修 士 論 文

スペクトルと基本周波数のイベント操作による音声モーフィングに関する研究

指導教官　赤木 正人 教授
審査委員主査　赤木 正人 教授
審査委員　小谷 一孔 助教授
審査委員　嵯峨山 茂樹 教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

910100
藤野 善行

2001 年 2 月

要 旨

人間と機械の接する場面が多くなってきて、人間にとって自然で柔軟性のあるインターフェースが強く望まれるようになってきた。音声はこれからのインターフェースの重要な要素であり、とりわけ音声の声質を制御する音声モーフィングは実用化が望まれるマンマシンインターフェース技術の一つである。

音声モーフィングに関する研究は幅広く行なわれている。これらの手法では、様々なモーフィング音声を作成するためには、元音声だけでなくモーフィング対象話者（目標話者）音声に関する膨大な量の音響的特徴量を含んだ音声データ（フルサイズデータ） が必要となり、その処理も複雑である。このことから、目標話者のフルサイズデータが入手困難な場合に有効な、すなわち目標話者の少量の音声データ（サブセットデータ）による頑健な音声モーフィングが強く望まれている。

本研究では、任意の話者に関するフルサイズデータが存在すると仮定し、それらを目標話者に関するサブセットデータを用いて音声モーフィングを行なう。また、用いたモーフィングパラメータがモーフィング音声に与える影響を検討する。

目 次

図 目 次

表 目 次

第 1 章

序論

音と人間は切っても切れない関係にある。人類の誕生から現代にいたるまで、また未来永劫に、人は音と共に生きる存在である。この音の歴史の新しいページを開いたのは、言葉の発明である。音声が生まれ、音楽が生まれ、音のコミュニケーションの新たな時代が始まった。現在においても、我々は日常のコミュニケーションの大半を音声を介して行なっている。この意味でも、音声は人間にとってもっとも親しみのある自然なメディアである。

人間と機械の接する場面が多くなってきて、人間にとって自然で柔軟性のあるインターフェースが強く望まれるようになってきた。音声はこれからのインターフェースの重要な要素であり、それは音声メディアの利用・処理技術にかかっているといえる。とりわけ、音声合成技術は比較的古くから応用されてきた。しかし、現在の合成音は機械的で無味乾燥であるとよくいわれる。特に個人性のある音声合成、とりわけ音声の声質を制御する音声モーフィングは実用化が望まれるマンマシンインターフェース技術の一つである。

本論文では、音声の生成機構に着目し、音声の調音構造に基づいたモデルによる音声モーフィングを試みる。特に、目標音声に関する少ない情報でのモーフィングを行ない、モーフィング音声とモーフィングパラメータとの関連を検討する。

1.1 背景

音声モーフィングとは

モーフィングとはコンピュータグラフィックスなど画像処理の分野で用いられている技術であり、映像中に描かれた物体 A を物体 B へと変形させるものである。これを音声に応用したものが「音声モーフィング」であるとされている。

音声モーフィングといってもその対象とする音響的特徴量によって様々な見方ができるが、本論文では、個人性・話者性に着目し、ある話者の音声の声質を別の話者の音声の声質に変化させる、複数話者間における声質変換を音声モーフィングと定義する。

音声の個人性・話者性に関する音響的特徴量については、まだ十分に解明はされていないが、音声波に個人差が生じるには、おおきく 2 つの原因があると考えられている。1 つめは、音声生成のための音声器官の解剖学的構造の差に基づく先天的なもの、2 つめは言語の習得過程で身につけた発声習慣、すなわち音声器官の動かし方の差に基づく後天的なものである。前者は主として音声の周波数構造上の個人差、つまりスペクトル包絡の異なりとして現れる。後者は主として音声の周波数構造の時間的変化つまりイントネーションやアクセントの差である。音声モーフィングでは、これら音響的特徴量を制御することが必要となってくる。

このような声質制御[1]機能を備えた音声合成システムの実現は、音声合成システム自体の普及のために非常に重要であると同時に、話者間での音声翻訳システムにおける話者適応技術にも非常に有用である。また、合成音声の多様化の研究は近年、声質変換や様々な発話様式音声分析・合成などを中心に盛んになってきている。特に声質変換など音声モーフィングによる話者性・個人性制御の解明は最も重要な知見の一つであり、このことは合成音声の多様化だけではなく、話者認識や音声知覚などの分野にも貢献をもたらすと考えられる。

音声モーフィングに関する過去の研究

これまでの音声モーフィングに関する研究は幅広く行なわれている。阿部 [1] は基本周波数とスペクトルの漸次変形による音声モーフィングを行なっている。この手法は、ある周波数を境に低域と高域のスペクトルの入れ替えによりモーフィングを行なっているものである。小坂 [2] は 定常な楽音や音声（母音）に Sinusoidal model を用いた音色補間を行なっている。また、滑らかで自然に変化するという条件を満たすモーフィング方式として、坂野ら [3] は包絡と音源の独立操作による手法を提案している。この手法では、周波数軸の非線型伸縮によりスペクトル包絡間の対応付けを行なっている。土屋ら [4] は包絡成分の補間方法として大室らが提案した積分スペクトル逆関数（IFIS）[5] を用いた音声モーフィングを行なっている。

また、ベクトル量子化による話者適応化法に基づいた音声モーフィングに関する研究も行なわれている [6]。これらは学習と変換の 2 つのステップから構成されており、学習の

[1] ここでの声質制御とは目標とする音声へと自由に声質変換を行なうことを意味する。

ステップで変換コードブックを作成し、変換のステップで作成された変換コードブックによりモーフィングを行なっている。

これらの手法では、様々なモーフィング音声を作成するためには、元音声だけでなくモーフィング対象話者（目標話者）音声に関する膨大な量の音響的特徴量を含んだ音声データ（フルサイズデータ）が必要となり、その処理も複雑である。また、これらの手法によるモーフィング音声は 2 話者のフルサイズデータ間を補間することによって得られる音声にとどまっており、声質制御が十分にできているかというと疑問である。ベクトル量子化による音声モーフィングでは、少量の学習データ用音声でのモーフィングが可能ではあるが、スペクトルの類似性や、音声の文脈上の一致などの制約を重視したものである。

以上のことから、目標話者のフルサイズデータが入手困難な場合に有効な、すなわち目標話者の少量の音声データ（サブセットデータ）による頑健な音声モーフィングが強く望まれている。目標話者のサブセットデータによる任意の話者のフルサイズデータの音声モーフィングが可能となれば、様々な音韻に対するモーフィング音声を生成することが容易となり、入力側の話者の声質を保ったまま翻訳後の文章を発声することが要求される自動翻訳電話などにも応用できる。また、サブセットデータの組み合わせによっては全く新たな音声を生成することも可能となり、幅広い声質制御も期待できる。

1.2 本研究の目的

本研究では、任意の話者に関するフルサイズデータが存在すると仮定し、それらを目標話者に関するサブセットデータを用いて目標話者への音声モーフィングを行なう。また、用いたモーフィングパラメータがモーフィング音声に与える影響を検討する。図 1.1にその概略図を示す。

