(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2007-114355
(P2007-114355A)
(43)公開日　平成19年5月10日(2007.5.10)

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

審査請求 未請求　請求項の数 22　ＯＬ　（全 14 頁）

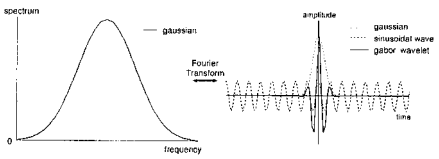
(21)出願番号　特願2005-304082 (P2005-304082)
(22)出願日　平成17年10月19日 (2005.10.19)

(71)出願人　504137912
国立大学法人　東京大学
東京都文京区本郷七丁目３番１号
(74)代理人　100103137
弁理士　稲葉　滋
(72)発明者　嵯峨山　茂樹
東京都文京区本郷七丁目３番１号　国立大学法人東京大学内
(72)発明者　槻　武也
東京都文京区本郷七丁目３番１号　国立大学法人東京大学内
(72)発明者　酒向　慎司
東京都文京区本郷七丁目３番１号　国立大学法人東京大学内

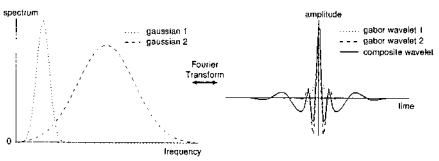
最終頁に続く

(54)【発明の名称】音声合成方法及び装置

(57)【要約】
【課題】
高品質の合成音声を提供すると共に、加工性に優れた音声合成手法を提供する。
【解決手段】
音声のスペクトル包絡を混合ガウス分布関数で近似することで少数のパラメータによって音声スペクトルを表現して分析パラメータを得る。そして、この混合ガウス分布関数の逆フーリエ変換であるGabor関数の重ね合わせを基本波形とし、それをピッチ周期ごとに配置して有声音を合成する。ピッチ周期をランダムにすれば無声音も合成できる。
【選択図】
図５



(a) 単一ガウス関数 (Single gaussian) のフーリエ変換



(b) 混合ガウス関数 (Gaussian mixture) のフーリエ変換

【特許請求の範囲】
【請求項１】
フレーム毎の音声のスペクトル包絡を、所定数の単峰性関数の混合分布で近似する音声分析により得られた音声スペクトル特徴量に基づいて、時間領域において前記単峰性関数に対応する関数を、所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置することを特徴とする音声合成方法。
【請求項２】
前記音声スペクトル特徴量は、単峰性関数の混合分布のモデルパラメータである、請求項１に記載の音声合成方法。
【請求項３】
前記モデルパラメータは、各単峰性関数の平均、分散、重みを含む、請求項２に記載の音声合成方法。
【請求項４】
前記モデルパラメータは、ＥＭアルゴリズムを用いて取得される、請求項３に記載の音声合成方法。
【請求項５】
前記基本波形は、前記混合分布を逆フーリエ変換したものに相当する、請求項１乃至４いずれかに記載の音声合成方法。
【請求項６】
前記混合分布は、所定数のガウス分布関数からなる混合ガウス分布であり、前記基本波形は、所定数のガボール関数を重畳してなる複合ガボール関数である、請求項１乃至５いずれかに記載の音声合成方法。
【請求項７】
前記音声合成方法は、フレーム毎の音声のスペクトル包絡を、所定数の単峰性関数の混合分布で近似して音声スペクトル特徴量を求める音声分析ステップを含む、請求項１乃至６いずれかに記載の音声合成方法。
【請求項８】
前記スペクトル包絡は、ラグ窓を用いた音声スペクトルの平滑化により取得される、請求項７に記載の音声合成方法。
【請求項９】
前記音声合成方法は、ピッチ抽出を含む、請求項１乃至８いずれかに記載の音声合成方法。
【請求項１０】
前記音声合成方法は、有声音/無声音の判定を含む、請求項１乃至９いずれかに記載の音声合成方法。
【請求項１１】
前記音声が有声音であり、前記駆動時点は、ピッチ周期ごとに設定される、請求項１乃至１０いずれかに記載の音声合成方法。
【請求項１２】
前記音声が有声音であり、前記駆動時点を設定するにあたり、複合波形の各成分ごとに重畳時点をずらしてピッチ周期で重畳する、請求項１乃至１１いずれかに記載の音声合成方法。
【請求項１３】
前記音声が有声音であり、前記駆動時点は、ピッチ周期内に複数ある、請求項１乃至１２いずれかに記載の音声合成方法。
【請求項１４】
前記音声が無声音であり、前記駆動時点は、ランダム間隔に設定される、請求項１乃至１３いずれかに記載の音声合成方法。
【請求項１５】
フレーム毎の音声のスペクトル包絡を、所定数の単峰性関数の混合分布で近似する音声

