高品質音声合成のための
スペクトル包絡の推定
及び変換に関する研究

Studies on Spectral Envelope Estimation
and Conversion
for High Quality Speech Synthesis

2006年2月

早稲田大学大学院 理工学研究科
情報・ネットワーク専攻
知覚情報システム研究

望月 亮

目 次

図 目 次

表 目 次

第1章　序論

1.1　背景

現在，任意のテキストを音声によって読み上げる音声合成システムは，ユーザの所望する情報を音声によって伝達する手段として活用されている．例えばカーナビゲーションシステムにおける目的地や周辺情報の案内，電子メールや Web ページの読み上げ，コールセンターでの CTI（Computer Telephony Integration）システムにおける自動応答など，近年ではその実用化の場面も増えている．音声合成によるテキストの読み上げが検討されるようになった 70 年代から 80 年代にかけては，音声をパラメータ化し，規則によって生成された韻律パタンに沿って音声を合成する，いわゆる「規則合成」が主流な方式であった．当時，音声合成システムを実現する音響処理技術としては，LPC（Linear Predictive Coding）[板倉 70] を代表とするパラメトリックな合成方式が盛んに検討されたが，その音質は不明瞭で，人間が発声する音声からはほど遠いものであった．80 年代後半になると，音声波形をパラメータ化せず，原波形レベル（またはそれに相当するレベル）で保存し，必要に応じて韻律変更を行うノンパラメトリックな合成方式が検討されるようになった．このアプローチによって合成音声の明瞭性は大幅に改善され，従来の機械的な音色に代わり，発話者の個人性が再現できるレベルになった．例えば PSOLA（Pitch Synchronous OverLap Add）法 [Moulines 90] はその代表的な方法であり，処理が簡単な上に，基本周波数の変更が小さい場合は音質の良い韻律変換が実現できた．近年では計算機の処理能力や記録媒体の性能向上に伴い，大量の音声データを取り扱うことができるようになった．そのため 90 年代半ばからは，大量の音声データを利用したコーパスベースの音声合成が主流となり，その音質は改善され，テキストの読み上げなどの用途では肉声感のある音声の合成が可能となった．

特に大規模な音声コーパスを用い，韻律変換をまったく行わない波形接続合成方式 [Black 95, Campbell 96] は，自然音声と比べてほとんど遜色の無い合成が可能である．

一方，合成によって高品質のテキスト読み上げが実現できるようになると，音声合成の次のターゲットとして，感情や態度，話者性，発話口調を自由に表出するための技術が要望されるようになった．例えば音声合成を音声対話システムの応答に使用する場合，ユーザとシステムとの自然なやり取りを実現するためには，単なる読み上げ口調ではなく，システムの発話意図や態度などを表出するための多彩なパラ言語表現が必要となる．また，アプリケーションによっては一つのシステム上で複数話者の音声を合成したいなどの要望がある．このため，90 年代後半になると，発話者の変換や音色・発声スタイルなどに多彩さを持たせるための取り組みが盛んに検討されるようになった．

音声合成によって多様な発話スタイルの合成を実現する手段としては，(1) 発話スタイルごとに音声コーパスを収録する，(2) 学習によって適応する，というアプローチが考えられる．前者のアプローチでは，波形接続合成方式を用いることで非常に音質の良い合成を達成できるが，発話スタイルごとに十分なカバレッジがある音声データベースを構築する必要があり，録音やラベル情報の付与に膨大な人手の作業が発生することを考えると効率的なアプローチとは言いがたい．そこで限られた音声データで発話の多様化を目指す後者のアプローチを考える．現時点では十分な適応・変換方法が存在しないため，変換処理によって音質劣化が際立ったり，ターゲットへの変換が不十分だったりという問題がある．しかし，この問題は今後検討が進むにつれて改善されることが期待できる．

今後，ユーザへの情報提供や機械とのインタフェースとして，ますます多くの場面で音声合成の利用が期待される．多様化が進むアプリケーションの中で，音声合成に対するユーザの要望を満たすためには，品質の高い音声を合成することは必要最低条件であり，加えて，合成に使用する音声データベースの制約を受けず，自由自在に多様な発話スタイルの合成を実現する技術が必要となってくる．このため，適応や変換処理によって表現の自由度を高めることができ，信号処理による音質劣化が極力発生しない音声合成方式が強く望まれる．

1.2　従来の合成方式

これまでに音質改善や表現の多様化を目的とし，数多くの音声合成に関する研究が進められてきた．合成音声の音質を向上させるためには，流暢で自然なイントネーションの発話を可能にする韻律パタンの推定も重要であるが，それに劣らず，合成音声独特の「ざらつき」や「こもり」などを無くすために，合成時の信号処理によって生じるスペクトル歪を減らすことが重要である．また，発話者の個人性や発話のスタイルを再現する場合も，アクセントや話速，イントネーションなどの韻律に関する特徴量の制御に加え，声質を決める特徴量，すなわちスペクトルを正しく再現することが必要である．そこで言語解析や韻律制御など，多岐にわたった音声合成に関連する技術の中で，本研究では実際に波形の生成・合成を行う音響処理技術に着目する．ここでは特に音質改善や発話の多様化を目指す上で重要なスペクトルの推定・制御技術について，従来の取り組みを考察する．

70 年代，音声合成を実現する技術として，線形予測分析 [板倉 70] が盛んに検討された．この方式は音声の生成モデルを信号処理で扱えるように一般化した代表的な方式であり，音声信号を入力の音源と，調音部を表す声道フィルタとに分離して考えるため，Source-Filter model とも呼ばれる．線形予測（LPC）による音声の分析は，フォルマントの抽出など，スペクトルの典型的な特徴を捉えるのに適した方法であり，パラメータ化するという点では非常に効率の良いデータ圧縮が可能である．このため，現在では音声符号化技術として，例えば携帯電話などのコーデックに応用されている．実際，この LPC を音声合成に利用する場合は，LPC 係数の代わりに PARCOR（PARtial autoCORrelation）係数や補間特性の優れた LSP（Line Spectrum Pair）パラメータ [板倉 79] が用いられる．LPC 分析で得られるスペクトルパラメータは，典型的なスペクトル形状を表す情報のみを持ち，微細構造はすべて音源情報に割り振られる．すなわち，LPC によって自然な音声を再現するためには，合成時に線形予測誤差（音源信号）を再現する必要がある．しかし，LPC による音声合成が盛んに検討された時代は，計算機や記録媒体などの制約により残差信号を何らかの手法でモデル化し，情報圧縮するのが一般的だった．このような背景から，LPC をベースとした合成方式では，合成時に詳細なスペクト

ル構造や揺らぎ情報が正しく再現されず，十分な品質の音声を合成できなかった．

準同型分析 [Oppenheim 69] によってケプストラムを求め，これをインパルス応答波形として合成に利用するケプストラム合成は，LPC を用いた合成方式と同じく，70 年代に検討が進められた方式である．この方式は，ある程度の長さを持つ窓関数で抽出した音声信号に対して，周波数分析した場合に観測される基本周期のハーモニクス成分を，ケフレンシー領域において取り除くことで滑らかなスペクトル包絡を得る．このケプストラム法によって音源と調音部とに分離した音声の生成モデルを考える場合，ケフレンシー軸における高次成分は音源信号に相当し，低次成分は声道特性に相当する情報とみなせる．しかし音声合成目的で利用する場合，音源はインパルス列を用いるのが一般的である．このため，ケプストラムによる合成も LPC の場合と同様，スペクトルの微細構造が失われてしまい，その音質は「こもり」や「ざらつき」を伴うものであった．

従来のケプストラム分析によって得られるスペクトル包絡は，基本周期成分を取り除くことで得られる包絡であるのに対して，PSE 法 [中島 88] は信用できるスペクトル情報が F_0 の整数倍の周波数にのみ存在することに着目し，この F_0 の高調波のピークを曲線で結ぶことによってスペクトル包絡を再現する方式である．また，通常の PSE 法では抽出が困難であった高域における高調波のピークについて，近似精度を改善した改良 PSE 法 [Tanaka 97] も検討されている．これらの方式は，安定したスペクトル特徴量を獲得するために，ある程度の長さを持つ分析窓によって波形抽出する必要があるが，分析窓長を長くすると特徴量が平滑化されるという問題が発生する．一般的に分析に用いる窓長とシフト幅を固定したフレーム分析では，特徴量抽出の安定性と音質とがトレードオフの関係にあり，その最適化が一つの課題となっている．

上述の合成方式に共通した課題として，分析過程においてスペクトルの微細構造が失われ，音質が劣化するという問題がある．すなわち，LPC を用いた方式の場合は残差信号をモデル化すことで，ケプストラムを用いた方式の場合は高次のケプストラム係数を取り除くことで，このような損失歪が発生する．また，分析対象の波形を抽出する際に少し長めの窓関数を用いると，スペクトル包絡の抽出は安定するが，特徴量が平滑化されるという問題が発生する．固定長の分析窓を用い

る場合，分析対象の音声の F_0 が低い場合でも数ピッチの周期波形が含まれるように，少し長めの窓関数を用いる．これは声道特性の変化が時間に対して緩やかな変化であることを仮定しているためであるが，実際の音声では数周期の間に F_0 が極端に変化する場合もあるため，固定長の分析窓でスペクトル変化のない定常区間のみを抽出するのは困難である．このような問題に対して，80 年代後半から検討が進められるようになった PSOLA 法は，非常に短時間の窓関数を利用し，ピッチ同期のフレームワークによって合成処理を進める方式である．この PSOLA 法では，当初，基本周期の 3 倍以上の長さを持つ分析窓によって波形を抽出し，周波数領域での補間によりスペクトル包絡を推定する方法が検討されていた [Charpentier 86]．一方，周波数領域でのスペクトル包絡推定を必要としない方法として，時間領域で直接合成に使用する短時間波形を獲得する TD（Time Domain）-PSOLA 法が検討されるようになった．この方法では基本周期の影響を含まない短時間波形を時間領域で得るために，基本周期の 2 倍という短い窓長のハニング窓を用いている [Hamon 89]．すなわち，声帯の 1 振動における応答波形を直接抽出することで，長めの窓関数を用いた場合に生じるスペクトル包絡の平滑化の問題を避けられる．この TD-PSOLA 法は，そもそも分離の困難な音源と声道特性とをあえて分離せず，抽出した短時間波形をそのままインパルス応答波形として用いることから，モデル化を行わない方式という意味で Null model，またはノンパラメトリックな合成方式と呼ばれる．このノンパラメトリックなアプローチによって生成された合成音声は，それまでのモデル化を行った合成方式と比較して格段に音質が良く，韻律の変更が小さい場合は，肉声感が再現できるレベルに至った．

一方で，モデル化は行うが，パラメータ化を行わないことで音質の良い合成を実現した方式も存在する．音声信号を複数の周期と位相の異なる正弦波の重み付け加算で表す Sinusoidal model [Quatieri 86] は，誤差最小化基準によって正弦波の振幅，周波数，位相パラメータを推定し，韻律変換を行う方式である．この方式はフーリエ変換による周波数分析を用いた場合と比べて，分析に使用する窓関数の影響を直接受けない．このため，各周波数成分の振幅推定が精度良く行え，短時間の分析シフトを用いることで，高品質の合成を実現している [Macon 96, George 97]．また，スペクトルを強い周期性が観測される低域成分と，非周期成分が支配

的である高域成分とに分離し，低域は Sinusoidal model によってモデル化し，高域は AR フィルタとノイズでモデル化する合成方式 [Stylianou 01] は，TD-PSOLA 法に勝る音質を実現している [Syrdal 98]．相補的な窓関数を用いて滑らかなスペクトル包絡を抽出し，聴覚的な知見に基づいて設計したオールパスフィルタによって音源を再現する合成方式 [Kawahara 99] では，シフト幅の細かいフレーム分析によって音質の良い合成を実現している．

上述で紹介した合成方式は，いずれも信号処理によって韻律変換を行う方式であるが，その中でも，ピッチ同期，または基本周期より細かい単位で分析処理を行い，加えて，特徴量のパラメータ化を避けた合成方式は，比較的高品質の合成を実現している．これらの方式は細かい単位で合成処理を行うため，スペクトル特徴量に対して詳細な適応や変換処理も期待できる．このため，現時点では波形接続合成方式と同レベルの音質は実現できないものの，将来，合成によって自由度の高い発話表現の実現を視野に入れると，これらの合成方式に対して，適応や変換処理を考慮しながら，音質改善に関する取り組みを進めて行くことは重要だと考えられる．

1.3　本研究の目的

現在，非常に音質の良い合成が可能な波形接続合成方式は，大規模な音声コーパスを使用し，韻律変換を行わないことで，信号処理によって生じる音質劣化を避けた方式である．しかし，この方式で複数の発話スタイルの合成を実現するためには，発話スタイルごとにデータベースの構築を行う必要があり，その作業は膨大な手間とコストがかかるため，現実的なアプローチとは言いがたい．限られた音声データで発話スタイルの制御・多様化を目指すという観点からは，少なくとも適応や変換処理が施せるレベルまで「音声信号処理」に踏み込んだ合成方式を検討する必要がある．この条件を満たす合成方式の一つとして，PSOLA 法が挙げられる．PSOLA 法は波形接続合成方式より韻律変換が可能という点で自由度が高く，特に変換率が低い場合は従来の線形予測を代表とするパラメトリックな合成方式よりも格段に音質が良いという長所を持つ．そこで本研究では，高品質の

音声合成が期待できる PSOLA 法をベースに，音質の改善，及び多彩な発話表現の実現に必要不可欠な要素技術を提案・検討する．

1.4　TD-PSOLA 法

PSOLA 法は当初，周波数領域でスペクトル包絡を抽出する方式 [Charpentier 88] が検討されていたが，検討が進むにつれ，時間領域で波形抽出する方式や，LPC と組み合わせて残差波形に対して処理を施す方式 [Edgington 96] など，いくつかのバリエーションが派生した．本研究では，音源入力をインパルスと仮定したとき，そのインパルス応答に相当する単位波形を時間領域で抽出する TD-PSOLA 法に着目する．

図 1.1 を用いて，TD-PSOLA 法による韻律変換処理を簡単に説明する．まず，原音声波形に対して，ピッチ同期分析を行うための基準位置となるピッチマークを付与する．従来では原波形レベルでのローカルピークをピッチマークとして用いるのが一般的である．続いて，基本周期の 2 倍の窓長を持つハニング窓を用いて単位波形の抽出を行う．この際，窓関数の中心がピッチマークに合うようにして波形抽出を行う．この単位波形抽出処理は，有声区間におけるすべてのピッチマークに対して行う．次に，この抽出した単位波形列を新たに所望する基本周期で重畳加算することによって F_0 変換音声を作成する．図 1.2 に示すように，合成音声のピッチを高くする場合は基本周期の間隔を短くすることになる．このとき，時間長を変更しない場合は同じ単位波形を繰り返し配列することで，元の時間長を維持する．逆にピッチを低くする場合は基本周期の間隔を長くして単位波形の配列を行う．元の時間長を保つ場合は，余分な単位波形を間引きすることになる．すなわち，PSOLA 法における F_0 の制御は，再配列する単位波形の間隔を変更することで行い，時間長の制御は単位波形の繰り返しや間引き配列によって行う．振幅に関しては，変換処理後の音声のエネルギーが，変換前のエネルギーを保存するように補正する．なお，無声子音や無声化母音など，ピッチマークが定義できない区間に関しては，固定長のシフト幅で便宜的にピッチマークを定義し，上述の要領で時間長の制御のみを行う．

本研究では，基本的に上述の TD-PSOLA 法に従い韻律変換を行う．従来の PSOLA 法では，単位波形の抽出に用いる窓関数とは別に，合成の際にも窓掛けを行う方法が検討されていたが，本研究では時間領域で抽出した単位波形をそのまま利用する．なお，原音声の基本周期に応じて窓長を決定する代わりに，合成ターゲットの基本周期に合った窓長を用いて単位波形を抽出する方法も考えられるが，ピッチを高い方へ変換する場合，元の基本周期の 2 倍よりも短い窓幅で単位波形を抽出すると，スペクトル歪が大きくなるものと考えられる．そこで本研究では，ピッチマークを基準にして，その前後のピッチマークまでを窓幅とする非対称のハニング窓を用いて単位波形の抽出を行う．

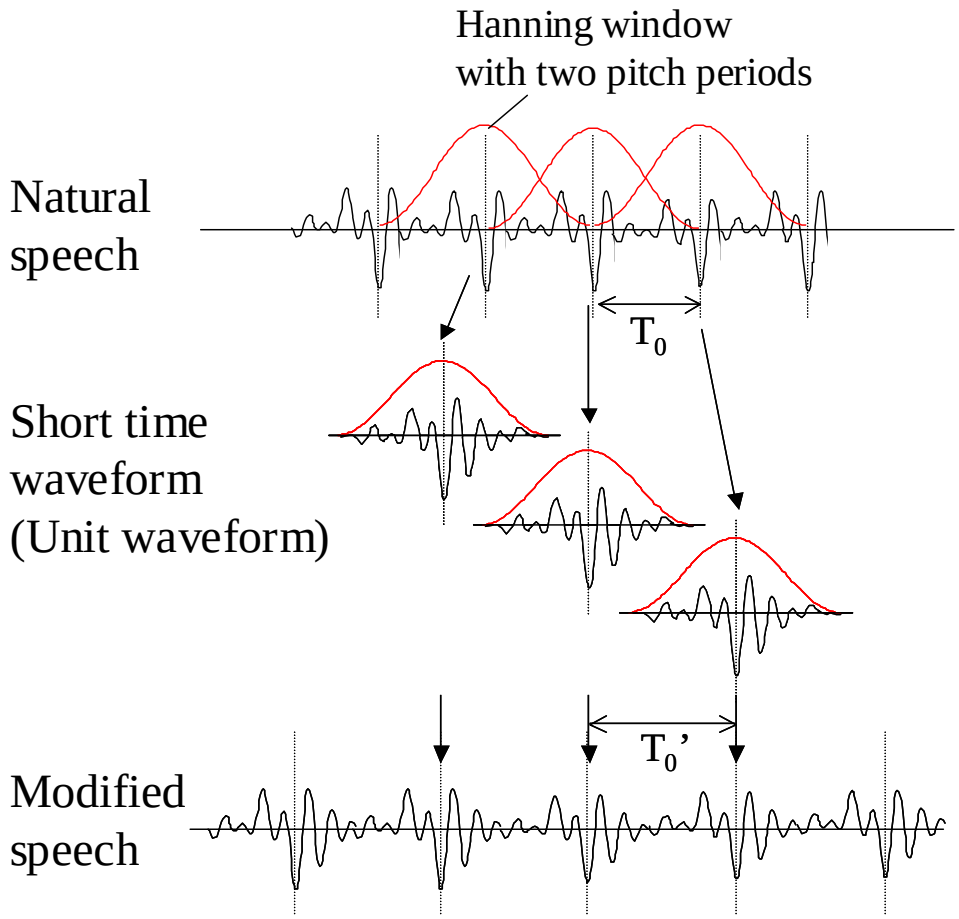


図 1.1　TD-PSOLA 法による F_0 変換

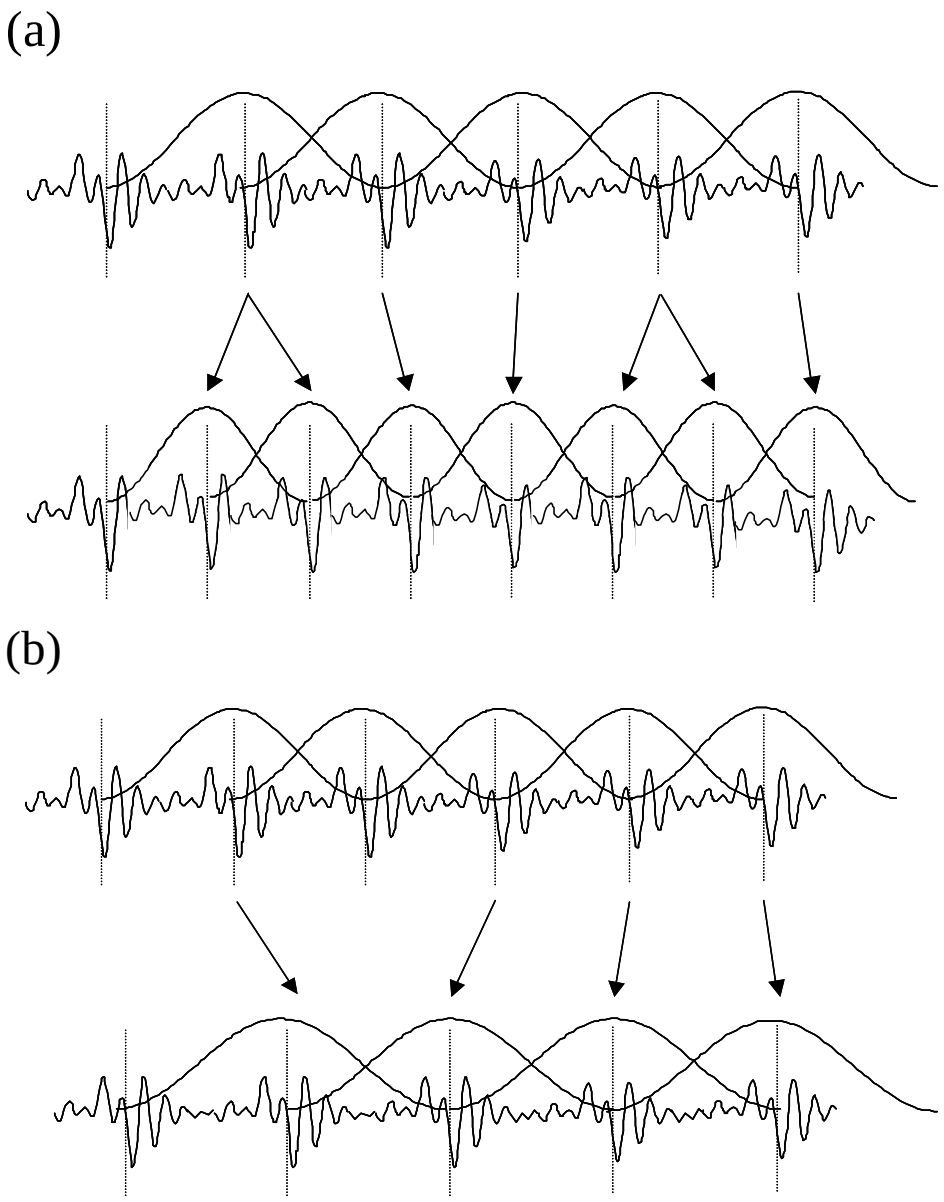


図 1.2　TD-PSOLA 法による時間長の調整:(a) ピッチを高くする場合，(b) ピッチを低くする場合

1.5　本論文の構成

本論文では次章以降，PSOLA 法をベースとした音質改善，及び声質の変換に関する取り組みについて扱う．第 2 章と第 3 章における取り組みは，PSOLA 法にもともと内在する問題を扱ったものであり，音質改善を目的としている．第 4 章における取り組みは，表現の変換を実現するための第一歩として，統計的な手法により合成音声の声質変換を試みる．

第 2 章では，歪の少ないスペクトル包絡の推定を目的とし，ピッチ同期で単位波形を抽出する方法について提案する．TD-PSOLA 法は短時間の窓関数を利用して基本周期の影響を含まない単位波形を抽出し，この単位波形を所望する基本周期で再配列することで，簡単に韻律変換を実現することが可能である．TD-PSOLA 法では，原波形レベルでのローカルピークが聴感上重要な役割を果たすという見解に基づいて，このローカルピークに窓関数の中心を合わせて単位波形を抽出するのが従来の一般的なアプローチであった．しかしながら，この原波形レベルでのローカルピークは，位相やフォルマントの影響によってピーク位置が暴れ，ピッチ同期分析が安定して行えないという問題が存在する．基本周期に同期した単位波形の抽出が安定して行えない場合，そのまま韻律変換処理を行うと，この区間で顕著な音質劣化が生じる．このため，TD-PSOLA 法ではピッチ同期分析を安定して進められることが必須条件となる．そこで本研究では，原音声からフォルマントや位相の影響を取り除いた位相等化残差波形を求め，このピーク位置をピッチマークとして用いる方法について検討する．また，提案方法によって決定したピッチマークを基準に，波形抽出位置を少しずつずらし，スペクトル歪が最小となる最適な波形抽出位置を実験的に探索する．ここで実験的に決定した波形抽出位置を用いて F_0 変換音声を作成し，試聴実験により最適な波形抽出位置とピッチマークとの関係について検証する．なお，ピッチマーク決定方法の頑健性についても F_0 変換音声の試聴実験によって評価する．

第 3 章では，ピッチ同期で抽出した単位波形の低域におけるスペクトル包絡を，スペクトル傾斜とピッチ変換率に応じて動的に再構成する方法について提案する．PSOLA 法によって韻律変換を行う場合，抽出した単位波形をそのまま利用すると

変換音声に著しい音質劣化が生じる場合がある．この音質劣化は原音声から抽出した単位波形のスペクトル包絡が，韻律変換後の環境に適合していないことが原因として考えられる．このスペクトルと韻律との不適合の問題の一つとして，PSOLA 法では元の F_0 より低域において，信頼できるスペクトル情報が得られないという問題が存在する．本来，周波数分析によって求められるスペクトルは，F_0 の整数倍にあたる高調波のみで構成される線スペクトルとなるのが理想である．しかし短時間の窓関数を用いて抽出した単位波形のスペクトルは，窓関数の漏れが隣接する高調波間で重畳され，滑らかなスペクトル包絡が形成される．このため，F_0 より高い周波数領域ではスペクトル包絡が観測される．一方，F_0 より低い帯域においては，窓関数の漏れの影響が観測されるのみで，正しいスペクトル情報が観測できない．この低域における問題により，PSOLA 法では F_0 を低い方へ変換した場合に音質劣化が顕著になっているものと考えられる．そこで本研究では，F_0 変換を行ってもスペクトル傾斜は保存されるという仮定に基づいて，動的に低域におけるスペクトル包絡を再構成することで，音質劣化を軽減する方法を検討する．提案方法によって生成した F_0 変換音声の試聴実験を行い，F_0 を低い方へ変換した場合の有効性について検証する．

第 4 章では，統計的な手法によってスペクトル特徴量をターゲットの環境に変換する際，その変換精度の向上を狙い，韻律情報を考慮したスペクトル変換モデルを提案する．音声合成によって多様な発話を実現するためには，音声収録時の発話スタイルから，ターゲットの発話スタイルへ変換する技術が必要となる．音声の発話スタイルや話者性を決定づける要因としては，話し口調やアクセントなど韻律的な特徴が重要であるが，それに劣らず，声質を決定するスペクトル包絡に関しても精度の良い再現が不可欠である．このスペクトル変換を実現するために，今まで統計的な手法を用いた様々な方法が検討されているが，従来のほとんどの方法では，変換元のスペクトルとターゲットのスペクトルとを 1 対 1 で対応付けし，写像関数を学習している．しかし，スペクトル変換を音声合成へ応用することを考えると，変換関数の入力には変換元のスペクトル以外にも，韻律や音素系列などのコンテキスト情報を利用することが可能である．特にスペクトルは韻律特徴量との間にある程度の相関があるため，変換モデルに韻律情報を考慮す

ることで変換精度の改善が期待できる．そこで本研究では，スペクトル変換を音声合成システムの枠組で利用することを前提に，韻律情報を活用したスペクトル変換モデルについて検討する．実際，提案するスペクトル変換方法を話者変換に応用し，物理評価，及び試聴評価によって韻律情報を用いることの有効性を確認する．更に，従来では変換モデルの学習に同一発話文を用いた方法が利用されていたが，非同一発話文を学習データに使って変換モデルを学習する方法についても検討する．

最後に第 5 章では，PSOLA 法をベースに進めた音質改善，及び声質変換に関する取り組みに対して結論を述べる．また，今後の課題についても考察する．

第2章　スペクトル歪を最小にする単位波形の抽出

2.1　はじめに

音声波形を声帯の1振動に対する応答波形とみなせる短時間波形（単位波形）列に分解し，それを再配列して韻律を制御するTD-PSOLA法は，従来のLPC法 [板倉 70] やPSE法 [中島 88] などよりも音質が良いため，近年の音声合成ではこの方法がよく用いられている．このTD-PSOLA法において韻律変換を行う際，音質劣化を避けるためには，まず単位波形の抽出によって生じるスペクトル歪を抑えることが重要である．従来では窓掛けによって波形形状が大きく崩れることを避けるため，原波形の局所的な振幅最大値（ローカルピーク）に窓の中心を合わせて単位波形を抽出するのが一般的であった．しかし，ローカルピークの位置はフォルマントや位相の影響によってばらつき，このばらつきのある位置を基準に単位波形の抽出を行うと，ピッチ変換音声に異音が生じる．この問題を回避するためには安定したピッチ同期分析が必要であり，ウェーブレット変換を用いて声門閉鎖点を推定する方法 [阪本 95] や，DP法によってピッチマークを選択する方法 [河井 95] などが検討されているが，いずれも若干の手修正を必要とする．また，EGG（Electro Glotto Graph）信号を使う方法 [Krishnamurthy 86] は，安定した声門閉鎖点の推定が期待できるが，音声収録と同時にEGG信号を収録する必要があるため，既存の録音音声に対して使える方法ではない．

