前記推定部は、各々の前記セグメントごとに、目標音声と前記融合部により生成された前記融合素片による音声との差を、当該セグメントの目標音声に対応する韻律情報と前記選択部により選択された前記音声素片の韻律情報とから推定し、推定された当該目標音声と融合素片による音声との差が大きいセグメントほど、強めのフォルマント強調度合いを推定することを特徴とする請求項１ないし５のいずれか１項に記載の音声処理装置。
【請求項９】

前記推定部は、前記複数のセグメントのそれぞれに対して、フォルマントごと又は複数に分割した周波数帯域ごとにフォルマント強調度合いを推定し、
前記フォルマント強調フィルタ部は、それぞれのフォルマント又は周波数帯域に対して推定されたフォルマント強調度合いに従って、フォルマント又は周波数帯域間で異なる強さのフォルマント強調を行うことを特徴とする請求項１ないし５のいずれか１項に記載の音声処理装置。
【請求項１０】
第１の取得部、第２の取得部、選択部、融合部、推定部及びフォルマント強調フィルタ部を備えた音声処理装置の音声処理方法であって、
前記第１の取得部が、目標音声に対応する音韻系列を合成単位で区切って得られる複数のセグメントを取得するステップと、
前記第２の取得部が、前記目標音声に対応する各々の前記セグメントの韻律情報を取得するステップと、
前記選択部が、各々の前記セグメントごとに、当該セグメントに対し、当該セグメントの前記韻律情報に基づいて、予め用意された複数の音声素片のうちから、複数個の音声素片を選択するステップと、
前記融合部が、各々の前記セグメントごとに、当該セグメントに対して選択された複数個の前記音声素片を融合することによって、融合素片を生成するステップと、
前記推定部が、各々の前記セグメントごとに、前記選択部により選択された複数個の前記音声素片に関する特徴量と、前記融合部により生成された前記融合素片に関する特徴量との少なくとも一方を用いて、当該セグメントに係る前記融合素片に対して行うべきフォルマント強調における強調度合いを推定するステップと、
前記フォルマント強調フィルタ部が、各々の前記セグメントごとに、当該セグメントに係る前記融合素片に対して、前記推定部が推定した前記強調度合い基づくフォルマント強調を行うステップとを有することを特徴とする音声処理方法。
【請求項１１】
前記音声処理装置は、生成部を更に備えるものであり、
前記音声処理方法は、前記生成部が、各々の前記セグメントについて前記フォルマント強調フィルタ部によりそれぞれ得られたフォルマント強調された前記融合素片に係る音声波形をもとにして、合成音声を生成するステップを更に含むことを特徴とする請求項１０に記載の音声処理方法。
【請求項１２】
前記音声処理装置は、出力部を更に備えるものであり、
前記音声処理方法は、前記出力部が、各々の前記セグメントについて前記フォルマント強調フィルタ部によりそれぞれ得られたフォルマント強調された前記融合素片をそのまま出力するステップを更に含むことを特徴とする請求項１０に記載の音声処理方法。
【請求項１３】
第１の取得部、第２の取得部、選択部、融合部、推定部及びフォルマント強調フィルタ部を備えた音声処理装置としてコンピュータを機能させるためのプログラムであって、
前記第１の取得部が、目標音声に対応する音韻系列を合成単位で区切って得られる複数のセグメントを取得するステップと、
前記第２の取得部が、前記目標音声に対応する各々の前記セグメントの韻律情報を取得するステップと、
前記選択部が、各々の前記セグメントごとに、当該セグメントに対し、当該セグメントの前記韻律情報に基づいて、予め用意された複数の音声素片のうちから、複数個の音声素片を選択するステップと、
前記融合部が、各々の前記セグメントごとに、当該セグメントに対して選択された複数個の前記音声素片を融合することによって、融合素片を生成するステップと、
前記推定部が、各々の前記セグメントごとに、前記選択部により選択された複数個の前記音声素片に関する特徴量と、前記融合部により生成された前記融合素片に関する特徴量

との少なくとも一方を用いて、当該セグメントに係る前記融合素片に対して行うべきフォルマント強調における強調度合いを推定するステップと、
前記フォルマント強調フィルタ部が、各々の前記セグメントごとに、当該セグメントに係る前記融合素片に対して、前記推定部が推定した前記強調度合い基づくフォルマント強調を行うステップとをコンピュータに実行させるためのプログラム。
【請求項１４】
前記音声処理装置は、生成部を更に備えるものであり、
前記プログラムは、前記生成部が、各々の前記セグメントについて前記フォルマント強調フィルタ部によりそれぞれ得られたフォルマント強調された前記融合素片に係る音声波形をもとにして、合成音声を生成するステップを更にコンピュータに実行させることを特徴とする請求項１３に記載の音声処理装置。
【請求項１５】
前記音声処理装置は、出力部を更に備えるものであり、
前記プログラムは、前記出力部が、各々の前記セグメントについて前記フォルマント強調フィルタ部によりそれぞれ得られたフォルマント強調された前記融合素片をそのまま出力するステップを更にコンピュータに実行させることを特徴とする請求項１３に記載の音声処理装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音声処理装置、音声処理方法及びプログラムに関する。
【背景技術】
【０００２】
任意の文章から人工的に音声信号を作り出すことを、テキスト音声合成という。テキスト音声合成は、一般的に、言語処理部、韻律処理部及び音声合成部の３つ段階によって行われる。
【０００３】
入力されたテキストは、まず言語処理部において、形態素解析や構文解析が行われ、次に韻律処理部において、アクセントやイントネーションの処理が行われて、音韻系列・韻律情報（基本周波数、音韻継続時間長、パワーなど）が出力される。最後に、音声合成部において、音韻系列・韻律情報から音声信号を合成する。そこで、音声合成部に用いる音声合成方法は、韻律処理部で生成される任意の音韻系列を、任意の韻律で音声合成することが可能な方法でなければならない。
【０００４】
従来、このような音声合成方法として、入力の音韻系列を分割して得られる複数の合成単位（合成単位列）のそれぞれに対して、入力された音韻系列・韻律情報を目標にして、予め記憶された大量の音声素片の中から音声素片を選択し、選択した音声素片を合成単位間で接続することによって、音声を合成する、音声合成方法（素片選択型の音声合成方法）が知られている。例えば、特許文献１に開示された素片選択型の音声合成方法では、音声を合成することで生じる音声合成の劣化の度合いを、コストで表すこととし、予め定義されたコスト関数を用いて計算されるコストが小さくなるように、音声素片を選択する。例えば、音声素片を編集・接続することで生じる変形歪み及び接続歪みを、コストを用いて数値化し、このコストに基づいて、音声合成に使用する音声素片系列を選択し、選択した音声素片系列に基づいて、合成音声を生成する。
【０００５】
特許文献１に開示された音声合成方法のように、音声を合成することで生じる音声合成の劣化の度合いを考慮して、大量の音声素片の中から適切な音声素片系列を選択することによって、音声素片の編集及び接続による音質の劣化を抑えた合成音声を生成することができる。
【０００６】

しかしながら、特許文献１に開示された素片選択型の音声合成方法には、部分的に合成音の音質が劣化する問題点がある。この理由は次のようなものである。

【０００７】

第１の理由は、予め記憶された音声素片が非常に多い場合であっても、様々な音韻・韻律環境に対して適切な音声素片が存在するとは限らないことである。

【０００８】

第２の理由は、人が実際に感じる合成音声の劣化の度合いをコスト関数が完全に表現できないため、必ずしも最適な素片系列が選ばれない場合があるからである。

【０００９】

第３の理由は、音声素片が非常に多いために予め不良な音声素片を排除しておくことが困難であり、また不良な音声素片を取り除くためのコスト関数の設計も難しいため、選択された音声素片系列中に、突発的に不良な音声素片が混入する場合があるからである。

【００１０】

そこで、合成単位当たり１つずつの音声素片を選ぶのではなく、合成単位当たり複数個の音声素片を選択し、これを融合することによって新たな音声素片を生成し、こうして生成された音声素片を使って音声を合成する方法が開示されている（特許文献２参照）。以下、この方法を「複数素片選択融合型の音声合成方法」と呼ぶ。

【００１１】

特許文献２に開示された複数素片選択融合型の音声合成方法では、合成単位毎に複数の音声素片を融合することによって、目標とする音韻・韻律環境に合う適切な音声素片が存在しない場合や、最適な音声素片が選択されない場合、不良素片が選択されてしまった場合でも、高品質な音声素片を新たに生成することができ、さらに、この新たに生成した音声素片を使用して音声合成を行うことで、前述した素片選択型の音声合成方法の問題点を改善することができ、より安定性を増した高音質の音声合成を実現することができる。

【００１２】

この複数素片選択融合型の音声合成方法においては、音声素片の融合による平均化の副作用によってスペクトル包絡が原音に比べて若干鈍る傾向があり、その結果、こもり感やブザー感が生じる場合がある。こうしたこもり感やブザー感の主観的な改善には、音声符号化や音声合成でよく用いられるようなフォルマント強調フィルタを、融合された素片に対して適用することが効果的である。

【００１３】

フォルマント強調フィルタは、入力音声波形のスペクトル包絡のフォルマントによる山・谷を強調したような音声波形を出力するフィルタで、適度な度合いでフォルマントを強調できれば、スペクトル包絡が鈍ったことによって生じるこもり感やブザー感を改善できる。一般的に、フォルマント強調フィルタは入力波形のスペクトル特性に応じてフィルタ特性を変える点では適応的だが、どの程度フォルマントを強調するかについては、適切な強調度合いを決めるための客観尺度が存在しないため、主観評価などによって実験的に決めるしかなく、ハイパーパラメータなどの値を外部から指定することによって制御することが多い。

【００１４】

そのため、複数素片選択型の音声合成方法で用いる場合には、フォルマントの強調度合いは、合成音声の主観的な音質が総合的に良くなるように、主観評価などによって実験的に決める。すなわち、フォルマントの強調度合いは、融合されたあらゆる素片に対して共通のものが適用される。

