修士論文

音声の構造的表象に基づく
音声合成技術に関する基礎的研究

2008 年 1 月 29 日

指導教員　峯松 信明 准教授

東京大学大学院新領域創成科学研究科
基盤情報学専攻　66324

齋藤 大輔

内容梗概

音声を直接の人間の発声によらず，人工的に作り出す「音声合成技術」は近年の目覚ましいコンピュータの発達に伴って発展してきた．これらは現在大規模な話者から収録した音声データに基づいたものが主流となってきている．近年の音声合成システムは TTS（Text-to-Speech）変換と呼ばれ，与えられたテキスト列を音響信号へと変換するものである．この際，大規模な学習データを用いて音韻列と音響信号との対応関係を統計的手法により学習する．このとき構築されるモデルは多くの場合，音そのもののモデルである．本研究では音韻列を入力して音に変換する方法論とは全く異なる観点から音声合成を実装することを試みる．

幼児の言語獲得のプロセスは音声模倣 (vocal imitation) と呼ばれるが，上記のように音そのものを模倣して言語を獲得する訳ではない．幼児が両親の音声の音響的実体そのものを模倣する事は声道形状の差異から不可能である．幼児は何らかの抽象化を通して音声模倣を行っていると考えられる．TTS の枠組みと同様，音響信号を話者不変の音韻列に変換し，各音韻を音に変換しているとの議論も可能であるが，発達心理学はこれを否定する．幼児は音韻的意識が希薄であり，語から個々の音韻を抽出する能力が完成するのは小学校入学前後といわれている．即ち前述した音声の抽象化と音韻による音声表象は独立であると考えられる．

幼児の音声模倣を説明しうる音声合成を考えた場合，幼児が音声模倣に際して参照している話者不変の抽象的事象を物理的，音響的に定義する必要がある．発達心理学は「幼児は単語全体の語形・音形（語ゲシュタルト）を獲得し，その後，個々の分節音を獲得する」と主張する．近年この「語ゲシュタルト」の音響的定義と考えられうる，話者に不変な音声の構造的かつ抽象的表象が提案されている．

本研究ではこの構造的表象に基づく音声合成技術についてその枠組みの提案を行った．提案する枠組みは，音の実体モデルを持ちテキストを入力とする従来の音声合成システムとは大きく異なる．発話全体の語形を考え，それに対して身体特性，収録機器の伝送特性を与える事で初めて，聞き手が聴取する音響信号が生成される．本枠組みは音声模倣のモデルとして解釈可能である．本研究では提案する枠組みに関してその基礎的検討を行った．まず提案する枠組みを実験的に定式化するのに先立ち，話者の声道長差異に起因する非言語的特徴を多次元特徴量空間の幾何学を用いて理論的・実験的に検証した．その結果これらの特徴が特徴量空間における回転性として表出することを明らかにした．次に提案する枠組みを従来の話者適応学習の技術と比較してその違いを示した．続いて提案する枠組みを音響空間の制約条件付き探索問題として定式化し，様々な照合条件を検討する事で，提案する枠組みによる日本語孤立母音系列の音声合成を実現した．さらに対象を連続発声の母音系列に拡張し，連続音声においても合成が可能である事を示した．最後に解析的手法を導入する事でより高速に上記のタスクを実現可能である事を示した．

目次

図目次

表目次

第1章

序論

1.1　本研究の背景

人間と他の動物とを区別する要素はいくつか存在する．そのなかでも人間が言語を獲得していることは最も特筆すべき要素の一つであるといえる．文化や文明の発展に人間の言語獲得が寄与した部分は決して小さいとは言えない．人間のコミュニケーションにとって言語，とりわけ音声言語は欠かせないものである．

人間のメディア活動を計算処理によって実現しようとする「メディア情報処理」は多岐にわたって活発な議論が行われている．人間のコミュニケーションという観点から捉えた時，人間のセンシング能力（五感）に直接対応した生成器官を有し，人間がみずからの身体を使って発する事のできるメディアが「音声」 である．特に音声を直接の人間の発声によらず，人工的に作り出す「音声合成技術」は古くから研究が行われており，最初の音声合成器は，オーストリアの von Kempelen によって 1791 年に発表されている [1, 2, 3]．この合成器は機械式のもので，音源にはリードと摩擦音の為の笛を用い，その振動に共鳴部で音色を付与する事で音声として出力される．この合成器によって 19 種の子音と 5 種の母音が合成できたといわれている．

近年の目覚ましいコンピュータの発達により，あらかじめ収録された人間の音声を蓄積，処理，利用して合成を行うことが主流となっている．特に近年の音声合成システムは，与えられたテキスト列を音響信号として出力する Text-to-Speech 変換システム (TTS) である．TTS では音韻列を音声の表象として考え，その上で漢字仮名混じり文と音韻列との対応関係，および音韻と音響信号との対応関係を統計的手法により学習する．このとき構築されるモデルは多くの場合，異音の音響モデル (triphone) ，即ち音そのもののモデルである．大規模なデータの使用や種々の統計的手法の発達によりこのような音声合成システムの品質は目覚ましい進歩を遂げている．

再び人間の音声言語活動に話を戻す．幼児の言語獲得のプロセスは音声模倣 (vocal imitation) と呼ばれる．このとき幼児は上記のように音そのものを模倣して言語を獲得する訳ではない．幼児が両親の音声の音響的実体そのものを模倣する事は身体特性，声道形状の差異から不可能である．父親の「おはよう」と母親の「おはよう」を真似しても同じ本人の「おはよう」となるように，幼児は何らかの抽象化を通して音声模倣を行っていると考えられる．ここで [おはよう] という音響信号を /おはよう/ という話者不変の音韻列に変換し，各音韻を音に変換しているとの議論も可能であるが，発達心理学はこれを否定する [4]．そもそも幼児は音韻的意識が希薄であり，語から個々の音韻を抽出する能力が完成するのは小学校入学前後といわれている [5]．即ち前述した音声の抽象化と音韻による音声表象は独立であると考えられる．

人間の行う柔軟なメディア処理の実装を「メディア情報処理」の本質的命題とした場合，幼児の音声模倣を如何にして説明するのか．その際，幼児が音声模倣に際して参照している話者不変の抽象的事象を物理的，音響的に定義する必要がある．発達心理学は「幼児は単語全体の語形・音形（語ゲシュタルト [6, 7]）を獲得し，その後，個々の分節音を獲得する」と主張する [8]． 近年，この「語ゲシュタルト」の音響的定義と考えられうる，話者に不変な音声の構造的かつ抽象的表象が提案されている [9]．これは音声の音響的実体そのものは直接用いず，実体間の関係性のみをモデル化することで，非言語性歪みに対して不変性を有する音声表象である．

1.2　本研究の目的

本研究の目的は音の実体モデルを持ちテキストを入力とする従来の音声合成システムとは大きく異なる，新しい音声合成の枠組みを構築する事である．そのために本研究では近年提案されている話者不変の音声表象である，音声の構造的表象に基づく音声合成システムについてその枠組

みを提案し，その技術の基礎的検討を行う．提案する枠組みは，発話全体の語形を考え，それに対して身体特性，収録機器の伝送特性を与える事で初めて，聞き手が聴取する音響信号が生成される [10]．本枠組みは音声模倣のモデルとして解釈可能である．

1.3　本論文の構成

本論文は，全 8 章で構成される．**第 2章**では，特に本研究で用いる音声情報処理の基礎的技術と従来の音声合成システムをいくつかの視点から論ずる．**第 3章**では，音声に不可避的に混入する非言語的特徴に関して，そのモデル化を行う．加えて，特に声道長変化に着眼し，その特徴量空間における幾何学的性質を理論的，実験的側面から述べる．**第 4章**では，上記の非言語的特徴を本質的に包有しない音声の不変表象である音声の構造的表象について説明し，本研究で提案する構造的表象に基づく音声合成技術の枠組みについて述べる．さらに従来の話者適応学習による声質変換との比較を行う．

第 5章からは，構造的表象に基づく音声合成の基礎的検討を行っていく．**第 5章**では，孤立発声された日本語 5 母音系列を対象として音声合成を行う．この際構造に基づく音声認識で提案されている幾つかの手法を合成システムに応用することを検討し，実験を行う．**第 6章**では，連続発声された日本語 5 母音系列へと対象を拡張し，実験を行う．**第 7章**では，提案する枠組みのさらなる高精度化に向けて解析的手法を導入し，その基礎的検討を行う．最後に**第 8章**で本論文をまとめ，今後の展望について述べる．

なお，本論文では付録が掲載されている．**付録 A**では，**第 7章**で述べる，解析的アプローチの実装の為の式変形について記述する．

第2章

音声情報処理の基礎理論と従来の音声合成技術

2.1　はじめに

本章では音声分析，音声認識や音声合成をはじめとする音声情報処理における基本的な技術の枠組みについて述べる．さらに従来の音声合成システムとして，特に人間の生成過程に立脚したシステムおよび Text-to-Speech 変換のうち，特に収録された人間の音声を出発点とするシステムに二つに大別し論ずる．この中で本研究において提案する枠組みの位置づけを示していく．

2.2　音響特徴量

2.2.1　ケプストラム

音声の特徴は特にその周波数特性であるスペクトルによって表現される．音声分析によって抽出される特徴量のうち，このスペクトル情報を表現しかつ最も扱いやすい特徴量として広く用いられているのがケプストラム（Cepstrum）である．音声分析によって音声波形からケプストラムを抽出する過程を図 2.1に示す．まず音声波形から数十ミリ秒程度を 1 つのフレームとして切り出し，その区間について短時間の離散フーリエ変換（Discreter Fourier Transform; DFT）を施し，その区間の周波数特性であるスペクトルを抽出する．その後，対数パワースペクトルに対して逆離散フーリエ変換（Inverse DFT; IDFT）を施して得られるのがケプストラムである．

ケプストラムは現在の音声情報処理の基礎である線形分離等価回路モデル（ソースフィルタモデル）に基づいている．ソースフィルタモデルでは，人間の音声生成の過程に基づき，人間の音声が声帯の振動による音源特性 $G(\omega)$ に対して，人間の声道における調音の特性 $H(\omega)$ を伝達関数として我々の耳に届いていると考える．すなわち周波数領域において生成される音声 $S(\omega)$ を以下の式で表す．

$$S(\omega) = G(\omega)H(\omega) \tag{2.1}$$

式 (2.1) の絶対値の対数をとりこれを逆フーリエ変換する．対数をとる処理により積関係を和の形で分離できる．

$$\begin{aligned} c(\tau) &= \mathcal{F}^{-1} \log |S(\omega)| \\ &= \mathcal{F}^{-1} \log |G(\omega)| + \mathcal{F}^{-1} \log |H(\omega)| \end{aligned} \tag{2.2}$$

このとき式 (2.2) における $c(\tau)$ が連続量としてのケプストラムである．IDFT によりケプストラムはベクトル表現となり，この低次項のみを DFT することでスペクトル包絡が得られる．これは式 (2.2) における $\mathcal{F}^{-1} \log |H(\omega)|$，すなわち声道の調音特性に対応する．スペクトル包絡の山の部分は声道の共鳴周波数に対応しフォルマント周波数と呼ばれる．音声の音韻的特徴はこのフォルマント周波数によく表れる．つまりケプストラムはスペクトル情報を効率的に表現するベクトル特徴量となる．

2.2.2　聴覚特性を反映したケプストラム

人間の音の高さに対する周波数分解能は低い周波数ほど細かく，高い周波数ほど粗い事が知られている．このような聴覚特性をケプストラム特徴量に反映させたものが幾つか存在する．

MFCC（Mel-Frequency Cepstrum Coefficient）はメル周波数と呼ばれる，人間の聴覚特性を反映した周波数軸上において等間隔に配置された三角窓を用意し，フィルタバンク分析を行う事で求められる．各窓毎に対応する周波数帯域のパワーを求め，窓の大きさの重みをつけて和をとる

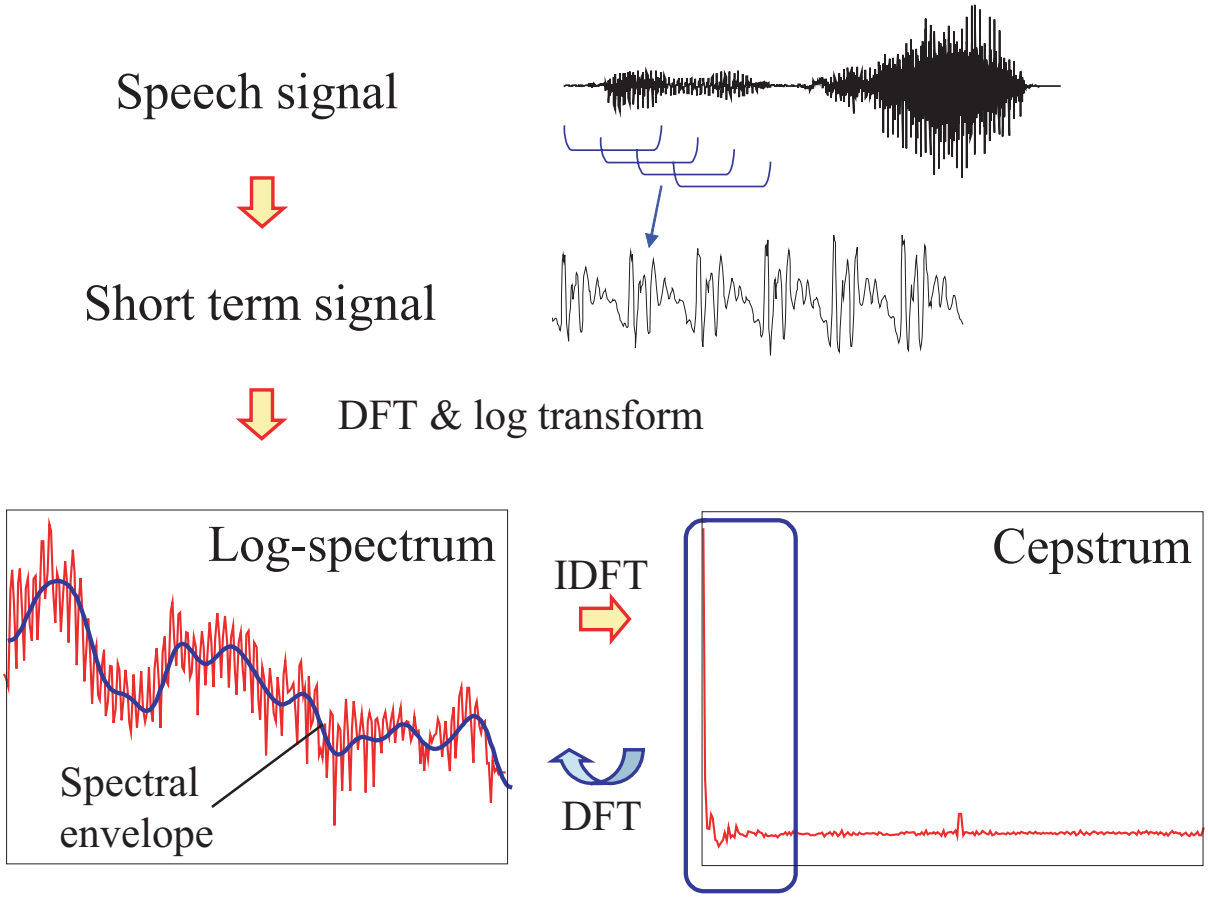


図 2.1: 音声分析

ことでメル化したスペクトルの離散情報が得られ，これに離散コサイン変換を施すことで MFCC が求められる．なおメル周波数は以下の周波数ウォーピングで求める．

$$f_{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.3}$$

メルケプストラム（Mel-Frequency Cepstrum; MCEP）は周波数ウォーピングとして低域強調効果のある全域通過関数を用いている．その後前節と同じ枠組みでケプストラム分析を行う事でメルケプストラムが求められる．なおメルケプストラムの係数から声道の周波数特性をあらわすディジタルフィルタを設計する事で，駆動音源から音声を合成する事ができ，音声合成の分野では特に広く使われている特徴量である [11].

2.2.3　Δ ケプストラム

スペクトルの時間軸に対する動的な特徴をとらえるため，ケプストラムベクトルの時間変化量である Δ ケプストラムがある．第 3章で詳しく述べるが，Δ ケプストラムは差分に基づく特徴量であるため，マイクの伝達特性の変化等に頑健で，時間変化量として動的特徴を表現するのに適していると考えられており，広く用いられている．

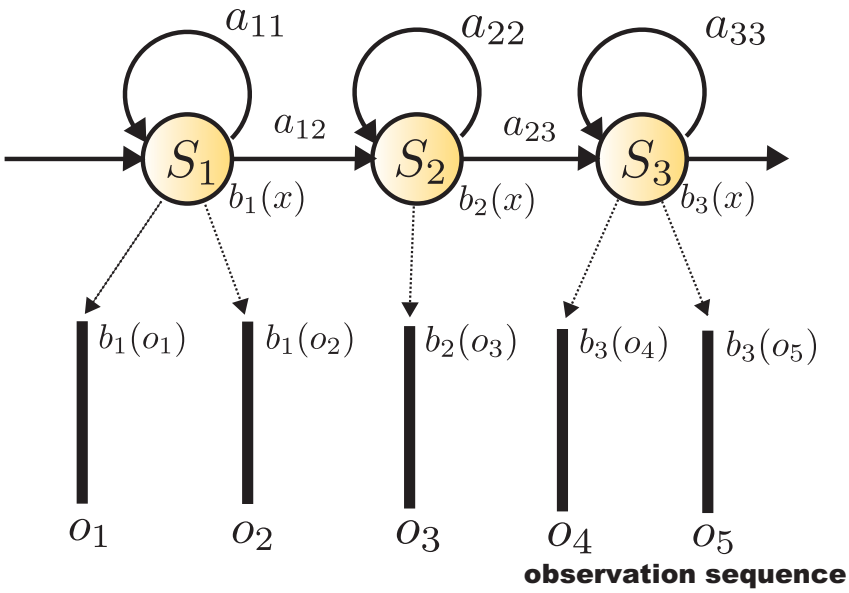


図 2.2: 隠れマルコフモデル（HMM）

2.3　HMM による音響特徴のモデル化

2.3.1　隠れマルコフモデル（HMM）

隠れマルコフモデル（HMM）は信号源間の状態遷移確率と信号源からの出力ベクトルの確率分布をパラメータとして持ち，状態遷移とベクトルの出力を繰り返す生成モデルである．このとき状態遷移確率が発声の時間的な揺らぎを，出力ベクトルの確率がスペクトルの揺らぎをうまく表現していると考えられる．図 2.2に HMM の構造を示す．図 2.2において S_i は i 番目の状態を，a_{ij} が S_i から S_j への遷移確率を表している．各状態 S_i はベクトル x を出力する確率 $b_i(x)$ をパラメータとして持つ．このとき $b_i(x)$ の分布系としてガウス分布に基づくものが広く用いられている．

2.3.2　HMM の学習

HMM において学習すべきパラメータは $\theta = \{a_i, b_i(x)\}$ であるが，これは最尤（Maximum Likelihood; ML）推定に基づいて行なわれる．即ち，学習データから音声特徴量の時系列データ X が観測されたとき，その尤度を最大化する θ を求める問題に帰着され，

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(X|\theta) \tag{2.4}$$

が求めるパラメータとなる．しかし，HMM の場合は隠れ変数[1]が存在し，式 (2.4) を解析的に解くのは困難である．このため，実際には式 (2.4) の局所最適解を求める Baum-Welch アルゴリズムが用いられる．

Baum-Welch アルゴリズムでは前向き変数 $\alpha_i(t)$，後向き変数 $\beta_i(t)$ と呼ばれる変数が登場する．

[1]外部から直接観測することができない変数．HMM の枠組みでは，データ系列 X が観測されたとき，その各々がどの状態から生じたものなのかまでを観測することはできない．

これらは，時刻 t における状態が i であれば 1，そうでなければ 0 をとる隠れ変数を z_{ti} とすると，

$$\alpha_i(t) = P(z_{ti}=1, x_1, \cdots, x_t|\theta) \tag{2.5}$$
$$\beta_i(t) = P(x_{t+1}, \cdots, x_T|z_{ti}=1, \theta) \tag{2.6}$$

と表すことができるものである．この前向き変数 $\alpha_i(t)$ 及び後向き変数 $\beta_i(t)$ を用いて，時刻 t における状態が i である確率 $\overline{z_{ti}}$ を，

$$\overline{z_{ti}} = P(z_{ti}=1|X, \theta) \tag{2.7}$$
$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N}\alpha_t(j)\beta_t(j)} \tag{2.8}$$

のようにして求めることができる．上記は，パラメータ θ と学習データからの音声特徴量の時系列データ X を用いれば，そのデータ系列の各々が特定の状態から生じた確率を求めることができることを意味する．式 (2.5) から式 (2.8) を用いれば，新しいパラメータを最尤推定によって求めることができる．例えば，出力確率 $b_i(x)$ の分布形として単一ガウス分布 $\mathcal{N}(x;\mu_i,\sigma_i^2)$ を用いる場合，パラメータ $\theta=\{a_i,\mu_i,\sigma_i^2\}$ に対して，新しいパラメータ $\hat{\theta}=\{\hat{a}_i,\hat{\mu}_i,\hat{\sigma}_i^2\}$ を，

$$\hat{a}_i = \frac{\sum_{t=1}^{T-1}\alpha_i(t)a_ib_i(x_{t+1})\beta_{i+1}(t+1)}{\sum_{t=1}^{T-1}\alpha_i(t)\beta_i(t)} \tag{2.9}$$
$$\hat{\mu}_i = \frac{\sum_{t=1}^{T}\overline{z_{t,i}}x_t}{\sum_{t=1}^{T}\overline{z_{t,i}}} \tag{2.10}$$
$$\hat{\sigma}_i^2 = \frac{\sum_{t=1}^{T}\overline{z_{t,i}}(x_t-\mu_i)^2}{\sum_{t=1}^{T}\overline{z_{t,i}}} \tag{2.11}$$

のようにして求めることができる．式 (2.9) から式 (2.11) の算出には，式 (2.5) から式 (2.8) を求めておく必要があり，一方，式 (2.5) から式 (2.8) の算出には，パラメータを式 (2.9) から式 (2.11) によって求めておく必要があるので，両者は互いに依存関係にある．しかしながら，このようにして得たパラメータ $\hat{\theta}$ は，パラメータ θ に対して常に，

$$P(X|\theta) \leq P(X|\hat{\theta}) \tag{2.12}$$

が成立するので，式 (2.5) から式 (2.8) の算出と，式 (2.9) から式 (2.11) の算出を繰り返す反復アルゴリズムによって，パラメータは局所最適解に収束する．

2.4　従来の音声合成システム

2.4.1　調音音声合成

人間の調音運動に着目し，調音運動のメカニズムを利用して行う音声合成を調音音声合成とよぶ．緒方ら [12] は，声帯を Ishizaka らの 2 質量モデル [13] によってモデル化し，さらに声道の形状特性を声道長 L とし，声道断面積 $A_n(n=1,\cdots,20)$ の 20 個の円筒形の管を接続したモデルで近似シミュレーションを行っている．

誉田ら [14] はコンピュータシミュレーションを用いることなく，実際にロボットに声帯（3 自由度），肺（1 自由度），そして調音器官である口唇（5 自由度），歯（1 自由度），舌（7 自由度）

鼻腔，軟口蓋（1 自由度）を実装する事によって，喉を震わせ口を動かして発話するロボットの開発を行っている．また聴覚のフィードバック，すなわち自分の声を耳で聞いて補正することによってより自然性の高い合成音声を目指している．

また Dang らは EMMA（Electromagnetic Midsagittal Articulographic）や MRI を用いて声道の動きのデータを取得し調音合成による子音も含めた音声合成を試みている [15].

調音音声合成は現在発声メカニズムの解明や，身体障害者の発声訓練支援などの応用が検討されている．特に人間の音声生成過程はソースフィルタモデルでは近似できない非線形な要因を含んでいる．そのような発声メカニズムの解明は音声工学の発展上大いに期待されるところである．またこれらが人間のメディア活動に密接に関わっている研究である点は特筆すべきである．

2.4.2　信号処理的音声合成

現在の音声合成は前節のような音声の生成を一から模擬する事はほとんどなく，あらかじめ収録した実際の音声を用いることが一般的である．このような現在の音声合成を本論文では，調音音声合成との対比として信号処理的音声合成と呼ぶ．この際に，信号処理の出発点となるのは，出力すべき信号を文字として抽象化した，いわば音韻列の集合である．

近年の音声合成システムは，与えられたテキスト列を音響信号として出力する Text-to-Speech 変換システム (TTS) と呼ばれるシステムである．TTS においては出力すべき単語列 W が与えられた時に適切な音声信号 X を出力する問題として考えられる．すなわち単語列 W が与えられた時に，その音声信号が X である事後確率 $P(X|W)$ を考え，これを最大化する $\hat{X}$ を求める．すなわち以下のように定式化される．

$$\hat{X} = \underset{X}{\operatorname{argmax}} P(X|W) \tag{2.13}$$

これは音声認識の逆問題として考えられる．このときどのような形で出力音声信号 X を出力するかによって信号処理的音声合成を二つの方式に大別することができる．

波形接続方式は人が発声した音声波形をそのまま蓄積しておき，必要に応じてつなぎ合わせて出力を行うものである．このとき接続単位としては音節，単語，diphone（二つ組の音素），音素などが考えられる．一般に接続単位が小さい方が少ないデータ量で柔軟な合成が可能になるが，接続コストの問題が顕著になる．一方で人の発声した音声波形をそのまま用いるため，上記の問題を解決した場合，非常に高音質の合成音が得られる．電車のアナウンス放送や近年話題の歌唱作曲アプリケーション等はこの方式によっている．この方式は近年のストレージの増大などを背景として，種々の研究報告がある [16, 17, 18].

一方パラメータ編集方式は人が発声した音声波形を分析してパラメータに変換した形で蓄積する．この時，単語とパラメータとの対応を音響モデルとしてモデル化しておく．パラメータから音声への変換を伴うため，必ずしも高品質の音声を得る事はできないが，この点に関する改善も数多くなされてきている．一方，特徴量空間における柔軟な操作が可能である点がパラメータ編集方式の利点といえる [19]．この時，生成モデルである HMM は，式 (2.13) の定式化に適しており，HMM に基づく音声合成が近年盛んに研究されている [20, 21].