上述のローカルピークのばらつきの問題は，フォルマントや位相の影響を含んだ原波形に対して処理を行うために発生していると考えると，これらの影響を取り除いた信号に対して処理を行えば，安定したピッチ同期分析が期待できる．そこで本研究では，ピッチ同期処理を行うための基準位置（ピッチマーク）を安定して

決定する方法として，位相等化残差駆動型線形予測モデル [誉田 84] に基づくピッチマーキング法を検討する．本章では，まず PE-RELP（Phase Equalized Residual Excited Linear Prediction）モデルに基づいて，ピッチマークを推定する具体的な方法について述べる．続いてピッチマークを基準に，最適な単位波形の抽出位置を音声信号モデルを用いて実験的に探索する．また，一様ピッチ変換音声の音質評価によって，提案方法で決定した単位波形抽出位置の妥当性を示す．更に単語データベースに対して全自動のピッチマーキング実験を行い，提案方法によるピッチマーキングの頑健性について検証する．

2.2　ピッチマーキング

従来，PSOLA 法では音声波形のパワーが集中するローカルピークを単位波形の抽出基準位置とする方法が用いられていた．その理由は，原波形のローカルピークは聴感上重要な役割を果たすという考えに基づいており，このピークを損なわないようにするため，窓関数の中心をピーク位置に合わせるようにしていた．しかし図 2.1 に示すように，原波形のローカルピーク位置は，フォルマントや位相の影響によって変動し，必ずしも基本周期に同期した位置とはならない．時間領域で抽出した単位波形を合成に使用する TD-PSOLA 法において，このようにばらついた位置を基準にして単位波形を抽出すると，ピッチ変換音声に位相の不連続が生じ，その影響で音質劣化が発生する．この波形抽出位置の誤りによって生じる音質劣化を避けるためには，従来のようにローカルピーク位置を波形抽出の基準にするのではなく，まず基本周期に同期した基準位置を安定して決定することが重要である．そこで PE-RELP モデルに基づいてピッチマーキングを行う方法について検討する．この方法は線形予測によってフォルマントの影響を取り除き，更に，予測誤差波形の位相を局所的に零位相化することで，パルス列を得ることができる．すなわち，フォルマントや位相の影響を受けることなく，ピッチ同期処理を行うための基準位置を決定できる．以下，PE-RELP モデルの概念について簡単に述べ，続いてこのモデルに基づいたピッチマーキング法の具体的な手順について述べる．

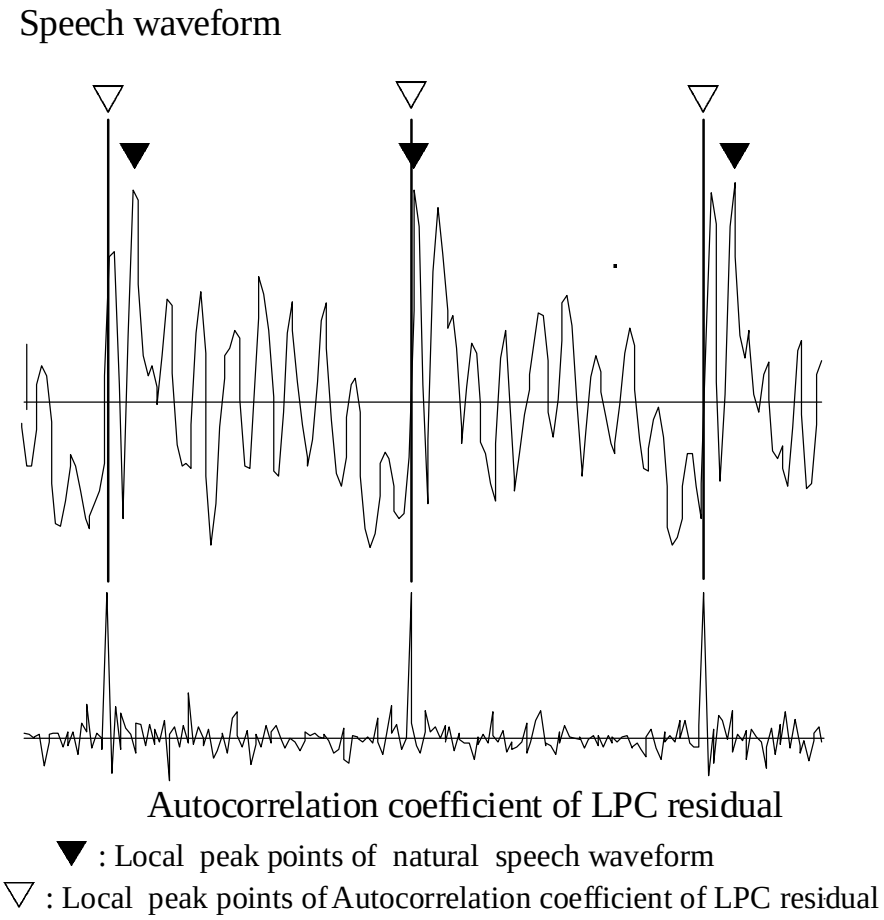
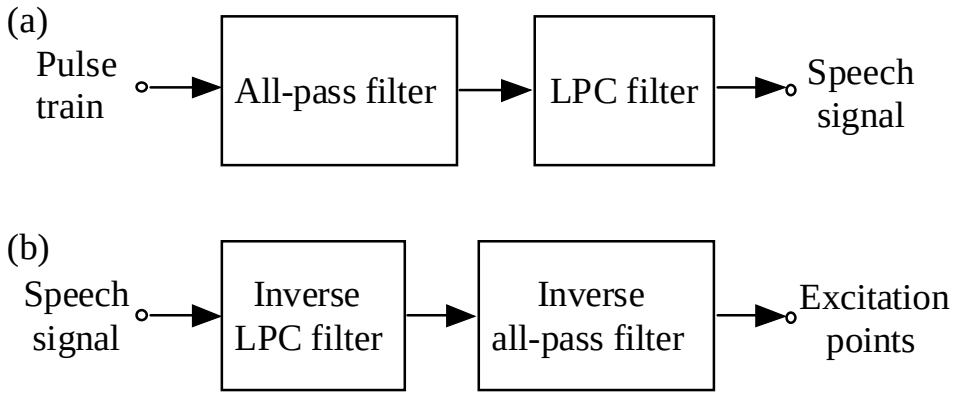


図 2.1　原波形レベルでのローカルピーク位置

2.2.1　位相等化残差駆動型線形予測モデル

図 2.2 に位相等化残差駆動型線形予測（PE-RELP）モデルを示す．このモデルは音声符号化のために提案されたもので，パルス列をオールパスフィルタに入力し，その出力を線形予測フィルタに通して音声信号を生成するモデルである．ここで用いられるオールパスフィルタは，基本周期程度の短時間の群遅延特性を与えるものである．一方，このモデルの分析過程では，音声波形を線形予測逆フィルタに通して残差波形を求め，逆オールパスフィルタを通すことによってパルス化された波形，すなわち位相等化残差波形を得る．ここで逆オールパスフィルタは，残

差波形の局所的な零位相化によってパワーを局所に集中させる役割を果たす．本研究では，ここで得られるパルス位置をピッチマークと定義し，単位波形を抽出する基準位置とする．



Phase equalized residual excited linear prediction model
(a) Synthesis process　(b) Analysis process

図 2.2　位相等化残差駆動型線形予測（PE-RELP）モデル: (a) 合成過程，(b) 分析過程

2.2.2　ピッチマークの決定

図 2.3 にピッチマークを推定する方法の概要を示す．ここでの目的は単位波形を抽出する基準点，すなわちピッチマークを決定することであり，位相等化残差波形そのものを求める必要はない．位相等化残差波形を求める処理は，短時間変形自己相関と等価な処理であるため，実際は短時間変形自己相関係数をピッチマークの推定に利用する．以下にピッチマークを決定する具体的な手順を示す．

1) ケプストラム法を用い，フレーム同期分析（窓幅 32ms，5ms シフト）により，各フレームにおける平均的な基本周波数 f_c を推定する．

2) f_c を用いて，非巡回型のピッチフィルタ [大村 95] を時間領域で構成し，ピッ

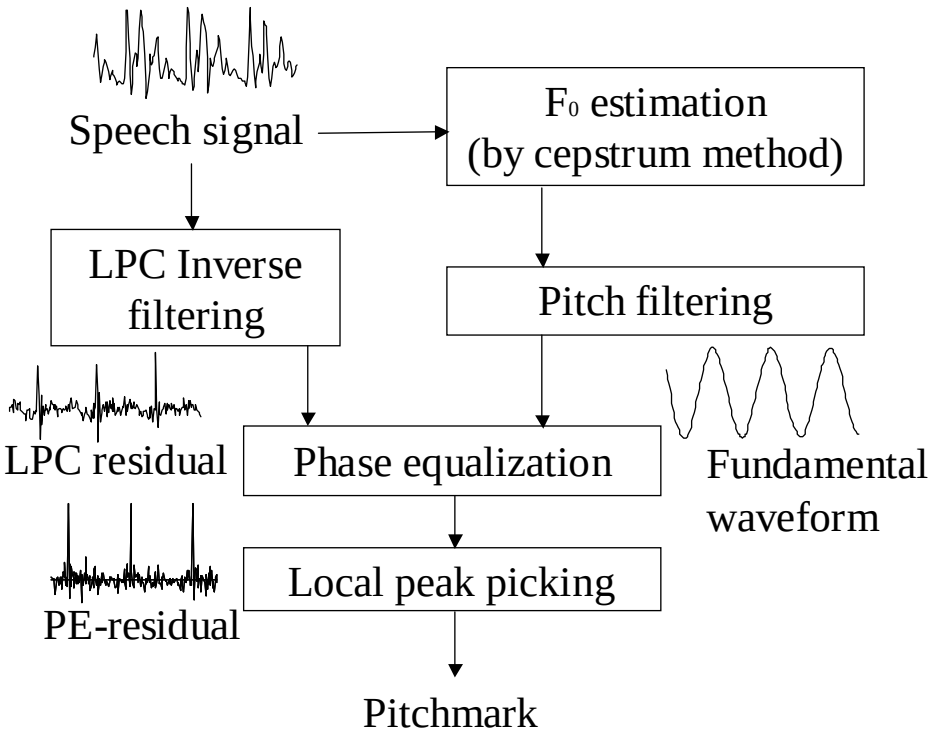


図 2.3　ピッチマーク推定方法の概要

チ基本波を求める．このピッチフィルタ（FIR フィルタ）のフィルタ係数は次式で与えられる．

$$W(n) = W_m(n) \cdot \cos(2n\pi \frac{f_c}{F_s}), \quad -T_c < n < T_c (= F_s / f_c) \tag{2.1}$$

ここで F_s は標本化周波数，$W_m(n)$ は窓関数を表す．この窓関数 $W_m(n)$ には，ブラックマンハリス窓やハニング窓などが用いられる．本研究では以下のハニング窓を用いた．

$$W_m(n) = \begin{cases} 0.5 + 0.5\cos(\pi n \frac{f_c}{F_s}) & (|n| \le T_c) \\ 0 & (|n| > T_c) \end{cases} \tag{2.2}$$

3) ピッチ基本波のパワーを用いて，あらかじめ実験的に設定した閾値により有声部と無声部の判別を行う．

4) 原波形に対して線形予測分析を行い，残差波形を求める．

5) 有声部において，残差波形の振幅絶対値が最大となる点をピッチマークの推定開始位置（初期ピッチマーク）E_i とする．

6) 図 2.4 に示すようにピッチ基本波のゼロクロスから基本周期 T_p を推定し，初期ピッチマーク E_i を中心に $m+1$ 点（$m < T_p$）の短時間残差波形を用いて，$T_p \pm m/2$ の範囲内で自己相関係数を計算する．この係数が最大となる位置を次のピッチマークとする．

7) すべての有声区間に対して，初期ピッチマーク E_i を中心とし，前向き，及び後ろ向きに上述の処理を繰り返し，順々にピッチマークを決定する．

残差波形の自己相関係数を用いることで，フォルマントや位相の影響によって生じる波形レベルでのピークのずれが除去され，安定したピッチマーキングが可能である．また，ピッチ基本波のゼロクロスによって，局所区間に対して推定した基本周期を用いることで，波形形状が急激に変化する部分においても，ピッチ同期分析が安定して行える．

2.3　音声信号モデルを用いた最適波形抽出位置の探索

提案する方法で求めたピッチマークは，ほぼ残差波形のピーク位置と一致する．このピッチマークは安定したピッチ同期処理を行うための基準位置となるが，原波形におけるローカルピーク位置とは一致しない．すなわち，ここで求めたピッチマークに窓関数の中心を合わせて単位波形を抽出すると，位相の不連続による音質劣化は避けられるが，ローカルピーク近傍の波形が窓掛けによって損なわれ，音質に悪影響を及ぼす危険性がある．図 2.5 に提案方法で付与したピッチマークの例を示す．実音声では，提案方法で求めたピッチマークよりもわずかに遅延した位置にローカルピークが観測される場合が多い．このことから，ピッチマークの

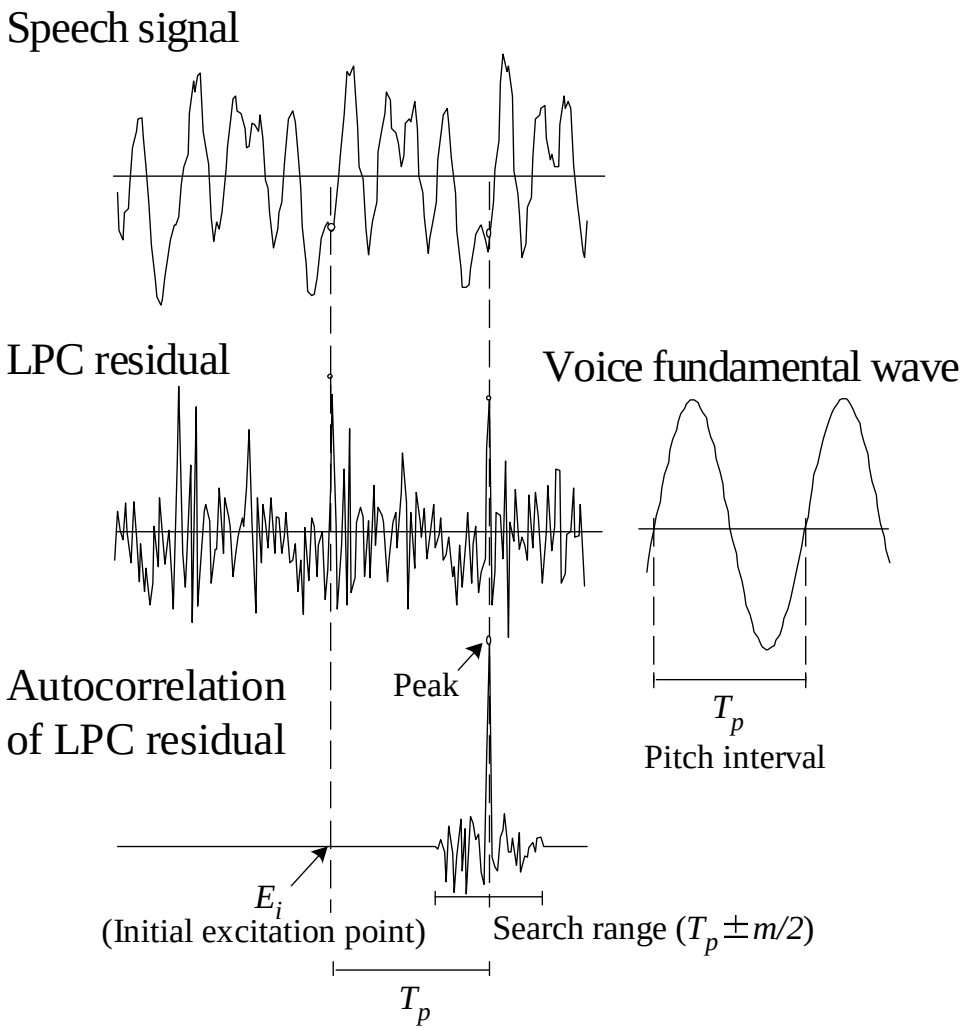


図 2.4　ピッチマークの決定方法

近傍にスペクトル歪が最小となる単位波形抽出位置が存在する可能性があると考えられる．そこでピッチマークを基準にして，このピッチマークからどれだけ遅延した位置にスペクトル歪が最小となる波形抽出位置が存在するのか，音声信号モデルによって生成した擬似音声を用いて実験的に探索する．また，同じく擬似音声を用いて，音声の F_0 や音韻の違いが，スペクトル歪とどのような関係にあるのか調査する．

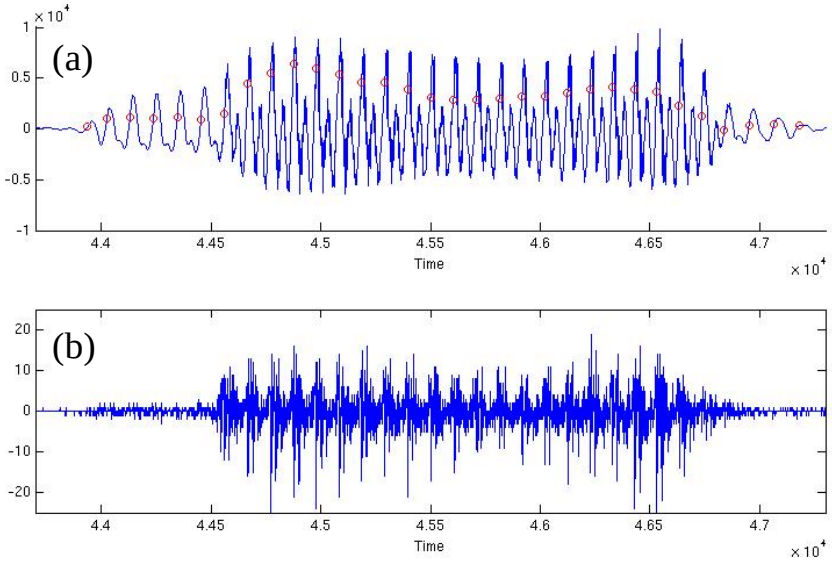


図 2.5　ピッチマークと原波形のローカルピークとの関係:(a) 原波形（音節/ne/）とピッチマーク（図中の○），(b) 残差波形

2.3.1　音声信号モデル

本実験では，インパルス応答の線形的な重畳加算によって音声を生成する簡単な信号モデル（音声信号モデル）を仮定し，擬似音声を生成する．この擬似音声の生成方法を図 2.6 に示す．インパルス応答の作成には，各音韻のスペクトル特徴量を持たせるために実音声を利用する．まず，ある音素の定常区間から 32ms のブラックマンハリス窓で波形を抽出し，フーリエ変換を行い，振幅特性を求める．こ

の際，ケプストラム分析によってリフタリングを行い，元の F_0 の影響を取り除く．位相特性に関しては最小位相条件を与え，因果性を満たすインパルス応答を生成する．このインパルス応答波形と，元の音声の基本周期で並べたインパルス列とを畳み込んで音声信号を生成する．

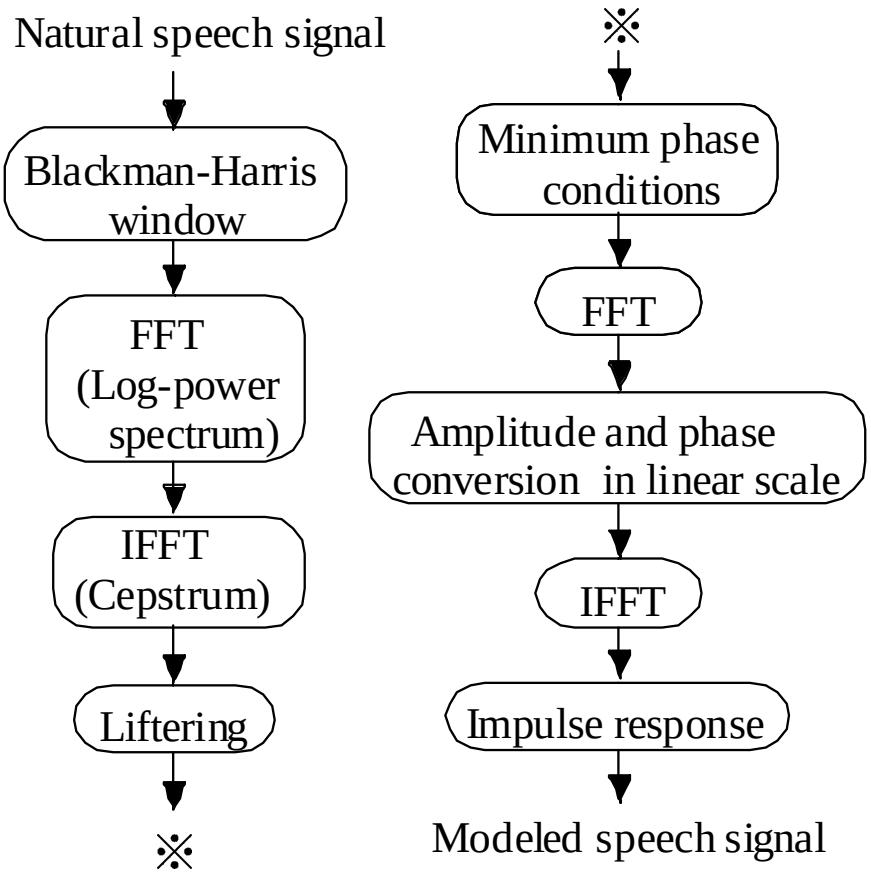


図 2.6　擬似音声を生成するための音声信号モデル

2.3.2　スペクトル歪の測定方法

ピッチマークから遅延した位置にスペクトル歪が最小となる波形抽出位置があるかどうかを調査するために，音声信号モデルを用いてスペクトル歪の測定を行った．本実験におけるスペクトル歪は，擬似音声から抽出した単位波形の対数スペクトルと，擬似音声を生成するのに使用したインパルス応答の対数スペクトルとのスペクトル距離によって定義する．すなわち，本実験におけるスペクトル歪は次式によって得られる．

$$D = \sqrt{\frac{\sum_{i=0}^{N-1}\{M(i) - R(i)\}^2}{N}} \tag{2.3}$$

ここで $M(i)$ は擬似音声の生成のために用意したインパルス応答の対数スペクトル，$R(i)$ はこのインパルス応答の重畳によって生成した擬似音声から PSOLA 法の要領で抽出した単位波形の対数スペクトルである．また，N は周波数ポイント数で，本実験では 1024 ポイントを用いた．スペクトル歪の測定は，擬似音声のピッチマークを基準に単位波形の抽出位置，すなわち窓関数の中心を 1 サンプルずつ遅延させ，各位置におけるスペクトル歪を測定した．この際，先行する応答波形が後続の応答波形に影響を与える可能性を考慮して，6 周期分の応答波形を重畳した擬似音声に対して，4 番目のピッチマークからスペクトル歪測定を開始する．なお，擬似音声から単位波形を抽出する際，PSOLA 法と同様に基本周期の 2 倍の窓長を持つハニング窓を用いた．本実験に用いた音声のサンプリング周波数は 16kHz，量子化ビット数は 16 ビットである．

2.3.3　最適な抽出位置の探索

女性話者が発声した単音節/ma/，/mi/，/mu/，/me/，/mo/について，単位波形の抽出によって生じるスペクトル歪を測定した．図 2.7 に，女声/ma/の母音定常部から生成した擬似音声を用いて，１サンプルずつ単位波形の抽出位置を変えてスペクトル歪を測定した結果を示す．この図から，窓関数の中心を少しずつ遅延させると，ピッチマークより若干遅延した位置でスペクトル歪が最小となり，ピッチマークのほぼ中間でスペクトル歪が最大となっていることがわかる．また，各音

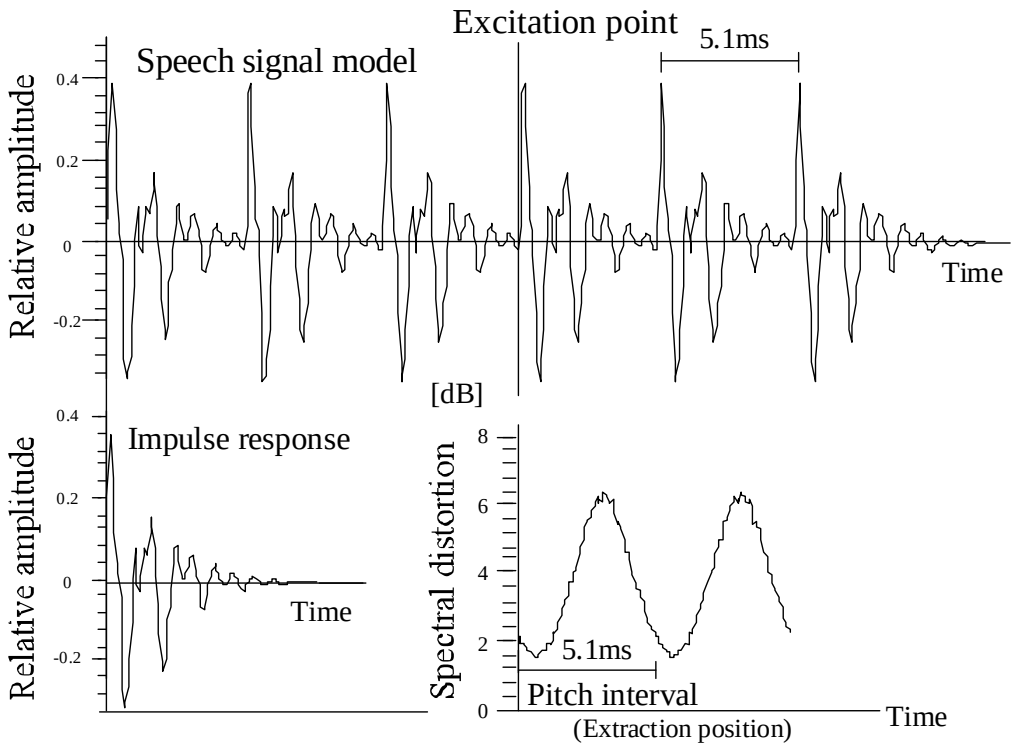


図 2.7　単位波形抽出位置とスペクトル歪との関係:女声/ma/の母音定常部からインパルス応答を作成した例

韻ごとに擬似音声を生成し，同様のスペクトル歪測定を行ったところ，音韻の違いによる傾向としては，第 1 フォルマントの周波数が低い狭母音/i/，/u/ではその他の母音よりスペクトル歪が全体的に大きくなる傾向が確認された．ただし，擬似音声を生成するのに利用した音声データの音韻の違いによって，スペクトル歪の測定結果に多少の差異はあるものの，全体としては，以下の傾向が確認された．

1) ピッチマークから基本周期の 10%～20%遅延したところに窓関数の中心を合わせると，スペクトル歪が最小となる．ただし，ピッチマークに窓関数の中心を合わせた場合と比較して，その差はそれ程顕著なものではない．

2) 基本周期の 50%程度遅延したあたりで，スペクトル歪が最大となる．

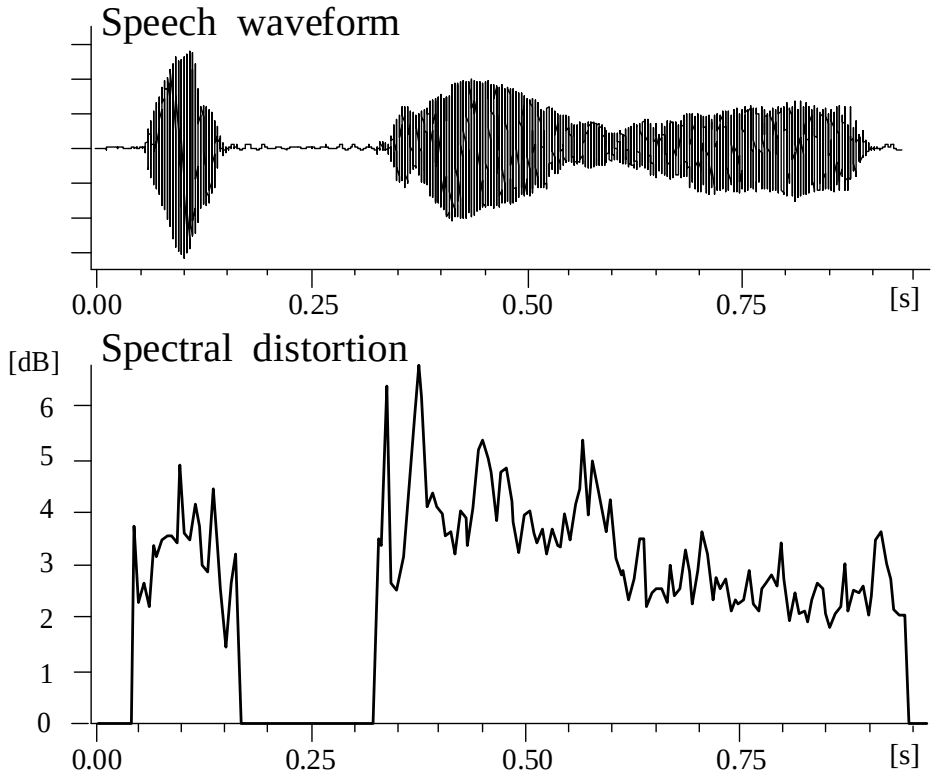


図 2.8　女声単語（高い声）のスペクトル歪: 平均 $F_0$280Hz, 最高 $F_0$460Hz

2.3.4　F_0 による影響

単位波形の抽出によって生じるスペクトル歪が，擬似音声を生成するのに利用した音声の F_0 によって，どのような影響を受けるのか調査した．女性話者が声の高さを変えて発声した単語音声についてスペクトル歪の測定を行った．図 2.8 は F_0 が高い場合，図 2.9 は F_0 が低い場合の結果である．この図において，上段は音声波形，下段は 5ms ごとに擬似音声を生成し，ピッチマークに窓の中心を合わせて抽出した場合のスペクトル歪を示している．その結果，F_0 の高い音声は低い音声と比較して，スペクトル歪が全体的に大きくなることがわかった．F_0 の高い音声は，インパルス応答波形が次のピッチマークまでの間に十分に減衰しないため，PSOLA 法のような短時間窓を用いると，特徴抽出が困難になるものと考えられる．

