Internal Use Only



TR-IT-0197

# 自然音声波形接続型音声合成システムにおける波形接続方式の検討

# A Method for concatenating segment on a multi-lingual speech re-sequencing synthesis system

坂野 秀樹
Hideki Banno

ニック・キャンベル
Nick Campbell

1996.9.24

話者性や発話様式等の特徴を損ねることなく音声を合成するための方法として自然音声波形接続型音声合成システムを使用することを考える。これは予め録音された音声データベース中の音素単位の音声波形を、何らの信号処理も行わずに接続し、連続音声として出力する方式である。この方式の利点は信号処理を行わないため原波形の特徴を減ずることなく音声を合成できることであるが、反面、接続点における波形の不連続性のために合成音の品質が低下する可能性があるという欠点がある。このような自然音声波形接続型音声合成システムの特徴を考慮に入れると、できるだけ原波形を加工せずしかも接続点における劣化が少ない合成方式が必要となってくる。我々はこれらの要求を満たす合成方式として、最近注目を浴びている音声モーフィングの技術を応用したものを提案する。これは接続点において波形のモーフィングを行い、接続点における波形の不連続性を緩和させるものである。この方式を用いることにより、接続点における波形の連続性を向上させることができ、より自然性の高い音声合成が可能となることを示す。

ⓒＡＴＲ音声翻訳通信研究所

ⓒATR Interpreting Telecommunications Research Laboratories

# 目次

# 第 1 章

# 序章

## 1　はじめに

これまでの音声合成システムでは、音声は比較的簡単な物理的なモデル化が可能であり、韻律とは切り離して考え、音素の系列のみから音声波形の合成が可能であるという暗黙の了解に基づいていた。これに対し我々は、音声波形は音響的及び韻律的な環境によって一意に定まるものであるとする立場から、音声合成システム自身が音声波形を生成することをせず、音響的及び韻律的な環境が最も適する音素単位の音声波形を何らの信号処理も行わずに接続し合成音を得るという手法を取っている。このため通常の音声合成システムより多くの音声データを必要とするが、極めて自然性の高い音声の合成が可能であることが確認されている [?]。

この方法の特長は、合成時に信号処理を行わずにそのまま波形を用いるため、話者性や発話様式等の特徴を保存したまま任意の音声を合成することができ、かつ選択された波形自体には劣化が全く無いことであるが、その反面、出力される音声は選択された音声波形の性質に依存するため、接続点において波形の不連続が生じ合成音の品質が低下する可能性があるという欠点がある。実際、合成音の品質はデータベースの作成方法や話者の特徴によって左右され、極端な場合では不連続感が強くとても自然とは言えない音声が合成されることもある。つまり、このような場合では波形自体の情報が保存されていることによる自然性よりも元々連続ではない音声波形を接続したことによる非自然性の方が、合成音により大きな影響を与えていると考えられる。

このような自然音声波形接続型音声合成システムの特徴を考慮に入れると、合成段階で、できるだけ波形自体を加工せず接続点における不連続性を緩和させるような手法を用いる必要があると考えられる。今回我々はこのような要求を満たす合成方式として、声質変換の応用の

一つである音声モーフィングの技術を応用したMORPHINGアルゴリズムを提案する。

音声モーフィングとはある話者の音声を別の話者の音声へと徐々に変化させることであり、ある音声波形をそれとは特徴の異なった音声波形へと変化させるような信号処理を行うことでそれを実現している。この音声モーフィングの考え方を自然音声波形接続型音声合成システムに適用し、単位音声波形の接続点において波形を補間することで、合成音の不連続性を緩和させることができると考えられる。本論文では、実際にこの手法を用いることにより、原波形の特徴を残したまま接続点における波形の連続性を向上させることができ、より自然性の高い音声を合成することが可能となることを示す。