今式 (2.13) について再考する．このとき出力される X には必ず特定の話者性が要求される．すなわち単語列 W に対して構築される音響モデルは必ず「ある話者」の音声信号でなければならない．不特定話者音響モデルを用いた音声合成も提案されているが [22]，これもたくさんの話者の声が平均化された「ある話者」にすぎない．

2.5　本研究の位置づけ

従来の音声合成システムとの比較から本研究の役割を位置づける．本研究では近年提案されている話者不変の音声表象に基づく音声合成システムを提案する．このときこの不変表象と音声信号を音響的に直接関連づけることはしない．幼児の音声模倣を考えた場合，母親や父親の声と自身の声には声道形状に起因する大きな差異があり，音韻的意識の希薄な幼児がこの対応を逐一学習しているとは考えにくい．よって提案する枠組みでは，不変表象に身体性を付与する事で初めて音響信号が得られる．式 (2.13) との対比のもとでは，ある話者不変な音響的単語表象（語ゲシュタルト）W と発話者の身体特性 B が与えられたとき，最尤な音声 $\hat{X}$ を出力する．すなわち以下のようになる．

$$\hat{X} = \underset{X}{\operatorname{argmax}}\, P(X|\boldsymbol{B}, W) \tag{2.14}$$

このとき，本研究で提案する枠組みは信号処理的音声合成を基本とするが，身体という調音音声合成的要素を明示的に伴う音声合成といえる．本研究はその基礎的検討段階であるため，直接的に身体性を取り扱うことはしないが，今後調音音声合成との関連が極めて重要になってくると考えられる．

2.6　まとめ

本章ではまず音声分析，音声認識や音声合成をはじめとする音声情報処理における基本的技術の枠組みについて述べた．さらに従来の音声合成システムを生成過程に立脚した調音音声合成と人間の音声信号をもとにした信号処理的音声合成に大別しその枠組みを論じた．さらに調音音声合成と信号処理的音声合成の架け橋としての本研究の位置づけを明確にした．

第3章

声道長に対する音響特徴量の依存性とその定量的分析

3.1　はじめに

前章では，音声分析，音声認識や音声合成をはじめとする音声情報処理における基本的技術の枠組みと従来の音声合成システムの概要について述べた．音声の特徴はその周波数特性であるスペクトルによって表現されるが，現在の音声情報処理においてこのスペクトル情報を効率的に表現するベクトル特徴量として，ケプストラムが広く用いられている．

一方でスペクトルに含まれる情報には，音韻のような言語的特徴に加えて，年齢，性別，声道長や収録する音響機器などの非言語的特徴も多分に含まれている．これらの特徴による影響は当然ケプストラムにも表出することになる．音声認識のように言語的特徴を捉える事を主眼とするアプリケーションの場合，特にこれらの非言語的特徴の取り扱いが重要になる．また音声合成においても，より自然性の高い音声を生成する上でこれらの特徴についての考察が不可欠である．

本章では，特に声道長の変化によって音響特徴量が受ける影響について，そのモデル化と従来手法によるアプローチを紹介する．さらに従来議論されてこなかった特徴量空間の幾何学的特性に着眼して，分析と実験を行った．

3.2　非言語的特徴のモデル化

峯松は音声に不可避的に混入する非言語的特徴について，数学的にモデル化している [23, 24, 25, 26]．以下，それについて説明する．音声に混入する非言語的特徴は主に加算性雑音，乗算性歪み，線形変換性歪みの三種類に分類される．加算性雑音は時間軸上の加算で表現される雑音である．テレビやラジオなどの背景雑音がその典型としてあげられる．これらの雑音は場所の移動等の対応が可能であり，不可避的なものではないといえる．特に不可避的な歪みとして考えられるのが後者の二つである．

乗算性歪みは，スペクトルに対する乗算で表現される歪みである．ケプストラム空間では，この種の歪みは加算演算 $\boldsymbol{c'} = \boldsymbol{c} + \boldsymbol{b}$ として表現される．マイクロフォンの音響特性がその典型例である．また話者の声道形状差異も一部近似的に乗算性歪みであると考えられる．音声は必ず発話者を伴い，音響機器によって収録されるため，これらの歪みは不可避である．

線形変換性歪みはケプストラム空間において行列 $\boldsymbol{A}$ による線形変換 $\boldsymbol{c'} = \boldsymbol{A}\,\boldsymbol{c}$ で表現される歪みである．スペクトル表現においては，話者の声道長差異や聴取者の聴覚特性差異は周波数ウォーピングとして考えられる．周波数ウォーピングはケプストラム空間において線形変換で記述されることが示されている [27, 28]．すなわち声道長差異や聴覚特性差異は近似的に線形変換性歪みとして扱うことができる．

以上をまとめると，音声の音響的実体に不可避的に混入する非言語的特徴は，ケプストラム空間においてアフィン変換 $\boldsymbol{c'} = \boldsymbol{A}\,\boldsymbol{c} + \boldsymbol{b}$ で表現される．これらの $\boldsymbol{A}, \boldsymbol{b}$ が話者や収録環境によって多様に変化し，音声の音響的実体に様々な歪みが混入する事になる．図 3.1はアフィン変換 $\boldsymbol{c'} = \boldsymbol{A}\,\boldsymbol{c} + \boldsymbol{b}$ によって，音声の対数スペクトルが変化する様子を示したものである．このとき行列 $\boldsymbol{A}$ は周波数方向（図中の水平方向）への変化となり，加算するベクトル $\boldsymbol{b}$ は対数パワー方向（図中の垂直方向）への変化として表れる．特に線形変換 $\boldsymbol{c'} = \boldsymbol{A}\,\boldsymbol{c}$ は主に声道長の長さの違いに起因する歪みであり，一方で声道長の伸縮は共鳴周波数を変化させる．すなわちケプストラム空間において，ある $\boldsymbol{A}$ による変換が声道長の長さの変化を端的に表す事になる．

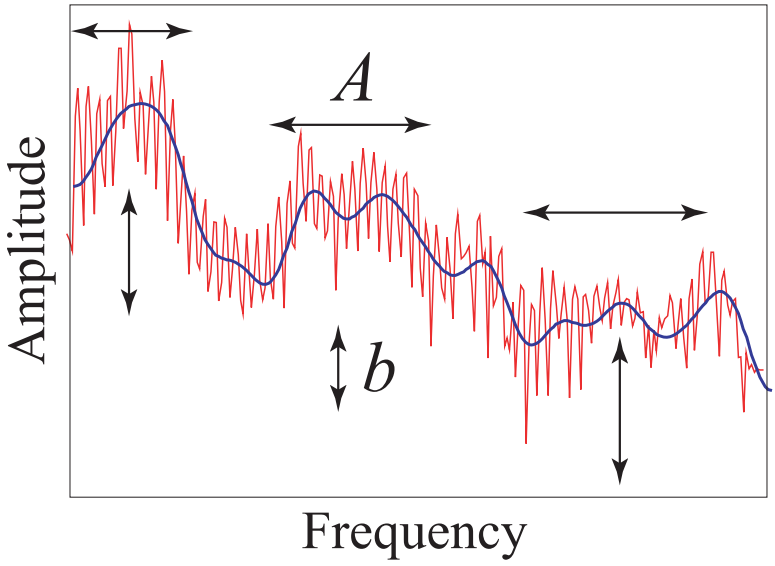


図 3.1: 非言語的特徴により変化する対数パワースペクトル

3.3　非言語的特徴への対応

前述の通り，音声は，音韻等の言語的特徴に加えて，年齢，性別，声道長や音響機器などの非言語的特徴を含んでいる．言語的特徴のみを精確に捉えることを音声認識システムの命題とした時，これらの非言語的特徴は音声に「歪み」を与えているといえる．これらの歪みは音声認識システムの性能に大きな影響を与えている．このような音声の多様な歪みに対応するため，大規模な学習話者による学習データを用いて，個々の音韻に対応する音響モデルを構築することが一般に広く行われている．これらは話者非依存 (Speaker Independent: SI) の音声認識システムと呼ばれている．一方，子供の音声に代表されるような特異な音声に対しては，話者非依存のシステムは時に全く性能を発揮できないことがある [29]．このような問題はミスマッチ問題と呼ばれ，この意味において話者非依存のシステムは “真に話者非依存” ではないといえる．

上記のようなミスマッチ問題に対して，話者正規化の技術が広く用いられている．話者正規化技術は，広く「話者性」として考えられている非言語的特徴の差異が音響特徴量に与える歪みに対して，その特徴量に何らかの演算を施す事でその影響を低減させる手法である．このような考えに基づく話者正規化技術は，特徴量の差分処理や微分演算に基づくもの，および変換に基づくものの二つに大別する事ができる．前者の例としてはケプストラム平均正規化 (CMN) や Δケプストラムの利用，後者として声道長正規化 (VTLN) がある．

CMN (Cepstrum Mean Normalization) では，ケプストラム特徴量の長時間平均をもとの特徴量から削除する [30]．CMN では二つのケプストラムの差を捉える事で，相対特徴量を導出する．あるケプストラム系列 $\boldsymbol{c}(t)$ に対して，伝送特性に対応する $\boldsymbol{b}$ を加算された系列 $\boldsymbol{c'}(t) = \boldsymbol{c}(t) + \boldsymbol{b}$ を考える．このとき $\boldsymbol{c}(t)$ および $\boldsymbol{c'}(t)$ のある時間 T に対する平均は以下のようになる．

$$\boldsymbol{E}(\boldsymbol{c}) = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{c}(t) \tag{3.1}$$

$$\boldsymbol{E}(\boldsymbol{c'}) = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{c}(t) + \boldsymbol{b}) \tag{3.2}$$

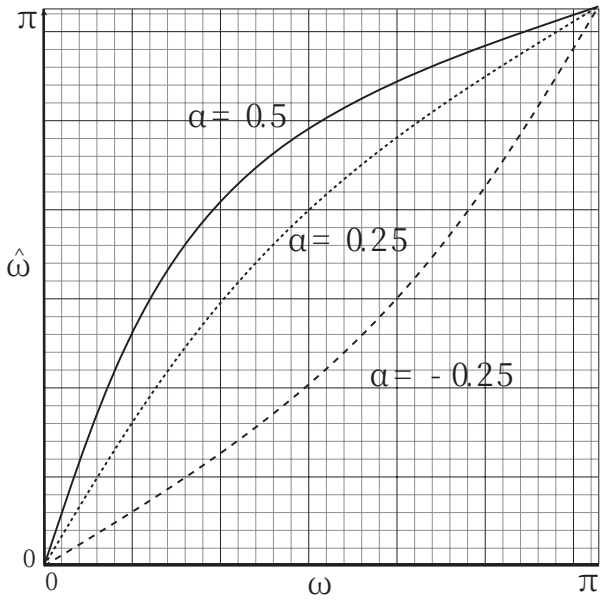


図 3.2: 周波数ウォーピング関数

故に，$\boldsymbol{E}(\boldsymbol{c})$ および $\boldsymbol{E}(\boldsymbol{c'})$ をそれぞれの元の系列から削除し，式 (3.1) を利用すると以下のようになる．

$$\boldsymbol{c}(t) - \boldsymbol{E}(\boldsymbol{c}) \equiv \boldsymbol{c}_{\boldsymbol{cmn}}(t) \tag{3.3}$$

$$\begin{aligned} \boldsymbol{c'}(t) - \boldsymbol{E}(\boldsymbol{c'}) &= (\boldsymbol{c}(t) + \boldsymbol{b}) - (\boldsymbol{E}(\boldsymbol{c}) + \boldsymbol{b}) \\ &= \boldsymbol{c}_{\boldsymbol{cmn}}(t) \end{aligned} \tag{3.4}$$

式 (3.3) および式 (3.4) から，CMN により音響機器の伝送特性に対応するスペクトルの乗算性歪み $\boldsymbol{b}$ が取り除かれていることがわかる．また $\boldsymbol{E}(\boldsymbol{c})$ は長時間平均をとることで音韻の違いを吸収し話者性を表していると考えられ，$\boldsymbol{c}_{\boldsymbol{cmn}}(t)$ を用いることで，声道形状の違いに対しても効果がある．一方で 2.2.3で述べたように，Δケプストラムも直前のフレームとの差を捉えており，乗算性歪みに頑健で音声の動的特徴を表現することから広く用いられる．

一方 VTLN(Vocal Tract Length Normalization) は声道長の違いを取り除くために用いられる [31]．前述の通りケプストラム空間における線形変換 $\boldsymbol{A}$ は声道長の違いに対応している．VTLN ではケプストラム空間における変換行列を推定する事で入力話者を標準声道長の話者の特徴へと変換する．VTLN は，変換行列の推定が正しく行えれば，ミスマッチ問題の解消に大きな効果があり，広く用いられている．また 2.2.2で述べた通り，聴覚特性の変化を低域強調の全域通過関数で表した場合，ケプストラムに対する線形変換となるため，これらの変化も同一の枠組みで議論ができる．

3.4　全域通過関数に基づく変換行列の実装

ケプストラム空間における変換行列 $\boldsymbol{A}$ を用いた話者正規化技術は広く用いられている．一方でこの変換行列の ケプストラム空間における幾何学的な性質は今まであまり議論されてこなかった．以下本章ではこの声道長変化を表す変換行列の幾何学的性質について，理論的および実験的に検

討する．また乗算性歪みや声道形状差異に効果のある Δパラメータに対して変換行列が与える影響についても言及する．

話者の声道長の変化は，音声のスペクトル表現における周波数ウォーピングとして考えることができる．今，周波数ウォーピングにおける変換前後の正規化角周波数を $\omega, \hat{\omega}$ $(0 \leq \omega, \hat{\omega} \leq \pi)$ とする．このとき $z = e^{j\omega}$, $\hat{z} = e^{j\hat{\omega}}$ として，ウォーピング関数として以下の 1 次全域通過関数を考える．

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{3.5}$$

このとき α は $|\alpha| < 1$ の実数であり，$\alpha < 0$ の場合，周波数軸が低域に変換され声道長は長くなる．$\alpha > 0$ の場合は，周波数軸は高域に変換され声道長が短くなる．以後 α をウォーピングパラメータと呼ぶ．図 3.2はウォーピングパラメータを変化させた場合の式 (3.5) の様子を示している．

以下，前述のスペクトルドメインにおける周波数ウォーピングをケプストラム空間における記述に置き換える．江森らは声道長の変化をケプストラム空間で記述し，これらのパラメータ推定に基づく声道長正規化を行っている [32, 33]．パワーを表現するケプストラムの 0 次項（c_0, $\hat{c}_0$）を考慮しない場合，周波数ウォーピングは以下の式で表現される．

$$\boldsymbol{\hat{c}} = \boldsymbol{A}\,\boldsymbol{c} \tag{3.6}$$

$$\boldsymbol{\hat{c}} = (\hat{c}_1\ \hat{c}_2\ \hat{c}_3\ \hat{c}_4 \cdots)^t$$

$$\boldsymbol{A} = \begin{pmatrix} 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots & \cdots \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \tag{3.7}$$

$$\boldsymbol{c} = (c_1\ c_2\ c_3\ c_4 \cdots)^t$$

さらに Pitz らによれば，式 (3.7) における行列 A の要素 a_{ij} はウォーピングパラメータ α を用いて以下のように表せる [28]．

$$a_{ij} = \frac{1}{(j-1)!} \sum_{m=\max(0,j-i)}^{j} \binom{j}{m} \times \frac{(m+i-1)!}{(m+i-j)!} (-1)^{(m+i-j)} \alpha^{(2m+i-j)} \tag{3.8}$$

ただし

$$\binom{j}{m} = \begin{cases} {}_jC_m & (j \geq m) \\ 0 & (j < m) \end{cases} \tag{3.9}$$

とする．

3.5　ケプストラム空間における回転性

3.5.1　2 次元ケプストラム空間における回転性

本節では，式 (3.7) の性質について幾何学的に考察する．簡単のためまず，ケプストラムの 1 次および 2 次の項のみに着目し，2 次元のケプストラム空間を考える．次に，n 次元のケプストラム空間へと議論を拡張する．

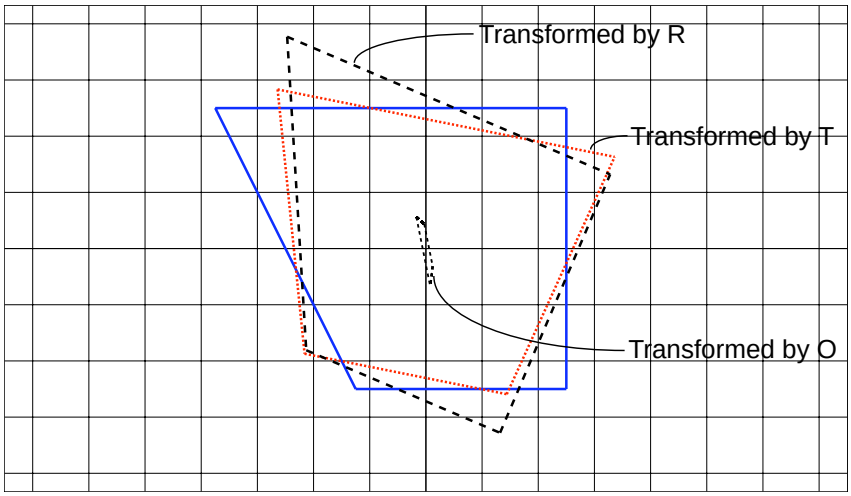


図 3.3: $\alpha = 0.2$ における行列 T，R，O による変換の様子

2 次元空間において，式 (3.6) における $\boldsymbol{c}$ から $\hat{\boldsymbol{c}}$ への変換は以下のように記述する事ができる．

$$\begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \end{pmatrix} = \begin{pmatrix} 1-\alpha^2 & 2\alpha-2\alpha^3 \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \tag{3.10}$$

今，式 (3.10) の変換行列を行列 $\boldsymbol{T}$ と呼ぶ事にする．このとき $\boldsymbol{T}$ は以下のように分解する事ができる．

$$\boldsymbol{T} = \boldsymbol{R} + \boldsymbol{O} \tag{3.11}$$

ただし

$$\boldsymbol{R} = \begin{pmatrix} 1-2\alpha^2 & 2\alpha(1-\frac{1}{2}\alpha^2) \\ -2\alpha(1-\frac{1}{2}\alpha^2) & 1-2\alpha^2 \end{pmatrix} \tag{3.12}$$

$$\boldsymbol{O} = \begin{pmatrix} \alpha^2 & -\alpha^3 \\ -\alpha & -2\alpha^2+3\alpha^4 \end{pmatrix} \tag{3.13}$$

とする．行列 $\boldsymbol{R}$ について着目すると，一次近似である $(1+t)^k \simeq 1+kt$ を用いる事で次式のように変形できる．

$$\boldsymbol{R} \simeq \begin{pmatrix} 1-2\alpha^2 & 2\alpha\sqrt{1-\alpha^2} \\ -2\alpha\sqrt{1-\alpha^2} & 1-2\alpha^2 \end{pmatrix} \tag{3.14}$$

$$= \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ -\sin 2\theta & \cos 2\theta \end{pmatrix} \quad (\alpha = \sin\theta) \tag{3.15}$$

式 (3.15) は行列 $\boldsymbol{R}$ が回転行列であることを示しており，$\boldsymbol{R}$ による変換は原点を中心として，全てのベクトルを時計まわりに 2θ 回転させる事になる．

一方行列 $\boldsymbol{O}$ は，行列 $\boldsymbol{T}$ による変換において，その影響が非常に小さいといえる．なぜならば $|\alpha| < 1$ であり，$\boldsymbol{O}$ のうち 3 つの要素が 2 次以上の高次項で構成されているためである．故に 2 次元平面における変換行列 $\boldsymbol{T}$ の性質は行列 $\boldsymbol{R}$ の性質でおよそ記述でき，すなわち $\boldsymbol{T}$ による変換は高い回転性を示す事になる．図 3.3は $\alpha = 0.2$ において，2 次元平面における台形が行列 $\boldsymbol{T}$，$\boldsymbol{R}$,

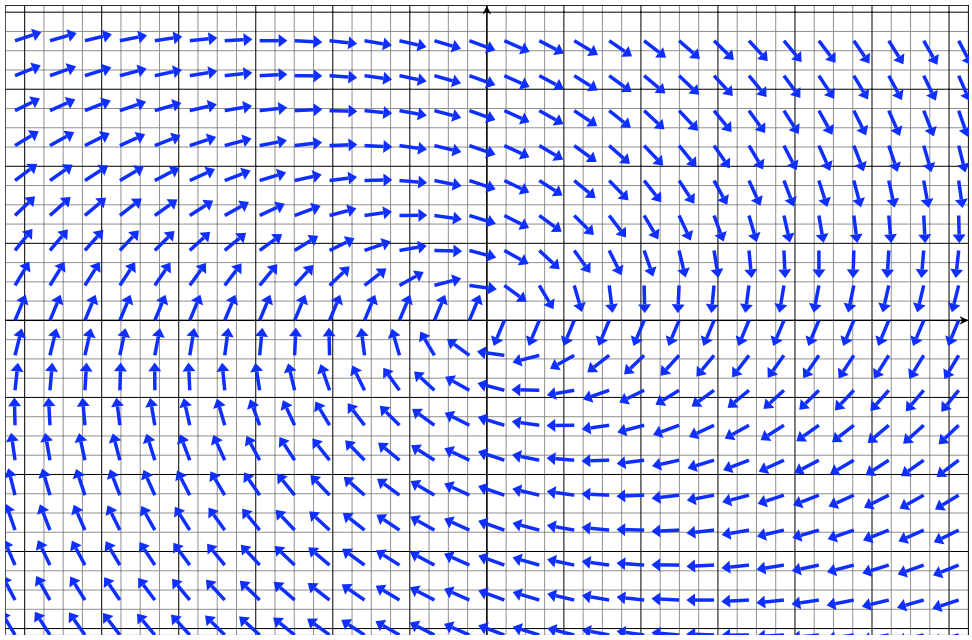


図 3.4: 式 (3.16) のベクトル関数のベクトル場表示 ($\alpha = 0.2$)

$\boldsymbol{O}$ によってどのように変換されるかを示している．原点付近に存在する小さな四角形が O によって変換されたものである．図 3.3により，明らかに $\boldsymbol{T}$ によって変換された台形が時計回りに回転しており，その様子が $\boldsymbol{R}$ による変換と非常に近い事がわかる．一方 $\boldsymbol{O}$ における変換は全ての点を原点付近へ圧縮しているといえ，その性質は零行列に近いため，$\boldsymbol{T}$ における $\boldsymbol{O}$ の影響は非常に小さい．さらに別の側面から行列 $\boldsymbol{T}$ の性質を記述する事を考える．図 3.4は以下のベクトル関数による 2 次元平面におけるベクトル場を表している．

$$\boldsymbol{y} = (\boldsymbol{T} - \boldsymbol{I})\boldsymbol{c} = \hat{\boldsymbol{c}} - \boldsymbol{c} \tag{3.16}$$

ここで $\boldsymbol{I}$ は 2 次元における単位行列を表している．行列 $(\boldsymbol{T} - \boldsymbol{I})$ は変換前後の差を表現しており，ベクトル $\boldsymbol{y}$ は 2 次元平面の各点における変換行列 $\boldsymbol{T}$ の影響を表している．図 3.4を見ると，式 (3.16) によるベクトル場は渦を描いている．このことは行列 $\boldsymbol{T}$ による変換が強い回転性を持つことを示唆している．

3.5.2　n 次元ケプストラム空間における回転性

n 次元空間については，前述の 2 次元空間のように変換行列からその回転性のみを抽出してくることは難しい．故に本節では n 次元における回転行列の一般的な定義に基づき，式 (3.7) の行列 $\boldsymbol{A}$ の幾何学的性質を記述する．一般に行列 $\boldsymbol{R}$ が以下の性質を満たすとき，$\boldsymbol{R}$ を回転行列と呼ぶ．

$$\boldsymbol{R}^t\boldsymbol{R} = \boldsymbol{R}\boldsymbol{R}^t = I \tag{3.17}$$

$$\det \boldsymbol{R} = +1 \tag{3.18}$$

以下，式 (3.17) および式 (3.18) が式 (3.7) の $\boldsymbol{A}$ に関して近似的に成立することを示す．$|\alpha| \ll 1$ が成立すると仮定すると，$\boldsymbol{A}$ は 2 次以上の高次項を無視して以下のように近似する事ができる．

$$\boldsymbol{A}_n = \begin{pmatrix} 1 & 2\alpha & 0 & \cdots & \cdots \\ -\alpha & 1 & 3\alpha & 0 & \cdots \\ 0 & -2\alpha & 1 & 4\alpha & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \tag{3.19}$$

このとき行列 $\boldsymbol{A}_n$ の要素 a_{ij} は以下のように表される．

$$a_{ij} = \begin{cases} 1 & (i = j) \\ \mathrm{sgn}(j-i) * j\alpha & (|i-j| = 1) \\ 0 & (\text{otherwise}) \end{cases} \tag{3.20}$$

ここで $\mathrm{sgn}(j-i)$ は $j-i > 0$ のとき $+1$ を，$j-i < 0$ のとき -1 を返す関数である．このとき $\boldsymbol{A}_n^t \boldsymbol{A}_n$ および $\boldsymbol{A}_n \boldsymbol{A}_n^t$ がおよそ単位行列となることを示す．

$$\boldsymbol{A}_n^t \boldsymbol{A}_n = \begin{pmatrix} 1+\alpha^2 & \alpha & -3\alpha^2 & 0 & \cdots \\ \alpha & 1+8\alpha^2 & \alpha & -8\alpha^2 & \cdots \\ -3\alpha^2 & \alpha & 1+18\alpha^2 & \alpha & \cdots \\ 0 & -8\alpha^2 & \alpha & 1+32\alpha^2 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \tag{3.21}$$

式 (3.21) の対角成分は $k \in \mathcal{R}$ として $1+k\alpha^2$ と表される．さらに $|i-j| = 1$ の要素は α，$|i-j| = 2$ の要素は $m \in \mathcal{R}$ として $m\alpha^2$，それ以外の要素は 0 となる．一方 $\boldsymbol{A}_n \boldsymbol{A}_n^t$ は以下で表される．

$$\boldsymbol{A}_n \boldsymbol{A}_n^t = \begin{pmatrix} 1+4\alpha^2 & \alpha & -4\alpha^2 & 0 & \cdots \\ \alpha & 1+10\alpha^2 & \alpha & -9\alpha^2 & \cdots \\ -4\alpha^2 & \alpha & 1+20\alpha^2 & \alpha & \cdots \\ 0 & -9\alpha^2 & \alpha & 1+34\alpha^2 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \tag{3.22}$$

式 (3.21)(3.22) ともに，$|\alpha| \ll 1$ の仮定のもと α^2 の項を無視する事ができる．この時，二つの行列積は対角成分が 1，$|i-j| = 1$ の成分が α となる三重対角行列と考える事ができる．このことから $\boldsymbol{A}_n^t \boldsymbol{A}_n$ および $\boldsymbol{A}_n \boldsymbol{A}_n^t$ は厳密には単位行列にはならないものの，高い直交性を持っているといえる．そのため近似的に式 (3.17) を満たしているといえる．

一方 $\boldsymbol{A}_n$ は三重対角行列であるため，その行列式を求める事ができる [34]．n 次の三重対角行列の行列式は再帰的に以下の形で求められる．

$$\det \boldsymbol{A}_n = a_{nn} \det \boldsymbol{A}_{n-1} - a_{n(n-1)} a_{(n-1)n} \det \boldsymbol{A}_{n-2}. \tag{3.23}$$

式 (3.19) より，$a_{nn} = 1$ および $a_{n(n-1)} a_{(n-1)n} \sim \alpha^2$ となる．ここで $|\alpha| \ll 1$ の仮定を用いれば，$\det \boldsymbol{A}_n \approx \det \boldsymbol{A}_{n-1} \approx \cdots \approx \det \boldsymbol{A}_1 \approx 1$ と再帰的に行列式を求める事ができる．よって $\boldsymbol{A}_n$ は近似的に式 (3.18) を満たす．

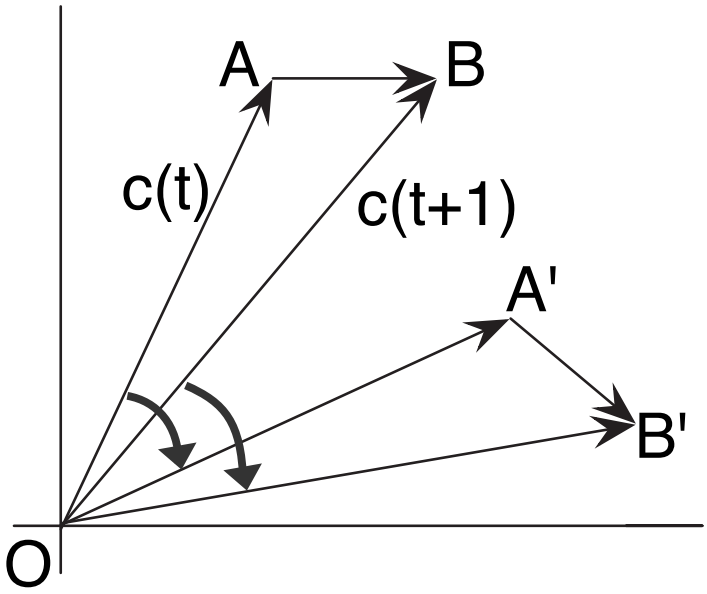


図 3.5: ケプストラムベクトルの回転とそのデルタベクトル

上記の議論から，式 (3.7) の行列 $\boldsymbol{A}$ による変換は n 次元空間において回転性を有しているといえる．しかし，上記の議論にはいくつかの近似を導入しているため，実験的に $\boldsymbol{A}$ の回転性を確かめる必要がある．

いま 2 次元空間で得られた図 3.4が n 次元空間においても観察されると仮定する．このとき行列 $\boldsymbol{A}$ による変換の性質が予測できる．図 3.4は行列 $\boldsymbol{T}$ による変換が全ての点を，およそ等しい回転角で回転させている事を示している．さらにこの回転角はおよそ α のみに依存している．このことからケプストラムベクトルは音素や性別に依存せずウォーピングパラメータのみに依存して近い回転角で変化することが示唆される．

一方，Δパラメータに関する性質も予測ができる．図 3.5は時刻 t および $t+1$ におけるケプストラムを c_t, c_{t+1} として，c_t, c_{t+1} および $\Delta c = c_{t+1} - c_t$ の回転の様子を示している．空間中の各ベクトルがおよそ等しい角度で回る場合，これらの Δパラメータも同様に回転することになる．これらは二つの Δベクトルに対して $\Delta\Delta$ ベクトルを考えた場合も同様である．即ち，空間中の各ベクトルがおよそ等しく回転する事はこれらの高次の Δベクトルも同様に回転する事を示している．3.3で述べたように，Δパラメータの導入は微分演算に相当し，音声認識における学習データと評価データのミスマッチ問題に対してある程度の効果を示す事が知られている．しかし，本節における考察により，空間における回転に相当する変換に対しては，Δパラメータの導入がおよそ効果がない事が示唆される．

3.6　分析再合成音声による実験

3.6.1　実験条件

前節で議論した声道長変化に伴うケプストラムの回転性が実際の音声でどのように表出するかを調べるため分析再合成音による評価実験を行った．

本実験では成人男女 1 名ずつの日本語 5 母音の連続発声/aiueo/の音声を用いた．これらに対し

表 3.1: 音響分析条件（3.6.1節）

サンプリング条件	16 kHz / 16 bit
フレーム窓	Hamming window
フレーム長	25 ms
シフト長	5 ms
ケプストラムパラメータ	MFCC (1–12)

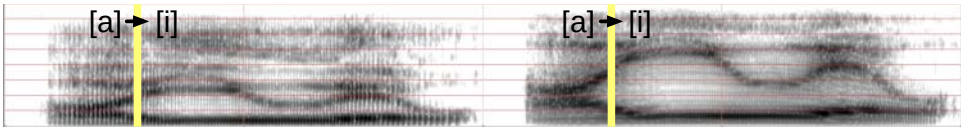


図 3.6: ウォーピング前後の音声のスペクトログラム

て STRAIGHT[35] を用いて周波数ウォーピングを行った．これらのウォーピング後の音声が声道長の異なる話者に対応する．これらの音声について表 3.1に示す音響分析条件で分析を行った．加えて ΔMFCC，ΔΔMFCC についても抽出した．作成した音声資料の同一時刻についてウォーピング前後のパラメータの比較を行った．比較については二つのパラメータベクトルのなす角を用いた．二つのベクトル a, b のなす角 θ については式 (3.24) で求められる．

$$\theta = \arccos \frac{a \cdot b}{|a||b|} \tag{3.24}$$

ここで $a \cdot b$ はベクトルの内積を，$|a|, |b|$ はベクトルのノルムを表す． 声道長変化との対応関係を明確にするため，周波数ウォーピングに際して式 (3.5) および式 (3.7) を直接的に用いず，下記の折れ線関数を用いて行った．

$$\hat{\omega} = \begin{cases} \frac{1}{m}\omega & (0 \le \omega < \frac{m}{1+m}\pi) \\ m(\omega - \pi) + \pi & (\frac{m}{1+m}\pi \le \omega \le \pi) \end{cases} \tag{3.25}$$

上記の折れ線関数は図 3.2を低域，高域ごとに二つの直線で近似したものである．式 (3.25) における m はウォーピング前後の音声の声道長の比を表している．式 (3.25) によってウォーピングを施した音声の例を図 3.6に示す．図 3.6において，左がウォーピング前，右がウォーピング後の音声である．色の濃い部分がフォルマントを表しているが，ウォーピングに伴い，これらが高周波数方向にシフトしていることがわかる．

式 (3.25) における m と α は 1 対 1 に対応させる事ができる．それぞれの α について，図 3.2におけるウォーピング関数を式 (3.25) で近似した場合の 2 乗誤差が最小となる m を求めた．これらの α と m との対応関係を図 3.7に示す．図 3.7によれば身長 180 cm の話者について $\alpha = 0.4$ のウォーピングは身長約 90 cm の話者に対応し，$\alpha = -0.4$ のウォーピングは身長約 360 cm の話者に対応する．

3.6.2　実験結果

図 3.8は母音の遷移部分（/a/ から /i/，/i/ から /u/，/u/ から /e/，/e/ から /o/）について，身長を横軸に，ウォーピング前のパラメータベクトルを基準とした式 (3.24) の角度を縦軸に