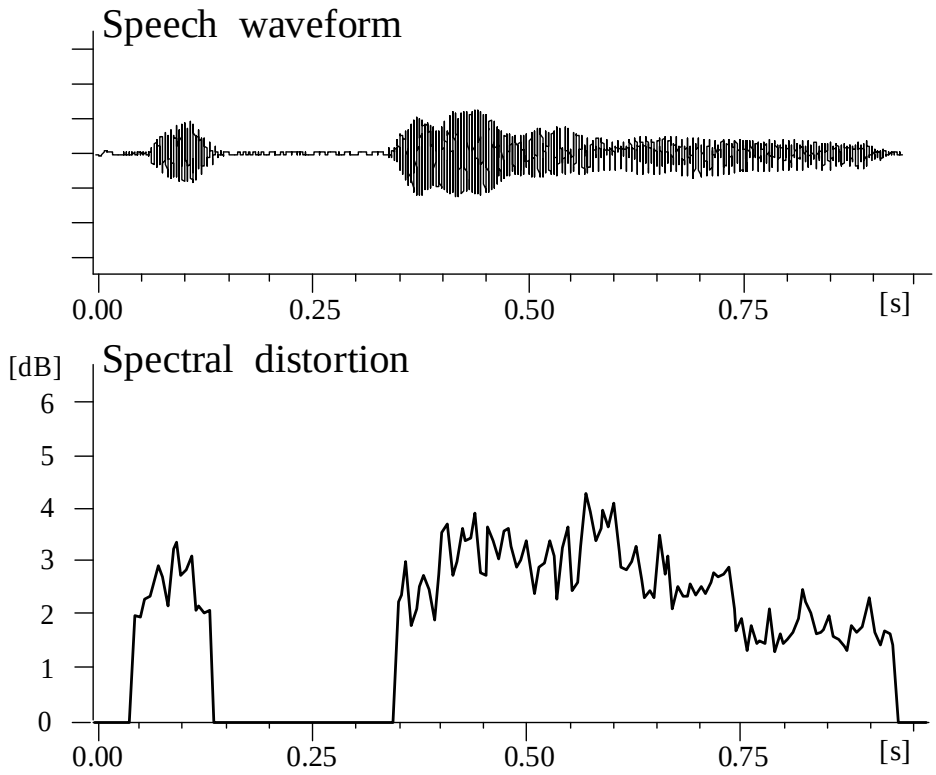


図 2.9　女声単語（低い声）のスペクトル歪: 平均 $F_0$220Hz, 最高 $F_0$300Hz

2.4　ピッチ変換音声の試聴評価

音声信号モデルを用いたスペクトル歪測定の結果，ピッチマークから基本周期の 10%〜20%遅延した位置に窓の中心を合わせて抽出した単位波形は，ピッチマークで抽出したものより，わずかながらスペクトル歪が小さくなることがわかった．そこで単位波形抽出位置の違いによるスペクトル歪の差が，聴感的にどの程度有効なのか検証するため，ピッチ変換音声の試聴実験を行った．試聴実験は，F_0 の高さが異なる単語音声について，単位波形の抽出位置を変えて一様ピッチ変換音声を生成し，その音質について評価した．

2.4.1　一様ピッチ変換音声の生成

実音声では，ピッチマークの前後における基本周期が必ずしも等間隔にはならない．そこで単位波形抽出には，ピッチマークの前後 2 区間にわたる非対称ハニング窓を用いる．一様ピッチ変換は，基本周期を一定比率で伸縮して単位波形を再配列する方法で行う．この時，時間長を原音声と合わせるために，単位波形の繰り返し配列や間引きを行う．なお，無声部についてはピッチ変換，及び時間長制御が不要なため，原音声をそのまま用いる．

2.4.2　音質評価

評価には，男性，女性話者各 1 名が声の高さを 3 段階（Very high, Normal, Very low）で発声した単語音声（16kHz サンプリング，16bit 量子化）を用いた．単位波形の抽出は，ピッチマーク，及びピッチマークから基本周期の 20%，40%，60%，80%，100%遅延した位置を窓関数の中心に合わせて行い，基本周期を 1.3 倍，及び 0.7 倍に変換した音声を作成した．評価はピッチマークに窓関数の中心を合わせて単位波形を抽出した場合の変換音声を基準に，その他の位置で単位波形を抽出した場合の変換音声を比較評価した．評価者は音声処理の研究，開発に従事する成人 10 人で，評価音声をヘッドホンで受聴し 7 段階評定尺度法（+3:非常に良い〜–3:非常に悪い）を用いて音質の善し悪しを 2 回ずつ評価した．

女声，男声の評価結果を図 2.10，図 2.11 に示す．F_0 の高い声（Very high voice）に関しては，単位波形の抽出位置を変えても，顕著な音質の差異はなかった．特に女声では差異がわずかであった．F_0 の高い声では応答波形が十分に減衰しないため，波形抽出位置に関わらず全体的にスペクトル歪が大きくなっているものと考えられる．しかし，F_0 の低い声（Very low voice）では，40%〜80%遅延した位置で単位波形を抽出した場合，音質の劣化が顕著に現れた．また，0%遅延位置（ピッチマーク）における評価と，20%遅延位置における試聴評価の差異は認められなかった．別途，0%〜20%遅延した区間について，5%刻みで波形抽出位置をずらしてピッチ変換音声を作成し，上述と同様に試聴実験を行ったが，この区間における波形抽出位置の違いは，聴感的な差として確認できなかった．試聴評価の結果に

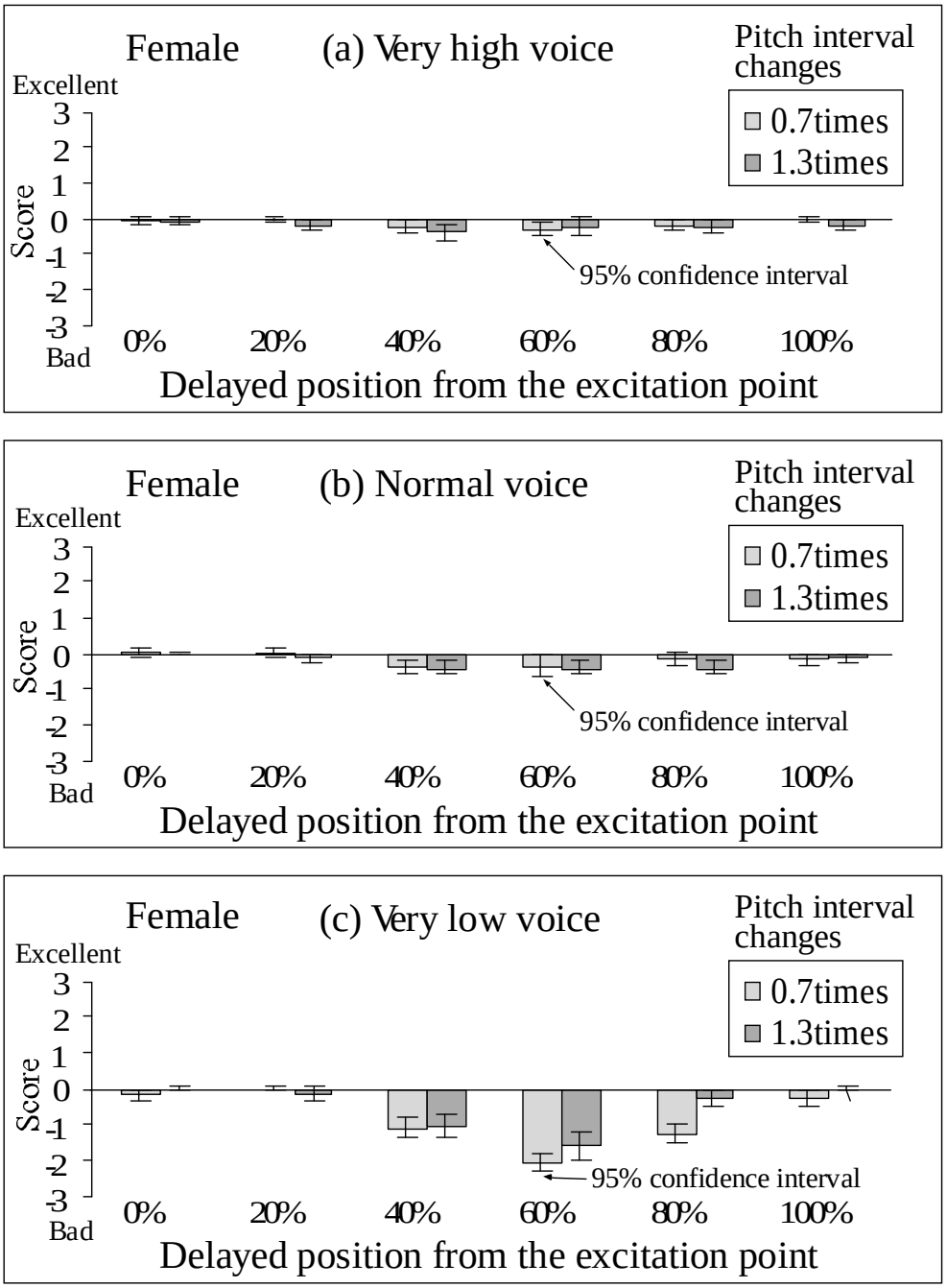


図 2.10　女声の試聴評価結果:声の高さ（平均 F_0，最高 F_0），(a)Very high voice（340Hz, 460Hz），(b)Normal voice （290Hz, 380Hz），(c)Very low voice （180Hz, 250Hz）

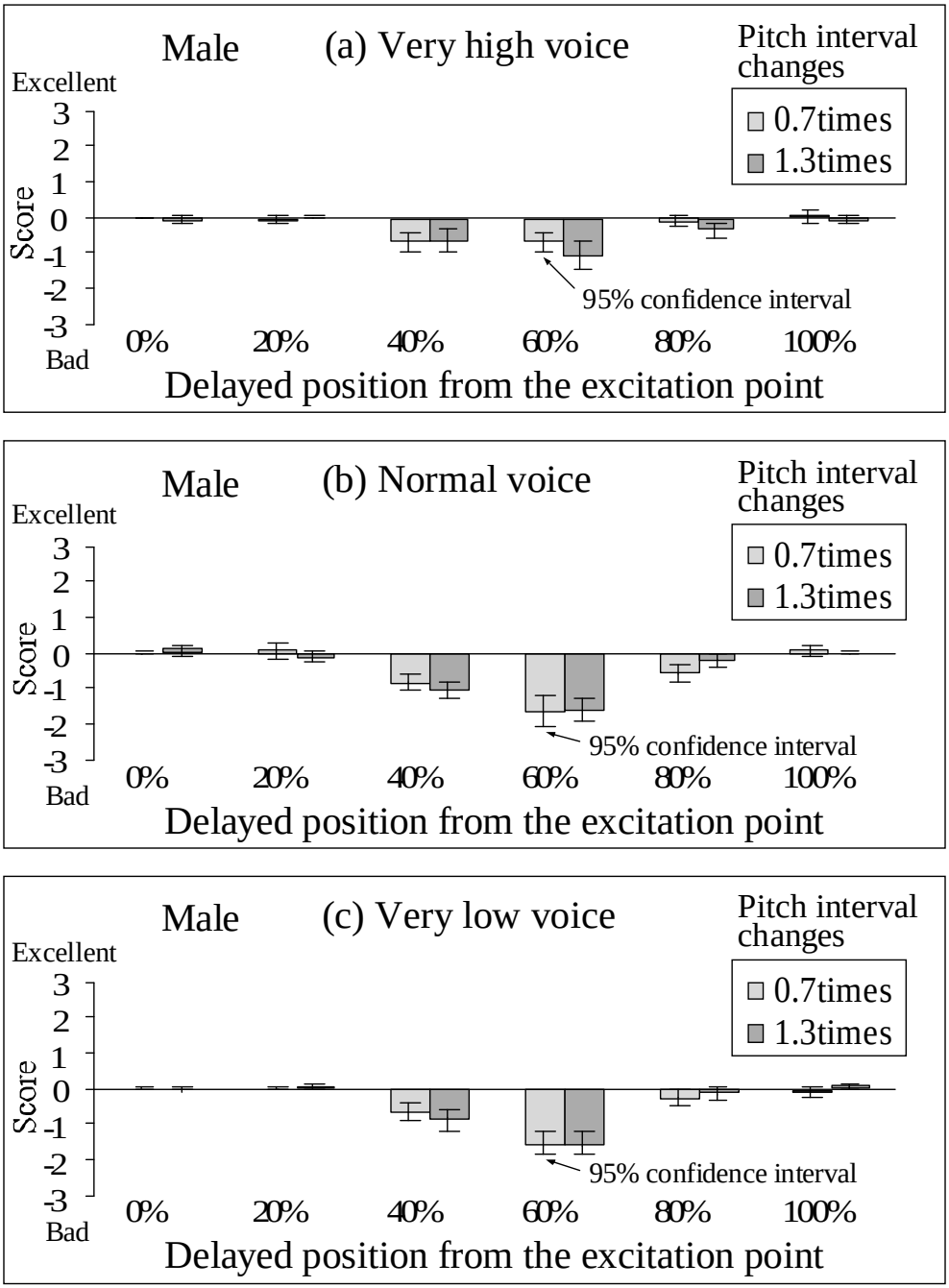


図 2.11　男声の試聴評価結果:声の高さ（平均 F_0，最高 F_0），(a)Very high voice（250Hz, 320Hz），(b)Normal voice （170Hz, 240Hz），(c)Very low voice （130Hz, 170Hz）

表 2.1　女声における試聴評価の検定結果: ピッチマークにおける評価結果を基準に有意水準 5%で検定，〇有意な差である，×有意な差ではない

遅延位置 %		20	40	60	80	100
Very high voice	1.3 倍	×	×	×	×	×
	0.7 倍	×	×	×	×	×
Normal voice	1.3 倍	×	×	×	×	×
	0.7 倍	×	×	×	×	×
Very low voice	1.3 倍	×	〇	〇	×	×
	0.7 倍	×	〇	〇	〇	×

表 2.2　男声における試聴評価の検定結果: ピッチマークにおける評価結果を基準に有意水準 5%で検定，〇有意な差である，×有意な差ではない

遅延位置 %		20	40	60	80	100
Very high voice	1.3 倍	×	×	〇	×	×
	0.7 倍	×	〇	×	×	×
Normal voice	1.3 倍	×	〇	〇	×	×
	0.7 倍	×	〇	〇	×	×
Very low voice	1.3 倍	×	〇	〇	×	×
	0.7 倍	×	〇	〇	×	×

ついて，平均値の差の検定を行ったところ，5%の危険率で 0%遅延位置と 20%遅延位置との有意差はないと判定された（表 2.1，表 2.2）．男声，及び女声における低い声（Very low voice）については，40%，60%遅延位置で，0%遅延位置との差が有意であることがわかった．音声信号モデルを用いたスペクトル歪の測定では，基本周期の 10%〜20%遅延した位置で単位波形の歪が最小となったが，試聴評価結果に対して検定を行ったところ，聴感上はピッチマークを窓関数の中心に合わせた場合と差がないことがわかった．

2.5　ピッチマーキングの頑健性

日本語単語音声データベースを用いて，提案方法による全自動のピッチマーキング実験を行い，ピッチマーキングの頑健性について評価した．ピッチマーキング実験には，VCV ／ CVC バランス単語セット WD-I [速水 85] に外来語を加え，女性，男性ナレータがそれぞれ発声した音声データベース DB1，DB2 と，音韻バランスのとれた 100 単語を女性ナレータが少し高めの声，及び非常に高い声の 2 段階で発声したデータベース DB3，DB4 を用いた．評価はこれらの単語音声データベースに対して全自動でピッチマーキングを行い，このピッチマークを用いて一様ピッチ変換音声を作成し，変換音声に異音が生じていないかどうかを試聴によって確認した．ピッチの変換率は元の基本周期に対して 0.7 倍，及び 1.3 倍とした．それぞれのデータベースの単語数，F_0，及び実験結果を表 2.3 に示す．DB1，DB2，及び DB3 については，ピッチ変換により音質劣化（若干のこもり感）を感じるものもあったが，位相の不連続による異音（ざらつき感，ごろつき感，ポップノイズ）は生じなかった．DB4 については，異音の生じた単語が 4 個あった．その原因として，以下のことが挙げられる．

1) 音声のわたり部分において，ピッチフィルタのバンド幅が F_0 の変動に追従しきれず，適切なピッチ基本波が求められなかったため，ピッチマークの推定誤りが発生した．

2) 非常に強いフォルマントが F_0 の倍音の帯域に存在し，原波形からは視察でも基本周期の推定が困難な音声が存在した（倍ピッチが検出された）．

上述の問題は，F_0 が極めて高い単語音声で生じたが，その他のデータベースでは生じなかった．従って，普通発声の単語音声を扱う限り，提案方法は極めて頑健なピッチマーキング法であると言える．

2.6　むすび

ピッチ同期処理を行うために必要なピッチマークを全自動で決定する方法を検討した．提案方法では，原波形のローカルピークを基準にする従来のアプローチ

表 2.3　ピッチマーキング実験に用いたデータベースと誤り率

データベース	単語数	F_0（平均，最高）Hz	誤り率%
DB1(女声)	569	280，430	0.0
DB2(男声)	569	150，290	0.0
DB3(女声)	100	340，480	0.0
DB4(女声)	100	470，640	4.0

と異なり，位相等化残差駆動型線形予測モデルに基づき，フォルマントや位相の影響を取り除くことで安定して単位波形抽出位置を決定できる．更に，ここで求めたピッチマークを基準にして，音声信号モデルを用いたスペクトル歪の測定，試聴評価実験による最適な単位波形抽出位置の検討，及び単語音声データベースを用いたピッチマーキング実験を行い，以下の結果を得た．

1) 音声信号モデルを用いて単位波形の最適な抽出位置について検討した結果，ピッチマークより基本周期の 10%〜20%遅延した位置にスペクトル歪が最小となる抽出位置が存在し，基本周期の約 50%程度遅延したあたりにスペクトル歪が最大となる抽出位置が存在することがわかった．

2) スペクトル歪測定実験から得た結果の妥当性を聴感的に評価するために，波形抽出位置を変えて生成した一様ピッチ変換音声の試聴実験を行った．その結果，F_0 の高い声では抽出位置を変えても顕著な音質の差はなく，F_0 の低い声では，40%〜60%遅延した位置で音質の劣化が認められた．擬似音声を用いたスペクトル歪の測定実験では，10%〜20%遅延した位置でスペクトル歪が最小となることがわかったが，試聴実験の結果からは聴感的な有意差はなく，窓関数の中心をピッチマークに合わせて単位波形を抽出するのが概ね妥当であることがわかった．

3) 女声，男声の単語音声データベースを用いて，ピッチマーキング実験を行った．その結果，極端に F_0 の高い音声では，ピッチマークの推定誤りによっ

て異音の生じる場合があったが，普通の高さで発声した音声を扱う限り，提案方法は極めて頑健なピッチマーキングが可能であることがわかった.

第3章　スペクトル傾斜に基づいた低域スペクトルの補正

3.1　はじめに

スペクトル包絡の推定の問題は，推定したスペクトルを音声合成に利用することを考えると，単純に厳密なスペクトルを推定するだけでは十分とは言えない．同じ音韻コンテキストで発声された音声であっても，発声時の韻律コンテキストによってそのスペクトル形状は異なる．すなわち，ある環境で録音された音声のスペクトルは，その音声が発声された環境において適切なスペクトルであって，異なる発声環境においては必ずしもそのままの状態で利用できる保証がない．音声合成での利用を考えた場合，問題は発話時のスペクトルを厳密に推定することだけでなく，むしろ合成時の環境に適したスペクトルを推定することが重要となってくる．

TD-PSOLA 法はハニング窓によって抽出した単位波形をそのまま再配列する方法であるため，抽出された単位波形は元の韻律環境に適したスペクトル情報を保存しているが，変換先の韻律環境は一切考慮されていない．ここで TD-PSOLA 法で用いる単位波形の振幅特性に着目すると，基本周期の 2 倍の長さを持つハニング窓によって抽出された単位波形のスペクトルは，窓関数の影響によって F_0 の高調波間のスペクトルが補間されるため，F_0 より高い周波数帯域では滑らかな包絡が得られる．しかしながら，元の F_0 より低い周波数帯域においては信頼できるスペクトル情報が存在しないため，適切にスペクトル包絡を再現できない．これは PSOLA 法によってピッチを低い方へ変換した場合に，音質劣化を引き起こす原因の一つとして考えられる．

上述のような問題を解決する手法としては，複数の韻律環境において事前に学

習を行い，合成時にターゲットの韻律に適したスペクトルへ変形する方法が考えられる．例えばスペクトル特徴量を F_0 の高さに応じてクラスタリングし，クラス間の写像関数を利用してスペクトルを変形する方法 [Tanaka 97] は，F_0 の違いによるスペクトル変化差分を入力の F_0 の条件に応じて変形し，元のスペクトルに加算することでスペクトル補正を実現している．この補正処理では，元のスペクトルが持つ微細構造は残され，緩やかな包絡成分のみが補正されることになる．Sinusoidal model を用いた合成でも，これと類似の補正方法が検討されている [Wouters 01]. また，多重回帰分析において，スペクトル特徴量と強い相関がある説明変数を用い，F_0 の変更に伴いターゲットのスペクトル特徴量を線形変換する方法 [峯松 99] も提案されている．しかしながら，これらの手法は事前に学習が必要であるため，同一話者において複数の韻律環境の学習データを持っていることが条件となる．このため，ある程度まとまった学習データを持たない場合には利用できないアプローチである．

一方，上述のアプローチが学習を必要とするのに対し，事前に学習することなく，動的にスペクトルを補正する方法も検討されている．スペクトル包絡をケプストラム分析によって推定し，その際，リフタリングによって切り離した基本周期成分をターゲットに合わせて線形伸縮し，この伸縮した基本周期成分を合成時に再びスペクトル包絡に加算する方法 [阿部 89, Takano 01] は，基本周期成分に関するスペクトルをターゲットに適応することで音質改善を実現している．また，スペクトルの位相特性に関しては，ピッチ変換率に応じて最適な群遅延を実験的に決定する方法 [坂野 00] や，元の波形形状を可能な限り崩さずに済む位相特性を波形レベルでの誤差最小基準に基づいて再帰的に求める方法 [Griffin 84] などが提案されており，後者においては振幅特性の補正と合わせて検討されている [阿部 89, 東 99]．しかしながら，これらの手法も元の F_0 より低域のスペクトル包絡が再現できないという問題に対して，その解決を明示的に進めた取り組みではない．そこで本研究では，元の F_0 より低域におけるスペクトル包絡の問題に着目し，この帯域におけるスペクトル包絡をスペクトル傾斜に基づいて再構成する方法を提案する．本章では，まず PSOLA 法を用いてピッチ変換した場合に生じる低域スペクトルの問題について詳細に説明する．続いてピッチ変換後も元のスペクトル傾斜を

保存するという考えに基づいて，低域のスペクトルを再構成する方法について検討する．最後に試聴評価を行い，提案する方法がピッチを低い方へ変換する場合に有効であることを示す.

3.2　単位波形のスペクトル包絡

PSOLA 法において，ハニング窓で抽出した単位波形は基本周期の影響を含まない声帯 1 振動に対するインパルス応答波形とみなすことができる．しかし，この周波数特性には，元の F_0 より低い周波数帯域で信頼できるスペクトル情報を持たないという問題がある．以下，この低域スペクトルの問題について簡単な信号モデルを仮定して説明する．また，PSOLA 法における窓関数の役割についても議論する.

3.2.1　低域スペクトルの課題

図 3.1 を用いて低域におけるスペクトルの問題について説明する．ここでは問題を簡単にするために，図 3.1(a) に示すような各周波数帯域でフラットな振幅特性を持つ単位波形 A を考える．この単位波形 A を T_0（$= 1/F_0$）間隔で繰り返し重畳して生成した周期波形 B の振幅特性は，図 3.1(b) のように F_0 とその整数倍の周波数のみ値を持つ理想的な線スペクトルとなる．この周期波形 B から，TD-PSOLA 法の要領で基本周期 T_0 の 2 倍の窓長を持つハニング窓で単位波形 C を抽出すると，その振幅特性は図 3.1(c) のようになる．この単位波形抽出処理は，図 3.1(b) の線スペクトルに，単位波形抽出に用いた窓関数の影響を畳み込んだことになるため，高調波間が窓関数の漏れによって補間され，滑らかな包絡が形成される．基本周期の 2 倍の窓長を持つハニング窓は，連続する二つの高調波の中間で振幅が半分に減衰する特性を持つため，高調波間を滑らかに補間するのに都合の良い特性と言える．この単位波形 C の振幅特性は，F_0 より高い帯域では元の単位波形 A の振幅特性とほぼ同様のフラットな特性が再現できている．一方で，F_0 より低い帯域においては，F_0 における振幅に対して畳み込まれた窓関数の影響が観測されるの

みで，上述のような補間が行われない．よって単位波形 C の振幅特性は，元の F_0 から低くなるにつれて減衰している．ここで T_0 より長い基本周期 T_0'（$> T_0$）で単位波形 A を繰り返し重畳した場合，その周期波形 D の振幅特性は図 3.1(d) のような線スペクトルになるのが理想である．しかし，TD-PSOLA 法では，基本周期 T_0 の 2 倍の窓長を持つハニング窓で抽出した単位波形 C を基本周期 T_0' で配列することになるので，この基本周期を変換した波形 E は，図 3.1(e) のようなスペクトル包絡を持つことになる．図 3.1(d) と図 3.1(e) を比べると，基本周期を変換した波形 E の振幅特性は，元の F_0 より低い変換先の F_0'（$= 1/T_0'$）で減衰しており，はじめに仮定したフラットな振幅特性が再現されていない．

上述では説明を簡単にするため，フラットな振幅特性を持つ単位波形を仮定して議論を進めたが，実際の音声波形の場合も基本周期の 2 倍のハニング窓で抽出した単位波形を元の基本周期より長い周期で再配列すると，これと同じ問題が発生する．このため，特に周波数が低くなる方へピッチ変換する場合は，元の F_0 より低い周波数帯域のスペクトル包絡を何らかの方法で補正する必要がある．

3.2.2　窓関数の補間特性

一般的に信号の周波数分析を行う場合，窓関数の特性には以下のことが望まれる．

1) メインローブの幅が狭い

2) サイドローブが十分に減衰する

しかしながら，メインローブの幅と，サイドローブの減衰はトレードオフの関係にあり，両者の条件を都合良く満たす窓関数は存在しない．一方，PSOLA 法において単位波形の抽出に利用する窓関数は，急峻なピークを持つハーモニクス構造を観測することが目的ではなく，むしろハーモニクス間を窓関数の漏れによって補間し，基本周波数の影響を含まない滑らかなスペクトル包絡を得ることが重要である．そこで基本周波数の整数倍において同じ振幅値を持つ線スペクトルを仮定し，この線スペクトルに基本周期の 2 倍の窓長を持つ窓関数の影響を畳み込み，再現されるスペクトル包絡を調査した．図 3.2 は基本周期の 2 倍の長さを持つハ