分析により得られた音声スペクトル特徴量に基づいて、時間領域において前記単峰性関数に対応する関数を、所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置することを特徴とする音声合成装置。
【請求項１６】
前記音声合成装置は、
フレーム毎の音声のスペクトル包絡を、所定数の単峰性関数の混合分布で近似して音声スペクトル特徴量を取得する音声分析部と、
時間領域において前記単峰性関数に対応する関数を、所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置する音声合成部と、
を有する、請求項１５記載の音声合成装置。
【請求項１７】
前記音声合成装置は、
フレーム毎の音声のスペクトル包絡を、所定数の単峰性関数の混合分布で近似する音声分析により得られた音声スペクトル特徴量を記憶する記憶部と、
時間領域において前記単峰性関数に対応する関数を、所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置する音声合成部と、
を有する、請求項１５、１６いずれかに記載の音声合成装置。
【請求項１８】
前記混合分布は、所定数のガウス分布関数からなる混合ガウス分布であり、前記基本波形は、所定数のガボール関数を重畳してなる複合ガボール関数である、請求項１５乃至１７いずれかに記載の音声合成装置。
【請求項１９】
フレーム毎の音声のスペクトル包絡を、所定数の単峰性関数の混合分布で近似する音声分析により得られた音声スペクトル特徴量に基づいて音声合成を行うためにコンピュータを、
時間領域において前記単峰性関数に対応する関数を、所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置する手段として機能させるための音声合成用コンピュータプログラム。
【請求項２０】
音声スペクトル特徴量に基づいて音声合成するためにコンピュータを、
フレーム毎の音声のスペクトル包絡を、所定数の単峰性関数の混合分布で近似する音声分析により得られた音声スペクトル特徴量を記憶する記憶手段と、
時間領域において前記単峰性関数に対応する関数を、所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置する手段と、
して機能させるための音声合成用コンピュータプログラム。
【請求項２１】
音声スペクトル特徴量に基づいて音声合成するためにコンピュータを、
フレーム毎の音声のスペクトル包絡を、所定数の単峰性関数の混合分布で近似する音声分析により音声スペクトル特徴量を取得する手段と、
時間領域において前記単峰性関数に対応する関数を、所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置する手段と、
して機能させるための音声合成用コンピュータプログラム。
【請求項２２】
請求項１９乃至２１いずれかに記載の音声合成用コンピュータを記録したコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、音声合成技術に関するものである。
【背景技術】

【０００２】
コンピュータにおける音声情報処理が進展するに伴い、音声合成においては、テキストをただ読み上げるだけにとどまらず、対話調の合成音声など様々な要求に適用可能な、高品質かつ多様なスタイルの音声を生成できる音声合成が待望されている。
【０００３】
PSOLA方式や波形接続型（非特許文献１）の音声合成手法は、十分なバラエティの音声素片のデータがあれば高品質な合成音声が期待できるが、データベースに含まれない条件の音声を合成したり、話者適応するような音声の特徴を操作する加工性は高くない。合成したい音声のスタイルに応じた音声データを補うことによって対処するとしても、様々な発話スタイルに対応したデータを収集することは困難が予想される。このため、感情音声や対話音声などを生成するには効率が悪いと考えられる。
【０００４】
これに対して、パラメトリックな音声合成手法の代表例であるフィルタ型の音声合成では、スペクトル包絡と微細構造を(近似的に)分離して扱う。そのため、F_0は任意に変化させられ、フィルタ特性を比較的少数のパラメータで制御して音声スペクトルを生成するため、加工性が高いと期待されている。フィルタ特性を与えるパラメータとしてLPC（非特許文献２）、PARCOR（非特許文献３）、LSP（非特許文献４）やケプストラム（非特許文献５）などが提案されており、それぞれ比較的品質が高い音声分析合成方式が確立されている。しかし、これらの方法ではフィルタパラメータと音声の声質やスタイルの間の関係が一意には定まらないため、音声の性質を自在に制御することは容易ではない。
【０００５】
フィルタ型の音声合成は、音声分析合成系として使われる場合はかなり高い品質を示す。しかし、分析時とは異なるF_0で駆動した場合など、一般に波形接続型音声合成に比べ音声品質が低い。その一因として、次に述べるフィルタの利得特性と時間特性に注目することができる。
【０００６】
全極型フィルタによる音声分析合成方式(LPC系)における有声音の分析合成について考察する。一般に音声スペクトル包絡の山と谷の間には数十dBに達する大きなレベル差(スペクトルダイナミックレンジ)があることが多く、これを少数のパラメータを用いたモデルで表現するために、十数次のような比較的次数が低い全極型フィルタを用いる。全極型フィルタは多重共振系であるが、このような理由によっておのおのの極の共振特性のQ値は、実際の声道の特性よりも大きな値をとる傾向がある。
【０００７】
このような周波数特性のフィルタの時間特性は、共振周波数の信号成分に対してQ値にほぼ比例した利得が生じるとともに、Q値にほぼ比例した時定数で出力振幅が立ち上り、減衰する。アクセント(ピッチ)を制御して音声を合成するような場合を考えると、分析時と異なるF_0で全極型フィルタを駆動し、たまたま駆動音源信号の倍音成分が高Q値の共振周波数に一致した場合などには、出力振幅の立ち上りにも減衰(立ち下がり)にも時間がかかり、その結果として合成音声の時間制御特性が悪くなる。そして、このような音が後続の音声に重畳することで、エコーが掛かっているような印象の「歯切れの悪い」音になる一因となっている可能性がある。
【０００８】
図1は、ある音声データにおいて、音素/o/に該当する区間をLPC分析して得た全極型フィルタに、１フレーム分の長さ(30msec)のインパルス列(有声音駆動に相当)を入力したときの出力波形である。入力に対して出力振幅は増大を続ける(定常状態に達するまでに時間が掛かる)とともに、入力が終了した後も数十msecにわたり出力が持続している。また、フィルタでは出力信号の利得がQ値に比例するため、その利得は駆動音原信号のピッチ周波数によって大きく変動する。このような現象のため、フィルタ型音声合成では合成音声のパワーを制御しにくい。
【０００９】