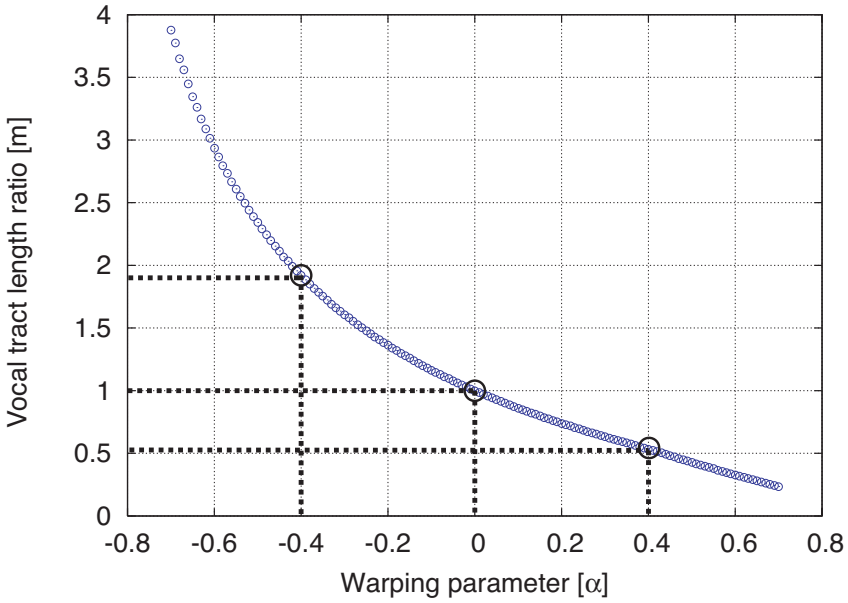


図 3.7: ウォーピングパラメータ α と声道長比 m との対応関係

表 3.2: 変化させた実験パラメータ

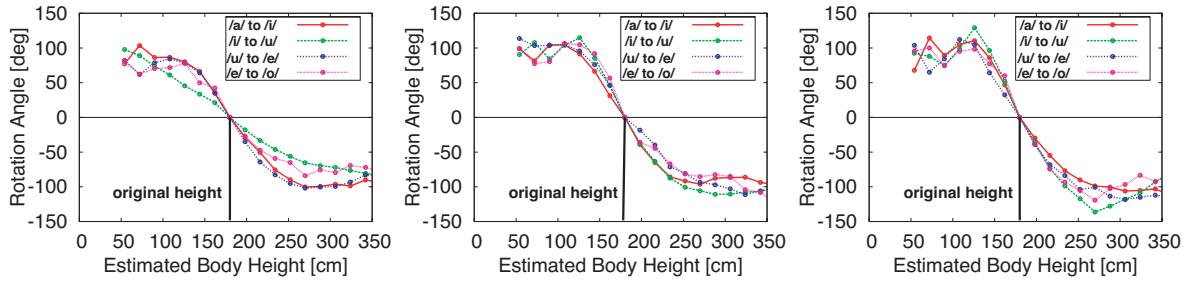
次元数	2, 3, 4, 6, 8, 12
ケプストラム	FFTCep, MelCep, MFCC
分析箇所	定常部 5 カ所, 遷移部分 4 カ所

プロットしたものである．(a) から (c) までが男性話者の，(d) から (f) までが女性話者のデータである．さらに図 3.8の左から順に MFCC，ΔMFCC，ΔΔMFCC に対応している．図 3.8によれば，性別，音韻にあまり依存しない回転性を示している事がわかる．加えて MFCC，ΔMFCC，ΔΔMFCC ともに強い回転性を示しており，微分演算の回数を増やす事でむしろ回転角が大きくなっている傾向がみられる．

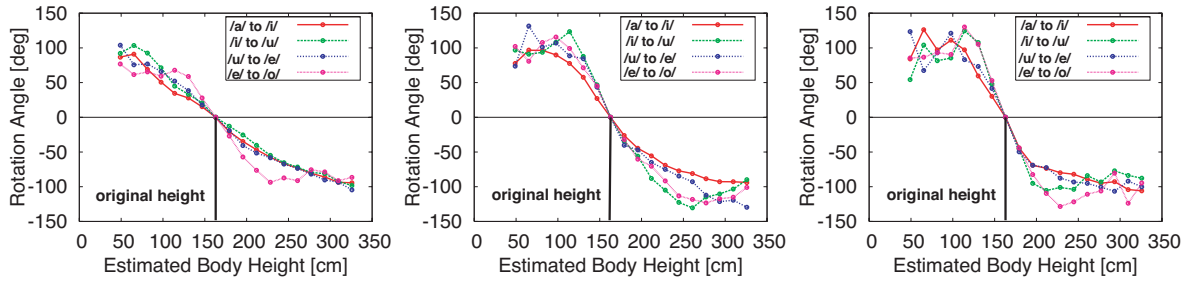
3.6.3　パラメータを変化させた場合の実験

通常，音声情報処理で用いられるケプストラムの次元数は 10–12 次元以上である．これはスペクトルの特徴を捉える上で最低限，この程度の次数が必要であるためである．一方本章における議論は 3.5.1，3.5.2に示したように，基本的に次元数に依存せずに成立する．また音声情報処理で用いられるいくつかのケプストラムパラメータについても共通の傾向がみられるものと考えられる．そこで前節の実験条件に加えてケプストラムパラメータの次元数，種類を変化させた場合の実験を行った．加えて母音の定常部分における回転性と遷移部分における回転性とを比較するため，音声の分析箇所として定常部の中心部分 5 箇所についても調べた．変化させたパラメータを表 3.2に示す．分析窓長や種類等その他の分析条件は表 3.1と同様である．

ケプストラムの種類を変化させた場合の結果を図 3.9に示す．その他の条件として，次元数は 12 のΔパラメータ，分析箇所は遷移部分 4 カ所，発話者は男性である．図 3.9によればどのパラメータでもおよそ同様の回転傾向がみられる．一方局所的にみれば (a):ΔMFCC の /a/ から /i/ の遷移部分における回転性が異なる．これは本実験で用いた MFCC の実装が高次ケプストラムの

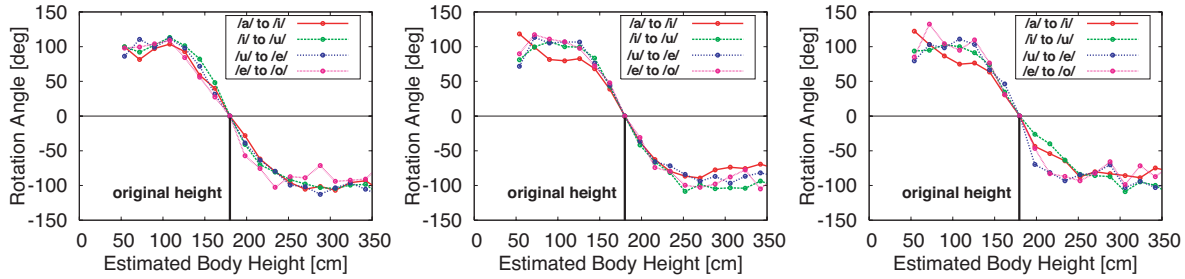


(a):MFCC (male)　(b):ΔMFCC (male)　(c):ΔΔMFCC (male)



(d):MFCC (female)　(e):ΔMFCC (female)　(f):ΔΔMFCC (female)

図 3.8: 身長と回転角との対応関係: (a)–(c): 男性話者（身長 180 cm）;(d)–(f): 女性話者（身長 163 cm） の女性話者のデータ

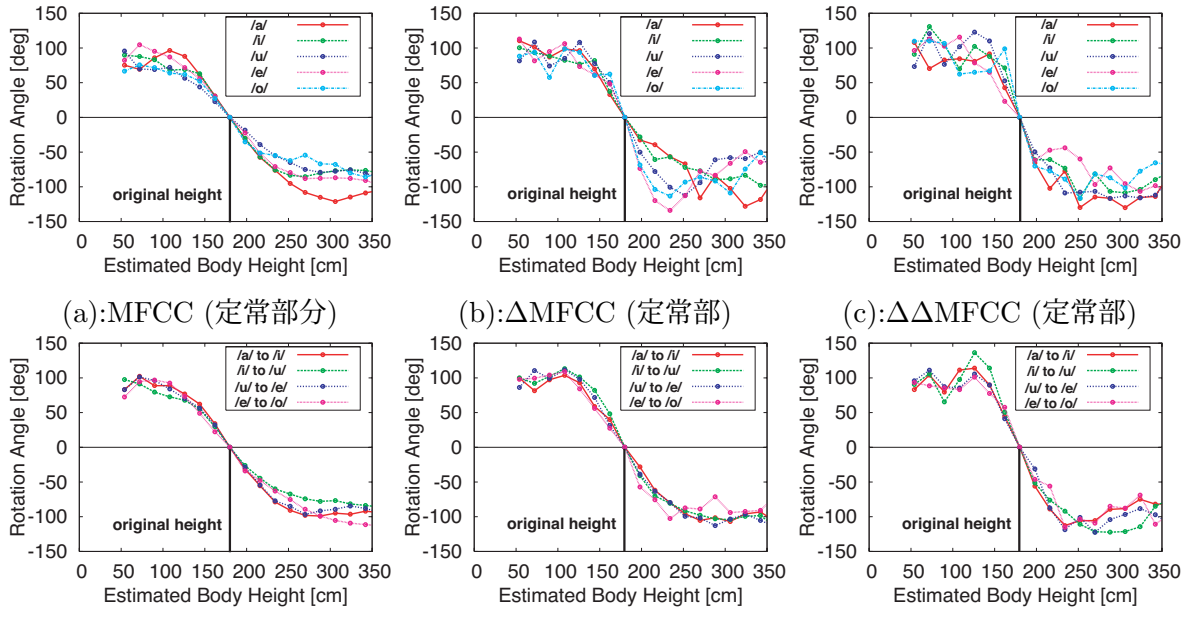


(a):ΔMFCC　(b):ΔMCEP　(c):ΔFCEP

図 3.9: 異なるケプストラムパラメータの回転性: (a): ΔMFCC; (b):ΔMCEP; (c):ΔFCEP;

ダイナミックレンジ拡張のため，高次項を大きくする処理が行われている為と考えられる．また (b):ΔMCEP と (c):ΔFCEP にはほとんど差がない事が分かる． 一方，遷移部分と定常部分とを比較した結果を図 3.10に示す．分析パラメータは MFCC，ΔMFCC，ΔΔMFCC の 3 種類，話者は男性である．図 3.10によれば，大局的な回転傾向は全てにおいてみられるが，(b) および (c) において値が乱雑に変化している事がわかる．これは定常部における Δ パラメータが非常に小さい値となり，結果的に乱雑な回転性を示したと考えられる．また (a) より定常部のケプストラムにおいても顕著な回転を示している事が分かる．

最後にケプストラムの次元数を変化させた結果を図 3.11に示す．分析パラメータおよび分析箇所は MFCC の定常部および ΔMFCC の遷移部分で，話者は男性である．図 3.11によれば，MFCC の定常部分 (a)–(c) において次元数が増加するにつれて，回転角が増加している様子がわかる．これは多次元において回転の自由度が高くなっていることを示している．一方遷移部分については大局的な回転傾向があるものの，次元数の増加に伴って，乱雑な回転が少なくなっている．これも回転の自由度が高くなる事により，声道長の変化に伴って連続的な回転変化が実現されたもの



(a):MFCC (定常部分)　(b):ΔMFCC (定常部)　(c):ΔΔMFCC (定常部)

(d):MFCC (遷移部)　(e):ΔMFCC (遷移部)　(f):ΔΔMFCC (遷移部)

図 3.10: 異なる分析箇所における回転性: (a)–(c):定常部分; (d)–(f):遷移部分

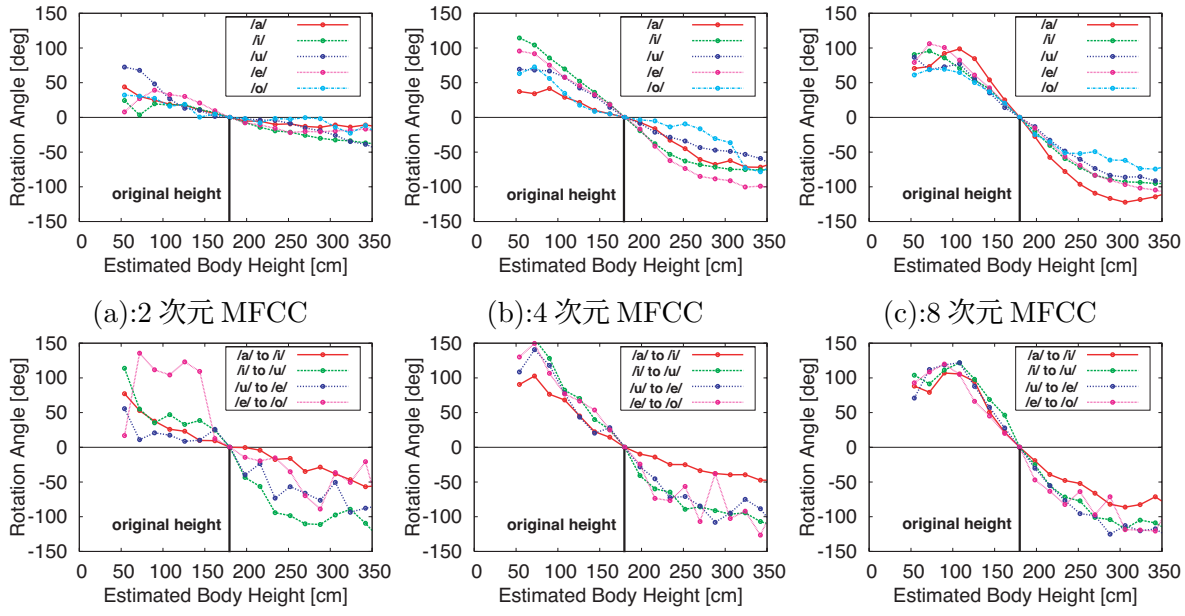
と考えられる．またこのことはケプストラムの低次元のみでスペクトルを表現することの限界を示しているとも考えられる．

3.7　考察

例えば図 3.8(b) に着目すると，身長 180 cm の話者とウォーピング後の身長 120 cm に対応する話者について，ΔMFCC のベクトルがおよそ直交してくることになる．すなわち成長に伴って，話者の音声のパラメータベクトルが回転し，いずれベクトルとして直交性を持つ事になる．このことからパラメータベクトルの回転度という指標が性別や年齢，身長といった声道長に依存する話者情報の表現として有用であり，話者の声道長の情報を推定する有効なパラメータである事が示唆される．一方，α の絶対値が小さい場合にはおよそ等しい回転傾向がみられるが，α の絶対値が大きく，すなわち話者の声道長の違いが大きくなる場合には音韻や性別によって回転角が異なる様子が観察される．これらは 3.5.2における議論において α の絶対値に基づく近似を多く含んでいるため，α の絶対値が大きい場合の回転が異なる様子を示している事が考えられる．しかしこれらの異なる回転性によって音韻や性別を表現できる可能性も示唆しており，今後詳細な検討を行っていく必要がある．また次元数が異なる場合や分析箇所によって異なる回転性を示しているため，これらの分析も行う必要がある．

一方，これらの結果が音声認識システムに与えている影響について考える．前述の通り，同一の音韻を表すパラメータベクトルが，身長 180 cm の話者と身長 120 cm の話者とではおよそ直交している．すなわちこれは成人で構築した音響モデルによって子供の音声を認識させるタスクに対応しており，今回の実験結果は従来の音声認識システムが子供の音声のような特異音声の認識を苦手とすることの一因を定量的な形で示したものといえる．

本章で述べたように，音声の動的特徴を表現する場合に Δケプストラムは広く用いられている．Δパラメータの導入は微分演算に相当し，マイクの伝送特性に代表されるような，スペクトルに対



(a):2 次元 MFCC　(b):4 次元 MFCC　(c):8 次元 MFCC

(d):2 次元 ΔMFCC　(e):4 次元 ΔMFCC　(f):8 次元 ΔMFCC

図 3.11: 異なる次元数における回転性: (a)–(c):MFCC 定常部分; (d)–(f): ΔMFCC 遷移部分

する乗算性歪みに起因するミスマッチ問題に対してある程度の効果を示す事が知られている．しかし本章で議論したような線形変換性歪みによるミスマッチ問題に対してはその効果が低い事が示唆される．動的特徴の表現として Δケプストラムのようなベクトル表現を用いる限り，本質的にこのような問題は不可避であると考えられる [36].

3.8　まとめ

本章では，音声に不可避に混入する非言語的特徴についてモデル化し，従来行われてきたこれらに対する対応手法について紹介した．加えて，特に声道長の変化に着目して，その特徴量空間における幾何学的特性について論じた．声道長の変化をケプストラム空間における線形変換として表現し，この変換行列の性質を数学的に議論する事で，その変化がケプストラム空間における回転性として表出することを示した．加えてこれらの回転性が音韻や性別，Δパラメータの次数におよそ依存しない事を確認した．またこれらの幾何学的性質を確認するため，周波数ウォーピングを施した分析再合成音を用いて，実際に声道長の変化が回転性として表出することを示した．

声道長変化による回転性はおよそ等しく作用しているが，音韻や性別，Δパラメータの次数による影響も多少見られた．一方，高次の回転行列については対角化などの操作によりその様子を詳細に記述することが可能である [37]．現在これらの分析を行っており，今後話者情報表現としての検討を行っていく．

一方で実験結果は音声の全体的，動的な特徴を捉える場合に，ベクトル表現を用いることの本質的な問題点を示している．以降本論文では，スカラー特徴量によって動的特徴を捉えることにより，非言語的特徴に対して本質的に不変な表象について議論し，その音声合成への応用を検討する．

第4章

音声の構造的表象とそれに基づく音声合成の枠組み

4.1　はじめに

前章では，音声にはケプストラム $\boldsymbol{c}$ に対するアフィン変換 $\boldsymbol{Ac}+\boldsymbol{b}$ によって表現される非言語的特徴が不可避的に混入することについて述べた．加えてこれらに対する従来のアプローチを紹介した．また特に変換行列 $\boldsymbol{A}$ で表される声道長の変化に着目し，声道長変化がケプストラム空間における回転性として表出することを理論的，実験的に示した．

音声模倣を考える上で，音声の全体的かつ動的な特徴を捉えることは必要不可欠である．しかし，ベクトル表現としての動的特徴量は本質的に上記の回転性の影響を受けると考えられる．すなわち前章での結果を基に論ずれば，音声の物理的実体をそのままモデル化する限り，身長 120 cm の兄の「おはよう」と身長 180 cm の父の「おはよう」の同質性を説明する事はできないといえる．両者の「おはよう」は音響特徴量としては直交し，絶対量としての関連は希薄であるためである．一方で幼児はこの同質性を容易に感覚することができる．

近年提案されている音声の構造的表象は，上記のような非言語的特徴による歪みに対して本質的に不変な音声表象である．構造的表象に基づいた全体的，動的特徴の表現はスカラー特徴に基づいており，上記の回転性の影響を受けない．すなわちこれは物理量の絶対座標を除外したモデル化である．本章では，この音声の不変表象について説明する．その出発点として，まず音声に内在する普遍構造について，その言語学的および発達心理学的な側面について言及し，幼児の音声模倣との関連について述べる．続いて音声の構造的表象について説明し，本研究において提案する，音声の構造的表象に基づく音声合成の枠組みについて述べる．

4.2　差異，対立に基づく言語体系

近代言語学の祖であるソシュールは言語に対して,「言語が含むのは，言語体系に先立って存在する観念でも音でもなく，ただ，その体系から生じる概念的差異と音的差異とだけである」と述べている [38]．このソシュールの言語哲学からヤコブソンらの「構造音韻論」へと発展した．これは弁別素性と呼ばれる，二つの音素を区別する音声的特徴を用いて，音素の違いや全ての音素間差異によって構成される幾何学的構造を議論する学問分野である．例えばヤコブソンはフランス語の母音体系を弁別素性を用いて幾何学的に構造化している [39]．図 4.1はヤコブソンの幾何学的音韻構造を示している．すなわち言語体系の本質は差異，対立にあると考えられている．

一方で，早川は，母親が「マンマ（ごはん）？」と聞いた際に，十一ヶ月の幼児が「パッパ」と述べた例を紹介し，[m] と [p] の音声的対立が既に成立していることを述べている [6]．さらに，ことばという体系の獲得には，単なる模倣や暗記によって獲得することはできず，差異，対立が必須であることについても言及している．さらに音韻体系や語彙体系が複雑に絡みあいながら獲得される過程について述べている．

加えて構造音韻論では同一の幾何構造が話者を問わず普遍的に存在すると主張する．また幼児の音声模倣や言語獲得においても当然父親，母親の発話から自身の音声による模倣，獲得を行う．そこに両親の話者性は伴わない．一方で音声の構造的表象は前章で行った非言語的特徴の数学的モデル化に基づいて，非言語性歪みを原理的に有しない物理表象として提案されてきた．この点から，音声の構造的表象は構造音韻論を物理的に実装したものと解釈でき，また本研究で提唱する音声合成の枠組みは幼児の音声模倣のモデル化に相当する．そしてこれらは全て，方向性を伴わない差異，対立をその基盤としている．

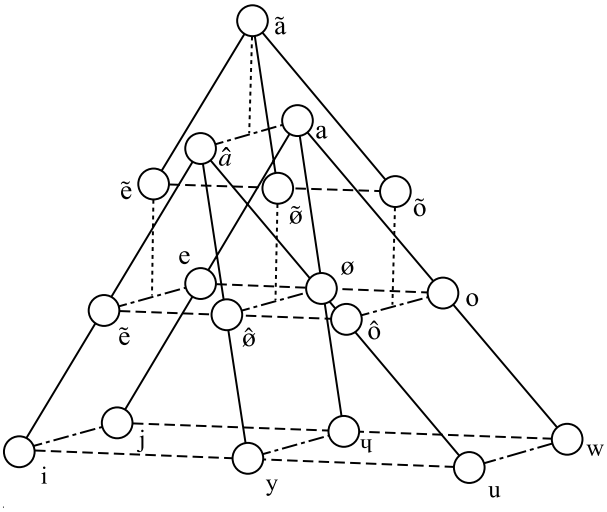


図 4.1: ヤコブソンの幾何学的音韻構造

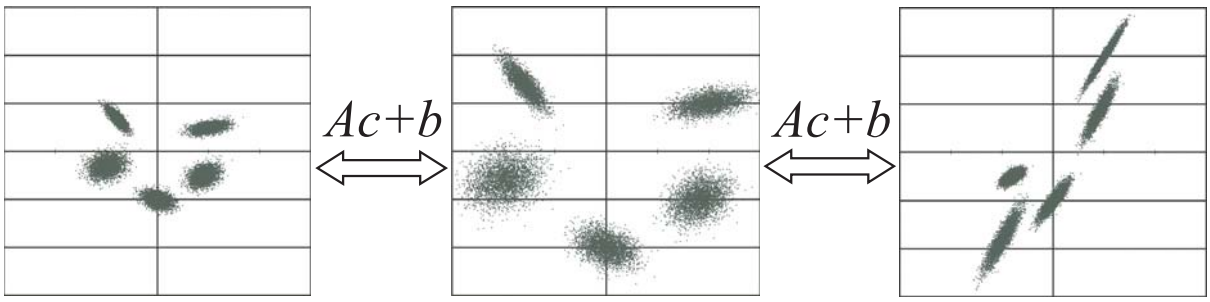


図 4.2: アフィン変換による分布の変化; これらは全て同一の構造である

4.3　音声の構造的表象

ユークリッド空間において N 角形の形状は ${}_NC_2$ 個の全ての頂点間距離を規定する事で，鏡像の曖昧性を除けば一意に定めることができる．即ち事象群に対して，全ての事象間距離を求めることでその事象群を構造的に表象することになる．しかしケプストラム空間において N 点の「**点間距離**」によって構造を規定した場合，その構造は非言語的特徴によって不可避に歪む．なぜなら，非言語的特徴はケプストラム空間におけるアフィン変換としてモデル化され，アフィン変換は特殊な場合を除けば，構造を歪ませる変換である為である．しかしこの不可避に歪む構造は空間自体を歪ませる事で不変構造として定義することができる．

峯松らは上記の不変構造の導出に関して，以下の構造不変の定理を示した [25].

構造不変の定理

意味のある記述が分布としてのみ可能な物理現象を考える．分布群に対して，全ての二分布間距離を求める（距離行列）．二分布間距離として，バタチャリヤ距離，カルバック・ライブラ距離，ヘリンガー距離などを用いた場合，各分布に対して単一の任意一次変換を施しても，二分布間距離は不変である．即ち距離行列は不変であり，その結果，構造も不変となる．

以下,「**分布間距離**」の一つである Bhattacharyya 距離 (以下 BD と記述) を考えた場合，任意

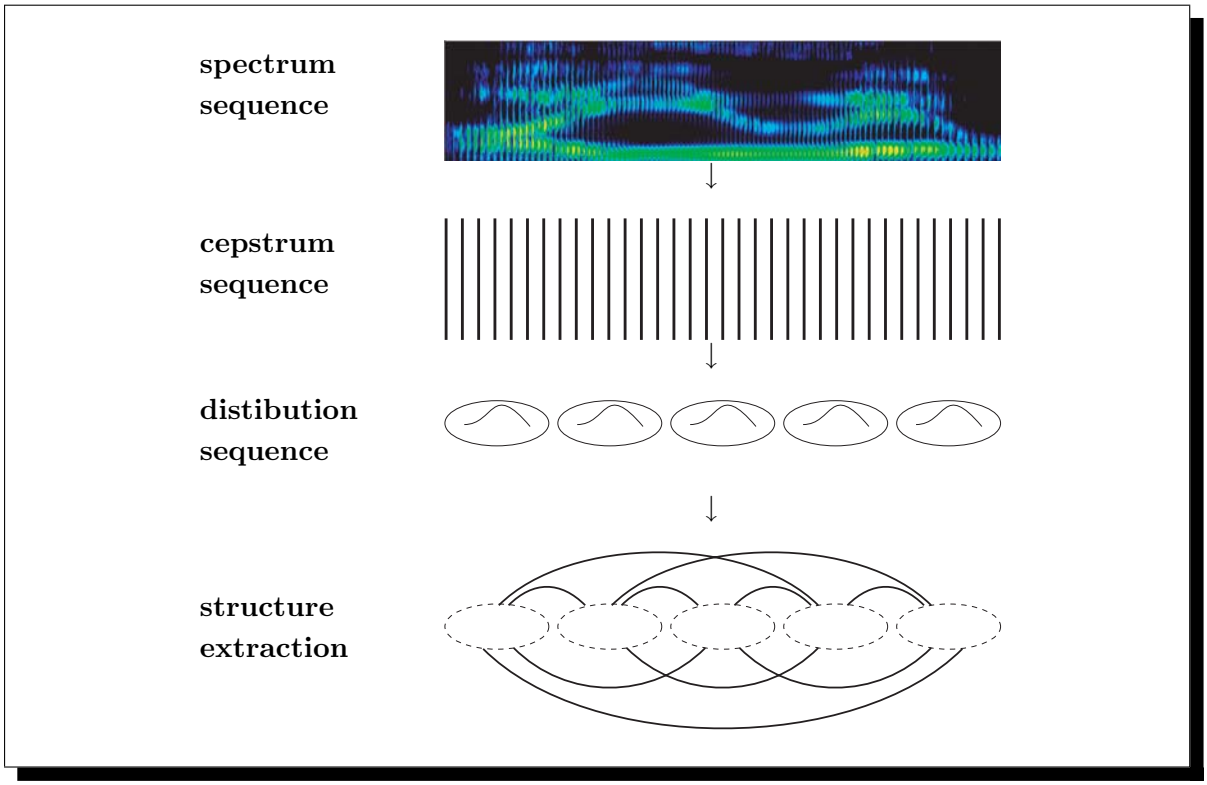


図 4.3: 音声からの構造的表象の抽出

の二つの分布の確率密度関数を $p_1(x), p_2(x)$ として以下で表される．

$$BD(p_1(x), p_2(x)) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}\, dx \tag{4.1}$$

このとき $0 \leq \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}\, dx \leq 1$ を確率として解釈すれば，BD は情報量尺度となり，その単位は [bit] である．二つの分布に対して共通のアフィン変換 $\boldsymbol{Ac} + \boldsymbol{b}$ を施した場合，BD は変換前後で不変となる．なおこの不変性は 1 対 1 対応のとれる非線形変換においても成立する [40]．図 4.2はいくつかのアフィン変換を 2 次元平面における点群に対して適用した様子を表している．これらは一見すると，その幾何学構造を大きく変形してしまっているように見える．しかし「**分布間距離**」を距離尺度として用いる事でこれらは同一の不変構造となる．これは構造不変の定理を満たす分布間距離が空間を歪める距離尺度であることに起因する．

非言語的特徴はケプストラム $\boldsymbol{c}$ に対するアフィン変換 $\boldsymbol{c'} = \boldsymbol{Ac} + \boldsymbol{b}$ で表される．即ちケプストラム空間において音響事象を分布として捉え，音響事象群を「**分布間距離**」のみによって定義することで，変換不変，即ち非言語性歪みにおよそ不変な構造を求める事ができる．このときアフィン変換 $\boldsymbol{Ac} + \boldsymbol{b}$ はこの音響的不変構造に対して，その絶対座標上の位置を回転（$\boldsymbol{A}$）とシフト（$\boldsymbol{b}$）によって変化させているにすぎないといえる．

4.4　一発声の構造化

一発声を一つの構造的表象で記述する場合を考える．図 4.3に一発声の音声からの構造的表象の抽出の流れを示す．音声の時系列信号は，まず短時間スペクトル系列からケプストラム系列へ

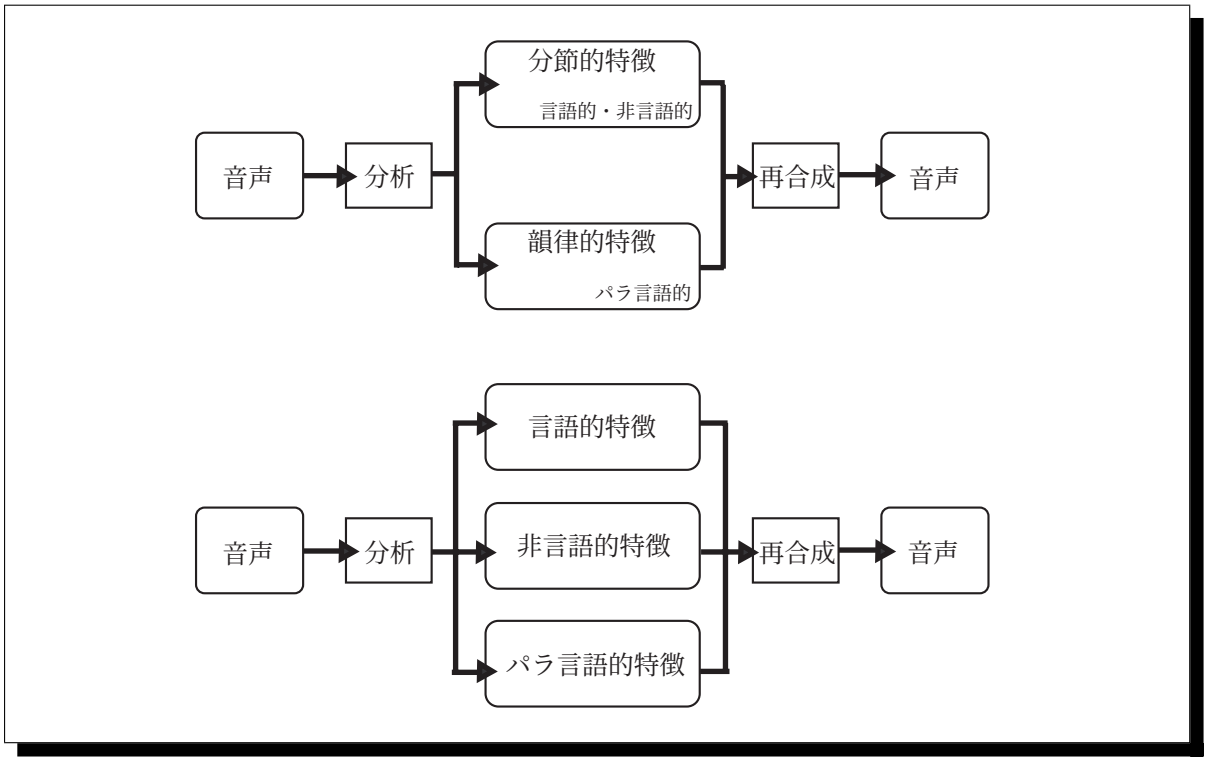


図 4.4: 従来の分析再合成系の枠組み（上）と提案する枠組み (下)

