(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4489371号
(P4489371)
(45)発行日　平成22年6月23日(2010.6.23)　(24)登録日　平成22年4月9日(2010.4.9)

(51)Int.Cl.　　F I
G10L 13/02 (2006.01)　　G10L 13/02 123
G10L 19/06 (2006.01)　　G10L 19/06 B

請求項の数 15　(全 14 頁)

(21)出願番号　特願2003-125758(P2003-125758)
(22)出願日　平成15年4月30日(2003.4.30)
(65)公開番号　特開2003-323200(P2003-323200A)
(43)公開日　平成15年11月14日(2003.11.14)
審査請求日　平成18年4月28日(2006.4.28)
(31)優先権主張番号　10/134281
(32)優先日　平成14年4月29日(2002.4.29)
(33)優先権主張国　米国(US)

(73)特許権者　392026693
株式会社エヌ・ティ・ティ・ドコモ
東京都千代田区永田町二丁目１１番１号
(74)代理人　100098084
弁理士　川▲崎▼　研二
(72)発明者　ワイ　チュウ
アメリカ合衆国、カリフォルニア州　９５１１２、サンノゼ、１７００、ノース・ファースト・ストリート　１３０
(72)発明者　コズロウ　ラシュキャリ
アメリカ合衆国、カリフォルニア州　９４５３９、フリーモント、１５２５、サラマンカ　コート

審査官　山下　剛史

最終頁に続く

(54)【発明の名称】合成音声を最適化する方法、音声合成フィルタを生成する方法、音声最適化方法及び音声最適化装置

(57)【特許請求の範囲】
【請求項１】
合成音声を最適化する方法において、
原音サンプルを使用して音声合成多項式の第１の係数を計算して、第１合成音声サンプルを生成する過程と、
前記原音サンプルと前記第１合成音声サンプルとの間の第１誤差を計算する過程と、
前記第１誤差の勾配を計算する過程と、
前記勾配を使用して前記音声合成多項式の第２の係数を計算し、第２合成音声サンプルを生成する過程と、
前記原音サンプルと前記第２合成音声サンプルとの間の第２誤差を計算する過程と、
前記第２誤差が前記第１誤差より小さい場合は前記の第２の係数を選択し、前記第２誤差が前記第１誤差より大きい場合は前記第１の係数を選択する過程と、
システムの不安定さに関して前記第２の誤差をテストし、前記テスト結果が不安定であるとなった場合は前記第１係数を選択するテスト過程と
を有することを特徴とする方法。
【請求項２】
請求項１に記載の方法において、
前記勾配に適用されたステップサイズを使用して、前記第２の係数が繰り返し計算され、
各繰り返しにおいて誤差が計算され、各繰り返しでの前記誤差が減少しなくなったら繰

り返しを終える
ことを特徴とする方法。
【請求項３】
請求項２に記載の方法において、
前記第２の誤差の勾配を計算する過程と、
前記第２の誤差の勾配を使用して前記音声合成多項式の第３の係数を計算して、第３の合成音声サンプルを生成する過程と
を更に有し、
前記第３の係数は、前記第２の誤差の勾配に適用されるステップサイズを使用して繰り返し計算され、
各繰り返しにおいて１つの誤差が計算され、
各繰り返しにおける誤差が減少しなくなると繰り返しを終了し、
前記原音サンプルと前記第３の合成音声サンプルとの間の第３の誤差を計算する過程と、
システムの不安定さに関して前記第３の誤差をテストする過程と、
前記テストにより不安定であるとされた場合、前記第２の係数を選択する過程と
を更に有することを特徴とする方法。
【請求項４】
請求項１に記載の方法において、
前記第２の係数は前記勾配に適用される適応ステップサイズを用いて計算され、
前記適応ステップサイズは前記勾配と前記第１の係数との関数である
ことを特徴とする方法。
【請求項５】
請求項１に記載の方法において、
前記テスト過程は更に、
前記第２誤差と前記第１誤差との間の差分を計算する過程と、
前記差分が終了閾値より大きいかどうかを判定する過程と
を有することを特徴とする方法。
【請求項６】
請求項１に記載の方法において、
前記勾配に聴覚重み付けを与える過程を
更に有することを特徴とする方法。
【請求項７】
請求項１に記載の方法において、
前記勾配に適用されたステップサイズを使用して前記第２の係数が繰り返し計算され、
各繰り返しにおいて１つの誤差が計算され、
前記誤差が各繰り返しで減少しなくなったら前記繰り返しを終了し、
前記勾配に聴覚重み付けを行う過程と
を更に有することを特徴とする方法。
【請求項８】
音声合成フィルタを生成する方法において、
第１合成音声を生成する過程と、
前記第１合成音声に基づいて第１誤差エネルギーを計算する過程と、
前記第１誤差エネルギーに基づいて誤差エネルギー勾配を計算する過程と、
前記誤差エネルギー勾配を使用して第２合成音声を生成する過程と、
システムの不安定性さに関して前記第２の合成音声をテストする過程と、
前記テストにより不安定であるとされた場合、前記第１の合成音声を選択する過程と
を有することを特徴とする方法。
【請求項９】
請求項８に記載の方法において、