これらの問題は決して特殊な状況ではなく、LPC系においてはしばしば起こりうる。実験的にそれを示すために、ある程度長い(1分程度)音声を用意し、LPC系で分析合成を行った。まず、時間制御特性を調べるための実験を行った。ピッチ周期を0.8倍から1.2倍まで0.02刻みで変更し、分析したフィルタに30msec間入力した。その後入力をせずに合成を続け、各フレーム、ピッチ周期で減衰時間を調べた。ただし、減衰時間は入力停止から合成音声のパワーが30dB低下するまでの時間と定義する。また、速い変化に追従するためパワーを10msec間の振幅の二乗和として定義した。図1においては、55msecが減衰時間である。そして、図2に減衰時間を5ms単位のヒストグラムで示した。分布が右に偏るほど、減衰時間が長くなりやすいと言える。さらに、利得特性を調べるためにピッチ周波数を同様に変化させて音声全体の合成を行い、有声区間の各フレームのパワーを調べた。同一のフレームで、駆動音源のピッチ周波数を変えることでパワーが変化するが、その最大になる場合と最小になる場合のパワーの差を図3にヒストグラムで示した。やはり分布が右へ偏るほど、利得の変化が大きいと言える。これらの結果より、LPCフィルタにおいて、時間特性の問題や利得が大きく変化する現象が確認できる。

【００１０】

以上の理由から、LPC系の分析合成では、原音声のピッチ周波数を用いれば比較的高い品質の分析合成音が得られるが、原音と異なるピッチ周波数で駆動すると品質が劣化する現象が見られると考えられる。この問題は、有極型フィルタ(巡回型ディジタルフィルタ)の本質に根ざす問題で解消は難しい。仮にそれを改善するためにQ値を下げると、包絡の山と谷のレベル差が形成できず明瞭性の低いbuzzyな印象の音が生成されてしまう。

【００１１】

CSM法（非特許文献６）では、線スペクトルモデルに基づく音声分析法であるCSM音声分析によって、フォルマント周波数にほぼ対応する複数個の正弦波周波数(CSM 周波数)を得る。そして、それらの周波数の正弦波の和を基本波形として、位相を基本周期ごとに0にリセットすることで音声を合成する。線スペクトルを広げる目的で振幅に指数関数減衰を乗じることも行われた。これは、巡回型フィルタを用いずにパラメトリックに音声合成が行える方式なので、振幅の制御は極めて容易であるため「歯切れのよい」音声合成が期待できる。しかし、CSM法は音声スペクトルを図4のように線スペクトルで近似することに相当するため、スペクトルの再現方法としては検討の余地が残っていた。

【特許文献１】特公昭６１－１３６００号

【非特許文献１】ニック・キャンベル，アラン・ブラック：“CHATR: 自然音声波形接続型任意音声合成システム，” 信号処理学会技術報告，vol.96,no. 39, pp. 45-52, 1996.

【非特許文献２】F. Itakura and S. Saito: “AnalysisSynthesis Telephony Based onthe Maximum Likelihood Method,” Proc. 6th Int. Congresson Acoustics, 1968

【非特許文献３】北脇信彦，板倉文忠，斉藤収三：“ PARCOR 形音声分析合成系における最適符号構成，” 電子通信学会論文誌，J61-A, pp.119-126, 1978

【非特許文献４】管村昇，板倉文忠：“線スペクトル対(LSP) 音声分析合成方式による音声情報圧縮，” 電子通信学会論文誌，J64-A, pp. 599-606,1981.

【非特許文献５】今井聖，北村正，竹谷博行：“2 次元ケプストラムを利用する音声分析，” 電子通信学会論文誌，J59-A, pp. 1096-1103, 1976.

【非特許文献６】嵯峨山茂樹，板倉文忠：“複合正弦波による音声合成，” 音声研究会資料，S79-39, pp.293-300, 1979.

【非特許文献７】ParhamZolfaghari, Tony Robinson, “Formant Analysis Using Mixture of Gaussians,” Proc.ICSLP 96, vol. 2, pp. 1229. 1232, 1996. .

【非特許文献８】嵯峨山茂樹，古井貞熙：“ラグ窓を用いたピッチ抽出の一方法，”電子情報通信学会全国大会予稿集，1235, Vol. 5, p. 263, 1978.

【非特許文献９】亀岡弘和，西本卓也，嵯峨山茂樹，“ 調波時間構造化クラスタリング(HTC) による音楽の音響特徴量同時推定，” 情報処理学会研究報告，2005-MUS-61-12, pp. 71-78, 2005.

【非特許文献１０】亀岡弘和，小野順貴，嵯峨山茂樹：“スペクトル包絡と調波構造の合

成関数モデルによる音声分析，” 日本音響学会2005 年秋季研究発表会講演論文集，2-6-4，2005.