ここで用いるサブセットデータは、あまりにそのデータ量が大きくなるとサブセットの意味が無くなってしまう。いかに少ない情報量で目標話者へとモーフィングを行なうかが重要である。

音声の声質は主にスペクトル包絡と基本周波数で表現される。また、音声の個人性は子音よりも母音に多く現れると言われている。よって、本研究では用いるモーフィングパラメータとして母音スペクトルと基本周波数に関連したパラメータを用いる。そのためには、1. 音声データをスペクトル構造成分と基本周波数成分に分離、かつそれらを独立に制御する必要がある。さらに、2. 母音スペクトルを独立に制御するには、得られたスペクトル構造を子音スペクトル成分と母音スペクトル成分に時間分解する必要がある。つまり、音声の生成機構に着目し、音声の調音構造に基づいたモデルを用いる必要がある。

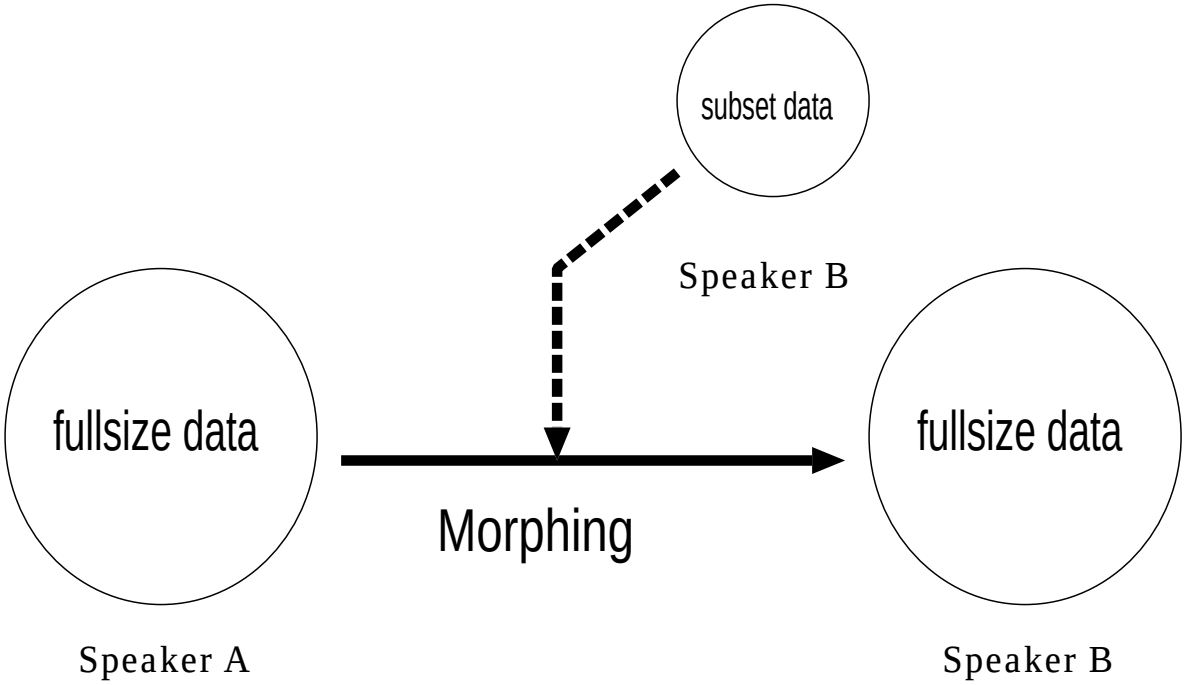

図 1.1: モーフィングの概略図

以上の条件を満たす手法として、本研究では音声分析変換合成法 STRAIGHT [7] と S^2BEL-TD [8] を用いて音声を分解・合成する。STRAIGHT により音声データをスペクトル成分と基本周波数成分に分離することが可能であり、基本周波数成分を独立に制御することが期待できる。さらに S^2BEL-TD ではスペクトル成分を子音スペクトル成分と母音スペクトル成分に分離することが可能であり、母音スペクトル成分を独立に制御することが期待できる。

最後に聴取実験を行ない、得られたモーフィング音声と用いたモーフィングパラメータとの関連を検討する。

1.3 本論文の構成

本論文の構成を以下に示す。

第 1 章では、本論文が対象としている音声モーフィングに関する研究分野の現状と問題点を指摘し、本論文の目的を明らかにする。

第 2 章では、本論文で用いる音声分析合成系 STRAIGHT と S^2BEL-TD の構造とその

有用性を説明する。

第 3 章では、音声モーフィングに用いるモーフィングパラメータを説明・検討し、本研究で用いるサブセットデータとしてのモーフィングパラメータを決定する。

第 4 章では、聴取実験を通じてモーフィング音声とそれに用いたモーフィングパラメータとの影響を調べる。

第 5 章では、全体の考察を行ない、第 6 章にて本論文で得られた結果を要約し、今後の課題を示す。

第 2 章

音声分析合成系

2.1 目的

本研究では、主に母音スペクトルと基本周波数を中心とした個人性情報を取り扱っていく。そのような合成音声を作成するためには、音声から個人性情報を抽出し、また母音成分の情報を独立に扱える音声分析合成系を用いる必要がある。このための音声分析合成系として、音声をスペクトルと基本周波数に分解でき、高品質な合成音声を合成できる STRAIGHT（Speech Transformation and Representation basedon Adaptive Interpolation of weiGHTed spectrogram）[7] を、スペクトルを子音成分と母音成分に分解できるテンポラルデコンポシション S^2BEL-TD (Spectral Stability Based Event Localizing Temporal Decomposition)[8] を採用する。

本章では、その構造を説明する。

2.2 STRAIGHT

STRAIGHT は STRAIGHT-core、SPIKES、TEMPO2 の 3 つ主要な部分から構成されている。

STRAIGHT-core は、音声の励振の周期性による干渉の影響のない時間周波数表現を抽出する方法である。その中心的なアイディアは基本周期、基本周波数を節点とする区分的線形関数による補間と等価な時間周波数領域の平滑化を行なうことにある。

SPIKES は、合成に用いる駆動音源の位相[1]特性を操作することにより、VOCODER 特有の buzzy な音色を軽減する方法である。ここでは、同一のパワースペクトルであっても

[1]正確には群遅延

群遅延を操作して時間的な微細構造を変えることで音色が変化することを利用している。

TEMPO2 は、2 つのフィルタ出力の微分の特性を基に、音声の基本周波数を推定する方法である。特別なフィルタ設計と搬送対雑音比（C/N 比）の組み合わせにより、基本周波数の推定が正確なものになっている。