前記誤差エネルギー勾配は、システム差分方程式から直接計算される
ことを特徴とする方法。
【請求項１０】
請求項９に記載の方法において、
前記第２合成音声から最小第２誤差エネルギーを繰り返し検索することにより、前記第２合成音声が生成される
ことを特徴とする方法。
【請求項１１】
請求項１０に記載の方法において、
前記第２合成音声は、前記誤差エネルギー勾配に適用される適応ステップサイズを使用して計算され、
前記適応ステップサイズは、前記誤差エネルギー勾配と前記第１合成音声の関数である
ことを特徴とする方法。
【請求項１２】
第１の線形予測係数を計算する過程と、
システム差分方程式から直接誤差エネルギー勾配を計算する過程と、
前記誤差エネルギー勾配から第２の線形予測係数を計算する過程と、
前記第１の線形予測係数の前記誤差エネルギーと、前記第２の線形予測係数の前記誤差エネルギーとを比較する過程と、
前記第１の線形予測係数および前記第２の線形予測係数のうち誤差エネルギーの少ないほうを選択する過程と、
システム不安定性さに関して前記第２の線形予測係数をテストし、前記テストにより不安定であるとされた場合、前記第１の線形予測係数を選択する過程と
を有することを特徴とする音声最適化方法。
【請求項１３】
請求項１２に記載の音声最適化方法において、
前記第２の線形予測係数を適応的に計算する過程を更に有する
ことを特徴とする音声最適化方法。
【請求項１４】
請求項１３に記載の音声最適化方法において、
前記誤差エネルギー勾配に聴覚重み付けを行う過程を更に有する
ことを特徴とする音声最適化方法。
【請求項１５】
原音サンプルを使用して音声合成多項式の第１の線形予測係数を計算する手段と、
システム差分方程式から直接誤差エネルギー勾配を計算する手段と、
前記誤差エネルギー勾配から第２の線形予測係数を計算する手段と、
前記第１の線形予測係数の前記誤差エネルギーと、前記第２の線形予測係数の前記誤差エネルギーとを比較し、前記誤差エネルギーの少ない方を選択する手段と、
システム不安定性さに関して前記第２の線形予測係数をテストし、前記テストにより不安定であるとされた場合、前記第１の線形予測係数を選択する手段と
を有することを特徴とする音声最適化装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は音声符号化に関し、特に線形予測係数の最適化に関する。
【０００２】
【従来の技術】
音声符号化（あるいは圧縮）と言う技術は、データ送信のために音声をデジタルデータに符号化する広く知られた技術である。符号化されたデータを受信した受信者側は、その音声を再現する。デジタル化された音声データは符号化後、後に音声に復号されるまでの間

、様々なデジタル記録メディアに保存することが出来る。
【０００３】
音声符号化システムは、他のアナログ符号化システムおよびデジタル符号化システムとは異なっている。アナログ符号化システムおよびデジタル符号化システムでは、音声を高いビットレートでダイレクトサンプリングして、サンプリングされた生データを受信機に送信する。ダイレクトサンプリングシステムは、原音を高品質に再現し、再現音質が重要な場合に好まれる。ダイレクトサンプリングシステムが使われている一般的な例としては、アナログの音楽レコードやカセットテープ、またデジタルの音楽ＣＤやＤＶＤがある。ダイレクトサンプリングシステムの欠点は、データを送信するのに広い帯域幅が必要であり、またデータの保存にも大きな記憶容量が必要なことである。これゆえ、原音からサンプリングされた生の音声データを送信する典型的な符号化システムでは、毎秒１２８０００ビットものデータレートが必要となることがある。
【０００４】
これに対して、音声符号化システムは、人間の音声発生の数学的モデルを使っている。発声モデル化の基本的技術は当技術分野で知られており、アメリカ音響協会の機関紙、１９７１年第５０巻で、Ｂ・Ｓ・アタル（B・S・Atal）とスザンヌ・Ｌ・ハナー（Suzanne・L・Hanauer）による「音声分析と音声波の線形予測による合成（Speech Analysis and Synthesis by Linear Prediction of the Speech Wave）」に説明されている。音声符号化システムで使われる人間の音声発生のモデルは、普通ソース・フィルター・モデルと呼ばれている。一般に、このモデルには、肺と声帯によって発生する空気の流れを表している励起信号と、声道（すなわち声門、口、舌、鼻腔と唇）を表している合成フィルタが含まれている。よって、肺と声帯が声道にパルス状の空気の流れを発生させるように、励起信号は合成フィルターへの入力信号として働く。そして、声道が肺と声帯からの空気の流れに変更を加えるように、合成フィルターが励起信号に変更を加える。この結果、出来上がった合成音声は、ほぼ原音を表すようになる。
【０００５】
音声符号化システムの長所は、ダイレクトサンプリングシステムと比べて、原音をデジタル化した形で送信するのに必要な帯域幅が、ずっと小さくなり得ることである。比べてみると、ダイレクトサンプリングシステムでは原音を表す生の音響データが送られているのに対し、音声符号化システムでは、数学的な音声モデルを再現するのに必要なわずかな量の制御データが送られているだけである。結果として、典型的な音声符号化システムを使えば、音声を送るのに必要な帯域幅をほぼ毎秒２４００～８０００ビットまで減らすことができる。
【０００６】
【非特許文献１】
アメリカ音響協会、機関誌、１９７１年第５０巻、Ｂ・Ｓ・アタル（B・S・Atal）とスザンヌ・Ｌ・ハナー（Suzanne・L・Hanauer）による「音声分析と音声波の線形予測による合成（Speech Analysis and Synthesis by Linear Prediction of the Speech Wave）」
【０００７】
【発明が解決しようとする課題】
音声符号化システムの欠点の１つは、ダイレクトサンプリングシステムに比べて、再現された音声の質がかなり低いことがあると言うことである。多くの音声符号化システムでは、受信者が正確に元の音声の内容を認知するのに十分な質は提供されている。しかし、いくつかの音声符号化システムでは、再現された音声は聞きやすくない。つまり、受信者はもともと話された言葉を理解する事は出来るが、音声の質が低かったり、不快であったりする。従って、より正確な音声生成モデルを提供する音声符号化システムが望まれている。
【０００８】
音声符号化システムの質を改善する１つの方法として認知されているものが、ラシュキャリ（Lashkari）等によるアメリカ特許出願０９／８０００７１号に説明されている。簡単

