Internal Use Only(非公開)

TR-I-0175

声質変換技術と高品質ピッチ変換法

Voice Quality Conversion and A High Quality Pitch Modification

梅田哲夫

UMEDA, Tetsuo

1990.7

内容梗概

物理的な面から音声の個人性の問題を取り上げそれらの制御可能な部分を利用した声質変換法の原理について述べた。また、高品質な声質制御を行う場合、従来、残差駆動による分析合成方式が利用されてきたが音源情報と声道特性の分離が不完全なため、ピッチ変更時に合成音の品質劣化を生じやすい。そこで、この品質劣化を低減する新しい変換方式を示す。

ＡＴＲ自動翻訳電話研究所

ATR Interpreting Telephony Reserch Laboratories

あらまし

我々が人の音声を聞いたときにまず感じるのは言うまでもなく発話者が伝えたい情報すなわち言語情報である。しかし、これ以外にも話者が男性か女性か、おおよそなん歳くらいか、またどんな感情で話しているかなども知ることができる。場合によれば、声の良い悪いについても感じ取ることができる。よく知っている人の声は短い音声、たとえば咳ばらいひとつでも誰のものか判ることがある。しかし、逆に親子や兄弟の声を取り違えてしまうこともある。

このように音声の担っている情報のうち、その人らしい、あるいは似ている、いないの個人性情報も大切な情報のひとつであり、音声の付加情報として声質に関係する重要な役割を果たしている。この個人性情報のうちには、育った環境で訓練されるような方言などを含めた社会的、職業的、教養的環境による話し方といったものを除いても、その人固有の声質というものがある。音声を取り扱う場合にはこの個人性情報を適当に処理することが不可欠である。すなわち、音声認識においては、発話者に対して個人差を出来るだけ減らして発音するように要求することは無理であるから、個人性の正規化をしたりあるいは、短時間の音声サンプルから個人性を抽出し、適応化することが必要になってくる。また、音声合成においては何等かの形で合成された音声に対して個人性情報を付加して自然性のある音声を生成したり、あるいは人間離れしていない音声を生成することが問題となる。また、話者識別や話者照合を行う場合は、音声認識ではむしろ妨害となっていた個人性情報を利用し、出来るだけ個人差の大きなパラメータを選択する必要がある。

第1部では、主として物理的な面から音声の個人性の問題を取り上げ、それらの制御可能な物理量を利用した声質変換の原理について述べる。

また、高品質な声質制御を行う場合、従来、残差駆動による分析合成方式が採用されてきたが、音源情報と声道特性の分離が不完全なため、ピッチ変更時に合成音の品質劣化を生じやすい。そこで、この品質劣化を低減する新しい変換方式を開発した。

まず、音声波形から複素ケプストラムを求め、これをピッチ周期の整数倍を取り出す櫛形リフタと逆リフタで分離し、それぞれの時間領域波形を求める。分離された音源情報については零位相化、声道特性についてはその逆の補償を施すこと

により、音源情報はパルス状になり、声道特性との分離が良くなる。この音源情報の位置を適応的に位相制御することによってピッチ同期で音声合成を行う複素ケプストラムによるピッチ変換法について第2部で報告する。

なお、本研究はＮＨＫ放送技術研究所の清山信正氏の協力によるところが大きい。

あらまし

もくじ

第１部　　声質変換技術

１．音声の個人性とは

明瞭に録音されたテープレコーダの再生音を聞くと、まるで当人がその場にいるかのように感じることがある。したがって音声波形の中には個人性の情報が含まれていることは確かである。また、短い音声でも個人を聞き分けられることを考慮にいれるといわゆる話し方だけに限らず、音声波形の内に存在する何れかの音響パラメータもその音響的特徴を表しているはずであるが、実際に個人性を担う音響パラメータが何であるかは今日までなお議論されているところであり簡単には見つけられないのが現状である。

音声の個人性は、発声に関係する諸器官（図１参照）の形や大きさに個人差があり、その調音器官の制御の仕方にも個人差があるところから生じるが、親子、兄弟が似ているように身体的特徴が似ていて育った環境が同じならば、その音声どうしもある程度似て来るのは当然である。一般的には音声の個人性といった場合にはこれらの調音器官の形状によって決まる音響的特徴の他に、個人個人の発音の癖や出身地による母音や子音の違いなどいわゆる言語学的な特徴までも含まれるが、本文では音声の個人性といった場合には音響学的特徴のみを指すものとする。しかし、会話音声、朗読音声等の自然音声を扱う場合には言語学的特徴を完全には分離しきれず、個人性を表す特定の音響的特徴を捜し出すことの困難の原因のひとつとなっている（図２参照）。また、発声者自身に関しても絶えず発声の仕方は変動している。もともと発声器官は声帯、舌、唇、鼻等軟らかい組織でできていてその動きも神経・筋肉等のアナログ量で制御されるものなので、不安定さは避けられない。例えば喜怒哀楽等の感情や環境雑音の大小等、発声者の置かれた環境によっても声の高低、口の開き方はかなり変わる。また、長時間かけて獲得した方言等を含んだパラメータでさえ数週間から数カ月で変動すると言われている[3]。この様に声の個人性の根拠や現象がかなりはっきりしている場合でさえ、いざ実際に個人識別をしようとしたり音響的特徴を抽出しようとすると大きな問題にぶつかってしまうのが現状である。

2．個人性を決める物理的特徴

音声が口や鼻を通じて外に放射されるには音源が必要であるが、母音の場合には声帯振動が、また子音の場合は舌や唇を使用しての呼気による摩擦音や、破裂音が各々の音源となる。線形予測分析合成技術の基本となっているモデルにおいてはこれらの２種類の音源がスイッチングされて用いられる場合が多い。しかし、自然音声においてはむしろこうしたモデルからのズレが個人性を決めている場合もある。また声帯は、呼気流の大きさ、声帯筋の緊張に応じて自励的な振動を続けることができる。これが声の高さ、強さに直接関係する音源信号として10kHz 以上まで高調波成分を持っ声帯波形となる。人による話者識別が平均ピッチ周波数をひとつの重要な手がかりとしていることは、同一母音を用いた話者識別の実験からも裏付けられている。また、ボコーダ等の情報圧縮を目的とした処理システムの場合ピッチ周波数や強さを数十ミリ秒のフレーム内で平均化しまたフレーム間で平滑化をしている事が多い。しかし、こうした平滑化されたパラメータで合成された単母音などから個人性や自然性が失われやすいこともよく経験することである。したがってピッチの平均からの外れ方にも個人性情報が含まれていると考えるべきである。この声帯波形が音響管としての共鳴特性を持った声道を通過して唇から放射されるまでに変形を受け音声波形となる。すなわち、人間の音声はピッチ周波数を基本波とする高調波成分からなり、その高調波成分の強さの包絡がその時点の声道の共振特性に対応している。

従って音声に含まれる情報は、個人性情報も含めて、声帯に関係する情報と声道に関係する情報の双方に含まれている。声帯と声道は音声を生成する器官の働きとしても独立でなく、さらに音声波形から声帯波だけを抽出したりあるいは声道の特性だけを分離する技術も現時点では完全ではないが、ある程度両者を分離して考えることができる。　すなわち、声帯特性として、

（１）ピッチ周波数の平均値

（２）ピッチパターン（音節単位程度の長さの時間的変化）

（３）ピッチ周波数のゆらぎ

（４）声帯波形

声道特性としては直接測定することが困難なため等価的なスペクトル情報を利用

して

（５）スペクトル包絡の形と傾斜

（６）ホルマント周波数（共鳴周波数）

（７）ホルマントパターン（時間的変化）

（８）平均スペクトル特性（文節、文章単位の長さの平均）

等に分離して個人性を考えることが多い。

しかし、これらのパラメータはそれぞれ互いに関連し合っていて単独に抽出するのが困難なだけでなく、どういった個人性特徴とどの程度関わっているかについても先に述べた様な変動要因によって単純には決められない。

男性、女性、少年、子供の声として聞こえるためには声の高さを決めているピッチ周波数とホルマントを作る共鳴管である声道の長さの間に一定の関係を保っていることが必要であると言われている[4]。これは管の断面積を任意に変えられる音響的声道模型を使ってシミュレートしたものである。これによると駆動源である声帯波形も声質を決定する重要な要因のひとつであることが分かる。例えば、駆動源として図３の三角波を採り波形の立ち下がり部の長さT3を短くすると張りのある声になり、長くすると鈍いもやもやした声になる。また、男性、女性、子供の声が最も自然に聞こえるためには声帯波形もそれぞれに合ったものにする必要がある。男性の声の場合、ピッチ周波数は80～150Hz 声道長17.5cm前後、T3は0.4～0.6mSecが自然な音声になる。女性の場合、ピッチは200～350Hz、声道長14cm前後、T3は0.5～1.0mSecの範囲が最も良く、短か過ぎると子供っぽくなる。子供の声の場合、ピッチは300～860Hz、声道長は9cm 前後、T3は0.2～ 1.0mSec で子供らしくなるが短いほど可愛らしくなるとしている。これらの関係を図４に示す。例えば男性らしく聞こえるためには同図のＭの範囲にあることが必要である。Ｂは少年、Ｆは女性、Ｃは子供の声の自然性を保つために必要な範囲であり、この範囲をはずれると人間離れした声になることを示している。実際の音声では、これら定常的な因子の他に、高さや、強さの不規則な変動がありそれらも個人性や自然性へ大きな寄与をしている。我々がある特定の母音を発声しようとする時には、その母音に固有な口の形を与えてやればよい。Ｘ線写真等を参考にして得られる日本語５母音に対するの標準的な声道形を図５に示す。縦軸は断面積（最大10cm2、最小0.6cm2） 横軸は声門から唇までの位置を示している。同図に示した形に相似であれば全長を8cm～24cm まで変

化させても充分な韻質が得られる。またこのときの音響伝達特性を図６に示した。声道長をｎ倍にした場合、周波数軸を１／ｎにしたものと相似の特性になるので横軸は周波数を声道長を１／４波長とした周波数で正規化したもので示してある。これからも分かるように母音の音韻性は男性、女性、子供にかかわらず音響伝達特性の共振の極（ホルマント）の位置の相対的な関係すなわち、声道の相対的な形（あごの開き、舌の位置）が決めていることが分かる。声道特性に関係して付与される個人性は、この標準形からのずれと、声道のダイナミックな制御の差から起こっていることが考えられる。