【発明の開示】

【発明が解決しようとする課題】

【００１２】

本発明は、高品質の合成音声を提供すると共に、加工性に優れた音声合成手法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】

【００１３】

本発明が採用した音声合成方法は、フレーム毎の音声のスペクトル包絡を所定数の単峰性関数の混合分布で近似する音声分析により得られた音声スペクトル特徴量に基づいて、時間領域において前記単峰性関数に対応する関数を所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置することを特徴とする。有声音の音声はほぼ周期波形で、周期的にある波形が繰り返される。本明細書では、その周期的に繰り返される波形を基本波形という。本発明では、基本波形は、前記音声スペクトル特徴量に基づいて求めることができる。加工（変化）された音声スペクトル特徴量に基づいて、基本波形を求めてもよい。

【００１４】

１つの好ましい態様では、前記音声スペクトル特徴量は、単峰性関数の混合分布のモデルパラメータの取得である。混合分布のパラメータから、目的とする基本波形を生成することができる。典型的には、前記モデルパラメータは、各単峰性関数の平均、分散、重みを含む。尚、単峰性関数の混合数をパラメータに含めて扱ってもよい。１つの好ましい態様では、前記モデルパラメータは、ＥＭアルゴリズムを用いて取得される。本明細書において、「EMアルゴリズム」は、非特許文献９で用いられているような実質的にEMアルゴリズムと等価であるアルゴリズムも含む意味で用いる。

【００１５】

時間領域における基本波形は、周波数領域における混合分布に対応すると考えられ、前記基本波形は、前記混合分布を逆フーリエ変換したものに相当する。本発明において、逆フーリエ変換は必須ではなく、時間領域におけるある関数と周波数領域におけるある関数との対応関係が既知であれば、周波数領域の関数のパラメータ（音声スペクトル特徴量）を用いて直接基本波形を計算することができる。１つの好ましい態様では、周波数領域におけるガウス分布関数と時間領域におけるガボール関数を対応させる。したがって、この場合、前記混合分布は、所定数のガウス分布関数からなる混合ガウス分布であり、前記基本波形は、所定数のガボール関数を重畳してなる複合ガボール関数である。

【００１６】

１つの態様では、前記音声合成方法は、フレーム毎の音声のスペクトル包絡を、所定数の単峰性関数の混合分布で近似し、音声スペクトル特徴量を求める音声分析ステップを含む。スペクトル包絡の取得は必須ではなく、予め取得され格納されているスペクトル包絡を分析してもよい。１つの態様では、前記スペクトル包絡は、ラグ窓を用いた音声スペクトルの平滑化により取得される。

【００１７】

１つの態様では、前記音声合成方法は、ピッチ抽出を含む。また、１つの態様では、前記音声合成方法は、有声音/無声音の判定を含む。

【００１８】

１つの態様では、前記音声が有声音であり、前記駆動時点は、ピッチ周期ごとに設定される。すなわち、基本波形を、ピッチ周期で配置することで、有声音を合成する。１つの態様では、駆動時点を設定するにあたり、複合波形の各成分ごとに重畳時点をずらしてピッチ周期で重畳する。有声音の場合も駆動時点の配置は、周期ごとには限定されない。１つの態様では、前記駆動時点は、ピッチ周期内に複数ある。LPCにおけるマルチパルス方式に倣って、大小の駆動時点を適切に配置して、そこに「複合Gabor関数」を配置しても良

い。これにより合成音声品質の向上が見込まれる。
【００１９】
１つの態様では、前記音声が無声音であり、前記駆動時点は、ランダム間隔に設定される。例えば、ランダム信号と複合Gabor関数の畳み込みが考えられる。また、無声音については、従来の無声音の生成法を採用してもよい。
【００２０】
本発明が採用した音声合成装置は、フレーム毎の音声のスペクトル包絡を所定数の単峰性関数の混合分布で近似する音声分析により得られた音声スペクトル特徴量に基づいて、時間領域において前記単峰性関数に対応する関数を所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置することを特徴とする。１つの好ましい態様では、前記混合分布は、所定数のガウス分布関数からなる混合ガウス分布であり、前記基本波形は、所定数のガボール関数を重畳してなる複合ガボール関数である。
【００２１】
１つの態様では、前記音声合成装置は、フレーム毎の音声のスペクトル包絡を所定数の単峰性関数の混合分布で近似して音声スペクトル特徴量を取得する音声分析部と、時間領域において前記単峰性関数に対応する関数を所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置する音声合成部と、を有する。
【００２２】
１つの態様では、前記音声合成装置は、フレーム毎の音声のスペクトル包絡を所定数の単峰性関数の混合分布で近似する音声分析により得られた音声スペクトル特徴量を記憶する記憶部と、時間領域において前記単峰性関数に対応する関数を所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置する音声合成部と、を有する。
【００２３】
本発明に係る音声合成方法は全てコンピュータによって実行することができる。また、本発明に係る音声合成装置は、コンピュータ（入力手段、出力手段、表示手段、演算手段、記憶手段、を含む）によって構成することができる。したがって、本発明は、さらに、本発明に係る音声合成をコンピュータに実行させるためのコンピュータプログラム、ないし、当該コンピュータプログラムを記録したコンピュータ読み取り可能な記録媒体に係る。
【００２４】
１つの態様では、本発明は、フレーム毎の音声のスペクトル包絡を所定数の単峰性関数の混合分布で近似する音声分析により得られた音声スペクトル特徴量に基づいて音声合成を行うためにコンピュータを、時間領域において前記単峰性関数に対応する関数を所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置する手段として機能させるための音声合成用コンピュータプログラム、である。
【００２５】
１つの態様では、本発明は、音声スペクトル特徴量に基づいて音声合成するためにコンピュータを、フレーム毎の音声のスペクトル包絡を所定数の単峰性関数の混合分布で近似する音声分析により得られた音声スペクトル特徴量を記憶する記憶手段と、時間領域において前記単峰性関数に対応する関数を所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置する手段と、して機能させるための音声合成用コンピュータプログラム、である。
【００２６】
１つの態様では、本発明は、音声スペクトル特徴量に基づいて音声合成するためにコンピュータを、フレーム毎の音声のスペクトル包絡を所定数の単峰性関数の混合分布で近似する音声分析により音声スペクトル特徴量を取得する手段と、時間領域において前記単峰性関数に対応する関数を所定数重畳させてなる複合関数を基本波形とし、前記基本波形を所定の駆動時点に配置する手段と、して機能させるための音声合成用コンピュータプログラム。
【発明の効果】