## 2 現在用いられている方法

現在ATRで用いられている自然音声波形接続型音声合成システム“CHATR”では接続点での劣化を抑えるために、選択された音声波形に対し最も波形形状の近い部分をある一定の窓長にて探索し、その窓の中点で波形の接続を行うという方式を採用している（DUMB+アルゴリズム）。しかし、この方法は単に切り出し点を調整するだけに相当するため、波形の特徴が著しく異なる場合には対処できないという欠点がある。

## 3 本論文の構成

本論文は5章より構成されており、各章の概要は以下の様になる。

まず、2章でMORPHINGアルゴリズムの概要を述べる。

次に3章では、MORPHINGアルゴリズムにおけるスペクトルのモーフィング方法について述べ、スペクトルのモーフィング方法による合成音の差異を試聴実験により検討する。

そして4章では実際にMORPHINGアルゴリズムを用いて合成した音声の評価を行う。ここではまず既存の合成方法（DUMB+アルゴリズム）との比較を行い、次に合成音に与えるモーフィングの効果について調査する。

最後に5章において以上で得られた結果を基に本論文の結論を述べる。

# 第 2 章

# MORPHING アルゴリズム

## 1　はしがき

我々の使用する音声合成システムでは、まず与えられた入力に対し、合成に必要とするすべての特徴を予測し、その特徴に最も近い音声波形をデータベース中から検索する。この処理によって選択された単位音声波形を合成段階で信号処理などを施さずに接続し、合成音声を得るわけだが、選択された音声波形の性質によっては波形の不連続が生じ合成音の品質が低下する可能性がある。

MORPHING アルゴリズムはこの様な合成段階における劣化を防ぐための一方法であり、単位音声波形接続点においてピッチ同期で波形の補間（モーフィング）を行い、波形の接続を滑らかにするというものである。MORPHING アルゴリズムにおいては単位音声波形接続点においてのみ処理を行うため劣化を最小限に抑えることができるという特徴がある。今回は音声の特徴パラメータの中でも特に音声のピッチとスペクトルに着目し、これらのパラメータの制御を行うことによって音声波形の補間を行った。

本章ではこの MORPHING アルゴリズムを実際の処理手順に従って述べる。なお、これらの処理は全て音声合成システムとして ATR で使用されている CHATR を用いて行う。

## 2　処理手順

MORPHING アルゴリズムの処理手順を以下に示す（図??）。

**ピッチマーキング**　MORPHING アルゴリズムではピッチ変更のアルゴリズムとして PSOLA 法を用いている。PSOLA 法とは、音声波形のピッチマーク位置を中心に波形を切り出した“ピッチ素片”を、合成する際の基本周波数に対応するピッチマーク間隔で配置することにより、任

意のピッチの音声を得る手法である。このPSOLA法を用いるためには、予めデータベース中の音声波形のピークの位置などにピッチマーキングをしておく必要がある。

CHATRにおいてはデータベース作成時にすでにピッチマーキング処理を行っているため、今回はそのデータを用いて以下の処理を行っている。

**ピッチ素片の切り出し** CHATRによって出力された単位音声波形に対し、1ピッチ単位で波形を切り出す。方法としてはピッチマーク位置を中心に、隣接するピッチマークとの間隔にて左右非対称窓で切り出しを行う。左右非対称窓は隣接する波形ピーク（ピッチマーク位置に相当）がピッチ素片内に入ってくるのを避けるために用いる。

**最適なピッチ素片の選択** 接続を行う二つの単位音声波形に対し、モーフィング開始時のピッチ素片及び終了時のピッチ素片を選択する。今回は波形形状の差異が最も少ないもの同士を最適なピッチ素片の組として選択した。

**有声・無声の判定** 予備実験により選択されたピッチ素片のどちらかが無声音であった場合は既存のアルゴリズムを用いて合成した方が良い結果が出たため、その場合はMORPHINGアルゴリズムを用いずに合成する。これは無声音が有声音へとモーフィングされる際にピッチ性を持ってしまうことが原因だと思われる。なお、有声・無声判定にはCHATRのデータベース中のデータを用いて行った。