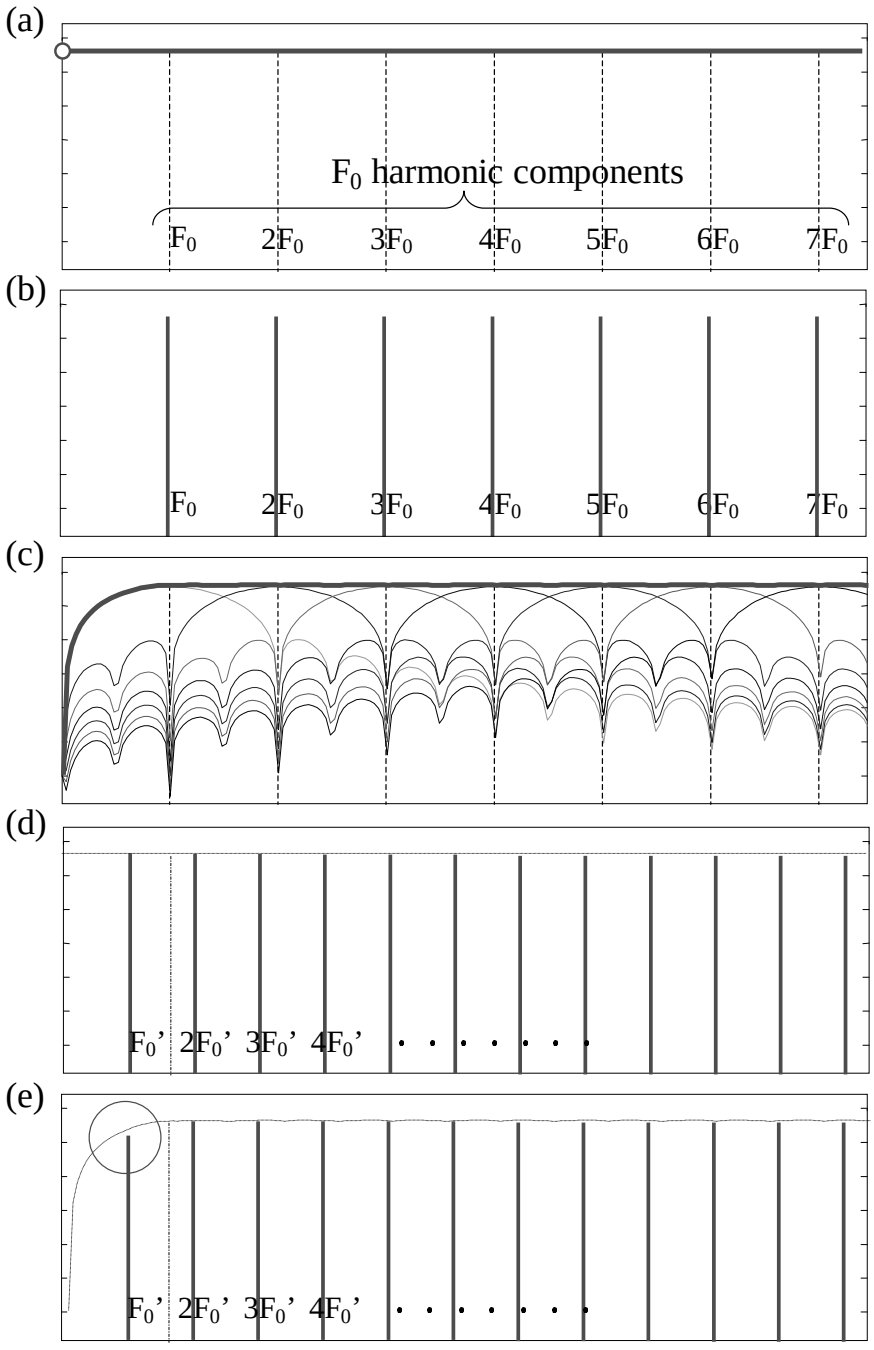


図 3.1　低域におけるスペクトル減衰の問題:以下の波形の振幅特性，(a) フラットな特性を持つ単位波形 A，(b) 単位波形 A から生成した周期波形 B，(c) 周期波形 B から抽出した単位波形 C，(d) 単位波形 A から生成した周期波形 D，(e) 単位波形 C から生成した周期波形 E

ニング窓の影響を重畳した場合の補間特性である．ハニング窓のメインローブの影響は，隣接するハーモニクスとの中間で半減し（およそ 6dB の減衰），隣接するハーモニクスでは十分に減衰してほとんど影響を与えない．このため隣接するハーモニクスにはさまれた周波数帯域では，窓関数の影響が重なることによってフラットな振幅特性が形成される．ここでハニング窓以外の代表的な窓関数として，矩形窓を用いた場合の補間特性を図 3.3 に示す．矩形窓はメインローブの幅が狭いため周波数分解能は優れているものの，補間によって滑らかなスペクトル包絡を得るという用途では利用できない．また，図 3.4 に示すブラックマンハリス窓の補間特性は，ハーモニクス間では安定した補間を実現しているように見えるが，メインローブ幅が広く，隣接するハーモニクスへ影響を与えるため，肝心なハーモニクス自体の振幅値が本来の値からずれてしまう．これらのことから，各ハーモニクス間の振幅特性を補間によって生成するという観点からは，ハニング窓の特性が好都合であることがわかる．

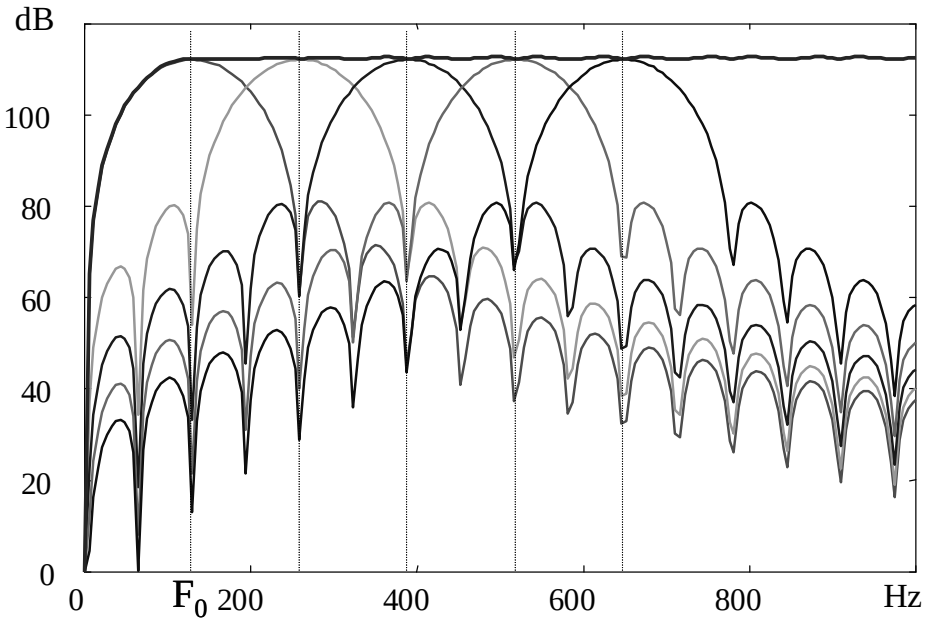


図 3.2　ハニング窓の補間特性

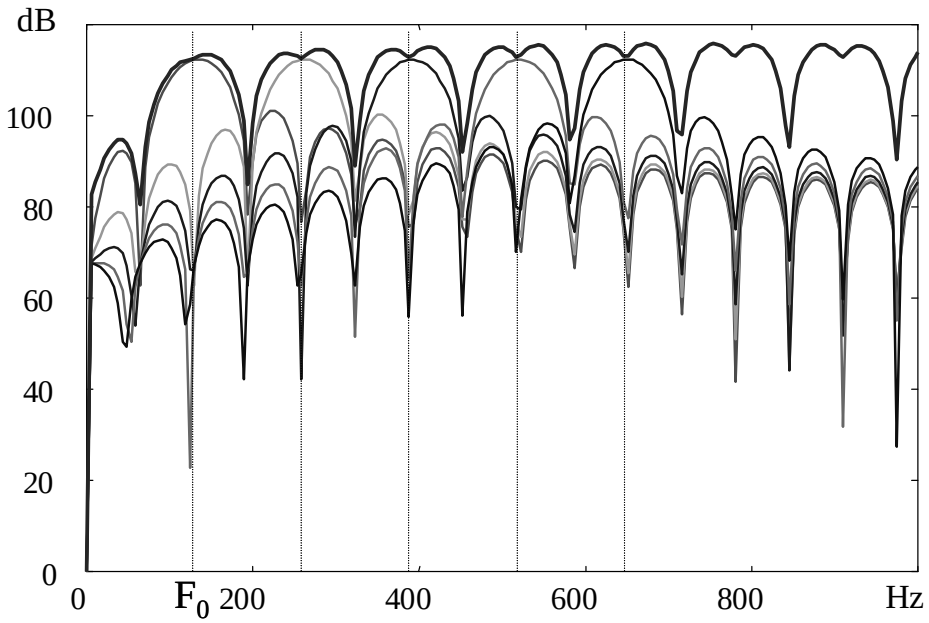


図 3.3　矩形窓の補間特性

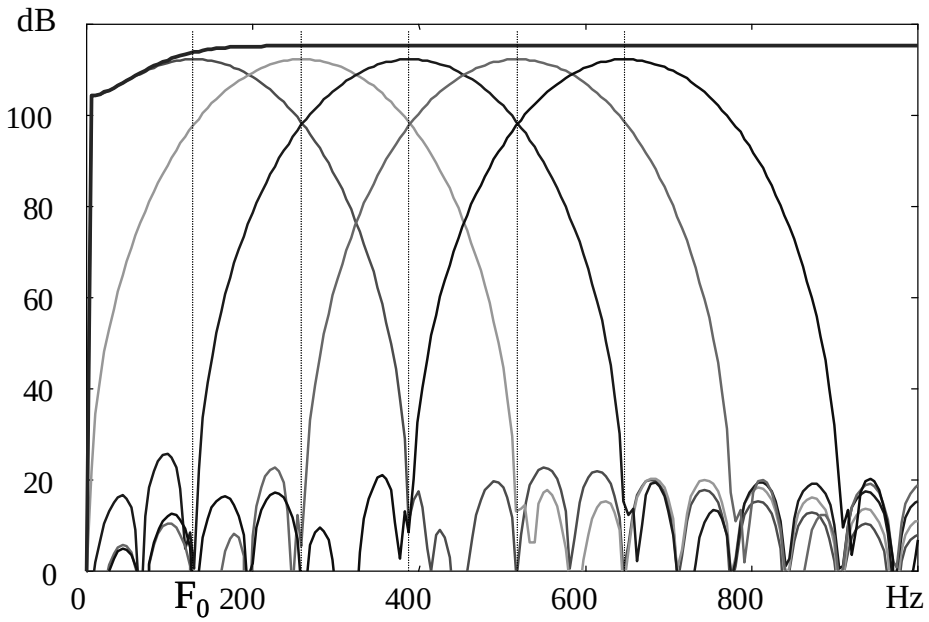


図 3.4　ブラックマンハリス窓の補間特性

3.3　スペクトル包絡の補正

本来，韻律環境が異なれば，スペクトル包絡も変わると考えるのが一般的であり，以前から韻律変換に伴いスペクトルをターゲットの韻律に適応する方法が検討されている．しかし，本研究ではF_0を変更しても声質は極端に変わらないという仮定のもと，PSOLA法において低域で発生している問題にのみ焦点を当て，スペクトルの補正方法を検討する．

PSOLA法の要領で抽出した単位波形は元のF_0より低い周波数帯域に信頼できるスペクトル情報を持たないため，学習などの方法を用いずにスペクトルを再現するには，何らかの仮説に基づいて生成することになる．本研究では上述のようにF_0の変化が声質に与える影響は少ないという仮定のもと，F_0変換後もスペクトル傾斜を一定に保つという考えに基づき，低域のスペクトルを新たに生成する方法を検討する．図3.5に，提案するスペクトル補正処理のフローを示す．このスペクトル包絡の補正処理は，PSOLA法によるピッチ変換の枠組の上で，単位波形に対して行う．以下，各処理について説明する．

3.3.1　スペクトル包絡の抽出

第2章で提案した方法によりピッチマークを求め，基本周期の2倍の窓長を持つハニング窓で単位波形を抽出する．この単位波形をフーリエ変換し，振幅特性（対数スペクトル）と位相特性を獲得する．ここで位相特性は，再び時間波形を求めるときに利用するため保存する．

3.3.2　スペクトル傾斜の推定

図3.6に示すように，単位波形のスペクトル包絡に対して，スペクトル傾斜を表す係数α（単位はdB/octave）をLMS（Least Mean Square）法によって求める．F_0より低い帯域には信頼できるスペクトル情報がないため，F_0から$F_s/2$の間で傾斜

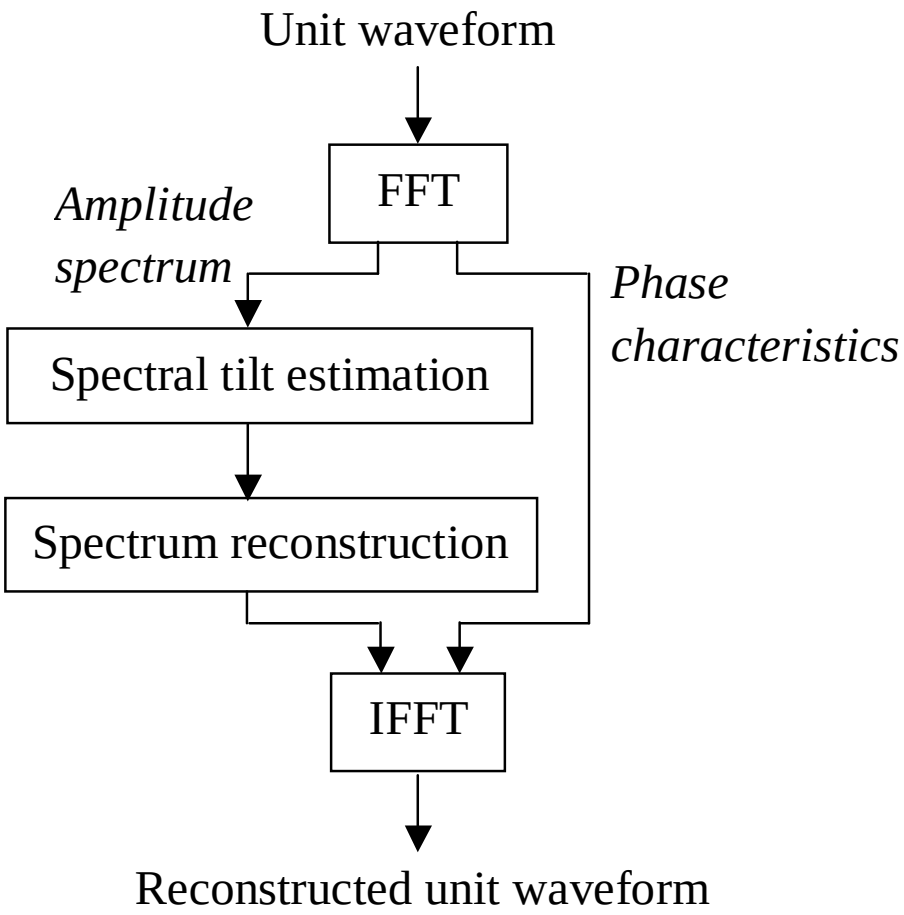


図 3.5　単位波形のスペクトル補正処理

を定義する（F_s は標本化周波数）．この係数 α は次式によって得られる．

$$\alpha = \frac{\sum_{\omega_i \in \Omega} \log_2 \frac{\omega_i}{\overline{\omega}} \cdot \ln \frac{|S(\omega_i)|}{\overline{|S(\omega)|}}}{\sum_{\omega_i \in \Omega} (\log_2 \frac{\omega_i}{\overline{\omega}})^2} \qquad (3.1)$$

$$\Omega = \{\omega_i \mid \omega_0 < \omega_i < \pi\}$$

ここで ω_0 は F_0 に対応した角周波数，ϖ は平均角周波数，$\overline{S(\omega)}$ はスペクトル包絡の平均振幅値である．

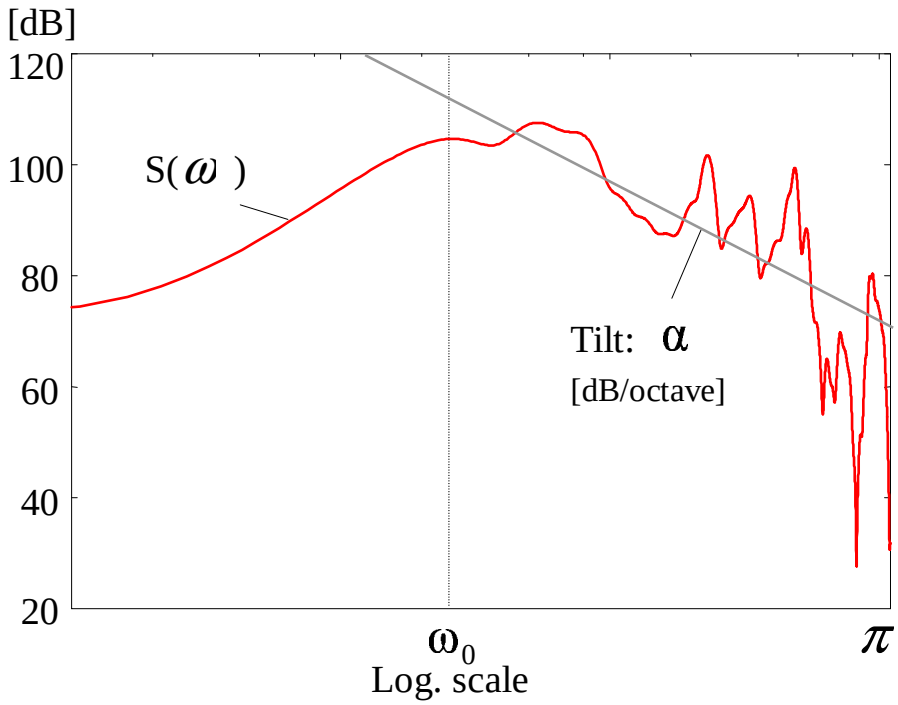


図 3.6　スペクトル傾斜の推定

3.3.3　スペクトルの再構成

図 3.7 を用いて，低域スペクトルの再構成処理を説明する．

1) F_0 より低域におけるスペクトル包絡を取り除く（図 3.7(a)）．

2) 図 3.7(b) において，F_0 におけるスペクトルの振幅値を起点とし，それより低域に向けて係数 α の傾きの線分を引く．この線分と変換先の F_0' とが交わる位置を A_1 とする（図 3.7(b)）．

3) F_0' に対して，最大振幅値が A_1 のハニング窓の振幅特性を求める．同様に，F_0' の高調波である $2F_0'$ に対しても，ハニング窓の振幅特性（最大振幅値は A_2）を求める．このハニング窓の振幅特性は，新たに変換先となる F_0' から，元の F_0 に最も近い F_0' の高調波（ただし，$\geq F_0$）まで求める（図 3.7(c) の例では $2F_0'$ まで）．

4) ここで F_0' とその高調波に対して求めたハニング窓の振幅特性を畳み込み，新しいスペクトル包絡を作成する．低域（図 3.7(c) の例では $2F_0'$ 以下の帯域）は新たに作成したスペクトル包絡を利用し，それより高域では元のスペクトル包絡を利用して，単位波形のスペクトル包絡を再構成する（図 3.7(d)）．

例として，母音/e/について，元のスペクトル包絡と，F_0 を提案方法で 1.0octave 下げた場合のスペクトル包絡を図 3.8 に示す．再構成されたスペクトル包絡は，ターゲットとなる F_0 の帯域が補強され，高域は元の形状が保存されているのがわかる．また，時間方向へも滑らかに遷移していることがわかる．

以下，上述の処理によって補正されるスペクトル包絡を数式によって表す．補正処理によって新たに構成するスペクトル $S'(\omega)$ は，変換先の F_0' に対する線スペクトルと，ハニング窓の周波数特性により，次式によって定義する．

$$S'(\omega) = \begin{cases} \displaystyle\sum_{i=1}^{N} A_i \cdot \frac{|W_i(\omega)|}{Wmax_i} & (\omega < \omega_0' N) \\ S(\omega) & (\omega_0' N \leq \omega) \end{cases} \tag{3.2}$$

$$Wmax_i = max \mid W_i(\omega) \mid, \qquad N = \left\lceil \frac{\omega_0}{\omega_0'} \right\rceil + 1$$

ここで ω_0' は変換先となる F_0' に対応する角周波数，i は高調波のサフィックス，A_i は i 番目の高調波の振幅値，$W_i(\omega)$ は i 番目の高調波に対して畳み込まれる窓関数の周波数特性である．なお，$\lceil x \rceil$ は x を超えない最大の整数を意味するフロア関数

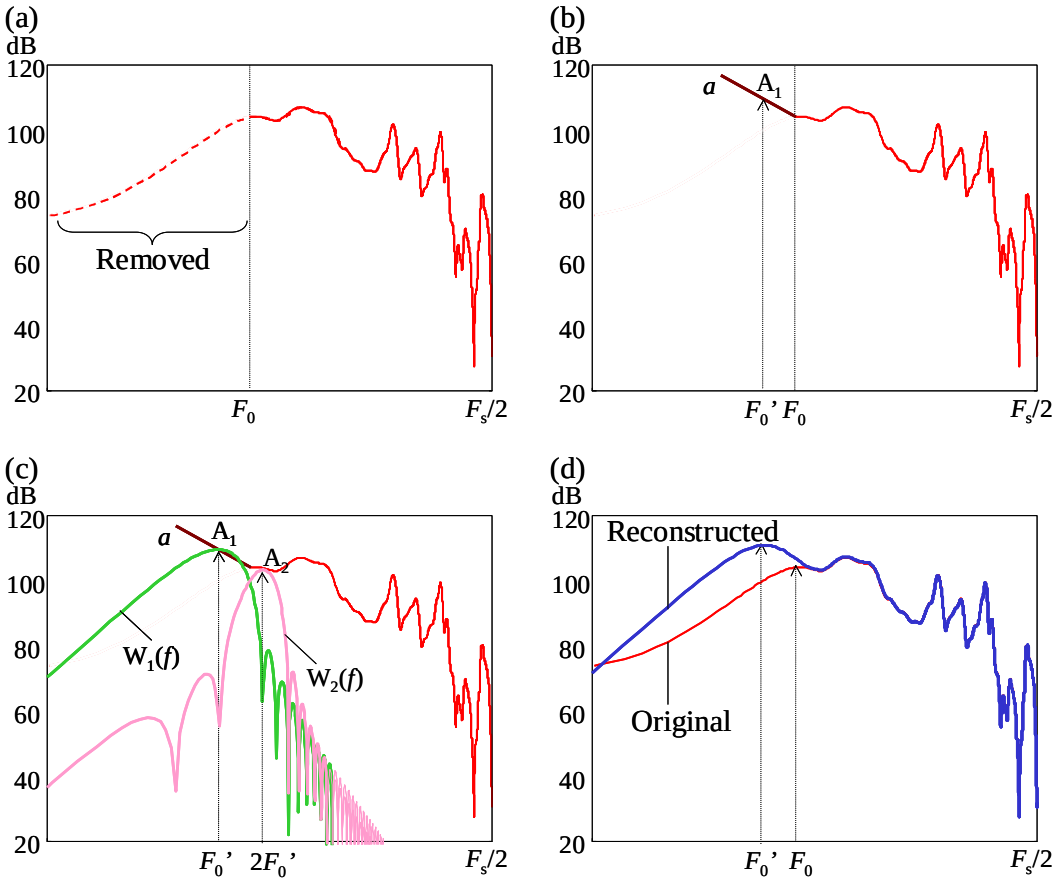


図 3.7　スペクトル包絡の再構成処理:(a) 元の低域スペクトルの削除，(b) 傾斜に基づいた振幅推定，(c) 窓関数の周波数特性の重畳，(d) スペクトル包絡の再構成

である．線スペクトルの振幅値 A_i は次式で与えられる．

$$A_i = \begin{cases} \exp\{\alpha \log_2(i\omega_0'/\omega_0)\} \cdot S(\omega_0) & (i < \omega_0/\omega_0') \\ S(i\omega_0') & (\omega_0/\omega_0' \leq i) \end{cases} \tag{3.3}$$

上式において，元の F_0 より低い周波数帯域における線スペクトルの振幅値は，スペクトル傾斜の係数 α と元の F_0 における振幅値 $S(\omega_0)$ とから計算されている．窓

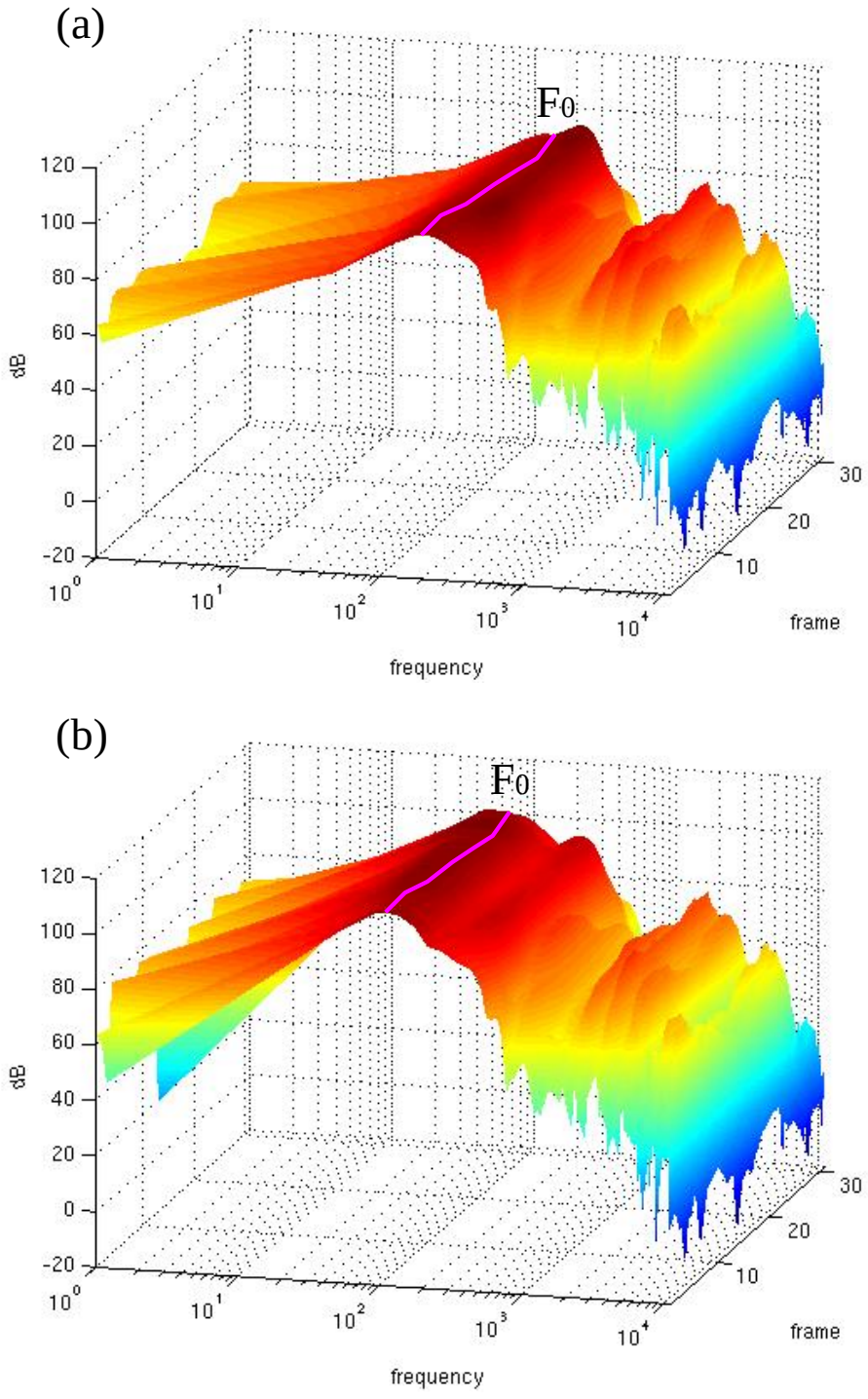


図 3.8　母音/e/のスペクトル包絡の時間変化:(a) 元のスペクトル包絡，(b) 再構成されたスペクトル包絡（F_0 を 1.0octave 下げた場合）

関数の周波数特性 $W_i(\omega)$ は，時間領域で以下のように設計することができる．

$$W_i(\omega) = F[w_i(t)] \tag{3.4}$$

$$w_i(t) = w_{han}(t, T_0') \cdot \cos(2\pi i t / T_0') \tag{3.5}$$

ここで T_0' は変換先となる基本周期，$F[\cdot]$ はフーリエ変換を意味する．また，ハニング窓 $w_{han}(t, \tau)$ は以下のように表せる．

$$w_{han}(t, \tau) = 0.5(1.0 + \cos(\pi t / \tau)) \qquad (|t| < \tau) \tag{3.6}$$

上述の要領でピッチを低い方へ変換した場合と，高い方へ変換した場合に再構成されるスペクトル包絡を図 3.9 に示す．ピッチを高い方へ変換する場合は，新しい F_0' における振幅値 A_i が元のスペクトル包絡 $S(\omega)$ から得られるため，元の F_0 より低い周波数帯域における振幅推定を必要としない．すなわち，スペクトル傾斜の推定は，ピッチを低い方へ変換する場合のみ必要となる．このことから，提案する低域スペクトルの再構成処理は，ピッチを低い方へ変換する場合に有効であると考えられる．

3.3.4　補正後の単位波形

低域補正を行ったスペクトル包絡 $S'(\omega)$ に対して，その逆フーリエ変換によって時間波形を得る．この際，元の単位波形の位相特性をそのまま利用する．上述の要領で新たに得られる単位波形を，ターゲットの T_0' 間隔で再重畳することにより，ピッチ変換を実現する．