【００２７】
本発明によれば、フィルタ型音声合成に比べて高品質の合成音声が得られる。本発明に係る音声合成手法は、音声スペクトル特徴量に基づいて音声合成を行うものであり、波形接続型音声合成に比べて加工性に優れる。したがって、本発明によれば、対話音声の生成に適した、高品質かつ多様なスタイルの音声を生成可能な音声合成が可能となる。
【発明を実施するための最良の形態】
【００２８】
本発明に係る音声分析合成手法について、１つの好適な実施形態である複合ウェーブレットモデル（ＣＷＭ：Composite Wavelet Model）に基づいて説明する。
【００２９】
［Ａ］基本波形の接続による音声合成
先ず、本発明に係る複合ウェーブモデルの前提となる基本波形の接続による音声合成について説明する。従来例で述べた方式における有声音の合成を、ピッチ周期のインパルス列を入力したある線形系と考えて、その線形系のインパルス応答により整理すると、波形接続型では音声波形のピッチ周期波形そのものをインパルス応答とするのに対し、全極型フィルタでは推定されたスペクトル包絡の逆Fourier変換が対応する。
【００３０】
これを基本波形の繰り返しとして解釈し比較すると、波形接続型におけるピッチ周期波形は、これを構成する基本正弦波とその多数の高調正弦波の重ね合わせととらえられるが、これら個々の振幅位相はピッチそのものに大きく依存するため、ピッチと独立した制御には適さない。
【００３１】
一方、CSM合成においてはほぼフォルマント周波数に対応する正弦波断片が、全極型フィルタにおいては単振動(二次系)のインパルス応答である指数型減衰正弦波が、それぞれ基本波形となっており、いずれもこれら基本波形の重ね合わせと解釈できる。これら基本波形に必要な性質は、音声のスペクトル包絡をよく近似するスペクトルをもつことである。この意味からは必ずしも巡回型フィルタの場合のような長い基本波形は必要ではなく、巡回型フィルタでは単に時間特性を悪化させる要因になっている。
【００３２】
したがって、パラメトリックでかつ時間特性が良い音声合成は、少なくとも有声音の合成においては、巡回型フィルタを用いず、スペクトル包絡の逆Fourier 変換をピッチ周期で繰り返し、それに希望する振幅を乗じる方法が有利である。
【００３３】
［Ｂ］基本波形のモデル化
合成音声の基本波形を少数の扱いやすいパラメータによって表現することができれば、合成音声の声質や感情を操作するなどの加工がしやすくなる可能性がある。その要求条件には、
（１）　多様な音声を少数のパラメータで表現できるパラメトリックな方式であること、
（２）　音声スペクトルの大きなダイナミックレンジを表現でき、かつQ値は低く抑えるために、巡回型フィルタによらない方式であること、
が要求される。
【００３４】
そこで、次のFourier変換公式に着目する。ωを周波数、tを時間、a,b, cを任意の実数とすると、
【数１】

$$\mathcal{F}\left[\frac{a}{2\sqrt{b\pi}}e^{-\frac{t^2}{4b}+jct}\right]=ae^{-b(\omega-c)^2} \qquad (1)$$

が成り立つ。すなわち、周波数領域のガウス分布関数は、図5(a)に示すように、時間領域ではガウス分布関数と正弦波の積であるGabor 関数で表される。ガウス分布関数はdB尺度

で見れば下に開いた放物線であり、これを共振特性と考えるとQ値を抑えつつ、かつ大きな山と谷を形成するのに都合がよい。これらの関数対は、スペクトル領域でも時間領域でも大きく拡がらない利点を持つ。これを音声のフォルマントに対応づけて考える。