**モーフィング継続時間の決定** モーフィング開始時のピッチ素片からどれだけのピッチ素片数で終了時のピッチ素片まで達するかを決定する。今回は、元の単位音声波形を何も加工せずにそのまま接続した時の長さにほぼ一致するように継続時間を決定した。

また、同時にそれぞれのピッチ素片毎にモーフィング率（ピッチ素片がどれだけモーフィング終了時のピッチ素片に近いかを示すもの）も決定しておく。今回モーフィング率は線形に変化させている。

**モーフィングの実行** スペクトルのモーフィングを行い、それぞれのモーフィング率に応じたピッチ素片を導出する。この処理によりモーフィング継続時間に応じた数の補間されたピッチ素片が作成される。スペクトルのモーフィングのアルゴリズムについては後述する。

**音声の合成** ピッチ素片間のスペクトルのモーフィングにより得られた一連のピッチ素片を、モーフィング開始時のピッチ素片から終了時のピッチ素片までのピッチが滑らかに補間されるようにピッチ間隔を調整し配置していくことで音声を合成する（PSOLA 法）。今回はピッチ間隔を線形に変化させることにした。

また、モーフィングを行っていない区間については切り出した波形をそのまま元のピッチ間隔で配置する。この処理によってモーフィングの行われた部分以外はそのままの波形を用いることになり、信号処理による劣化を最小限に抑えることができると考えられる。

これら一連のモーフィング処理により単位音声波形間のピッチ及びスペクトルの補間が行われ、それらの不連続による劣化を回避することが可能になると考えられる。

## 3 本章のまとめ

本章では、自然音声波形接続型音声合成システムの波形接続点での劣化を防ぐ手法として MORPHING アルゴリズムを提案し、その概要を示した。次章では MORPHING アルゴリズムにおけるスペクトルのモーフィング方法についての検討を行う。