と変換される．得られたケプストラム系列もまた時系列信号であるが，これを適当な時間区間において音響事象の分布としてとらえ，その分布の時系列へと変換する（このとき各分布に対応する時間長は分布によって異なる）．これら系列中の各分布に対して全ての組み合わせの分布間距離を求めることで一発声が構造化される．なお構造的表象はパラメータとしてケプストラムを要求する訳ではない．スペクトル包絡とケプストラムは FFT で変換でき，FFT も線形変換であるため，スペクトル表現における議論も可能である．

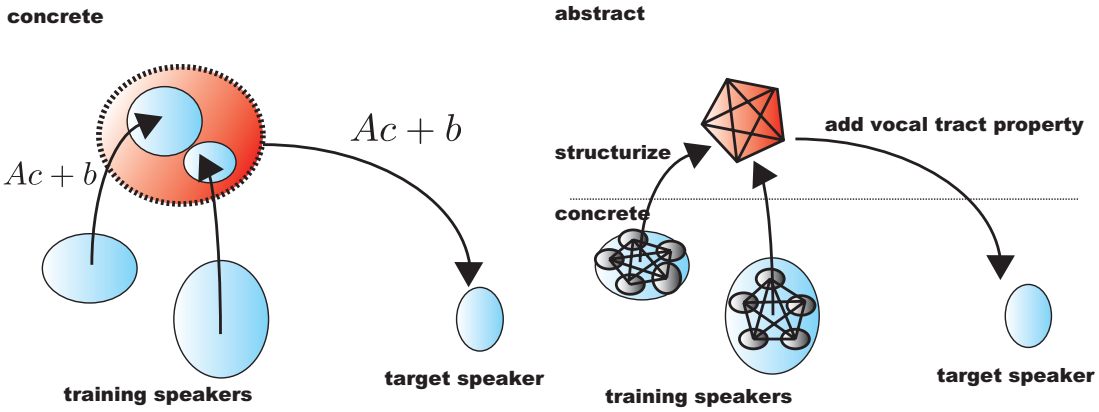
4.5　構造的表象に基づく音声合成

4.5.1　非言語的要因をも分離する分析再合成系

従来の音声合成では，学習話者による数百〜数千文の音声試料を学習データとして異音（シンボル）と音との対応を学習する．そして入力としてテキスト（異音列）を与えた場合に音ストリームを出力する．この場合得られるのは学習話者の声である．

幼児の音声模倣では両親の音声を模倣しても，自らの声で言葉を返す．第 1章で述べたように，幼児の音声模倣では両親の発声全体の語形を真似，自らの声で返していると考えられている．

本研究で提案する音声合成の枠組みは，生成対象の語形に対して，発声者の身体性（声道形状特性）を与えることで初めて音が生まれるという合成系である．分析再合成系でこの考えを示すと図 4.4のようになる．従来の分析再合成系では音声を分節的特徴（主にはスペクトル包絡に対応し，言語情報・非言語情報を伝搬）と韻律的特徴（主にピッチ，パワー，継続長に対応し，パラ言語情報を伝搬）に分解する．これは音声生成のソースフィルタモデルに由来する．一方提案する枠組みはこれをさらに細分化し，言語的特徴，非言語的特徴，パラ言語的特徴に分ける枠組み



(a): 話者適応学習における枠組み　(b): 構造的表象における枠組み

図 4.5: 音声模倣を実現する声質変換の枠組みの比較

である．この時，言語的特徴とパラ言語的特徴を与えても音は生成されない．生成する話者は非言語的特徴の担い手であるからである．この担い手の音響特性（具体的には声道形状），更には伝送媒体のチャネル特性が与えられて初めて，聞き手が聴取できる音響信号が生まれる．このことは幼児が両親の発話全体の語形を獲得し，自らの発声器官を使って言葉を発する過程をモデル化したものといえる．

4.5.2　構造的表象を介した音声模倣のモデルと話者適応学習との比較

ここで，構造的表象に基づいた音声模倣の解釈について説明する．音声模倣では両親の発声全体からその語形を抽出し，この語形を幼児自身の声で真似ることになる．これを両親の音声から幼児の音声への声質変換という観点から考えてみる．声質変換は入力音声をその音声と発話スタイルや話者の異なる音声へと変換する技術である [41]．この時，構造的表象に基づいた音声合成，音声模倣の枠組みは複数の話者から共通のモデルを介して変換したい声（ここでは幼児の声）へと変換しているとも解釈可能である．一方，共通のモデルを媒介とするという考え方は，話者適応学習 (SAT) にも見る事ができる．SAT (Speaker Adaptive Training) は複数の話者の音声データを得た場合，各話者のケプストラム毎に個別のアフィン変換 $\boldsymbol{Ac}+\boldsymbol{b}$ を行い，その上でモデルの学習を行うことで「架空の特定話者」のモデルを構築する [42]．その上で変換したい声に対する，この「特定話者」モデルからのアフィン変換を MLLR (Maximum Likelihood Linear Regression) の枠組みで推定し [43]，推定された変換を用いて目的音声への変換を実現する．

二つの枠組みを図 4.5に示す．これらの二つは，模倣すべき音声に対してある種の抽象化を施して共通のモデルを構築し，そのモデルを用いて解釈するという観点からみると共通点がある．SAT に基づく声質変換では，模倣（変換）すべき音声を抽象化する共通のモデルは「架空の特定話者」の音響的・絶対的なモデルであるとして捉え，このモデルに対する変換を通して変換対象の音声を獲得する（図 4.5(a)）．このとき模倣対象の音声，共通モデル，変換対象の音声は全て音響特徴量空間上に存在し，これらの間を特徴量空間におけるアフィン変換 $\boldsymbol{Ac}+\boldsymbol{b}$ を通して行き来することになる．このとき全ての音声およびモデルは非言語的特徴を不可避的に内包する．一方，本研究で提案する音声の構造的表象に基づく音声合成，音声模倣では模倣すべき音声を抽象化する共通のモデルは，非言語的特徴を取り除いた抽象的・相対的なモデルとして捉え，このモデルに対して，変換対象の音声に対応する非言語的特徴（身体特性）を与えることでその音声を

獲得する（図 4.5(b)）．このとき模倣対象の音声および変換対象の音声は音響特徴量空間上に存在するが，共通モデルは全く別の抽象空間に存在しているといえる．そして身体特性の付与，除去を通してこれらのモデルの間を行き来することになる．なおこのような抽象化により，抽象空間に存在する構造を介して音のストリームと身体運動（手や舌の動き）のストリームを結びつける，メディア変換も検討可能である．

両者を比較した時，幼児が両親の声を「架空の特定話者」の声として解釈し，それに変換をかけることで自らの声を生成するという解釈は果たして妥当だろうか．本研究では工学的実装の容易さではなく，人間の言語獲得のシミュレーションという立場をとり，図 4.5(b) の枠組みでの音声合成を考える．

4.6　まとめ

本章では，言語学的，発達心理学的見地との関連を述べながら，近年提案されている音声の音響的普遍構造について説明した．一発声から構造的表象を抽出する枠組みについても併せて述べた．さらにこの音声の不変表象に基づく音声合成についてその枠組みを提案した．さらに提案する枠組みと話者適応学習に基づく声質変換とを比較し，その特徴を明確にした．

音声の構造的表象は，原理的に非言語的特徴を有しない音声表象である．音声認識システムにおいては，非言語的特徴を除いた抽象空間において認識することが可能であり，すでに成果をあげている [44, 45]．一方音声合成は，最終的な目的が音声の出力であるため非言語的特徴を必ず必要とする．そのため本研究で提案する構造的表象からの音声合成には身体特性の除去に加えて，身体特性を付与することが必要となる．本研究では身体特性の付与を実験的にモデル化し，提案する枠組みの基礎検討として日本語孤立母音系列および連続母音系列の合成を行った．次章よりこれらの実験について述べる．

第5章

音声の構造的表象からの孤立５母音系列の合成

5.1　はじめに

本章より構造的表象に基づく音声合成に関して，その実験的な定式化と合成実験について述べていく．前章で述べた通り，構造的表象に基づく音声合成は身体特性を新たに付与する処理が必要となってくる．その際，構造的表象によって表現される模倣対象（音声模倣における母親に相当）の音声に加えて発話対象話者（音声模倣における幼児に相当）の身体特性をどのように表現するかが問題となる．本研究では構造的表象より得られる制約条件の下でのケプストラム空間の解探索問題としてその定式化を行う．このとき対象単語の語形に加えて，発話者の身体特性をどのように初期条件として導入するかが重要になってくる．

本章では日本語の孤立 5 母音系列を合成対象とした合成タスクを考える．その際，上記で述べたケプストラム空間の探索問題としての定式化を行い，種々の条件下で実験を行った．この際の初期条件の付与をケプストラムの絶対量を部分的に用いる事で行った．またこれらの結果について聴取実験による主観評価を行った．

5.2　探索問題としての定式化

5.2.1　ケプストラム空間の解探索

構造的表象に基づく音声合成を考える場合，音響的実体を得るためには，身体特性を付与する事が必要不可欠になる．この身体特性は人間の音声の場合，声道形状そのものである．すなわち本研究の枠組みは話者の声道形状を構造的表象に付与することにより音声を出力する枠組みとして解釈できる．即ち構造を抽象空間から実空間へと導く枠組みである．

音ストリーム生成時に与える声道形状のパラメータとして調音器官の制御パラメータが考えられる．しかし調音パラメータは複雑であり，ケプストラム空間との対応関係も明確ではない [46, 47]．そのため，今回は提案する音声合成の枠組みの基礎的検討として，ケプストラム空間における制約条件を満たす解の探索問題として定式化する．

今，構造的表象の途中までが声として出力された場面を考える（事象 $s_{t-n}, \cdots, s_{t-1}$ までが出力済み）．このとき次の出力は事象 s_t を声にする操作で得られるが，これを，$s_{t-n}, \cdots, s_{t-1}$ の音的実体 $o_{t-n}, \cdots, o_{t-1}$ からの距離制約を使ってケプストラム空間を探索して求める．このことは，構造的表象により距離関係が与えられた語形に対して，いくつかの音的実体 $o_{t-n}, \cdots, o_{t-1}$ を用いる事で，ケプストラム空間に次の構造要素である s_t を定位させていると解釈可能である．すなわち絶対座標系を有しない物理表象である構造に対して，音響特徴量空間の絶対座標を復帰させていると捉えることもできる．提案する解探索による音声合成の枠組みを図 5.1に示す．この枠組みは構造的表象を制約条件，既に発声された音響事象を初期条件とした探索問題として定式化されている．しかしこのままでは探索空間が膨大となるため空間に制限を加えていく．

5.2.2　探索空間の制限

音響特徴量としてのケプストラムベクトルは 12〜25 次元程度の多次元ベクトルである．そのため適切な解探索のためには，探索対象となる音響空間に適切な制限を加える必要がある．今回，発声全体の平均と分散を用いて探索空間を制限することを考える．

単一の 1 次元ガウス分布においては平均を μ，標準偏差を σ とした時，$[\mu - 3\sigma, \mu + 3\sigma]$ の区間の値をとる確率は約 97%となるため，統計学ではこの区間を全空間として扱うことが多い．よって本研究では各々の次元について探索範囲をこの $\pm 3\sigma$ 区間に限定する．

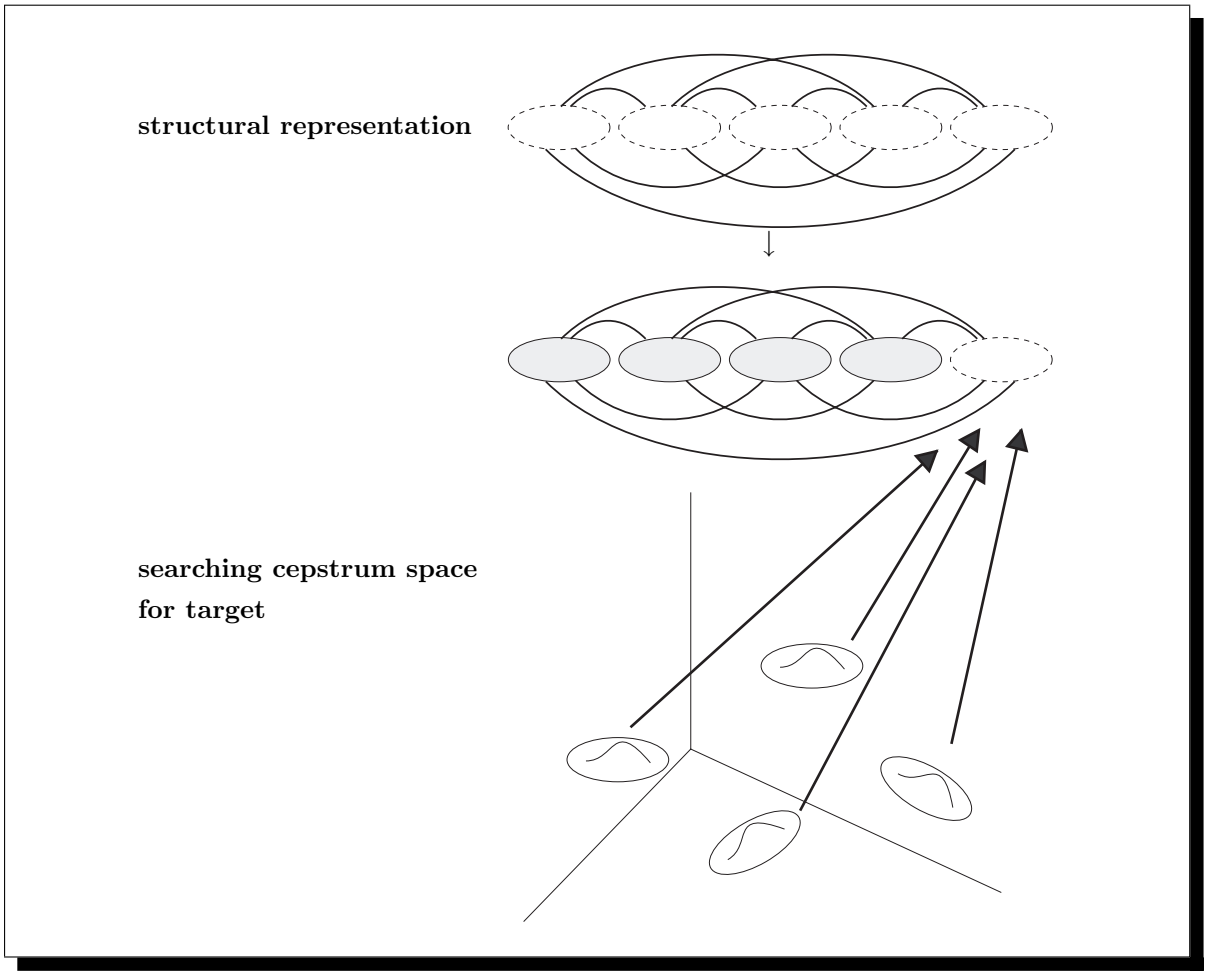


図 5.1: 構造的表象を制約とする解探索による音声合成の枠組み

上記のように探索空間を制限する事は，調音器官の運動しうる範囲の制約をケプストラム空間上で記述していることに相当する．即ち前節における構造の定位と併せれば，解探索のアプローチによっても，空間への構造定位と探索空間の制限によって，発話に身体性を与えるという提案する枠組みを示すことができることを意味している．

5.2.3　特徴量空間分割による余剰空間の制限

構造的表象を用いた音声認識において，構造の “過剰な不変性” のために異なる単語を同一とみなす問題が指摘されている．すなわち絶対座標を伴わない多次元空間における幾何構造は自由度が非常に高く，場合によっては異なる単語から多次元特徴量空間で構造的表象を抽出した場合でも，その幾何構造が一致してしまうことを意味している．このような問題に対して，朝川らは特徴量空間を分割することで話者の違いに対してのみ適切に不変性が成立するような制約条件を導入している [48, 49]．以下この点について述べる．

図 5.2は 2 次元の幾何構造と 1 次元の部分空間への射影構造を示したものである．図 5.2において，(a)(b) に描かれている五角形の幾何構造は 2 次元空間において合同である．しかしそれぞれ x および y 軸への射影を考えれば，これらの各軸上での構造は異なる事になる．この議論を多次元に拡張する事は容易であり，本研究では以降，射影する部分空間の次元数 m をブロックサイズ

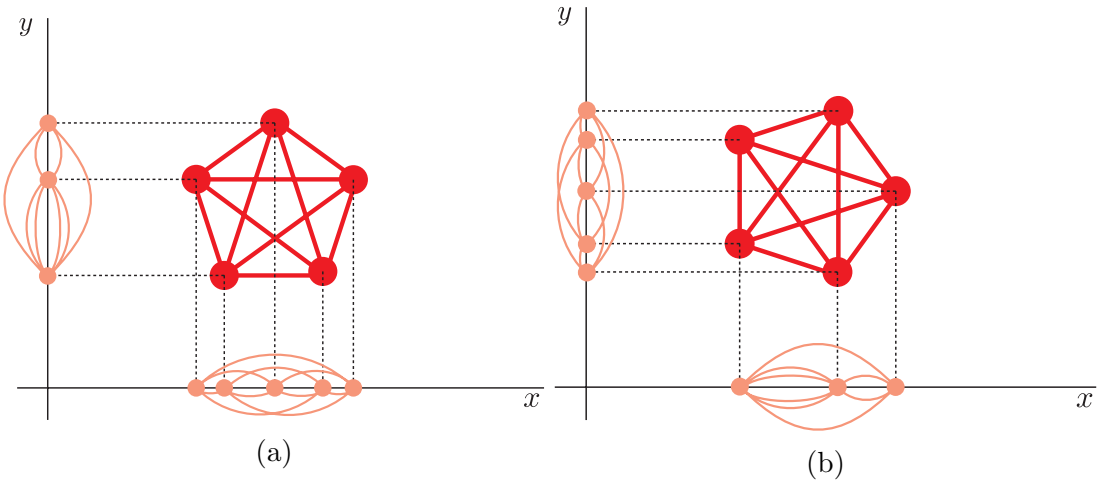


図 5.2: 部分空間における構造の不一致: 上記の (a)(b) は 2 次元空間における構造は同一だが，射影した各軸上での構造は異なる

と呼ぶ．上記の特徴量空間分割手法は，部分空間における幾何構造をも照合制約として加え，部分空間でも構造不変性が満たされているとすることで，多次元空間における幾何構造の自由度を下げ，“過剰な不変性” を抑制していることになる．

また本論文では第 3章において，音声に加わる線形変換性歪みについて，ケプストラム空間において幾何学的に強い回転性を表出することに言及してきた．最適なブロックサイズのもとにおいて，上記の特徴量分割は話者性に起因する回転に対してのみ適切な制約を与える効果を持つ．

さらに本研究における音声合成の枠組みにおいて，特徴量分割を行う事は探索の高速化に直結する．一般に n 次元空間で線形探索を行う場合，各次元における空間解像度[1]を r は $O(r^n)$ となる．一方，特徴量分割を行った場合，各部分空間を独立に探索することになるため，その計算量は $O(\frac{n}{m}r^m)$ となり，$m < n$ のため大幅に計算量を改善することができる．よって本研究でも特徴量空間分割の手法を採用し，各部分空間ごとに構造推定および探索を行う．

5.3　構造の照合と正規化

5.3.1　音声認識における構造間差異の表現

音声合成を探索問題として定式化した場合，模倣の成立・不成立を表す評価関数を定義しなければならない．この際，模倣すべき語の構造的表象と，現在の探索点群と初期条件の事象群が張る構造との差異を定量的に表現する必要がある．構造的表象を用いた音声認識において，既にこの指標が定義されている．以下これについて述べる．

二つの M 角形幾何構造を考える．このとき M 個の頂点 $(P_1, \ldots, P_M, Q_1, \ldots, Q_M)$ について二つの幾何構造を回転（$\boldsymbol{A}$）とシフト（$\boldsymbol{b}$）で近づけ，対応する頂点間距離の和（$\sum_{i=1}^{M} \overline{P_iQ_i}^2$）の最小値を求めることで二つの幾何構造の差異を定義する．二つの幾何構造を N 次元ユークリッド空間内で考えれば，両構造の重心を O として以下の式で導出される．

$$\sum_{i=1}^{M}(\overline{OP_i}^2 + \overline{OQ_i}^2) - 2\sum_{i=1}^{N}\sqrt{\alpha_i} \tag{5.1}$$

[1] ここでは探索時に調べる各次元あたりの点の数を表す．

ここで α_i は N 次正方行列 S^tTT^tS の i 番目の固有値を表す．S, T は二つの幾何構造の頂点構造を保存した行列であり $S = (\overrightarrow{OP_1}, \ldots, \overrightarrow{OP_M}), T = (\overrightarrow{OQ_1}, \ldots, \overrightarrow{OQ_M})$ となる．しかし構造的表象は空間を歪ませる「**分布間距離**」によって構成され，ユークリッド空間には存在しない．よって式 (5.1) を直接用いることはできない．一方，分布間距離としてバタチャリヤ距離の平方根を用いた場合，回転（$\boldsymbol{A}$）およびシフト（$\boldsymbol{b}$）後の $\sum |\theta_i|$ ($\theta_i = \angle P_iOQ_i$) が十分小さければ，以下の近似式が成立することが示されている [50].

$$\sqrt{\frac{1}{M^2}\sum_{i<j}(\overline{P_iP_j} - \overline{Q_iQ_j})^2} \simeq \sqrt{\frac{1}{M}\sum_{i}(\overline{OP_i} - \overline{OQ_i})^2} \tag{5.2}$$

$$\simeq \sqrt{\frac{1}{M}\sum_{i}\overline{P_iQ_i}^2} \tag{5.3}$$

このとき式 (5.2) の左辺は幾何構造を表す $M \times M$ の距離行列[2]の上三角部分によって構成されるベクトル（以下これを「構造ベクトル」と定義する）間のユークリッド距離に対応する．すなわち構造間の差異は距離行列のみで表現でき，距離行列によって導かれる構造ベクトルのユークリッド距離によって定義可能である．以下ではこれを二構造間の「**構造歪み**」と呼ぶ.

5.3.2　構造ベクトルの類似度尺度

本研究では，音声認識において議論されている「**構造歪み**」に加えて新たな指標も検討する．今二つの構造を表す構造ベクトルをそれぞれ $\boldsymbol{a}$，$\boldsymbol{b}$ とする．ここで二つの構造ベクトルの “**向き**” をその類似度評価に利用する．構造ベクトルの方向は，調音努力[3]に関係なく語形を表現していると考えられ，2 つの構造ベクトルのなす角を測る事で調音努力を考慮せずに語形の類似度を評価できると考えられる．そこで類似度指標を s として

$$s = \frac{\boldsymbol{a}\cdot\boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|} \quad (0 \le s \le 1) \tag{5.4}$$

と定義する．ただし $\boldsymbol{a}\cdot\boldsymbol{b}$ はベクトルの内積を，$|a|$ はベクトルのノルムを表す．以降この指標を二構造間の「**構造類似度**」と呼ぶ.

5.3.3　部分構造の歪み最小化

構造的表象は，言語的特徴と非言語的特徴を原理的に分離できる（図 4.4）．一方，パラ言語的特徴については一部構造的表象の中に内包される．即ちこれは発話スタイルなどのパラ言語情報によって構造が変形することを意味し，[51] では調音努力の差が構造のサイズを変化させる様子を示している．上記の探索の枠組みで音声模倣をモデル化する際，すでに出力された過去の状態がなす構造の影響を考慮しなければならない．これは幼児の音声模倣の例において，母親と幼児の調音努力が異なっていることを意味している．このときすでに出力された過去の状態が張る構造（以下これを「部分構造」と呼ぶ）に母子間で差異が生じている．

初期条件における部分構造と制約条件におけるそれとの差異を最小にする操作を考える．ここでは模倣対象となる構造のサイズのみを変化させて，初期条件の音響事象の張る構造との差異を最小にする．このとき構造ベクトルにおいて初期条件の音響事象による構造ベクトルの模倣対象

[2] ここでは各頂点間の距離を要素とする行列を表す.
[3] ここでは個々の音韻をよりはっきり区別しようとすることを調音努力が大きいと定義する.

表 5.1: 音響分析条件（5.4.1節）

サンプリング条件	16 bit / 16 kHz
フレーム窓	Hamming window
フレーム長	25 ms
シフト長	5 ms
ケプストラムパラメータ	Mel cepstrum (1–12) [$\alpha = 0.42$]

に対する正射影を考えればよい．即ち既に音が生成された部分に関して，模倣対象の部分構造ベクトルを $\boldsymbol{a}$，初期条件の張る構造ベクトルを $\boldsymbol{b}$ として

$$\boldsymbol{a'} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}|^2}\boldsymbol{a} \tag{5.5}$$

なる操作を行う．これは部分構造について二人の話者の構造サイズを正規化していることに相当する．また部分構造の類似度が高い場合は，$s \simeq 1$ として式 (5.4) を代入すれば，$\boldsymbol{a'} = \frac{|\boldsymbol{b}|}{|\boldsymbol{a}|}\boldsymbol{a}$ となり，二人の話者の構造サイズを統一していることになる．

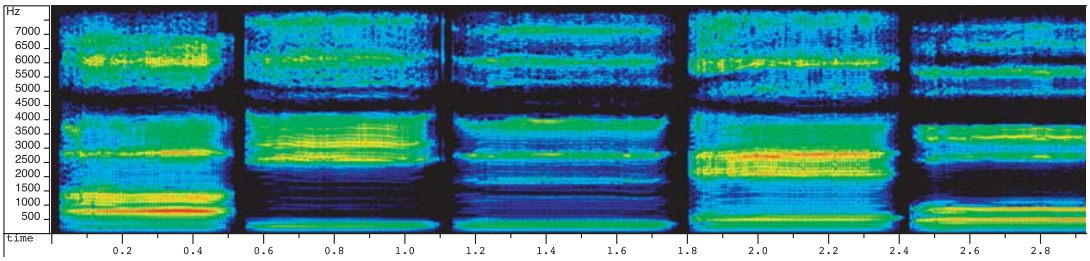
5.4　メルケプストラムを用いた実験

5.4.1　実験方法

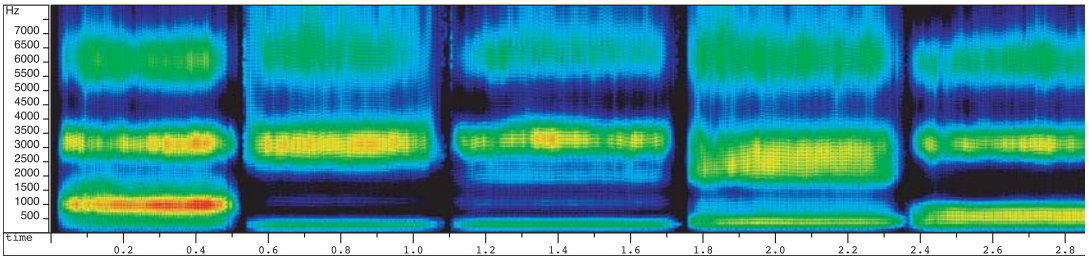
提案する枠組みによって音声合成が可能であることを確認するため，日本語 5 母音の孤立発声（/a/, /i/, /u/, /e/, /o/）を用いて実験を行った．成人男女各 2 名（それぞれ話者 M1，M2，F1，F2 とする）の日本語 5 母音の発声を収録した．これらの発声について表 5.1に示す音響分析条件でケプストラム分析を行った．同時にそれぞれの発声のピッチ，パワー，継続長についても分析した．これらのパラメータを用いて以下に示す手順で構造抽出と解探索を行った．

1. 各母音発声を平均ベクトルと対角共分散行列で表される多次元ガウス分布として，最尤推定を行い各パラメータをモデル化する．
2. ブロックサイズを 1 とし，分割された各特徴量空間で分布間距離を求めて構造を抽出する．
3. 5 母音のうち n 個を既知，残りを探索対象とする．本実験では分散項は既知とし，平均ベクトルを探索範囲で変化させる．この際 5.3.3で述べた部分構造歪みを最小化しておく．既知の母音数 n を変化させて実験を行う．
4. 線形探索を行い，構造類似度を最大化するように解を決定する．
5. 得られたケプストラムの解と事前に抽出したピッチ，パワー，継続長から音声を合成する．

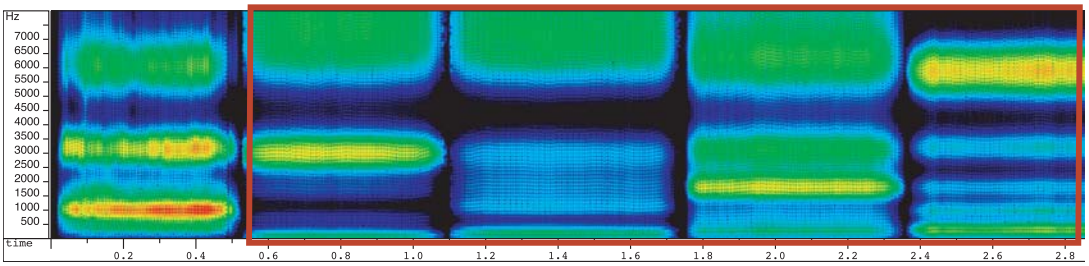
上記の実験について，構造提供話者を収録した上記の 4 名，探索における初期条件を提供する話者を上記のうちの男女 1 名ずつ（M1，F1）とした．なお構造提供話者と初期条件提供話者が同一の場合は異なる発声を用いた．幼児の音声模倣の枠組みでいえば，構造提供話者は母親に，初期条件提供話者は幼児に対応することになる．状態生成は身体性を与える事に相当するため，構造提供話者と初期条件提供話者が等しい場合は，自身の過去の発声に対して語形のみで復唱することに相当する．一方構造提供話者と初期条件提供話者が異なる場合はまさに音声模倣のモデルとして解釈可能である．また探索における空間解像度について，1 次元あたり，既知母音数が 4 の場合 200 点，3 の場合 100，2 の場合 50，1 の場合 25 とした．



(a):原音声



(b):分析再合成音



(c):提案手法の合成音

図 5.3: 合成実験の結果; (c) において枠囲いされた部分が探索による音声

5.4.2　実験結果

実験によって得られた合成音声の一例を図 5.3に示す．図 5.3には本実験に得られた音声と，比較のため初期条件提供話者の原音声，および再合成音声のスペクトログラムを示している．発声内容は孤立発声の連結（/a/ /i/ /u/ /e/ /o/）であり，(a) が原音声，(b) が分析再合成音，(c) が得られた音声である．

(c) において，先頭の/a/を既知の状態として与えており，残りの 4 母音は探索によって得られたものである．(a) と (b) とを比べると再合成によってスペクトルが平滑化されている事がわかる．一方 (b)(c) を比べると，およそ概形を再現していることがわかる．ここでこの実験では (a) や (b) の語形の情報は用いておらず，他者の語形情報といくつかの自身の絶対量をもとにケプストラム空間を探索し，合成すべき音声を推定していることに注意する．

5.4.3　主観評価実験

合成実験で得られた音声について，主観評価実験を行った．今回は孤立発声の実験結果について聴取実験を行った．提案する枠組みで評価すべき項目は以下の二つであると考えられる．

表 5.2: 聴取実験の実験条件

(a): 音韻性評価（1）の為の聴取実験条件

被験者	日本人成人男性 4 名
聴取実験サンプル	提案手法による合成音 240 サンプル
	分析再合成音による比較音声 120 サンプル
実験法	ヘッドホン聴取による書き取りテスト

(b): 話者性評価（2）の為の聴取実験条件

被験者	日本人成人男性 4 名
聴取実験サンプル	提案手法による合成音 24 サンプル
	対応する模倣対象の分析再合成音 24 サンプル
	対応する合成対象の分析再合成音 24 サンプル
実験法	上記 24 組のサンプルによる ABX 法

1. 本手法で得られた音声について，合成すべき音韻が正しく生成されたかどうか.
2. 本手法で得られた音声について，その話者性が構造抽出話者のものではなく身体性を与えた話者のものとして知覚されるかどうか.