に説明すると、この解決法は、原音サンプルと合成音声サンプルとの間の合成化誤差を最小化する方法である。この音声符号化システムで分かった難しい問題の１つは、合成化誤差がかなり非線形であるということである。このことで、この問題が数学的に難しくなっている。この問題を解くこの難しさは、合成フィルター多項式の解を、多項式の係数の代わりに使うことで克服されている。よって、合成フィルター多項式の解を探すための解最適化アルゴリズムが、そこで説明されている。

【０００９】

上述の解法および当業者に知られている他の最適化解法に関して解決されない問題は、原音をエンコードするのに必要なコンピュータの処理能力が大きいことである。当業者なら分かるであろうが、原音を符号化するのに使用される様々な計算式を計算するのに、音声符号化システムでは、ＣＰＵ（central processing unit）やＤＳＰ（digital signal processor）を使う必要がある。しばしば、携帯電話などの携帯ユニットで音声符号化が行われる場合、ＣＰＵやＤＳＰは内臓のバッテリから電気を得ている。よって、通常、音声符号化のために利用可能な計算能力は、ＣＰＵやＤＳＰまたはバッテリ容量によって制限をうけることとなる。この問題は、どの音声符号化システムにおいても共通したものであるが、最適化アルゴリズムを使用するシステムにおいて、特に重大である。概して、最適化アルゴリズムは、通常の符号化アルゴリズムに加えて、追加の数式計算を含めることで、より質の高い音声を提供することができる。しかしながら、非効率な最適化アルゴリズムでは、ＣＰＵやＤＳＰとして、値段が高く、計算能力があり、大きなものが必要になってしまう。非効率な最適化アルゴリズムでは、バッテリも余計に使用し、バッテリライフが短くなってしまう。よって、音声符号化システムでは、効率的な最適化アルゴリズムが望まれている。

【００１０】

【課題を解決するための手段】

本発明は、音声生成の数学的モデルを最適化するためのアルゴリズムを提供する。この最適化アルゴリズムは、畳み込みを使用せず、また係数を解領域に変換することなしに、合成フィルタ多項式の係数を計算する。このように、係数の最適化に関する計算効率が改善される。原音と合成音声との間の誤差の勾配を使用して、係数がシステム差分方程式から直接計算される。システムの不安定さをテストするために終結閾値が使用され、不安定であるとされた場合、最適化アルゴリズムは止められ、最後の係数が返される。

【００１１】

【発明の実施の形態】

図を参照して説明する。図１に、より正確に原音をモデル化するために合成誤差を最小化する音声符号化システムを示す。図１には音声の合成による分析（ＡｂＳ（analysis-by-synthesis））システムが示されている。このシステムは、普通ソース・フィルター・モデルと呼ばれている。当技術分野でよく知られているように、ソース・フィルター・モデルは、人間の音声発生を数学的にモデル化する。このモデルで普通使われる仮定では、音声を生成している人間の音声発生メカニズムは、短い期間またはフレームで（例えば２０～３０ｍｓの分析フレーム）で、変化しないものとしている。更にまたこのモデルは、隣り合う期間の間に人間の音声発生メカニズムは変化するものと仮定している。このシステムでモデル化された物理的メカニズムには、声帯、声門、口、舌、鼻腔そして唇によって起こされる気圧変化が含まれている。よって、音声復号化器は、そのモデルを再現し、各期間用に制御データを少数セット使用するだけで、原音を再生成する。このように、通常の音声送信システムと異なり、原音の生サンプルデータは符号化器から復号化器へは送られない。この結果、送信されたり、記憶されたりするするデジタル符号化されたデータ（つまり、バンド幅やビット数）は、典型的なダイレクトサンプリングシステムが必要とするのよりずっと少なくなる。

【００１２】

図１において、デジタル化された元の音声１０が励起モジュール１２に送られている。励起モジュール１２は、この原音の各サンプルｓ（ｎ）を解析して、励起関数ｕ（ｎ）を生

成する。励起関数ｕ（ｎ）は通常、連続したパルス信号であり、この連続したパルス信号は、声帯によって声道に突発的に放出される肺からの空気の流れを表している。原音サンプルｓ（ｎ）の性質によるが、励起関数ｕ（ｎ）は、有声音１３か１４か無声音１５かである。

【００１３】

音声符号化システムでの再生音質を改善する方法として、有声音の励起関数ｕ（ｎ）をより正確にする方法がある。今まで、励起関数ｕ（ｎ）は、決まったパルス間隔Ｐと大きさＧを持つパルス列１３であった。当業者に知られている様に、この大きさＧと間隔Ｐを隣り合う期間で変えるようにしてもよい。大きさＧと間隔Ｐが固定されている今までのものに比べて、励起パルス１４のパルスの大きさと間隔を変化させることで励起関数ｕ（ｎ）を最適化すると、よりよい音声合成がなされることが明らかにされている。この改良は、米国電気電子技術者協会（ＩＥＥＥ）の音響、音声、信号処理に関する国際会議（１９８２年、６１４頁～６１７頁）の、ビシュヌ・Ｓ・アタル（Bishnu・S・Atal）とジョエル・Ｒ・レムデ（Joel R. Remde）による、「低ビットレートにおける自然な音声を生成するためのＬＰＣ励起の新しいモデル（A New Model of LPC Excitation For Producing Natural-Sounding Speech At Low Bit Rates）」に説明されている。