ここで基本周期を T_0 から T_0'（$> T_0$）へ変換する際，オリジナルの単位波形をそのまま用いる場合と，提案方法でスペクトル包絡を補正した単位波形を用いる場合について考える．図 3.10(a) に示すように，オリジナルの単位波形を配列した場合は，連続する二つの単位波形の間隔が開き，その間の波形の振幅が極端に小さくなってしまう．一方，提案方法では図 3.10(b) に示すように，低域のエネルギー

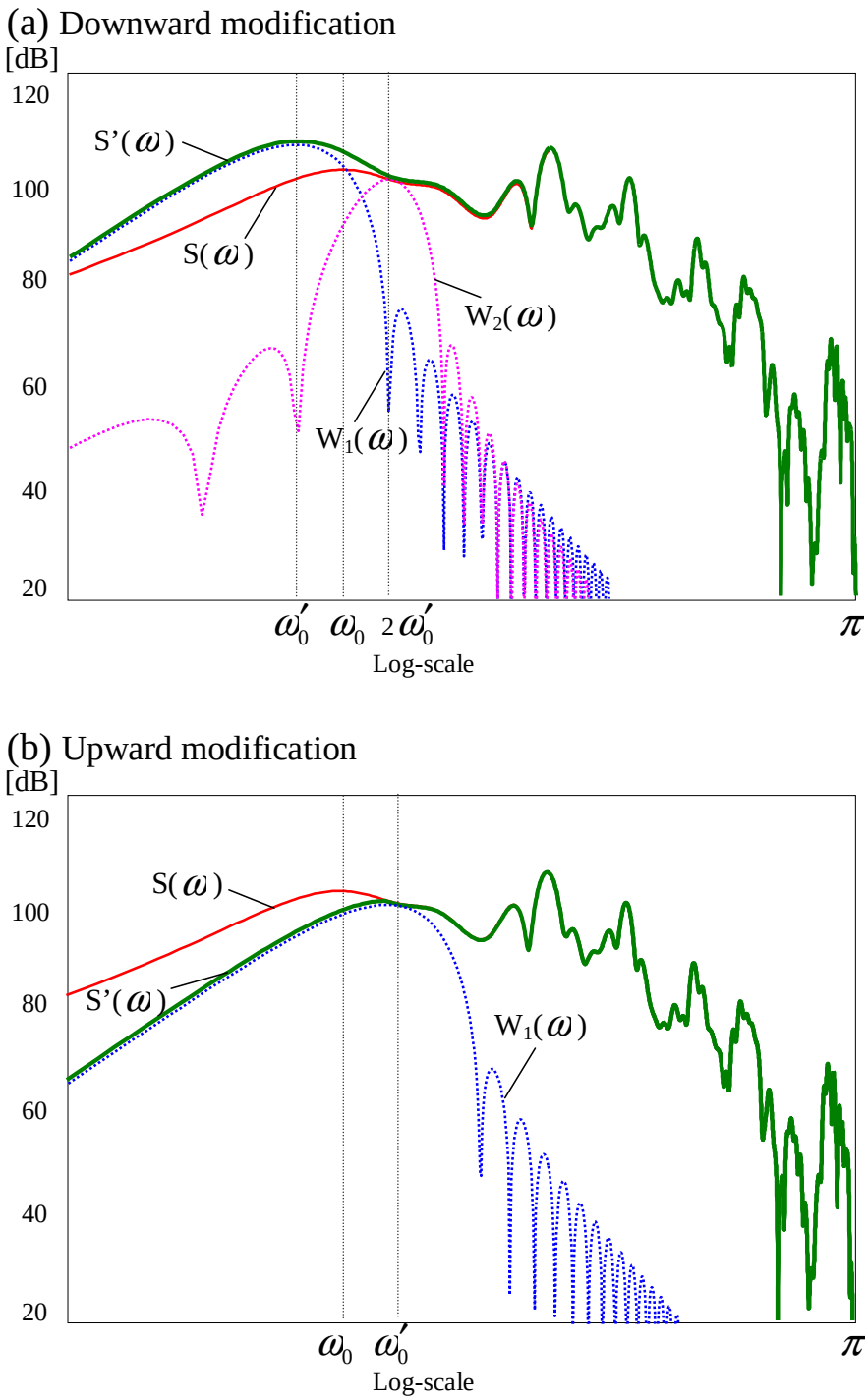


図 3.9　補正後のスペクトル包絡:(a) ピッチを低い方へ変換した場合，(b) ピッチを高い方へ変換した場合

が充実し，そのような問題が回避できる．なお，提案方法で生成した波形は，オリジナルの単位波形を時間軸方向に線形伸張した波形に似ているが，以下の点で単純な線形伸張波形とは異なる．

1) 時間軸での波形の線形伸張処理は，フォルマントの位置などを無視してすべての帯域のスペクトルを低い方へシフトすることになるため，音韻性や話者性が変化してしまう危険性を伴う．一方，提案方法では元の F_0 より低い帯域のみが補正されるだけなので，音質を変えてしまう危険性は極めて少ない．

2) 時間軸で波形を線形伸張する場合，その伸張率は主観評価などの実験結果に基づいて決定する必要があるが，評価者の好みによって最適な伸張率が異なるものと予想される．また，音韻や韻律環境の違いによっても最適な伸張率は異なることが考えられるため，一義的に決定するのが困難である．一方，提案方法ではスペクトル傾斜という物理量を保持するという考えに基づいており，元の F_0 と変換先の F_0' が決まれば，補正処理は一義的に決定できる．

3.4　スペクトル補正音声の音質評価

提案するスペクトル補正の聴感的な有効性を確認するため，ピッチを低くする方へ変換した音声の試聴実験を行った．評価に用いた音声試料は，女性ナレータ（F_0 レンジは 170〜350Hz）が発声した 6 文（3〜4sec. 程度）で，変換率はピッチが低くなる方へ，0.0 から 1.0 octave まで 0.2 octave 刻みでシフトさせた．評価は提案するスペクトル補正を行ったピッチ変換音声と，補正を行わなかったピッチ変換音声とを一対にし，順番をランダムにして評価者に提示した．評価者 8 名は，評価音声をヘッドホンで受聴し，音質が自然と感じられる方を強制的に選択した．なお，評価音声は 22050Hz サンプリング，16bit 量子化で収録されている．

評価結果を図 3.11 に示す．ピッチの変換率が 0.2octave までは，補正処理を行うことの有効性が認められなかったが，それ以上の変換率では，提案方法でスペクトル補正した音声の方が好まれる結果となった．この結果から提案方法は変換率が大きくなるにつれて有効であることがわかった．また，評価者からは，提案方

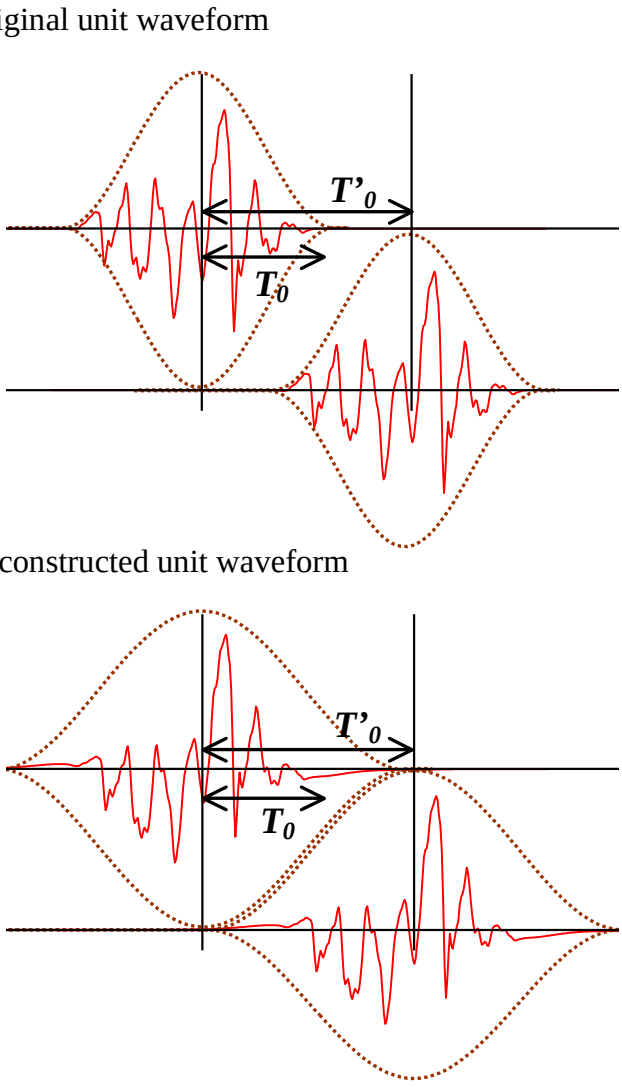


図 3.10　単位波形の再配列:(a) オリジナルの単位波形，(b) 提案方法でスペクトルを再構成した単位波形

法によって変換した音声は「かすれ感」が軽減されているというコメントを得ることができた.

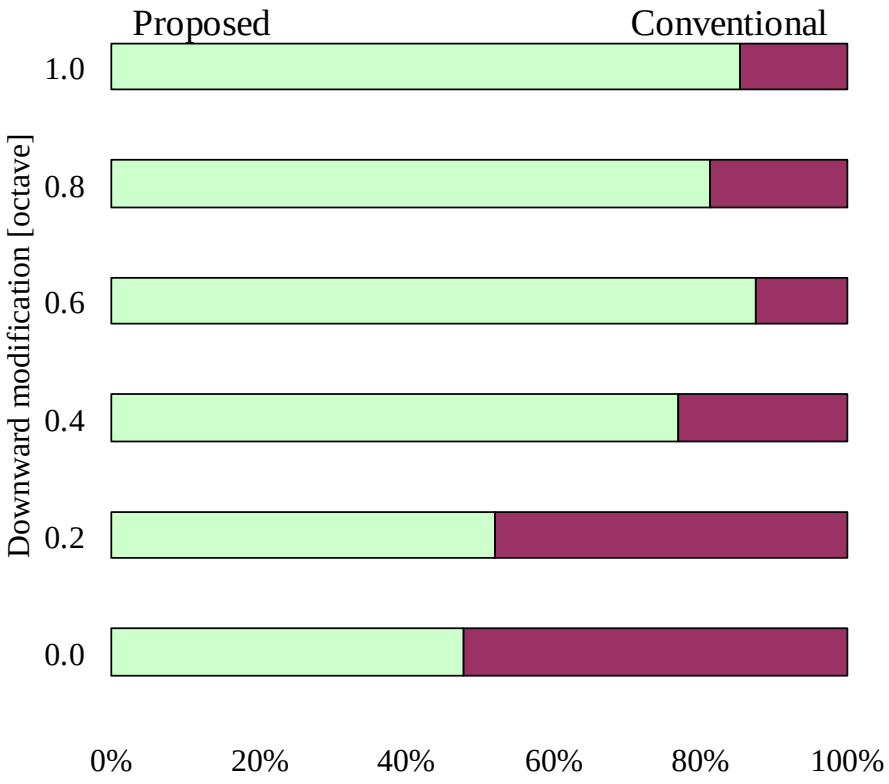


図 3.11　音質の比較評価結果:スペクトル補正処理の有無による音質の比較

3.5　考察

上述の試聴評価では，ピッチを低い方へ変換した場合のみを評価対象とした．ピッチを高い方へ変換した場合についても予備的な試聴実験を行ったが，補正処理を施した場合と，そうでない場合とで聴感上の差が確認できなかった．そこで提案方法によってピッチを高い方へ変換した場合，スペクトルにどのような変化が現れるのか調査した．図 3.12 は，ピッチを低い方へ変換した場合と，高い方へ変換した場合について，ピッチ変換後の音声から再びハニング窓によって抽出した単

位波形のスペクトル包絡である．図 3.12(a) に示すように，ピッチを低い方へ変換した場合，提案方法によって変換された音声は低域が補強されており，補正処理を行わなかった場合のスペクトル包絡とは低域が明らかに異なる．この低域の補強が上述の試聴評価における音質改善につながっているものと考えられる．一方，ピッチを高い方へ変換した場合は，図 3.12(b) に示すように補正処理を行った場合と，行わなかった場合とで，スペクトルに顕著な違いが見受けられない．ピッチを高い方へ変換する場合は，ピッチ変換後のスペクトル包絡を再現するのに必要なスペクトル情報がすべて元の単位波形のスペクトル包絡から得られるため，補正処理自体が必要ないものと考えられる．このことから，ピッチを低い方へ変換する場合にのみ補正処理を行えば良いことがわかった．

提案するスペクトル補正方法では，F_0 が変わってもスペクトル傾斜が一定であるという仮定に基づいて処理を行ったが，実際は F_0 の変化に応じて，スペクトル傾斜も変わるものと考えられる．ここで F_0 の変化とスペクトル傾斜の変化との間にある程度の相関があれば，変換先の F_0 に適したスペクトル傾斜を定式化できる可能性がある．そこで女性話者が孤立発声した母音について，F_0 とスペクトル傾斜との関係を調査した．母音/a/から単位波形を抽出し，その F_0 とスペクトル傾斜の頻度分布を図 3.13 にプロットした．単位波形のスペクトル傾斜は，F_0 によらず，およそ −10 dB/octave 付近に分布していることがわかる．この図からは，F_0 が高くなればスペクトル傾斜もそれに応じて線形的に変化するというような傾向は観察できない．この分布に対して回帰直線を求めたが，F_0 変化に対してスペクトル傾斜はほぼ一定値（傾きのない直線）となった．日本語 5 母音について調査を行ったが，いずれの場合も F_0 とスペクトル傾斜との間に強い相関は確認できなかった．以上のことから，スペクトル傾斜は声質と密接な関係があるが，F_0 に対して必ずしも線形的に変化するものではなく，単純なモデルでの定式化は困難であると考えられる．

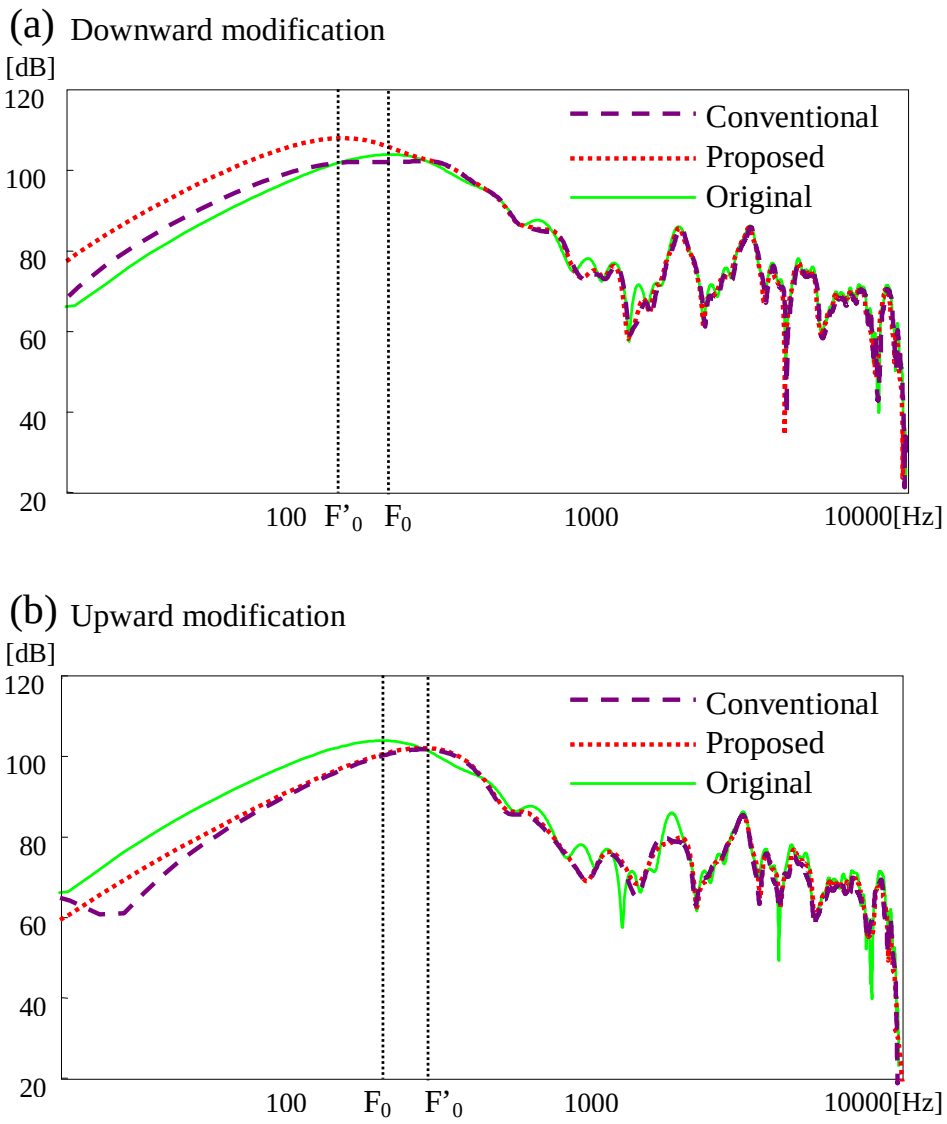


図 3.12　ピッチ変換後のスペクトル包絡:(a) 低い方へ変換した場合，(b) 高い方へ変換した場合

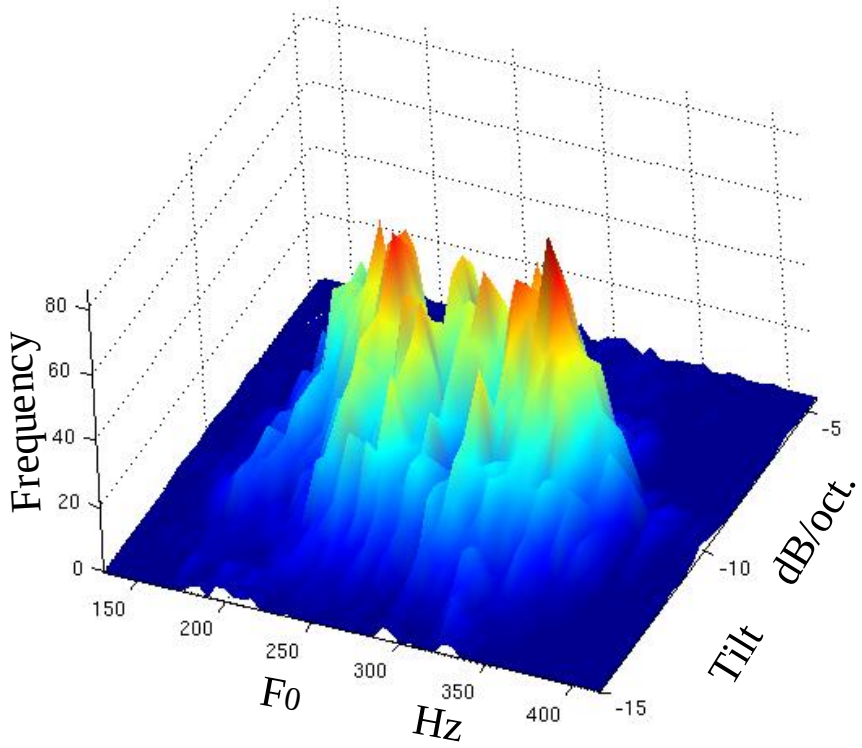


図 3.13　単位波形における F_0 とスペクトル傾斜の頻度分布:女声/a/から抽出した約 15000 個の単位波形を調査

3.6　むすび

PSOLA 法で用いる単位波形は，窓掛けの影響によって F_0 の高調波間のスペクトルが補間されるため，F_0 より高い周波数帯域では滑らかなスペクトル包絡が得られる．一方で，F_0 より低い周波数帯域には信頼できるスペクトル情報が存在しないため，F_0 を低くする方へ変換した場合，従来では低域のスペクトル包絡が正しく再現されていなかった．本研究ではこの低域におけるスペクトルの問題を解決するため，ピッチ変換後も元のスペクトル傾斜が保持されるという仮定に基づき，単位波形のスペクトルを F_0 の変換率に応じて再構成する方法を検討した．この提案方法によってピッチ変換音声を作成し，音質に関して試聴評価を行った．そ

の結果，ピッチを低い方へ 0.4octave 以上変換した場合に，提案するスペクトル補正処理の有効性が確認できた.

第4章　韻律特徴量を考慮したスペクトル変換

4.1　はじめに

近年，コーパスベースの音声合成方式によって高品質の合成が可能となった．しかし，現状合成できるのは，単調な読み上げ口調の音声に限られており，今後は感情や態度，話者性，発話口調の自由な表現を可能とすることが期待されている．このような多様な発話スタイルや話者性を制御するためには，韻律制御のみならず，スペクトルを所望の発話環境に変換する技術が強く望まれる．

一般的に高品質を目指した音声合成では，音色に統一感のある音声を合成するために，一人の話者がある一定の発話様式で発声した音声データを用いる．現在主流である大規模の音声データベースを用いたコーパスベースの音声合成では，発話者の肉声らしさが再現できる反面，あらかじめ大量の音声データを準備する必要があり，実現には膨大な労力を必要とする．特に波形接続合成方式によって複数の発話スタイルを実現するアプローチ [Iida 03] では，発話スタイルや感情ごとに音声データを収録する必要があり，その実現は容易でない．

上述の問題を解決するため，ターゲットの発話スタイルや話者に関する少量の学習データを用いて，適応によってターゲットの環境へ変換する方式が検討されている．コードブックマッピングによって 2 話者間の写像関数を作成し，スペクトル特徴量をターゲットへ変換する方法 [Abe 88] は，オフラインでクラスごとに代表の変換関数を学習し，オンラインで最適な変換関数を選択してスペクトルの変換を行う．同様のアプローチで，LSF パラメータと音源信号の二つに分離した信号を共に変換する方法 [Arslan 97] も検討されている．その他に，スペクトル特徴量を周波数軸方向へ非線形伸縮する変換方法 [Valbret 92, Maeda 99] や，コードブッ

クマッピングによってターゲットのフォルマント位置を推定しておき，このターゲットのフォルマントに向けて元のフォルマントをシフトする変換方法 [Mizuno 95] などが提案されている．また，話者変換という観点では，2 話者間の変換に関するもの以外に，複数話者の音声データを用いて，補間によって新しい話者の音声を生成する方法 [Iwahashi 95] が検討されている．HMM（Hidden Malcov Model）を用いた音声合成方式 [Tokuda 95, Masuko 96] では，あらかじめ複数の話者の学習データによって平均声モデルを生成しておき，MLLR（Maximum Likelihood Linear Regression）[Gales 96] によって HMM の適応を行い，ターゲット話者の音声を合成する方法 [Tamura 01] が提案されている．

近年では，GMM（Gaussian Mixture Model）を用いたスペクトル変換 [Stylianou 95] に関連する多くの研究が進められている．GMM を用いたスペクトル変換では，クラスごとに対応付けられた変換元と変換先のスペクトルの写像関数に対して，入力されたスペクトルが各クラスに帰属する確率に応じて重み付けを行い，最適な変換関数を生成する．このため入力に対して特定の変換関数を選択するベクトル量子化による方法 [Abe 88] より，安定した変換が可能であるという長所を持つ．しかしながら，この GMM を用いた方法においても，統計処理による過剰なスペクトルの平滑化が音質劣化の原因となっており，これを解決するために，周波数軸方向への非線形スペクトル伸縮をする方法 [Toda 01]，変換元と変換先の分散に関して強い相関を仮定し，MAP（Maximum A Posteriori）法によってパラメータを推定する方法 [Chen 03]，変換先のスペクトル微細構造を推定する方法 [Kain 01]，コードブック，または学習データから微細構造を選択する方法 [Ye 04]，選択後にスムージングによって安定化させる方法 [Suendermann 05]，発話内の分散自体が保持されるように分散のモデルを導入した方法 [Toda 05] などが提案されている．また，最近ではスペクトル特徴量の変換に加えて，基本周期に観測される Jitter を変換する方法 [Verma 05] や，Glottal formant をスペクトル包絡から分離し，独立のモデルで変換する方法 [Qin 05] などが検討されている．

一方，ある環境で発声された音韻のスペクトルは，発声時の基本周波数や発話コンテキストと密接な関わりがあることが一般的に知られている．すなわち，同じ音素であっても，発声される基本周波数の違いや，前後の音素の影響によってス

ペクトルが変化することになる．従来の GMM を用いたスペクトル変換方式では，変換元のスペクトルと変換先のスペクトルとを 1 対 1 で対応付けることで写像関数を定義していたが，モデル化の際に明示的に韻律情報を用いることで，更に精度の良い変換モデルが期待できる．特にスペクトル変換を音声合成の枠組で利用することを想定した場合，合成する音素系列情報や合成に利用するターゲットの韻律情報は与えられるため，これらの情報を有効に活用することが望まれる．そこで本研究では，規則合成での利用を前提とし，韻律情報を特徴量に用いた GMM によるスペクトル変換方式について検討する．本章では，まず GMM を用いたスペクトル変換方式について説明し，その応用である韻律情報を用いた変換方式を提案する．続いて提案するスペクトル変換方式を話者変換に応用する．変換モデルの学習には，より多くの韻律コンテキストを含む学習データを確保するために，トライフォン単位で結合ベクトルを作成する方法について検討する．提案方法によってスペクトル変換した音声の変換精度を評価するために，ケプストラム距離を用いた物理評価を行う．また，試聴評価を行い，音質，及び話者性について，提案方法の有効性を示す．なお，非同一発話文で構成される学習データを用いた場合についても，変換モデルの学習を検討する．

4.2　GMM を用いたスペクトル変換モデル

GMM を用いたスペクトル変換方法について説明する．この変換方法では，変換元と変換先との写像関係を GMM によって定義する．GMM のモデルパラメータは，変換元の特徴量とターゲットの特徴量との結合ベクトルを用いる JDE（Joint Density Estimation）法 [Kain 98] によって推定する．

$$\boldsymbol{z} = [\boldsymbol{x}^T \boldsymbol{y}^T]^T, \tag{4.1}$$

ここで $\boldsymbol{x}$ は変換元のスペクトルに関する特徴ベクトル，$\boldsymbol{y}$ は変換先のスペクトルに関する特徴ベクトルである．この結合ベクトルの確率密度分布は，GMM によっ

て以下のように表すことが可能である．

$$P(\boldsymbol{z}) = \sum_{i=1}^{q} \alpha_i N(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad \sum_{i=1}^{q} \alpha_i = 1, \alpha_i \geq 0, \tag{4.2}$$

ここで $N(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ は平均ベクトル $\boldsymbol{\mu_i}$ と共分散行列 $\boldsymbol{\Sigma}_i$ によって表される正規分布であり，q は混合数，α_i は i 番目の分布の重みを表す．

$$N(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{p/2} |\ \boldsymbol{\Sigma_i}\ |^{1/2}} \exp\{-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_i)^T \Sigma (\boldsymbol{z} - \boldsymbol{\mu}_i)\}, \tag{4.3}$$

また，平均ベクトル $\boldsymbol{\mu_i}$ と共分散行列 $\boldsymbol{\Sigma}_i$ は，以下のように表せる．

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^x \\ \boldsymbol{\mu}_i^y \end{bmatrix}, \tag{4.4}$$

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy} \\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy} \end{bmatrix}, \tag{4.5}$$

ここで $\boldsymbol{\mu}_i^x$ と $\boldsymbol{\mu}_i^y$ は入力と出力の平均ベクトル，$\boldsymbol{\Sigma}_i^{xx}$，$\boldsymbol{\Sigma}_i^{xy}$，$\boldsymbol{\Sigma}_i^{yx}$，$\boldsymbol{\Sigma}_i^{yy}$ は共分散行列を表す．GMM のパラメータは EM（Expectation Maximization）アルゴリズムによって推定する．入力となるスペクトル特徴量 $\boldsymbol{x}$ が与えられたとき，尤度関数 $\log P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ を最大にする $\boldsymbol{y}$ は，次式で与えられる．

$$\boldsymbol{y} = F(\boldsymbol{x}) = \sum_{i=1}^{q} h_i(\boldsymbol{x})[\boldsymbol{\mu}_i^y + \boldsymbol{\Sigma}_i^{yx} \boldsymbol{\Sigma}_i^{xx-1} (\boldsymbol{x} - \boldsymbol{\mu}_i^x)], \tag{4.6}$$

すなわち，上式はスペクトル特徴量 $\boldsymbol{x}$ を $\boldsymbol{y}$ に変換する式と考えることができる．なお，ここで $h_i(\boldsymbol{x})$ は $\boldsymbol{x}$ の i 番目の混合成分における事後確率を表す．

$$h_i(\boldsymbol{x}) = \frac{\alpha_i N(\boldsymbol{x}; \boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^{xx})}{\sum_{j=1}^{q} \alpha_j N(\boldsymbol{x}; \boldsymbol{\mu}_j^x, \boldsymbol{\Sigma}_j^{xx})}. \tag{4.7}$$

4.3　韻律情報を考慮したスペクトル変換モデル

ある音素におけるスペクトルの特徴が，発話時の基本周波数や前後の音素コンテキストに依存すると仮定すると，変換元と変換先のスペクトルの対応は，相互の発話環境を考慮した上でマッピングすることが望まれる．提案するスペクトル変換方式は，基本的に変換元と変換先の両方のスペクトル特徴量を GMM でモデル化する方法に基づき，韻律特徴量をモデルのパラメータに加えることで，変換精度の向上を狙った方式である．まず，モデルの学習に利用する結合ベクトル $\boldsymbol{z}$ は，変換元と変換先の特徴量ベクトルによって以下のように構成する．