2.2.1 STRAIGHT-core

STRAIGHT-core の重要なアイディアは、有声音に見られる周期的な励振を、直接には観測できない仮想的な時間周波数曲面を時間周波数領域で組織的にサンプリングする役割を担うものであると解釈するところにある。STRAIGHT-core の原理は、この解釈の下、サンプリングされた限られた局所的情報から曲面を復元するために、2 次の cardinal B-spline の基底関数を平滑化関数として用いていることにある。ここで、基底関数を補間関数ではなく平滑化関数として用いることで、雑音を誤差に強い形で周期性の影響を選択的に除去することを担っている。実際、後で説明する TEMPO2 の結果と併せると、STRAIGHT-core で求められる有声音のスペクトルは、雑音源で駆動される場合に比べてけた違いに小さな誤差を有することが示されている。

信号モデル

音声を、常に周波数の変動する基本波とそれにほぼ同期したイベントに駆動される高次の周波数成分からなる信号であると考える。

$$s(t) = \sum_{k \in N} \alpha_k(t) \sin\left(\int_{t_0}^{t} k \left(\omega_0(\tau) + \omega_k(\tau) \right) d\tau + \phi_k \right) \tag{2.1}$$

ここで、$\omega_0(t)$ は、基本波の角周波数、$\omega_k(t)$ は、k 番目の高次調波成分の角周波数を表わす。また、$\alpha_k(t)$ は、それぞれの成分の強さを表わし、ϕ_k は、k 番目の高次調波成分の初期位相を表わす。この信号の短時間フーリエ変換は、調波構造と調波間の干渉のため、周波数方向に $f_0(t) = \omega_0(t)/2\pi$、時間方向に $\tau_0 = 1/f_0$ のほぼ周期的な構造を有することになる。

時間方向の位相干渉の効果の軽減

実行的な長さが 1 基本周期上でサイドローブが十分に軽減しているような時間窓を用いれば、分析位置による短時間スペクトルの変動の解析は、隣接する調波の相互作用を考えるだけで良い。例えば、次のように定義される Gauss 型時間窓は、そのような窓の一

例である。ここで η は、窓の時間方向の伸長の程度を示すパラメータである。

$$w_G(t) = e^{-\pi(\frac{t}{\eta t_0})^2} \sin\left(\pi \frac{t}{t_0}\right) \tag{2.2}$$

このような窓を用いて周期信号を分析すると、周期的にパワースペクトルが零となる部分が出現する。この零となる部分を埋めて時間的に変動しないパワースペクトルを得ることが最初のステップである。パワースペクトルが零となるのは、調波と調波の中間の周波数で上の調波の位相を π だけ回転させるように作った相補的な窓 $\omega_c(t)$ を用いて計算した短時間スペクトルが零の部分で最大値をもつことと等価である。

$$\omega_c(t) = \omega(t) \sin\left(\pi \frac{t}{\tau_0}\right) \tag{2.3}$$

時間方向に伸長した時間窓（$\eta > 1$）で得られたスペクトル $P_0(\omega, t)$ とその相補的な窓から求められたスペクトル $P_c(\omega, t)$ とを、次のような加重和として合成することにより、時間方向での周期的変動のないスペクトル $P_\tau(\omega, t)$ が求められる。

$$P_\tau(\omega, t) = \sqrt{P_0^2(\omega, t) + \xi(\eta) P_c^2(\omega, t)} \tag{2.4}$$

ここで、$\xi(\eta)$ はスペクトルの時間方向の分散を最小にする混合係数である。なお、時間方向に少し引き延ばすだけで、$P_\tau(\omega, t)$ の時間方向の周期的変動は実質的に無視することができる。

周波数方向の平滑化

基本周波数に応じて適応的に変化する次のような 2 次の cardinal B-spline 基底関数 $h_t(\omega)$ を周波数方向の平滑化関数とする。

$$h_t(\omega) = 1 - \left| \frac{\omega}{\omega_0(t)} \right| \tag{2.5}$$

ここで、$\omega_0(t) = 2\pi f_0(t)$ であり、$-\omega_0(t) \leq \omega \leq \omega_0(t)$ である。$P_r(\omega, t)$ をこの平滑化関数を用いて次式により平滑化することで、周期的な励振の影響が除かれた時間周波数表現 $S(\omega, t)$ が得られる。

$$S(\omega, t) = \sqrt{g^{-1}\left(\int_D h_t(\lambda, t) g(|P_r(\omega - \lambda, t)|^2) d\lambda\right)} \tag{2.6}$$

ここで D は、平滑化関数の定義域を表わす。式（2.6）の中の $g()$ は、平滑化操作によって保存すべき量を定めるのに利用される。

最適な平滑化関数

前説で説明した原理を直接適用しただけでは、再合成音の品質はあまり良くない。これは、時間窓による周波数方向の平滑化と平滑化関数 $h_t(\omega)$ による平滑化が重なることにより、過剰な平滑化が行なわれてしまうためである。最適平滑化関数は、spline 関数の性質を利用すると、窓関数の周波数表現と 2 次の cardinal B-spline 基底関数の畳み込みを基本周波数の間隔で標本化した系列をインパルス応答とみなしたときの逆フィルタの対応を計算することで求めることができる。

2.2.2 SPIKES

SPIKES のアイディアは、パワースペクトルに変化を与えずに時間構造を制御するため、オールパスフィルタを用いたことと、オールパスフィルタの位相（群遅延特性）を三角関数と周波数重みによってモデル化したことにある。このモデルを用いることで、見通しの良い時間構造の操作が可能となる。

音声の再合成

STRAIGHT-core により求められた時間周波数表現から音声を再合成する方法として、複素ケプストラムを介して最小位相インパルス応答を求め、位相調整して再配置する方法について説明する。このような方法を用いることにより、基本周波数 $f_0(t)$ の精密な制御と、時間的微細構造に依存する音色の制御が可能となる。

この方法よる音声の変換と合成は、形式的には次のように表わすことができる。合成された音声波形を $y(t)$ とする。

$$y(t) = \sum_{t_i \in Q} \frac{1}{\sqrt{G(f_0(t_i))}} v_{ti}(t - T(t_i)) \tag{2.7}$$

$$v_{ti}(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} V(\omega, t_i)\Phi(\omega) \exp(j\omega(t)) d\omega \tag{2.8}$$

$$\text{where} \quad T(T_i) = \sum_{t_k \in Q, k<i} \frac{1}{G(f_0(t_k))} \tag{2.9}$$

ここで、Q は合成のための駆動信号を置く位置の集合であり、$G()$ は基本周波数の変換を表わす。オールパスフィルタ $\Phi(\omega) = \Phi_1(\omega)\Phi_2(\omega)\Phi_3(\omega)\Phi_4(\omega)$ の特性を次節で説明するように操作することにより、基本周波数と音色が制御される。

また、$A()$、$u()$、$r()$ をそれぞれ振幅、周波数、時間軸の変換としたとき、$V(\omega, t_i)$ は変換された振幅スペクトル $A((S(u(\omega), r(t)), r(t))$ から次のようにして複素ケプストラム