【特許文献１】特開２００１－２８２２７８公報

【特許文献２】特開２００５－１６４７４９公報

【発明の開示】

【発明が解決しようとする課題】

【００１５】

しかしながら、音声素片の融合によるスペクトル包絡の鈍り具合は、通常、合成単位に

よって異なり、一様ではない。例えば、合成単位に対して選ばれた複数の素片が類似のスペクトル包絡を持つ場合は、融合してもさほどスペクトル包絡は鈍らないと考えられるが、フォルマントの位置が素片間で大きく異なるなど、選ばれた音声素片のスペクトル包絡がそれぞれ異なる特徴を持つ場合には、融合するとスペクトル包絡が鈍ってしまう可能性が高い。

【００１６】

このような状況において、全音声素片に対して同じ強調度合いのフォルマント強調フィルタを一様に適用すると、融合によってスペクトル包絡が大きく鈍った箇所にはフォルマント強調の程度が不十分であるのに対し、逆に融合によるスペクトル包絡の鈍りが小さい箇所はフォルマントが強調されすぎて人工的な音になる問題がある。

【００１７】

本発明は、上記事情を考慮してなされたもので、こもり感やブザー感が少なく、かつ人工的でない高音質な合成音声を生成できる音声処理装置、音声処理方法及びプログラムを提供することを目的とする。

【課題を解決するための手段】

【００１８】

本発明に係る音声処理装置は、目標音声に対応する音韻系列を合成単位で区切って得られる複数のセグメントを取得する第１の取得部と、前記目標音声に対応する各々の前記セグメントの韻律情報を取得する第２の取得部と、各々の前記セグメントごとに、当該セグメントに対し、当該セグメントの前記韻律情報に基づいて、予め用意された複数の音声素片のうちから、複数個の音声素片を選択する選択部と、各々の前記セグメントごとに、当該セグメントに対して選択された複数個の前記音声素片を融合することによって、融合素片を生成する融合部と、各々の前記セグメントごとに、前記選択部により選択された複数個の前記音声素片に関する特徴量と、前記融合部により生成された前記融合素片に関する特徴量との少なくとも一方を用いて、当該セグメントに係る前記融合素片に対して行うべきフォルマント強調における強調度合いを推定する推定部と、各々の前記セグメントごとに、当該セグメントに係る前記融合素片に対して、前記推定部が推定した前記強調度合い基づくフォルマント強調を行うフォルマント強調フィルタ部とを備えたことを特徴とする。

【発明の効果】

【００１９】

本発明によれば、こもり感やブザー感が少なく、かつ人工的でない高音質な合成音声を生成できる。

【発明を実施するための最良の形態】

【００２０】

以下、図面を参照しながら本発明の実施形態について説明する。

【００２１】

（第１の実施形態）

本発明の第１の実施形態に係るテキスト音声合成装置（音声処理装置）について説明する。

【００２２】

図１に、本実施形態に係るテキスト音声合成を行うテキスト音声合成装置（音声処理装置）の全体構成例を示す。

【００２３】

図１に示されるように、本実施形態のテキスト音声合成装置は、テキスト入力部１、言語処理部２、韻律処理部３、音声合成部４を備えている。

【００２４】

テキスト入力部１は、テキストを入力する。

【００２５】

言語処理部２は、テキスト入力部１から入力されるテキストの形態素解析・構文解析を

行い、これら言語解析により得られた言語解析結果を韻律処理部３へ出力する。
【００２６】
韻律制御部３は、該言語解析結果を入力し、該言語解析結果からアクセントやイントネーションの処理を行って、音韻系列及び韻律情報を生成し、生成した音韻系列及び韻律情報を音声合成部へ出力する。
【００２７】
音声合成部４は、該音韻系列及び韻律情報を入力し、該音韻系列及び韻律情報から音声波形を生成して出力する。
【００２８】
以下、音声合成部４を中心に、その構成及び動作について詳細に説明する。
【００２９】
図２に、本実施形態の音声合成部４の構成例を示す。
【００３０】
図２に示されるように、音声合成部４は、音韻系列・韻律情報入力部４１、音声素片記憶部４２、素片選択部４３、素片融合部４４、フォルマント強調フィルタ部４５、フォルマント強調度合い推定部４６、素片編集・接続部４７、音声波形出力部４８を備えている。
【００３１】
音韻系列・韻律情報入力部（以下、情報入力部と略記する。）４１は、音声合成部４への入力として、韻律制御部３から音韻系列・韻律情報を受理する。
【００３２】
音声素片記憶部（以下、素片記憶部と略記する。）４２は、大量の音声素片を蓄積している。また、素片記憶部４２は、それら蓄積されている音声素片の全てについて、それぞれ、当該音声素片に対する音韻・韻律環境を併せて蓄積している。
【００３３】
素片選択部４３は、素片記憶部４２に蓄積された音声素片の中から、複数の音声素片を選択する。
【００３４】
素片融合部４４は、素片選択部４３により選択された複数の音声素片を融合して、新たな音声素片（以下、「融合素片」とも呼ぶ。）を生成する。
【００３５】
フォルマント強調フィルタ部４５は、（次のフォルマント強調度合い推定部４６により推定された、強調の程度に応じて）素片融合部４４により生成された音声素片に対して、フォルマント強調を行う（すなわち、フォルマント強調された融合素片を生成する）。
【００３６】
フォルマント強調度合い推定部４６は、フォルマント強調フィルタ部４５においてフォルマントを強調する程度を推定する。
【００３７】
素片編集・接続部４７は、フォルマント強調フィルタ部４５から得られた音声素片を韻律変形及び接続して、合成音声の音声波形を生成する。
【００３８】
音声波形出力部４８は、素片編集・接続部４７で生成した音声波形を出力する。
【００３９】
なお、情報入力部４１～音声波形出力部４８の各部の機能は、コンピュータに格納されたプログラムに実現できる。
【００４０】
次に、図２の音声合成部４の各ブロックについて詳しく説明する。
【００４１】
＜情報入力部＞
まず、情報入力部４１は、韻律制御部３から入力された音韻系列・韻律情報を、素片選

択部４４へ出力する。音韻系列は、例えば、音韻記号の系列である。また、韻律情報は、例えば、基本周波数、音韻継続時間長、パワーなどである。

【００４２】

以下、情報入力部４１に入力される音韻系列、韻律情報を、それぞれ、入力音韻系列、入力韻律情報と呼ぶ。

【００４３】

＜素片記憶部＞

次に、素片記憶部４２には、合成音声を生成するときに用いられる音声の単位（以下、「合成単位」と称する。）で、音声素片が大量に蓄積されている。

【００４４】

ここで、「合成単位」とは、音素あるいは音素を分割したもの（例えば、半音素など）の組み合わせ、例えば、半音素、音素（Ｃ、Ｖ）、ダイフォン（ＣＶ、ＶＣ、ＶＶ）、トライフォン（ＣＶＣ、ＶＣＶ）、音節（ＣＶ、Ｖ）、などであり（ここで、Ｖは母音、Ｃは子音を表す。）、また、これらが混在しているなど可変長であってもよい。

【００４５】

また、「音声素片」は、合成単位に対応する音声信号の波形もしくはその特徴を表すパラメータ系列などを表すものとする。

【００４６】

図３に、素片記憶部４２に蓄積される音声素片の例を示す。図３に示すように、素片記憶部４２には、各音素の音声信号の波形である音声素片が、当該音声素片を識別するための素片番号とともに記憶されている。これらの音声素片は、別途収録された多数の音声データに対して音素毎にラベル付けし、ラベルにしたがって音素毎に音声波形を切り出したものである。

【００４７】

また、素片記憶部４２には、大量の音声素片とともに、各音声素片に対応した音韻・韻律環境が蓄積されている。

【００４８】

ここで、「音韻・韻律環境」とは、対応する音声素片にとって環境となる要因の組み合わせである。要因としては、例えば、当該音声素片の音素名、先行音素、後続音素、後々続音素、基本周波数、音韻継続時間長、パワー、ストレスの有無、アクセント核からの位置、息継ぎからの時間、発声速度、感情などがある。

【００４９】

また、素片記憶部４２には、上記の他、音声素片の始端・終端でのケプストラム係数など、音声素片の音響特徴のうち音声素片の選択に用いる情報も蓄積されている。

【００５０】

以下では、素片記憶部４２に蓄積される音声素片の音韻・韻律環境と音響特徴量とを総称して、「素片環境」と呼ぶ。

【００５１】

図４に、素片記憶部４２に蓄積される素片環境の例を示す。図４に示す環境記憶部４３には、素片記憶部４２に蓄積される各音声素片の素片番号に対応して素片環境が記憶されている。ここでは、音韻・韻律環境として、音声素片に対応した音韻（音素名）、隣接音韻（この例では、当該音韻の前後それぞれ２音素ずつ）、基本周波数、音韻継続時間長が記憶され、音響特徴量として、音声素片始終端のケプストラム係数が記憶されている。

【００５２】

なお、これらの素片環境は、音声素片を切り出す元になった音声データを分析して抽出することによって得られる。また、図４では、音声素片の合成単位が音素である場合を示しているが、半音素、ダイフォン、トライフォン、音節、あるいはこれらの組み合わせや可変長であってもよい。

【００５３】

＜素片選択部＞

次に、図２の音声合成部４の動作を詳しく説明する。

【００５４】

図２において、情報入力部４１を介して素片選択部４３に入力された音韻系列は、素片選択部４７において、合成単位毎に区切られる。以下、この区切られた合成単位を、「セグメント」と呼ぶ。

【００５５】

素片選択部４３は、入力された入力音韻系列と入力韻律情報を基に、素片記憶部４２を参照し、各セグメントに対して、それぞれ、融合する複数個の音声素片の組み合わせを選択する。

【００５６】

このとき素片選択部４３は、各音声素片候補を用いて音声を合成した場合の合成音声と目標音声との歪みができるだけ小さくなるように、融合する音声素片の組み合わせを選択する。ここでは、素片選択部４３は、一般の素片選択型音声合成方法や従来の複数素片選択融合型音声合成方法と同様に、音声素片の選択の尺度として、各音声素片候補を用いて音声を合成した場合の合成音声と目標音声との歪みの大きさを間接的に表すコストを用い、このコストができるだけ小さくなるように、融合する音声素片の組み合わせを選択する。