【００３５】

したがって、図6に示すように、音声スペクトル包絡を、混合ガウス分布関数モデル(GMM)で近似すれば、GMMで表されたスペクトル包絡から、基本波形を生成することができる。(振幅)スペクトル包絡を図5(b)のように複数のガウス分布関数の重ね合わせによって近似した場合には、基本波形は複数のGabor関数の重ね合わせとなる。このため、本手法を複合正弦波モデル(Composite Sinusoidal Modeling)に倣って、正弦波の代わりにGabor Waveletの重ね合わせを基本波形とするという意味で、複合ウェーブレットモデル(ＣＷＭ：Composite Wavelet Model)と名付ける。尚、通常、GMMはGaussian Mixture Modelの略で、混合ガウス分布密度モデルを意味し、その積分値は1に等しくなければならない。しかし、本明細書において、GMMは、スペクトル(パワースペクトルあるいは0位相化した振幅スペクトル)のモデルとしての混合ガウス分布関数モデルを意味するものとする。

【００３６】

［Ｃ］EMアルゴリズムを用いたＧＭＭの近似による音声分析法

少数のガウス分布関数でスペクトル包絡の近似を行って音声スペクトル特徴量（平均、分散、重み）を取得することで、各混合成分の平均がフォルマント周波数に、分散がフォルマントの広がりに対応することが期待でき、分析パラメータ（音声スペクトル特徴量）によって音声のフォルマント構造を直接操作できる可能性がある。これにより、フォルマント音声合成同様に音声学の知見を活かした声質変換の点で有利であると考えられる。また、逆に多数のガウス分布関数でスペクトル包絡の近似を行う場合には、加工は難しくなるが近似の精度がよくなり音声品質が向上することが期待できる。

【００３７】

非特許文献７には、音声スペクトルのフォルマント分析のためにスペクトル包絡を混合ガウス分布関数で近似する手法が開示されている。本発明における音声分析においては、非特許文献７に開示された手法を用いることもできる。しかし、分布密度関数推定に関するEM(Expectation-Maximization)アルゴリズムがパワースペクトルのモデル化にそのまま使用できるかどうかは自明でない。それについては、非特許文献９で議論されており、EMアルゴリズムと同型のアルゴリズムにより、観測したスペクトルに対するモデルパラメータのKL尺度(Kullback-LeiblerL情報量と同型の関数間の擬距離)を最小化(あるいは極小化)することができることが示されている。ここでは、その原理に基づいて、EMアルゴリズムに同型なアルゴリズムに基づいて、分析フレーム単位の音声スペクトルのGMM推定によりスペクトルパラメータを抽出する。本明細書において、「EMアルゴリズム」は、非特許文献９で用いられているような実質的にEMアルゴリズムと等価であるアルゴリズムも含む意味で用いる。

【００３８】

また、非特許文献７では、GMM化が包絡でなくピッチ構造に収束する場合を指摘している。本実施例では、自己相関関数にラグ窓を掛けてフーリエ変換することにより平滑化パワースペクトルを得て用いることによりその問題を回避している。ラグ窓を用いたスペクトル包絡の計算については、特許文献１を参照することができる。

【００３９】

［Ｄ］ＣＷＭを用いた音声分析合成手順

本発明の音声合成法の手順を示す。本発明に係る音声合成法は、予め実行される音声分析ステップと、その分析結果の蓄積・伝送・加工などを経て行なわれる音声合成ステップと、から構成されている。以下に、音声分析ステップの１つの好ましい態様、音声合成ステップの１つの好ましい態様を例示する。

【００４０】

［Ｄ－１］分析系の手順

（１）フレーム毎に音声スペクトル特徴量を計算する。

その詳細は、例えば、
（1a）音声波形の差分処理を行う（例えば、高域強調フィルタを通す）；
（1b）短時間ごとに音声波形を切り出しデータ窓(Hamming窓など)を掛ける；
（1c）音声信号の自己相関関数を求める；
（1d）自己相関関数に窓(ラグ窓)を掛ける；
（1e）フーリエ変換する(FFTなどのアルゴリズムによる)；
（1f）周波数点ごとに平方根を求める（これにより零位相化された平滑化振幅スペクトルが求まる）；
ことで行う。
ここでの音声スペクトルの計算については、特許文献１を参照することができる。また、音声スペクトルの計算（音声スペクトル包絡の取得）については、その他の公知の様々な手法を採用することができる。

【００４１】
（２）フレームごとの音声スペクトルを混合ガウス関数で近似する。
その詳細は、たとえば、
混合ガウス関数(GMM)のモデルパラメータ(平均、分散、重み)を、適当な初期値から出発して、EMアルゴリズムに類似したアルゴリズムにより求める。
すなわち、混合数をmとして、
各平均μ_i(i=1,2,3,...,m)、
各分散σ_i^2 (i=1,2,3,...,m)、
各重みw_i(i=1,2,3,...,m)、
がスペクトル分析結果である。
ここで用いたEMアルゴリズムに類似したアルゴリズムの詳細については非特許文献９を参照することができる。音声スペクトル包絡をGMMで近似する手法は、ここで述べたものに限定されるものではなく、非特許文献７に記載された手法、その他の手法を用いても良い。例えば、非特許文献１０に記載されたスペクトル包絡推定を用いることも可能である。

【００４２】
目的に応じて、さらに有声音/無声音の判定、有声音の場合は基本周波数（ピッチ周波数）を求めて、分析結果に追加してもよい。有声音/無声音の判定とF_0推定には、既存のピッチ抽出手法を利用することができる。また、ラグ窓を用いたピッチ抽出については、非特許文献８を参照することができる。