$h_t(q)$ を介して求めた、最小位相インパルス応答のフーリエ変換である。ただし、q はケフレンシーを表わす。

$$V(\omega, t) = \exp\left(\frac{1}{\sqrt{2\pi}}\int_0^\infty h_t(q)\exp(j\omega q)dq\right) \tag{2.10}$$

$$h_t(q) = \begin{cases} 0 & (q < 0) \\ c_i(0) & (q = 0) \\ 2c_i(q) & (q > 0) \end{cases} \tag{2.11}$$

$$\text{and} \quad c_i(q) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\exp(j\omega q)\log A d\omega \tag{2.12}$$

オールパスフィルタの設計

オールパスフィルタの第 1 成分 $\Phi_1(\omega)$ は、次式で表わされる時間遅れに相当する直線位相成分である。$\Phi_1(\omega)$ は、基本周波数の精密な制御に用いられる。

$$\Phi(\omega) = \exp(j\omega f_s T_d) \tag{2.13}$$

ここで、$\omega = 2\pi f/f_s$ は正規化角周波数を表し、T_d は時間遅れを表す。離散時間系での実装にあたっては、n の任意の整数とすると、正規化角周波数が 2π のときの値が $2n\pi$ であるとの拘束条件を満たすように、次式を用いる。

$$\Phi_1(\omega) = \exp\left(j\left(\pi f_s T_d + p(\omega)\right)\right) \tag{2.14}$$

$$P(\omega) = \begin{cases} \frac{2\pi a}{1+\exp((\omega+\pi)/\omega_\omega)} & \omega \geq 0 \\ \frac{2\pi a}{1+\exp((\omega-\pi)/\omega_\omega)} & \omega < 0 \end{cases} \tag{2.15}$$

$$a = \lceil f_s T_d \rceil - f_s T_d \tag{2.16}$$

この式では、ナイキスト周波数での位相の不連続を、指数関数を利用して滑らかに接続することにより、特異点の影響の時間領域での局在化を図っている。$f_\omega = f_s\omega_\omega/2\pi$ は、位相の不連続を滑らかにつなぐ区間（遷移帯域）の幅を表す。

なお、$\Phi_2(\omega)$、$\Phi_3(\omega)$、$\Phi_4(\omega)$ についても、$\Phi_1(\omega)$ の場合と同様にナイキスト周波数において位相が連続になるように補正を行なっている。

2.2.3 TEMPO2

TEMPO2 は、帯域フィルタの中心周波数とフィルタ出力の瞬時周波数を周波数から周波数への写像とみなし、信号の主要な正弦波成分の周波数を、このような写像の安定な平衡点に対応する瞬時周波数として求める方法である。

基本周波数推定のために瞬時周波数を使うには、推定に先立って分離され選択される基本波成分が必要である。これは、log 周波数軸に沿って等しい間隔を保つフィルタからなる帯域通過フィルタ、特別に設計されたインパルス応答と選択機構によって行なわれる。フィルタのインパルス応答 $\omega_s(t,\lambda)$ は、STRAIGHT-core で示したガボール関数（式 2.2）と cadinal B-spline 基底関数（式 2.5）の畳み込みによって得られる。

フィルタの中心周波数 λ からフィルタ出力の瞬時周波数 $\omega_c(t_i,\lambda)$ へ等しく写像される点（不動点）の集合 $A(t)$ は次のように定義される。

$$A(t) = \{\lambda | \omega_c(t_i,\lambda) = \lambda, \omega(t_i,\lambda-\epsilon) - (\lambda-\epsilon) > \omega_c(t_i\lambda+\epsilon) - (\lambda+\epsilon)\} \tag{2.17}$$

ϵ は任意の小さな定数を表す。さらに、$\Lambda(t)$ から、基本周波数に対応する点を選択しなければならない。STRAIGHT では C/N 比を推定し、これが最も低い不動点を用いることにより基本周波数を推定している。

STRAIGHT を用いる理由

STRAIGHT を用いる理由は、文音声からの物理的特徴のスペクトル・基本周波数を得るためである。スペクトルは STRAIGHT-core により、基本周波数は TEMPO2 により計算される。これらは独立に操作・制御が可能である。さらに、SPIKES により高品質な合成音声を生成できることから STRAIGHT を採用した。本論文で用いたモーフィングによる合成音は、全て STRAIGHT により作成した。

2.3 スペクトルパラメータ（LSF）

STRAIGHT で得られたスペクトルをスペクトルパラメータに変換し、S^{2}BEL-TD を用いてスペクトルパラメータを時間変化パターン（イベント関数）とスペクトルの安定する位置におけるスペクトル情報（イベントターゲット）に分解する。S^{2}BEL-TD で分解するスペクトルパラメータとして、以下の理由により LSF（Line Spectral Frequencies）を用いる。

- LSF はその性質が線形補完性に優れており、また TD（テンポラルデコンポジション）で用いられているスペクトルパラメータのうち、より歪みが少なく再現性がよいパラメータとして報告されていることから、スペクトルパラメータとして LSF を用いることにする。

スペクトルから LSF へ変換する方法は以下の通りである。

1. STRAIGHT で得られるパワースペクトル

   STRAIGHT で得られる振幅スペクトル $X[k]$、$0 \leq k \leq N-1$ を用いてパワースペクトル $S[k]$ を計算する。

$$S[k] = |X[k]|^2, \quad 0 \leq k \leq N-1 \tag{2.18}$$

2. 相関関数の導入

   パワースペクトルからフーリエ逆変換することによって相関関数を求めると

$$R[n] = \frac{1}{N} \sum_{k=0}^{N-1} S[k] \exp\left\{ j \frac{2\pi kn}{N} \right\} \tag{2.19}$$

   となる。今、この相関関数を有する過程 $x(t)$ が全極型のフィルタ（次数 L）からの出力と仮定すれば、フィルタの係数を $\{a_l^L\}, l = 1, 2, \ldots, L, 0 < L < N/2$ として、

$$P_L = R[0] - \sum_{l=1}^{L} a_l^L R[l] \tag{2.20}$$

   と書ける。ここで、P_L は誤差である。P_L が最小となるようにフィルタの係数 $\{a_l^L\}, l = 1, 2, \ldots, L$ を決める。このときのフィルタ係数 $\{a_l^L\}, l = 1, 2, \ldots, L$, は LPC の予測係数と一致する。

3. LSF へ

   予測係数 $\{a_l^L\}, l = 1, 2, \ldots, L$ を用いて、次のような Z^{-1} の多項式を作る。

$$\begin{cases} A_L(Z) = 1 - \sum_{l-1}^{L} a_l^L Z^{-1} \\ B_L(Z) = Z^{-(L+1)} A_L(Z^{-1}) \end{cases} \tag{2.21}$$

   これを用いれば、LSF が計算できる。

   LSF への計算は式（2.21）より、

$$P(z) = A_L(z) - B_L(x) \tag{2.22}$$

$$Q(z) = A_L(z) + B_L(x) \tag{2.23}$$

   となる。ここで、$P(z)$、$Q(z)$ はそれぞれ反対称な係数、対称な係数をもつ $(p+1)$ 次の多項式を表している。式（2.21）から、L を偶数と仮定すると $P(z)$、$Q(z)$ はそれぞれ次のように因数分解される。

$$P(z) = (1 - z^{-1}) \Pi_{i=2,4,\cdots,l} (1 - 2z^{-1} \cos \omega_i + z^{-2}) \tag{2.24}$$

$$Q(z) = (1 + z^{-1}) \Pi_{i=1,3,\cdots,l-1} (1 - 2z^{-1} \cos \omega_i + z^{-2}) \tag{2.25}$$