【００５７】

ここで、「目標音声」とは、音声を合成する際の目標となる（仮想的な）音声、すなわち、入力された音韻の並びと韻律を実現し、かつ、理想的に自然な音声をいう。

【００５８】

最初に、コストについて説明する。

【００５９】

合成音声の目標音声に対する歪みの度合いを表すコストには、大きく分けて、目標コストと接続コストの２種類のコストがある。

【００６０】

目標コストは、コストの算出対象である音声素片（対象素片）を目標の音韻・韻律環境で使用することによって生じるコストである。

【００６１】

接続コストは、対象素片を隣接する音声素片と接続したときに生じるコストである。

【００６２】

具体的には、次の通りである。

【００６３】

目標コストとしては、音声素片が持つ基本周波数と目標の基本周波数の違い（差）によって生じる歪み（基本周波数コスト）、音声素片の音韻継続時間長と目標の音韻継続時間長の違い（差）によって生じる歪み（継続時間長コスト）、音声素片が属していた音韻環境と目標の音韻環境の違いによって生じる歪み（音韻環境コスト）などがある。　接続コストとしては、音声素片境界でのスペクトルの違い（差）によって生じる歪み（スペクトル接続コスト）や、音声素片境界での基本周波数の違い（差）によって生じる歪み（基本周波数接続コスト）などがある。

【００６４】

コストを用いて、一セグメント当たり複数個の音声素片を選択する方法については、どのような方法を用いても構わない。

【００６５】

例えば、特許文献２に開示された方法を用いても良い。ここでは、この選択方法の概要について、図５の処理手順例を参照しながら、一セグメント当たりＭ個の音声素片を選ぶ場合について説明する。

【００６６】

まず、ステップＳ１０１において、素片選択部４３は、入力された音韻系列を、合成単位毎のセグメントに分割する。ここで、分割されたセグメントの数をＮとする。

【００６７】
次に、ステップＳ１０２において、素片記憶部４２に記憶されている音声素片群の中から、各セグメントにつき１つずつの音声素片の系列を選択する。このときの選択においては、入力された目標の音韻系列・韻律情報と、素片記憶部４２の音声素片環境の情報を基に、系列としてのコストの総和（トータルコスト）が最小となるような音声素片の系列（最適素片系列）を求める。この最適素片系列の探索は、動的計画法（DP(dynamic programming)）を用いることで、効率的に行うことができる。
【００６８】
次に、ステップＳ１０３において、セグメント番号を表すカウンターｉに、初期値「１」をセットする。
【００６９】
次に、ステップＳ１０４において、セグメントｉに対する複数の音声素片候補の各々に対してコストを算出する。このときに用いるコストには、当該音声素片候補での目標コストと、当該音声素片候補の前後のセグメントの最適音声素片（最適素片系列に含まれる音声素片）と当該音声素片候補との接続コストとの和を用いる。
【００７０】
次に、ステップＳ１０５において、ステップＳ１０４で算出したコストを用いて、セグメントｉについて、コストの小さい上位Ｍ個の音声素片を選択する。
【００７１】
次に、ステップＳ１０６において、カウンターｉがＮ以下かどうかを判定する。
【００７２】
カウンターｉがＮ以下である場合（ステップＳ１０６のＹＥＳ）には、ステップＳ１０７に進んで、カウンターｉの値を１つ増やした後に、ステップＳ１０４に進んで、次のセグメントに係る処理を行う。
【００７３】
カウンターｉがＮに達した場合（ステップＳ１０６のＮＯ）には、この素片選択の処理を終了する。
【００７４】
このように、素片選択部４４は、各セグメントに対してＭ個ずつの音声素片を選択し、選択した音声素片を分離部４５に出力する。
【００７５】
素片選択部４４においてセグメント当たり複数個の音声素片を選択する方法は、上記した方法に限定する必要はなく、コストであっても、コスト以外であっても、何らかの評価尺度の下で、適切な音声素片の組を選べる方法であれば、いかなる方法を用いても良い。
【００７６】
＜素片融合部＞
素片融合部４４は、それぞれのセグメント毎に、素片選択部４３から入力された複数個の音声素片を融合して、新たな音声素片を生成する。
【００７７】
音声素片を融合する方法については、どのような方法を用いても構わない。
【００７８】
例えば、特許文献２に開示された方法を用いても良い。ここでは、この方法について図６及び図７を参照しながら説明する。
【００７９】
図６は、一つのセグメントに対する複数個の音声素片の波形を融合して、新たな音声波形を生成する手順を示すフローチャートである。図７は、あるセグメントに対して選択された３つの音声素片からなる素片組み合わせ候補（図中、６０）を融合して、新たな音声素片（図中、６３）を生成する例を示す図である。
【００８０】
まず、ステップＳ２０１において、（ある一つのセグメントについて）選択されたそれ

ぞれの音声素片からピッチ波形を切り出す。

【００８１】

ここで、「ピッチ波形」とは、その長さが音声の基本周期の数倍程度で、それ自身は基本周期を持たない比較的短い波形であって、そのスペクトルが音声信号のスペクトル包絡を表すものである。

【００８２】

このようなピッチ波形を抽出する方法には、どのような方法が用いられても良いが、その一つの方法として、基本周期同期窓を用いる方法があり、ここでは、この方法が用いられる場合を例にとって説明する。

【００８３】

具体的には、それぞれの音声素片の音声波形に対して基本周期間隔毎にマーク（ピッチマーク）を付し、このピッチマークを中心にして、窓長が基本周期の２倍のハニング窓で窓掛けすることによって、ピッチ波形を切り出す。図７のピッチ波形系列６１は、素片組み合わせ候補６０の各音声素片から切り出して得られたピッチ波形の系列の例を示している。

【００８４】

次に、ステップＳ２０２において、それぞれの音声素片に対するピッチ波形の個数が、音声素片間で同一になるように、ピッチ波形の数を揃える。

【００８５】

このときに、揃える対象となるピッチ波形の数は、目標の音韻継続時間長の合成音声を生成するために必要なピッチ波形数とするが、例えば、最もピッチ波形数の多いものに揃えても良い。

【００８６】

ピッチ波形の少ない系列は、系列に含まれるいくつかのピッチ波形を複製することによってピッチ波形数を増やし、ピッチ波形の多い系列は、系列中のいくつかのピッチ波形を間引くことによってピッチ波形数を減らす。図７のピッチ波形系列６２は、ピッチ波形の数を６つに揃えた例を示している。

【００８７】

次に、ステップＳ２０３において、ピッチ波形数を揃えた後、それぞれの音声素片に対応するピッチ波形系列中のピッチ波形を、その位置毎に融合することによって、新たなピッチ波形系列を生成する。

【００８８】

例えば、図７で生成された新たなピッチ波形６３に含まれるピッチ波形６３ａは、ピッチ波形系列６２のうち、６番目のピッチ波形６２ａ，６２ｂ，６２ｃを融合することによって得られる。このようにして生成された新たなピッチ波形系列６３を、融合された音声素片とする。

【００８９】

ここで、ピッチ波形を融合する方法としては、例えば、次のような方法がある。

【００９０】

第１の方法は、単純にピッチ波形の平均を計算する方法である。

【００９１】

第２の方法は、ピッチ波形間の相関が最大になるよう時間方向に各ピッチ波形の位置を補正してから平均化する方法である。

【００９２】

第３の方法は、ピッチ波形を帯域分割して、帯域毎にピッチ波形間の相関が最大になるようピッチ波形の位置を補正して平均化した結果を、帯域間で足し合わせる方法である。

【００９３】

いずれの方法を用いても良いが、本実施形態では、最後に説明した第３の方法を用いる場合を例にとって説明する。

【００９４】

素片融合部４４は、上記した方法を用いて、各セグメントについて、複数の音声素片を融合して新たな音声素片を生成し、フォルマント強調フィルタ部４５に出力する。

【００９５】

＜フォルマント強調フィルタ部＞

さて、上記のように融合によって生成された音声素片の音声波形は、融合の影響によって、融合元の音声素片の波形よりもスペクトル包絡がなまってしまい、いくつかのフォルマントが弱められてしまった結果、明瞭感が下がってしまうことが多い。そこで、フォルマント強調フィルタ部４５は、素片融合部４４から入力された融合素片に対して、フォルマントを強調するためのフィルタリングを行い、素片編集・接続部４７に出力する。

【００９６】

ここで用いるフォルマント強調フィルタとしては、例えば、J. Chenらの文献(J. Chen, etc., 「Adaptive Postfiltering for Quality Enhancement of Coded Speech」, IEEE Trans. Speech and Audio Processing, vol. 3, Jan 1995)（以下、文献３と呼ぶ。）によって開示されているものを、用いることができる。

【００９７】

こうしたフォルマント強調フィルタを、融合素片の音声波形に対して適用することによって、スペクトル包絡中のフォルマントを強調し、融合による明瞭性の低下を補償することが可能である。

【００９８】

フォルマント強調フィルタの概要を、文献３で開示されているフォルマント強調フィルタを例に用いて説明する。文献３で開示されているフォルマント強調フィルタは、数式（１）のような伝達関数を持つフィルタである。

【数１】

$$H(z) = G\frac{1-P(z/\beta)}{1-P(z/\alpha)}(1-\mu z^{-1}), \quad 0<\beta<\alpha<1 \qquad \cdots(1)$$