図７は、経験豊かなＮＨＫの男性アナウンサー１０名と一般の人５名の各々が発声した／青い空／を資料として一対比較法に基づいて音声の類似性判断した結果を多次元尺度構成法により２次元上に配置したものである[5]。本図では、聴覚上の類似度を図の上の距離で表している。1～10はアナウンサーの音声、11～15は一般の人の音声である。Ⅰ軸は右に行くほど明瞭な声で左へ行くほど口ごもる感じになる。またⅡ軸は上方は高い音声、下方は低い音声になる。図に見られるようにⅡ軸を境にしてアナウンサーと一般の人の声はきれいに分離して聞かれている。さらにアナウンサーの音声は個々には個性はあっても声質は似ていることを示している。この１５名の話者のピッチパターン（ピッチ周波数の変動の仕方）を調べて見ると、アナウンサーの方がピッチ周波数の変動幅が大きい。平均ピッチはアナウンサー平均で138Hz 、一般の人で124Hz で大きな差は無いが、ピッチ変動の最大値と最小値の差と、それを平均ピッチで割った変動の程度を見ると、アナウンサー平均で差は126Hz 92.5% の変動、一般の人で約半分の75Hz 60%の変動の変動であった。従ってピッチ周波数の動きの中にアナウンサーを一般の人と区別する要因のひとつがある。ホルマント周波数の動きを見るためにアナウンサーと一般の人とそれぞれ１名の第１、第２ホルマントの平面上に軌跡を書いた例を図８に示す。実線で示したアナウンサーのホルマント軌跡は波線の一般の人の軌跡と較べて激しい動きをしており、声道の形で言えば口をはっきりと動かしていることに相当する。

平均的な声道特性の違いを比較するために発声時間全体にわたって声道特性の平均を取ってみるとアナウンサーの音声は3～4kHz 付近の中域のレベルが一般的に高い。一般の人の場合、全帯域にわたって平坦に近い。オペラの歌声では3kHz付近に強い成分が現れ、これが歌声の響きに強く影響していると言われていることと符合

する現象である[6]。ピッチパターンの動きと平均ピッチ、ホルマントの動き、平均スペクトル包絡の４つのパラメータの話者間の類似度に対して聴感上の話者間の距離との相関をとると、ピッチパターンに対して0.40、平均ピッチに対して0.24、ホルマントパターンに対し0.52、平均スペクトル包絡について0.32の相関係数がそれぞれ得られ[6]、ピッチパターン、ホルマントパターンの動的特徴量との相関が高いことからアナウンサーの声の印象はこれら動的パラメータの中によくその特性が現れていると言える。この様に個人性について様々な研究がされている[7][8]。音源の声帯特性と声道特性のどちらにより多くの個人性情報が含まれているかを調べた研究も多いが音声資料として定常母音を用いた場合と短文を用いた場合とでは逆の結果が出ている場合もある。従って、ある特定のパラメータが個人性の情報を担っているのではなく、様々なパラメータの中に分散していることが推察される。実際に個人性の判断は、すでに述べたように発声の仕方、判断する人の着目するポイント、話者、発話内容などによってそのつど最重要視する音響パラメータが変化しているものと思われる。

３．音声の個人性判断

声道特性と音源パラメータを独立に制御することができると、個人性とこれらのパラメータの間の関係をさらに詳しく調べることができる。また、これによって声の質を制御できるようになる。

３．１　声質変換用分析・合成システムにおける音響パラメータの制御方法[9]

先に述べたように声道の特性は音声波形のスペクトル包絡に反映している。分析・合成技術を始めとして最近の音声情報処理技術の進歩により音声信号からスペクトル包絡（つまり声道特性）とその駆動音源波形を分離しまた再合成することができるようになった。その代表的なものは、声道逆フィルターと呼ばれる全極形モデルに基づく線形予測分析法、またはその改良版である。この分析結果である線形予測係数を用いることによって声道特性と等価なディジタルフィルターを構成することができる。またこの逆特性を音声信号に掛けることによって駆動信号（残差信号）を得ることができる。

図９に従来から用いられてきた声質変換用分析・合成方式のブロック図を示す。

自然音声は、たとえ定常的な母音の中心部でも全く同じ波形が繰り返すことは無い。１ピッチ毎の変化が自然性または個人性に寄与していることが考えられるため、この方式ではあらかじめピッチを抽出しておいて分析と合成を各ピッチ区間毎に同期させることも行っている。

図１０にピッチ周波数の制御法の原理を示す。分析は１ピッチ区間毎に行う。合成時には、各ピッチ区間に対しピッチ区間を長くする場合には残差信号の終わりに必要な長さだけ零信号を追加し、短くする場合には後ろから必要な長さだけ削除する。この残差信号を声道フィルターに通すことによってスペクトルは保存されたままの音声を再合成できる。この分析合成方式によっても大幅にピッチを変化させると音質劣化が避けられないが、かなり高品質で変換できる方法も提案されている。（第２部参照のこと）

声道特性は予測係数を係数とする高次多項式で表現できるディジタルフィルターとしてモデル化でき、そのホルマント周波数はこの多項式の根に対応している。そのためホルマント周波数やバンド幅を変えるためにはその根を周波数特性の変化に応じて変更し、その根が高次多項式の根となるように係数を変更するとその係数に基づいたディジタルフィルターが新しい声道特性となる。図１１にこのようにして変化させた音声の例を示す。

３．２　音響パラメータの変化と音声の個人性保存[10]

上で述べた様なシステムを利用すれば声道とピッチの音響パラメータを独立に制御できるので、それぞれを変化させた場合の音声の個人性の変化を調べることができる。図１２に全ホルマント周波数、全ホルマントのバンド幅、平均ピッチ周波数をそれぞれ独立に変化させた場合、個人性情報がどれくらい残っているかを調べたものを示す。ホルマント周波数の一様シフトに対して個人性は敏感である。これらの事実を利用することによって音声から個人性を消去することができる。３．に述べた音響的声道モデルによるシミュレーションの結果、相似的な形を保てば、かなりの範囲にわたってモデルのサイズを変えても音韻性保存されることが分かっている。個人毎に声道長は異なるが調音の様式は基本的に同じようになされている事を示唆している。調音を同じにし、声道長を変えるとホルマント周波数は対数周波数軸上で一様に平行移動する。個人の声道長が短期間で変化することは無いので人工

的な平行シフトに対して個人性は急速に失われたものと思われる。

4．音声変換システムの機能の分類

以上のまとめとして、上述した声質変換システムの機能と技術の現状及び考えられる応用例を示す。

（1）声の高さや抑揚の制御

これはもっとも基本的な機能の一つでありできるだけ高品質に行えることが望ましい。この技術はほぼ完成している。

（2）発声速度の変換

言葉としての自然さと速度との関連が問題である。この技術はほぼ完成しており、プロソディ制御のための時間調整等に利用可能と思われる。

（3）個人性の消去、男女声変換

音響レベルではほぼ完成したが、言語的特徴も制御することが必要になってくると考えられる。番組取材源の秘密保護に利用できる。

（4）明瞭性の向上や病的音声の改善

現在でもある程度可能だがケースに応じた処理方法の蓄積と自動化が必要。

（5）アクセントの修正や単語の差替え

音声データベース、言語学的辞書などを用いて自動化が必要になる。

（6）声の印象や感情表現の制御

聴感上の印象や感情表現と物理量（パラメータ）との関連の研究がさらに重要である。これは新しい音声編集・効果装置としての応用が考えられる。

（7）個人性の入れ替えや正規化

個人性の音響的・言語学的特徴の厳密な抽出と制御が必要で研究的課題が多い。正規化による不特定話者の音声認識等に応用できる。

（8）編集合成・規則合成の出力部

自然な品質を持つ音声素片の結合方法が研究課題である。これにより情報放送・自動翻訳の音声出力部へ応用可能となる。

現在、パラメータの変更の程度がある程度大きくなっても自然性がより良く保てる声質変換システムの研究が続けられている[11]。

５．参考文献

[1]鈴木（訳）：”音声の線形予測”コロナ社(1980)

[2]白井：”音声認識技術とその応用”テレビジョン学会誌Vol.41,No.8,P.716(1987)

[3]古井、板倉、斉藤：”長時間平均スペクトルによる話者認識”電子通信学会論文誌 Vol.55-A,No.10,P.549(1972)

[4]梅田、寺西：”声の韻と声質”日本音響学会誌第22巻第4号P.195(1966)

[5]桑原、大串：”アナウンサー音声の音響的特徴”電子通信学会論文誌Vol.J66-A,No.6,P.545(1983)

[6]辰巳、藤崎：”歌声の音響的特徴”日本音響学会音声研究会資料(1978-04)

[7]古井：”音声の個人性パラメータの時期的変動と話者認識”電子通信学会論文誌 Vol.57-A,No.12,P.880(1974)

[8]鈴木：”音声と話者の相関について”日本音響学会誌第41巻第12号P.985(1985)

[9]桑原：”声の個人性に関する諸問題”電子情報通信学会誌Vol.70,No.4,P.345(1987)

[10]桑原、大串：”ホルマント周波数・バンド幅の独立制御と個人性判断”日本音響学会 音声研究会資料(1984-40)

[11]都木、梅田：”ピッチ変更時の歪をスペクトル領域で補正する声質変換”信学技報　SP87-111(1988.1)

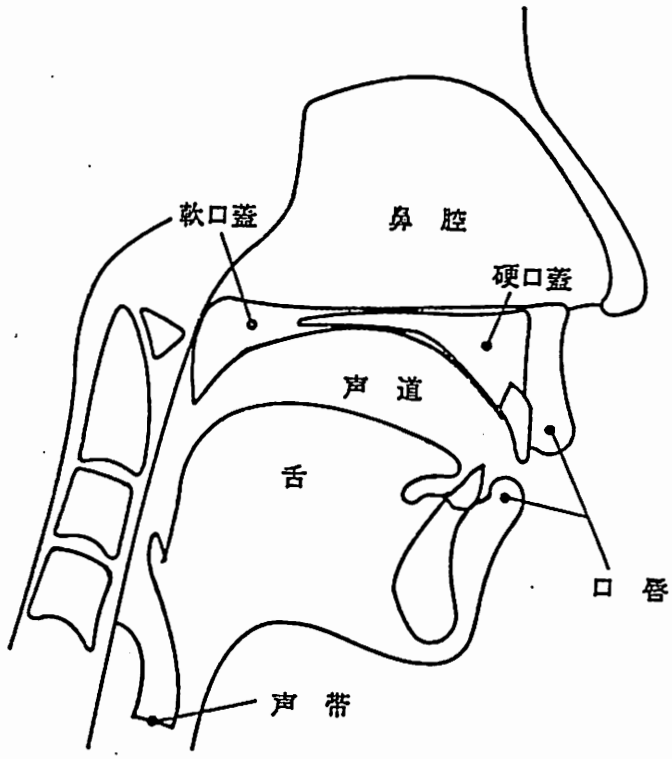


図1　発声に関係する諸器官[1]