【００１４】

この最適化技術では、原音ｓ（ｎ）を符号化するための計算量が増えるが、最近のコンピュータは励起関数ｕ（ｎ）の最適化に十分な計算能力があるので、重大な欠点ではない。この改良でのもっと重大な問題は、可変励起パルス１４のデータを送信するのに必要な帯域幅が余計に必要であると言うことだ。この問題を解決する方法として、米国電気電子技術者協会（ＩＥＥＥ）の音響、音声、信号処理に関する国際会議（１９８５年、９３７頁～９４０頁）の、マンフレッド・Ｒ・シュレッダー（Manfred R. Schroeder）とビシュヌ・Ｓ・アタル（Bishnu・S・Atal）による「符号励起線形予測化（CELP）：低ビットレートにおける高品質音声（Code-Excited Linear Prediction (CELP): High-Quality Speech At Very Low Bit Rates）」に説明されている符号化システムがある。

【００１５】

この解決法では、多くの最適化された関数を分類して、関数ライブラリすなわちコードブックを作成する。そして、符号化励起モジュール１２は、原音ｓ（ｎ）に最も近い合成音声を生成する最適化された励起関数をコードブックから選択する。そして、コードブック内の最適な項目を特定するコードが復号化器に送られる。復号化器は送られてきたコードを受信し、対応するコードブックにアクセスし、選択された最適な励起関数ｕ（ｎ）を再生成する。

【００１６】

励起モジュール１２は、無声音１５の励起関数ｕ(ｎ)も生成することが出来る。無声音１５の励起関数ｕ(ｎ)は、話者の声帯が開いて、突発的な空気の流れが声道に起こされた時に使われる。多くの励起モジュール１２は、この状態をモデル化するのに、パルスでなく白色ノイズ１５（すなわちランダム信号）を有する励起関数ｕ（ｎ）を生成する。

【００１７】

次に、合成フィルター１６は、声道のモデル化と、声帯からの空気の流れに対する、声道の効果のモデル化を行う。普通は、合成フィルター１６には、声道の様々な形を表す多項式を使う。多項式のパラメータつまり係数は通常、入力音声信号を使用して見積もりがなされる、また線形予測係数と呼ばれる。

【００１８】

上述のアタル（Atal）とレムデ（Remde）によると、合成フィルター１６は次の数式で表すことができる。

【数１】

$$H(z) = G/A(z)$$

ここで、Ｇは音声の大きさを表している利得項である（簡潔のため、以降の式では利得項Ｇは省略する）。Ａ（ｚ）はＭ次の多項式であり次の式で表される。

【数２】

$$A(z) = 1 + \sum_{i=1}^{M} a_i z^{-i}$$

【００１９】

多項式Ａ（ｚ）の次数は用途によって変わる。サンプリングレート８ｋＨｚの場合、１０次の多項式が通常使用される。合成フィルタ１６で決定される合成音声ｓｓ（ｎ）と励起関数ｕ（ｎ）との関係は次の式で定義される。

【数３】

$$ss(n) = u(n) - \sum_{i=1}^{M} a_i ss(n-i)$$

【００２０】

ここで、表記「ｓｓ」は、本願の優先権主張の基礎である米国特許出願においては「ｅ」の上に「∧」を載せた表記となっていたものであるが、本願明細書においてそのような表記を用いることが困難であったため、その代わりに採用されたものである。従って、これ以降における表記「ｓｓ」は、実際には「ｅ」の上に「∧」を載せたものを表していると解されることを望む次第である。

【００２１】

数３はシステム差分方程式とも呼ばれる。通常、この多項式の係数$a_1 \cdots a_M$は、この分野で線形予測符号化（ＬＰＣ）として知られる技術を使って計算される。ＬＰＣに基づく技術では、トータル予測誤差E_pを最小にすることにより、多項式の係数$a_1 \cdots a_M$を計算する。これにより、サンプル予測誤差e_p（ｎ）が次の式により定義される。

【数４】

$$e_p(n) = s(n) + \sum_{i=1}^{M} a_i s(n-i)$$

【００２２】

トータル予測誤差E_pは、次の式によって定義される。

【数５】

$$E_p = \sum_{i=0}^{N-1} e_p^{\,2}(i)$$

ここで、Ｎはサンプルの数で表される分析フレームの長さである。多項式の係数$a_1 \cdots a_M$は、よく知られた数学的方法を用いて、トータル予測誤差E_pを最小化することにより解くことが出来る。

【００２３】

多項式の係数$a_1 \cdots a_M$を計算するＬＰＣ技術に関する問題の１つは、トータル予測誤差だけが最小化されることである。このように、ＬＰＣ技術では、原音ｓ（ｎ）と合成音声ｓｓ（ｎ）との間の誤差を最小化していない。従って、サンプル合成誤差e_s（ｎ）は次の式で定義できる。

【数６】

$$e_s(n) = s(n) - ss(n)$$

【００２４】
トータル合成誤差Ｅ$_s$は、合成誤差エネルギーＪとも呼ばれるが、次の式で定義される。
【数７】

$$J = E_s = \sum_{n=0}^{N-1} e_s^2(n) = \sum_{n=0}^{N-1} (s(n) - ss(n))^2$$

Ｎはサンプルの数で表される分析フレームの長さである。上で述べたトータル予測誤差Ｅ$_p$のように、合成誤差エネルギーＪは、最適フィルターの係数ａ$_1$…ａ$_M$を計算するために最小化されなければならない。しかしながら、この技術の問題は、数３で表される合成音声ｓｓ（ｎ）によって、合成誤差エネルギーＪがかなり非線形な関数になり、数学的に扱うのが難しいことである。
【００２５】
最適化アルゴリズムの中には、この難しさを、係数ａ$_1$…ａ$_M$の代わりに、多項式Ａ（ｚ）の解を使用することで避けているものもある。合成フィルタ１６の安定性に関して制御することができるが、この解決法では、多くの計算能力を要求する。更に、解領域最適化は、畳み込みに関連したインパルス応答に基づいた勾配を通常計算する。畳み込みでは、先行する音声サンプルの履歴を考慮することなしに、システムのゼロ状態応答がわかるだけである。
【００２６】
他の最適化アルゴリズムと比較して、フィルタ係数ａ$_1$…ａ$_M$を解領域に変換することなしに、フィルタ係数ａ$_1$…ａ$_M$を最適化するため、合成誤差エネルギーＪの勾配が直接計算されるようにできる。よって、数３、数６、数７を使用して、合成誤差エネルギーＪの勾配は次式で与えられる。
【数８】

$$\frac{\partial J}{\partial a_i} = \sum_{n=0}^{N-1} \frac{\partial}{\partial a_i}(s[n] - ss[n])^2 = \sum_{n=0}^{N-1} 2(s[n] - ss[n])\left(-\frac{\partial ss[n]}{\partial a_i}\right)$$