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x}_p^T & \boldsymbol{x}_s^T & \boldsymbol{y}_p^T & \boldsymbol{y}_s^T \end{bmatrix}^T, \tag{4.8}$$

ここで $\boldsymbol{x}_p$ と $\boldsymbol{x}_s$ はそれぞれ変換元のスペクトルと韻律情報を表すベクトル，$\boldsymbol{y}_p$ と $\boldsymbol{y}_s$ は変換先のスペクトルと韻律情報を表すベクトルである．提案方法では，$\boldsymbol{x}_p$，$\boldsymbol{x}_s$，$\boldsymbol{y}_p$ を入力として，ターゲットのスペクトル $\boldsymbol{y}_s$ を求めることを考える．結合ベクトル $\boldsymbol{z}$ の確率密度関数は，GMM によって以下のように定義される．

$$p(\boldsymbol{z}) = \sum_{i=1}^{q} \alpha_i N(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad \sum_{i=1}^{q} \alpha_i = 1, \alpha_i \geq 0, \tag{4.9}$$

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^{x_p T} & \boldsymbol{\mu}_i^{x_s T} & \boldsymbol{\mu}_i^{y_p T} & \boldsymbol{\mu}_i^{y_s T} \end{bmatrix}^T, \tag{4.10}$$

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{x_p x_p} & \boldsymbol{\Sigma}_i^{x_p x_s} & \boldsymbol{\Sigma}_i^{x_p y_p} & \boldsymbol{\Sigma}_i^{x_p y_s} \\ \boldsymbol{\Sigma}_i^{x_s x_p} & \boldsymbol{\Sigma}_i^{x_s x_s} & \boldsymbol{\Sigma}_i^{x_s y_p} & \boldsymbol{\Sigma}_i^{x_s y_s} \\ \boldsymbol{\Sigma}_i^{y_p x_p} & \boldsymbol{\Sigma}_i^{y_p x_s} & \boldsymbol{\Sigma}_i^{y_p y_p} & \boldsymbol{\Sigma}_i^{y_p y_s} \\ \boldsymbol{\Sigma}_i^{y_s x_p} & \boldsymbol{\Sigma}_i^{y_s x_s} & \boldsymbol{\Sigma}_i^{y_s y_p} & \boldsymbol{\Sigma}_i^{y_s y_s} \end{bmatrix}, \tag{4.11}$$

ここで $N(\boldsymbol{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ は平均 $\boldsymbol{\mu}_i$，分散 $\boldsymbol{\Sigma}_i$ の正規分布を表し，q は混合数，α_i は重みを表す．続いて変換元の韻律情報とスペクトルに加え，変換先の韻律情報が与えられると仮定すると，入力と出力の関係から，結合ベクトル $\boldsymbol{z}$ を以下のように定義しなおすことができる．

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{v}^T & \boldsymbol{w}^T \end{bmatrix}^T, \tag{4.12}$$

$$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{x}_p^T & \boldsymbol{x}_s^T & \boldsymbol{y}_p^T \end{bmatrix}^T, \quad \boldsymbol{w} = \boldsymbol{y}_s, \tag{4.13}$$

すなわち，$\boldsymbol{z}$ の確率密度関数を構成するパラメータは，

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^v \\ \boldsymbol{\mu}_i^w \end{bmatrix}, \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{vv} & \boldsymbol{\Sigma}_i^{vw} \\ \boldsymbol{\Sigma}_i^{wv} & \boldsymbol{\Sigma}_i^{ww} \end{bmatrix}, \tag{4.14}$$

と置き換えることができる．$\boldsymbol{\mu}_i^v$，$\boldsymbol{\mu}_i^w$ は入力と出力の平均ベクトル，$\boldsymbol{\Sigma}_i^{vv}$，$\boldsymbol{\Sigma}_i^{vw}$，$\boldsymbol{\Sigma}_i^{wv}$，$\boldsymbol{\Sigma}_i^{ww}$ は共分散行列を表す．ここで韻律とスペクトルとの相互の相関を扱うため，これらの共分散行列は全共分散行列を用いる．各 GMM パラメータは，最尤推定により求めることができる．上述のモデルパラメータを用いて，変換関数 $F(\boldsymbol{v})$ は，各分布で定義された写像関数の重み付け和で表される．

$$F(\boldsymbol{v}) = \sum_{i=1}^{q} h_i(\boldsymbol{v})[\boldsymbol{\mu}_i^w + \boldsymbol{\Sigma}_i^{wv}\boldsymbol{\Sigma}_i^{vv-1}(\boldsymbol{v} - \boldsymbol{\mu}_i^v)], \tag{4.15}$$

ここで i 番目の正規分布における事後確率 $h_i(\boldsymbol{v})$ は次式で与えられる．

$$h_i(\boldsymbol{v}) = \frac{\alpha_i N(\boldsymbol{v}; \boldsymbol{\mu}_i^v, \boldsymbol{\Sigma}_i^{vv})}{\sum_{j=1}^{q} \alpha_j N(\boldsymbol{v}; \boldsymbol{\mu}_j^v, \boldsymbol{\Sigma}_j^{vv})}. \tag{4.16}$$

提案方法は韻律情報を考慮して GMM の各正規分布の重みを推定することができる．また，韻律情報とスペクトルとの相関関係を考慮してターゲットのスペクトルを推定できるため，従来の韻律特徴量を用いない方法と比較して自由度の高い変換が期待できる．なお，スペクトルと基本周波数とを一緒に変換する方法 [En-Najjary 04] は，スペクトルと基本周波数とを合わせてターゲットの空間に変換する方式であって，ターゲットの任意の韻律に対して最適なスペクトルを推定するという方法にはなっていない．一方，提案する変換方法は，ターゲットの韻律情報を与えると，その韻律に対して最適なスペクトルを推定することができる方法となっている．

4.4　話者変換への応用

韻律情報を用いたGMMによるスペクトル変換方式を話者変換に応用する．韻律に関する特徴量を用いる提案方法では，元話者とターゲット話者とのモデル学習を行う際に，韻律コンテキストに関して豊富な学習データを用意することが望まれる．従来の研究では同一発話文セットを用いて，対応する音素位置のみで学習データを収集する方法が用いられてきた．この方法は音素コンテキストが完全に一致する対応データを用いるため，スペクトルの時間変化に対しても安定した学習が期待できるが，その反面，類似の韻律コンテキストを持つ発話間での対応データしか得られないため，韻律に関してバライエティのある学習データを獲得するという点では効率が悪い．そこで学習データに韻律のバライエティを増やすために，トライフォン単位の対応付けによって学習データを獲得する方法について検討する．以下，まずスペクトル変換モデルの学習方法について説明し，続いて話者変換のためのスペクトル変換処理について説明する．

4.4.1　学習データの収集方法

GMMの学習には，元話者の音声とターゲット話者の音声とから生成した結合ベクトルを用いる．あらかじめ音素境界ラベルが付与された元話者，及びターゲット話者の音声データをトライフォン単位に分割し，同じ音素並びを持つトライフォンのグループに分類する．続いて各トライフォンのグループにおいて，元話者とターゲット話者のすべてのデータの組み合わせをとり，その中心音素において結合ベクトルを生成する．最終的にトライフォンの中心音素で作られた結合ベクトルを音素ごとにまとめなおし，各音素のGMMを学習する．

図4.1の例では，中心音素が/a/のトライフォンに着目して元話者とターゲット話者のデータの組み合わせをとり，中心音素/a/における結合ベクトルを作成している．この音素/a/のGMMは，中心音素が/a/の結合ベクトルをすべて用いて学習する．元話者とターゲット話者の音素位置の対応付けは事前に手作業で付与された音素境界ラベルを利用する．音素内での元話者とターゲット話者の学習データの

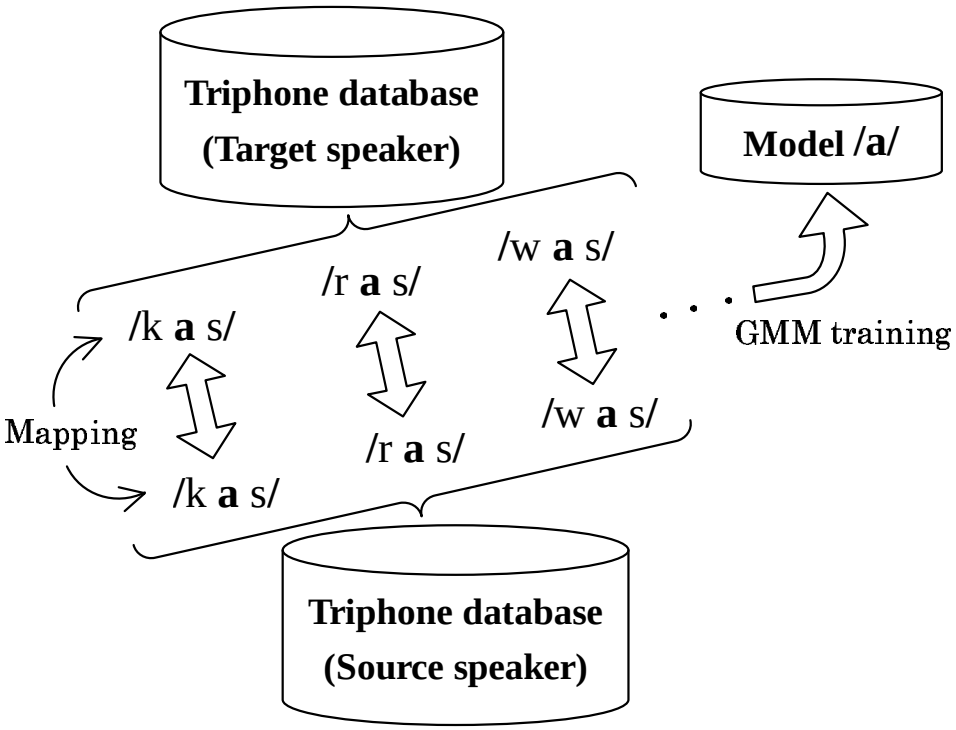


図 4.1　音素ごとの学習データの収集:音素/a/に関する学習データの収集例

対応付けは，図 4.2 に示すようにピッチ同期で行う．すなわち，第 2 章で提案したピッチマーキング法によりピッチマークを求め，このピッチマークを基準に抽出した単位波形を元話者とターゲット話者とで対応付ける．この際，単位波形の対応付けは，単位波形から求めた 32 次のケプストラムを用いて DP マッチングにより行う．以上のようにして元話者とターゲット話者との対応付けを行い，それぞれの組み合わせからスペクトル，韻律に関する特徴量を抽出して結合ベクトルを生成する．

提案する学習データの収集方法では，話者間の写像を学習するのに必要な結合ベクトルをその前後の音素コンテキストを考慮して生成している．その一方で，前後の音素コンテキストの違いは，GMM の混合分布によって表現されるものと仮定し，GMM の学習は音素ごとに行う．上述のトライフォン単位で対応付けして音素ごとの変換モデルを学習する方法は，従来の同一発話文セットの代わりに非同一

発話文セットを用いる場合にも利用可能である．また，同一発話文セットを用いる場合は，従来の発話文単位での対応付けで得られる学習データを包含するため，安定したスペクトル変換モデルの構築を達成した上で，韻律情報を用いることによる更なるスペクトル変換精度の改善が期待できる．

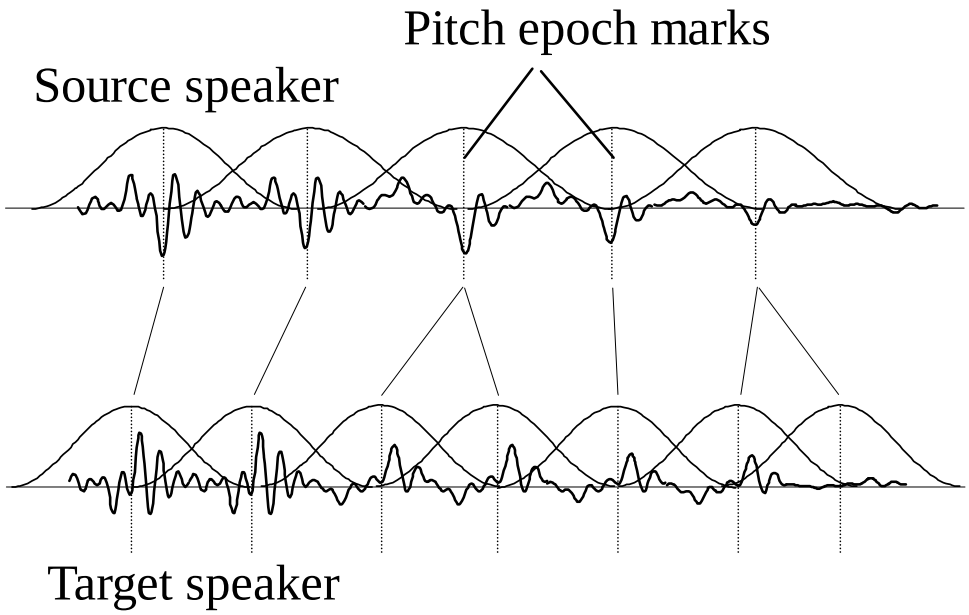


図 4.2　結合ベクトル作成のための単位波形の対応付け

4.4.2　スペクトル変換処理

話者変換のためのスペクトル変換処理のフローを図 4.3 に示す．元話者の音声に対して，TD-PSOLA 法と同じ要領でピッチマークを基準に基本周期の 2 倍の窓長を持つハニング窓で単位波形を抽出し，その単位波形に対して 32 次のケプストラムを計算する．このケプストラムを元話者のスペクトル特徴量 $\boldsymbol{x}_s$ とする．一方，元話者の音声から抽出する韻律特徴量 $\boldsymbol{x}_p$ として，対数 F_0，対数 F_0 の差分値，対数パワー，音素時間長を用いる．ここで F_0 はピッチ同期分析で得られるピッチマークの間隔から求めた瞬時 F_0 を用い，音素時間長は音素ごとに平均値で正規化した

値を用いる．元話者の特徴量 $\boldsymbol{x}_s$，$\boldsymbol{x}_p$ に加え，ターゲットとして与えられる韻律特徴量 $\boldsymbol{y}_p$ を用いて入力ベクトルを生成する．続いて音素ごとに生成した GMM の変換モデルを用いて，入力ベクトルからターゲットのケプストラム $\boldsymbol{y}_s$ への変換を行う．上述の要領で得られたターゲットのケプストラムに対して，位相特性を付加して時間波形を獲得する．位相特性に関してもターゲット話者の特性に近づけることが望まれるが [Ye 04]，本研究では元話者の位相特性をそのまま用いた．

4.5　ケプストラム距離を用いた変換精度の測定

提案する韻律情報を用いたスペクトル変換モデルの変換精度を評価するために，話者変換音声を用いて物理評価を行った．評価尺度にはスペクトル変換により話者を変換した音声と，ターゲット話者の自然音声とのケプストラム距離を用いた．以下，まず物理評価の実験条件について説明し，トライフォン単位の対応付けで学習した場合と，従来の発話文単位で学習した場合の 2 条件において，韻律情報を用いた提案方法の有効性を評価する．また，韻律特徴量を個別に加えた場合，どの特徴量が変換精度の改善に有効であるのか調査する．更に，非同一発話文セットを用いてモデル学習を行った場合についても，変換精度の評価を行う．

4.5.1　変換モデルの学習条件

3 人の女性ナレータが読み上げ口調で発声した同一発話文セットを用いて，各話者間の相互変換を行うスペクトル変換モデルを作成した．学習には元話者とターゲット話者とで，同じ発声内容の文で構成される同一発話文セットを用いた．学習に用いた文章は音素バランスを考慮した 50 文である．提案方法では韻律特徴量として，対数 F_0，対数 F_0 の差分値，対数パワー，音素時間長を用いた．評価データにはモデル学習に使用した文章とは別の同一発話文セット 50 文を用いた．評価は 3 人の話者の相互変換を行い，計 6 通りの変換音声についてケプストラム距離を求め，その平均値を評価結果として集計した．ケプストラム距離は，音素ごとに変換した 32 次のケプストラムと，対応するターゲット話者の自然音声のケプス

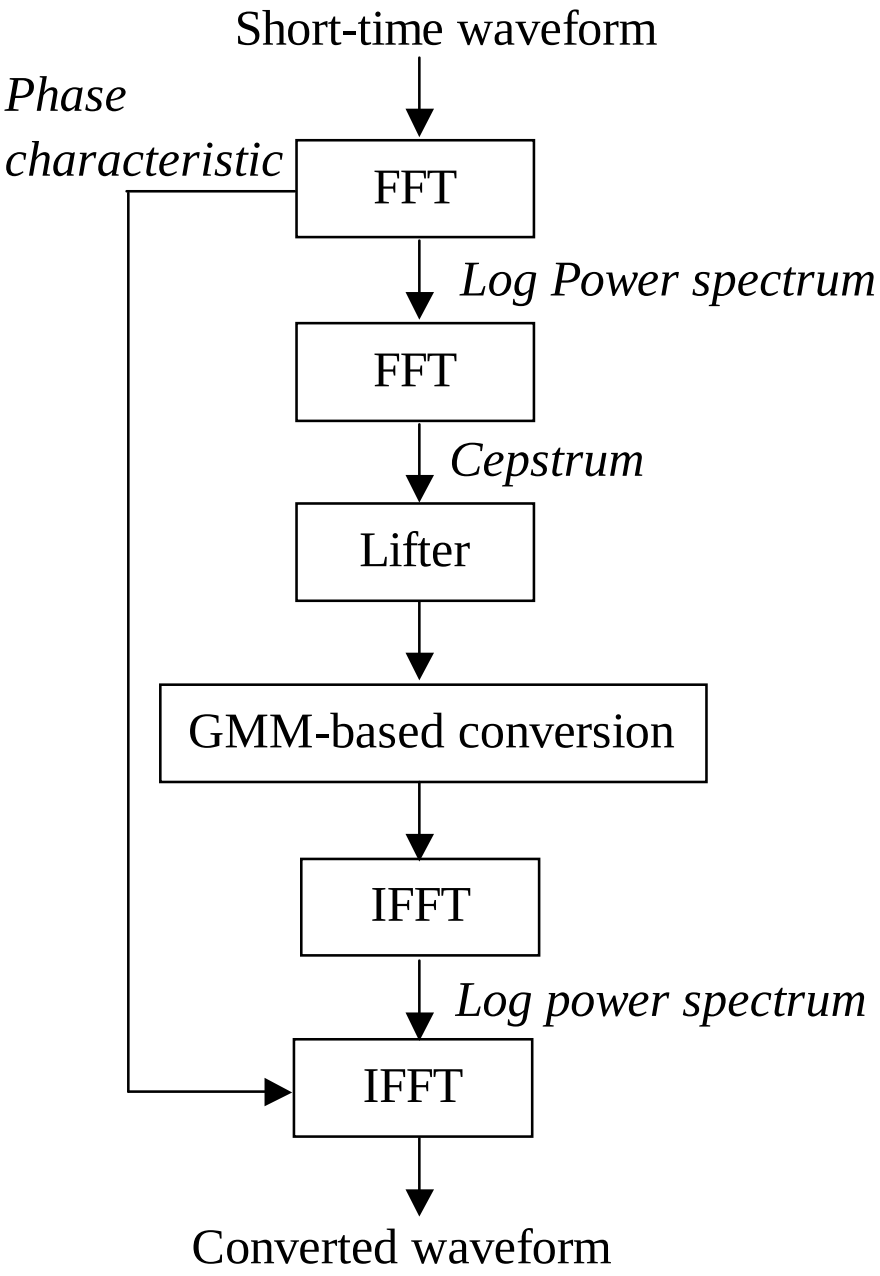


図 4.3　PSOLA 法をベースとしたスペクトル変換処理

表 4.1　同一発話文セットにおける各母音の学習データ数:それぞれ５０文を使用, (a) 発話文単位で対応付け学習した場合，(b) トライフォン単位で対応付け学習した場合

vowel	a	i	u	e	o
(a)Utterance	280	187	193	142	287
(b)Triphone	1544	678	957	476	1167

トラムから計算した．無声子音に関しては F_0 が得られないなどの制約があり，韻律情報を用いることの有効性が顕著でないと判断し，評価対象から外した．計 17 の有声の音素に対して変換モデルを学習した．また，GMM の混合数は全体の評価結果が最適化されるように実験的に策定し，トライフォン単位で対応付け学習した場合は，母音における混合数を 16，有声子音における混合数を 4 で統一した．評価実験で用いた音声は，22050Hz サンプリング，16bit 量子化で収録されている．

4.5.2　学習データ収集方法ごとの評価

スペクトル変換モデルに韻律情報を用いた場合と，用いなかった場合の変換精度について，トライフォン単位の対応付けで学習した場合と，従来の発話文単位の対応付けで学習した場合の 2 条件について評価を行った．参考のため，表 4.1 に各母音の学習データ（音素）数を示す．この表から，トライフォン単位で対応付けすることで，従来の文単位での対応付けより，およそ 3〜5 倍の学習データを集められることがわかる．これにより，従来よりも韻律に関してバライエティのある学習データの収集が期待できる．ここで従来の発話文単位の対応学習を用いた場合の GMM の混合数は，トライフォン単位で学習データを集めた場合の半分に設定した．図 4.4 に，それぞれの学習条件において韻律情報を用いた場合と用いなかった場合の平均ケプストラム距離，及びその 95%の信頼区間を示す．今回の実験では，いずれの学習データの収集方法を用いた場合でも，提案する韻律情報を

用いたスペクトル変換方法の有効性が確認できた．また，韻律情報を用いたスペクトル変換は，トライフォン単位での対応付け学習を行うことにより，従来の発話文単位の学習より変換精度が若干改善できた．その一方で，韻律情報を用いない従来のスペクトル変換の場合は，トライフォン単位での対応付け学習を用いることで，わずかながら変換精度が劣化する結果となった．このことから，トライフォン単位での対応付け学習は，韻律に関してバライエティのある学習データを収集することができるため，韻律情報を明示的に用いている提案方法においては有効だったと考えられる．逆にトライフォン単位での対応付けは，韻律コンテキストの異なる対応関係を含んだ学習データが生成されるため，韻律情報を用いない従来方法には悪影響を及ぼしたものと考えられる．

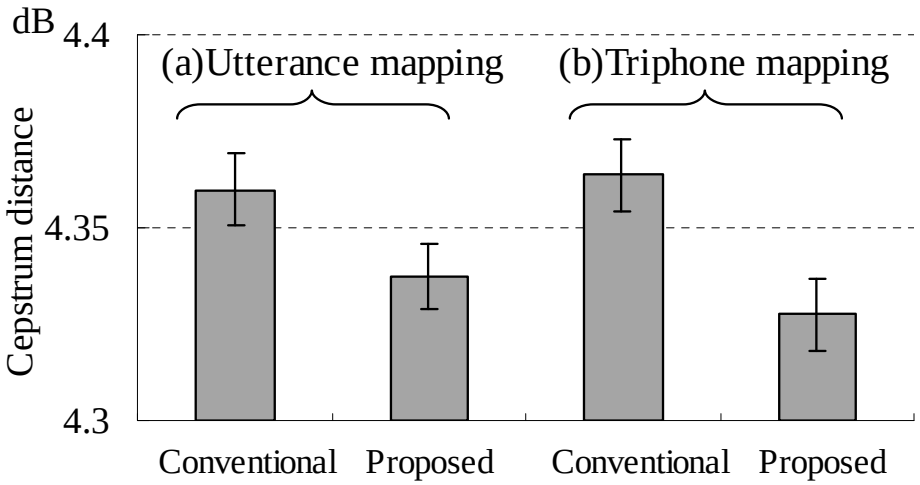


図 4.4　各学習方法を用いた場合の韻律情報の有効性:(a) 発話文単位で対応付け学習した場合，(b) トライフォン単位で対応付け学習した場合

4.5.3　各音韻における評価

音韻の違いによる変換精度のばらつきを調査するため，上述の要領で測定したケプストラム距離を母音ごとに集計しなおした．韻律を考慮した変換モデルを用

いた場合と，考慮していない変換モデルを用いた場合の結果を図 4.5 に示す．この結果は，トライフォン単位の対応付け学習によって生成した変換モデルを使用した場合のものである．本研究ではスペクトル変換モデルを各音素ごとに生成しているが，この結果からすべての母音において韻律情報を用いることの有効性を確認することができた．例として，母音/a/について変換した音声のスペクトル包絡を図 4.6 に示す．図 4.6(a) はターゲット話者の自然音声におけるスペクトル包絡であり，このスペクトル包絡と比較すると，変換音声のスペクトル包絡は十分にターゲットに近いとは言えないが，韻律情報を用いた場合（図 4.6(b)）の方が，韻律情報を用いない場合（図 4.6(c)）より，わずかながらフォルマントにおけるスペクトルの強弱がはっきりしている．提案方法で変換した音声は，従来方法と比較してスペクトルの過剰な平滑化が改善できているものと考えられる．

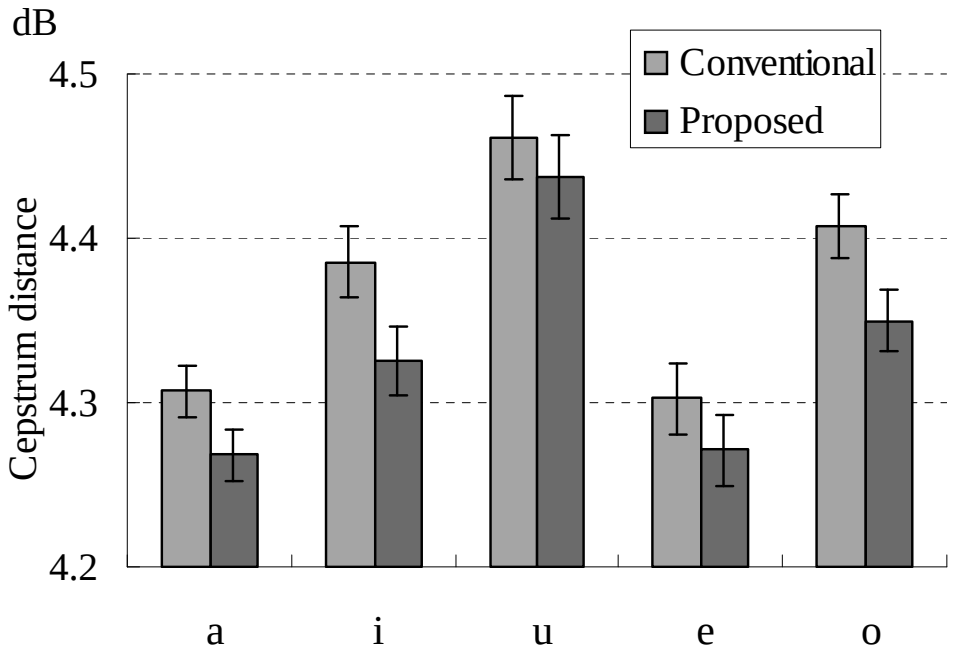


図 4.5　母音ごとの平均ケプストラム距離

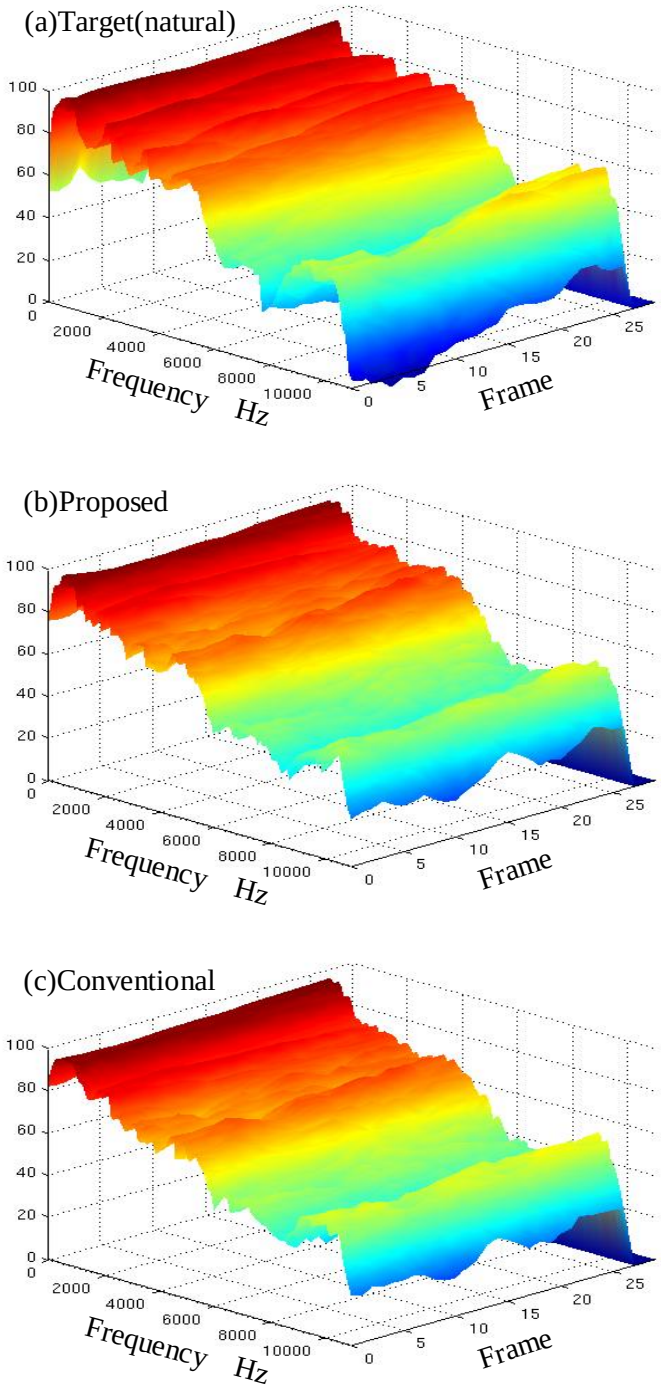


図 4.6　話者変換した母音/a/のスペクトル包絡:(a) ターゲット話者（自然音声），(b) 韻律を考慮したモデルで変換，(c) 韻律を考慮していないモデルで変換