1. について評価するため，身体性を与えた話者の分析再合成音と探索によって得られた母音によって構成される 5 モーラ単語を作成し書き取りテストによる了解度評価を行った．この際「5 モーラ中に 5 母音が一度ずつ含まれる」という単語知識による影響を除くため，分析再合成音によって 5 母音の重複を許して作成した 5 モーラ単語を実験セットに加えた．さらに聴取実験に出現する単語が「日本語 5 母音による 5 モーラ単語で出現の重複は許す」と被験者に予め伝えた.

2. について評価するために，制約条件提供話者の分析再合成音および初期条件提供話者の分析再合成音を比較対象とした ABX 法により合成音声の個人性を評価した．ABX 法とははじめに二つの刺激 A, B を呈示し，次に提示した刺激 X が A および B のいずれに近いかを判断させる手法である．この時聴取者の解答が用意した正解と一致した割合を調べることで評価を行う．既知母音数と制約条件提供話者，初期条件提供話者の異なる音声サンプルを 24 通り作成し，それぞれ同一単語で ABX 法を実施した．1.，2. それぞれの聴取実験の条件を表 5.2に示す.

上記聴取実験の結果を表 5.3に示す．表 5.3 (a) は書き取りテストの結果を示したものである．ここで正解とは 4 人の被験者のうち 3 人以上の被験者の解答と，模倣対象の単語とが一致した場合を表している．分析再合成音の結果は単語知識の解消の為に用いた，重複出現を許す 5 モーラ単語の結果である．これにより既知母音数を 1 として合成した場合でも 15 %ほどが単語完全一致で再現可能である事がわかる．さらに 1 モーラの誤りを許容して正解に加えた場合，大幅に正答率が上昇し，既知母音数が 1 の場合でも約 4 割の単語が正しく合成できた事になる.

表 5.3 (b) は話者性評価実験の結果を示したものである．表 5.3 (b) は被験者 4 名のうち 3 人以上，または全員一致で合成対象話者に近いと評価した数を示している．これにより既知母音の数に依存せず，模倣対象の話者性を除去し，合成対象の話者性を再現できている事がわかる.

5.4.4　考察

合成実験，および主観評価実験の結果について考察する．表 5.3 (a) より，合成された単語の了解度としては既知母音数が 4 のとき 48 %と分析再合成の値 (67 %) を下回った．おもに了解度が

表 5.3: 聴取実験の結果（5.4.3節）

(a):書き取りテストの結果

	総単語数	正答率（5 モーラが全て一致）	正答率（1 モーラの誤りを許容）
分析再合成音	120	0.67	0.83
既知母音数 1	40	0.15	0.38
既知母音数 2	80	0.20	0.44
既知母音数 3	80	0.43	0.81
既知母音数 4	40	0.48	0.95

(b):話者性評価の結果

	全員一致	3 人以上一致
全体	18/24	24/24
既知母音数 1	5/6	6/6
既知母音数 2	4/6	6/6
既知母音数 3	4/6	6/6
既知母音数 4	5/6	6/6

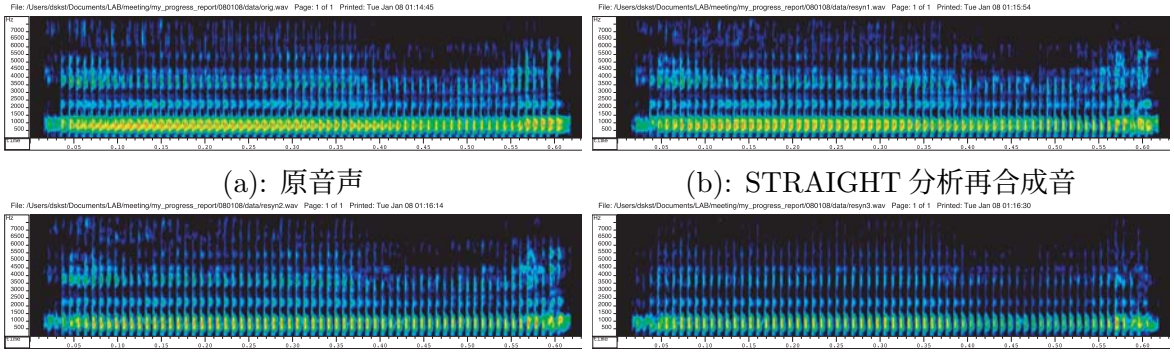
低くなる原因として以下が考えられる．

1. 解探索における解像度の問題
2. 解候補の評価における局所解の影響
3. ケプストラムの次元数の問題

1. は各次元の全探索を行う際の量子化の解像度の問題である．この解像度を高く設定する事で正しい解が得られると考えられるが，計算時間とのトレードオフとなる為，最適な値を検討する必要がある．2. は解候補の評価において本来異なる音に対して語形が類似していると判断してしまうことを意味している．式 (5.4) によって得られる，構造類似度による評価は，調音努力の影響を考慮しない評価尺度である．そのためある程度の発話スタイルのばらつきを抑える事ができる反面，構造をケプストラム空間に定位させる絶対量が少ない場合，調音努力が大きく異なる単語を等しいと判断する可能性を持っている．そのため構造歪みの最小化基準も併せて考慮した評価基準を検討する必要がある．また図 5.3のスペクトルのように，今回は再合成において分散項を考慮していないため全体的にブザー音のような音声が出力されてしまっている．このことも了解度を下げた要因であると考えられる．

なお主観評価実験では/i/と/u/を取り違えた間違いが目立った．また音韻交代のような現象も多く見られた．これらは定性的ながら人間が聴取において全体的特徴を利用していることを示唆するものといえる．このような現象が孤立母音の連結単語においても観察されたことになる．

表 5.3 (b) により，話者性の影響がおよそ取り除かれていることがわかる．この結果より特徴量空間の分割が有効に機能していることが示唆される．一方，今回はブロックサイズを 1 としてケプストラム空間を 1 次元ずつの部分空間に分割したが，ケプストラムの各次元の相関を考慮し，ブロックサイズを拡大して最適な特徴量空間の分割を行う事で，話者性の影響を除去したうえで合成音声の品質を向上させる事ができると考えられる．次節以降これらの観点をふまえ，より高品質な分析合成系である STRAIGHT[35] に基づいて実験を行い，さらなる考察を行う．



(a): 原音声　(b): STRAIGHT 分析再合成音

(c): ケプストラム分析後の再合成音　(d): ケプストラムの分散項を 0 とした時

図 5.4: STRAIGHT 分析による再合成音のスペクトル

5.5　STRAIGHT ケプストラムを用いた実験

5.5.1　STRAIGHT ケプストラム

STRAIGHT は河原らによって提案されている高品質な分析合成の枠組みである [35]．特に STRAIGHT 分析によって得られる時間周波数構造である STRAIGHT スペクトルは従来のスペクトル情報から音源に起因する微細構造を取り除いた特徴である．このため STRAIGHT スペクトルは特に聴取実験の刺激を作成する際のスペクトルの操作に対して音質の劣化が少なく，非常に優れた分析合成系であると言われている．北村や齋藤らは STRAIGHT スペクトルをもとにケプストラムを導出し，ケプストラムドメインでの操作を行った後にスペクトルを再度導出する事で，高品質の聴取実験の刺激音を作成している [52, 53]．また HMM に基づく音声合成にも用いられ [54]，音声合成システムの共通評価会である Blizzard Challenge においてもその有効性を示している [55, 56].

図 5.4は男性話者の母音/a/の発声を STRAIGHT の枠組みで再合成した場合のスペクトログラムを示している．(a) は原音声，(b) は STRAIGHT 分析を行った後 STRAIGHT スペクトルと音源情報から再合成したものである．一方 (c) は STRAIGHT 分析で得られた STRAIGHT スペクトルを IFFT によって STRAIGHT ケプストラムに変換，40 次元までのリフタリングを施した後スペクトルを再度求めて再合成した音声である．さらに (d) は 40 次元の STRAIGHT ケプストラムの時系列情報を ML 推定でガウス分布として状態化，その平均値で全てのケプストラムを置き換えた後，(c) と同様 STRAIGHT の合成系で求めた音声である．これらを見るとケプストラム分析を行った後の音声もほとんど問題なく原音声を再現している事がわかる．聴感上の差異もほとんどみられない．なお (d) におけるスペクトルのばらつきは音源によるものである．

5.5.2　ブロックサイズおよび照合条件を変化させた実験

前節におけるメルケプストラムを用いた実験では，ブロックサイズおよび構造の照合条件を固定して実験を行っていた．一方前節の考察の結果から，これらの条件も少なからず探索速度や音質に影響していると考えられる．そこで STRAIGHT ケプストラムを用いた上で，特徴量空間分割におけるブロックサイズおよび構造の照合条件を変化させて実験を行った．

実験の枠組みを述べる．前節における実験ではメルケプストラムを用いたが，STRAIGHT 分析の結果得られた対数パワースペクトルに対して IFFT を行うことで 40 次の STRAIGHT ケプス

表 5.4: STRAIGHT ケプストラムによる合成音の聴取実験結果; 括弧内はブロックサイズを表す

	総単語数	構造類似度 (1)	構造歪み (1)	構造歪み (2)	構造歪み (4)
分析再合成音	120	1.00			
既知母音数 1	40	0.05	0.05	–	–
既知母音数 2	80	0.18	0.20	–	–
既知母音数 3	80	0.49	0.45	0.40	–
既知母音数 4	40	0.73	0.83	0.73	0.63

トラムを得た．この時フレームのシフト長は 1 ms となっている．同時に発声のピッチ，パワー，継続長も STRAIGHT の分析をもとに得た．

まず照合条件による違いを調べるため，ブロックサイズが 1 のもとで二つの照合条件によって合成を行った．一つは前節で用いた構造類似度の最大化，もうひとつは音声認識で用いられている構造歪みの最小化である．さらにブロックサイズによる違いを調べるため，構造歪み最小化の照合条件のもとで，ブロックサイズを 1, 2, 4 と変化させて合成を行った．この際部分空間の各次元に重複はないものとした．探索時間を考慮し，ブロックサイズ 2 の場合，既知母音数 4 で解像度 200，3 で解像度 50 点とした．またブロックサイズ 4 のとき，既知母音数 4 で解像度を 50 とした．これらの合成音声について，前節とは異なる成人男性 5 名による音韻性評価の聴取実験を実施した．聴取実験の条件は上記の条件変化に応じてサンプル数が変わる以外は前節と同様である．

照合条件を変化させた場合の実験結果を表 5.4に示す．なお正答率は 5 名の聴取者のうち 3 名以上が 5 モーラ全て一致させた場合を正答として算出している．数値が存在しない欄については探索時間が膨大となるため実験を行っていない箇所である．まず STRAIGHT 分析を用いる事によって分析再合成音の了解度は 100%となった．すなわち再合成系の音質に大幅な改善がみられたことがわかる．一方ブロックサイズを 1 とし，照合条件を変化させた場合，既知母音数が少ない場合には了解度に差はみられなかったが，既知母音数が 4 の場合，構造歪み最小化を基準とした方が，高い了解度の音声が合成されたことになる．また探索に要する時間については，構造歪み最小化基準の方が 5〜6 倍の速度改善が確認できた．これは構造歪み最小化基準の場合，構造ベクトルのユークリッド距離はベクトルの各次元の差の二乗を足し合わせていくことで算出する．このため構造歪みの算出過程の途中で暫定の最小値を上回った場合に，枝刈りを実行することで探索空間を制限できている点に起因すると考えられる．また，ブロックサイズを増加させた場合，今回の結果からは了解度の向上は確認されなかった．これは探索問題の場合，探索すべき母音数の増加およびブロック数の増加によって計算量が増大するため，解像度を低くせざるを得ないためと考えられる．

5.5.3　主観評価実験のエラー分析

以下，聴取実験の結果について，話者の組み合わせや推定する母音の種類，および推定母音以外で誤っているケースなどに場合わけをして，その分析を行う．

まず構造抽出話者と初期条件提供話者の違いによる影響について，構造抽出話者と初期条件提供話者が同一の場合，および異なる場合に分けた時の聴取実験の結果を表 5.5に示す．表 5.5(a) は構造抽出話者と初期条件提供話者が同じ（ただし発話は異なる）場合，表 5.5(b) は構造抽出話者と初期条件提供話者が異なる場合である．表 5.5によると同一話者間での合成の方が別話者間での合成よりも了解度が高い傾向が見られた．一方ブロックサイズが大きい場合には両者の差は小さ

表 5.5: 同一話者間，別話者間に分類した場合の聴取実験結果; (a): 同一話者間での合成音の聴取実験結果（復唱）; (b): 別話者間での合成音の聴取実験結果（模倣）;

(a): 同一話者間での合成音の聴取実験結果

	総単語数	構造類似度 (1)	構造歪み (1)	構造歪み (2)	構造歪み (4)
既知母音数 1	10	0.10	0.10	–	–
既知母音数 2	20	0.40	0.55	–	–
既知母音数 3	20	0.95	0.85	0.95	–
既知母音数 4	10	1.00	1.00	1.00	0.70

(b): 別話者間での合成音の聴取実験結果

	総単語数	構造類似度 (1)	構造歪み (1)	構造歪み (2)	構造歪み (4)
既知母音数 1	30	0.00	0.03	–	–
既知母音数 2	60	0.10	0.08	–	–
既知母音数 3	60	0.33	0.32	0.22	–
既知母音数 4	30	0.63	0.77	0.63	0.60

表 5.6: 同一性別間，異性別間に分類した聴取実験結果; (a): 同一性別間での合成音の聴取実験結果; (b): 異性別間での合成音の聴取実験結果;

(a): 同一性別間での合成音の聴取実験結果

	総単語数	構造類似度 (1)	構造歪み (1)	構造歪み (2)	構造歪み (4)
既知母音数 1	20	0.10	0.10	–	–
既知母音数 2	40	0.30	0.38	–	–
既知母音数 3	40	0.73	0.70	0.70	–
既知母音数 4	20	0.95	1.00	0.95	0.70

(b): 異性別間での合成音の聴取実験結果

	総単語数	構造類似度 (1)	構造歪み (1)	構造歪み (2)	構造歪み (4)
既知母音数 1	20	0.00	0.00	–	–
既知母音数 2	40	0.05	0.03	–	–
既知母音数 3	40	0.25	0.20	0.10	–
既知母音数 4	20	0.50	0.65	0.50	0.55

くなっている．これはブロックサイズが大きくなる事によって解像度が低くなる為に同一話者間で探索音の合成品質が低下している点が一つの原因であると考えられる．しかし異なる話者間ではそれほど了解度が低下しておらず，多次元空間における構造の不変性が有効に働いている可能性を示唆している．

加えて構造抽出話者と初期条件提供話者が同一の性別である場合，および異性別である場合に分類した結果を表 5.6に示す．表 5.6によれば，話者の違いと同様，同一性別間の方が異性別間よりも了解度が高い傾向が見られている．一方異性別間でブロックサイズが 2 の場合およびブロックサイズが 4 の場合を比較すると，解像度が低くなっているにも関わらず了解度に向上傾向がみられることが分かる．このことから 5.2.3で述べた特徴量分割について，その効果および適用範囲，最適なブロックサイズなどを今後検討していく必要がある．

表 5.7: 推定母音によって分類した聴取実験結果; 既知母音数は 4

	総単語数	構造類似度 (1)	構造歪み (1)	構造歪み (2)	構造歪み (4)
/a/	8	0.63	0.75	0.63	0.75
/i/	8	0.88	0.75	0.50	0.50
/u/	8	0.88	1.00	0.88	1.00
/e/	8	0.63	1.00	1.00	0.88
/o/	8	0.63	0.63	0.63	0.50

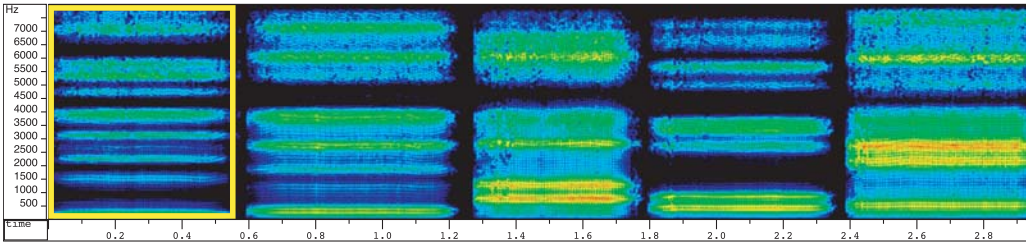
加えて推定する母音の種類による違いについて，既知母音数が 4 の場合を対象としてそれぞれの母音毎に分類した了解度を表 5.7に示す．表 5.7によれば，構造類似度の最大化基準および構造歪みの最小化基準のいずれの場合についても，/u/ または /e/ が合成対象の場合が高い了解度を示している事がわかる．これはブロックサイズが大きい場合も同様である．すなわち/u/ および /e/ については調音可能な音響空間上で他の母音が決まっていれば，その母音との関係だけから導出可能である事を示している．一方で /i/ および /o/ に関してはブロックサイズの増大に伴って了解度が低下する傾向が見られた．

また聴取実験において推定音以外の母音を誤った例について考える．ブロックサイズが 2 以上の場合にはそのような例は観察されなかったが，ブロックサイズが 1 の場合に少ないながらそのような例がみられた．なお聴取実験の被験者のうち 3 名以上が推定母音以外を間違えた場合をそのような事例として数えている．図 5.5に被験者が推定母音以外を間違えた場合の合成音声および対応する分析再合成音のスペクトルを示す．(a) および (b) は構造類似度最大化基準による男性話者 M1 の/iuaoe/ の合成音声および分析再合成音，(c) および (d) は構造歪み最小化基準による女性話者 F1 の/uoiea/ の合成音声および分析再合成音である．また枠囲いされた部分が探索によって得られた音声である．(a) においては被験者が最初の 2 モーラを /uu/，/ue/，/ie/ と間違えるケースが目立った．これは 1 モーラ目の音が異なる事によって 2 モーラ目の聴取が影響を受けている事を示唆するものと言える．一方，(c) においては前半 3 モーラについて，被験者が /oau/，/oae/と間違えるケースであった．第 2 モーラの/o/ については再合成音であるが，第 1 モーラを/o/ と判断してしまうことで，後続モーラの聴取に影響が表れたことを示している．これらの結果は，人間が単語の聴取において語の関係性を考慮している事を示唆するものといえる．

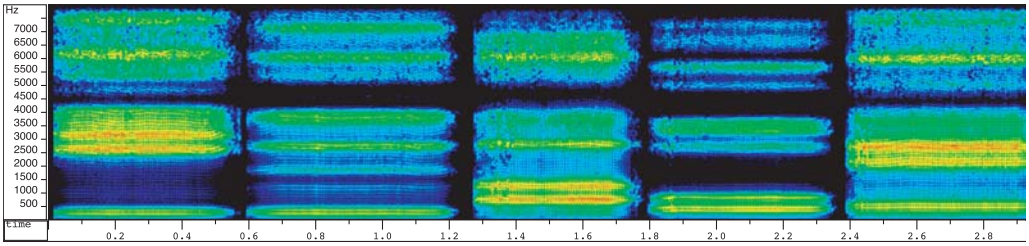
5.6　まとめ

本章では提案する構造的表象に基づく音声合成によって，日本語孤立 5 母音系列の合成を試みた．ケプストラム空間の探索問題としての定式化を通して，不変構造に身体性を付与するという枠組みによって合成音声を作成し，聴取実験の結果，最大で 83%の了解度を得る事ができた．一方で同一性別間，異性別間で了解度が異なる傾向がみられ，それぞれ最適なブロックサイズが異なることが示唆された．また推定音の種類によって了解度に差がみられた．加えて聴取実験結果の分析から，人間が聴取において語の関係性を考慮している事が示唆された．

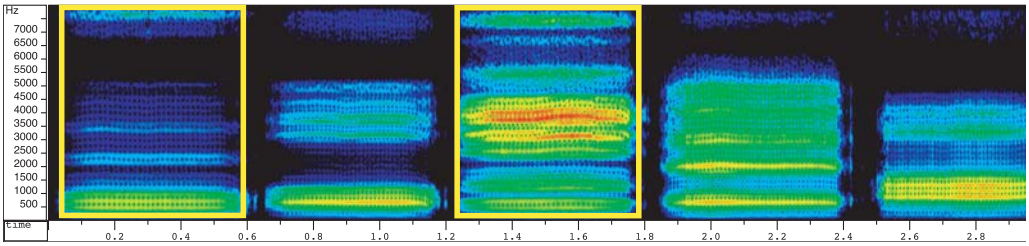
幼児が語ゲシュタルト全体を真似るという観点からみれば，今回の孤立の 5 母音を合成するというタスクは厳密には音声模倣のモデル化とはいえない．しかし本章での合成では，5 母音によって構築される音声の構造的表象を制約条件として母音の合成を実現しており，また語の体系を模倣するという観点から意義のある実験と考えられる．また近年音の体系に基づく外国語学習支援



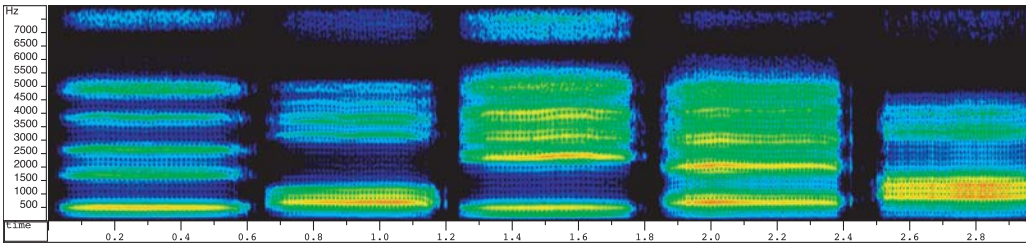
(a):合成音声（既知母音数 4，構造類似度基準による合成）



(b):分析再合成音 /iuaoe/（(a) に対応）



(c):合成音声（既知母音数 3，構造歪み基準による合成）



(d):分析再合成音/uoiea/（(c) に対応）

図 5.5: 被験者が推定音以外を誤った音声; 枠囲いされた部分が探索による音声

という観点からの CALL システムが提案されている [57]．これらのシステムにおける学習者へのフィードバック技術として本研究における合成系が応用可能であると考えられる．

第6章

音声の構造的表象からの連続５母音系列の合成

6.1　はじめに

前章において，構造的表象に基づく音声合成技術の基礎的検討として，孤立 5 母音系列に関して行った実験について報告した．本章では合成対象をさらに拡張し，連続的に発声された日本語 5 母音系列の合成を考える．これはまさしく語ゲシュタルトに基づく音声模倣のモデルとして考えられる音声合成である．この際，調音結合を伴い連続的に変化する音声から，分布としての時間幅を持った音響事象時系列を得なければならない．すでに構造的表象に基づく音声認識では HMM の学習アルゴリズムに基づく分布状態系列の導出が行われている [58]．[58] では HMM の学習アルゴリズムとして変分ベイズ法 [59] を用いている．変分ベイズ法はベイズ推定の枠組みにおけるパラメータの事後分布を近似的に求める手法として提案されている．ベイズ推定では推定におけるあらゆるパラメータを確率変数として取り扱い，それらに関する期待値をとり，周辺化する事で，データ量が少ない場合でも頑健な推定を実現する事ができる．一方で全てのパラメータについて最適な事後分布を求めるためその状態推定には時間を要する．加えて音響事象数がある程度大きい場合（例えば 5 母音の連続発声に対して 25 状態など）には，Baum-Welch アルゴリズムによって状態推定を行い，分布の推定に MAP（最大事後確率）推定を用いた手法より若干優位であるだけに留まっている．そのため，本研究における音声合成の実験に関しては後者の手法を採用し，実験を行った．

6.2　HMM を用いた連続音声の構造化

6.2.1　Baum-Welch アルゴリズムを用いた HMM の学習

本章では音声の音響事象数を $N = 25$ として取り扱う．定性的には 1 母音あたりに 5 状態ほどが割り当てられていると考える事ができる．ただし調音結合の影響などがあるため，厳密に 5 状態とは言えない．今便宜上一次元のケプストラムの場合について考える．このとき，連続発声から HMM を用いて分布系列を推定する枠組みを図 6.1に示す．ここで，$X = \{x_1, x_2, \ldots, x_T\}$ は連続音声から求めたケプストラム系列，$\theta = \{a_i, \mu_i, S_i \mid i = 1, \ldots, N\}$ は HMM のパラメータであり，a_i は状態 i から $i+1$ への状態遷移確率，μ_i および S_i はそれぞれ状態 i の出力確率密度分布の平均と精度である．連続発声からケプストラム系列を求め，この一発声のみから HMM のパラメータを推定（学習）させる．各状態と音響事象が対応しているので，各状態の出力確率密度分布をもって音響事象の分布とする．このとき，本章冒頭で述べたとおり，少量のデータを用いた Baum-Welch アルゴリズムでの状態推定は頑健に行われない可能性がある．しかし今回は $N = 25$ に固定し，十分音響事象数が多い場合に，少ない誤差で状態推定が可能なものとする．ただし各状態の出力確率密度分布に関しては，データ量に起因する推定誤差が発生する可能性があるため，最大事後確率推定を用いた頑健な分布推定を行う．

6.2.2　音響事象分布の最大事後確率推定

最大事後確率（Maximum A Posteriori; MAP）推定は入力データ量が少ない時，ML 推定より頑健なパラメータ推定が行われる推定手法である．

ML 推定では，入力データ X が得られたとき，パラメータ θ を以下の式によって推定する．

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(X|\theta) \tag{6.1}$$

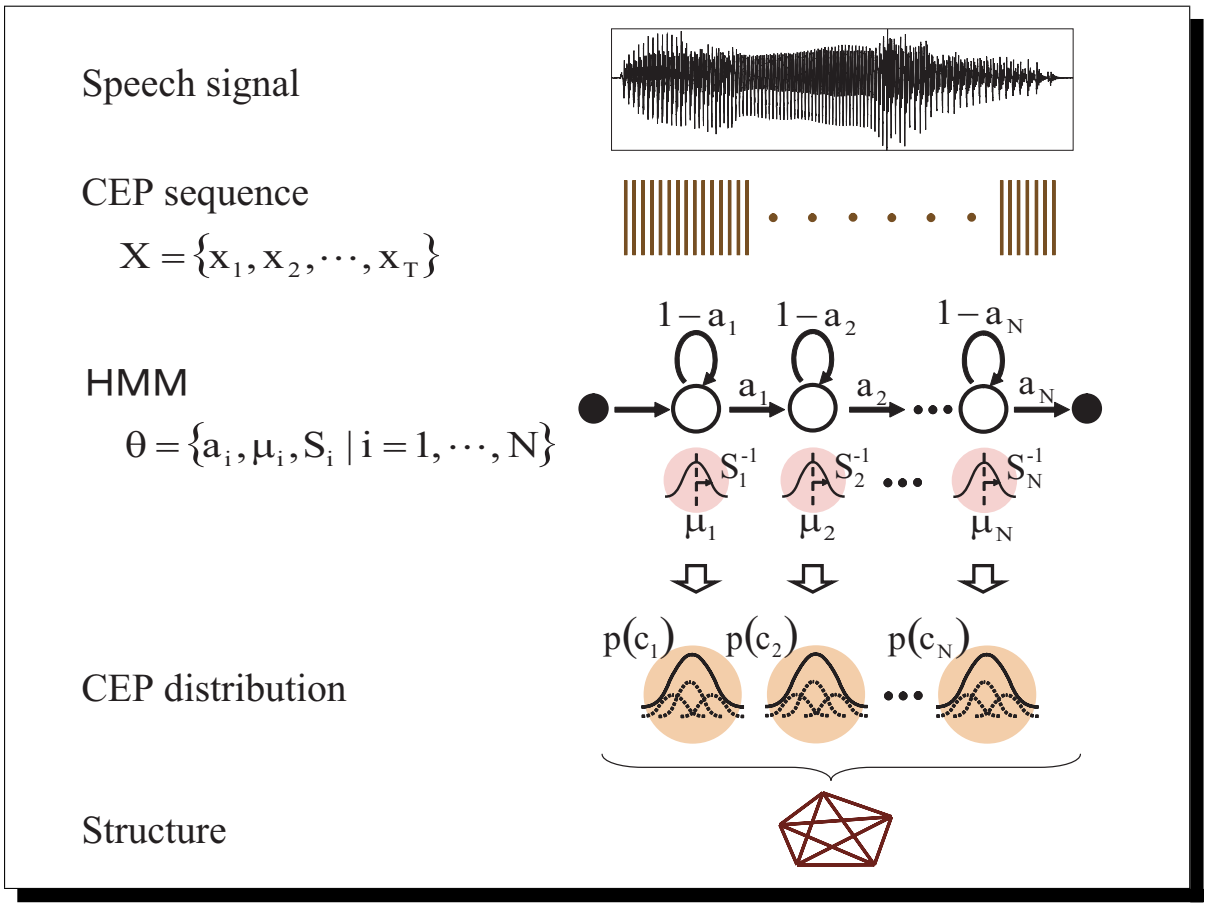


図 6.1: HMM を用いた連続音声の構造化の枠組み

これに対して，MAP 推定では θ もある確率密度分布に従って分布する確率変数とみなし，X を得た後パラメータが θ である事後確率を最大化する．即ち，

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(\theta|X) \tag{6.2}$$

によってパラメータ θ の推定を行なう．ここで，ベイズの定理により式 (6.2) は，次のように変形される．

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(X|\theta)P(\theta) \tag{6.3}$$