【００９９】

ただし、Ｐ（ｚ）は、数式（２）で表される。ここで、a_iは入力波形を線形予測分析したときのｉ番目の線形予測係数（ＬＰＣ）を表し、Ｍは線形予測次数である。

【数２】

$$P(z) = \sum_{i=1}^{M} a_i z^{-i} \qquad \cdots(2)$$

【０１００】

数式（１）における１／［１－Ｐ（ｚ／α）］は、α＝１の場合は、線形予測フィルタを表し、入力波形のＬＰＣスペクトルと同じ周波数応答を持つ。αを小さくすると、ＬＰＣスペクトルを鈍らせたような周波数応答になり、０に近づくにつれ、フラットな周波数応答になる。よって、入力波形のスペクトル中のパワーの大きい周波数成分は、より大きくなり、パワーの小さい周波数成分は、より小さくなるため、スペクトル中の山・谷を強調する効果を持つ。また、一般的な音声のスペクトル包絡には、低域から高域に向かって負の傾斜が見られるため、１／［１－Ｐ（ｚ／α）］の周波数応答は、全体的に、同様の負の傾斜を持つ。すなわち、スペクトルの山・谷を強調する効果に加え、副作用としてローパス特性を持っている。そこで、［１－Ｐ（ｚ／β）］および［１－μｚ$^{-1}$］の項によって、このローパス特性を補正する。［１－Ｐ（ｚ／β）］は、ＬＰＣスペクトルの極と同じ周波数に零点を持つフィルタであり、１／［１－Ｐ（ｚ／α）］でのスペクトルの傾斜を補償する効果を持つ。一方、［１－μｚ$^{-1}$］は、単純なハイパスフィルタで、残っているスペクトルの傾きを無くすよう調整するための項である。なお、Ｇは、フィルタリング前後でパワーが変化するのを防ぐためのパワー調整用のゲインであり、文献３で開示されている方法により、入力波形に応じて自動で決めることができる。

【０１０１】
このフォルマント強調フィルタでは、パラメータαを変えることによって、フォルマント強調の度合いを変えることができる（ただし、αの値に応じて、ローパス特性を補償するような適切なβ、μも決める必要がある）。αが１に近いほど強調の度合いが強く、αが小さくなるにつれ強調の度合いが弱まり、αが０．５以下になるとほとんど強調されない。どの程度フォルマントを強調すべきかは音声波形の特徴によって異なるが、これを決めるための客観尺度が存在しないため、通常、音声符号化や音声合成においてフォルマント強調フィルタを用いる場合には、フォルマント強調の度合いは主観評価などによって実験的に求める。

【０１０２】
しかしながら、複数素片選択融合型の合成方法においては、融合によるスペクトル包絡の鈍り具合がセグメントごとに大きく変わり得るため、１文など全体に対して同じパラメータを適用すると、融合によってスペクトル包絡が大きく鈍った箇所にはフォルマント強調の程度が不十分であるのに対し、逆に融合によるスペクトル包絡の鈍りが小さい箇所はフォルマントが強調されすぎて人工的な音になるという問題がある。

【０１０３】
そこで、本実施形態では、融合されてできたそれぞれの音声素片に対し（あるいは、それぞれの融合素片の各ピッチ波形に対し）、適切なフォルマント強調の度合いをフォルマント強調度合い推定部４６で推定し、フォルマント強調フィルタ部４５は、推定されたフォルマント強調度合いに応じてフォルマント強調フィルタの係数を変える。すなわち、融合素片ごとに（あるいは、融合素片のピッチ波形ごとに）、フォルマント強調度合いを適応的に制御する。ここで、フォルマント強調度合い推定部４６から与えられるフォルマント強調度合いは、例えば、０（強調無し）から１００（フォルマント強調フィルタの制御可能な範囲で最も強い強調）まで連続的に変化するようなものでもよいし、また、例えば、０（強調無し）から４（非常に強く強調）までの５段階で指定できるような離散的なものであってもよい。上述の文献３で開示されているフォルマント強調フィルタを用いる場合は、フォルマント強調度合い推定部４６で推定されたフォルマント強調度合いが大きい場合はαの値を１に近づけ、逆にフォルマント度合いが小さい場合はαを０．５に近づける。βおよびμの値もαの値に応じて変えるが、各αの値に対して適切なβとμの値は、実験的に求めることが可能である。

【０１０４】
また、フォルマント強調度合い推定部４６で推定されたフォルマント強調度合いを、フィルタ係数に具体的に反映するためのマッピングは、主観評価などによって実験的に得ることができる。

【０１０５】
本実施形態においては、文献３で開示されているフォルマント強調フィルタを用いる場合について説明したが、フォルマントが強調でき、フォルマントの強調度合いがパラメータなどで制御できるフォルマント強調フィルタであれば、いかなるものでも用いることができる。

【０１０６】
＜フォルマント強調度合い推定部＞
フォルマント強調度合い推定部４６は、素片選択部４３や素片融合部４４から与えられた融合素片や融合元の複数の音声素片の情報を元に、融合素片に対して適切なフォルマント強調度合いを推定し、推定したフォルマント強調度合いをフォルマント強調フィルタ部４５に出力する。

【０１０７】
前述のように、ある波形に対して適切なフォルマント強調度合いを決めるような客観尺度は存在しないが、融合素片と融合元の複数の音声素片の間でスペクトル包絡に関する特徴量を比較することによって、音声素片の融合によってどの程度スペクトル包絡が鈍ったかをある程度見積もることは可能である。そこで、フォルマント強調度合い推定部４６で

は、融合によるスペクトル包絡の鈍り具合を以下のような方法で推定し、これに基づいてフォルマントの強調度合いを決める。

【０１０８】

融合によるスペクトル包絡の鈍りが大きいほど、融合元の各音声素片と融合素片との間でスペクトル包絡の形状の差が大きくなると考えられる。そこで、融合元の各音声素片と融合素片との間でのスペクトル包絡の形状の差を見積もることができれば、音声素片の融合によるスペクトル包絡の鈍り具合を推定できると考えられる。

【０１０９】

スペクトル包絡の特徴を表すパラメータとしては、ケプストラムやＬＳＰ（線スペクトル対）などがある。以下では、ケプストラムの一つであるメル周波数ケプストラム係数（ＭＦＣＣ）を用いて、融合元の各音声素片と融合素片の間でのスペクトル包絡の形状の差を間接的に見積もる場合を例によって説明する。

【０１１０】

ＭＦＣＣは、音声認識の分野で広く用いられている音響特徴量で、音声合成においても上述の「スペクトル接続コスト」の評価尺度としてよく用いられる。ＭＦＣＣは、人間の聴覚特性を考慮した特徴量で、低い次元でもスペクトル包絡の特徴を良く表せる利点も持つ。ＭＦＣＣの低次の係数はスペクトル包絡の概形を、高次の係数はスペクトル包絡の細部を表現する。素片１と素片２のｉ次のＭＦＣＣをそれぞれc_{1i}、c_{2i}とすると、数式（３）により、素片１と素片２との間のＭＦＣＣ距離が算出できる。

【数３】

$$D_{MFCC} = \sum_{i=1}^{p} (c_{1i} - c_{2i})^2 \qquad \cdots (3)$$

【０１１１】

ただし、ｐはＭＦＣＣの次元を表す。

【０１１２】

なお、本例においては、ＭＦＣＣの次元は２０次程度とする。

【０１１３】

次に、このＭＦＣＣ距離を使って、音声素片の融合によるスペクトル包絡の鈍り具合を推定する方法について説明する。

【０１１４】

図８に、この場合の処理手順の一例を示す。

【０１１５】

ここで、融合素片の元になった融合元の素片数はＮとする。

【０１１６】

まず、融合素片のＭＦＣＣ（ｃ０）を算出する（ステップＳ３０１）。

【０１１７】

次に、カウンターｉを１に、D_{sum}を０に初期化して（ステップＳ３０２、ステップＳ３０３）、ステップＳ３０４に進む。

【０１１８】

ステップＳ３０４では、融合元のＮ個の音声素片のうち、ｉ番目の音声素片のＭＦＣＣ（ｃｉ）を算出する。

【０１１９】

次に、c_0とc_iとの間のＭＦＣＣ距離（D_i）を、数式（３）を用いて算出する（ステップＳ３０４）。

【０１２０】

次のステップＳ３０５では、算出されたD_iをD_{sum}に加算して、ステップＳ３０７に進む。

【０１２１】

ステップＳ３０７では、カウンターｉがＮ以下であるかを判定する。

【０１２２】

カウンターｉがＮ以下である場合（ステップＳ３０７のＹＥＳ）には、ステップＳ３０８に進んで、カウンターｉの値を１つ増やした後に、ステップＳ３０４に進んで、次の音声素片に係る処理を行う。

【０１２３】

カウンターｉがＮに達した場合（ステップＳ３０７のＮＯ）には、ループ処理を終了し、ステップＳ３０９に進む。

【０１２４】

ステップＳ３０９では、D_{sum}をＮで割ることによって、平均ＭＦＣＣ距離（D_{mean}）を求め、全ての処理を終了する。

【０１２５】

本実施形態では、このようにして求めた平均ＭＦＣＣ距離を、融合によるスペクトル包絡の鈍り具合を反映する評価尺度として用いる。すなわち、平均ＭＦＣＣ距離が小さいほどスペクトル包絡の鈍り具合が小さく、平均ＭＦＣＣ距離が大きいほどスペクトル包絡の鈍り具合が大きいとして、平均ＭＦＣＣ距離をそのままスペクトル包絡の鈍り具合とするか、平均ＭＦＣＣ距離の分布などに基づいて何らかの変換を行って得た値をスペクトル包絡の鈍り具合とする。

【０１２６】

次に、このようにして得たスペクトル包絡の鈍り具合に基づいて、フォルマント強調度合いを求める必要があるが、スペクトル包絡の鈍り具合が大きいほど強いフォルマント強調を施すべきと考えられるため、ここでは、スペクトル包絡の鈍り具合が増すとともに単調増加するような関数（ただし、フォルマント強調度合いが離散値の場合は、階段状に変化）を用いてフォルマント強調度合いに変換する。関数の形状については、例えば、スペクトル包絡の鈍り具合に対して途中まで線形に増加し、ある閾値を超えるとフォルマント強調度合いの上限値をとるようなものであっても良いし、シグモイド関数のように増加率がスペクトル包絡の鈍り具合に応じて変化するような形状のものであっても良く、それらの関数のパラメータ（傾き、など）は実験的に適切なものを得れば良い。