ここでｉ＝１～Ｍである。
数３を使用して、係数ａ$_1$…ａ$_M$に関する合成音声の勾配は、次式で表される。
【数９】

$$\frac{\partial ss[n]}{\partial a_i} = -\sum_{j=1}^{M} a_j \frac{\partial ss[n-j]}{\partial a_i} - ss[n-i]$$

ここで、ｎ＝０～Ｎ－１、かつｉ＝１～Ｍである。
係数ａ$_1$…ａ$_M$は、長さＮのフレームだけで有効であると仮定される。このように、係数は、ｎ＝０～Ｎ－１のみで存在し、この間隔を外れたところでは、音声は、係数ａ$_1$…ａ$_M$から独立している。
【００２７】
この最適化アルゴリズムの利点の１つは、数９に示される合成音声の偏導関数が、回帰的な方法により効率よく計算できることである。従って、数９の偏導関数は、以下の２次元配列で示される。
【数１０】

$$D[n,i] = \frac{\partial ss[n]}{\partial a_i}$$

ここで、ｎ＝０～Ｎ－１、およびｉ＝１～Ｍである。

配列D[ｎ，ｉ]は、以下の繰り返しコードＡを使用して計算される。

【００２８】

＜繰り返しコードＡ＞

```
For n=1 to N-1
D[n, i] = -ss[n-i]
For j = 1 to M
If n-j = 0
D[n, i] = -a_j * D[n-j, i]
Next j
Next n
```

【００２９】

繰り返しコードＡを使用して計算された偏導関数を数８に代入すると、合成誤差エネルギーの勾配を得ることができる。

合成誤差エネルギーの勾配ベクトルは、次式で得ることができる。

【数１１】

$$\nabla J = \left[\frac{\partial J}{\partial a_1} \quad \frac{\partial J}{\partial a_2} \quad \cdots \quad \frac{\partial J}{\partial a_M} \right]^T$$

係数$a_1 \cdots a_M$のベクトルは、次式でも定義される。

【数１２】

$$a = [a_1 \quad a_2 \cdots\cdots a_M]^T$$

最適化係数は、次式を使用して計算できる。

【数１３】

$$a_{new} = a - \mu \nabla J$$

ここで、μは正の値を持ちステップサイズとして知られる。よって、係数の新たなベクトルは、勾配に対して負の方向に移動することによって計算される。ステップサイズμの大きさは、最適化プロセスのスピードと安定性とを変えるために、増やされたり減らされたりする。

【００３０】

最適化アルゴリズムで使用されるフローチャートを図２と図３に示す。このフローチャートは、図１の合成フィルタ１６および合成フィルタ最適化部１８によって実行されるものである。図２に示すように、最適化アルゴリズムへ入力されるのは、原音ｓ、励起関数ｕ、線形予測係数$a_1 \cdots a_M$、そして合成フィルタのメモリである（ステップ２２）。CalculateErrorEnergy関数を呼び出して、数７を使用して、合成誤差エネルギーＪが計算される（ステップ２４）。変数OldErrorEnergyが、次に初期化されて、合成誤差エネルギーＪが代入される（ステップ２４）。次に、CalculateGradient関数を呼び出して、合成誤差エネルギーの勾配が、数８もしくは繰り返しコードＡを使用して計算される。

【００３１】

図３に示すGradientDescent関数を呼び出すことで、変数ErrorEnergyが、計算される（ステップ２８）。変数ErrorEnergyは、更新された係数$a_1 \cdots a_M$を使用した合成誤差エネルギーＪを表している。それに対し、変数OldErrorEnergyは、以前の係数$a_1 \ldots a_M$を使った場合の合成誤差エネルギーを表している。次に、ErrorEnergyは、OldErrorEnergyと比較される。これは、合成誤差エネルギーＪの変化が、終了閾値Termination_Threshold以下であるかどうかを決定するものである（ステップ３０）。合成誤差エネルギーＪの変化が、終了閾値Termination_Threshold以下でない場合は、OldErrorEnergyにErrorEnergy

が代入され、合成誤差エネルギーＪの新たな勾配、新たな係数$a_1 \cdots a_M$、そして新たな合成誤差エネルギーＪが、ステップ３２、２６、２８で計算される。合成誤差エネルギーＪの変化が、終了閾値Termination_Threshold以下である場合、最適化アルゴリズムは終了し、係数の修正された線形予測値が返される（ステップ３４）。