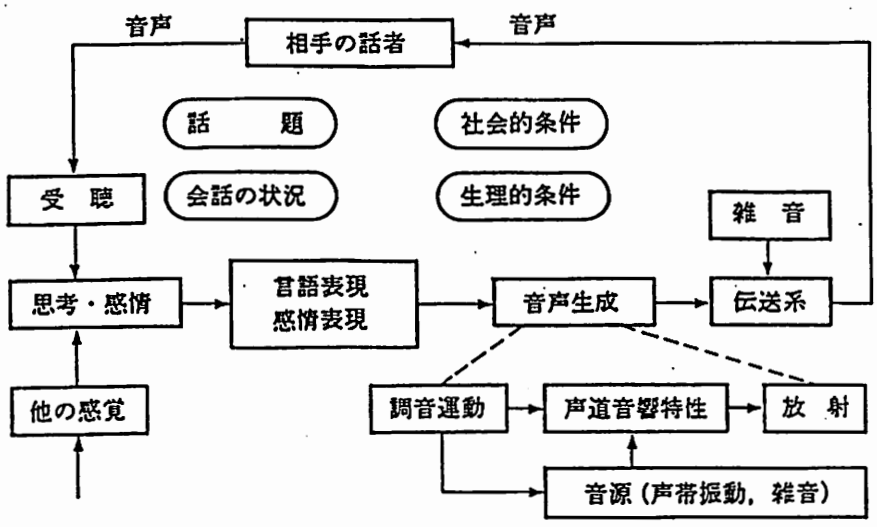


図2　音声の不安定さに関係する様々な要因[2]

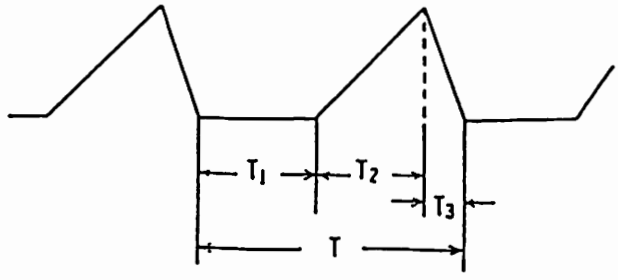


図3　声道モデルを駆動する声帯波[4]

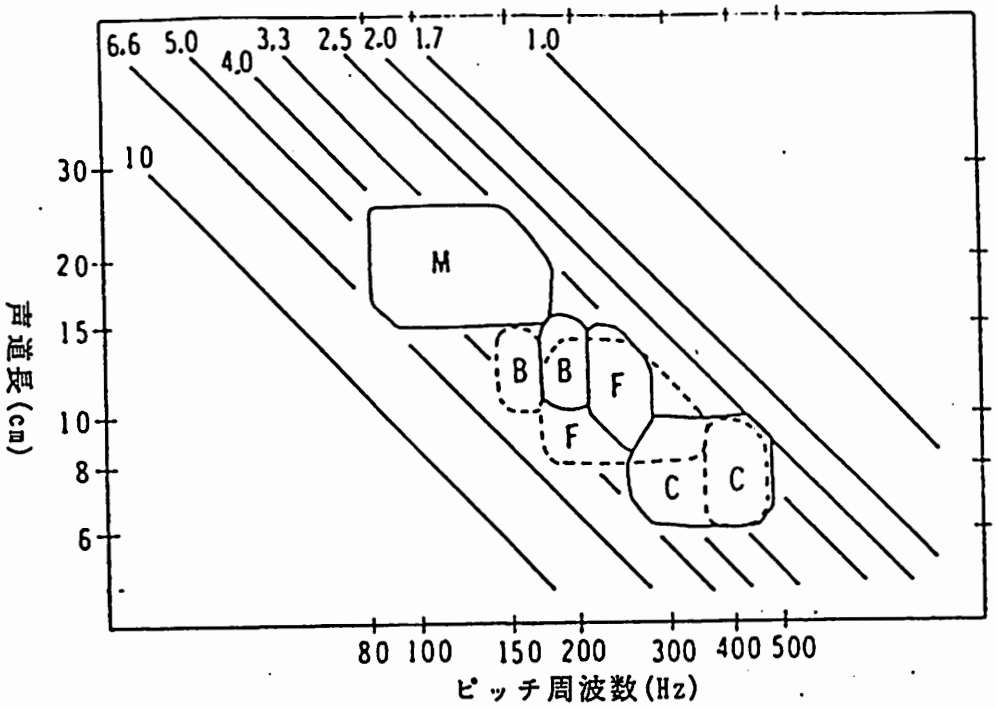


図4　声道長、ピッチ周波数、声帯波形が声質に及ぼす影響[4]
Mは男性、Fは女性、Bは少年、Cは子供の声に適した範囲を示す。実線はT3の立ち下がり0.4mSec、波線は1.0mSecの場合である。

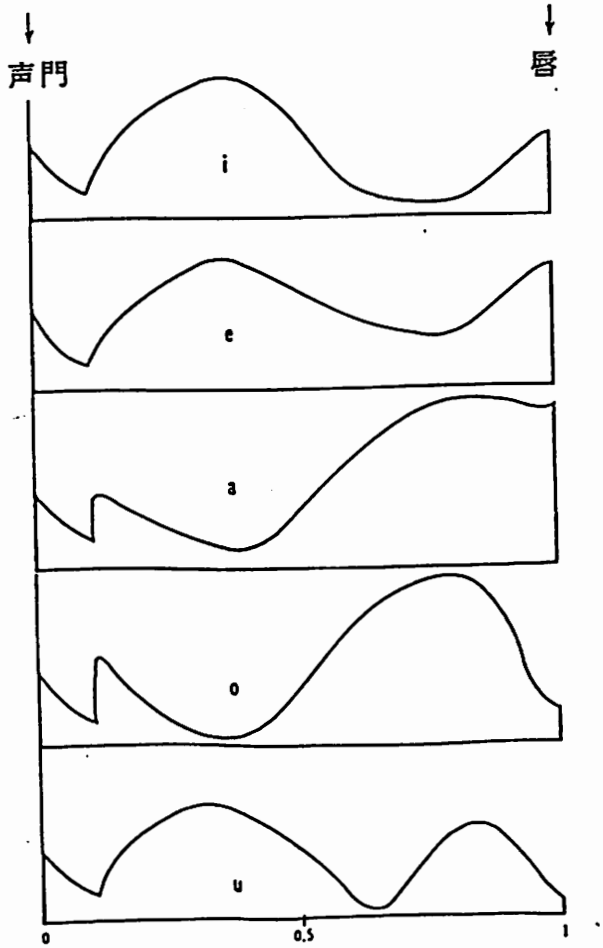


図5　標準的日本語5母音に対する声道形[4]

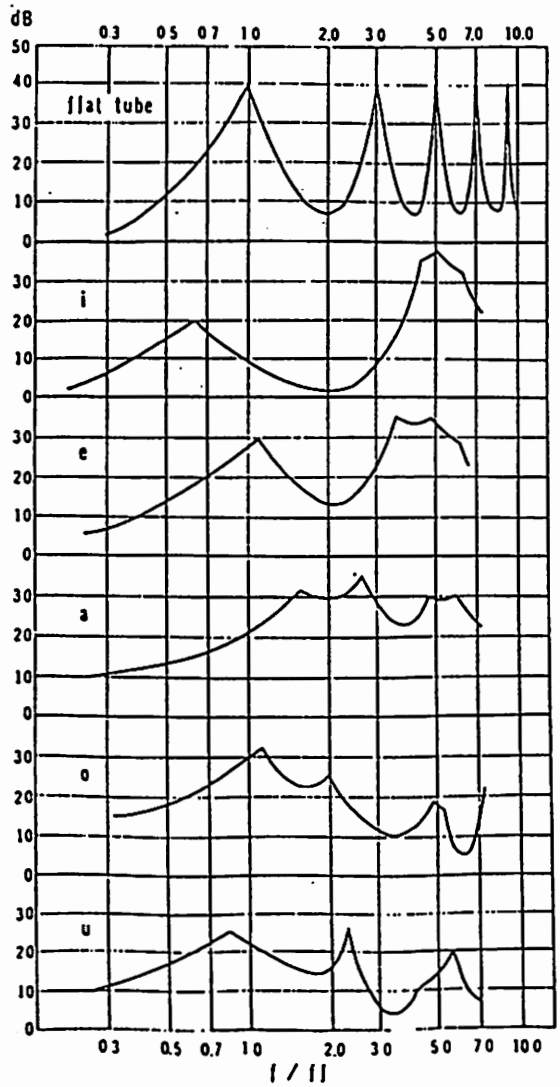


図6　標準的日本語5母音に対する声道伝達特性[4]
声道長を1/4波長とする周波数で正規化。

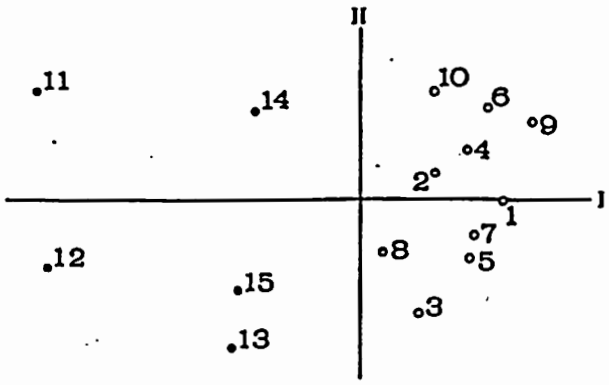
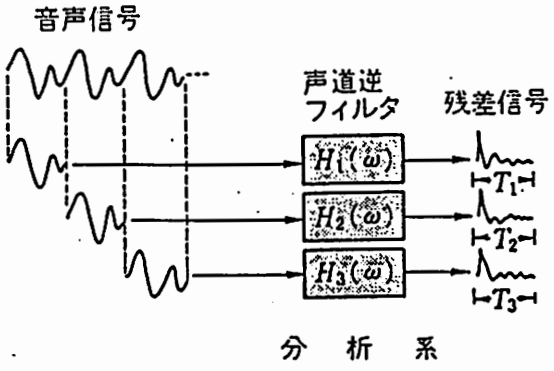