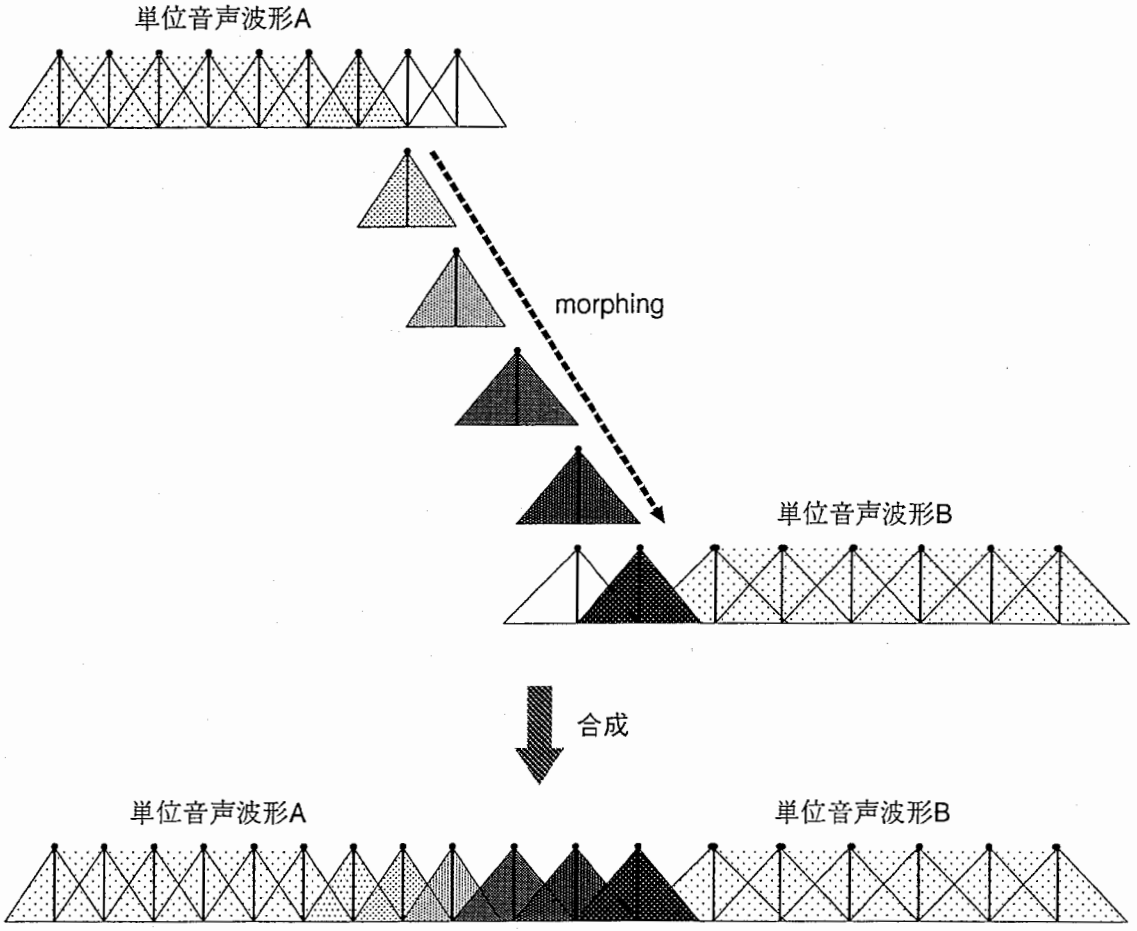


図 2.1: MORPHING アルゴリズム

# 第 3 章

# スペクトルのモーフィング方法

## 1　はしがき

本章では MORPHING アルゴリズムを用いて音声を合成する際のスペクトルのモーフィング方法について検討を行う。今回はスペクトルのモーフィング方法としてスペクトル包絡の補間によるもの及び波形自体の補間によるものの二つを用いた。

スペクトル包絡の補間によるモーフィングを用いた場合は、スペクトル包絡のピークが移動する形でモーフィングが行われるため、人間の発声の変化の様子と比較的似ており、聴感上も自然な音声が合成される可能性があると考えられる。しかし、こういった処理は非常に複雑であり、処理時間には膨大な時間を必要とする上に、複雑さ故の劣化が発生する可能性もある。これに対し波形自体の補間によるモーフィングでは自然性を損ねる可能性は否定できないが、処理時間は短くて済むという利点がある。

本章ではこれらのアルゴリズムを示し、それらを用いて合成した音声についての検討を行う。

## 2　モーフィング処理

以下に今回スペクトルのモーフィング方法として用いる二つの手法の概要を示す。

ここでは、モーフィング開始時のピッチ素片を A、モーフィング終了時のピッチ素片を B とし、モーフィング率は決定されているとするとした場合の、あるモーフィング率を持つピッチ素片の導出方法について説明する。実際には、モーフィング率を徐々に変化させそれに応じたピッチ素片を導出し、PSOLA 法でそれらを配置することにより波形の補間を行う。

## 2.1 スペクトル包絡の補間によるモーフィング

スペクトル包絡の補間によるモーフィングとは、周波数領域においてスペクトル包絡の対応付けを行い、その情報を用いて補間することでスペクトル包絡を自然に変化させる方式である。その処理手順は次のようになる（図??）。ここでは文献[?]のアルゴリズムを用いている。

**スペクトル包絡の抽出** まず、原波形のフーリエ変換から振幅スペクトルと位相スペクトルを求める。振幅スペクトルからFFTケプストラムを求め、その低ケフレンシー成分のフーリエ変換からスペクトル包絡を、高ケフレンシー成分から残差成分を抽出する。

**スペクトル包絡の対応付け** 周波数領域におけるDPマッチングによりAとBのスペクトル包絡の対応付けを行う。この処理によってほぼAとBのスペクトル包絡のピーク同士が対応付けられることになる。

**スペクトル包絡のモーフィング** 対応付けられたスペクトル包絡をモーフィング率に応じて変化させることによってスペクトル包絡のモーフィングを行う。この処理によってAとBの中間的なスペクトル包絡が計算されることになる。このスペクトル包絡は逆フーリエ変換を施すことによってケフレンシー領域に変換しておく。