4.5.4　各韻律パラメータの評価

提案方法において，特に有効な韻律特徴量が何であるのかを示すために，各韻律特徴量を個別に加えて生成した変換モデルを用いて物理評価を行った．図 4.7 に，韻律特徴量なし，音素時間長，対数パワー，対数 F_0，対数 F_0 とその差分値，対数 F_0 と対数パワー，及びすべての特徴量を用いた場合の平均ケプストラム距離とその 95%の信頼区間を示す．この結果から，特に対数 F_0 を加えることの有効性が顕著であり，続いて対数パワーが有効であることがわかった．ただし，対数 F_0 に更に対数 F_0 の差分を加えても，それ以上の顕著な改善は確認できなかった．また，音素時間長を単独で用いた場合も顕著な改善は確認できなかった．なお，すべての特徴量を用いた場合が，他の条件の場合よりも若干良い結果となったため，以下，本章における試聴評価では，すべての韻律特徴量を用いた場合を提案方法として用いる．

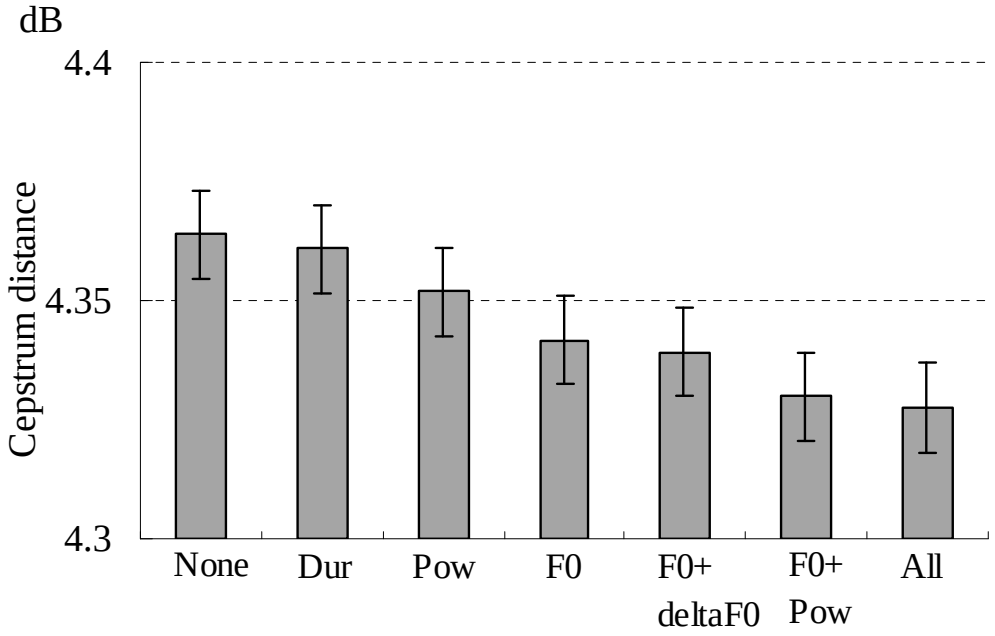


図 4.7　各韻律パラメータの影響:各韻律情報を個別，または組み合わせて用いた場合の平均ケプストラム距離

表 4.2　異なる学習文セットにおける各母音の学習データ数:それぞれ５０文を使用，(a) 同一発話文セット，(b) 非同一発話文セット

vowel	a	i	u	e	o
(a)	280	187	193	142	287
(b)	1466	576	937	442	1032

4.5.5　非同一発話文による学習の評価

提案するトライフォン単位の対応学習は，従来の発話文単位の学習と異なり，学習データが十分にあれば同一発話文セットでなくても学習が可能である．上述の実験では同一発話文セットを用いて学習を行っているが，ここでは非同一発話文セットを用いて学習した場合に，どの程度安定した変換が可能なのか調査した．元話者とターゲット話者との学習データに非同一発話文セット，及び比較のために同一発話文セットを用いて，日本語 5 母音について変換モデルを生成した．学習に用いた文章の数は同一発話文セット，非同一発話文セットともに 30，50，70 文である．ただし，同一発話文セットを用いる場合は，従来の発話文単位での対応学習を用いた．表 4.2 に 50 文を用いたときの各母音における学習データ（音素）数を示す．非同一発話文セットを用いた場合でも，トライフォン単位での対応学習を行うことで，従来の発話文単位で対応学習する場合より多くの学習データを集められることがわかる．図 4.8 に，各方式における平均ケプストラム距離を示す．この結果から，非同一発話文セットを用いた場合でも，韻律情報を考慮した変換モデルを使用することで，従来の同一発話文セットで学習した場合より良好な結果を得ることができた．

4.6　話者変換音声の試聴評価

韻律情報を考慮したスペクトル変換モデルの聴感的な有効性を示すために，話者変換音声を作成し，音質と話者性について試聴評価を行った．以下，音質のプ

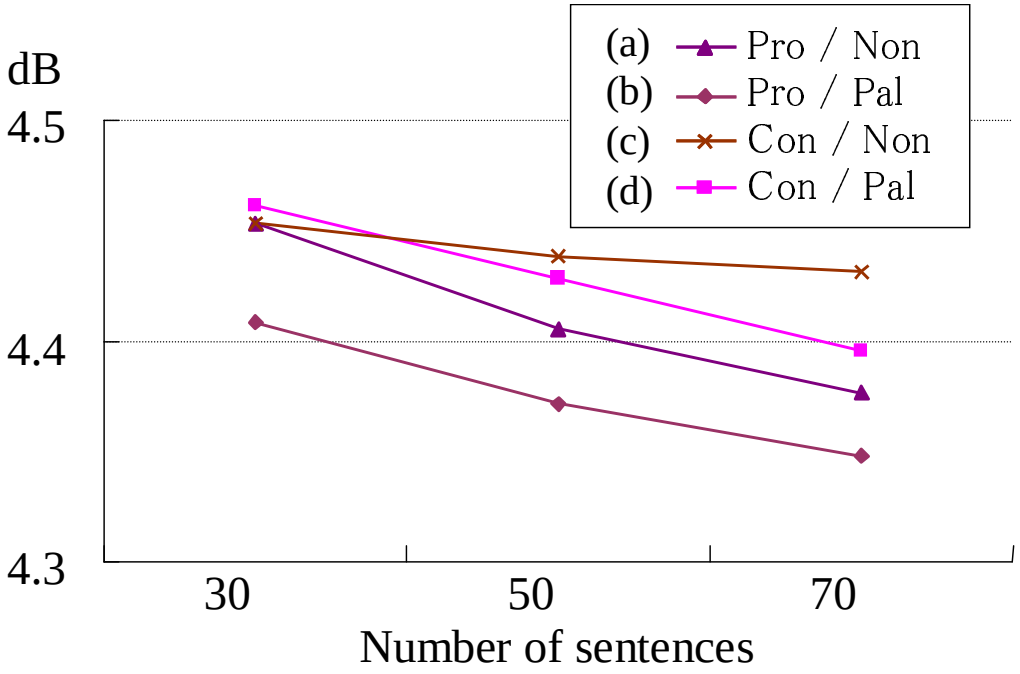


図 4.8　非同一発話文セットを用いた場合の平均ケプストラム距離:(a) 韻律を考慮した変換モデル+非同一発話文セット，(b) 韻律を考慮した変換モデル+同一発話文セット，(c) 韻律を考慮していない変換モデル+非同一発話文セット，(d) 韻律を考慮していない変換モデル+同一発話文セット

リファレンス評価，及び話者の識別評価について述べる．また，非同一発話文セットを用いて学習した変換モデルにより作成した変換音声についても，音質，及び話者性について評価する．

4.6.1　音質の比較評価

韻律情報を考慮した変換モデルを用いた場合と，考慮していない変換モデルを用いた場合の変換音声について，音質の差を比較評価した．評価音声の生成には，図 4.9 に示す音声合成の枠組を利用する．F_0 や継続時間長などの韻律変換にはピッチ同期で単位波形を重畳する TD-PSOLA 法を用いる．スペクトル変換は，上述の物理評価で構築したスペクトル変換モデルを用いて，音素ごとに元話者からターゲット話者への変換を行う．ここで音声合成の枠組を利用して生成した音声を試

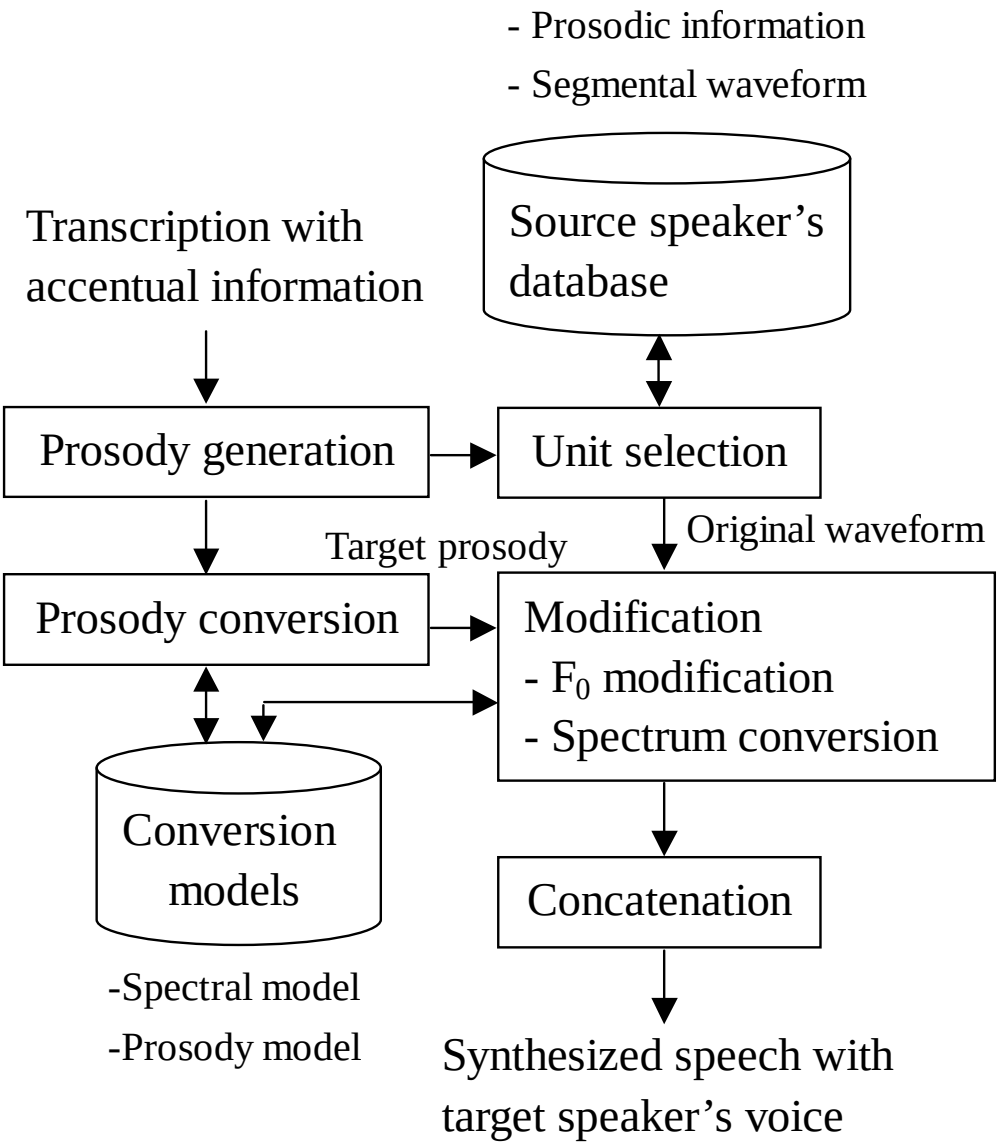


図 4.9　話者変換機能を備えた音声合成システム

聴評価に用いる場合，その音質は，スペクトル変換の影響だけでなく，(1) 素片選択，(2) 接続処理，(3) 韻律変換で生じる歪の問題が加わる．このため純粋にスペクトル変換のみの性能を評価することが困難となる．そこで今回の評価では，元話者のデータベースにある音声を合成することで，合成に使用する素片セットはクローズの条件とし，スペクトル変換モデルの学習データと評価音声とはオープンの条件となるようにした．これにより素片選択と接続処理は行わないのと等価になり，韻律変換とスペクトル変換処理の影響だけを含んだ評価音声が生成できる．また，ターゲットの韻律パタンを規則によって生成した場合，評価者が音質ではなく，韻律の不自然性に着目して判断する危険性があるため，元話者の韻律パタンをターゲット話者のレンジに合うようにシフトし，それをターゲットの韻律パタンとして用いた．韻律情報を考慮した変換モデルを用いた場合と，考慮していない変換モデルを用いた場合の評価サンプルを一対にして，順番をランダムにして評価者に提示し，どちらの音質が良いか判断させた．評価サンプルは各話者 10 文に対して女性話者 3 人の相互変換を行い，韻律情報の有無についてそれぞれ 60 サンプルを用意した．二つの異なる学習方法（トライフォン単位の対応学習，発話文単位の対応学習）に関して上述の要領で評価サンプルを作成し，それぞれ評価を行った．評価者は 7 人である．

図 4.10 に，(1) 従来の発話文単位で対応学習した場合，(2) トライフォン単位で対応学習した場合のプリファレンススコアとその 95%の信頼区間を示す．この結果から，いずれの学習方法を用いた場合においても，韻律情報を用いたスペクトル変換方法の有効性を示すことができた．

4.6.2　話者の識別評価

ABX 評価により，スペクトル変換音声の話者性について評価を行った．A，B には元話者とターゲット話者の自然音声を提示し，判断対象となる X には以下の条件で合成した音声を用いた．

1. 韻律情報を考慮していない変換モデルを用いた場合
2. 韻律情報を考慮した変換モデルを用いた場合

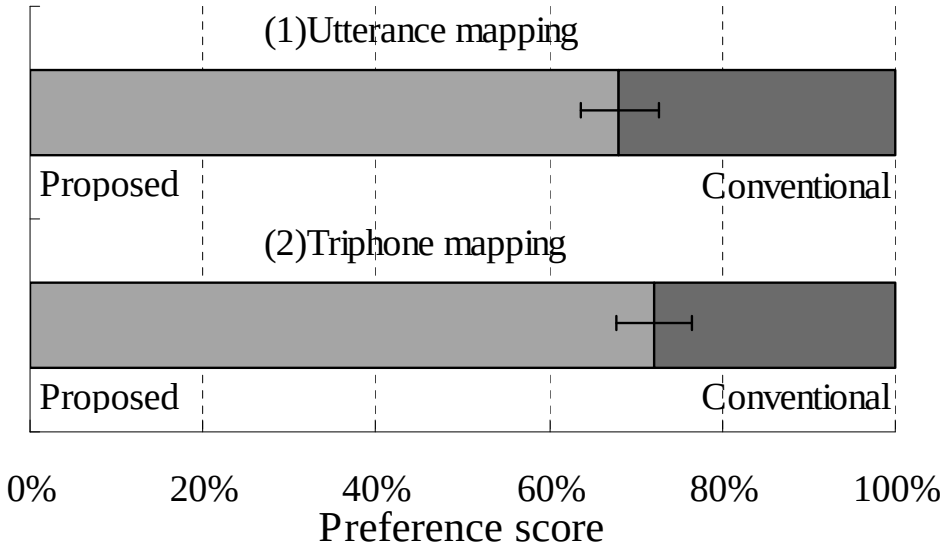


図 4.10　話者変換音声の音質評価結果:(1) 発話文単位の対応学習を行った場合，(2) トライフォン単位の対応学習を行った場合

3. 韻律パタン（F_0，パワー）のみをターゲット話者のレンジに変換した場合

この ABX 評価においても，音声合成の枠組を用いることで生じる素片選択や接続処理の歪の影響を無視できるように，元話者のデータベースに存在する文章を合成した．ここでターゲット話者の自然な韻律パタンを与えてしまうと評価者が韻律に着目して話者性を判断してしまうため，音素時間長は変更せず，元話者の F_0 とパワーをターゲット話者のレンジに合わせるようにした．また，規則合成へ応用することも考え，元話者，及びターゲット話者のいずれとも異なる第三者の韻律パタンを用いた場合についても評価を行った．ターゲットとして与える韻律パタン（F_0 とパワー）に関しては以下の二つの条件を考慮した．

1. 元話者の韻律パタンをターゲット話者のレンジに合わせたもの

2. 第三者の韻律パタンをターゲット話者のレンジに合わせたもの

なお，評価に使用したスペクトル変換モデル，評価者，評価文に関する条件は，上述の音質の比較評価と同じである．図 4.11 に話者の識別率と 95%の信頼区間を示

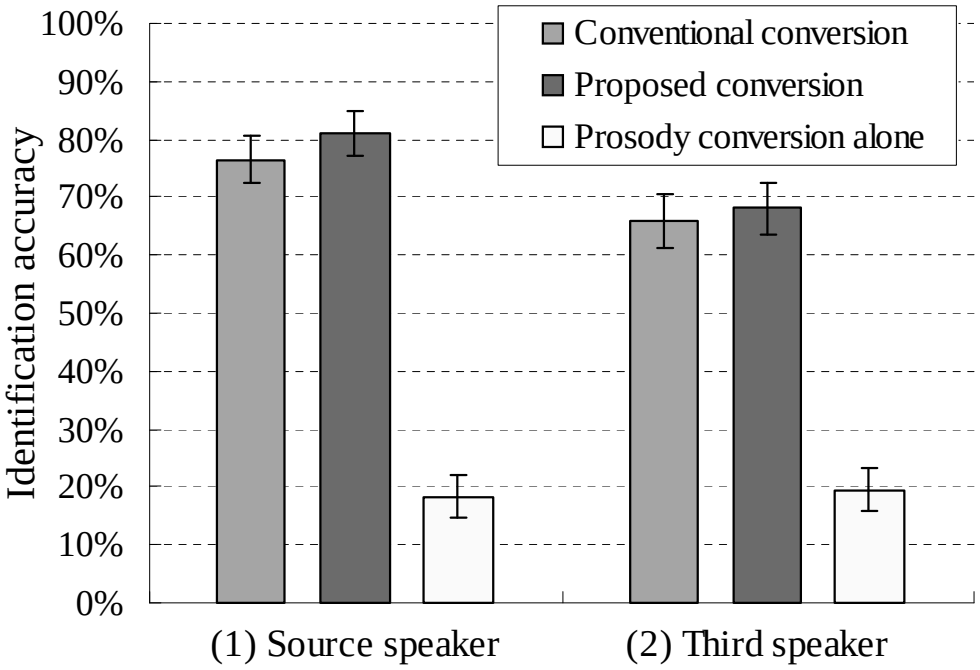


図 4.11　話者変換音声の話者判別評価結果: ターゲットの韻律パタンとして以下のものを付与，(1) 元話者の韻律パタンをターゲット話者のレンジに合わせたもの，(2) 第三者の韻律パタンをターゲット話者のレンジに合わせたもの

す．韻律情報を考慮した変換モデルを用いた場合，従来方法よりも高い識別結果を得ることができた．これは提案方法を用いることで変換音声の音質が改善され，話者の識別が容易になったためと考えられる．ただし，第三者の韻律パタンを与えた場合に関しては，話者識別率が低くなり，韻律情報の有無による差も若干縮まった．これは元話者の F_0 をターゲット話者のレンジに変換する処理が一定倍率の変換であるのに対して，第三者の F_0 パタンへ変換する処理は，合成する位置によって F_0 の変換率が異なり，これが素片間で音質劣化のむらを作り，識別を困難にしたものと考えられる．以上の結果から，韻律情報を考慮したスペクトル変換方法によって，話者識別率に関しても改善できる可能性が確認できたが，規則合成に応用することを考えた場合は，ピッチ変換処理で生じる音質劣化をより軽減するための検討が望まれる．

4.6.3　非同一発話文セットに関する試聴評価

非同一発話文セットを用いて学習した変換モデルによって話者変換音声を作成し，音質の比較評価，及び話者性の識別評価を行った．変換モデルは非同一発話文セット 50 文を用い，母音に関しては各母音ごとに変換モデルを作成した．また，有声子音に関しては音素を区別しないで一つの変換モデルを作成した．元話者の音声に対して，変換モデルを音素ごとに切り替えてスペクトル変換を行い，評価音声を作成した．この際，音素境界情報は手作業で与えている．無声子音は変換せず，元話者の波形をそのまま用いた．なお，ターゲットの韻律パタンは，元話者の韻律パタンをターゲット話者のレンジに合わせて使用した．評価サンプルは学習データに含まれない 10 文で，評価者は 7 人である．

図 4.12 に音質に関して一対比較を行った結果，図 4.13 に ABX 評価によって話者の識別実験を行った結果を示す．この結果から，以下のことがわかった．

1) 韻律を考慮した変換モデルは，学習データに非同一発話文セットを用いた場合でも，韻律を考慮しない従来の変換モデルより音質の良い変換音声を作成できる．

2) 変換モデルの条件が同じ場合（学習データの違いのみを単純に比較した場合）は，非同一発話文セットよりも，同一発話文セットを用いて学習した方が良い．

3) 音質の評価，及び話者の識別評価の結果からは，ケプストラム距離を用いた物理評価の結果と同様の傾向が確認できる．ただし，話者の識別評価の結果に関しては，各方式間でそれほど顕著な差はない．

4.7　考察

上述の試聴実験の結果から，提案方法によって変換した音声は，従来の韻律を考慮しない場合よりも，音質の良い変換音声を得られることがわかった．しかしながら，変換音声には「かすれ」や「こもり」などが発生し，原音声と比較する

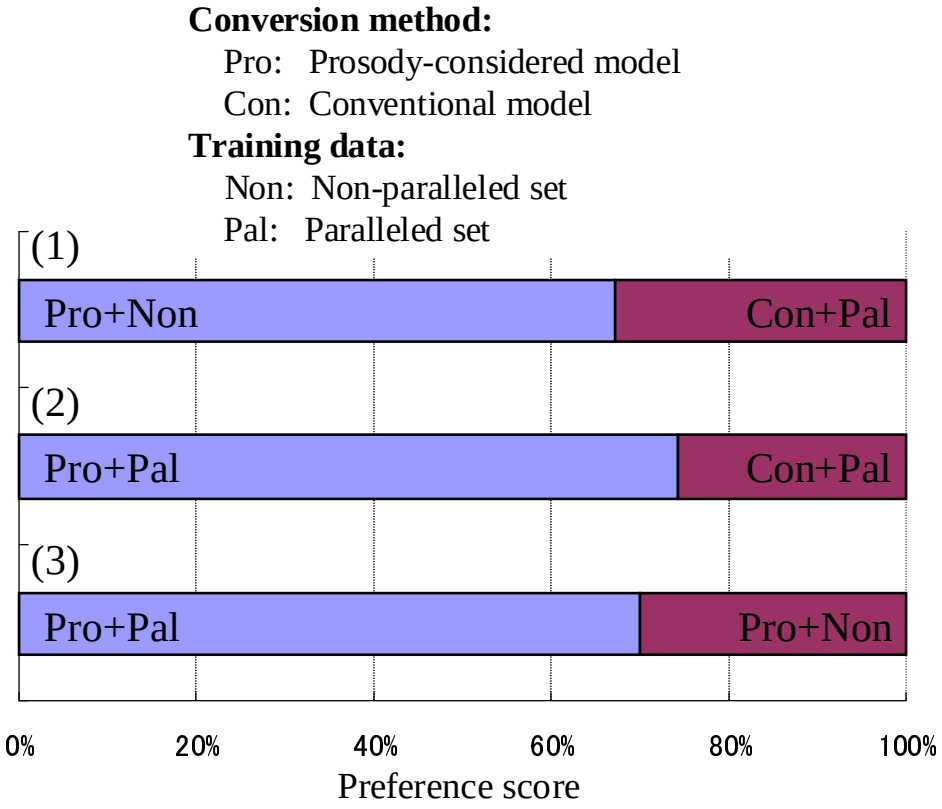


図 4.12　非同一発話文セットを用いた場合の音質評価結果: (1) 非同一発話文セット（韻律あり）と同一発話文セット（韻律なし）の比較，(2) 韻律考慮の有無，(3) 学習データ（同一発話文セット，非同一発話文セット）の違い

と，その音質は十分とは言いがたい．この音質の問題は，変換，及び合成にケプストラムの高次項を用いていないことや，元話者の位相情報をそのまま使用していることが原因として考えられる．音声の生成モデルを考えた場合，ケプストラムの高次項や位相はいずれも音源情報と密接に関係し，これらの特徴量を考慮したスペクトル変換の取り組みでは音質改善が報告されている [Ye 04]．本研究においても，このような取り組みを合わせて行うことで，更なる音質改善が期待できる．

(a)Prosody only: No spectral conversion
(b)Con+Pal: Conventional model trained using paralleled set
(c)Pro+Non: Prosody-considered model trained using non-paralleled set
(d)Pro+Pal: Prosody-considered model trained using paralleled set

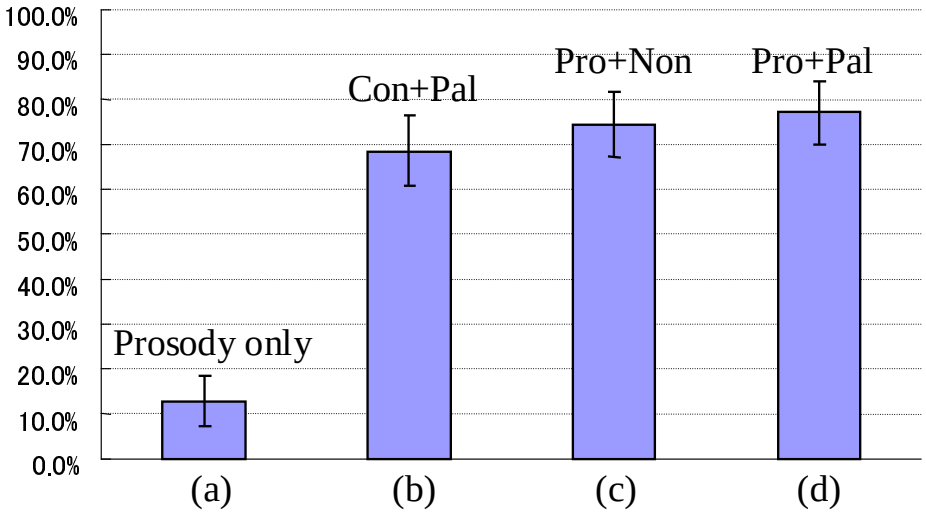


図 4.13　非同一発話文セットを用いた場合の話者判別評価結果: (a) スペクトル変換なし，(b) 同一発話文セット+韻律なし，(c) 非同一発話文セット+韻律あり，(d) 同一発話文セット+韻律あり

4.8　むすび

従来の GMM を用いたスペクトル変換では，スペクトルからスペクトルへの 1 対 1 の変換を対象としていたが，規則合成での利用を想定した場合，音素コンテキストや韻律情報を利用することが可能である．そこで本研究では，変換元と変換先の韻律情報を利用して GMM を学習し，精度の良いスペクトル変換方法について検討した．提案するスペクトル変換方式を音声合成の枠組の上で話者変換に応用し，物理評価，及び試聴評価を行った．ケプストラム距離を用いた物理評価により，提案方法による変換精度の有効性を示すことができた．また，従来方法と提案方法で変換した音声の試聴評価を行ったところ，韻律情報を利用した提案方法では，音質，及び話者性の両方において従来方法より良好な評価結果を得ることができた．また，非同一発話文セットを学習に用いた場合でも，提案する韻

律を考慮した変換モデルを用いることで，従来の同一発話文で対応学習する方法より，精度の良いスペクトル変換を行えることが確認できた.