図７　アナウンサーと一般の人の音声の聴覚的な類似度の２次元配置[5]
1～10はアナウンサー
11～15は一般の人



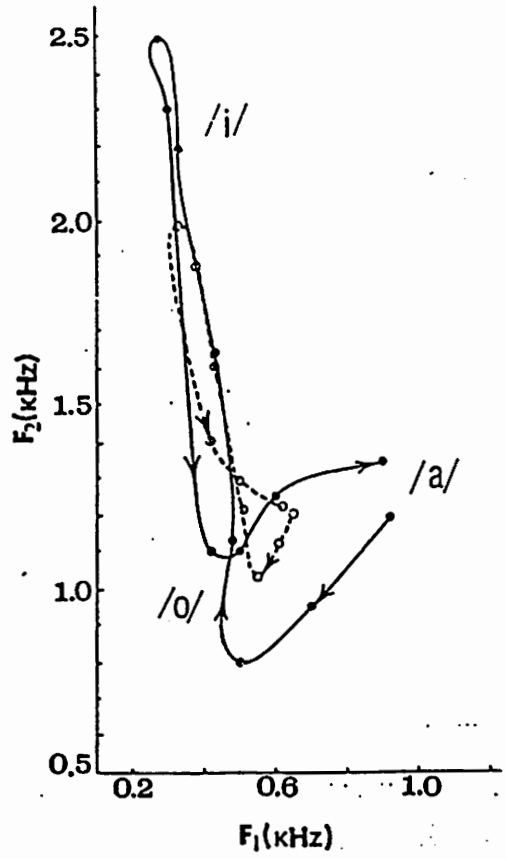


図８　アナウンサー（実線）と一般の人（破線）の第１、第２ホルマントの動き／青い空／[5]

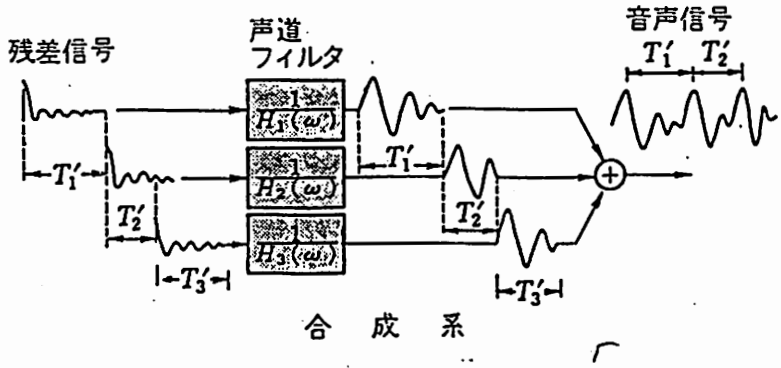


図１０　ピッチ周波数制御法[9]

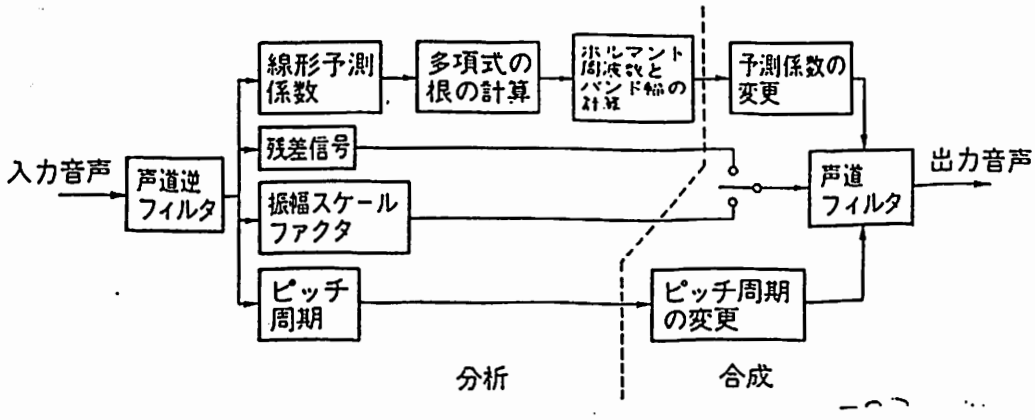


図９　声質変換用分析・合成システム[9]

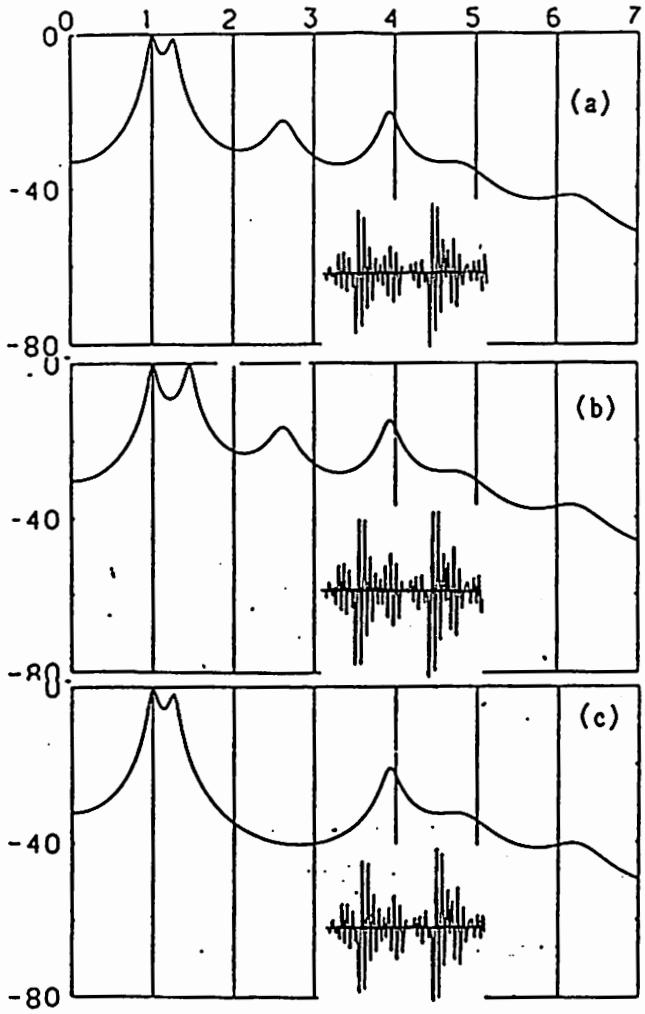


図１１　ホルマントの特性を変更した例[10]
(a)オリジナル音声
(b)第２ホルマントを10%シフト
(c)第３ホルマントの帯域幅を10倍したもの

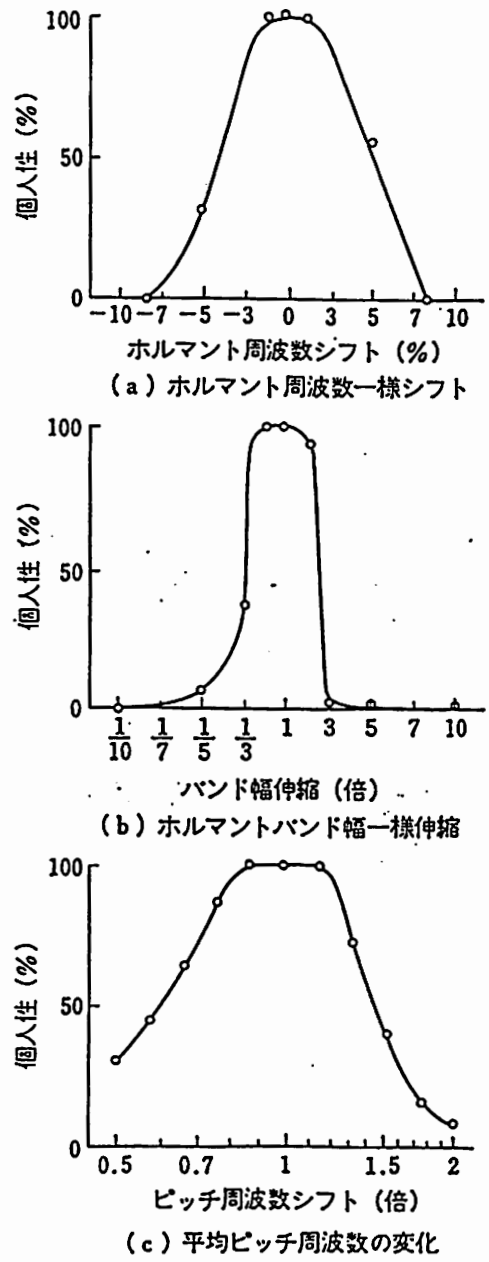


図１２　音響パラメータを変更したときの個人性の変化[9]

第２部　　複素ケプストラムによる高品質ピッチ変換法

１．まえがき

音声処理技術の基礎として、音声の個人性制御や音声制御を行う場合、できるだけ品質劣化の少ない高品質な合成が望まれている。また、音声の規則合成や録音編集方式を行う場合にもアクセントやイントネーションといった微妙な声質制御の研究を行うための基礎技術として高品質なピッチ変換方法が不可欠である。また、一応のテキスト合成ができるようになると、より自然で肉声に近い合成音への要求が強まってきた。

これらの合成器に用いられている規則合成法は、音素やＣＶ等の音節、あるいはＶＣＶ等の単位を音声結合素片としているのが一般的である。一般には波形やパラメータの辞書の許容容量の制限から、小数の素片をもとに結合時に韻律や調音結合を考慮して計算によって求められた物理パラメータに変更を加えることが多い。これらのパラメータにもとづいて分析合成方式で高品質な制御を行う場合、従来、線形予測により得られた声道フィルタを予測残差波形で駆動する、いわゆる残差駆動による分析合成方式が採用されてきた。しかし、音源情報と声道特性の分離が不完全であったり、本質的に関連を持っていたりするため、ピッチ変更時に残差波形の周期を変更することにより、同時に残差に含まれる音色の情報にも変化を与えることになり、このスペクトル的歪みに起因する合成音の品質劣化がさけられない。例えば、この残差波形には周期成分以外にも音韻性や、個人性等の音質を支配する情報もかなり含まれており、分析時と異なるピッチ周期で合成を行うために残差波形の打ち切りや零詰めを行うことによって情報の欠落が生じると考えられる。そのために、分析の際声道特性と駆動音源であるピッチパルス列に分離する時に駆動音源をできるだけパルス状にしようという試みもなされている。例えばＮＴＴでは、残差波形に１ピッチ区間毎に位相等価処理を施し、エネルギーを原点付近に集中させ打ち切り零詰めの影響を軽減する方式を提案した。しかし、この方式では残差波形が充分にはパルス状にならずピッチを大きく変化させた場合に影響がでるものと

考えられる。また、なによりもＬＰＣ残差波形のみを聴取してみると充分に元の音韻が理解できる程音韻情報が残っていることも問題である。ＡＴＲにおいてもパワーケプストラムのリフターとして、ピッチ周期に対応する部分を取り出す櫛型リフターを用いピッチの高調波の影響だけを取り出し、周波数領域においてピッチ変換する方式を提案した。しかし、折り返しにより高域スペクトルの補完や、パワーケプストラム法による位相情報の欠落、フレーム同期処理を行っている等が音質の限界につながっている様である。　本報告ではこのような音源情報と声道特性を分離する分析合成方式における品質劣化を低減するために、