【００３２】

図３にGradientDescent関数を示す（ステップ３６）。GradientDescent関数は、CalculateErrorEnergyを呼び出し、数７を使用して、合成誤差エネルギーＪを計算することで開始する（ステップ３８）。変数OldErrorEnergyに合成誤差エネルギーＪが代入される（ステップ３８）。次に、合成誤差エネルギーＪの勾配ベクトルの大きさが計算され、変数Ｇノルムが割り当てられる。また、係数$a_1 \ldots a_M$の大きさが計算され、変数Ａノルムが割り当てられる（ステップ４０）。変数Ｇノルム、Ａノルム、および変数ステップサイズに割り当てられた所定の値を使用して、適応ステップサイズμが計算される（ステップ４２）。新たな線形予測係数a_{new}が、数１３を使用して計算される（ステップ４４）。次に合成誤差エネルギーＪが、CalculateErrorEnergyを呼び出すことで、数７を使用して計算される（ステップ４６）。次に変数ErrorEnergyに、合成誤差エネルギーＪが代入される（ステップ４６）。次に、合成誤差エネルギーＪに増減があるかを調べるために、OldErrorEnergyがErrorEnergyと比較される（ステップ４８）。合成誤差エネルギーＪが減少している場合、変数OldErrorEnergyに変数ErrorEnergyが代入され、線形予測係数$a_1 \ldots a_M$が新たな線形予測係数a_{new}へと更新される（ステップ５０）。それから、新たな線形予測係数a_{new}と合成誤差エネルギーＪが計算される（ステップ４４、４６）。一方、合成誤差エネルギーＪが増大していた場合、GradientDescent関数は終了して、現在の線形予測係数$a_1 \cdots a_M$とOldErrorEnergyが返される（ステップ５２）。

【００３３】

合成モデルが決定され、係数$a_1 \cdots a_M$が最適化されたら、モデル化のための制御データが、送信又は保存のために量子化されてデジタルデータへとなる。量子化には業界で標準となっている方法が多くある。ある例では、量子化された制御データは、１０つの合成フィルタ係数$a_1 \cdots a_M$、励起パルスの大きさを表す利得値Ｇを１つ、励起パルスの周波数のためのピッチ間隔を１つ、有声１３もしくは無声１５の励起関数ｕ（ｎ）を示す指示子を１つ、含んでいる。よって、この例では、各音声フレームの最後で、１３の異なる変数を送信する必要がある。しかしながら、他の制御データを送るようにしてもよい。例えば、ＣＥＬＰエンコーダにおいては、使用する最適化励起関数ｕ（ｎ）を特定するコードブックインデックスも送信される。通常ＣＥＬＰエンコーダでは、制御データは、合計８０ビットへと量子化される。この例では、最適化を算入した合成音声ｓｓ（ｎ）を、毎秒８０００ビット（８０ビット／フレーム÷０．０１秒／フレーム）のバンド幅で送ることができる。

【００３４】

エンコードに使用できるコンピュータの能力によるが、合成化モデルをより正確にするために、追加のエンコードシーケンスも可能である。これらのシーケンス例を、図１に点線で示してある。例えば、励起関数ｕ（ｎ）は、合成モデルにおけるエンコードの間、様々な段階で再最適化を行うことができる。更に、数式とアルゴリズムは、特定のアプリケーションのために、変更することもできる。

【００３５】

合成音声の主観的品質を更に改善する方法として、聴覚重み付けを使用する方法がある。この場合、合成誤差エネルギーＪは、次式によって、聴覚重み付けも使用して定義される。

【数１４】

$$J = \sum_{n=0}^{N-1} \{e[n] * h_w[n]\}^2 = \sum_{n=0}^{N-1} \{(s[n] - ss[n]) * h_w[n]\}^2$$

ここで、$h_w[n]$は、聴覚重み付けフィルタのインパルス応答である。
数１４に畳み込み処理を行うことで、合成誤差エネルギーは次式になる。
【数１５】

$$J = \sum_{n=0}^{N-1}\left\{\sum_{k=0}^{n} h_w[k]s[n-k] - \sum_{l=0}^{n} h_w[l]ss[n-l]\right\}^2$$
$$= \sum_{n=0}^{N-1}\left\{\left(\sum_{k=0}^{n} h_w[k]s[n-k]\right)^2 - 2\left(\sum_{k=0}^{n} h_w[k]s[n-k]\right)\left(\sum_{l=0}^{n} h_w[l]ss[n-l]\right) + \left(\sum_{l=0}^{n} h_w[l]ss[n-l]\right)^2\right\}$$

【００３６】
次に、数１５を合成フィルタ係数$a_1 \cdots a_M$に関して微分することにより、合成誤差エネルギーの偏導関数は、以下の式になる。
【数１６】

$$\frac{\partial J}{\partial a_i} = \sum_{n=0}^{N-1}\left\{-2\left(\sum_{k=0}^{n} h_w[k]\left(s[n-k] - ss[n-k]\right)\right)\left(\sum_{l=0}^{n} h_w[l]\frac{\partial}{\partial a_i}ss[n-l]\right)\right\}$$

聴覚重み付け誤差も、次式で定義することができる。
【数１７】

$$e_w[n] = e[n] * h_w[n]$$

数１０、数１４、数１６、数１７、を使用して、合成誤差エネルギーの偏導関数は、以下の式のようになる。
【数１８】

$$\frac{\partial J}{\partial a_i} = \sum_{n=0}^{N-1}\left\{-2e_w[n]\left(D[n,i] * h_w[n]\right)\right\}$$

したがって、最適化アルゴリズムは、数１４を数７の代わりに使い、更に数１８を数８の代わりに使って、更に改善される。
【００３７】
当業者に明らかなように、この最適化アルゴリズムは、合成フィルタ多項式Ａ（ｚ）を最適化するのに必要な計算を著しく減らす。よって、エンコーダの効率が著しく改善される。あるいはこの効率化を、合成音声ｓｓ（ｎ）の質を改善するのに使用することもできる。従来の最適化アルゴリズムを使用する場合、各サンプルを合成音声にするのに必要な計算が多かった。しかし、改善された最適化アルゴリズムは、解領域を使用することなしに、合成誤差エネルギーＪの勾配をシステム差分方程式から直接計算することで、合成音声ｓｓ（ｎ）を計算するのに必要な計算量を減らしている。本発明は、様々な音声エンコーダに適用することができるが、ラシュキャリ等による米国特許出願０９／８０００７１号に記載の音声エンコーダに適用した場合は、すでに性能の改善が認められている。
【００３８】
この最適化アルゴリズムは多くの利点を持っている。例えば、解領域解決法が通常使う畳み込みを、この最適化アルゴリズムは使用しないので、ゼロ状態応答とゼロ入力応答とを含む、システムのトータル応答も考慮されることになる。この最適化アルゴリズムはまた、各繰り返しにおいて、合成誤差エネルギーを所定の終了閾値でテストすることで、不安定さの問題を避けている。よって、テスト結果が、システムが不安定であるとなった場合、最適化アルゴリズムは終了して、最後に最適化された線形予測係数が使われる。適応ス