第5章　結論

高品質の音声を合成するという立場から考えると，大規模な音声コーパスを利用した波形接続合成方式が現在の主流な合成方式であるが，この方式で合成できる音声は，音声コーパス収録時の発話スタイルに限定されるため，表現力の多様化という観点からは自由度が低い．本研究では限られた音声データによって発話スタイルの制御を目指すという立場から，高品質の合成が期待できる PSOLA 法を採用し，音質の改善，及び表現変換のための要素技術の提案・検討を行った．

第 2 章では，単位波形をピッチ同期で安定して抽出するためのピッチマーキング法を提案した．提案方法では音声の生成イベントの一つである声門閉鎖点，すなわち残差波形のローカルピークに着目し，変形自己相関によってこのローカルピークを逐次推定する方法について検討した．これにより，従来問題となっていた Jitter や Simmer によって生じる波形の揺らぎの影響を受けることなく，安定したピッチ同期分析を可能とした．また，音声信号モデルを用いて，ピッチマークを基準にスペクトル歪が最小となる波形抽出位置を実験的に探索した．その結果，ピッチマークから基本周期に対して 10%～20%程度遅延した位置にスペクトル歪が最小となる波形抽出位置があることがわかった．ただし，ピッチ変換音声を用いた試聴評価からは，窓関数の中心をピッチマークに合わせて単位波形を抽出するのが概ね妥当であることがわかった．更に提案方法の頑健性を評価するため，F_0 レンジの異なる複数の音声データベースに対してピッチマーキング実験を行った．その結果，F_0 が 600Hz を超える極端に高い音声を除けば，安定したピッチマーキングが可能であることを確認した．TD-PSOLA 法では，基本周期の 2 倍の窓長を持つハニング窓で単位波形を抽出し，これを再配列することで韻律変換を行っているが，これは安定した単位波形の抽出が可能となって初めて成立する．従来の原波形レベルでのローカルピーク抽出では，Jitter や Simmer の影響で安定した波

形抽出が出来なかったことを考えると，本研究で提案したピッチマーキング法，及びスペクトル歪を最小にする波形抽出位置の探索は，TD-PSOLA 法を実現する上で，必要不可欠な要素技術の一つとして位置付けられる．また，提案方法をベースにして，更に安定した波形抽出に取り組んだ研究も報告されている [峯松 00].

第 3 章では，ピッチ変換率に応じて動的に低域のスペクトル包絡を補正する方法について提案した．本来，周波数分析によって求められる理想的なスペクトルは，F_0 とその高調波のみで構成される線スペクトルであるが，実際は分析窓の影響で線スペクトルの代わりにスペクトル包絡が形成される．周波数分析の観点からは，現存する窓関数は一長一短の特徴を持ち，分析対象や用途に応じて適切な窓関数が用いるのが一般的であるが，スペクトル包絡を補間によって再現するという観点では，ハニング窓が適していることを確認した．また，単位波形のスペクトルは，F_0 より低い帯域に信頼できる情報がないため，正しいスペクトル包絡を再現できないという問題が存在する．これに対して，本研究ではピッチ変換を行ってもスペクトル傾斜は一定に保たれるという仮定に基づいて，動的に低域のスペクトル包絡を再構成する方法について検討した．その結果，ピッチを低い方へ 0.4octave 以上変換した場合，提案するスペクトル補正方法が有効であることを確認した．この低域スペクトルの問題は，PSOLA 法に限らず，音源と声道特性とを明確に分離しないでスペクトルを推定する方法の共通の問題である．この低域スペクトルの問題に対する他の解決方法としては，事前に複数の F_0 環境において学習を行い，合成時の F_0 に応じてスペクトルを変形，または選択するアプローチが考えられる．以前から，学習を用いてスペクトル包絡を推定する研究は多数あるが，低域の問題に焦点を当て，オンラインで利用可能な情報だけを使用してスペクトルを再現するという取り組みは例が無い．提案方法は学習を利用する統計的なアプローチと比較すると，必ずしも正解のスペクトルを再現できる方法ではないが，統計処理で進めた場合には深く議論されずに片付けられてしまう問題に対して，その本質に一歩踏み込んでモデル化を進めた取り組みと言える．

第 4 章では，GMM を用いたスペクトル変換方式において，韻律情報を考慮して，スペクトルをターゲットの空間へ変換する方法を提案した．スペクトル変換を音声合成へ応用することを考えた場合，ターゲットの韻律，及び合成する音素

情報は既知情報として与えられる．そこで本研究では，変換元の韻律とスペクトル，及び変換先の韻律が与えられたときに，モデルの尤度関数が最大となるターゲットのスペクトルを推定する方法について検討した．本研究では提案するスペクトル変換方法の応用例として，複数の女性話者の音声データを用いて，話者変換実験を行った．物理評価，及び試聴評価を行った結果，音質，及び話者性の面でも韻律情報を明示的に用いることの有効性を確認することができた．従来では，変換元のスペクトルと韻律とが与えられた場合に，モデルの出力確率を最大にするターゲットのスペクトルと韻律とを同時に求める方法 [En-Najjary 04] が検討されているが，提案方法では，ターゲットの韻律は外部から与えられるという前提のもと，ターゲットのスペクトルのみを求める変換関数を定義した．また，従来の研究では，変換元からターゲットへの 1 対 1 のスペクトル変換が目的であったため，音素を特に考慮せず，同一発話文を用いて一つの変換モデルを生成する方法がほとんどであったが，提案方法では音素系列も外部から与えられるという前提のもと，音素ごとにモデル構築を行った．上述のように，提案方法は音声合成システムへの応用を考慮し，システムから与えられる情報（音素系列，ターゲットの韻律）を有効活用した方式となっている．また，本研究では効率良く多くの学習データを集めるために，トライフォン単位での対応付けを行い，その中心音素において変換モデルを構築する方法について検討した．その結果，同一発話文セットを用いた場合，従来の文単位の対応付けで得られる学習データを包含するため，安定かつ変換精度の良いモデルの構築ができた．また，この学習方法は非同一発話文セットに対しても学習が可能であり，韻律情報を用いることで従来方法より優れた変換精度を得ることができた．従来の非同一発話文セットを用いた学習として，既存の変換関数を適応によって利用する方法 [Mouchtaris 04] が知られているが，提案方法では JDE 法 [Kain 98] によって直接 2 話者間の変換モデルを容易に学習できるという長所を持つ．

以上の取り組みの結果，PSOLA 法によって合成音声の音質改善，及び発話スタイルの変換を実現できる可能性を示すことができた．ノンパラメトリックな合成方式である PSOLA 法は，音源と声道特性とをモデルによって明示的に分離していないため，取り扱うスペクトル特徴量には両方の情報が含まれている．例えば音

源と声道特性とをそれぞれ独立に制御するモデルの場合，合成時の組み合わせによっては不整合を発生する危険性があるが，PSOLA 法ではこのような問題が回避できる．また，特徴量を音源と声道フィルタとに分離したモデルでは，上述の不整合の問題を避けるために特徴量のスムージングなどを行うが，このような処理はスペクトルの微細構造を損失することになる．従って，音源とフィルタとの安定した分離モデルが実現できない限り，これらの情報をまとめて扱う PSOLA 法は，安定性という点でも有効なアプローチである．

しかし，今後も少ない音声データで，更に自由度の高い表現を目指すという立場をとるなら，やはり音源と声道フィルタとに分離したモデルを考え，それぞれの特徴量に対して最適な方法で適応や変換処理を施すことが期待されるアプローチである．その実現には，まずパラメータ制御という観点で音源と声道フィルタとの独立性が高く，加えて，合成時に両者の整合性の良い分離モデルを考案することが必要であろう．その際，従来のようにスペクトルの微細構造を損失するようなモデル化を行うと音質劣化が発生するが，微細構造が自然に再現できるモデルを検討することで，高品質で表現の豊かな音声合成が期待できる．

謝辞

本研究を進めるにあたり，多くの御指導，御助言をくださいました早稲田大学理工学部 小林哲則教授に感謝の意を表します.

本論文をまとめるにあたり，貴重な御意見をくださいました早稲田大学 白井克彦総長，早稲田大学スポーツ科学部 誉田雅彰教授，早稲田大学大学院国際情報通信研究科 匂坂芳典教授に感謝いたします.

本研究を進めるにあたり，早稲田大学理工学部 知覚情報システム研究室にて御意見，御討議を頂いた 松坂要佐氏（現，日本学術振興会特別研究員），藤江真也助手，小川哲司助手をはじめ，試聴実験等に御協力くださいました同研究室の卒業生，在学生の皆様に感謝いたします．特に音声合成の研究を共に進め，活発な御討議を頂いた大久保雅史氏（現，トヨタ自動車 (株)），ならびに猿渡サイモン氏に深く感謝いたします.

本研究の一部は，松下電器産業 (株) における業務として推進させて頂きました．早稲田大学において修学する機会を与えてくださいました松下電器産業 (株) ＡＶコア技術開発センター 岡村和男所長，ならびに同センター UI グループ第 5 チーム石田明チームリーダーに感謝いたします．同社における音声合成開発プロジェクトにおいて，修学期間中の開発業務を補助して頂いた礒野敏幸氏をはじめ，同プロジェクトのメンバーの皆様に感謝いたします．また，早稲田大学にて研究をはじめる切っ掛けを与えてくださいました蓑輪利光氏（現，同社 R&D 知財センター）に感謝いたします.

本研究で利用した音声データの大部分は，(株) 松研において収録を行ったものです．音声収録にあたり，御助言，御協力くださいました (株) 松研 山中正人部長をはじめ，同部門のスタッフの皆様，研究用音声を御提供くださいましたナレータの皆様に感謝いたします.

なお，本研究の一部は，明治大学理工学部において推進したものです．当時，御指導くださいました明治大学理工学部 本多高教授，ならびに新居康彦氏（現，金沢工業大学教授）に感謝いたします．

最後に，学生時代からの長い研究生活を可能としてくださいました両親や，本研究を進めるにあたり，精神面で支えて頂いた妻，友人に感謝いたします．

参考文献

[Abe 88] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice Conversion Through Vector Quantization,” Proc. ICASSP88, vol.1, pp.655-658, April 1988.

[阿部 89] 阿部匡伸, 田村震一, 桑原尚夫, “FFT スペクトルからの信号再生法による音声変換手法,” 信学論 D-II, vol.72, no.8, pp.1180-1186, Aug. 1989.

[Arslan 97] L.M. Arslan and D. Talkin, “Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum,” Proc. EUROSPEECH97, vol.3, pp.1347-1350, Sept. 1997.

[坂野 00] 坂野秀樹, 陸金林, 中村哲, 鹿野清宏, 河原英紀, “時間領域平滑化群遅延による位相制御を用いた声質制御方式,” 信学論 D-II, vol.83, no.11, pp.2276-2282, Nov. 2000.

[Black 95] A.W. Black and N. Campbell, “Optimising selection of units from speech databases for concatenative synthesis,” Proc. EUROSPEECH95, vol.1, pp581-584, Sept. 1995.

[Campbell 96] N. Campbell, “CHATR: A High-Definition Speech Re-Sequencing System,” Proc. Third ASA/ASJ Joint Meeting, pp.1223-1228, Dec. 1996.

[Charpentier 86] F.J. Charpentier and M.G. Stella, “Diphones synthesis using an overlap-add technique for speech waveforms concatenation,” Proc. ICASSP86, vol.11, pp.2015-2018, April 1986.

[Charpentier 88] F.J. Charpentier and E. Moulines, “Text-to-speech algorithms based on FFT synthesis,” Proc. ICASSP88, vol.1, pp.667-670, April 1988.

[Chen 03] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, “Voice Conversion with Smoothed GMM and MAP Adaptation,” Proc. EUROSPEECH2003, vol.4, pp.2413-2416, Sept. 2003.

[Edgington 96] M.D. Edgington and A. Lowry “Residual-based Speech Modification Algorithms for Text-to-Speech Synthesis,” Proc. ICSLP96, vol.3, pp.1425-1428, Oct. 1996.

[En-Najjary 04] T. En-Najjary, O. Rosec, and T. Chonavel, “A Voice Conversion Method Based on Joint Pitch and Spectral Envelope Transformation,” Proc. ICSLP2004, vol.2, pp.1225-1228, Oct. 2004.

[George 97] E.B. George and M.J.T. Smith, “Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model,” IEEE Trans. Speech Audio Process., vol.5, no.5 pp.389-406. Sept. 1997.

[Gales 96] M.J.F. Gales and P.C. Woodland, “Mean and variance adaptation within the MLLR framework,” Computer Speech and Language, vol.10, no.4, pp.249-264, 1996.

[Griffin 84] D.W. Griffin and J.S. Lim, “Signal estimation from modified short-time Fourier transform,” IEEE Trans. Acoust. Speech Signal Process., vol.32, no.2, pp.236-243, April 1984.

[Hamon 89] C. Hamon, E. Moulines, and F. Charpentier, “A Diphone Synthesis System Based on Time-Domain Prosodic Modifications of Speech,” Proc. ICASSP89, vol.1, pp.238-241, May 1989.

[速水 85] 速水悟, 田中和世, 横山晶一, 太田耕三, “研究用音声データベースのためのＶＣＶ／ＣＶＣバランス単語セットの作成,” 電総研彙報, vol.49, no.10, 1985.

[東 99] 東弘人, 川又政征, “振幅スペクトルからの音声合成法におけるピッチ変換法,” 信学技報 SP99-89, pp.25-30, Oct. 1999.

[誉田 84] 誉田雅彰, 守谷健弘, “位相等価処理を用いた音声符号化,” 音声研資, S84-05, pp.33-40, April 1984.

[Iida 03] A. Iida, F. Higuchi, N. Campbell, and M. Yasumura, “A corpus-based speech synthesis system with emotion,” Speech Communication, vol.40, no.1-2, pp.161-187. April 2003.

[板倉 70] 板倉文忠, 斉藤収三, “統計的手法による音声スペクトル密度とホルマント周波数の推定,” 信学論 A, vol.53, pp.35-42, 1970.

[板倉 79] 板倉文忠, 菅村 昇, “ＬＳＰ音声合成器の原理と構成”, 音声研資, S79-46, pp.349-356, Nov. 1979.

[Iwahashi 95] N. Iwahashi and Y. Sagisaka, “Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks,” Speech Communication, vol.16, no.2, pp.139-151, Feb. 1995.

[Kain 98] A. Kain and M.W. Macon, “Spectral Voice Conversion for Text-to-Speech Synthesis,” Proc. ICASSP98, vol.1, pp.285-288, May 1998.

[Kain 01] A. Kain and M.W. Macon, “Design and Evaluation of a Voice Conversion Algorithm Based on Spectral Envelope Mapping and Residual Prediction,” Proc. ICASSP2001, vol.2, pp.813-816, May 2001.

[Kawahara 99] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds,” Speech Communication, vol.27, no.3-4, pp.187-207, April 1999.

[河井 95] 河井恒, 山本誠一, “基本周波数および音素持続時間を考慮した音声合成用波形素片データセットの作成,” 信学技報 SP95-7, pp.47-52, May 1995.

[Krishnamurthy 86] A. K. Krishnamurthy, “Two-channel speech analysis,” IEEE Trans. Acoust. Speech Signal Process., vol.34, no.4 pp.730-743, Aug. 1986.

[Macon 96] M.W. Macon and M.A. Clements, “Speech Concatenation and Synthesis Using an Overlap-Add Sinusoidal Model,” Proc. ICASSP96, vol.1, pp.361-364, May 1996.

[Maeda 99] N. Maeda, H. Banno, S. Kajita, K. Takeda, and F. Itakura, “Speaker conversion through non-linear frequency warping of STRAIGHT spectrum,” Proc. EUROSPEECH99, vol.2, pp.827-830, Sept. 1999.

[Masuko 96] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, “Speech synthesis from HMMs using dynamic features,” Proc. ICASSP96, vol.1, pp.389-392, May 1996.

[峯松 99] 峯松信明, 中川聖一, “F_0 変化に伴うスペクトル変動に対する分析とモデル化,” 音響学会誌, vol.55, no.3, pp.165-174, March 1999.

[峯松 00] 峯松信明, 中川聖一, “PSOLA 分析合成に基づく F_0 変換音声の品質向上に関する実験的検討,” 信学論 D-II, vol.83, no.7, pp.1590-1599, July 2000.

[Mizuno 95] H. Mizuno and M. Abe, “Voice Conversion Algorithm Based on Piecewise Linear Conversion Rules of Formant Frequency and Spectrum Tilt,” Speech Communication, vol.16, no.2, pp.153-164, Feb. 1995.

[Mouchtaris 04] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, “Non-Parallel Training for Voice Conversion by Maximum Likelihood Constrained Adaptation,” Proc. ICASSP2004, vol.1, pp.1-4, May 2004.

[Moulines 90] E. Moulines and F. Charpentier, ”Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” Speech Communication, vol.9, no.5-6, pp.453-467, Dec. 1990.

[中島 88] 中島隆之, 鈴木虎三, “パワースペクトル包絡 (PSE) 音声分析・合成系,” 音響学会誌, vol.44, no.11, pp.824-832, Nov. 1988.

[大村 95] 大村浩, 田中和世, “基本波フィルタリング法による精細ピッチパターンの抽出,” 音響学会誌, vol.51, no.7, pp.509-518, July 1995.

[Oppenheim 69] A.V. Oppenheim, “Speech analysis-synthesis based homomorphic filtering,” Journal Acoust. Soc. Am, vol.45, no.2, pp.458-465, 1969.

[Qin 05] L. Qin, G. Chen, Z. Ling, and L. Dai, “An Improved Spectral and Prosodic Transformation Method in STRAIGHT-based Voice Conversion,” Proc. ICASSP2005, vol.1, pp.21-24, March 2005.

[Quatieri 86] T.F. Quatieri and R.J. McAulay, “Speech transformations based on a sinusoidal representation,” IEEE Trans. Acoust. Speech Signal Process., vol.34, no.6, pp.1449-1464, Dec. 1986.

[阪本 95] 阪本正治, 斉藤隆, 鈴木和洋, 橋本泰秀, 小林メイ, “波形重畳法を用いた日本語テキスト音声合成システムについて,” 信学技報 SP95-6, pp.39-45, May 1995.

[Stylianou 95] Y. Stylianou, O. Cappe, and E. Moulines, “Statistical methods for voice quality transformation,” Proc. EUROSPEECH95, vol.1, pp.447-450, Sept. 1995.

[Stylianou 01] Y.Stylianou, “Applying the Harmonic plus Noise Model in Concatenative Speech Synthesis,” IEEE Trans. Speech Audio Process., vol.9, no.1, pp21-29, Jan. 2001.

[Suendermann 05] D. Suendermann, A. Bonafonte, and H. Ney, “A Study on Residual Prediction Techniques for Voice Conversion,” Proc. ICASSP2005, vol.1, pp.13-16, March 2005.

[Syrdal 98] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, “TD-PSOLA versus Harmonic Plus Noise Model in Diphone Based Speech Synthesis,” Proc. ICASSP98, vol.1, pp.273-276, May 1998.

[Takano 01] S. Takano, K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima, “A Japanese TTS system based on multiform units and a speech modification algorithm with harmonics reconstruction,” IEEE Trans. Speech Audio Process., vol.9, no.1, pp.3-10, Jan. 2001.

[Tamura 01] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR,” Proc. ICASSP2001, vol.2, pp.805-808, May 2001.

[Tanaka 97] K. Tanaka and M. Abe, “A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F_0,” Proc. ICASSP97, vol.2, pp.951-954, April 1997.

[Toda 01] T. Toda, H. Saruwatari, and K. Shikano, “Voice Conversion Algorithm Based on Gaussian Mixture Model with Dynamic Frequency Warping of STRAIGHT Spectrum,” Proc. ICASSP2001, vol.2, pp.841-844, May 2001.

[Toda 05] T. Toda, A.W. Black, and K. Tokuda, “Spectral Conversion Based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter,” Proc. ICASSP2005, vol.1, pp.9-12, March 2005.

[Tokuda 95] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, “An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features,” Proc. EUROSPEECH95, vol.1, pp.757-760, Sept. 1995.

[Valbret 92] H. Valbret, E. Moulines, and J.P. Tubach, “Voice transformation using PSOLA technique,” Proc. ICASSP92, vol.1, pp.145-148, March 1992.

[Verma 05] A. Verma and A. Kumar, “Introducing Roughness in Individuality Transformation through Jitter Modeling and Modification,” Proc. ICASSP2005, vol.1, pp.5-8, March 2005.

[Wouters 01] J. Wouters and M.W. Macon, “Control of Spectral Dynamics in Concatenative Speech Synthesis,” IEEE Trans. Speech Audio Process., vol.9, no.1, pp.30-38, Jan. 2001.

[Ye 04] H. Ye and S. Young, “High Quality Voice Morphing,” Proc. ICASSP2004, vol.1, pp.9-12, May 2004.

研究業績

本論文に関連する研究業績

論文

(1) 望月亮, 大久保雅史, 小林哲則, “韻律情報を用いたスペクトル変換方式の検討,” 電子情報通信学会誌, D-II, vol.88, no.11, pp.2269-2276, Nov. 2005.

(2) R. Mochizuki and T. Kobayashi, “A low-band spectrum envelope reconstruction method for PSOLA-based F0 modification,” IEICE Trans. INF.&SYST., vol.87, no.10, pp.2426-2429, Oct. 2004.

(3) 望月亮, 新居康彦, 西村洋文, 本多高, “駆動点同期型ピッチ波形抽出法”, 音響学会誌 53 巻 10 号, pp.772-778, Oct. 1997.

講演

(4) R. Mochizuki and T. Kobayashi, “A low-band spectrum envelope modeling for high quality pitch modification,” Proc. ICASSP2004, SP-P9.5 vol.1 pp.645-648, May 2004.

(5) Y. Arai, R. Mochizuki, H. Nishimura, and T. Honda, “An excitation synchronous pitch waveform extraction method and its application to the VCV-concatenation synthesis of Japanese spoken words,” Proc. ICSLP96, vol.3 pp.1437-1440, Oct. 1996.

(6) 望月亮, 小林哲則, “GMM によるスペクトル変換モデルの非パラレルコーパスを用いた学習,” 音響学会講演論文集 3-6-20, Sep. 2005.

(7) 望月亮, 小林哲則, “ＰＳＯＬＡ法における音質改善のための低域スペクトル包絡の補正方法,” 音響学会講演論文集 2-Q-4, pp.319-320, Sep. 2003.

(8) 望月亮, 本多高, 新居康彦, “駆動点同期型ピッチ波形抽出法の頑健性評価,” 電子情報通信学会総合大会 D-14-1, pp.243, March 1997.

(9) 望月亮, 三浦成充, 本多高, 新居康彦, 蓑輪利光, “ピッチ波形抽出位置と一様ピッチ変換音声の音質との関係,” 音響学会講演論文集 2-P-22, pp.345-346, Sep. 1995.

(10) 新居康彦, 西村洋文, 吉田博子, 蓑輪利光, 望月亮, 本多高, “ピッチ波形抽出位置の検討,” 信学技報 SP95-8, pp.53-59, May 1995.

(11) 望月亮, 本多高, 新居康彦, 吉田博子, 蓑輪利光, “短区間変形自己相関係数を用いたピッチ波形抽出法の検討,” 音響学会講演論文集 3-4-6, pp.285-286, March 1995.

その他の研究業績

論文

(1) 大久保雅史, 望月亮, 小林哲則, “心的態度表現に寄与する韻律／スペクトル包絡特徴の評価,” 電子情報通信学会誌, D-II, vol.88, no.2, pp.441-444, Feb. 2005.

(2) 望月亮, 蓑輪利光, “平滑化特徴ベクトルを用いたアクセント句の F0 パターン選択方法,” 電子情報通信学会誌, D-II, vol.87, no.2, pp.475-486, Feb. 2004.

(3) R. Mochizuki, Y. Arai, and T. Honda, “A study on the word synthesis method by using the VCV-balanced word database,” J. Acoust. Soc. Japan (E)21, pp.17-24, Jan. 2000.

講演

(4) T. Minowa, R. Mochizuki, and H. Nishimura, “Improving the naturalness of synthetic speech by utilizing the prosody of natural speech,” Proc. IC-SLP2000, vol.1 pp.609-612, Oct. 2000.

(5) R. Mochizuki, Y. Arai, and T. Honda, “A study on the natural-sounding Japanese phonetic word synthesis by using the VCV-balanced word database that consists of the words uttered forcibly in two types of pitch accent,” Proc. ICSLP98, vol.5 pp.2011-2014, Dec. 1998.

(6) Y. Arai, R. Mochizuki, and T. Honda, “A study on natural-sounding Japanese phonetic word synthesis based on the pitch waveform concatenation,” Proc. ICA98, vol.1 pp.267-268, June 1998.

(7) 大久保雅史, 望月亮, 小林哲則, “HMM 素片選択を用いた話者変換方式の検討,” 信学技報 SP2004-139, pp.13-18, Jan. 2005.

(8) 蓑輪利光, 望月亮, “テキスト音声合成に対する大規模コンテキストの利用に関する一考察,” 信学技報 SP2002-24, pp.1-6, May 2002.

(9) 蓑輪利光, 望月亮, 西村洋文, 釜井孝浩, “韻律のベクトルを利用した音声合成方式,” 信学技報 SP2000-4, pp.25-31, May 2000.

(10) 望月亮, 西村洋文, 蓑輪利光, 新居康彦, “波形接続合成に用いるＶＣＶ素片データベースの構築方法,” 信学技報 SP99-1, pp.1-8, May 1999.

(11) 大久保雅史, 望月亮, 小林哲則, “波形重畳型音声合成におけるHMMを用いた素片選択,” 音響学会講演論文集 3-2-14, pp.343-344, Sep. 2004.

(12) 大久保雅史, 望月亮, 小林哲則, “心的態度表現における韻律的／分節的特徴の影響,” 音響学会講演論文集 2-P-23, pp.375-376, March 2004.

(13) 大久保雅史, 望月亮, 蓑輪利光, 小林哲則, “波形重畳型音声合成における心的態度の再現性評価,” 情報科学技術フォーラム FIT2003, vol.2, pp.285-286, Sep. 2003.

(14) 望月亮, 蓑輪利光, “属性ベクトルを用いた F0 パタン選択方法の検討,” 音響学会講演論文集 3-10-21, pp.369-370, Sep. 2002.

(15) 蓑輪利光, 望月亮, “コーパスサイズに最適な韻律制御方法の検討,” 音響学会講演論文集 2-10-19, pp301-302, March 2002.

(16) 望月亮, 蓑輪利光, “波形重畳型の合成方式に用いる代表ピッチ波形生成方法の検討,” 音響学会講演論文集 2-1-4, pp.181-182, Sep. 2000.

(17) 望月亮, 西村洋文, 蓑輪利光, 新居康彦, “ターゲットピッチパターンに着目したＶＣＶ素片データベースの検討,” 音響学会講演論文集 2-3-3, pp.231-232, March 1999.

(18) 星野弘行, 岩田邦弘, 本多高, 望月亮, 新居康彦, “波形接続合成におけるＶＣＶ素片の音質に及ぼす影響について,” 音響学会講演論文集 2-P-18, pp.303-304, March 1999.

(19) 望月亮, 西村洋文, 蓑輪利光, 釜井孝浩, “韻律ベクトルを用いた高音質規則合成方式,” 音響学会講演論文集 1-3-22, pp.227-228, Sep.-Oct. 1999.

(20) 望月亮, 新居康彦, 星野弘行, 岩田邦弘, 本多高, “０型および１型アクセントによる強制発声ＶＣＶバランス単語データベースを用いた高音質単語音声合成の検討,” 音響学会講演論文集 2-P-2, pp.289-290, Sep. 1998.

(21) 望月亮, 西村洋文, 蓑輪利光, 新居康彦, “波形接続合成に用いるＶＣＶマルチ素片データベースの検討,” 音響学会講演論文集 2-P-1, pp.287-288, Sep. 1998.

(22) 西村洋文, 望月亮, 蓑輪利光, 釜井孝浩, “素片韻律と韻律テンプレートを利用した音声合成システム,” 音響学会講演論文集 2-P-3, pp.291-292, Sep. 1998.

(23) 蓑輪利光, 西村洋文, 望月亮, “合成音声のリズムの自然性向上の検討,” 音響学会講演論文集 2-P-8, pp.301-302, Sep. 1998.

(24) 新居康彦, 望月亮, 疋田和彦, 米本清, “周波数変換単語音声の高齢難聴者による聞き取り評価,” 音響学会講演論文集 1-8-1, pp349-350, Sep. 1998.

(25) 西村洋文, 望月亮, 新居康彦, “ピッチマーク補間方法の検討,” 音響学会講演論文集 2-P-2, pp297-298, March 1998.

(26) 新居康彦, 望月亮, 疋田和彦, “一様ピッチ変換音声の難聴者による聞き取り評価,” 音響学会講演論文集 2-8-10, pp.409-410, March 1998.

(27) 望月亮, 新居康彦, 久野圭督, 本多高, “ピッチ波形の間引きと補間方法の検討,” 音響学会講演論文集 1-P-20, pp.343-344, Sep. 1997.

(28) 望月亮, 本多高, 新居康彦, 西村洋文, “ＶＣＶ波形接続合成のためのピッチ変換法の検討,” 音響学会講演論文集 2-4-7, pp.233-234, Sep. 1996.

(29) 望月亮, 本多高, 新居康彦, 蓑輪利光, “スペクトル補正によるピッチ変換音声の音質改善,” 音響学会講演論文集 2-P-18, pp.337-338, March 1996.

(30) 望月亮, 本多高, 新居康彦, 蓑輪利光, “単語音声データベースのＶＣＶ連鎖波形を接続する任意単語生成法の検討,” 電子情報通信学会総合大会 D-697, March 1995.