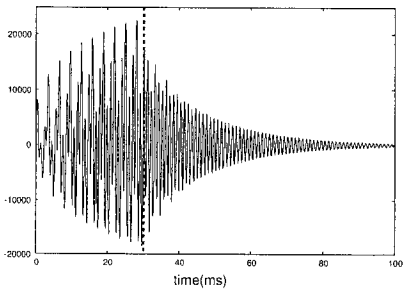
【００４３】
［Ｄ－２］合成系の手順
（１）駆動時点を決定する。有声音の場合はピッチ周期ごと、無声音の場合はランダムに、駆動時点を決める。
本明細書において、有声音の合成の場合は、基本波形が周期的に繰り返されるが(実際は基本波形は徐々に形を変えて行くが)、その基本波形を配置する位置を駆動時点と呼ぶ。ＣＷＭ合成の場合は、ＣＷＭ基本波形は原点を中心にした左右対称波形であるが、これを時間軸上に周期的に配置して、周期波形を作る。そのような、基本波形の中心を置く時点のことを駆動時点と呼ぶ。

【００４４】
（２）駆動時点に対応するフレームの分析で得られた(あるいはそれを加工した)混合ガウス関数の逆フーリエ変換に相当する「複合Gabor関数」(複数のGabor関数を重畳したもの)を、駆動時点に配置する。フレームごとのガウス関数の平均μ_i、分散σ_i^2、重みw_i (但しｉ = 1,…,m)から、GMMの逆フーリエ変換に対応するGabor関数の重みつき和を求める。これを、ピッチ周期間隔で周期的に配置して音声合成出力とする。
（２－１）Gabor関数はすべて中心を揃えるのでなく、適度にずらせば、ピークを下げつつ全体のエネルギーを増す(波高率の改善)ができる。同時に、合成音声波形の位相を調整して品質を向上させられる可能性もある。

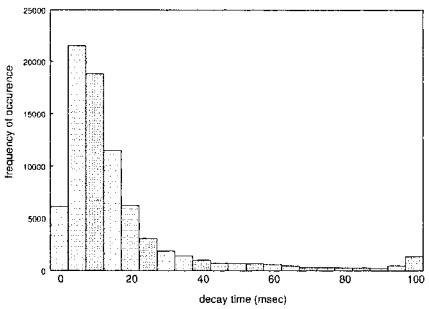
（２－２）有声音の場合も周期ごととは限らずに、LPCにおけるマルチパルス方式に倣って、大小の駆動時点を適切に配置して、そこに「複合Gabor関数」を配置して良い。合成音声品質の向上が見込まれる。駆動時点に相当するマルチパルスを、単一パルスから複合パルス、さらにランダムパルスまで連続的に変化させることで、有声音から無声音までを連続的に生成することができ、より滑らかに自然な音声を合成することができる。
（２－３）無声音の場合は、ランダム信号と複合Gabor関数の畳み込みでも良い。その他、無声音の生成法のみ従来手法を用いるなど、各種の変形が考えられる。
【００４５】
［Ｅ］音声合成実験
本発明の音声分析合成手法の有効性を確認するために、分析合成によって音声が再現されるかを確認した。また、従来法の問題点の解決に向けて、LPC法との比較を行った。
【００４６】
［Ｅ－１］実験条件
まず、本発明に係る音声合成法及びGMMによる近似の動作検証のために、本発明の手法によって音声を低次元のパラメータに分析し、パラメータから合成を行った。実験にはATR音声データベースより3－5秒程度の女性話者による文音声を5 程度選び、用いた。サンプリング周波数16kHz、サンプルサイズ16bitの音声に対して、ラグ窓法によるスペクトル包絡の抽出を行った。さらに、スペクトル包絡を5個のガウス関数の和に近似した。したがって、分析パラメータは１フレームにつき15次元である。今回は、F_0はSnack Sound Toolkit(“The Snack Sound Toolkit,”http://www.speechkth.se/snack/）付属のF_0抽出ツールによって抽出した。また、フレーム長30ms、フレームシフト10msで分析した。
【００４７】
まず、ピッチ周期や分析パラメータに変更を加えず、音声を合成した。無声音については、ランダムなピッチ周期を与える方法で合成した。そして聴取による比較の他、スペクトルの比較を行った。さらに、ピッチ周期や分析パラメータの平均を0.7倍－1.3倍程度に変化させ、音声を合成し、音声として破綻していないか聴取によって確認を行った。後者は、フォルマント周波数を変更したことに相当する。
【００４８】
本発明に係る音声合成手法により時間特性が改善することを示すため、従来例で記載したLPCフィルタについての実験と同様の実験を行い、時間特性と利得特性を調べた。
【００４９】
［Ｅ－２］実験結果と考察
聴取実験によって、良好な音声が合成されることを確認したが、背景にブザー的な雑音が聴かれた。図8に「うれしいはずが...」の冒頭部分の原音声と提案法の合成音声のスペクトルを示す。この図から分かるように、合成音声はかなり原音声の特徴を再現できているが、基本波形をゼロ位相化しているためにエネルギーの集中が著しくなっていることがわかる。図8に本発明に係る音声合成手法により合成される「あ」の音の一部を示す。原音声とは明らかに異なる波形を持つが、スペクトルはほぼ同じである。ピッチ周期やフォルマント周波数を変更する試験を行ったところ、いずれの条件においても破綻することなく音声を合成することができた。図9および図10に本発明に係る手法の時間特性と利得特性を示す。図2および図3との比較より、本発明に係る手法によって時間特性が改善し、かつ利得が安定したことがわかる。
【産業上の利用可能性】
【００５０】
本発明は、高音質かつ多様な声質の音声を生成可能な音声合成技術を提供するものであり、歌声合成、会話音声や感情音声の生成、音声対話システム、カーナビゲーションシステム、HMM音声合成系と組み合わせた擬人化エージェントやロボット、視覚障害者支援などの様々な場面において利用が可能である。
【図面の簡単な説明】
【００５１】