ここで，$P(\theta)$ は入力データ X が与えられていない時点での θ の事前分布である．すなわち事前知識を利用して，入力データを得た場合の事後確率が最大となるパラメータを推定する事で頑健な推定を行っている事になる．

構造的表象を用いた音声認識では，すでに [60] をもとに，MAP 推定に基づく頑健な構造化の手法を提案している [44, 58]．以下その手法を連続発声の場合について述べる．

分散共分散行列は全て対角とする．上記の HMM の状態とケプストラムベクトル系列のデータとの対応付けがとれているとする．このとき各状態番号毎に分布推定する話者とは別の話者のデータを用いて事前知識とする．これらを各状態番号に対応する発話区間（便宜上ここでは一発声とよぶ）毎にガウス分布化する（合計 M 個の発声）．MAP 推定に用いるパラメータは以下の通り

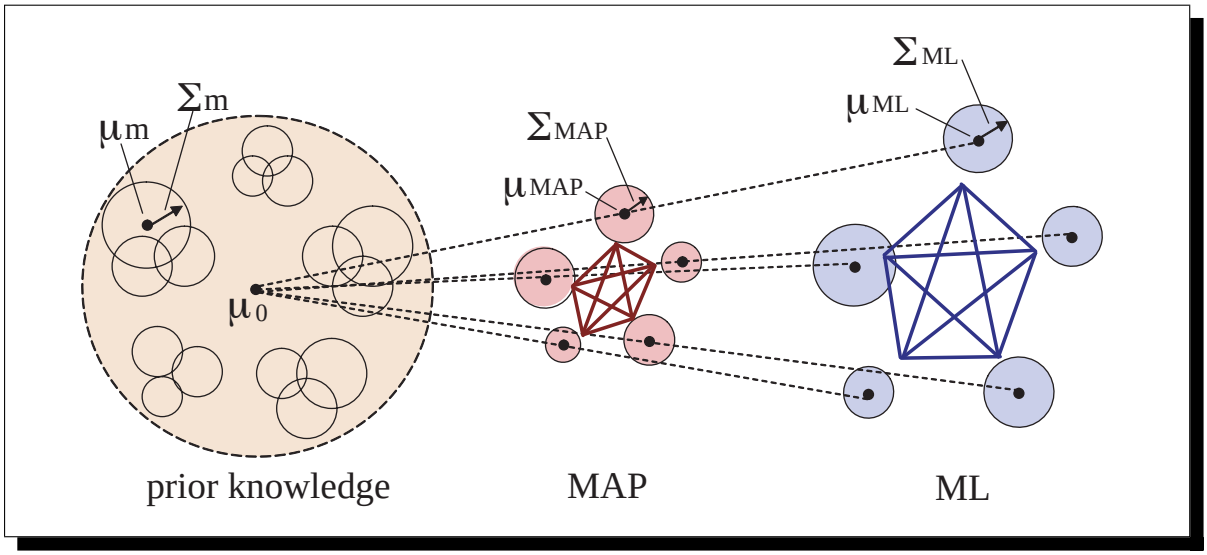


図 6.2: 音声事象分布の最大事後確率推定

である．

$$
\begin{aligned}
\mu_m &: m \text{ 番目の発声の平均ベクトル} \\
\Sigma_m &: m \text{ 番目の発声の対角共分散行列} \\
\mu_0 &: \{\mu_m\} \text{ の平均 } (= \tfrac{1}{M}\textstyle\sum_{m=1}^{M} \mu_m) \\
\Sigma_0 &: \{\Sigma_m\} \text{ の平均 } (= \tfrac{1}{M}\textstyle\sum_{m=1}^{M} \Sigma_m) \\
S_\mu &: \{\mu_m\} \text{ の対角共分散行列} \\
&\quad (= \tfrac{1}{M}\textstyle\sum_{m=1}^{M} (\mathrm{DIAG}(\mu_m - \mu_0))^2) \\
\Omega &: = \Sigma_0 S_\mu^{-1} \\
\mu_{ML} &: \text{入力発声の平均ベクトル（ML 推定）} \\
\Sigma_{ML} &: \text{入力発声の対角共分散行列（ML 推定）}
\end{aligned}
$$

ここで，DIAG(x) は，ベクトル x の要素を対角成分に並べた対角共分散行列である．これらを用いて，MAP 推定では一状態に対応する入力発声の分布を以下のように推定する．

$$\mu_{MAP} = \hat{\mu_0} \tag{6.4}$$

$$\Sigma_{MAP} = \hat{B}\hat{A}^{-1} \tag{6.5}$$

ここで，

$$\hat{\mu_0} = \Omega(\Omega + nE)^{-1}\mu_0 + n(\Omega + nE)^{-1}\mu_{ML} \tag{6.6}$$

$$\hat{B} = B + \frac{n}{2}\Sigma_{ML} + \frac{n}{2}\Omega(\mathrm{DIAG}(\mu_{ML} - \mu_0))^2(\Omega + nE)^{-1} \tag{6.7}$$

$$B = E \tag{6.8}$$

$$\hat{A} = A + \frac{n}{2}E \tag{6.9}$$

$$A = \Sigma_0^{-1} \tag{6.10}$$

表 6.1: 音響分析条件（6.3.1節）

サンプリング条件	16 bit / 16 kHz
フレーム窓	Hamming window
フレーム長	25 ms
シフト長	10 ms
ケプストラムパラメータ	Mel cepstrum (1 to 12) [$\alpha = 0.55$]

である．μ_{MAP} は μ_0 と μ_{ML} の内挿値をとり，n の増加につれて μ_{ML} に近づく．音響事象分布の MAP 推定の様子を図 6.2に示す.

6.3　実験

6.3.1　実験条件

連続音声において，孤立母音時と同様，探索に基づく構造からの音声合成の実験を行った．対象音声として，男性話者 2 名の連続発声/aiueo/ を用いた．これらの発声について表 6.1に示す音響分析条件でケプストラム分析を行った．同時にそれぞれの発声のピッチ，パワー，継続長についても分析した．これらのパラメータを用いて以下に示す手順で構造抽出と解探索を行った.

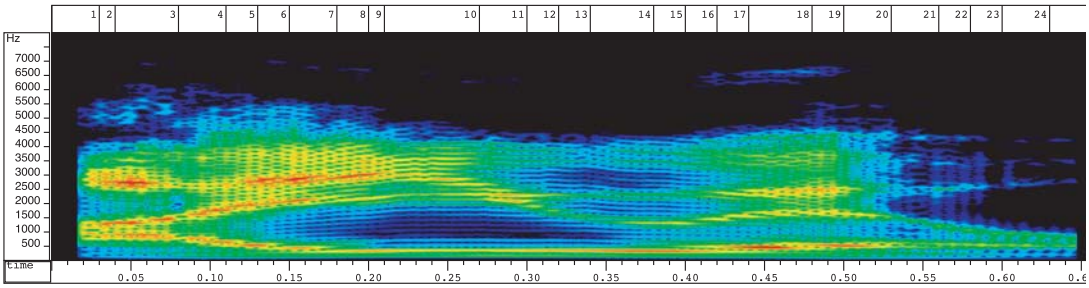
1. 本章で示した Baum-Welch による HMM の学習アルゴリズムと MAP 推定を利用し，ケプストラム系列から 25 状態の分布系列への変換を行う．なおこの際には表 6.1のケプストラムパラメータに加え ΔE および Δ メルケプストラムもあわせて用いている.
2. 分布間距離を求めて構造（25×25 の距離行列）を抽出する.
3. 本実験では分散項は既知とし，平均ベクトルを探索範囲で変化させる．25 状態のうち 21 状態を既知とし，前章と同様，部分構造歪みを最小化しておき残りの 4 状態を推定する.
4. 構造類似度の最大化を照合基準として解を決定する.
5. 得られたケプストラムの解と事前に抽出したピッチ，パワー，継続長から音声を合成する.

なおブロックサイズについては 1 として実験を行った.

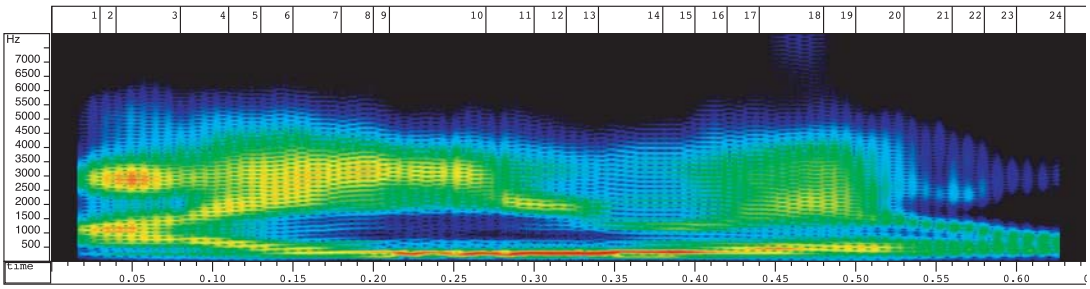
6.3.2　実験結果

実験によって得られた合成音声の一例を図 6.3に示す．図 6.3には本実験に得られた音声と，比較のため初期条件対象話者の原音声および分析再合成音のスペクトログラムを示している．(a) が原音声，(b) が分析再合成音，(c) が提案手法による合成音である.

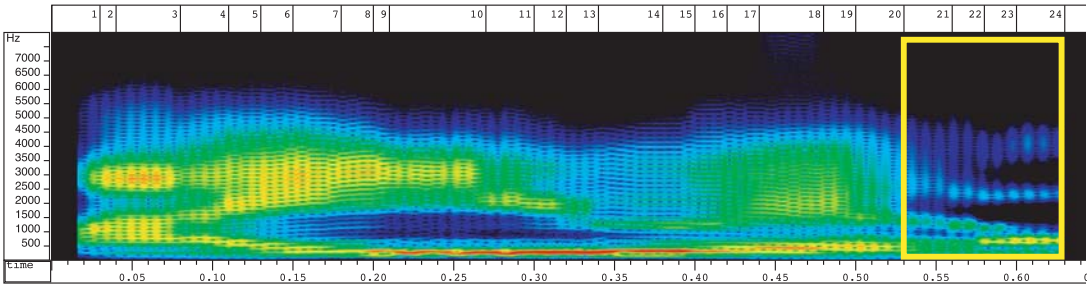
(c) においては先頭から 21 状態を既知として与えており，末尾の 4 状態は探索によって得られたものである．まず (a) と (b) についてスペクトルの概形を比べてみると，ケプストラム分析後の再合成によってスペクトルの山にいくらかの平滑化がなされていることがわかる．孤立母音の場合と異なるのは，語全体のスペクトル概形の変化は保存されているため，聴感上の差異があまり変わらない点である．一方 (b) と (c) について比較する．最後の 2 状態について若干差はあるものの，末尾の/o/にあたる部分がおよそ再現されていることがわかる．ここでこの実験では (a) の音声の語形情報は用いておらず，他者の語形情報といくつかの自身の絶対量をもとにケプストラム空間を探索し，合成すべき音声を推定していることに注意する.



(a):原音声



(b):分析再合成音



(c):提案手法の合成音

図 6.3: 合成実験の結果．(c) で枠囲いされた部分が探索による音声．

上記の 4 状態を推定する問題は定性的にはおよそ 1 母音を推定していることになる．このとき予備的な聴取実験では，既知の音響事象数を 4 とした場合の孤立母音の合成音よりも品質の向上がみられた．これは今回の実験が 21 状態からの強い制約条件を受けており，これによって音響空間への構造の定位が容易になったことが要因であると考えられる．この結果はあくまで予備的な聴取によるものであるが，幼児が個々の音に分割しているのではなく，語ゲシュタルト全体を真似ていることを示すものであると考えられる．

6.4　STRAIGHT ケプストラムを用いた実験

6.4.1　実験方法

5.5と同様，STRAIGHT の分析合成系に基づくケプストラムを用いて実験を行った．また 4 状態の推定位置および初期条件提供話者の性別の影響について調べるため，種々の条件における合成を行った．

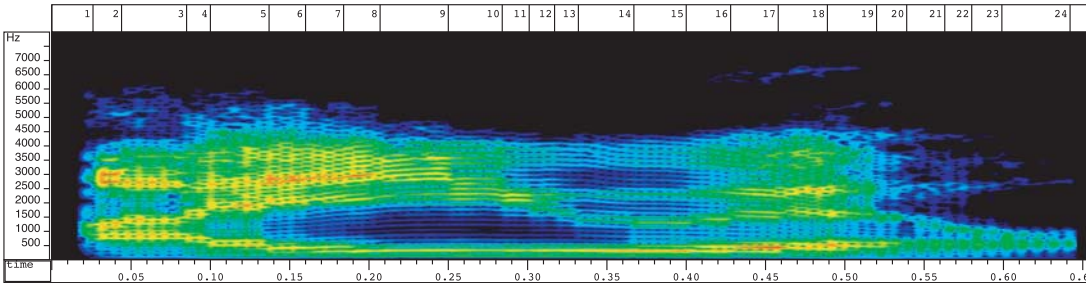
パラメータとして，STRAIGHT 分析によって得られた平滑化スペクトルに対してケプストラム分析を行い，40 次元のケプストラムを得た．その後，6.3.1と同様の手順で，構造抽出を行った．探索に関しては第 5章で議論した，構造歪みの最小化を基準として探索を行った．また構造抽出話者を成人男性 1 名，初期条件提供話者を前記の成人男性と異なる成人男女 1 名ずつとした．合成単語は日本語 5 母音の連続発声/aiueo/ である．

推定位置における影響について，成人男性間における合成結果を図 6.4，図 6.5に示す．図 6.4(a) には比較のため初期条件提供話者の分析再合成音声を示している．図 6.4および図 6.5の結果によれば，スペクトルの概形は推定位置に依存せず，ほぼ正しく再現されていることがわかる．このとき制約条件としては異なる成人男性話者の音声を用いていることに注意する．一方，予備的な聴取実験の結果，聴感上もこれらの合成音は違和感なく /aiueo/ と知覚された．またこれらの結果のうち，図 6.5(c) のような末尾部分の推定においては，あまり正確にスペクトルが再現されておらず，聴感上の違和感も認められた．これは，単語末においては，各話者毎の発話スタイルの違いが表出しやすく，その影響が構造に基づく音声合成に影響したものと考えられる．

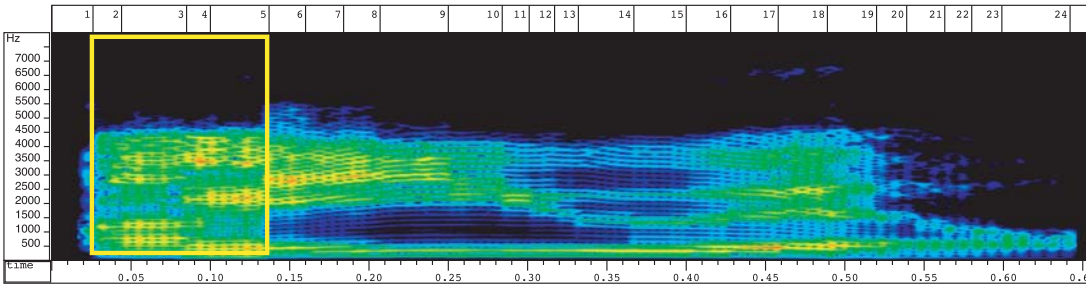
一方，初期条件提供話者の性別が異なる場合について，その結果を図 6.6に示す．図 6.6には参考のために初期条件提供話者の女性の分析再合成音，および構造提供話者の男性の分析再合成音を示している．図 6.6(a) および (c) は大きくスペクトルが異なる．一方 (c) の構造をもとに推定された (b) は，高域部分は異なるものの，(a) の低域部分をよく再現している．また聴感上も (b) は問題なく/aiueo/ と知覚された．このことから，性別に関わらず話者性をうまく消失し音声模倣が実現されている事が分かる．

6.5　まとめ

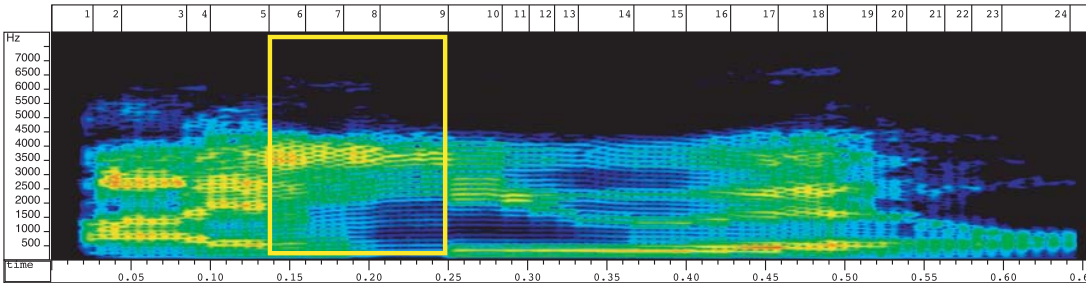
本章では，第 5章で行った孤立 5 母音系列の構造的表象からの音声合成を日本語 5 母音連続発声の場合に適用し実験を行った．その際，連続発声をより適切にモデル化するため，構造的表象による音声認識ですでに検討されている，HMM を用いた連続発声の構造化を本研究にも適用した．結果として，およそ 1 母音相当という小さなタスクであるが，自然性の高い音声を合成する事ができた．本章における実験はまさに音声模倣のモデル化ということができ，今後より多くの状態を構造的表象から生成することを検討していく必要がある．



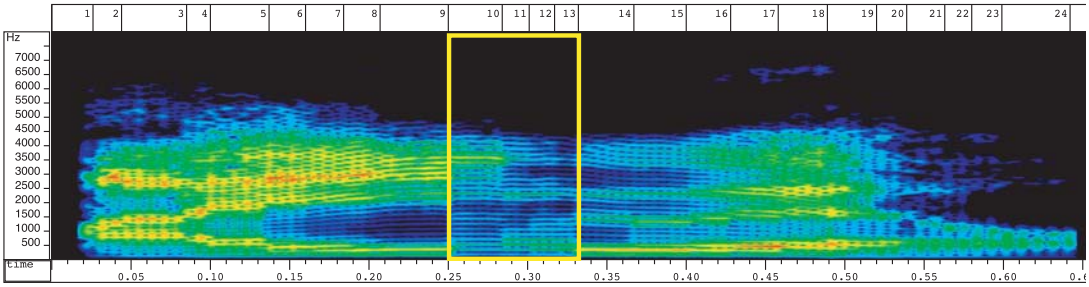
(a):分析再合成音



(b):提案手法の合成音（状態 2–5 が推定部分）

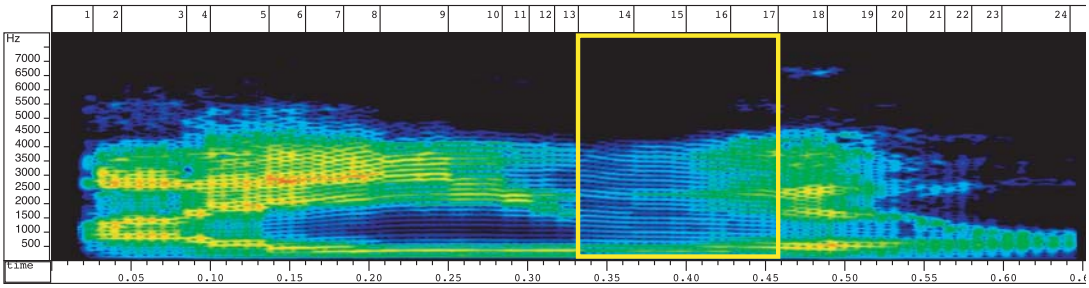


(c):提案手法の合成音（状態 6–9 が推定部分）

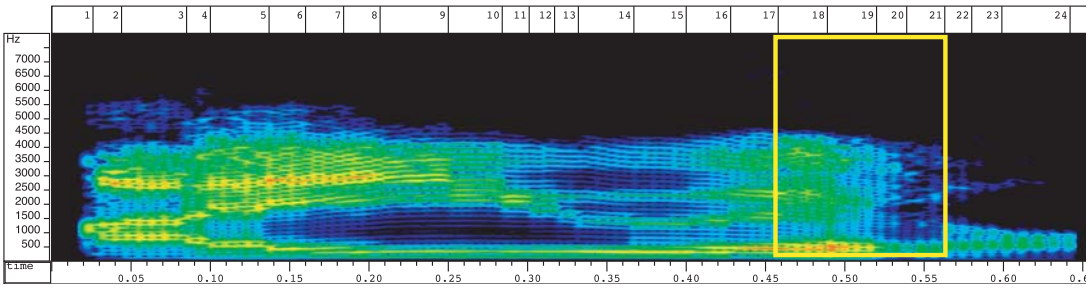


(d):提案手法の合成音（状態 10–13 が推定部分）

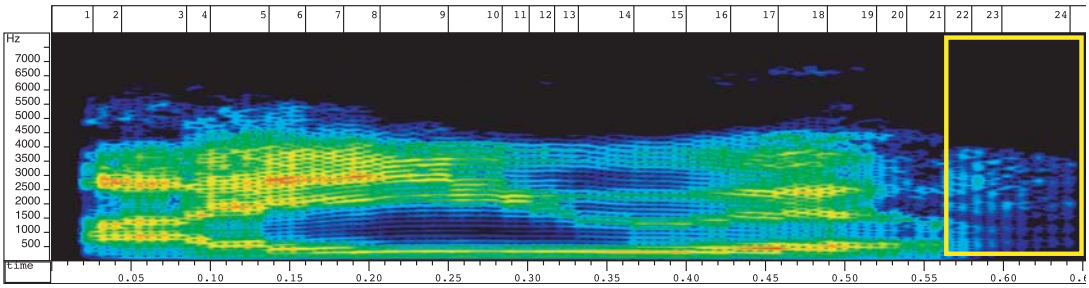
図 6.4: 異なる推定位置の合成結果 (1): 枠囲いされた部分が探索による音声.



(a):提案手法の合成音（状態 14–17 が推定部分）

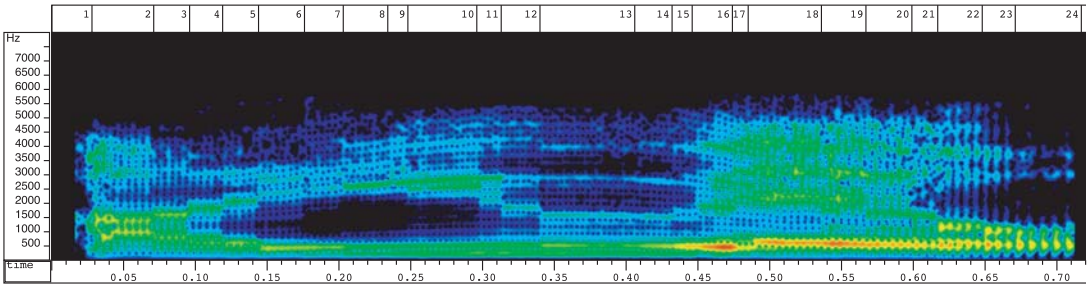


(b):提案手法の合成音（状態 18–21 が推定部分）

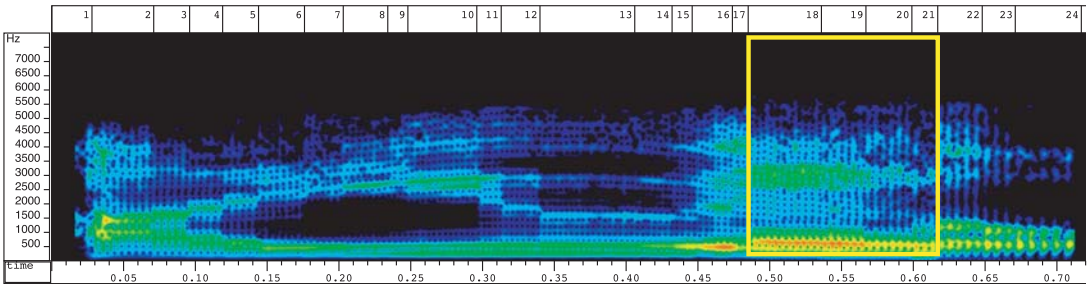


(c):提案手法の合成音（状態 22–25 が推定部分）

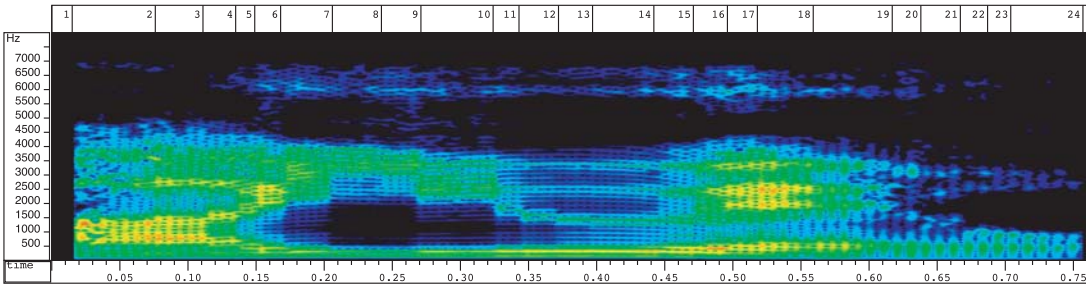
図 6.5: 異なる推定位置の合成結果 (2): 枠囲いされた部分が探索による音声.



(a):女性話者の分析再合成音



(b):初期条件提供話者（女性）の合成音



(c):構造提供話者（男性）の合成音

図 6.6: 異性別間の合成結果: 枠囲いされた部分が探索による音声.

第7章

解析手法による構造的表象からの音声合成の高精度化

7.1　はじめに

前章までにおいて，探索問題としての定式化を通した，日本語孤立母音系列および日本語連続母音系列の音声合成実験の結果を示してきた．探索問題としての定式化の問題点として，探索時間と了解度面での合成品質との間にトレードオフがあるという問題があった．すなわち元の空間全体の次元数を n，ブロックサイズを m，空間解像度を r としたとき，線形探索としての定式化では $O(\frac{n}{m}r^m)$ の計算量を必要とし，ブロックサイズが大きくなると指数関数的に計算量が増大する．しかし幼児の音声模倣に着眼すれば，模倣時に自身の発声できる候補を全探索しながら次に出力する音を模索しているとは考えにくい．幼児や当然我々成人も身体特性としての制約と過去の発声，調音の制約に基づいてより少ない候補の中から発声をしていると考えられる．

本章ではこのような考えに立ち，解析手法に基づき解候補を予め導出することで構造的表象を満たす音響事象を高速に推定する手法について提案し，その効果を実験的に検証する．

7.2　解析手法に基づく解候補の導出

7.2.1　楕円体の軌跡

今，第 4章で用いたバタチャリヤ距離（式 (4.1)）を再掲する．

$$BD(p_1(x), p_2(x)) = -\ln \int_{-\infty}^{\infty} \sqrt{p_1(x)p_2(x)}\, dx \tag{7.1}$$

ここで二つの確率分布 $p_1(x), p_2(x)$ がガウス分布 $\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)$ で表現されている時，式 (7.1) は以下のようになる．

$$BD(p_1, p_2) = \frac{1}{8}\mu_{12}^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} \mu_{12} + \frac{1}{2}\ln \frac{|(\Sigma_1 + \Sigma_2)/2|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}} \tag{7.2}$$

ただし $\mu_{12} = \mu_1 - \mu_2$ とする．このとき，式 (7.2) の右辺第一項は二つのガウス分布の平均ベクトルの差 μ_{12} を変数とする二次形式である．加えて Σ_1, Σ_2 は分散共分散行列であり，その要素は全て正となる．分散項が 0 でなければ，$((\Sigma_1 + \Sigma_2)/2)^{-1}$ は対角化可能で正の固有値を持ち，式 (7.2) の右辺第一項は正値二次形式となる [61]．このとき式 (7.2) において μ_1 を変数とし，その他を定数と見なせば，式 (7.2) が描く空間上の軌跡は μ_2 を中心とする多次元の楕円体となる．ただし以下の不等式の成立を仮定している．

$$BD(p_1, p_2) - \frac{1}{2}\ln \frac{|(\Sigma_1 + \Sigma_2)/2|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}} > 0 \tag{7.3}$$

式 (7.3) の第二項は二つの確率分布の広がり具合に大きな差が見られる場合にその値が大きくなる．孤立発声や連続発声の定常部ではそれぞれの音響事象は式 (7.3) をみたすものと考えられる．

よって音響事象 p_2 との関係性に着目して p_1 を導出する場合，p_1 は式 (7.2) の描く軌跡上にその中心を見出す事ができる．今 N 個の事象によって構造的表象が与えられている場合，式 (7.2) に基づた音響事象 p_1 に関する方程式を $N-1$ 個求める事ができる．故に p_1 以外の音響事象 $p_2, \ldots, p_N$ が初期条件として与えられていればこれらの連立方程式を解く事で p_1 の平均ベクトル μ_1 を導く事ができる．