【０１２７】

なお、本実施形態においては、融合によるスペクトル包絡の鈍り具合を推定する方法の一例として、上記のＭＦＣＣを用いる方法を例にとって説明したが、スペクトル包絡の形状の差を適切に見積もれる音響パラメータであれば、どのようなものを用いてもよい。例えば、ＬＳＰ係数の二乗誤差を用いても良いし、ＦＦＴ（高速フーリエ変換）によって得られたＦＦＴスペクトルを確率分布のように見なすことによって、確率分布の差を計算するのによく用いられるＫＬ距離（Kullback-Leibler距離）を算出して、これを用いても良い。

【０１２８】

また、融合によるスペクトル包絡の鈍り具合を推定する方法として、素片選択部４３で算出された目標コストを用いる方法も考えられる。融合元の複数の音声素片がいずれも適切な音韻・韻律環境から選ばれた場合、目標コストは小さくなり、かつ、融合によるスペクトル包絡の鈍り具合も小さくなると考えられる。逆に、目標の音韻・韻律環境と異なる音声素片ばかりが選ばれた場合、目標コストは大きくなり、融合によるスペクトル包絡の鈍り具合も大きくなると考えられる。そこで、融合によるスペクトル包絡の鈍り具合を表す一つの指標として、融合元の音声素片が選ばれた際の目標コストを用いてもよいと考えられる。この方法は、前述の音響パラメータを用いる方法よりは間接的だが、非常に単純である。

【０１２９】

フォルマント強調度合い推定部４６は、上述のようにして推定した、融合によるスペクトル包絡の鈍り具合を、フォルマント強調フィルタ部４５に出力する。

【０１３０】

＜素片編集・接続部＞
素片編集・接続部４７は、フォルマント強調部４５から渡されたセグメント毎の音声素片を、入力韻律情報に従って変形して接続することによって、合成音声の音声波形を生成する。
【０１３１】
図９は、素片編集・接続部４７での処理を説明するための図である。図９には、フォルマント強調部４５から入力された、音素「ａ」「Ｎ」「ｓ」「ａ」「ａ」の各合成単位に対する音声素片を、変形・接続して、「ａＮｓａａ」という音声波形を生成する場合を示している。
【０１３２】
この例では、有声音の音声素片はピッチ波形の系列で表現されている。一方、無声音の音声素片は、フレーム毎の波形として表現されている。
【０１３３】
図９の点線は、目標の音韻継続時間長に従って分割した音素毎のセグメントの境界を表し、白い三角は、目標の基本周波数に従って配置した各ピッチ波形を重畳する位置（ピッチマーク）を示している。
【０１３４】
図９のように、有声音については音声素片のそれぞれのピッチ波形を対応するピッチマーク上の重畳し、無声音については各フレームの波形をセグメント中の各フレームに対応する部分に貼り付けることによって、所望の韻律（ここでは、基本周波数、音韻継続時間長）を持った音声波形を生成する。
【０１３５】
以上のように本実施形態によれば、素片融合によるフォルマントの鈍り具合に応じて、セグメントごとに適切な強さのフォルマント強調を行うので、こもり感やブザー感が少なく、かつ人工的でない高音質な合成音声を生成できる。
【０１３６】
（第２の実施形態）
本発明の第２の実施形態に係るテキスト音声合成を行うテキスト音声合成装置（音声処理装置）について説明する。
【０１３７】
第１の実施形態では、音声素片の融合処理およびフォルマント強調の処理に大きな計算量を要するため、ＣＰＵスペックが比較的低いミドルウェア向けの応用などには適用が向かないこともあり得る。
【０１３８】
そこで、本実施形態では、音声素片の融合およびフォルマント強調の処理を予め行った音声素片をオフラインで作成しておき、実際の動作時には、こうして作成された音声素片から適切な音声素片を選択して接続するだけの処理で合成波形を生成する。
【０１３９】
本実施形態に係るテキスト音声合成装置の全体構成例は、図１と同様であり、テキスト入力部１、言語処理部２、韻律処理部３、音声合成部４を備えている。
【０１４０】
図１０に、本実施形態の音声合成部４の構成例を示す。
【０１４１】
以下、図１０を参照しながら、本実施形態について、第１の実施形態と相違する点を中心に説明する。
【０１４２】
図１０に示されるように、本実施形態の音声合成部４は、情報入力部４１、素片記憶部４２、素片選択部４３、素片編集・接続部４７、音声波形出力部４８を備えている。
【０１４３】
第１の実施形態（図２）と比較すると、本実施形態の音声合成部４は、図２の素片融合

部４４、フォルマント強調フィルタ部４５、フォルマント強調度合い推定部４６が省かれている。

【０１４４】

また、本実施形態の素片記憶部４２には、後述の方法によって生成された融合済みの音声素片が格納されている。

【０１４５】

第１の実施形態の素片選択部４４が各セグメントに対して複数個ずつの音声素片を選択するのに対し、本実施形態の素片選択部４４は、各セグメントに対して１つずつの融合済み音声素片の最適系列を選択する。

【０１４６】

素片選択部４４の動作としては、例えば第１の実施形態で図５のフローチャートを用いる場合と比較すると、本実施形態では、図５のフローチャートのうち、ステップＳ１０１とステップＳ１０２だけを実行すればよい。もちろん、各セグメントに対して１つずつの融合済み音声素片の最適系列を選択する方法は、これに限られるものではなく、種々の方法が可能である。

【０１４７】

なお、素片編集・接続部４７および音声波形出力部４８の動作は、第１の実施形態のものと同様である。

【０１４８】

次に、音声素片記憶部４２に格納する融合済みの音声素片を学習する方法について、図１１及び図１２を参照しながら説明する。

【０１４９】

本実施形態では、融合済みの音声素片を作成する融合済み音声素片作成部５を用いる。融合済み音声素片作成部５は、図１０のテキスト音声合成装置に含まれても良い。この場合、テキスト音声合成に供するための「フォルマント強調された融合素片」の生成時には、図１の音声合成部４を融合済み音声素片作成部５に置き換えた構成で用いれば良い。

【０１５０】

また、融合済み音声素片作成部５は、テキスト音声合成装置に含まれなくても良い。この場合、例えば、融合済み音声素片作成部５を、独立した音声処理装置（テキスト音声合成に供するための「フォルマント強調された融合素片」を生成する音声処理装置）として構成しても良い。この場合、独立した音声処理装置は、図１の音声合成部４を融合済み音声素片作成部５に置き換えた構成にすれば良い。

【０１５１】

図１１に、融合済み音声素片作成部５の構成例を示す。

【０１５２】

融合済み音声素片作成部５の構成は、第１の実施形態の音声合成部４の構成とほとんど同じであるため、ここでは相違する点について説明する。

【０１５３】

融合済み音声素片作成部５は、第１の実施形態の音声合成部４の素片編集・接続部４７および音声波形出力部４８の代わりに、音声素片出力部４９を持つ。第１の実施形態の音声合成部４の素片編集・接続部４７および音声波形出力部４８は、フォルマント強調部４５から入力された各セグメントに対する音声素片を接続して、入力テキストに対する合成波形を生成するのに対し、音声素片出力部４９は、フォルマント強調部４５から入力された音声素片をそのまま出力する。

【０１５４】

すなわち、融合済み音声素片作成部５は、音声素片（フォルマント強調された融合素片）を、図１０のテキスト音声合成装置の音声素片記憶部４２へ出力し、音声素片（フォルマント強調された融合素片）は、音声素片記憶部４２に記憶される。

【０１５５】

次に、音声素片記憶部４２に格納する融合済みの音声素片を学習する手順について説明

する。
【０１５６】
図１２に、この場合の処理手順の一例を示す。
【０１５７】
まず、ステップＳ５０１において、融合済み音声素片作成部５を備えたテキスト音声合成装置又は独立した音声処理装置に対して、大量の文を入力する。
【０１５８】
次に、ステップＳ５０２において、入力された各文の各セグメントに対して生成された融合済み音声素片が、融合済み音声素片生成部５から出力される。
【０１５９】
次に、ステップＳ５０３において、外部から指定された音声素片記憶部４２に格納する音声素片の総数のうち、それぞれの素片種別に対して幾つずつ配分するかを決める。
【０１６０】
ここで、素片種別とは、音声素片の音韻環境などで分類された種別を指す。例えば、素片種別／ａ／は、音素／ａ／に対応する音声素片のこととする。
【０１６１】
各素片種別に何個ずつ素片を配分するかは、各素片種別の音声素片の出現頻度などに応じて決める。例えば、素片種別／ａ／の素片が素片種別／ｕ／の素片よりも出現頻度が高い場合は、素片種別／ａ／に多めの素片を配分することとする。
【０１６２】
素片種別ｉに配分する音声素片の個数をN_iとする。
【０１６３】
次に、ステップＳ５０４において、素片種別番号ｉに初期値１をセットする。
【０１６４】
次に、ステップＳ５０５において、素片種別ｉの融合済み音声素片を、ステップＳ５０２で出力された素片種別ｉの音声素片の中から、出現頻度が上位のものをN_iずつ抽出する。
【０１６５】
次に、ステップＳ５０６において、ｉと素片種別数を比較する。
【０１６６】
ｉが素片種別数以下である場合（ステップＳ５０６のＹＥＳ）には、ステップＳ５０７に進んで、ｉの値を１つ増やし、そして、ステップＳ５０５～ステップＳ５０６を繰り返す。
【０１６７】
ｉが素片種別数を超えている場合（すなわち、全ての素片種別に対する処理が完了している場合）（ステップＳ５０６のＮＯ）には、全ての処理を終了する。
【０１６８】
上記のようにして抽出した融合済み音声素片を、音声素片記憶部４２に格納する。
【０１６９】
ここで、音声素片記憶部４２に格納するために選択する音声素片の個数は、トータルでの音声素片サイズと合成音声の音質とのトレードオフで、任意に決めることができる。より多くの音声素片を選択して格納すれば、サイズは大きくなるが、合成音声の音質を高くすることができ、音声素片の数を減らせば、合成音声の音質は犠牲になるが、サイズを小さくすることができる。
【０１７０】
なお、上記では出現頻度の高い素片を抽出する方法を説明したが、音声素片の両端で算出したメルケプストラムなどの音声素片の特徴量を用いて抽出しても良い。
【０１７１】
この場合、各素片種別に対して出力された融合済み音声素片をそれぞれ、音声素片の特徴量を用いてクラスタリングし、分割された各クラスタの中心（セントロイド）に最も近