**残差成分、位相スペクトルのモーフィング** 残差成分、位相スペクトルについてもモーフィング率に応じて変化させる。残差成分についてはケフレンシー領域でモーフィング率に応じてAとBの和を取ることで、位相スペクトルについては周波数領域でモーフィング率に応じてAとBの和を取ることでモーフィングを行う。

**モーフィング後のピッチ素片の導出** ケフレンシー領域でモーフィング後のスペクトル包絡と残差成分との和を取り、フーリエ変換することで、モーフィング後の振幅スペクトルが導出される。これにモーフィング後の位相スペクトルを付与し、モーフィング後のピッチ素片を導出する。

## 2.2 波形自体の補間によるモーフィング

波形自体の補間によるモーフィングはピッチ素片をそのまま足すだけであり、処理時間はスペクトル包絡の補間によるものに比べ大幅に削減できる。実際の処理は、ピッチ素片のパワーをモーフィング率に応じて変化させることによってモーフィングを行っている（図??）。

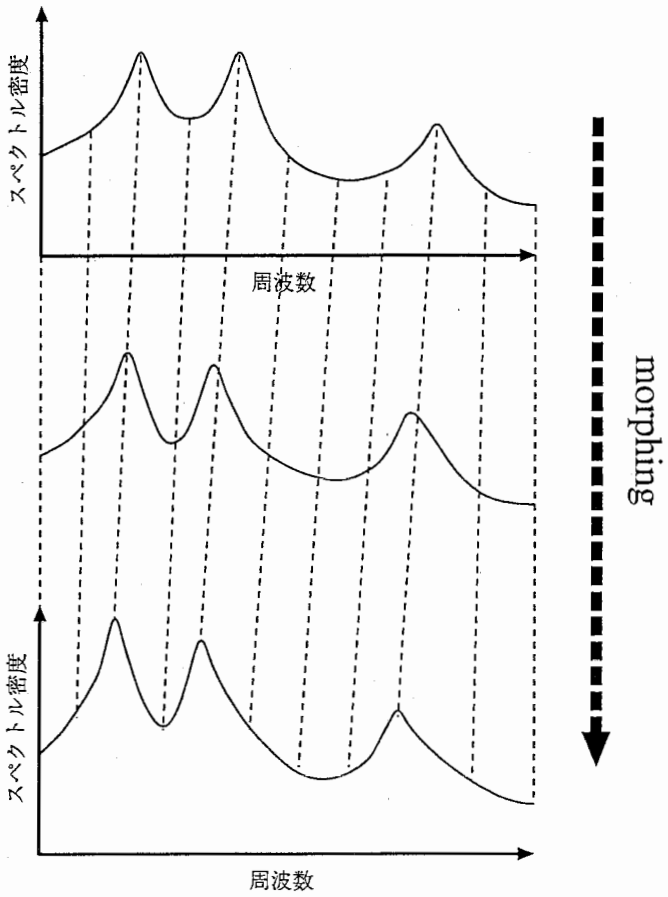


図 3.1: スペクトル包絡の補間によるモーフィング

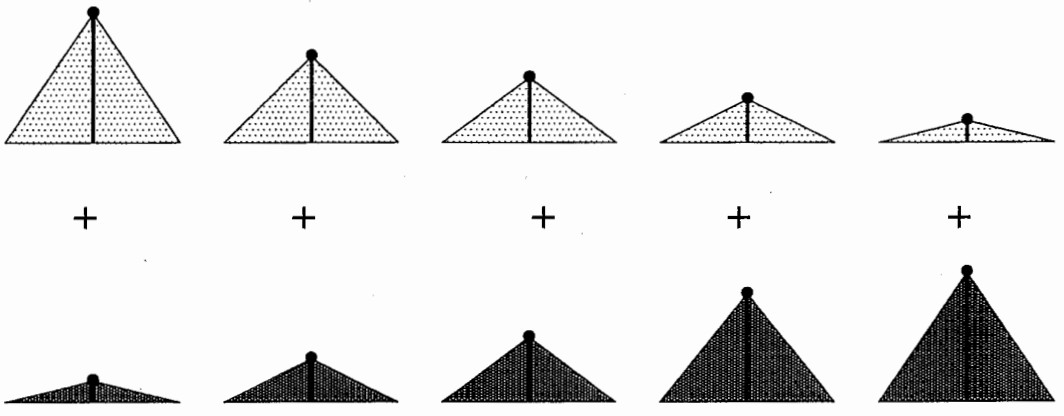

図 3.2: 波形自体の補間によるモーフィング

# 3 実験1：スペクトルのモーフィング方法による合成音の差異

スペクトルのモーフィング方法としてスペクトル包絡のモーフィングを用いた場合及び波形自体の補間によるモーフィングを用いた場合における合成音の差異を調査する。

## 3.1 実験方法

スペクトル包絡の補間及び波形自体の補間によるMORPHINGアルゴリズムを用いて様々な音声を合成し、それらを試聴により比較した。

## 3.2 実験結果

試聴の結果、両者の差はほとんど感じられず、ヘッドホン受聴時に多少の違いが分かる程度であった。また、スペクトログラムでもほとんど違いは見られなかった。

## 3.3 考察

スペクトルのモーフィング方法による合成音の差異がほとんど見られないのは、モーフィング継続時間が、元の波形長にほぼ一致するように選ばれているため、短いことが主な原因だと思われる。実際、モーフィング継続時間を比較的長い時間に設定すると、波形自体の補間によるモーフィングを用いた場合では不自然さが目立つようになった。しかし、音声を合成する際にモーフィング継続時間を長く設定する場合はほとんど無いため、処理時間のことを考えると波形自体の補間によるモーフィングで十分である。よって、次章以降ではスペクトルのモーフィング方法として波形自体の補間によるモーフィングを用いることとする。