①複素ケプストラムに櫛歯型リフタを適用し、パルス列と単位応答に分離する。

②パルス列を零位相化し単位応答にはその逆の補償を施す。

③ピッチ同期で処理をする。

以上を基本とするピッチ変換法を用いた新しい分析合成法を提案する。

2．本方式の特徴（原理）

本方式は音声の大部分を占める有声音区間の変換に関するものである。

分析合成方式でピッチ変更を行う場合、従来の残差駆動方式では、図１(a)のように予測残差波形の打ち切りや零づめを行うことによって情報が欠落し音質劣化を生じる。これは、予測残差波形の周波数特性はおおよそ平坦とはいっても音源情報以外の成分がかなり含まれているためと考えられる。

線形予測分析合成手法は声帯波とそれによって駆動される音響管というモデルにもとづいて主に音響管の共鳴特性を抽出する手法として確立されている。ピッチ変更等が必要となる音声合成や声質変換等の場合には残差信号が合成音声の音色に及ぼす影響を別に考慮する必要があった。

しかし、分析の際、音源情報と声道特性に分離する時に、音源情報としてはピッチ周波数情報だけが抽出されて音色を含まないパルス状になり、声道特性としては音声波の音色に関わる情報がすべて残りこれにはピッチ情報は含まないというような分析方法が望ましい。

本方式では複素ケプストラム分析で音声波形を音源情報と声道特性に分離したものに、更に零位相化とその逆の補償を施し、音源情報はパルス化しそれ以外の成分を声道特性に負わせることによって、図1(b)のようにピッチを変化させてもスペクトルのミスマッチが少なくなり高品質なピッチ変換を可能とする。

本方式は図2のように前処理、波形分離部、合成部、後処理の4つの部分に分かれている。

前処理では無音、無声、有声区間の判別およびピッチ区間の判定を行う。

波形分離部では複素ケプストラムを用いて1ピッチ毎の音源情報および声道特性を分離する。

波形合成部ではピッチパルス列に所望の変更を加え(ピッチ変換および時間長制御)、それに基づいて音源情報と声道特性をたたみ込んだ1ピッチ毎の波形を重ね合わせる。

後処理ではピッチ変更や時間長制御による歪みの除去およびゲイン調整を行い、得られた波形と前処理で記録した無音、無声区間と接続することによって最終出力を求める。

3. 本方式における処理

以下図2各部の処理について詳細に説明する。

3. 1 前処理

前処理として、以下のような処理で予め、無音、無声、有声区間の判別およびピッチ区間の判定を行っておく。[1]

無音区間の判別

まず、入力音声をパワーに基づき有声区間と無声区間との判別を行う。

無声区間の判別

次に有声区間に対してPARCOR分析と零交差分析を行い、無声区間と有声区間の判別を行う。

ピッチ区間の決定

有声区間に対しては、まず共分散法による4～5msの短区間線形予測分析を行い、Wongらの方法により声門体積波形を推定する。この波形および原波形をローパスフィルタリングした波形のそれぞれのピーク間隔のうち、正しいと思われる方を採用してピッチ周期を決定する。求められたピッチ周期を基に、原波形のレベルが急に大きくなる点の手前で、かつ、レベルが0に近いサンプル点を開始点とし、次のピッチの開始点の1サンプル手前を終了点として、ピッチ区間を決定していく。

以上のような方法で無声区間の長さ及び無声区間の波形はそのまま記録する。有声区間に対してはピッチ区間を予め決めておく。

これは波形分離部から得られるピッチパルスにより適応的にトラッキングしていくことによっても比較的簡単に実施できる。適応的に行う場合、分析窓の位置と長さを制御してピッチの中心の前後1／2ピッチ、計1ピッチの長さの原音声波形と次に示す分離された単位応答波形の2乗誤差の最小値を与える分析窓を選択することとする。この選択は1ピッチ毎に行なう。窓の選択範囲は、窓長については3ピッチから2ピッチの長さまでの窓を10種類用いた。また窓位置については各窓長についてピッチ中心から1ピッチ長前から半ピッチ長前までの間の5種類の位置を窓の開始点として用いた。したがって1ピッチ毎に50種類の窓について分離を試み、それぞれの単位応答の中から原音声波形に最も近いものを選択する。

3．2　波形分離部

以下図3のフローチャートに沿って波形分離部の説明を行う。

複素ケプストラム分析（図3①）

1)有声音区間内の対象とするピッチ区間に対　して、1ピッチ区間長だけ前の位置からピッチ区間長の 2倍から3倍のハニング窓をかけ波形を切り出す。（図4）

2)そのデータの後半に零詰めを行い、解析ポイント長を1024ポイントにしてＦＦＴを使って複素ケプストラムを求める。[2]

複素ケプストラムを求める際には次のような点に注意する必要がある。

【位相アンラップ】

位相アンラップの方法としてはTriboletの位相微分による方法等いろいろなものが報告されているが[3]、声道特性が微妙に変化する自然音声に適用するといった点や、方法によっては計算量、前提条件などに縛られることを考慮して、今回は次のような極めて簡単な方法をとった。[4]

複素スペクトルから位相を求める。この位相は2πを法とする主値となっている。これから真値を再生するため、真値の隣接した2つのサンプル値間の位相差の絶対値が充分小さくなるように、データに0系列を付加し解析ポイントを増す。これにより、ＦＦＴにより求められるスペクトル周波数間隔が充分密にとれているので、隣合う主値の位相差がπより大きいスペクトル間では位相のラッピングがおこっているとして2πを加減する。

【直線位相成分】

リフタでケフレンシ帯域の一部を除去し波形再生する場合には直線位相による歪みが再生波形に悪影響を及ぼす。そこで、ケプストラムに変換する時に一旦直線位相成分の除去　を行い、リフタで分離した後、時間波形を再生するときに音源情報成分を含む方にこれを再び補償することとした。

櫛形リフタによる分離（図３②）

複素ケプストラムをピッチ周期の整数倍を取り出す櫛形リフタとその逆リフタで分離する。図５のように予め求めておいたピッチ周期にもとづいた間隔の正負のケフレンシ方向に対称なリフタとなっている。[5]

各々の時間波形再生（図３③）

リフタで分離されたケプストラムのそれぞれの時間波形を求める。この際、音源情報を担うものをパルス列、声道特性を担うものを単位応答と呼ぶ。パルス列の方には複素ケプストラムを求める際に除去した直線位相を付加する。ところで、この時点でパルス列と単位応答をたたみ込み演算すると元の切り出し波形に戻すことができる。すなわち情報は失われていない。

パルス列の零位相化、単位応答の逆補償（図３④）

再生されたパルス列には零位相化、単位応答にはその逆の補償を施す。零位相化および逆補償について次に説明する。[6]

【零位相化】

複素ケプストラムをリフタリングすることによって分離抽出されるパルス列のパルスがフレーム内で出現する位置は、音声波形に対する分析窓の位置および長さによって不確定になることがある。

そこで、音声波形に対して、パルスの位置を確定しながらピッチ同期を行うために零位相化処理を行う。これは、また、パルスのピーク化処理でもあるため、ピッチ変更時に行う打ち切りや零詰めの処理による影響をあらかじめ減らす効果もある。

1)分離されたパルス列のフレーム全体の波形をh(n)として、この中で最大のパルスを中心にピッチ区間長のハニング窓でこのパルスを切り出す。これをg(n)とする。

2)対象ピッチ区間の開始点を原点O'としてパルスg(n)からFFTによりの位相成分argG(z)を求める。これをh(n)から周波数毎に除去する。すなわち、h(n)のz変換をH(z)とすると

$$H'(z) = |H(Z)| e^{j(\arg H(z) - \arg G(z))}$$

これにより、パルス列は時間領域でシフトされ、最大パルスが所望のピッチ区間開始点O'に一致し、エネルギーがO'に集中する。ここで、この最大パルスの零位相化は時間領域でのシフトだけではなく波形のピーク化の効果も持っている。さらにフレーム内の次のパルスのピーク化も行っているのでピッチトラッキングが容易になる。

【逆補償】

一方、合成時に正しい波形を得るには1フレームに対して得られる単位応答に零位相化処理の逆の補償を施してやらねばならない。すなわち、単位応答f(n)のz変換F(z)に対して、

$$F'(z) = |F(Z)| e^{j(\arg F(z) + \arg G(z))}$$

を行う。この操作により、逆補償した単位応答f'(n)と零位相化したパルス列h'(n)を畳込む合成時に元の波形が再生できる。

ピッチパルスと単位応答の記録（図3⑤）

零位相化パルス列のフレームから対象ピッチ区間の開始点を中心としてピッチ区間長のハニング窓で最大パルスを切り出し、このピッチパルスをフレ

ームを代表する駆動音源とし、逆補償された単位応答とともに記録しておく。分析区間を移動して①から⑤の処理を続ける。以上の操作で各フレームにつき音源情報としての零位相化されたピッチパルスと声道特性としての逆補償された単位応答波形を得る。ここで零位相化されたピッチパルスはそれぞれのフレームでピッチ区間の開始点に位置しており、フレーム毎に並べていくことにより、ピッチパルス列を構成する。これを試聴しても予測残差波形とは異なり、音韻性がまったく感じられない。

3．3　波形合成部

次に図6のフローチャートに沿って波形合成部について説明する。

ピッチ変更（図6⑥）

ピッチ区間毎にピッチ周期に所望の変更を加える。このままだと各ピッチ区間に対するピッチ変換倍率 r が $r<1$ の場合は必然的に時間長が短くなり、$r>1$ の場合は時間長が長くなる。

時間長制御（図6⑦）

ピッチ変更後の発声速度を保つため、ピッチ区間毎に間引きと繰り返しによって時間長に所望の変更を加える。1カ所で集中して間引きや繰り返しの操作を加えると波形の連続性が崩れてしまうので、一様変換の場合にはピッチ変換倍率 $r<1$ では $r/(1-r)$ ピッチに一回繰り返し、$r>1$ では $r/(r-1)$ ピッチに一回間引きの操作を加える。[7]

逆に、発声速度を変化させる場合には母音区間で間引き及び繰り返しの比率を変えて行う。

ピッチパルスと単位応答の畳み込み（図6⑧）

各フレームを代表するピッチパルスと単位応答をピッチ区間毎にたたみ込む。この際、各ピッチパルスと単位応答の関係が分析時の対応関係を崩さないようにする。これが、各ピッチ区間を代表する1ピッチ分の時間波形となる。