ただし、ω_i は以下の関係を満たすように順序付けるものとする。

$$0 < \omega_1 < \omega_2 < \cdots < \omega_l - 1 < \omega_l$$

この因数分解に現れる係数 $\omega_1, \omega_2, \cdots, \omega_l$ を LSF と呼び、これらを用いることにより LSF 上での補間が可能となる。なお、LSF 係数の次数は 30 次を用いた。これは、対象とする音声データのサンプリング周波数が 8 kHz であるためである。

2.4 S^{2}BEL-TD

音声の個人性は子音よりも母音にあると言われており、また音声はそのほとんどが母音によって支配されている。以上のことをふまえ、本論文では母音成分におけるスペクトルに着目し、その操作による音声モーフィングを試みる。ここで、音声中の母音成分に関するスペクトルパラメータを独立に扱えるように音声を分解する必要があるが、そのためには音声を構成する調音構造に基づいたモデルを用いる必要がある。これらの要求を満たすモデルとして、本研究では S^{2}BEL-TD（Spectral Stability Based Event Localizing Temporal Decomposition）[8] を採用する。

以下、S^{2}BEL-TD の構造を説明する。

2.4.1 TD モデル

得られるスペクトルパラメータ $\boldsymbol{y}(n)$ の時間変化の様子を時間の関数で表わすことを考える。与えられた発話中には K 個のイベントがあるとすると、それらは k 番目のイベントターゲット $\boldsymbol{a}_k$ とそのイベントの時間変化を表わすイベント関数 $\phi_k(n)$ で記述できる。

$$\hat{\boldsymbol{y}}(n) = \sum_{k=1}^{K} \boldsymbol{a}_k \phi_k(n), \quad 1 \leq n \leq N \tag{2.26}$$

式（2.26）を行列表示すると、以下のようになる。

$$\hat{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{\Phi} \quad \hat{\boldsymbol{Y}} \in R^{P \times N}, \boldsymbol{A} \in R^{P \times K}, \boldsymbol{\Phi} \in R^{K \times N}$$

ここで、P はスペクトルパラメータの次数である。式（2.26）において、イベントターゲットとイベント関数は未知であり、Temporal Decomposition 分析はこれらを決定していくことでもある。

発話中における各イベントは、時間と共に徐々に増加、減少していき、それらは隣同士重なりあう。よって、時間変化パターンを表わすイベント関数には以下の特性が考えられる。

- 各イベントには始まりと終りの時間が存在する、すなわち時間間隔が存在する
- 各イベントはその存在期間においては非負で表わせる
- 各イベントは実際の発話における音声生成と同様、ゆるやかな増加・減少で表わせる

以上のポイントを数学的に制限することにより、イベント関数を決定していく。

S^{2}BEL-TD の計算ステップは以下のようになる。まず、S^{2}BEL-TD のステップを以下に示す。

1 イベントターゲットの決定

$$\boldsymbol{A}^{(0)} = \left[\boldsymbol{a}_k^{(0)}\right]_{1\leq k\leq K}$$

2 イベント関数の決定

$$\boldsymbol{\Phi}^{(0)} = \left[\phi_k(n)^{(0)}\right]_{1\leq k\leq K, 1\leq n\leq N}$$

3 イベントターゲットとイベント関数の反復再構成

$$(\boldsymbol{A}^{(0)}, \boldsymbol{\Phi}^{(0)}) \Rightarrow (\boldsymbol{A}^{(1)}, \boldsymbol{\Phi}^{(1)}) \Rightarrow \cdots (\boldsymbol{A}^{(S)}, \boldsymbol{\Phi}^{(S)})$$

右肩の数字は繰り返し数を表わしている。

2.4.2 イベントターゲット

発話音声にはスペクトル安定点が存在し、それは音声イベントの位置づけとしてのヒントになる。スペクトル安定点とそれを表わすパラメータはそれぞれイベント位置とイベントターゲットとみなすことができる。このように、スペクトルの安定点からイベントを抽出する手法を Spectral Stability Based Event Localizing Temporal Decomposition (S^{2}BEL-TD) という。

区間 $[n-M, n+M]$ 中におけるスペクトルパラメータ $y_i(n)$ の回帰直線の勾配 $c_i(n)$ と Spectral Feature Transition Rate (SFTR) $s(n)$ を考える。

$$c_i(n) = \frac{\sum_{m=-M}^{M} m y_i(n+m)}{\sum_{m=-M}^{M} m^2}, \quad 1 \leq i \leq P \tag{2.27}$$

$$s(n) = \sum_{i=1}^{P} c_i(n)^2, \quad 1 \leq n \leq N \tag{2.28}$$

$s(n)$ の局所的最小値はフレーム内の局所スペクトル安定点を表わすことになる。よって、これらの点はイベントの位置を、またそのスペクトルパラメータによってできるベクト

ルはイベントターゲットを表わすことになる。よって、$s(n)$ の局所最小値をそれぞれ $n_1, n_2, \cdots, n_K (n_1 < n_2 < \cdots < n_K)$ とすると、最初のイベントターゲットの行列 $\boldsymbol{A}^{(0)}$ は以下のようになる。

$$\begin{aligned} \boldsymbol{A}^{(0)} &= \left[\boldsymbol{a}_1^{(0)}\ \boldsymbol{a}_2^{(0)}\ \cdots\ \boldsymbol{a}_K^{(0)}\right] \\ &= \left[\boldsymbol{y}(n_1)\,\boldsymbol{y}(n_2)\ \cdots\ \boldsymbol{y}(n_K)\right] \end{aligned}$$

イベント数 K とイベント位置 $n_1 < n_2 < \cdots < n_K$ は SFTR 分析から得ることができる。よって、SFTR での窓の大きさ $2M$ は S^{2}BEL-TD のアルゴリズムにおいてはイベントの数と位置だけに関係のあるパラメータである。

2.4.3 イベント関数

音声イベントはある時間間隔でしか存在しないため、イベント関数にも同様のことが言える。このことは、イベント関数 $\phi_k(n)$ を次のような重み関数 $w_k(n)$ を用いて評価できることを示す。

$$w_k(n) = \begin{cases} n_{k-1} - n, & 1 \le n < n_{k-1} \\ 0, & n_{k-1} \le n \le n_{k+1} \\ n - n_{k+1}, & n_{k+1} < n \le N \end{cases}$$

$$\boldsymbol{w}_k = [w_k(1)\, w_k(2)\ \cdots\ w_k(N)]$$

重み関数行列 $\boldsymbol{W}$ は次のように表わせる。

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_K \end{pmatrix} \in R^{K \times N}$$