しかしながら、スペクトル包絡の補間によるMORPHINGアルゴリズムの劣化がほとんど見られなかったことは、今後例えば調音結合を自ら作成しなければならない場合など（極端に貧弱なデータベースを用いての合成）に効果的に使用できる可能性があることを示唆しているとも言えよう。

# 4 本章のまとめ

本章ではスペクトルのモーフィング方法として、スペクトル包絡の補間によるモーフィングと波形自体の補間によるモーフィングを提案し、両者の比較を行った。実際にそれぞれの手法を用いて音声を合成し比較を行ったが、両者の差異はほとんど見られず、処理時間のことを

考えると波形自体の補間によるモーフィングで十分であることが分かった。

次章では、スペクトルのモーフィング方法として波形自体の補間によるものを用い、MORPHINGアルゴリズムの評価を行う。

# 第 4 章

# MORPHING アルゴリズムの評価

## 1　はしがき

本章では MORPHING アルゴリズムの評価を行う。まず、MORPHING アルゴリズムを既存の合成方式と比較し、MORPHING アルゴリズムの有効性を示す。次に、モーフィングを行うパラメータの中でどのパラメータの変更が最も効果があるかを調査する。

## 2　実験 2：既存の合成方式との比較

MORPHING アルゴリズムの有効性を示すため、既存の合成方式と比較し検討を行う。

### 2.1　実験方法

MORPHING アルゴリズムと既存の合成方式を用いて様々な音声を合成し、それらを試聴により比較した。今回は既存の合成方式として前述の DUMB+ アルゴリズムを用いている。

### 2.2　実験結果

試聴の結果、DUMB+ アルゴリズムでは接続時の単位音声波形の性質が著しく異なる場合、不連続感は強く劣化が非常に目立つが MORPHING アルゴリズムを用いた場合は滑らかになることが分かった。

また、DUMB+ アルゴリズム用いて合成した音声のスペクトログラムを図??、MORPHING アルゴリズムを用いて合成した音声のスペクトログラムを図?? に示してあるが、これらを比較しても MORPHING アルゴリズムを用いた場合の合成音が滑らかになっていることが分かる。