重ね合わせ（図6⑨）

前段で得られた時間波形を周期や時間長を変えて得られたピッチパルス列に基づいて重ね合わせ合成音声を得る。

今回、位相アンラップには極めて簡単な方法を用いているので、１フレームについて数ヵ所のアンラップミスが生じている可能性がある。そのためフレーム内での周期的な成分と声道特性の分離にも誤りが生じ、各フレームを代表するピッチパルスと単位応答をたたみ込んだ時間波形には、この誤分離に起因する裾引きが発生することがある。（計算量は増えるが、適応的な窓位置、窓長といったものを採用することによって、音源情報と声道特性の誤分離は解消できる。）そのために、重ね合わせる際に、特に同じ波形を間引いたり、繰り返す場合に接続部で歪みを生じ相関性のノイズが発生する可能性がある。そこで、この相関性のノイズを抑えるために合成窓をかける。

各ピッチ区間に対して、それに同期させてピッチ変更後のピッチ区間長の１ピッチ分前をスタート点として長さがピッチの2.5倍のハニング窓をかけて裾引きを抑えて、隣合ったフレームの波形を重ね合わせる。これにより波形の連続性が保たれ、スムーズな音声が得られる。

この合成窓については、予備実験によって２倍から３倍の範囲の内から2.5倍の窓を選んだ。

３．４　後処理

ここでは、重ね合わせで得られたピッチ変換および時間長制御を施した合成音声に対して、ピッチ同期ではなく、一定のフレーム周期Ｌで歪みの軽減およびゲイン調整を行う方法について述べる。ただし、ここでＬは概ね対象とする音声波形の平均ピッチ区間長に近い値とする。また、以下の説明において、Ｎは512ポイントである。[1]

まず、Ｎサンプルのデータを矩形窓で切り出し、この波形を x(m)とする。それとともに、その中心から前後Ｌサンプルの範囲のデータの自乗和Ｐ₀を求めておく。

櫛形ろ波による歪みの軽減

先に述べたようなピッチ変更による歪みを調波構造の上で軽減するために櫛形ろ波を行う。[8] x(m)に対して、次式の櫛形ろ波を行いx'(m)をもとめる。

$$x'(m)=h_1 x(m-k_p)+h_0 x(m)+h_1 x(m+k_p) \qquad m=1,\cdots,N$$

ここでK_pはフレーム内におけるピッチ変更後の平均ピッチ区間長であり、$i \leqq 0$または$i > N$の場合の $x(i)$の値は0とする。また、

$h_0 = 0.5(1+\alpha)$

$h_1 = 0.25(1-\alpha)$

である。αは櫛形ろ波の谷の深さを決める定数であり、ピッチの変化量が大きいほど小さい値にする。

ゲイン調整

櫛形ろ波により得られた $x'(m)$のＮ点の波形の内、中心の２Ｌ点の範囲のデータの自乗和P'_0を求め、これが先に求めたP_0に等しくなるようにゲイン調整を行う。この２Ｌ点のデータをハニング窓で切り出し、これを記録しておく。

フレームをＬポイント後ろへシフトし、同様な一連の処理を行った後、２Ｌ点の波形の前半のＬ点と直前のフレームの後半のＬ点とを重複加算する。

以下有声区間が終るまで同じ操作を繰り返す。

以上のような後処理によって複素ケプストラムで分離しきれなかった成分による、ピッチ変更時の聴感的な歪みが軽減され、ゲインの調整も行われる。

このようにして得られたピッチ変更された波形を前処理で記録した無音、無声部分と接続することにより出力音声を得る。

４．ピッチ変更音声の品質評価

単語音声を対象に、本方式を含む３つの方式でイントネーションを変えずに単語全体にわたって一様に平均ピッチを変更した音声を作成し、聞き取り実験によってその音質の心理評価を行い、本手法の有効性を検討した。

４．１　ピッチ変換音声の作成

実験に用いた方式は残差駆動、零位相化残差駆動および本方式の３つの方式である。

ピッチ変更の操作は有声音区間のみであり、無声子音区間および無声区間は原波形をそのまま接続した。ピッチ周波数は単語全体にわたって一様に、一単語につき原音声に対して±１オクターブの変化範囲を0.2オクターブずつ、原音声の高さを含めて、１方式について11個、３方式で計33個のピッチ変換音声を作成した。

残差駆動、零位相化残差駆動方式の場合には対象ピッチ区間について20mSのフレームで自己相関法を用いた18次の線形予測分析により、声道フィルタおよび予測残差波形を各ピッチ区間毎に求めた。

３方式とも発生速度が変わらないように適宜ピッチ区間を間引くまたは繰り返すことにより時間長を制御した。

以下各方式について記す。

残差駆動

各ピッチ区間について得られた残差波形の後半部分を直接打ち切るか零詰めを行うことによりピッチ周期を変更し、この残差波形で同一のピッチ区間について得られた声道フィルタを駆動した。

零位相化残差駆動

予測残差波形に対して各ピッチ区間毎に零位相化を行い、それを駆動音源として上と同様にピッチ周期を変更し、声道フィルタを駆動した。

本方式

３．で述べたとおりである。

４．２　音声の品質評価実験方法

３方式による一様ピッチ変換音声に対して歪み感や雑音感に着目して評価を行った。

評定者は成人男女計６名であり、１音声について１評定者が10回の評定を行った。音声は防音室内おいてスピーカーで呈示した。

評価実験開始時にまず原音声を繰り返し提示した後、３方式について各11段階の高さの一様ピッチ変更音声33個の音声と原音声を合わせた計34個の音声をランダムな順序で呈示した。

評定者はそれぞれについて、実験開始時に聴取した原音声に対する品質の相

対劣化度を７段階で評価した。[9]

４．３　評価実験結果

評価結果を図７に示す。(a)は母音のみ、(b)は無声、有声子音を含む音声に対する結果であり、話者 M1、M2 は男声、F1、F2 は女声である。

横軸は平均ピッチ周波数の変化量をオクターブ単位で表したものである。0oct.はピッチ不変で分析合成系を通過する場合を示し、±１オクターブの範囲で0.2オクターブ刻みになっている。

横軸は全評定者による平均評価値を系列範疇法で処理し、原音声を0とした場合の心理的距離である。値が大きいほど良好な品質であることを示す。また評定のカテゴリーの境界を細い実線で示し、評価値を右側に記した。

例えば、原音声の高さ0オクターブに関しては、本方式と残差駆動方式は心理的距離のスケールで原音声と1以上は離れていないので、原音声とどちらが優位であるかは判断できないことを示している。

それぞれのグラフを見比べてみると、話者や言葉の違いによって多少異なるが、全体的にみて、±0.6オクターブぐらいまでは本手法が残差駆動や零位相化残差駆動法に比べて優位である。

５．むすび

本方式によれば、従来の残差駆動におけるピッチ変換時に生ずる予測残差の打ち切り、零詰めによる品質劣化に相当する歪みを防ぐことができる。

また、平均ピッチ周波数変更合成音に関して行った心理評価実験の結果により本方式の有効性について検討した。

参考文献

[1]都木、梅田：”ピッチ変更時の歪をスペクトル領域で補正する声質変換”、信学技報、SP87-111(1988-01).

[2]A.V.Oppenheim, " Digital Signal Processing", Englewood Cliffs,NJ: Prentice-Hall,1975.

[3] J.M.Tribolet, " A new phase unwrapping algorithm",IEEE Trance.Vol. ASSP-25,No.2,pp.170-177,April 1977

[4]森下巌、小畑秀文：”信号処理”、pp.158-173,計測自動制御学会（昭57）

[5]A.V.Oppenheim,R.W.Schafer and T.G.Stockham,Jr.:"Nonlinear Filtering of Multiplied and Convolved Signals"IEEE Trans.Vol.AU-16,No.3,pp.437-466,September 1968

[6]清山信正、梅田哲夫：”複素ケプストラムによる周期成分波形を零位相化することによる基本周波数抽出法”、音学講集2-3-14,pp.257-258(平元-10).

[7]清山信正、桑原尚夫、梅田哲夫：”複素ケプストラム分析合成方式における時間長制御”音学講集3-4-8,pp.283-284(平2-03).

[8]J.S.Lim, A.V.Oppenheim and L.D.Braida: "Evaluation of an adaptive comb filtering method for enhancing speech degradated by white noise addition"IEEE Trans.Vol.ASSP-26,No.4,pp.354-358,August 1978.

[9]都木徹、梅田哲夫：”ピッチ変更時の歪をスペクトル領域で修正する声質変換方式とその品質評価”,信学論誌(A),J73-A,No.3,pp.387-396,(平2-03)

[10]梅田哲夫：”ＬＰＣ残差波形の零位相化による基本周波数抽出法”,音学講集 - - ,pp.211-212,(昭62-10).