イベント関数 $\phi(n)$ はその中心 n_K 付近に存在しているが、その長さ、すなわち存在距離はわからない。しかし、k 番目のイベントはそのとなりのイベント位置、すなわちとなりのイベントの安定点 n_{k-1}、n_{k+1} を越えては振る舞いは小さくなるので、その範囲では徐々にその大きさは減少していく。よって、$w_k(n)$ は上の式のように表わせる。これはイベント関数の時間的な振る舞いの自由度が n_{k-1}、n_{k+1} の内側では大きく、外側では小さく制限していることになる。

次に、以下のような汎関数 $J(\boldsymbol{\phi}_n, \lambda)$ を考える。

$$J(\boldsymbol{\phi}(n), \lambda) = \sum_{i=1}^{P} (y_i(n) - \hat{y}_i(n))^2 + \lambda \sum_{k=1}^{K} w_k(n)^2 \phi_k(n)^2, \quad 1 \le n \le N \tag{2.29}$$

ここで、λ は重み係数、$\boldsymbol{\phi}(n)$ は以下の通りである。

$$\boldsymbol{\phi}_(n) = [\phi_1(n)\ \phi_2(n)\ \cdots\ \phi_K(n)\ ]^T,\quad 1 \leq n \leq N$$

また、$y_i(n)$、$\hat{y}_i(n)$ はそれぞれスペクトルベクトル $\boldsymbol{y}(n)$、$\hat{\boldsymbol{y}}(n)$ の i 番目の要素である。

$\boldsymbol{\phi}(n)$ は $J(\boldsymbol{\phi}(n), \lambda)$ が最小になるようなものが選ばれる。すなわち、

$$\frac{\partial J(\boldsymbol{\phi}(n), \lambda)}{\partial \phi_r(n)} \sum_{i=1}^{P} 2\left(\sum_{k=1}^{K} a_{ik}\phi_k(n) - y_i(n)\right) a_{ir} + 2\lambda w_k(n)^2 \phi_r(n) = 0 \tag{2.30}$$

$$\sum_{i=1}^{P} a_{ir}\left(\sum_{k=1}^{K} a_{ik}\phi_k(n)\right) + \lambda w_k(n)^2 \phi_r(n) = \sum_{i=1}^{P} a_{ir} y_i(n),\quad 1 \leq r \leq K \tag{2.31}$$

行列表示すると、次の結果が得られる。

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\phi}(n) + \lambda \boldsymbol{W}_n^T \boldsymbol{W}_n \boldsymbol{\phi}(n) = \boldsymbol{A}^T \boldsymbol{y}(n)$$

$$\boldsymbol{\phi}(n) = \left(\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{W}_n^T \boldsymbol{W}_n\right)^{-1} \boldsymbol{A}^T \boldsymbol{y}(n),\quad 1 \leq n \leq N \tag{2.32}$$

よって、最初のイベント関数行列 $\boldsymbol{\Phi}^{(0)}$ は次のように決定できる。

$$\boldsymbol{\Phi}^{(0)} = (\boldsymbol{\phi}(1)\ \boldsymbol{\phi}(2)\ \cdots\ \boldsymbol{\phi}(N)) \tag{2.33}$$

重み係数 λ の値は、シミュレーションの結果に基づいて選ばれる。最初の値は $\lambda^{(0)}$ とする。

2.4.4 反復再構成

反復再構成をすることにより、イベント関数の形状の改良、TD の精度向上、またイベントターゲットの改良が見込める。図 2.1に示すように、最初のイベント関数は Major-lobe と Minor-lobes、すなわち負の波形が存在する。反復再構成することにより、Minor-lobes を減少、かつ Major-lobe を自由に振る舞えるようにする。通常、4、5 回の繰り返しでイベント関数の形状が得られる。

イベント関数の再構成

イベント関数の再計算は 2.4.3章での手順を用いるが、より最適な重み関数と重み係数の選択は後で述べる。

l、S をそれぞれ反復ステップ数、最終的な反復回数とすると、イベント関数の再構成は以下のような形になる。

$$\left(\boldsymbol{A}^{(l-1)}, \boldsymbol{\Phi}^{(l-1)}\right) \rightarrow \boldsymbol{\Phi}^{(l)},\quad 1 \leq l \leq S$$

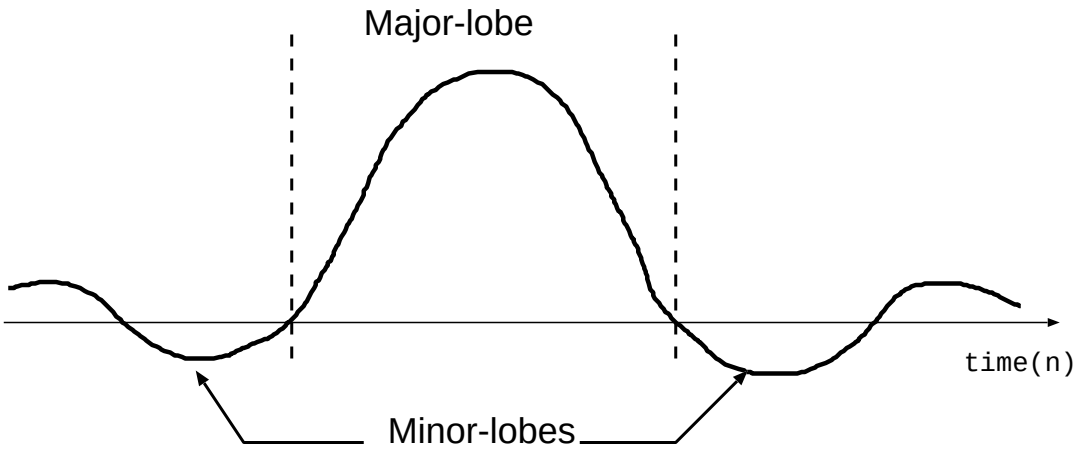

図 2.1: イベント関数

- 最適重み関数：最適な重み関数は以下の式のようになる。

$$w_k^{(l)}(n) = \begin{cases} l_k^{(l-1)} - n, & 1 \le n < l_k^{(l-1)} \\ 0, & l_k^{(l-1)} \le n \le r_k^{(l-1)} \\ n - r_k^{(l-1)}, & r_k^{(l-1)} < n \le N \end{cases} \tag{2.34}$$

  ここで、$l_k^{(l-1)}$、$r_k^{(l-1)}$ はイベント関数 $\phi_k(n)^{(l-1)}$ の Major-lobe の境界点を表わしている。よって、分析が繰り返されると Major-lobe は伸縮をうながし、Minor-lobe は減少することになる。