しかし、有声・無声が誤って判定された場合やピッチが倍ピッチもしくは半ピッチに誤っ

て判定された場合にはMORPHINGアルゴリズムによる合成音の劣化が顕著となる。

## 2.3 考察

試聴の結果より、DUMB+アルゴリズムでは接続時の単位音声波形の性質が著しく異なる場合に、不連続感の強い音声が合成されることが分かる。これは波形自体の情報が保存されていることによる自然性よりも元々連続ではない単位音声波形を接続したことによる非自然性の方が、合成音により大きな影響を与えていると考えられる。それに対し、MORPHINGアルゴリズムを用いた場合は滑らかな音声が合成されており、人工的に作り出したものではあるが、不連続性を緩和させるのに成功している。こういった意味ではMORPHINGアルゴリズムはより“自然”な音声の合成が可能だと言える。

また、MORPHINGアルゴリズムは有声・無声誤り、ピッチ誤りに脆弱であることが分かったが、これらはデータベース作成時に生じるものであり、人間の手によって修正を加えることで改善できるため、それほど問題ではないと言える。

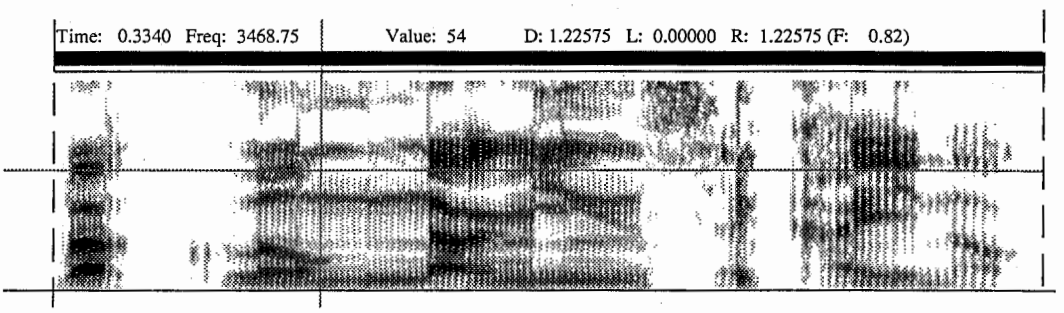


図 4.1: DUMB+ による合成音のスペクトログラム

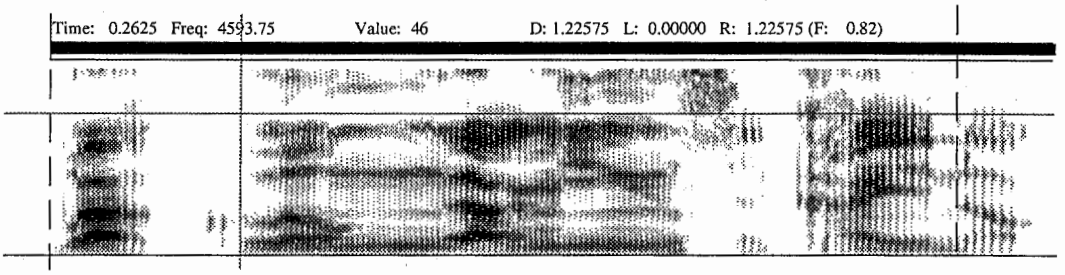


図 4.2: MORPHING による合成音のスペクトログラム

## 3 実験3：合成音に与えるモーフィングの効果

前節では既存の合成方式との比較によりMORPHINGアルゴリズムの有効性を示したが、ここではMORPHINGアルゴリズムがなぜ効果的なのかを検討する。

MORPHINGアルゴリズムを用いた場合にはスペクトル、ピッチ、パワーの三つのパラメータを制御することになるが、それらの中でどのパラメータの変更が最も合成音に影響を与えているかを調査する。

### 3.1 実験方法

まず、スペクトル、ピッチ、パワーの三つのパラメータの中の一つを選択し、モーフィング継続時間の中間においてそのパラメータの入れ換えを行い、その他のパラメータについては通常にモーフィングを行った音声を合成する。この処理を施すことによって、選択したパラメータについてはモーフィングを行っていない、すなわちそのパラメータが急激に変化する音声が合成されることになる。