い素片を抽出する。クラスタリングにおけるクラスタ数は、各素片種別に配分する素片数に応じて決める。

【０１７２】

出現頻度に基づいて素片を抽出する場合は、出現頻度が低いコンテキストに対して適切な素片が抽出されない可能性があり、入力テキストによっては音質が大きく劣化してしまう可能性があるが、本方法によって素片を抽出した場合、特徴量空間をできるだけ広く覆うような音声素片のセットが抽出できるため、出現頻度に基づいて抽出した場合より安定した合成音が生成できる。

【０１７３】

以上のように本実施形態によれば、複数の音声素片を融合する処理とフォルマント強調の処理を予めオフラインで行うので、第１の実施形態よりも少ない計算量で実現でき、ＣＰＵスペックが比較的低いミドルウェア向けなどの応用にも適用可能である。

【０１７４】

また、合成音声の音質とのトレードオフで、格納する音声素片のトータルのサイズもスケーラブルに決めることができる。

【０１７５】

（第３の実施形態）

本発明の第３の実施形態に係るテキスト音声合成装置について説明する。

【０１７６】

本実施形態は、フォルマント強調度合い推定部４６の推定方法が、第１の実施形態で説明した例とは相違するものであり、以下、この相違点を中心に説明する。

【０１７７】

第１の実施形態では、フォルマント強調度合い推定部４６でフォルマント強調度合いを推定する方法として、融合元の各音声素片と融合素片の間でのスペクトル包絡の差を算出することによって推定する方法を説明したが、融合元の各音声素片と融合素片との間でのスペクトル包絡の差と、融合によるスペクトル包絡の鈍り具合の間には、高い相関はあると考えられるものの、直接的な関係があるわけではない。そこで、スペクトル包絡の鈍り具合を、より直接的に求められる方法があれば、より確度の高い推定を行うことが可能と考えられる。

【０１７８】

その一つの方法として、線形予測極（ＬＰ極）を用いる方法が考えられる。ＬＰ極は、数式（２）のＰ（ｚ）について（１－Ｐ（ｚ））を＝０とおいたときに得られる解（複素数）のことで、この解のｚ平面上での位置と単位円との関係から、各フォルマントの周波数とバンド幅を推定することができる。それぞれの極が各フォルマントに対応すると考えられ、ｉ番目の極に関して、極と原点を結ぶ線の角度をθ_i、極と原点の距離をr_iとした場合、ｉ番目のフォルマントの周波数F_iおよびバンド幅BW_iは、数式（４）のように推定できる。

【数４】

$$F_i = \frac{\theta_i}{2\pi T_s}, BW_i = \frac{-\ln(r_i)}{\pi T_s} \qquad \cdots (4)$$

【０１７９】

このようにして推定した各フォルマントの周波数とバンド幅を用いれば、スペクトル包絡のうち、特にフォルマントに関する鈍り具合がより正確に推定できると考えられる。

【０１８０】

以下、ＬＰ極から推定される各フォルマントのバンド幅を用いて、スペクトル包絡の鈍り具合を推定する方法の一例を、図１３を参照しながら説明する。

【０１８１】

図１３に、ＬＰ極から推定される各フォルマントのバンド幅を用いてスペクトル包絡の鈍り具合を推定する手順の一例を示す。

【０１８２】

まず、ステップＳ６０１において、融合素片のＬＰ極を算出する。具体的には、融合素片の音声波形に対してＬＰＣ分析を行い、得られた線形予測係数を係数に持つ数式（２）のＰ（ｚ）について、（１－Ｐ（ｚ））＝０とおいたときの解を得る。

【０１８３】

次のステップＳ６０２では、融合元の音声素片それぞれに対するＬＰ極を、ステップＳ６０１と同様の方法で算出する。

【０１８４】

次に、ステップＳ６０３では、フォルマントバンド幅比率の和R_{sum}を０に、ステップＳ６０４では、用いたＬＰ極の個数N_{LP}を０に、ステップＳ６０５では、カウンターｉを１に、それぞれ初期化して、ステップＳ６０６に進む。

【０１８５】

ステップＳ６０６では、融合素片のｉ番目のＬＰ極が実軸上（すなわち虚数項が０）かどうかを判定し、実軸上である場合（ステップＳ５０６のＹＥＳ）には、ステップＳ６２０に進んで、カウンターｉの値を１つ増やした後に、再びＳ６０６に進む。

【０１８６】

これは、実軸上のＬＰ極がフォルマントには対応しない（スペクトル包絡全体の形状に寄与）ため、実軸上である場合については、ステップＳ６０７以降の処理をスキップし、フォルマントに対応したＬＰ極のみを考慮するためのものである。

【０１８７】

ＬＰ極が実軸上でない場合（ステップＳ６０６のＮＯ）には、ステップＳ６０７に進む。

【０１８８】

ステップＳ６０７では、N_{LP}の値を１つ増やした後に、ステップＳ６０８に進む。

【０１８９】

ステップＳ６０８では、融合素片のｉ番目のＬＰ極に対するフォルマントのバンド幅ＢＷｉを、数式（４）を用いて算出する。

【０１９０】

次のステップＳ６０９では、融合元の音声素片のフォルマントに関するバンド幅の和$BW_{i_org_sum}$を０に初期化し、ステップＳ６１０に進む。

【０１９１】

ステップＳ６１０では、カウンターｋを１に初期化して、ステップＳ６１１に進む。

【０１９２】

ステップＳ６１１では、融合元の音声素片（計N_{fused}個）のうちｋ番目の音声素片（「音声素片ｋ」と呼ぶ。）について、この音声素片のＬＰ極の中で、融合素片のｉ番目のＬＰ極が表すフォルマントに対応するようなＬＰ極を選択する。具体的には、音声素片ｋのＬＰ極の中で、融合素片のｉ番目のＬＰ極に最も近いものを選択する。ＬＰ極の間の距離については、例えば数式（５）（文献“Goncharoff, etc., 「Interplation of LPC spectra via pole shifting.」, IEEE ICASSP, Detroit, MI, Vol.1, pp.780-783, 1995”参照）を用いて算出できる。ただし、p_iはＬＰ極の複素数表現、r_iはＬＰ極と原点の距離を表し、Ｄ（ｐ０，ｐ１）がＬＰ極p_0とp_1の距離を表す。

【数５】

$$D(p_0,p_1) = \begin{cases} |\ln(\frac{p_1}{p_0})|\{\frac{\ln((1-r_0^2)/(1-r_1^2))}{\ln(r_1/r_0)}\}, & r_0 \neq r_1 \\ |\ln(\frac{p_1}{p_0})|\{2r^2/(1-r^2)\}, & r = r_0 = r_1 \end{cases} \quad \cdots (5)$$

【０１９３】
この数式（５）によって、融合素片のｉ番目のＬＰ極との距離を、融合元の音声素片のＬＰ極のそれぞれについて算出し、最も距離が小さいＬＰ極を選択する。
【０１９４】
次のステップＳ６１２では、ステップＳ６１１で選択されたＬＰ極に対するバンド幅$BW_{i_org_k}$を、数式（４）を用いて算出する。
【０１９５】
次に、ステップＳ６１３において、ステップＳ６１２で算出した$BW_{i_org_k}$を$BW_{i_org_sum}$に加算する。
【０１９６】
続いて、ステップＳ６１３において、カウンターｋが融合元の音声素片数N_{fused}以下かどうかを判定する。
【０１９７】
カウンターｋがN_{fused}以下である場合（ステップＳ６１３のＹＥＳ）には、ステップＳ６１９に進んで、カウンターｋの値を１つ増やした後に、ステップＳ６１１からのステップを繰り返す。一方、カウンターｋがN_{fused}を超える場合（ステップＳ６１３のＮＯ）には、ステップＳ６１５に進む。
【０１９８】
ステップＳ６１５では、$BW_{i_org_sum}$をN_{fused}で割ることによって、融合素片のｉ番目のＬＰ極に対応するような、融合元の各音声素片のＬＰ極についての、フォルマントのバンド幅の平均値$BW_{i_org_mean}$を算出する。
【０１９９】
次のステップＳ６１６では、ステップＳ６１５で算出した$BW_{i_org_mean}$に対する、融合素片のｉ番目のＬＰ極のバンド幅ＢＷｉの比率を、フォルマントバンド幅比率の和R_{sum}に加算する。
【０２００】
続いて、ステップＳ６１７では、カウンターｉが、N_{maxLP}という設定値以下かどうかを判定する。
【０２０１】
ここで、N_{maxLP}は、フォルマントの鈍り具合を推定するのに用いるＬＰ極の個数の最大値を表す。
【０２０２】
この値は、例えば、ＬＰＣ分析での分析次数の１／２などに設定する。
【０２０３】
カウンターｉがN_{maxLP}以下である場合（ステップＳ６１７のＹＥＳ）には、ステップＳ６２０に進んで、カウンターｉの値を１つ増やした後に、ステップＳ６０６からの処理を繰り返す。一方、カウンターｉがN_{maxLP}を越える場合（ステップＳ６１７のＮＯ）には、ステップＳ６１８に進む。
【０２０４】
ステップＳ６１８では、フォルマントバンド幅比率の和R_{sum}を、用いたＮＰ極の個

数N_{LP}で割ることによって、フォルマントバンド幅比率の平均値R_{mean}を算出し、全ての処理を終了する。

【０２０５】

本実施形態では、上記のような方法で算出したフォルマントバンド幅比率の平均値R_{mean}を、音声素片の融合によるスペクトル包絡の鈍り具合を表す尺度として用いる。この値は、フォルマントのバンド幅がほぼ変わらずスペクトル包絡がほとんど鈍らなかった場合には１に近い値、フォルマントのバンド幅が融合元の音声素片より広がってスペクトル包絡が鈍った場合には１より大きい値となり、スペクトル包絡の鈍り具合が大きければ大きいほど大きな値になると考えられる。そこで、本実施形態においては、R_{mean}が１以下の場合は強調無しで、１より大きい場合は強調度合いが単調増加するような何らかの関数を用いることによって、このR_{mean}からフォルマント強調度合いを算出することとする。