- 汎関数 $J(\boldsymbol{\phi}(n), \lambda)$ のバランス：l 回目の繰り返しステップでの重み関数 $\lambda^{(l)}$ は汎関数 $J(\boldsymbol{\phi}(n), \lambda)$ の 2 つの誤差の項のバランスをとるように、$l-1$ 回目のステップから得られる結果、すなわち $\boldsymbol{\Phi}^{(l-1)}$ と $\boldsymbol{A}^{(l-1)}$ をもとに決定される。

$$\lambda^{(l)} = \sigma \times \left(\frac{\sum_{n=1}^{N} \sum_{i=1}^{P} \left(y_i(n) - \hat{y}_i^{(l-1)}(n) \right)^2}{\sum_{n=1}^{N} \sum_{k=1}^{K} w_k^{(l)}(n)^2 \phi_k^{(l-1)}(n)^2} \right)$$

  ただし、σ はバランス比を表わす定数である。

反復数 l ステップでのイベント関数行列 $\boldsymbol{\Phi}^{(l)}$ は以下のように計算できる。

$$\boldsymbol{\phi}(n)^{(l)} = \left(\boldsymbol{A}^{(l-1)^T} \boldsymbol{A}^{(l-1)} + \lambda^{(l)} \boldsymbol{W}_n^{(l)^T} \boldsymbol{W}_n^{(l)} \right)^{-1} \cdot \boldsymbol{A}^{(l-1)^T} \boldsymbol{y}(n), \quad 1 \le n \le N \tag{2.35}$$

ただし、

$$\boldsymbol{W}_n^{(l)} = \mathrm{diag} \left[w_1^{(l)}(n)\ w_2^{(l)}(n)\ \cdots\ w_K^{(l)}(n) \right]$$

よって、イベント関数行列は以下のようになる。

$$\boldsymbol{\Phi}^{(l)} = \left(\boldsymbol{\phi}(1)^{(l)}\ \boldsymbol{\phi}(2)^{(l)}\ \cdots\ \boldsymbol{\phi}(N)^{(l)}\right)$$

イベントターゲットの再構成

イベントターゲットの再構成は、元スペクトルパラメータと合成スペクトルパラメータとの二乗誤差をターゲットベクトルで最小化する問題として扱える。繰り返し数 l のイベントターゲットは繰り返し数 l のイベント関数から決定できる。

$$\boldsymbol{\Phi}^{(l)} \rightarrow \boldsymbol{A}^{(l)},\quad 1 \leq l \leq S$$

元スペクトルパラメータと l 次合成スペクトルパラメータの二乗誤差は次のように表現できる。

$$E_i^{(l)} = \sum_{n=1}^{N} \left(y_i(n) - \sum_{k=1}^{K} a_{ik}^{(l)} \phi_k^{(l)}(n) \right)^2,\ 1 \leq i \leq P$$

a_{ir} に関する偏微分を行なうと、以下の式が導かれる。

$$\sum_{k=1}^{K} a_{ik}^{(l)} \sum_{n=1}^{N} \phi_k^{(l)}(n)\phi_r^{(l)}(n) = \sum_{n=1}^{N} y_i(n)\phi_r^{(l)}(n) \tag{2.36}$$

ただし、$1 \leq r \leq K$、$1 \leq i \leq P$ である。これより、l ステップのイベントターゲット行列は次のように表わせる。

$$\boldsymbol{A}^{(l)} = \left[a_{ik}^{(l)}\right]_{1 \leq i \leq P,\ 1 \leq k \leq K}$$

反復再構成の終りと収束

上の 2 つのステップは $\mathrm{MLC}^{(l)}$（Minor lobe content）があらかじめ設定していた値、例えば 1 %になるまで繰り返される。l 反復数における $\mathrm{MLC}^{(l)}$ は以下の式で表わされる。

$$MLC^{(l)} = \sqrt{\frac{\sum_{k=1}^{K}\sum_{n=1}^{N} \phi_k^{(l)}(n)^2 c_k^{(l)}(n)}{\sum_{k=1}^{K}\sum_{n=1} N\phi_k^{(l)}(n)^2}} \times 100\ \%$$

ただし、

$$c_k^{(l)}(n) = \begin{cases} 0, & l_k^{(l)} \leq n \leq r_k^{(l)} \\ 1, & \text{otherwise} \end{cases}$$

また、l 反復数における元スペクトルパラメータと合成スペクトルパラメータとの rms 誤差を定義する。

$$E_{\mathrm{rms}}^{(l)} = \sqrt{\frac{1}{NP}\sum_{n=1}^{N}\sum_{i=1}^{P}\left(y_i(n) - \hat{y}_i^{(l)}(n)\right)^2}$$

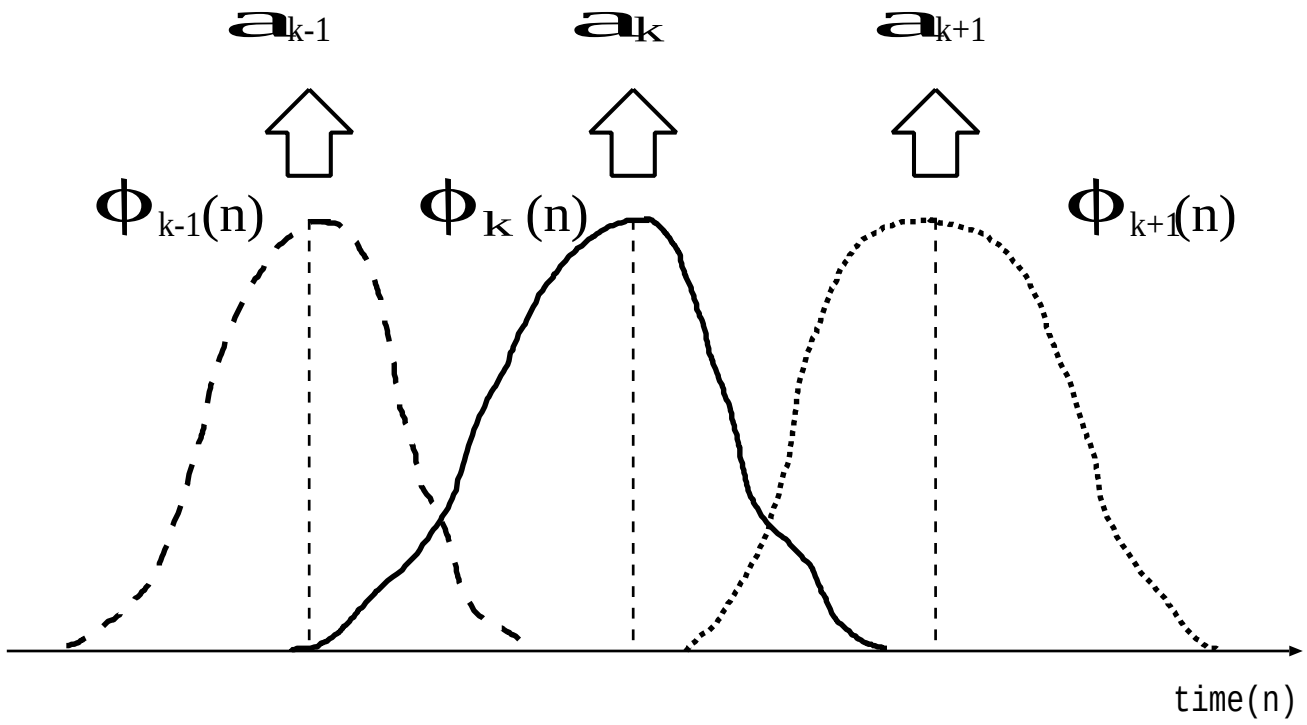

図 2.2: イベントターゲット a_k とイベント関数 $\phi_k(n)$

反復ステップ数 l における $\mathrm{MLC}^{(l)}$ の収束と $E_{\mathrm{rms}}^{(l)}$ は反復再合成の手法において、重要な特性である。