このようにして合成された、あるパラメータについてのモーフィングを行っていない音声と、通常にMORPHINGアルゴリズムを用いて合成された音声を試聴により比較し、選択したパラメータをモーフィングすることによる合成音への影響を調べる。

### 3.2 実験結果

ピッチ、パワーについてはモーフィングを行った場合に多少の改善が見られる程度であり、スペクトルのモーフィングによる効果が三つのパラメータの中では最も大きいことが分かった。

### 3.3 考察

パワーの効果については、MORPHINGアルゴリズムでは最初にピッチ素片を選択する際に波形の形状の差異によって探索を行っているため、パワーの近いもの同士が選択される可能性があり、今回の結果からだけではパワーの補間の効果が無いと断言することはできないが、ピッチについては隣接する波形間でピッチの補間を行わなくてもそれほど劣化はしないと考えて良い。ピッチに関しては、それよりもピッチマーク位置がある程度のピッチ間隔で並ぶことが重要だと考えられる。今回比較に用いたピッチについてモーフィングを行わない場合でも、一定の範囲内でピッチマーク位置が配置されており、その条件をほぼ満たしているためそれほ

ど差が見られなかったのだと思われる。それに対し、例えば前節で比較に用いたDUMB+アルゴリズムなどは、波形の切り出し点を変更してしまうため、接続点付近においてピッチマーク位置がある程度のピッチ間隔で並ぶことは全く保証されない。そのためDUMB+アルゴリズムを用いた場合にはその時点で劣化することも考えられる。

いずれにせよ、MORPHINGアルゴリズムを用いることにより、スペクトルの不連続性を緩和させることができ、それによって合成音の品質が向上することがこの結果から読み取れる。

## 4 本章のまとめ

本章では、まずMORPHINGアルゴリズムを既存の合成方式と比較し、MORPHINGアルゴリズムを用いた場合は単位音声波形の接続点での変化が滑らかな音声の合成が可能であることが示された。

次に、モーフィングを行うパラメータの中でどのパラメータの変更が最も効果があるかを調査し、MORPHINGアルゴリズムはスペクトルの不連続性を緩和させるのに特に効果があることが分かった。

# 第 5 章

# 結論

本論文では自然音声波形接続型音声合成システムにおける合成方式としてMORPHINGアルゴリズムを提案した。

まず、実際の処理手順に従ってMORPHINGアルゴリズムの概要を示した。

次にスペクトルのモーフィング方法として、スペクトル包絡の補間によるモーフィングと波形自体の補間によるモーフィングの二つを提案した。両者の比較を行う実験の結果、ほとんど差が見られず、処理時間のことを考えると波形自体の補間によるモーフィングで十分であることが分かった。

また、MORPHINGアルゴリズムを用いた場合においては、既存の合成方式を用いた場合よりも“自然”な音声の合成が可能であり、MORPHINGアルゴリズムの有効性が示された。

そして、MORPHINGアルゴリズムを用いることによって、スペクトルの不連続性を緩和させることができ、それが合成音の品質の向上に貢献していることが分かった。

# 謝辞

本研究を進めるにあたりまして、数多くの貴重な助言及び協力を頂きました樋口宜男室長を始めとするATR音声翻訳通信研究所第二研究室の皆様に深く感謝致します。

また、実務訓練の機会を与えて下さった奈良先端科学技術大学院大学の鹿野清宏教授及びATR音声翻訳通信研究所の山崎泰弘社長に心から感謝致します。

# 参考文献

[1] ニック・キャンベル, アラン・ブラック: “CHATR：自然音声波形接続型任意音声合成システム”, 電子情報通信学会技術研究報告,SP96-7,pp.45–52(1996-05).

[2] 坂野秀樹, 武田 一哉, 板倉 文忠: “包絡と音源の独立操作による音声モーフィング”, 電子情報通信学会技術研究報告,SP96-6,pp.39–44(1996-05).