テップサイズも、最適化のスピードを改善するために使用される。更に、最適化アルゴリズムの計算効率を改善する反復アルゴリズムを使用して、合成誤差エネルギーの勾配を計算することもできる。この最適化アルゴリズムの他の利点は、合成音声の質を更に改善するために、聴覚重み付けを使うことができることである。

【００３９】

図４は、最適化アルゴリズムによる結果例を示す図である。図４は、原音サンプル、Ｇ．７２９エンコーダを使用した場合の合成音声、勾配降下最適化アルゴリズムを使用したＧ．７２９エンコーダを使用した場合の合成音声、それぞれの音声波形を示している。当業者は理解するであろうが、このＧ．７２９エンコーダは、様々な音声エンコーダの質を比較するために、研究者などに使用される標準化された音声エンコーダである。図から分かるように、勾配降下最適化アルゴリズムを使った場合の合成音声が、Ｇ．７２９エンコーダだけで生成された合成音声よりも、原音に一致している。

【００４０】

他のテスト結果では、最適化アルゴリズムにより良好になった質および効率が示されている。例えば、あるテストにおいて、最適化アルゴリズムの付属したものと付属していないＧ．７２９を使用して、男性と女性の標準化された音声データを符号化した。１０ミリ秒毎に部分ＳＮ比（ＳＳＮＲ：segmental signal to noise ratio measurements）を計測すると、Ｇ．７２９エンコーダのみの場合、７．０６ｄＢ　ＳＳＮＲであり、解領域最適化アルゴリズム付属のＧ．７２９エンコーダを使用した場合、７．３３ｄＢ　ＳＳＮＲとなり、説明を行ってきた勾降下最適化アルゴリズム付属のＧ．７２９エンコーダでは、７．３４ｄＢ　ＳＳＮＲとなった。当業者には当然であるが、ＳＳＮＲ計測値が高いということは通常、合成音声が良好な聴覚品質を有しているということである。更に、勾配降下最適化アルゴリズムの計算量は、解領域最適化アルゴリズムに比べて、およそ２０分の１～３０分の１である。このテストは聴覚重み付けを使用せずに行った。聴覚重み付けを行うと、勾配降下最適化アルゴリズムでは、部分信号対重み付けノイズ比（ＳＳＷＮＲ：segmental signal to weighted noise ratio）の計測値が１４．１９ｄＢ　ＳＳＷＮＲとなるが、普通のＧ．７２９エンコーダでは、１４．０４ｄＢ　ＳＳＷＮＲである。

【００４１】

本発明の好ましい実施形態をここに説明したが、本発明はこれに限定されず、本発明の趣旨から外れずに変形することが可能である。本発明の範囲は、特許請求の範囲によって決まるものであり、文言上でも均等上でも特許請求の範囲内にある装置および方法は、本発明に含まれるものである。

【００４２】

【発明の効果】

以上説明したように、本発明によれば、より正確な合成音声を提供する音声符号化システムが提供される。

【図面の簡単な説明】

【図１】　音声の合成による分析システムのブロック図である。

【図２】　フィルタ係数のための最適化アルゴリズムのフローチャートである。

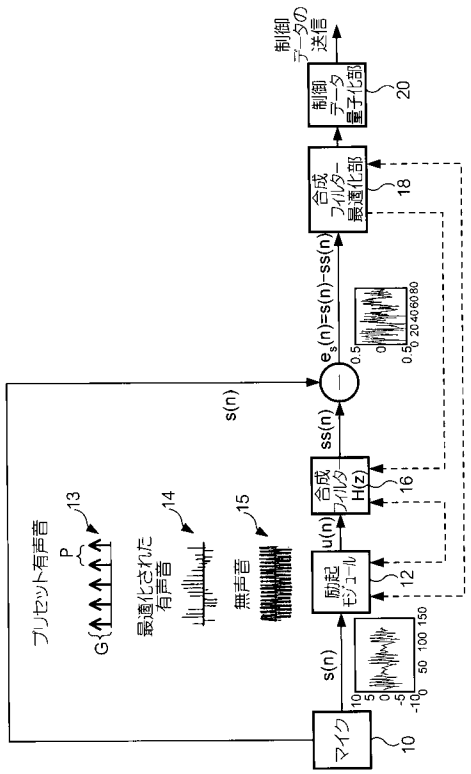
【図３】　誤差勾配を見つけるための最適化アルゴリズムで使用される、勾配降下関数のフローチャートである。

【図４】　原音の波形と、Ｇ．７２９エンコーダを使用した場合の合成音声の波形と、勾配降下最適化を行うＧ．７２９エンコーダを使用した場合の合成音声の波形とを比較している図である。