【図１】LPC フィルタ出力の時間特性の例を示す。
【図２】LPC フィルタ出力の時間特性の傾向を示す。
【図３】LPC フィルタ出力の利得特性の傾向を示す。
【図４】CSM 法による音声スペクトルの線スペクトル表現を示す。
【図５】ガウス関数のFourier 変換対を示す。
【図６】GMM による音声スペクトルの近似の例を示す。
【図７】原音声(上) および合成音声(下) のスペクトログラム例を示す。
【図８】原音声(上) および合成音声(下) の波形例を示す。
【図９】本発明に係る手法の時間特性を示す。
【図１０】本発明に係る手法の利得格差を示す。

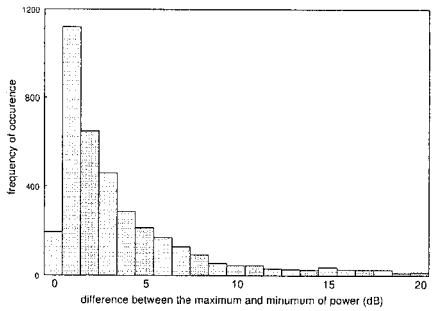
【図１】



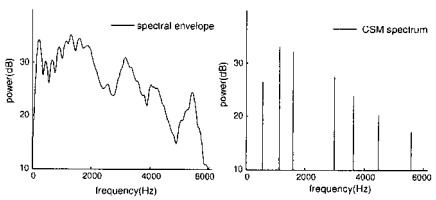
【図２】



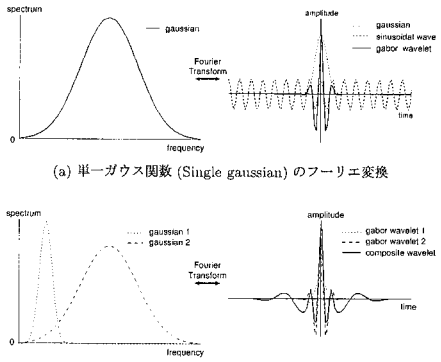
【図３】



【図４】



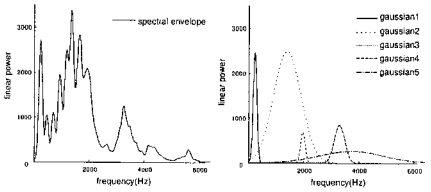
【図５】



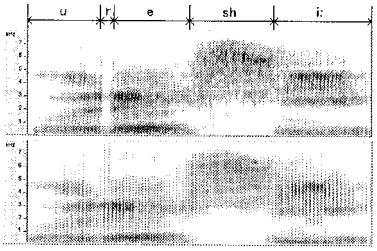
(a) 単一ガウス関数 (Single gaussian) のフーリエ変換

(b) 混合ガウス関数 (Gaussian mixture) のフーリエ変換

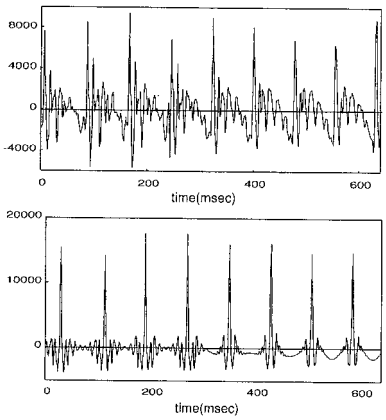
【図６】



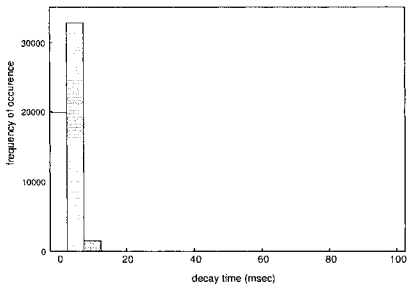
【図７】



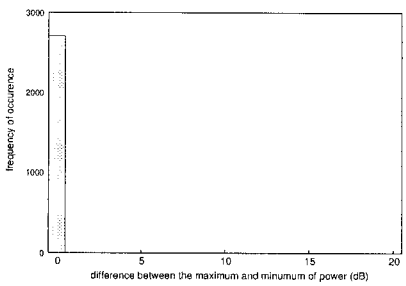
【図８】



【図９】



【図１０】



フロントページの続き

(72)発明者　松本　恭輔
東京都文京区本郷七丁目３番１号　国立大学法人東京大学内
(72)発明者　西本　卓也
東京都文京区本郷七丁目３番１号　国立大学法人東京大学内