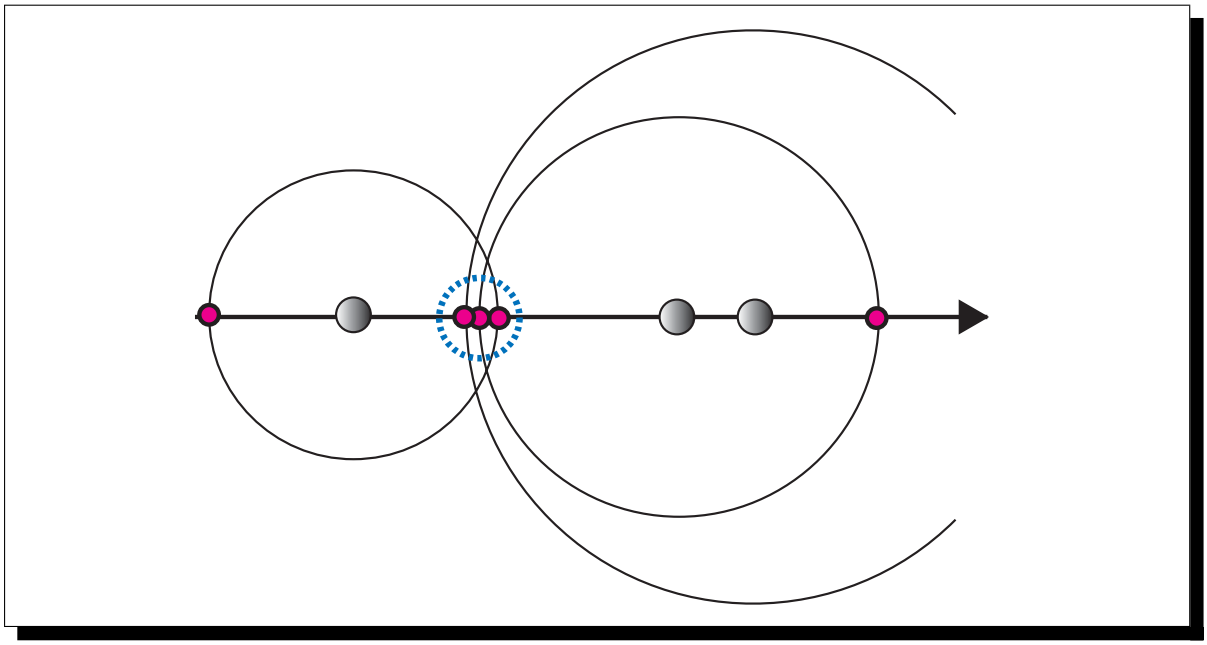


図 7.1: 解析手法に基づく解の導出（1 次元の場合）

7.2.2　ブロックサイズが 1 の場合

以下，具体的なブロックサイズにおいて解析手法に基づく解候補の導出を考えていく．いまブロックサイズを 1 とし二つの音響事象を A, B とする．いま 1 次元のある部分特徴量空間において二つの音響事象が 1 次元ガウス分布 $\mathcal{N}(c_a, V_a), \mathcal{N}(c_b, V_b)$ で表されているとする．これらを式 (7.2) に代入し，c_a について解くと以下のようになる．

$$c_a = c_b \pm \sqrt{(V_a + V_b)\left(BD - \frac{1}{2}\ln\frac{|(V_a + V_b)/2|}{|V_a|^{\frac{1}{2}}|V_b|^{\frac{1}{2}}}\right)} \tag{7.4}$$

このとき音響事象 A を音響事象 B との関係から推定する場合，その候補は対称性をもって二つ導出されることになる．よっていくつかの既知母音を用いて音響事象 A を一つに絞り込む事が必要になる．もし二人の話者の間で構造的表象が完全に一致している場合は，他の全ての音響事象を中心とする軌跡は一点を通ることになる．しかし構造的表象は発話スタイルなどのパラ言語的特徴によって変形を伴うため，これらの軌跡が必ずしも一点を通る事はない．ただしこれらの軌跡は求めるべき音響事象の “**真値**” の近傍を通る事になる．よって N 個の既知の音響事象 $p_1, \ldots, p_N$ から一つの音響事象 q を推定する流れは以下のようになる．下記では混乱しない範囲で便宜上事象 p_i とその平均 μ_i を同一の p_i で表記する．

1. 式 (7.4) に基づいて音響事象 p_i からの関係に基づく候補点 q_{i1}, q_{i2} を求める．もし式 (7.3) が成立しない場合，$q_{i1} = q_{i2} = p_i$ とする．
2. 各 i 毎に q_{i1}, q_{i2} を選択し，得られた N 個の点群についてその平均と分散を求める．
3. 2. を 2^N 通り全て試し，その分散が最小となる組み合わせについて，その平均を解とする．

この事を模式的に示したものが図 7.1である．よってブロックサイズ 1 のとき，音響事象を求めるには最低 2 つの既知の事象が与えられればよい．このとき調べる点は既知の事象数 N に対して 2^N でよいことになり，線形探索に比べて大幅な高速化が期待できる．

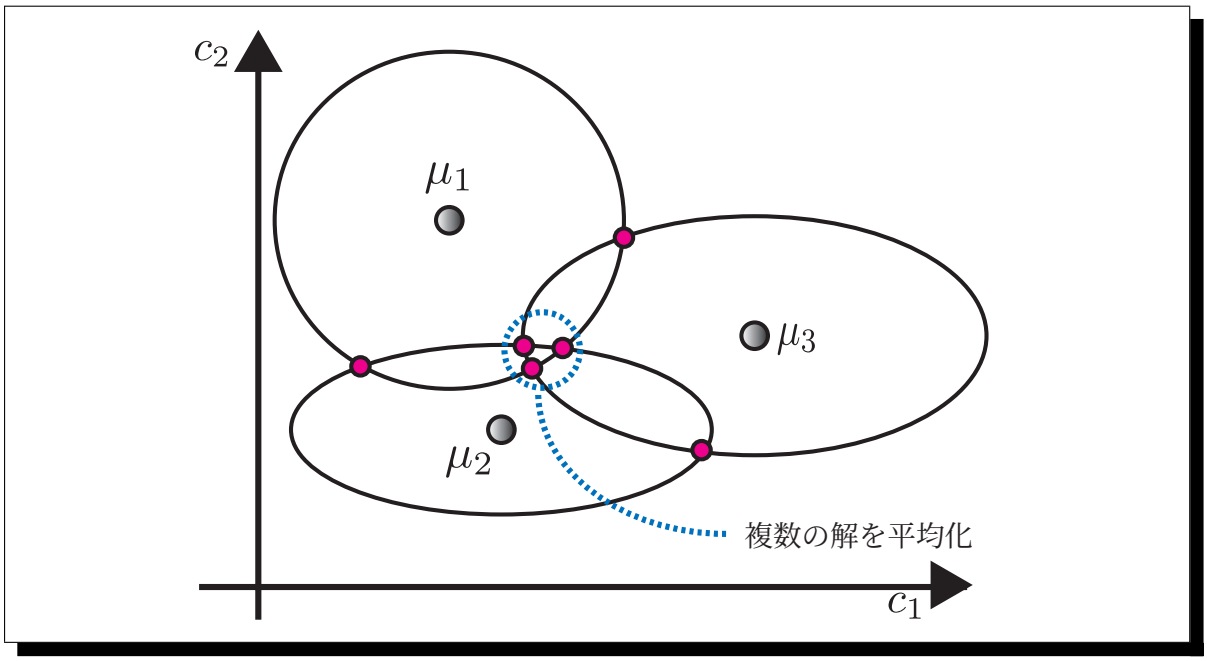


図 7.2: 解析手法に基づく解の導出（2 次元の場合）

7.2.3　ブロックサイズが 2 の場合

ブロックサイズが 2 の場合を考える．ある 2 次元の部分特徴量空間で初期条件として二つの音響事象 $A = \mathcal{N}(\boldsymbol{a}, \Sigma_a), B = \mathcal{N}(\boldsymbol{b}, \Sigma_b)$ が与えられているとする．この時音響事象 p の平均 $\mu = (c_x, c_y)$ を求めたいとする．ただし p の分散共分散 Σ は対角で既知とし $\mathrm{diag}(\Sigma) = (V_x, V_y)$ とする．ただし $\mathrm{diag}(\Sigma)$ は行列 Σ の対角成分によるベクトルを表す．また同様に $\mathrm{diag}(\Sigma_a) = (V_{ax}, V_{ay}), \mathrm{diag}(\Sigma_b) = (V_{bx}, V_{by})$ とする．このとき音響事象 A, B と p とのバタチャリヤ距離 BD_a, BD_b を用いると μ に対して以下の連立方程式が成立する．

$$\begin{cases} BD_a - \dfrac{1}{2}\ln\dfrac{|(\Sigma+\Sigma_a)/2|}{|\Sigma|^{\frac{1}{2}}|\Sigma_a|^{\frac{1}{2}}} = \dfrac{1}{4(V_x+V_{ax})}(c_x-a_x)^2 + \dfrac{1}{4(V_y+V_{ay})}(c_y-a_y)^2 \\ \\ BD_b - \dfrac{1}{2}\ln\dfrac{|(\Sigma+\Sigma_b)/2|}{|\Sigma|^{\frac{1}{2}}|\Sigma_b|^{\frac{1}{2}}} = \dfrac{1}{4(V_x+V_{bx})}(c_x-b_x)^2 + \dfrac{1}{4(V_y+V_{by})}(c_y-b_y)^2 \end{cases} \tag{7.5}$$

これは 2 次元において楕円の交点を求めることに相当する．この時二つの楕円の交点は一般には 2 つ，長軸および短軸の配置により最大で 4 つ求まる．そのため一つの音響事象を求める為にはさらに方程式が必要となる．一般に n 次元空間において n 個の超楕円体だけでは交点をただ一つに定めることはできない．よって n 次元における一つの音響事象の定位には少なくとも $n+1$ 個の音響事象が必要となる．

ブロックサイズが 1 の場合と同様，提案するアプローチについて述べる．既知の音響事象数を n とすると，連立方程式は ${}_nC_2 = m$ 個求まる．今これらを EQ_i $(i = 1, \ldots, m)$ と表記する．EQ_i は一つの連立方程式を表す．このとき一つの音響事象 q を推定する流れは以下のようになる．

1. 連立方程式 EQ_i から求まる候補点 $p_{i1}, \ldots, p_{i4}$ を求める．これらの個数は変動する．また式 (7.3) を満たさない方程式を含む場合はその方程式の中心音響事象を p_{i1}（両方の方程式が満

表 7.1: 聴取実験の結果（7.3.2節）; 括弧内はブロックサイズ，表の値は全モーラ一致の正答率を表す.

	総単語数	線形探索 (1)	線形探索 (2)	提案手法 (1)	提案手法 (2)
既知母音数 1	40	0.05	—	—	—
既知母音数 2	80	0.20	—	0.14	—
既知母音数 3	80	0.45	0.40	0.44	0.28
既知母音数 4	40	0.83	0.73	0.65	0.73

たさない場合は p_{i1}, p_{i2}）とする．共に楕円が存在するが解が求まらない連立方程式の場合はその連立方程式 EQ_i からは候補点を求めない.

2. 各 i について毎に候補点を選択し，得られた最大 m 個の点群の平均と分散共分散を求める.
3. 2. を最大で 4^m 通り繰り返し，分散共分散行列の行列式が最小になる平均を解とする.

提案するアプローチを図 7.2に示す．計算量は $m = {}_nC_2$ によって支配的である．すなわち既知の事象数が少なければ高速な推定が可能となる．一方でその時に得られる点群の数が小さくなるため，頑健な推定との間に若干のトレードオフが存在することになる.

7.3　孤立音声による実験

7.3.1　実験方法

提案手法の有効性について調べるため，日本語 5 母音の孤立発声（/a/, /i/, /u/, /e/, /o/）を用いて実験を行った．成人男女各 2 名（それぞれ話者 M1，M2，F1，F2 とする）の日本語 5 母音の発声を収録した．これらの発声について STRAIGHT[35] に基づくスペクトル分析を行い，このスペクトルから 40 次のケプストラムを得た．同時に発声のピッチ，パワー，継続長も STRAIGHT の分析を基に得た.

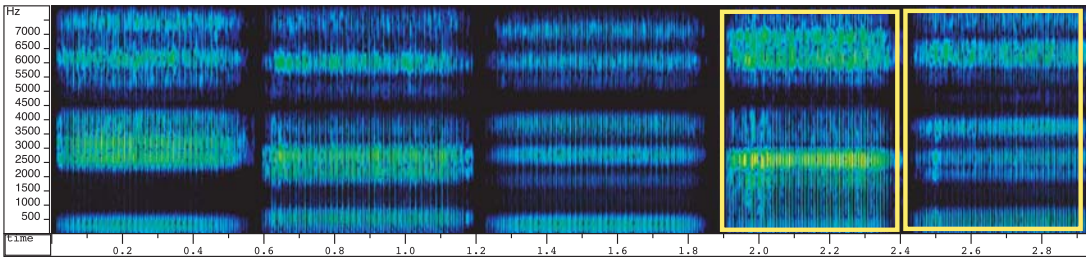
これらのパラメータから図 4.3の流れで構造を抽出した．さらに既知母音の数および次元分割のブロックサイズを変化させ，従来の線形探索に基づく手法および提案手法で音響空間探索と音声の合成を行った．構造を提供する話者は上記の 4 名，初期条件を提供する話者を上記のうち男女 1 名ずつ（M1, F1）とした．幼児の音声模倣の枠組みでいえば，構造提供話者は母親に，初期条件提供話者は幼児に対応する.

また音質について評価するため，生成したサンプルと初期条件とした既知母音によって構成される 5 モーラ単語を作成し成人男性 5 名による書き取りテストを行った．単語知識の影響を除くため，あらかじめ被験者に 5 モーラの出現の重複を許す旨を伝えた.

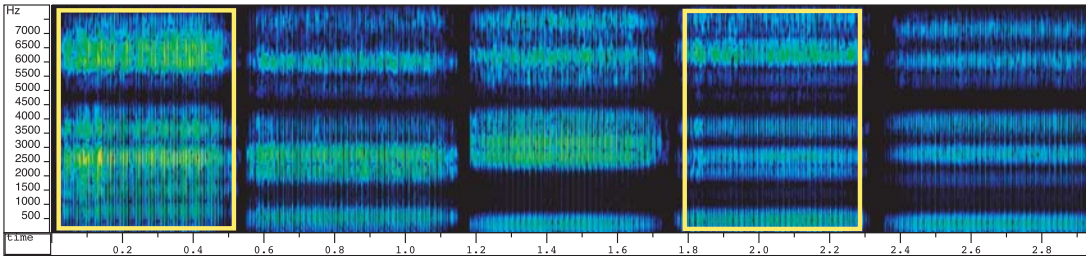
7.3.2　実験結果

実験によって得られた合成音声の一例を図 7.3に示す．図 7.3は構造提供話者として話者 F2，初期条件提供話者として話者 M1 の音声を用いている．探索および解析によって母音 /a/, /o/ を推定した．図 7.3(a) は従来の線形探索による結果，図 7.3(b) は提案手法による結果を示している．このとき両手法により得られるスペクトルにほとんど差がない事がわかる.

一方計算時間について比較する．筆者の実験環境において，ブロックサイズを 1 とした場合，従来手法では既知母音数と話者の組み合わせを変化させた 240 セットを作成するのに，3 時間程度



(a): 従来手法による合成音 /ieuao/



(b): 提案手法による合成音 /aeiou/

図 7.3: 合成音声の一例

の時間を要していた．一方提案する手法では同様の条件において，30 秒程度で終了する．およそ 20000 倍の高速化を実現したことになる．

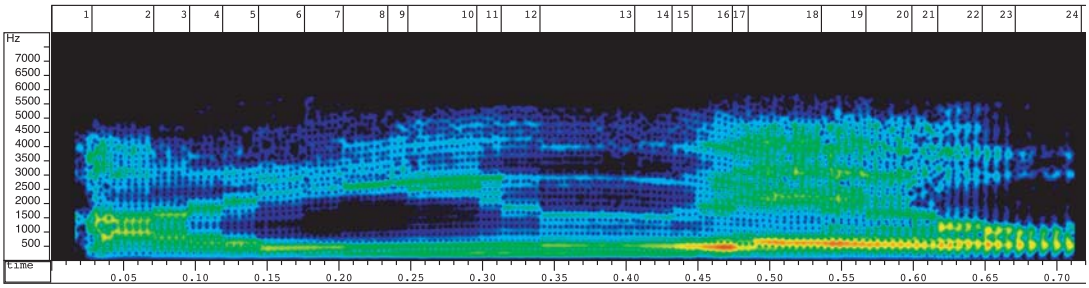
続いて聴取実験の結果を表 7.1に示す．括弧内の数字は特徴量分割におけるブロックサイズを表している．各欄の数字は 5 名の話者のうち 3 名以上が 5 モーラ全て一致させた場合を正答として正答率を求めている．数値の存在しない欄については探索時間が膨大または方程式の個数が不足しているため実験を行っていない．表 7.1 から，提案手法による合成音の了解度は従来の線形探索の結果を若干下回る結果となった．しかし一方で提案手法内では，既知母音数が 4 の時にブロックサイズを大きくすることによる了解度の改善が確認された．

従来手法ではブロックサイズが大きい場合，探索時間とのトレードオフにより各次元を低い解像度で探索しなければならない．一方提案手法では，解像度，探索時間の両面で従来手法のような制約を受けないため，このことが有効に作用していることがわかる．一方了解度の改善が，既知母音数が多い場合に限定的であるのは，前述した既知母音数に応じた候補の点群の数が少なくなることに起因すると考えられる．またさらなる発展手法として，得られた解候補をもとに構造による制約を満たすまで繰り返し演算を行う等により合成品質を向上させることがが可能であると考えられる．

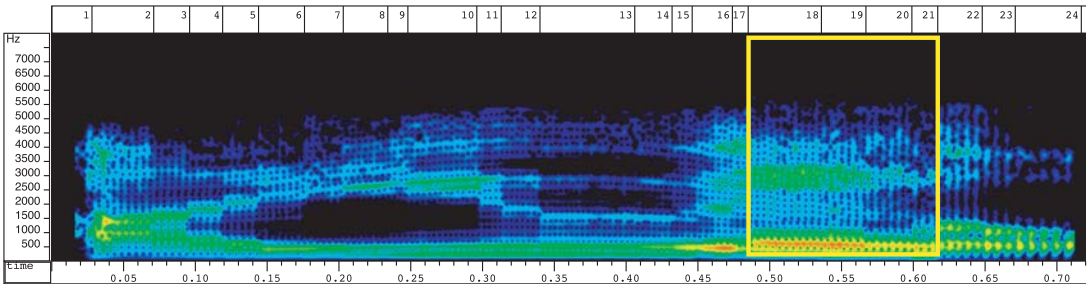
7.4　連続音声による実験

7.4.1　実験方法

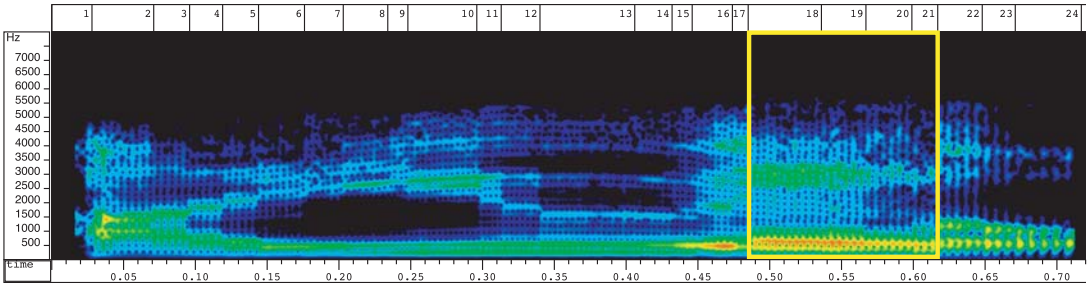
解析手法の導入による高速化によって，第 6章における連続音声の合成において，より少ない初期条件から合成が可能になると考えられる．そこで連続音声においても同様に解析的手法を導入し，実験を行った．音響分析条件は孤立母音の場合と同様であり，連続音声における状態推定は第 6章と同様の条件で行った．初期条件として，(1) 21 状態および (2) 5 状態 の二通りを与えた．(1) は探索に基づく合成と同様の条件であり，(2) は計算量の関係から従来手法では合成困難



(a): 初期条件提供話者の分析再合成音



(b): 探索手法による合成音



(c): 解析手法による合成音

図 7.4: 連続音声における解析的手法による結果（初期条件 21 状態）．枠囲いされた部分が推定音;(a): 分析再合成音; (b): 探索手法による合成音; (c): 解析手法による合成音;

だったものである．加えて (2) については，およそ 1 母音相当の連続した状態を与えた場合および発話全体の中から飛び石のように疎らに 5 状態を与えた場合の二つについて実験した．

7.4.2　実験結果

ブロックサイズを 1，初期条件を 21 状態とした場合について，従来手法および分析再合成音と比較した結果を図 7.4に示す．構造提供話者は成人男性，初期条件提供話者は成人女性，合成単語は/aiueo/ である．図 7.4(a) は初期条件提供話者の分析再合成音，図 7.4(b) は従来の探索による手法，図 7.4(c) が解析的手法による合成音である．このとき (b) および (c) についてはほぼ同様のスペクトルが再現されている事が分かる．これらは他の推定箇所の場合でも同様であった．一方，計算時間については提案手法の方が探索手法よりも倍程度の時間がかかった．これは探索手法における解像度が 25 と低いため平等な比較ではないが，複数の解候補を一つにまとめあげる際に既知の状態数に応じて計算量が指数的に増大する問題が顕著に現れているといえる．このこと

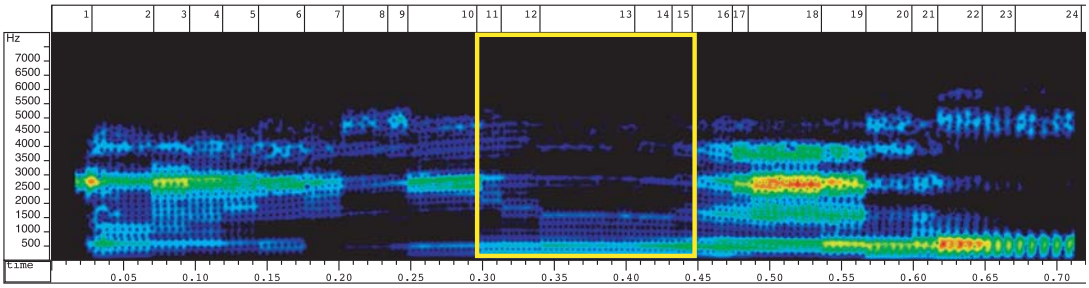
から品質と計算量のトレードオフを考慮して最適な数の初期条件を与える必要があるといえる．

一方，ブロックサイズ 1，初期条件を 5 状態とした場合についての結果を図 7.5に示す．構造提供話者は成人男性，初期条件提供話者は成人女性，合成単語は/aiueo/ である．図 7.5においては，枠囲いされた部分が初期条件として与えた部分を表している．初期条件提供話者の分析再合成音である (d) と比較した場合，(a) および (b) においてはスペクトルの様子が異なる一方で，(c) ではおよそ正しく再現されている事が分かる．これは (a)(b) では初期条件として与えた音声が音響空間上において非常に近いため，候補から解を導出する際に対称な音声を捉えてしまうことが原因と考えられる．一方，(c) では，個々の初期条件が音響空間上である程度離れているため，適切にその他の状態を推定できていると考えられる．この結果はいくつかのターゲットとなる音響事象と連続的な動きの情報のみから，連続発声を適切に合成できることを示すものである．なお計算量は既知の音響事象の数に依存するため，今回の合成は 4 状態を推定する場合よりも高速に実現されている．

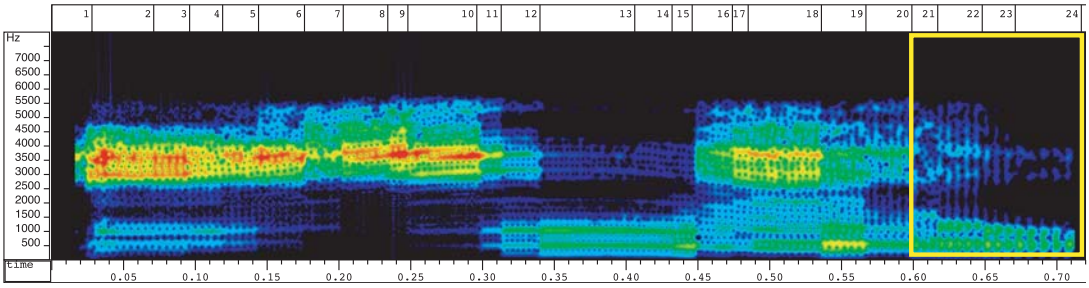
またブロックサイズを 2，初期条件を 5 状態とした場合の結果を図 7.6に示す．構造提供話者は成人男性，初期条件提供話者は成人女性，合成単語は/aiueo/ である．図 7.6において，(a) のスペクトルは (c) とは異なるが，予備的な聴取において，聴感上は違和感なく/aiueo/ と知覚された．一方 (c) と比較すると若干異なる話者性が感覚された．これは連続する状態だけでは構造の定位が不十分な一方で，ブロックサイズの制約を変化させることで構造に付与する話者性が異なってくることを示唆している．一方 (b) はいくつかのターゲットを与える事で，(c) に非常に近いスペクトルを再現できていることが分かる．なお計算量については，ブロックサイズが 1 の場合よりも多くの時間を要した．これは先に議論した通り，ブロックサイズの増加により方程式の組み合わせが増大することに起因している．

7.5　まとめ

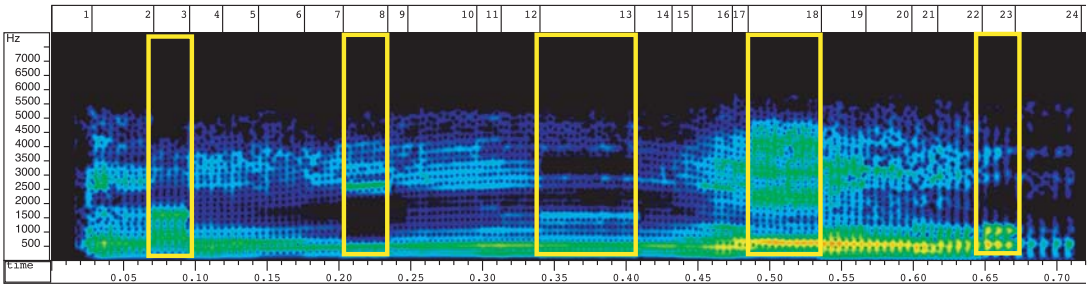
本章では解析手法を導入する事で提案する枠組みにおける音響空間探索を高速に実現する手法を提案した．孤立 5 母音系列の合成においては，合成音声の了解度の面では探索手法よりも若干劣る結果となった．しかし一方で従来の手法に対して約 20000 倍の高速化を実現した．さらに従来の探索問題では明らかにできなかった，ブロックサイズの増加による合成音声の了解度向上を確認する事ができた．加えて初期条件の数によって計算時間と合成品質との間にトレードオフがあることも確認された．連続 5 母音系列の合成においては，少数の初期条件を基に連続発声全体の合成が可能である事を示した．なお現在従来の探索問題において照合条件としている構造歪みや構造類似度は考慮していない．今後はこれらの基準を考慮したうえで，繰り返し演算などによって品質を向上させることができると考えられる．また本章での議論により，より多次元のブロックサイズにおいても提案する枠組みの検討が可能である事が理論的に示された．今後は本章での手法をより多次元のブロックサイズへと拡張し，より最適なパラメータの検討を行う必要がある．



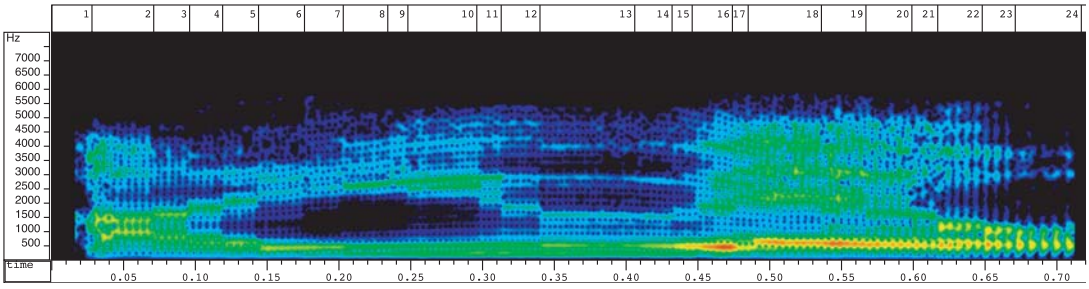
(a): 連続する状態を与えた場合（状態 11–15）



(b): 連続する状態を与えた場合（状態 21–25）

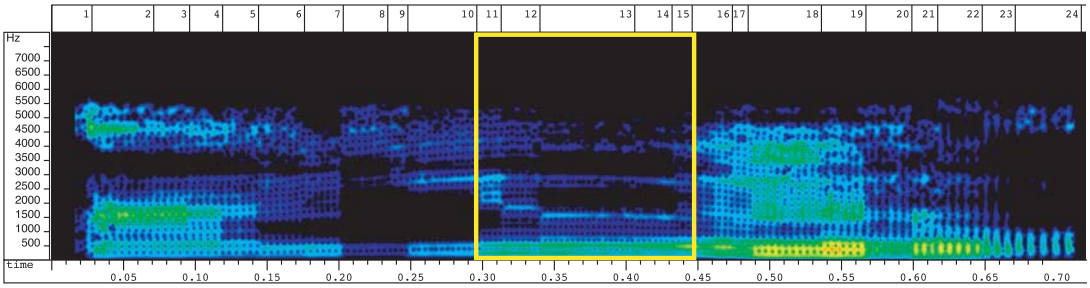


(c): 飛び石的に状態を与えた場合

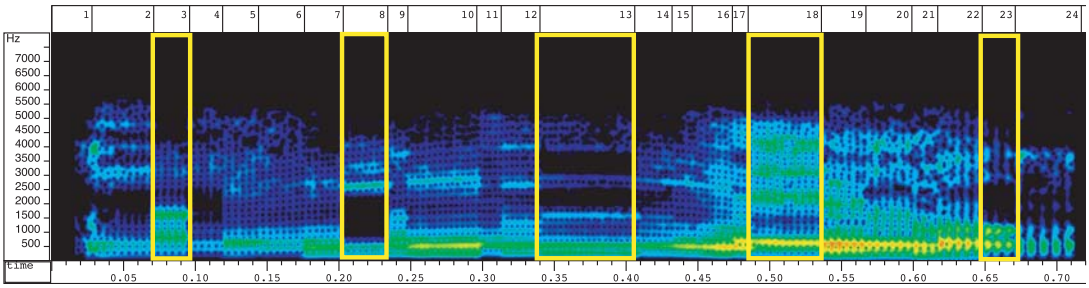


(d): 初期条件提供話者の分析再合成音

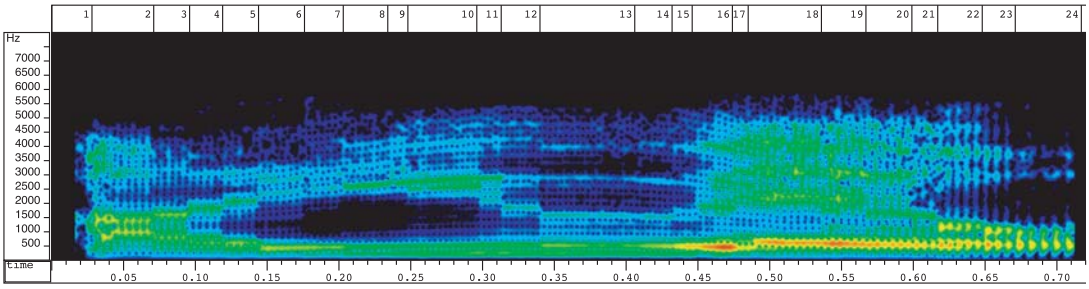
図 7.5: 連続音声における解析的手法による結果（初期条件 5 状態）．枠囲いされた部分が**初期条件**; (a): 連続する状態を与えた場合（状態 11–15）; (b): 連続する状態を与えた場合（状態 21–25）; (c): 飛び石的に状態を与えた場合; (d): 分析再合成音