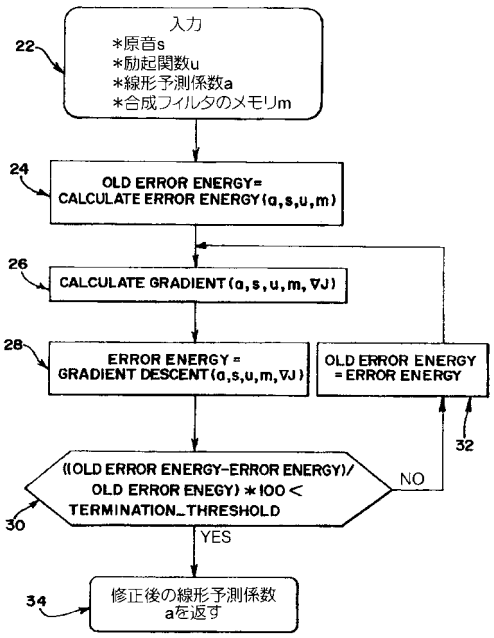
【符号の説明】

１０……マイク、１２……励起モジュール、１６……合成フィルタ、１８……合成フィルタ最適化部、２０……制御データ量子化部

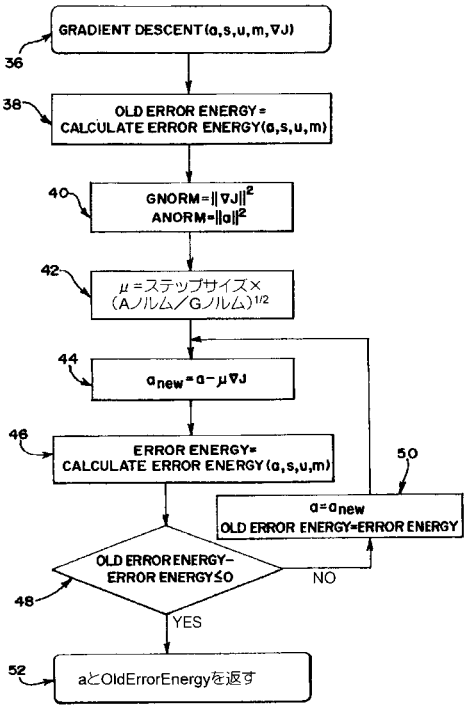
【図１】


制御データの送信
制御データ量子化部
20
合成フィルター最適化部
18
$e_s(n)=s(n)-ss(n)$
合成フィルタ H(z)
16
ss(n)
s(n)
u(n)
励起モジュール
12
s(n)
マイク
10
プリセット有声音
P
G
13
最適化された有声音
14
無声音
15

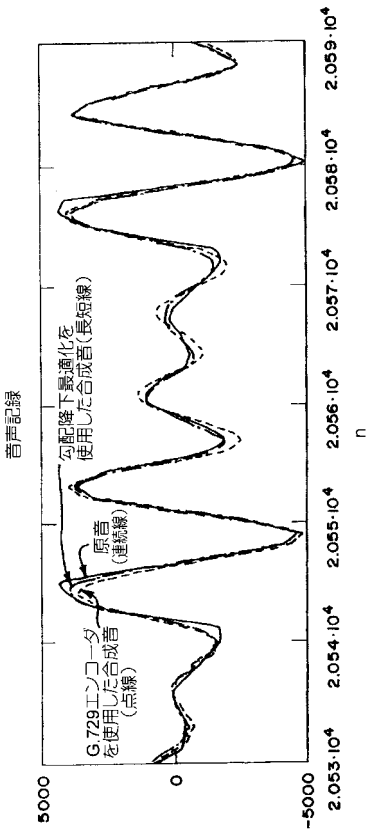
【図２】


入力
*原音s
*励起関数u
*線形予測係数a
*合成フィルタのメモリm
22
OLD ERROR ENERGY=
CALCULATE ERROR ENERGY(a,s,u,m)
24
CALCULATE GRADIENT(a,s,u,m,∇J)
26
ERROR ENERGY =
GRADIENT DESCENT(a,s,u,m,∇J)
28
OLD ERROR ENERGY
= ERROR ENERGY
32
((OLD ERROR ENERGY-ERROR ENEGY)/
OLD ERROR ENEGY)＊100＜
TERMINATION_THRESHOLD
30
NO
YES
修正後の線形予測係数
aを返す
34

【図３】


GRADIENT DESCENT(a,s,u,m,∇J)
36
OLD ERROR ENERGY=
CALCULATE ERROR ENERGY(a,s,u,m)
38
GNORM=‖∇J‖²
ANORM=‖a‖²
40
μ=ステップサイズ×
(Aノルム／Gノルム)^1/2
42
$a_{new}=a-\mu\nabla J$
44
ERROR ENERGY=
CALCULATE ERROR ENERGY(a,s,u,m)
46
a=anew
OLD ERROR ENERGY=ERROR ENERGY
50
OLD ERROR ENERGY-
ERROR ENERGY≦0
48
NO
YES
aとOldErrorEnergyを返す
52

【図４】


音声記録
勾配降下最適化を
使用した合成音(長短線)
原音
(連続線)
G.729エンコーダ
を使用した合成音
(点線)
5000
0
-5000
2.059·10⁴
2.058·10⁴
2.057·10⁴
2.056·10⁴
2.055·10⁴
2.054·10⁴
2.053·10⁴
n

フロントページの続き

(56)参考文献 特開昭６２－１１１２９９（ＪＰ，Ａ）
特開平０２－２７２５００（ＪＰ，Ａ）
特開２００２－０７３０９７（ＪＰ，Ａ）
特開２００２－３２８６９２（ＪＰ，Ａ）
小池伸一，”絶対値誤差の勾配に基づく適応制御ステップサイズサインアルゴリズムの解析”，1997年電子情報通信学会総合大会講演論文集，A-4-46（1997-03），p.177
小池伸一，”勾配法を用いた適応制御ステップサイズサイン－サインアルゴリズムの解析”，1997年電子情報通信学会基礎・境界ソサイエティ大会講演論文集，A-4-14（1997-08），p.84

(58)調査した分野(Int.Cl.，ＤＢ名)
G10L 13/00-13/08,19/00-19/14
CiNii