【０２０６】

このように、融合素片と融合元の各音声素片に対して推定されたフォルマントのバンド幅を用いることによって、スペクトル包絡形状の差を用いる場合よりも、融合によるスペクトル包絡の鈍り具合をより直接的に求められるので、フォルマント強調度合いをより高い確度で推定することが可能である。

【０２０７】

（第４の実施形態）

本発明の第４の実施形態に係るテキスト音声合成装置について説明する。

【０２０８】

本実施形態は、フォルマント強調度合い推定部４６の推定方法が、第１、第３の実施形態で説明した例とは相違するものであり、以下、この相違点を中心に説明する。

【０２０９】

第３の実施形態においては、融合素片の各ＬＰ極に対して求めたフォルマントバンド幅比率を平均化することによって、スペクトル包絡全体での鈍り具合を推定しているが、実際にはスペクトルの鈍り具合がフォルマント毎で異なる場合も考えられる。そこで、各ＬＰ極に対して求めたフォルマントバンド幅比率（以下、R_iとする。）をそのまま用いることによって、フォルマントごとに強調度合いが異なるようなフォルマント強調を行うことも可能である。

【０２１０】

ここで、融合素片のｉ番目のＬＰ極をp_iとすると、数式（２）のＰ（ｚ）に関して、数式（６）のように表せる。

【数６】

$$1-P(z)=\prod_{i=1}^{M}(1-p_i z^{-i}) \qquad \cdots(6)$$

【０２１１】

Ｈ（ｚ）＝１／（１－Ｐ（ｚ））という伝達関数を持つフィルタ（線形予測フィルタ）に、ＬＰＣ分析したときの予測残差を入力すると完全に元の波形が再現できるが、上記のp_iをｚ平面上の単位円に近づくように変更したフィルタに予測残差を入力すると、ｉ番目のＬＰ極に対応するフォルマントのバンド幅が狭まり、結果的に、このフォルマントを強調することができる。すなわち、R_iに応じて適切にp_iを変更したフィルタをフォルマント強調フィルタとして用いれば、フォルマントごとに適切なフォルマント強調を行うことができる。

【０２１２】

図１４に、本実施形態のフォルマント強調フィルタ部４５の構成例を示す。

【０２１３】

ＬＰＣ分析部４５１は、入力された波形に対してＬＰＣ分析を行い、算出されたＬＰＣをＬＰＣ変形部４５２に、予測残差を線形予測フィルタ部４５３に出力する。

【０２１４】

ＬＰＣ変形部４５２では、フォルマント強調度合い推定部４６から入力された各ＬＰ極に対するフォルマントバンド幅比率R_iに応じてＬＰＣ係数を変形し、この変形されたＬＰＣ係数を線形予測フィルタ部４５３に与える。

【０２１５】

線形予測フィルタ部４５３では、ＬＰＣ変形部４５２から与えられたＬＰＣ係数をフィルタ係数に用いて、ＬＰＣ分析部４５１から入力された予測残差をフィルタリングすることによって、フォルマント強調された波形を出力する。

【０２１６】

なお、ＬＰＣ変形部４５２においては、まず、入力されたＬＰＣ係数から数式Ｐ（ｚ）を得た後、（１－Ｐ（ｚ））を数式（６）のように因数分解することによってＬＰ極p_iを得る。

【０２１７】

次に、ＬＰ極p_iをR_iに応じて変更する。

【０２１８】

例えば、数式（７）のように変更すれば、フォルマントのバンド幅は１／R_i倍となり、融合元の音声素片での平均的なフォルマントのバンド幅に近づくようバンド幅を狭めることが可能である。

【数７】

$$p_i = |p_i|^{\frac{1}{R_i}-1} p_i \qquad \cdots (7)$$

【０２１９】

このような方法でR_iに応じて変更したＬＰ極p_iを数式（６）に代入して、この数式を展開することによって、変形されたＬＰＣ係数を得ることができる。

【０２２０】

本実施形態においては、融合素片および融合元の各音声素片に対して求めたＬＰ極を用いてフォルマントごとに強調度合いを変える方法を説明したが、この方法以外にも、フォルマントごとあるいは周波数帯域によって強調度合いが変わるようなフォルマント強調を行うことも可能である。

【０２２１】

例えば、フォルマント強調度合い推定部４６において、融合素片および融合元の各音声素片の波形を複数の周波数帯域に分割し、それぞれの帯域においてスペクトル包絡の鈍り具合を推定することによって、それぞれの帯域でのフォルマント強調度合いを推定する。そして、フォルマント強調フィルタ部４５において、融合素片の波形を帯域分割して得た各周波数帯域の波形に対し、フォルマント強調度合い推定部４６から入力された各帯域の強調度合いに従ってフォルマント強調した後、周波数帯域間で波形を足し合わせれば、各周波数帯域でのスペクトル包絡の鈍り具合に応じたスペクトル強調を行うことが可能である。

【０２２２】

なお、以上の各機能は、ソフトウェアとして記述し適当な機構をもったコンピュータに処理させても実現可能である。

また、本実施形態は、コンピュータに所定の手順を実行させるための、あるいはコンピュータを所定の手段として機能させるための、あるいはコンピュータに所定の機能を実現させるためのプログラムとして実施することもできる。加えて該プログラムを記録したコンピュータ読取り可能な記録媒体として実施することもできる。

【０２２３】

なお、本発明は上記実施形態そのままに限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で構成要素を変形して具体化できる。また、上記実施形態に開示されている複数の構成要素の適宜な組み合わせにより、種々の発明を形成できる。例えば、実施形態に示される全構成要素から幾つかの構成要素を削除してもよい。さらに、異なる実施形態にわたる構成要素を適宜組み合わせてもよい。

【図面の簡単な説明】

【０２２４】

【図１】本発明の一実施形態に係るテキスト音声合成装置の構成例を示すブロック図

【図２】同実施形態に係る音声合成部の構成例を示すブロック図

【図３】同実施形態に係る音声素片記憶部に蓄積される音声素片の例を示す図

【図４】同実施形態に係る音声素片記憶部に蓄積される素片属性情報の例を示す図

【図５】音声素片の選択手順の一例を示すフローチャート

【図６】音声波形を融合して新たな音声波形を生成する手順の一例を示すフローチャート

【図７】選択された３つの音声素片からなる素片組み合わせ候補を融合して新たな音声素片を生成する例について説明するための図

【図８】音声素片の融合によるスペクトル包絡の鈍り具合を推定する手順の一例を示すフローチャート

【図９】同実施形態に係る素片編集・接続部での処理を説明するための図

【図１０】同実施形態に係る音声合成部の他の構成例を示すブロック図

【図１１】同実施形態に係る融合済み音声素片作成部の構成例を示すブロック図

【図１２】融合済みの音声素片を学習する手順の一例を示すフローチャート

【図１３】フォルマント強調度合いを推定する手順の一例を示すフローチャート

【図１４】同実施形態に係るフォルマント強調フィルタ部の構成例を示すブロック図

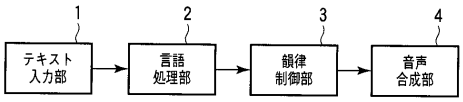
【符号の説明】

【０２２５】

１…テキスト入力部、２…言語処理部、３…韻律処理部、４…音声合成部、４１…音韻系列・韻律情報入力部、４２…音声素片記憶部、４３…素片選択部、４４…素片融合部、４５…フォルマント強調フィルタ部、４６…フォルマント強調度合い推定部、４７…素片編集・接続部、４８…音声波形出力部、４９…音声素片出力部、５…融合済み音声素片作成部、４５１…ＬＰＣ分析部、４５２…ＬＰＣ変形部、４５３…線形予測フィルタ部

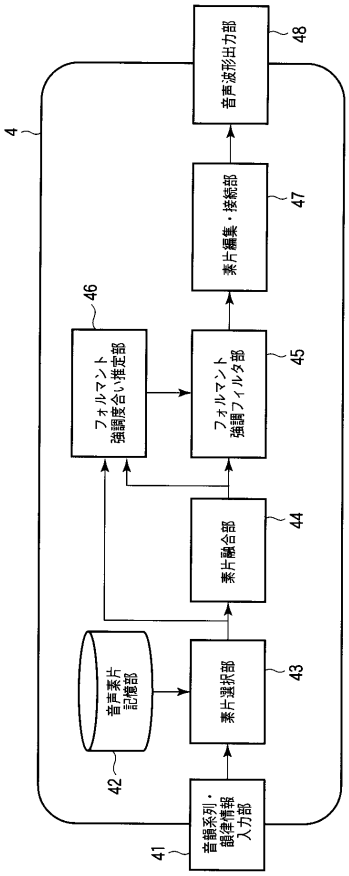
【図１】

図１



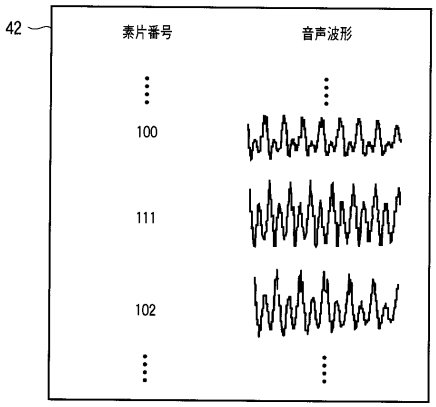
【図２】

図２



【図３】

図３



【図４】

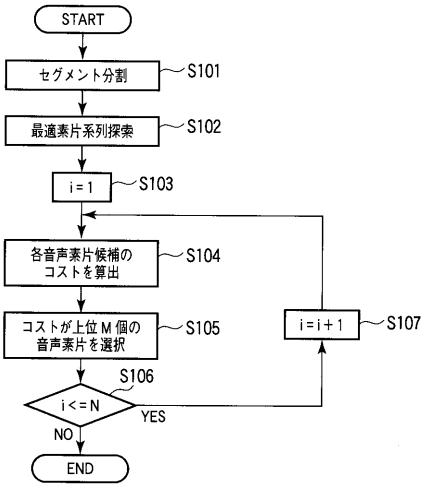
図４

42

素片番号	当該音韻	隣接音韻 前後2音韻ずつ	基本周波数	音韻継続時間長	ケプストラム係数 始端	ケプストラム係数 終端
…	…	…	…	…	…	…
100	/a/	/-//k/, /i//m/	221Hz	83msec	2.54, 0.24, …	2.49, 0.18, …
101	/a/	//a//m/, /k//e/	296Hz	125 msec	2.33, 0.28, …	2. 55, 0.22, …
102	/i/	/o//k/, /r//u/	240Hz	61msec	2.54, -0. 35, …	2.23, 0. 02, …
…	…	…	…	…	…	…