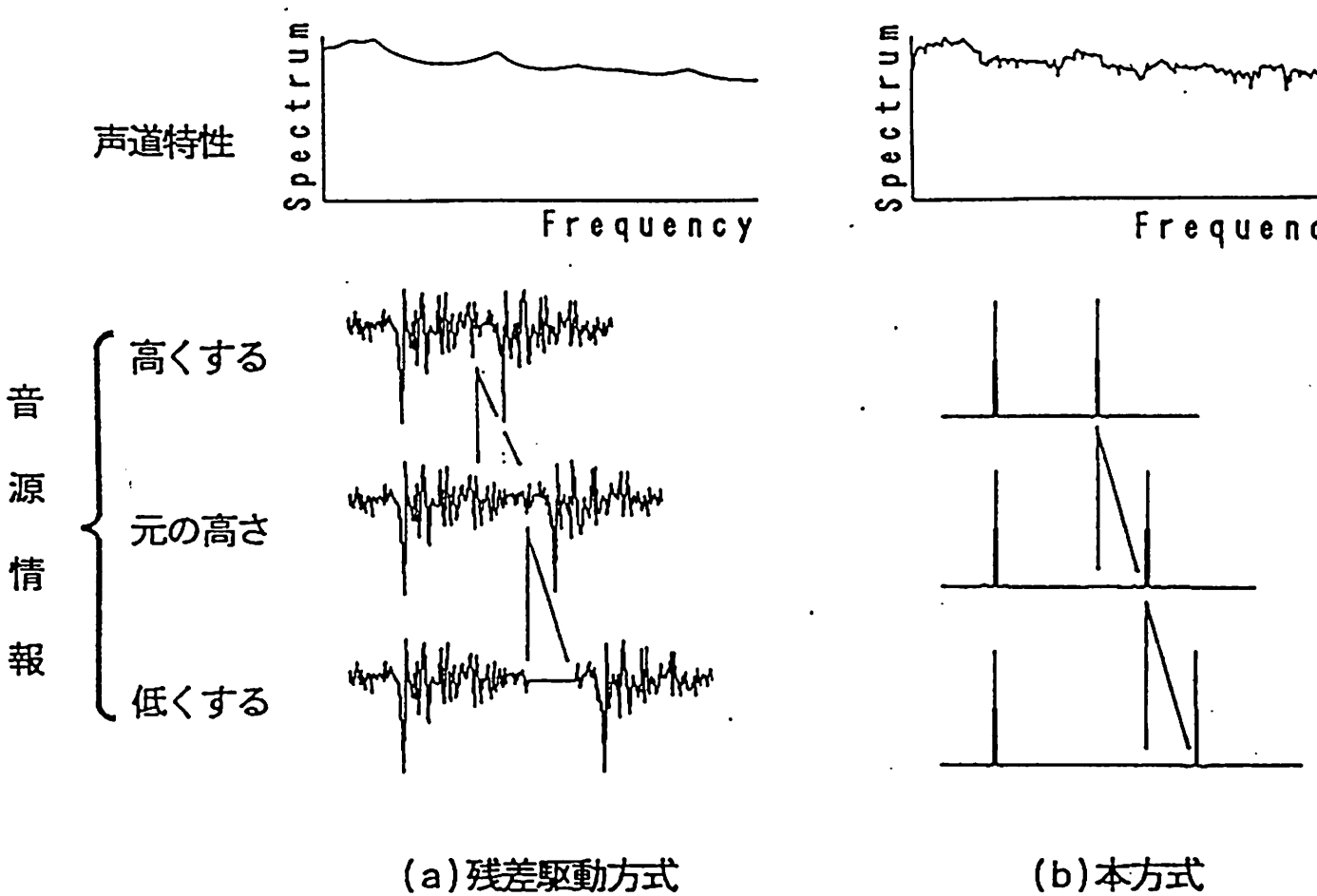


図1　残差駆動方式と本方式の比較

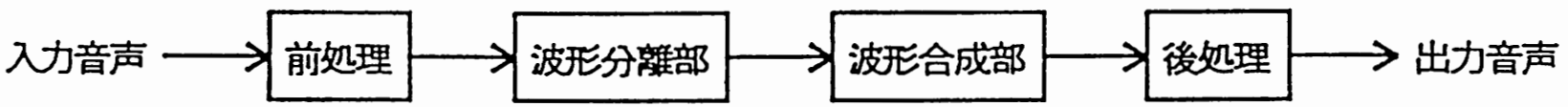


図２　本方式の概略

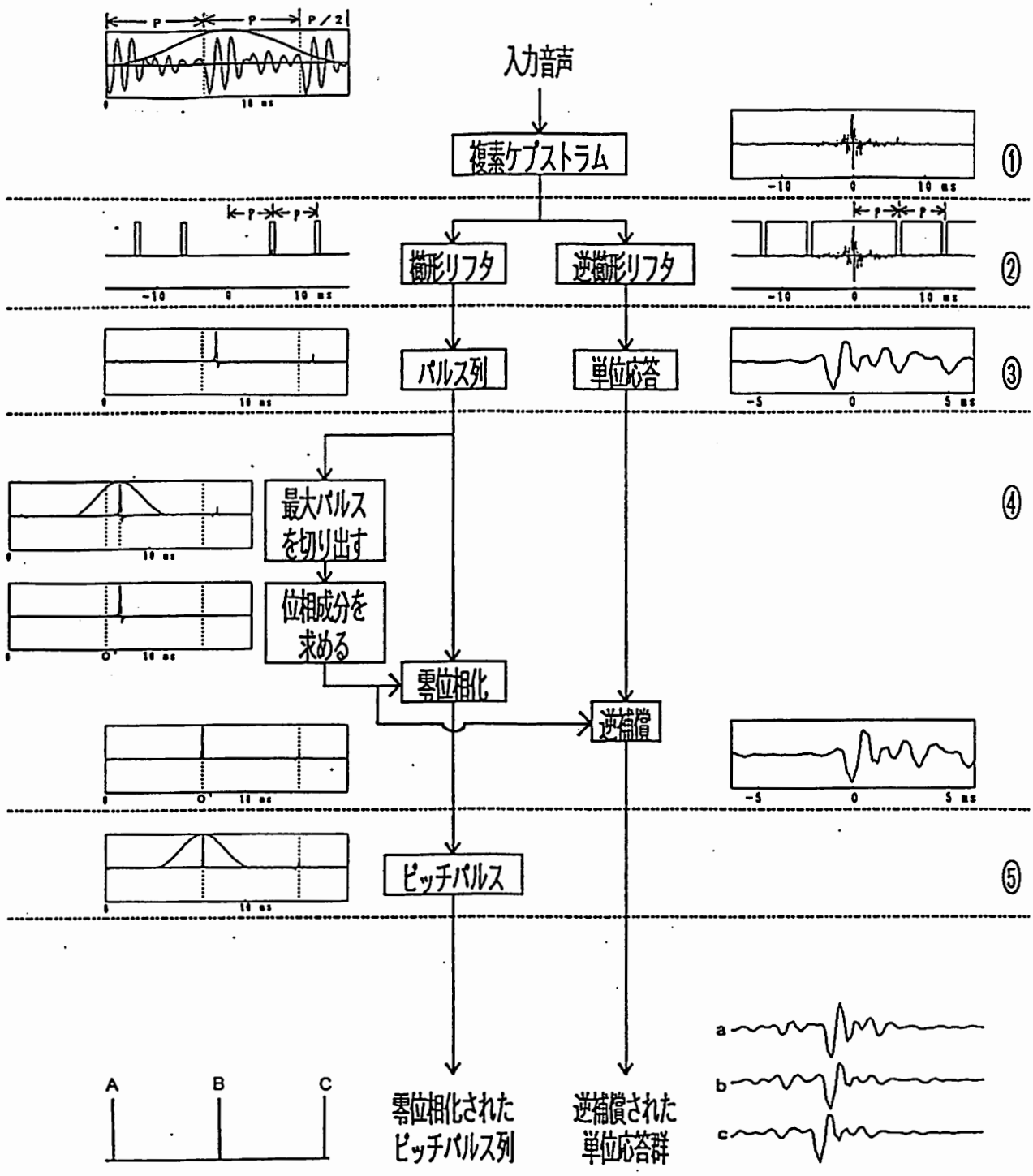

入力音声
複素ケプストラム
櫛形リフタ
逆櫛形リフタ
パルス列
単位応答
最大パルスを切り出す
位相成分を求める
零位相化
逆補償
ピッチパルス
①
②
③
④
⑤
A
B
C
零位相化されたピッチパルス列
逆補償された単位応答群
a
b
c