(a): 連続する状態を与えた場合（状態 11–15）



(b): 飛び石的に状態を与えた場合



(c): 初期条件提供話者の分析再合成音

図 7.6: 連続音声における解析的手法による結果（ブロックサイズ 2，初期条件 5 状態）．枠囲いされた部分が**初期条件**; (a): 連続する状態を与えた場合（状態 11–15）; (b): 飛び石的に状態を与えた場合; (c): 分析再合成音

第8章

結論

8.1　本研究の成果

本研究では，幼児の音声模倣のより妥当なモデルとしての音声合成という観点から出発し，近年提案されている音声の構造的表象を通して従来の音声合成とは大きく異なる音声合成技術の枠組みを提案し，その基礎的検討を行った．本章では本研究の成果および展望について整理する．

第 1章において,「メディア情報処理」における音声情報処理技術の位置づけを確認するとともに，より柔軟な，より人間に近い音声コミュニケーションを計算機によって処理する上での問題提起を行った．その上で従来の音声合成システムの枠組みと幼児の音声言語獲得の出発点である音声模倣との間に非常に大きな乖離があることを示した．

第 2章においては従来の音声合成システムとして，人間の生成過程に立脚した上でその調音運動を再現するシステム，および近年の主流となっている，あらかじめ収録された人間の音声を出発点として，その処理によって音声合成を実現するシステムの二つに大別して述べた．その上で本研究で提案する枠組みのあるべき位置づけを示した．

第 3章において，音声に不可避的に存在する非言語的特徴のモデル化について述べた．特に人間の声道長の違いに着眼し，音響特徴量空間における幾何学的特性を理論的に示した．さらに高品質の分析再合成音による実験を行い，特徴量空間における話者性表出の一検討を行った．第 3章における成果は，従来の音声情報処理において方向成分を直接的に扱うことの問題点を明らかにし，また特徴量空間における多次元の幾何学的特性を情報記述に利用するという新しいアプローチを導入し，話者情報表現としての可能性を示した点にある．

第 4章においては，第 3章で述べた音声に不可避的に含まれる非言語的特徴を，本質的に包含しない音声の不変表象について述べ，その特徴を示した．さらにこの話者不変の音声の構造的表象に基づく音声合成技術の枠組みについて述べた．加えて，現在広く用いられている話者適応学習の技術と比較し，幼児の音声模倣を説明する上でのアプローチの違いを明確にした．第 4章での問題提起は，従来技術と提案する枠組みとの共通点および相違点を明らかにすることであり，より柔軟な「メディア情報処理」研究の発展のための礎となることを期待するものである．

第 5章において，提案する構造的表象に基づく音声合成技術の基礎的検討として，日本語孤立 5 母音系列の合成に関して行った実験について述べた．本研究において提案する音声合成の枠組みは，不変表象に対して身体特性を付与することで初めて音声が生成される枠組みであり，この基礎検討として，音響特徴量空間における制約条件つき解探索問題として定式化を行い，提案する枠組みによって，孤立 5 母音系列のいくつかの事象を初期条件とした状態で残りの音声を合成可能である事を示した．第 5章における枠組みは，語形全体を真似るという観点から厳密には音声模倣のモデル化とは言えないが，語の体系を真似るという観点から体系全体を真似る事によって実体を獲得できる事を示したことになる．

第 6章においては，提案する枠組みによる音声合成を日本語連続 5 母音系列に拡張して行った実験について述べた．構造的表象に基づく音声認識において提案されている HMM に基づく分布状態系列へのモデル化と最大事後確率推定による安定した分布推定によって，連続音声においても提案する枠組みによって音声合成が可能である事を示した．

第 7章において，前章で用いてきた手法のさらなる高精度化を目的とした，解析的手法の導入による高速な解空間探索に関する実験について述べた．音響事象間の関係が満たす距離関係を方程式として再解釈し，解析的に解候補を導出することで，約 20000 倍の高速化が可能となる事を示した．さらに連続発声が少数の音響事象を初期条件として合成可能であることを示した．

本研究の成果は従来のシステムとは大きく異なる枠組みを提案し，その基礎的検討において枠組みの実現可能性を示したことにある．

8.2　今後の展望

本研究の今後の課題を述べる．本研究は基礎的検討の段階にあるため多くの課題が残されているといえる．以下基礎技術としての課題，アプリケーション応用としての課題，メディア情報処理としての課題に分けて述べる．

基礎技術としての課題

本論文を通して，音声合成において扱うパラメータとして，スペクトル情報であるケプストラムのみを取り扱った．すなわちその他のピッチ，継続長，パワーなどの韻律的特徴は既知のものとした．しかしこれらの特徴も単語全体を表象するという観点から非常に重要であり，これらの特徴を本枠組みの中でどのように扱っていくのかが課題となる．また今回ブロックサイズ決定など探索問題の定式化による制限を受けていた．第 7章における議論により今後これらのパラメータをより最適なものにしていく必要がある．また構造的表象における音声認識においては構造統計モデルの構築など，いくつかの有用な技術が検討されている．今回これらの検討を行っていないため音声認識におけるいくつかの技術を音声合成においても適用できないかを検討せねばならない．最後に子音を含めた合成や文などのより長単位での音声合成は，単語獲得など幼児の言語活動とも関連が深く今後検討していかなければならない．

アプリケーション応用としての課題

本研究のアプリケーションとしての発展は二つの可能性をもっている．一つは単語の語ゲシュタルトの観点から見た幼児の音声模倣のシミュレーション，もう一つは語の体系の観点からみた CALL システムへの応用である．本論文中で主張したように本研究における枠組みは幼児の音声模倣を妥当に説明しうる．そのため人間の音声言語活動を記述する観点からシミュレーションを行えるようにすることは，福祉利用まで視野にいれたアプリケーション応用ということができる．一方，ソシュール，ヤコブソンの考えに立てば語学はその音の体系を学ぶ必要があり，すでにその観点からの研究が行われている．本研究の合成技術は学習者へ音響的にフィードバックできるという側面があり，その応用が望まれるところである．

メディア情報処理としての課題

最後に大きな課題をあげたい．本研究の出発点となっている構造的表象は音声の差を捉える事で抽象化された表象となっている．すなわち構造的表象の枠組みはその他のメディア，さらにはメディア情報処理全体への応用が可能であると考えられる．メディア変換の可能性については 4.5.2で触れた．ここでは例として映像メディアへの応用を考えてみる．映像を見てその内容を理解する行為は 20 世紀からの所産であるが，映像は画像の系列ではなくその全体を捉える事で新たな意味を生み出すメディアである [62]．このようなメディア特性の類似性を一つの表象を出発点として扱う事はメディア情報処理としての大きな課題である．近年のメディア処理は「マルチメディア」と呼ばれているが個々のメディアストリームを単に並列に並べるだけではなく，抽象化・統合を通したシームレスなメディア情報処理が必要となっている．構造的表象を中心としたメディア情報処理技術の検討は,「シームレスなマルチメディア」の可能性の一つとなりうるだろう．

謝辞

本研究ならびに本論文の執筆にあたり，多大なる御指導，御鞭撻を賜りました指導教員の峯松信明准教授ならびに広瀬啓吉教授に深く感謝いたします．研究開始当初，研究室においても全く未開の領域であった本研究の音声合成の枠組みがこのように論文の形になるのは，両先生の懇切丁寧な御指導と適切な御助言が大きな支えとなりました．また，日頃の研究室活動を支えてくださった高橋登技官，秘書の笠島恵さん，楠本由香里さん，武田祥子さんに深く感謝いたします．

本研究を進めていく上で，博士課程の朝川智氏には数多くの鋭いご指摘やご意見を頂きました．また博士課程の学生として後輩を適切に指導していく姿など，研究者としての振る舞いも様々勉強になりました．深く感謝いたします．研究のアプローチについて理論的かつ多角的な視点から様々なご意見を頂いた，特別研究員の喬宇博士にも深く感謝いたします．本論文の第 3 章における理論や第 7 章は喬氏のご意見を出発点として展開する事ができたものです．

また共に研究を始め，今共に修士課程を修了する鎌田圭氏，下村直也氏に深く感謝いたします．二人と共に様々な事を語り笑い飛ばしながら，切磋琢磨して研究に励む事ができました．

柏の研究室生活におきましては，大学院入学当初から博士課程の平野宏子氏，当時修士課程の上西康太氏（現在 NTT 勤務），同じく当時修士課程の越智景子氏（現在情報理工学系研究科博士課程）に様々なご支援を頂きました．深く感謝いたします．皆様のお陰で楽しく大学院生活をスタートすることができたと思っています．また現在柏での研究室生活を共に過ごしている博士課程の鎌田敏明氏，馬学彬氏，修士課程の印南圭祐氏，國越晶氏，松浦良氏にも深く感謝いたします．日々の生活を楽しく快適なものにできているのは皆様のお陰です．

また本郷オフィスにおいても研究室の皆様に大変お世話になりました．特に大学院 1 年目，当時博士課程の八木裕司氏（現在日立製作所勤務）には積極的かつ真剣に研究と息抜きに取り組む姿勢を身をもって教えて頂きました．深く感謝いたします．また広瀬・峯松両先生指導のもと筆者の学年は多くの同期に恵まれました．石井英資氏，稲垣貴彦氏，篠田知宏氏，デゲル・エルハン氏，ナジャンド・アリ氏，ナリニョオ・ホアン氏，三輪周作氏，レボルダオ・アントニオ氏が同期であった事は研究の大きな支えとなりました．深く感謝いたします．また全ての方のお名前をここに挙げる事は叶いませんが，広瀬・峯松研究室のみなさまには日頃から大変お世話になりました．深く感謝いたします．

本研究に関連する口頭発表ならびにポスター発表において，数多くの研究者の皆様に議論に参加していただき，貴重なご意見を頂く事ができました．この場を借りて深く感謝いたします．

また日々の生活において，数多くの友人，中学・高校の同級生や先輩後輩，学部時代の部活の同期や先輩後輩などこれまで交流のあった皆様に深く感謝いたします．日々の生活の活力，時に研究の着想も皆様との交流にその原点があると考えています．

最後にここまで育てて頂いた両親と弟・和輝への謝意を表したいとおもいます．どうもありがとうございました．

2008 年 1 月 29 日

齋藤 大輔

参考文献

[1] L. R. Rabiner and R. W. Schafer: “Digital processing of speech signals, ” Prentice Hall, 1978.

[2] 古井貞熙: “ディジタル音声処理,” 東海大学出版会, 1985.

[3] Museum of Speech Analysis and Synthesis, `http://mambo.ucsc.edu/psl/smus/smus.html`

[4] 内田伸子編: “発達心理学キーワード,” 有斐閣双書, 2006.

[5] 天野清: “子どものかな文字の習得過程,” 秋山書店, 1986.

[6] 早川勝廣: “言語獲得と育児語,” 月刊言語, vol. 35, no. 9, pp.62–67, 2006.

[7] N. S. トルベツコイ, “音韻論の原理,” 岩波書店, 1958.

[8] 加藤正子: “特集にあたって”, コミュニケーション障害学, vol. 20, no. 2, pp.84–85, 2003.

[9] N. Minematsu, T. Nishimura, K. Nishinari and K. Sakuraba:“Theorem of the invariant structure and its derivation of speech Gestalt,” *Proc. Int. Workshop on Speech Recognition and Intrinsic Variations,* pp.47–52, 2006.

[10] 峯松信明, 西村多寿子, 櫻庭京子, 朝川智, 齋藤大輔: “孤立音 [あ] を聞いて/あ/と同定する能力は音声言語に必要か？,” 電子情報通信学会技術研究報告, SP2007–30, pp.37–42, 2007.

[11] 今井聖, 住田一男, 古市千枝子: “音声合成のためのメル対数スペクトル近似（MLSA）フィルタ,” 電子情報通信学会論文誌 (A), vol.J66-A, no. 2, pp.122–129, 1983.

[12] 緒方公一, 増矢拓郎:“Web アプリケーションとしての声道音響管モデルに基づく母音合成システムの開発,” 日本音響学会誌 No.62 vol.3, pp.199–207, 2006.

[13] K. Ishizaka and J. L. Flanagan:“Synthesis of voiced sounds from a two-mass model of the vocal cords,”, *Bell Syst. Tech. J.,* 51, pp.1233–1268, 1972.

[14] 誉田雅彰, 西川員史, 高西淳夫, 廣谷定男, 持田岳美:“人間型発話ロボット–喉を震わせ口を動かして発話するロボット–,” 日本音響学会誌, no.61 vol.2, pp.91–96, 2005.

[15] J. Dang, K. Honda: “Estimation of vocal tract shape from sounds via a physiological articulatory model,” *J.Phonetics,* vol.30, pp.511–532, 2002.

[16] 西澤信行, 河井恒:“素片接続型音声合成における最良優先探索に基づく素片選択,” 電子情報通信学会技術研究報告, SP2005-161, pp67–72, 2006.

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

[18] W. Hamza, R. Bakis, Z. W. Shuang and H. Zen: “On building a concatenative speech synthesis system from the Blizzard Challenge Speech Databases,” *Proc. EUROSPEECH,* pp. 97–101, 2005.

[19] T. Nose, J. Yamagishi, T. Masuko and T. Kobayashi: “A style control technique for HMM-based expressive speech synthesis,” *IEICE Trans. Inf. & Syst.,* vol.E90-D, no. 9, pp.1406–1413, 2007.

[20] A. Black, P. Taylor and R. Caley: “The festival speech synthesis system, ” `http://www.festvox.org/festival/`

[21] M. Schröder and J. Trouvain:“The German text-to-speech synthesis system MARY: A tool for research, development and teaching,” *International Journal of Speech Technology,* vol.6, pp.365–377,, 2003.

[22] J. Yamagishi, H. Zen, T. Toda and K. Tokuda:“Speaker-independent HMM-based speech synthesis system — HTS-2007 system for the Blizzard Challenge 2007.” *Proc. Blizzard Challenge 2007*, 2007.

[23] N. Minematsu: “Yet another acoustic representation of speech sounds,” *Proc. ICASSP*, pp.585–588, 2004.

[24] N. Minematsu: “Mathematical evidence of the acoustic universal structure in speech,” *Proc. ICASSP*, pp.889–892, 2005.

[25] 峯松信明, 西村多寿子, 西成活裕, 櫻庭京子: “構造不変の定理とそれに基づく音声ゲシュタルトの導出,” 電子情報通信学会技術報告, SP2005-12, pp.1–8, 2005.

[26] 峯松信明, 西村多寿子: “音声の相対音感 〜音声と音楽の同質性に関する一考察〜,” 電子情報通信学会技術報告, SP2005-131, pp.121–126, 2005.

[27] M. Pitz, S. Molau, R. Schlüter and H. Ney: “Vocal tract normalization equals linear transformation in cepstral space,” *Proc. EUROSPEECH*, pp.1445–1448, 2001.

[28] M. Pitz and H. Ney: “Vocal tract normalization equals linear transformation in cepstral space,” *IEEE Trans. Speech and Audio Processing*, vol. 13, pp.930–944, 2005.

[29] M. Russel and S. D’Arcy: “Challenges for computer recognition of children’s speech,” *CD-ROM of SLaTE2007*, 2007.

[30] B. Atal: “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” *J.Acoust.Soc.America,* vol. 55, pp. 1304–1312, 1974.

[31] E. Eide and H. Gish: “A parametric approach to vocal tract length normalization,” *ICASSP96,* vol. 1, pp. 346–348, 1996.

[32] 江森正, 篠田浩一: “音声認識のための高速最ゆう推定を用いた声道長正規化,” 電子情報通信学会論文誌 D-II, vol. J83-D-II, no. 11, pp.2108–2117, 2000.

[33] T. Emori and K. Shinoda: “Rapid Vocal Tract Length Normalization usgin Maximum Likelihood Estimation,” *Eurospeech2001,* pp. 1649–1652, 2001.

[34] R.A. Horn and C.R. Johnson: “Matrix Analysis,” Cambridge University Press, 1985.

[35] H. Kawahara *et al.*:“Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” *Speech Communication,*, vol. 27, pp. 187–207, 1999.

[36] 峯松信明, 松井健, 広瀬啓吉: “声に内在する音響的普遍構造とそれに基づく音声コミュニケーション,” 第三回「話し言葉の科学と工学」ワークショップ論文集, pp. 143–150, 2004.

[37] 高橋友和, Lina, 井手一郎, 目加田慶人, 村瀬洋: “高次元回転行列の補完とその応用,” Visual Computing / グラフィックスと CAD 合同シンポジウム, 2007.

[38] フェルディナン・ド・ソシュール:“一般言語学講義,” 岩波書店, 1940.

[39] ローマン・ヤコブソン: “構造的音韻論,” 岩波書店, 1996

[40] 峯松信明, 朝川智, 広瀬啓吉:“線形・非線形変換不変の構造的情報表象とそれに基づく音声の音響モデリングに関する理論的考察,” 日本音響学会春季講演論文集, 1-P-12, pp. 147–149, 2007.

[41] 大谷大和, 戸田智基, 猿渡洋, 鹿野清宏:“固有声に基づく声質変換のための話者正規化学習法,” 電子情報通信学会技術研究報告, SP2006-40, pp.37–42, 2006.

[42] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul: “A compact model for speaker-adaptive training,” *Proc. ICSLP,* vol.2, pp.1137–1140, 1996.

[43] C. J. Leggetter and P. C. Woodland: “Maximum likelihood speaker adaptation of continuous density hidden markov models,” *IEEE Trans. Speech and Audio Processing,* vol. 3, no. 5, pp. 357–366, 1995.

[44] 村上隆夫, 丸山和孝, 峯松信明, 広瀬啓吉:“音声の構造的表象を用いた用いた日本語母音系列の自動認識,” 電子情報通信学会技術研究報告, SP2005-14, pp.13–18, 2005.

[45] 朝川智, 村上隆夫, 峯松信明, 広瀬啓吉: “音声の構造的表象に基づく日本語母音系列連続発声の認識,” 電子情報通信学会技術研究報告, SP2006-105, pp.119–124, 2006.

[46] 錦戸信和, 党建武, “調音モデルを用いた特異発話状態の調査,” 日本音響学会春季講演論文集, 1-Q-28, pp.319–320, 2007.

[47] 錦戸信和, 党建武, “GMM を用いた通常発話状態と特異発話状態の弁別,” 日本音響学会秋季講演論文集, 1-P-18, pp.442–444, 2007.

[48] 朝川智, 峯松信明, 広瀬啓吉: “音声の構造的表象を用いた音声認識における特徴量空間分割とその効果,” 日本音響学会秋季講演論文集, 3-Q-10, pp. 229–232, 2007.

[49] S. Asakawa, N. Minematsu and K. Hirose: “Multi-stream parameterization for structural speech recognition,” *Proc. ICASSP,* 2008 (to appear).

[50] 峯松信明: “音声の音響的普遍構造の歪みに着眼した外国語発音の自動評定,” 電子情報通信学会技術研究報告, SP2003-180, pp.31–36, 2004.

[51] N. Minematsu, S. Asakawa, and K. Hirose:“Para-linguistic information represented as distortion of the acoustic universal structure in speech,” *Proc. ICASSP*, vol. 1, pp.261-264, 2006.

[52] 北村達也, 齋藤毅: “単母音の音響特徴量の変化が個人性知覚に与える影響,” 電子情報通信学会技術研究報告, SP-2006-167, 2007.

[53] 齋藤毅, 北村達也: “3 連続母音に含まれる個人性の知覚に寄与する音響特徴量,” 電子情報通信学会技術研究報告, SP2006-166, 2007.

[54] H. Zen and T. Toda: “An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005,” *Proc. INTERSPEECH,* pp.93–96, 2005.

[55] A. W. Black and K. Tokuda: “The Blizzard Challenge —2005: evaluating corpus-based speech synthesis on common datasets,” *Proc. INTERSPEECH,* pp.77–80, 2005.

[56] C. L. Bennett: “Large scale evaluation of corpus-based synthesizers: results and lessons from the Blizzard Challenge 2005,” *Proc. INTERSPEECH,* pp.105–108, 2005.

[57] 朝川智, 峯松信明, 広瀬啓吉: “音声の構造的表象に基づく英語学習者発音の音響的分析,” 電子情報通信学会論文誌, vol. J90-D, no.5, pp.1249–1262, 2007.

[58] 村上隆夫: “音声の構造的表象を用いた音声認識に関する基礎的研究,” 東京大学情報理工学系研究科 修士論文, 2006.

[59] 上田修功:“ベイズ学習 [I], [II], [III], [IV],” 電子情報通信学会誌, vol. 85, no. 4 (pp. 265–271), no. 6 (pp. 421–426), no. 7 (pp.504–509), no.8 (pp.633–638), 2002.

[60] C. H. Lee, C. H. Lin and B. H. Juang: “A study on speaker adaptation of the parameters of continuous density hidden Markov models,” *IEEE Trans. Signal Processing,* vol. 39, no. 4, pp. 806–814, 1991.

[61] 金谷健一: “これなら分かる最適化数学—基礎原理から計算手法まで,” 共立出版, 2005.

[62] ダニエル・アリホン:“映画の文法,” 紀伊國屋書店, 1980.

発表文献

[1] 峯松信明, 西村多寿子, 朝川智, 櫻庭京子, 齋藤大輔: “要素論から全体論へ　～全体から入る音声情報処理への招待～,” 情報処理学会音声言語情報処理研究会, 2007–SLP–67–14, pp.75–80, 2007.

[2] 峯松信明, 西村多寿子, 櫻庭京子, 朝川智, 齋藤大輔: “孤立音 [あ] を聞いて/あ/と同定する能力は音声言語に必要か？,” 電子情報通信学会音声研究会, SP2007–30, pp.37–42, 2007.

[3] 齋藤大輔, 朝川智, 峯松信明, 広瀬啓吉: “音声の構造的表象を入力とした音声合成に対する基礎的検討,” 日本音響学会秋季講演論文集, 1–P–2, pp.399–402, 2007.

[4] 竹内京子, 齋藤大輔, 峯松信明, 広瀬啓吉: “フランス語鼻母音における鼻音性の知覚,” 日本音響学会秋季講演論文集, 1–2–7, pp.497–498, 2007.

[5] 齋藤大輔, 朝川智, 峯松信明, 広瀬啓吉: “デルタケプストラムの声道長依存性に関する実験的検討,” 日本音響学会秋季講演論文集, 3–3–13, pp.169–170, 2007.

[6] 齋藤大輔, 朝川智, 峯松信明, 広瀬啓吉: “構造的表象からの音声生成に関する基礎的検討,” 電子情報通信学会音声研究会, SP2007–80, pp.55–60, 2007.

[7] 齋藤大輔, 松浦良, 朝川智, 峯松信明, 広瀬啓吉: “ケプストラムの声道長依存性に関する幾何学的考察,” 電子情報通信学会音声研究会, SP2007–128, pp.189–194, 2007.

[8] 齋藤大輔, 松浦良, 鎌田敏明, 朝川智, 峯松信明, 広瀬啓吉: “ケプストラムの声道長依存性に対する定量的分析とその応用,” 日本音響学会春季講演論文集, 1–Q–19, 2008（発表予定）.

[9] 竹内京子, 齋藤大輔, 峯松信明, 広瀬啓吉: “フランス語鼻母音のカテゴリー知覚の考察,” 日本音響学会春季講演論文集, 3–7–2, 2008（発表予定）.

[10] 齋藤大輔, 朝川智, 峯松信明, 広瀬啓吉: “構造的表象からの音声合成における音響空間探索の高速化,” 日本音響学会春季講演論文集, 3–Q–16, 2008（発表予定）.

[11] D. Saito, R. Matsuura, S. Asakawa, N. Minematsu, K. Hirose: “Directional dependency of cepstrum on vocal tract length,” Proc. Int. Conf. Acoustics, Speech, & Signal Processing (ICASSP’2008), 2008 (to appear).

[12] N. Minematsu, T. Nishimura, D. Saito, S. Asakawa, Y. Qiao: “Holistic and prosodic representation of the segmental aspect of speech,” Proc. Int. Conf. Speech Prosody, 2008 (to appear).

[13] 齋藤大輔, 朝川智, 峯松信明, 西村多寿子, 広瀬啓吉: “音声の不変表象に基づく語ゲシュタルトの物理的解釈とそれに基づく幼児の音声模倣の実装,” 人工知能学会第 22 回全国大会, 2008（発表予定）.

付録 A

多次元ニュートン法を用いた音響事象の高速推定

A.1　1 変数のニュートン法

いま関数 $f(x)$ が与えられているとき，$f(x)=0$ を満たす x を求めることを考える．ただし x はスカラーである．真の解に近い初期値 x_0 を与え，x_i を以下の式で更新する．

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)} \quad (i=0,1,\ldots) \tag{A.1}$$

ただし $f'(x)$ は $f(x)$ の導関数である．このとき十分に値の小さい収束判定指数 ε に対して，以下を満たすまで式 (A.1) を繰り返し計算する．

$$|x_{i+1} - x_i| \leq \varepsilon \tag{A.2}$$

このとき x_i は真値に 2 次収束することが知られている．

A.2　多次元ニュートン法による高次連立方程式の解法

未知数が n 個存在する，以下の高次連立方程式を考える．

$$f_i(x_1, x_2, \ldots, x_n) = 0 \quad (i=1,\ldots,n) \tag{A.3}$$

このとき，式 (A.1) を多次元に拡張する．以下表記の都合上，変数の右肩に $x_j^{(\nu)}$ の形で反復回数 ν を記す．更新する未知数をベクトルと考え，$\boldsymbol{x}=(x_1,\ldots,x_n)$ とする．さらに更新ベクトルを $\Delta\boldsymbol{x}=(\Delta x_1,\ldots,\Delta x_n)$，関数群 $f_i(\boldsymbol{x})$ をベクトル $\boldsymbol{F}(\boldsymbol{x})=(f_1(\boldsymbol{x}),\ldots,f_n(\boldsymbol{x}))$ で表すとき，式 (A.1) は以下のように拡張される．

$$\begin{cases} \boldsymbol{x}^{(\nu+1)} & = \boldsymbol{x}^{(\nu)} - \Delta\boldsymbol{x}^{(\nu)} \\ \boldsymbol{J}(\boldsymbol{x}^{(\nu)})\Delta\boldsymbol{x}^{(\nu)} & = \boldsymbol{F}(\boldsymbol{x}^{(\nu)}) \end{cases} \tag{A.4}$$

ただし $\boldsymbol{J}(\boldsymbol{x})$ はヤコビ行列とよばれ，その (i,j) 要素が以下の値となる．

$$J_{ij} = \frac{\partial f_i}{\partial x_j} \tag{A.5}$$

このとき十分に小さい収束判定指数 ε に対して以下の条件を満たすまで式 (A.4) を繰り返し計算する．

$$|\boldsymbol{x}^{(\nu+1)} - \boldsymbol{x}^{(\nu)}| \leq \varepsilon \tag{A.6}$$

A.3　ブロックサイズ 2 における音響事象の推定

A.3.1　更新式の導出

式 (7.5) について c_x を x に，c_y を y に置き換え，変形すると以下のようになる．

$$\begin{cases} \dfrac{1}{4(V_x+V_{ax})}(x-a_x)^2 + \dfrac{1}{4(V_y+V_{ay})}(y-a_y)^2 - BD_a + \dfrac{1}{2}\ln\dfrac{|(\Sigma+\Sigma_a)/2|}{|\Sigma|^{\frac{1}{2}}|\Sigma_a|^{\frac{1}{2}}} = 0 \\ \dfrac{1}{4(V_x+V_{bx})}(x-b_x)^2 + \dfrac{1}{4(V_y+V_{by})}(y-b_y)^2 - BD_b + \dfrac{1}{2}\ln\dfrac{|(\Sigma+\Sigma_b)/2|}{|\Sigma|^{\frac{1}{2}}|\Sigma_b|^{\frac{1}{2}}} = 0 \end{cases} \tag{A.7}$$

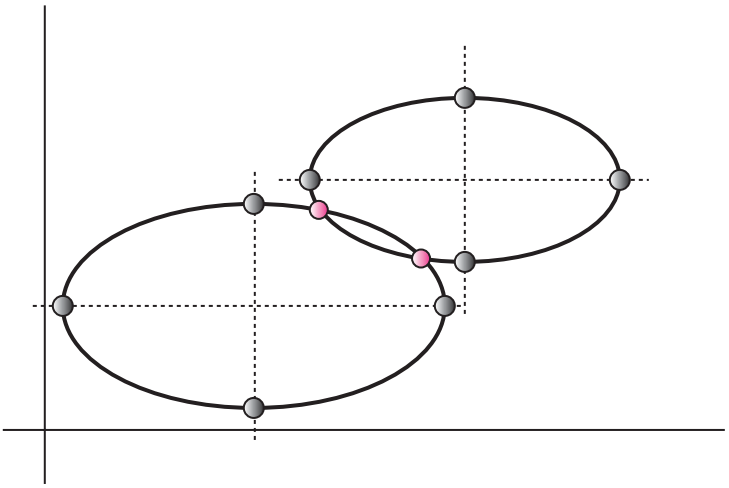


図 A.1: 初期値の設定

今，式 (A.7) の第 1 式左辺を $f(x,y)$，第 2 式左辺を $g(x,y)$ とする．それぞれ簡易的に f,g と表記し，偏導関数を f_x, f_y および g_x, g_y とすると更新に必要な関数はそれぞれ以下のようになる．

$$f_x = \frac{1}{2(V_x + V_{ax})}(x - a_x) \tag{A.8}$$

$$f_y = \frac{1}{2(V_y + V_{ay})}(y - a_y) \tag{A.9}$$

$$g_x = \frac{1}{2(V_x + V_{bx})}(x - b_x) \tag{A.10}$$

$$g_y = \frac{1}{2(V_y + V_{by})}(y - b_y) \tag{A.11}$$

このとき式 (A.4) の第 2 式を $\Delta\boldsymbol{x}$ について解くことで更新ベクトルを求める．二次元の場合，ヤコビ行列の逆行列と f,g を用いて更新することができる．すなわち更新ベクトル $(\Delta x, \Delta y)$ は以下のようになる．

$$\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \frac{1}{f_x g_y - f_y g_x} \begin{pmatrix} g_y & -f_y \\ -g_x & f_x \end{pmatrix} \begin{pmatrix} f \\ g \end{pmatrix} \tag{A.12}$$

A.3.2　初期値の設定

適切な初期値の設定について述べる．このとき以下の 4 つの方程式をそれぞれ解き，それぞれの (x,y) を求める．

$$f(x, a_y) = 0 \tag{A.13}$$

$$f(a_x, y) = 0 \tag{A.14}$$

$$g(x, b_y) = 0 \tag{A.15}$$

$$g(b_x, y) = 0 \tag{A.16}$$

上記の操作は図 A.1において，各楕円とその長軸，短軸の交点を設定していることに他ならない．この時，より解に近い点，すなわちもう一方の楕円に近い点を選択すれば解はおよそ収束する．