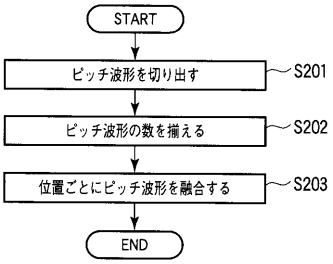
【図５】

図５



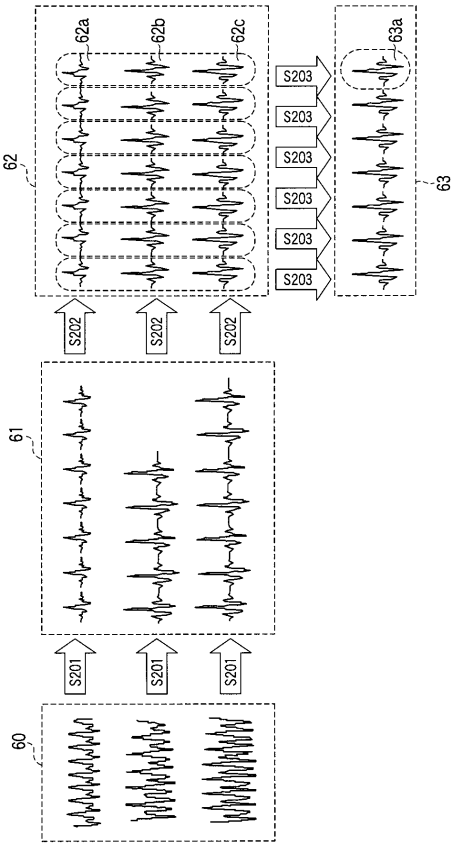
【図６】

図６



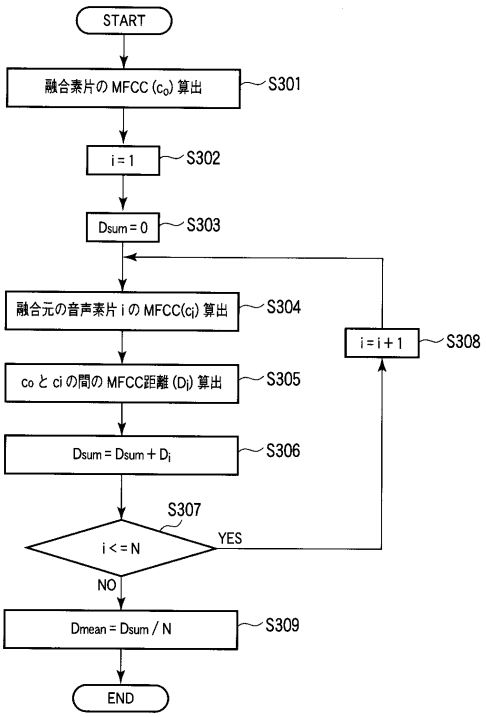
【図７】

図７



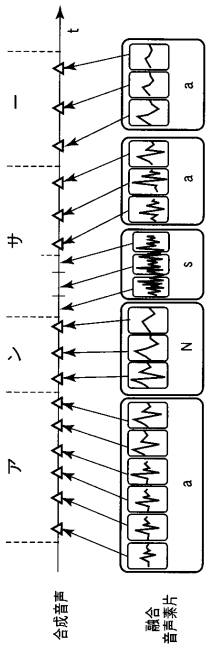
【図８】

図８



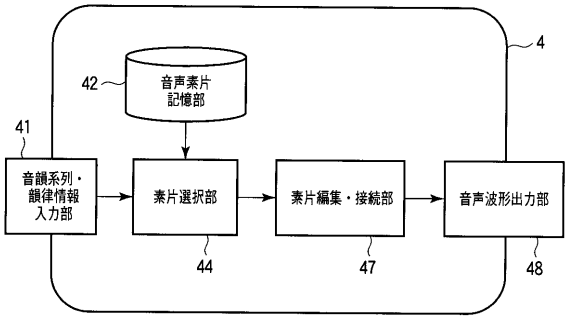
【図９】

図９



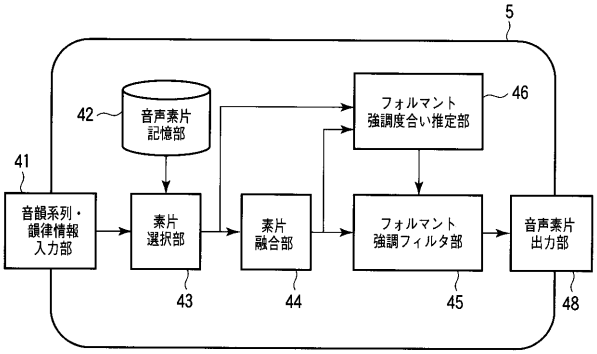
【図１０】

図 10



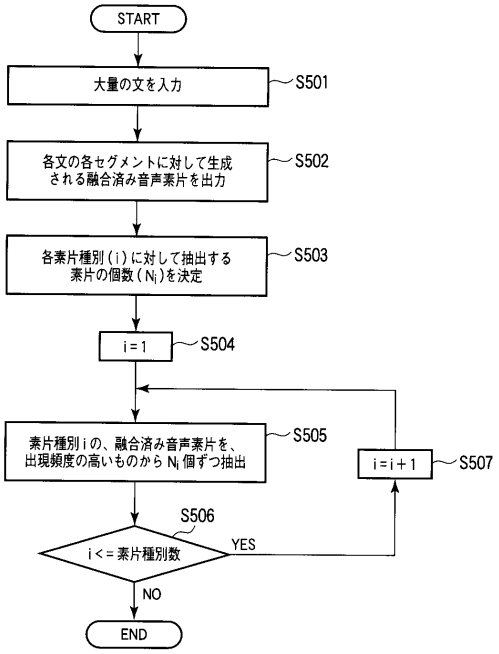
【図１１】

図 11



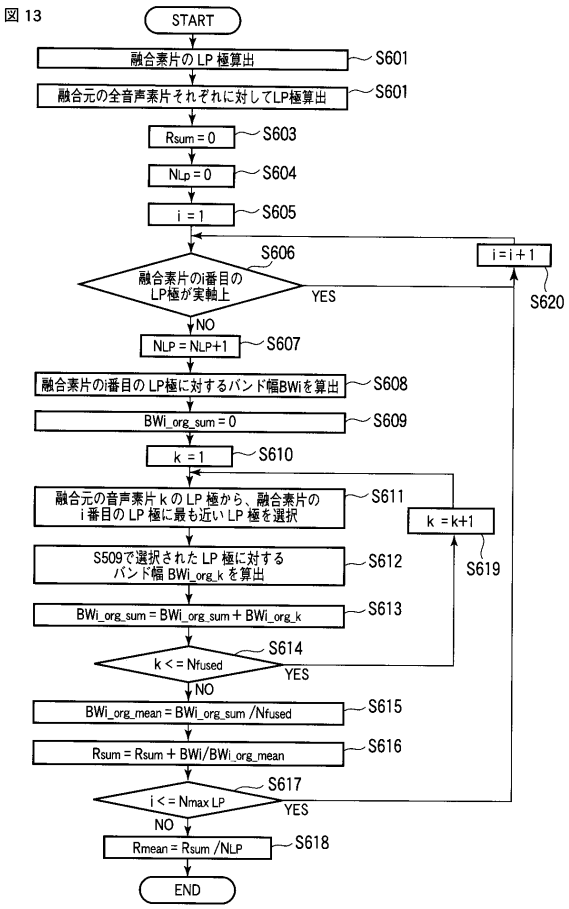
【図１２】

図 12



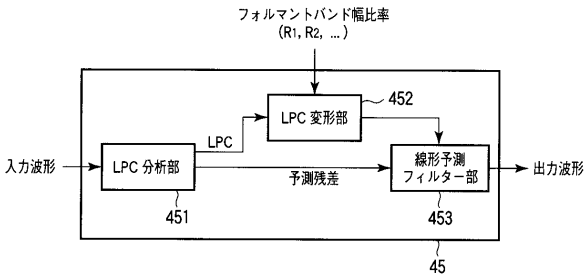
【図１３】

図 13



【図１４】

図 14



フロントページの続き

(74)代理人 100095441
弁理士 白根 俊郎
(74)代理人 100084618
弁理士 村松 貞男
(74)代理人 100103034
弁理士 野河 信久
(74)代理人 100119976
弁理士 幸長 保次郎
(74)代理人 100153051
弁理士 河野 直樹
(74)代理人 100140176
弁理士 砂川 克
(74)代理人 100100952
弁理士 風間 鉄也
(74)代理人 100101812
弁理士 勝村 紘
(74)代理人 100070437
弁理士 河井 将次
(74)代理人 100124394
弁理士 佐藤 立志
(74)代理人 100112807
弁理士 岡田 貴志
(74)代理人 100111073
弁理士 堀内 美保子
(74)代理人 100134290
弁理士 竹内 将訓
(74)代理人 100127144
弁理士 市原 卓三
(74)代理人 100141933
弁理士 山下 元
(72)発明者 森田 眞弘
東京都港区芝浦一丁目１番１号 株式会社東芝内
(72)発明者 籠嶋 岳彦
東京都港区芝浦一丁目１番１号 株式会社東芝内
(72)発明者 平林 剛
東京都港区芝浦一丁目１番１号 株式会社東芝内