図３　波形分離部

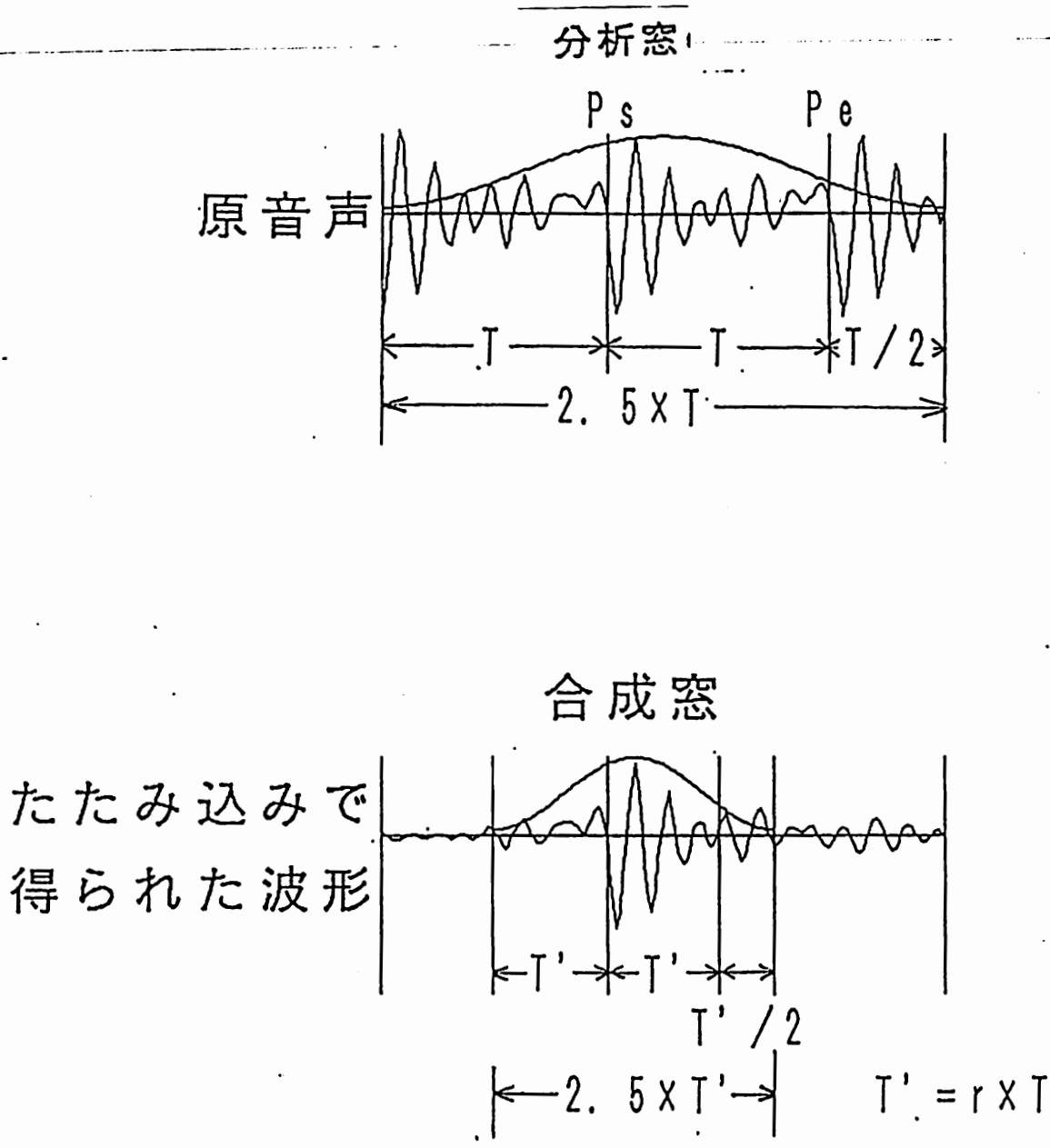


図４　分析窓のかけ方

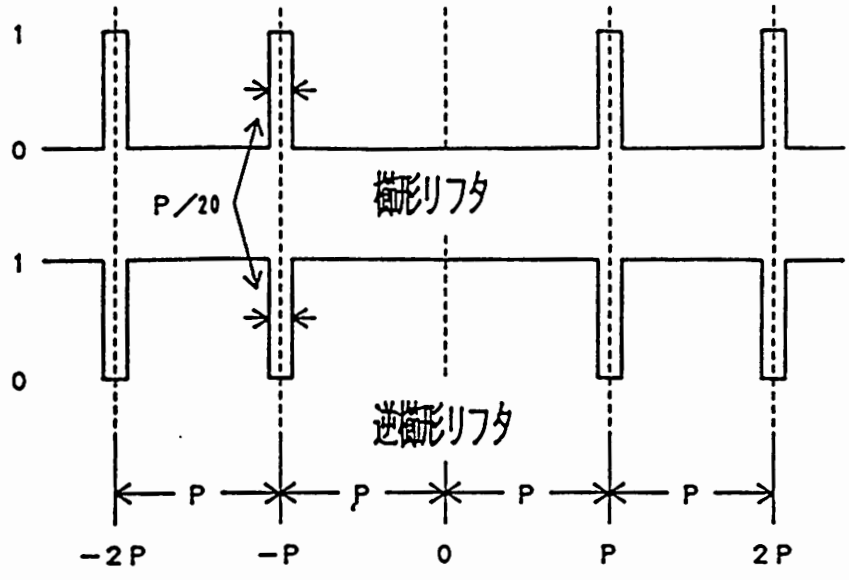


図５　櫛形リフタと逆櫛形リフタ

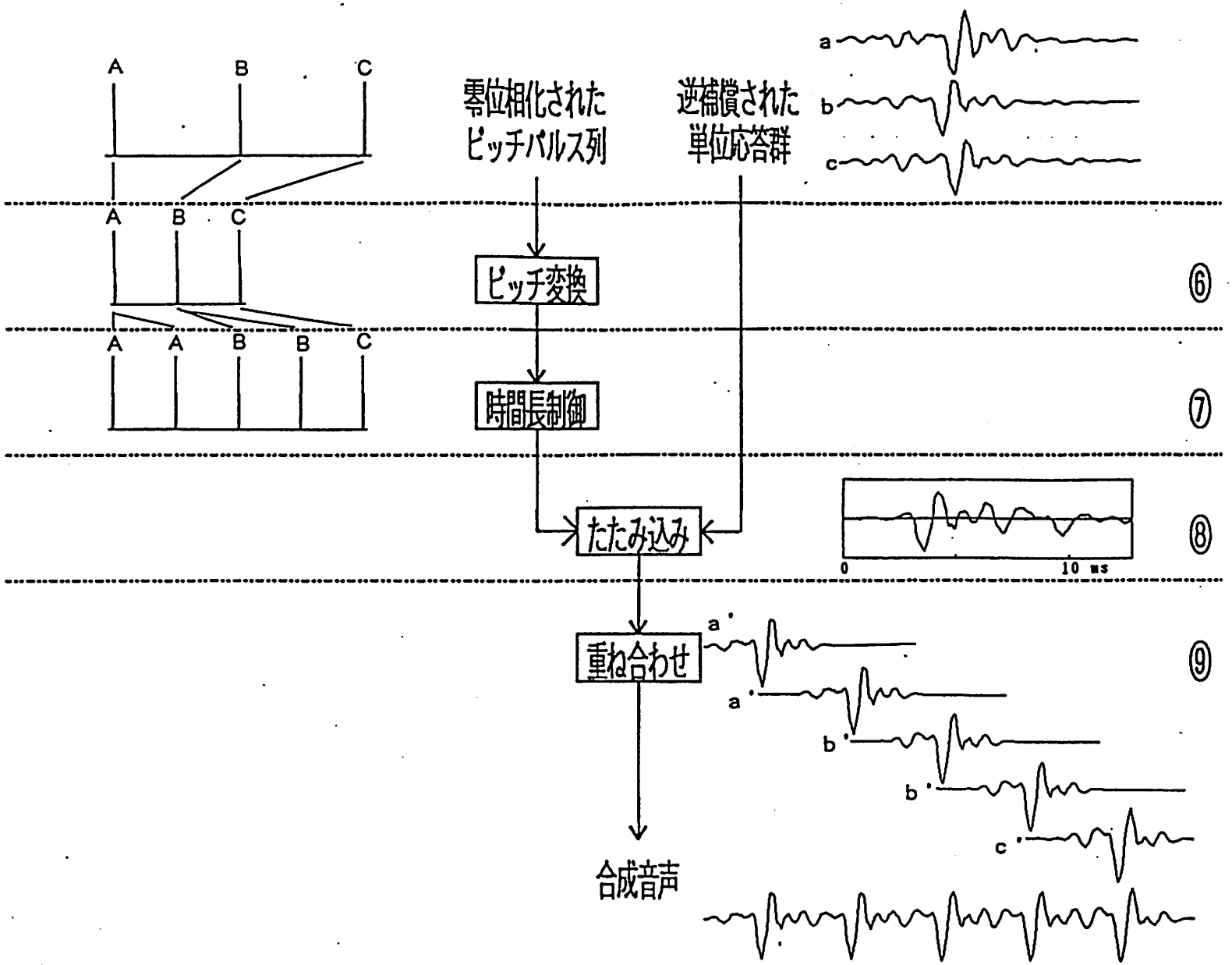
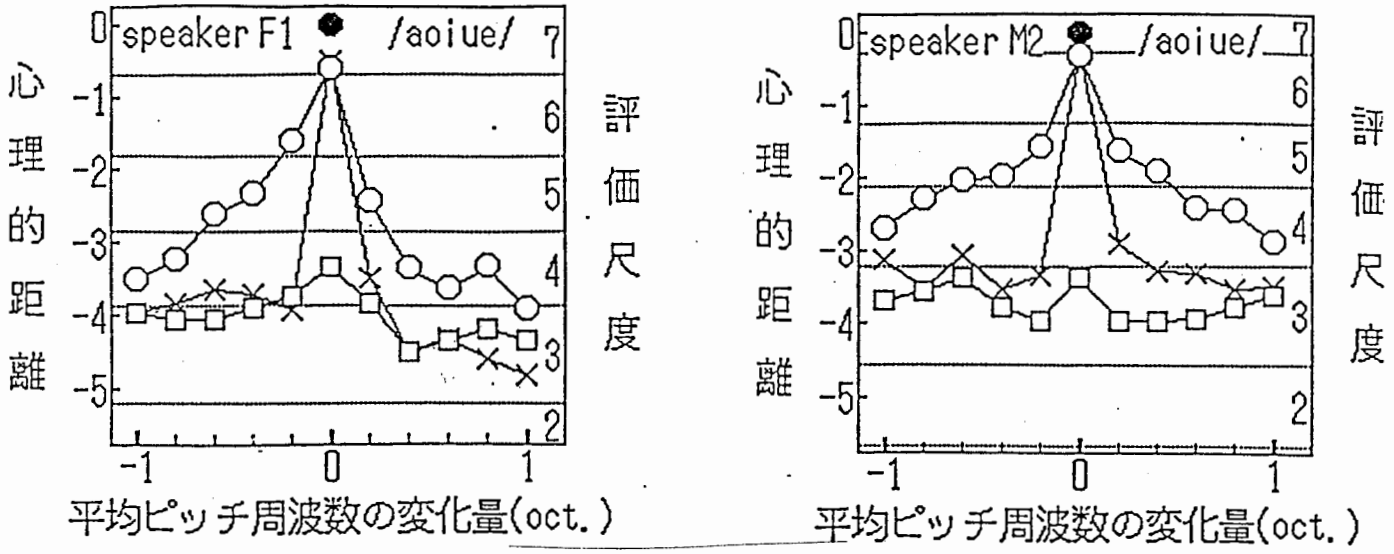
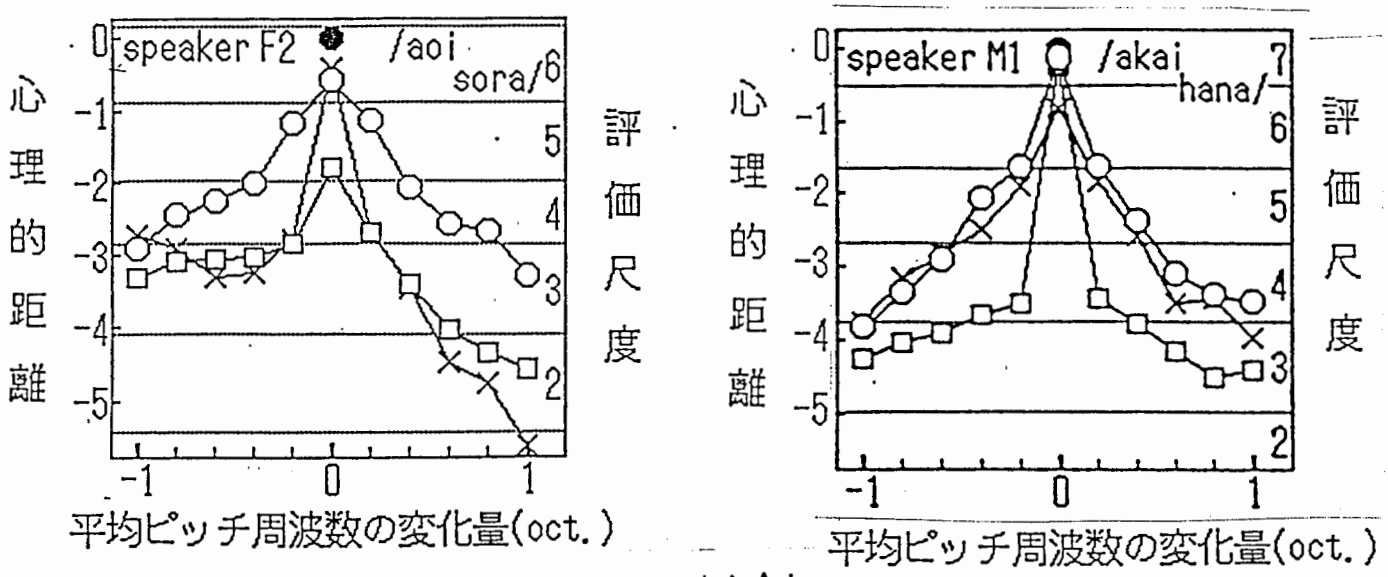

A
B
C
零位相化された
ピッチパルス列
逆補償された
単位応答群
a
b
c
ピッチ変換
⑥
時間長制御
⑦
たたみ込み
⑧
0
10 ms
重ね合わせ
⑨
合成音声

図６　波形合成部



(a)母音のみ



(b)子音も含む

● 原音声
○ 複素ケプストラム
□ 零位相化残差
× 残差駆動

図7　品質心理評価結果