2.4.5 S^{2}BEL-TD を用いる理由

イベントターゲットとイベント関数はスペクトル変化の安定点に現れる。イベントターゲットとイベント関数の模式図を図 2.2に示す。

スペクトルパラメータがスペクトル変化の安定点 k で $\boldsymbol{a}_k$ として与えられ、これがイベントターゲットになる。すなわち、イベントターゲットはイベント位置におけるスペクトルパラメータを表しているが、これらは音声の有声区間中での子音・母音成分の安定点での値を意味している。また、イベント関数 $\phi_k(n)$ はスペクトル変化の安定点、すなわちイベント位置 k から $k+1$ に時間が移動する時に、前後のスペクトルの混合する割合を時間的に示したものを意味している。

このように、S^{2}BEL-TD で音声を分解することにより得られるイベントターゲットはそれぞれ子音・母音成分の安定点でのスペクトルパラメータで表現されるため、このうち

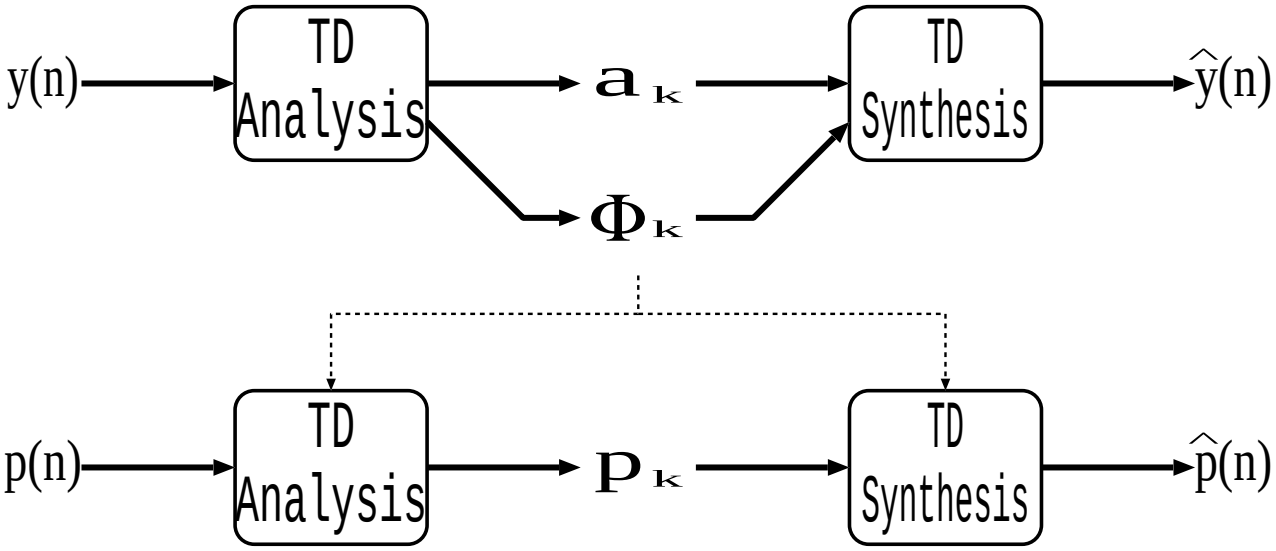

図 2.3: S^{2}BEL-TD によるスペクトルパラメータのイベント表現

の母音区間に対応するイベントターゲットを用いることにより母音を独立に制御することが可能となる。

また、STRAIGT で得られた基本周波数 $p(n)$ に関しても、イベント関数 $\phi_k(n)$ と基本周波数ターゲットを用いて次のように再現することができる（図 2.3）。

$$\hat{p}(n) = \sum_{k=1}^{K} p_k \phi_k(n), \quad 1 \leq n \leq N \tag{2.37}$$

ここで、$\hat{p}(n)$ と p_k はそれぞれ再現された基本周波数と k 番目の基本周波数ターゲットである。これにより、各母音における基本周波数成分を扱うことが可能となる。

このように、イベントターゲットがイベント位置における子音スペクトル成分、母音スペクトル成分、基本周波数成分を示し、かつ独立に分解、制御をすることができるため、S^{2}BEL-TD を採用した。

2.5 まとめ

本章では、音声分析合成系である STRAIGHT および S^{2}BEL-TD について述べてきた。STRAIGHT では、基本周波数を TEMPO2 により、スペクトルを STRAIGHT-core により抽出した。

表 2.1: STRAIGHT の分析条件

分析窓長	40 ms
分析シフト幅	1 ms
FFT 長	1024

S^2BEL-TD では、イベント位置に対するスペクトル、基本周波数のパラメータおよび時間変化に分解することができた。

なお、本研究での STRAIGHT による分析条件は、表 2.1 の通りである。

第 3 章

モーフィングパラメータ

3.1 目的

本論文では、サブセットデータによる音声モーフィングを試みる。より少量のサブセットデータで効果的な音声モーフィングを行なうためには、どのモーフィングパラメータをサブセットデータとして選択するかが重要な問題となってくる。

本節では、音声モーフィングを行なうための具体的なパラメータを取り上げ、その内容を説明する。

3.2 モーフィングを行なう音声データ

話者 A（フルサイズデータ）

モーフィングに使用する話者 A のフルサイズデータとしては、ATR 音声データベース男性話者 mms に関する情報を用いた。音声データ半母音や拗音を含まず、母音または有声・無声子音で構成された単語「そびえる」を採用した。

目標話者 B（サブセットデータ）

モーフィングに使用するサブセットデータを得るための話者 B には、24 歳の男性の大学院学生を用いた。話者 B に関する音声データの録音は防音室で行なった。マイクロホン（SONY C-536P）からの距離を約 15cm に保って発話させた音声を DAT レコーダ（SONY TCD-D10 PRO II）に入力し、サンプリング周波数 48kHz で録音した。この音声を 8kHz にダウンサンプリングして WS に保存した。表 3.1 に録音条件を示す。

表 3.1: 録音条件

機器	メーカー、機種
マイクロホン	SONY C-536P
DAT レコーダ	SONY TCD-D10 PRO 2
マイクロホンアンプ	SONY AC-148F
サンプリング周波数	48 kHz

なお、本研究で用いる音声データはすべてラベル付け[1]がしてあるものと仮定する。以下、音声モーフィングを行なうためのモーフィングパラメータを説明する。

3.3 単独発話母音イベント

文の構成単位は単語であり、単語は音節からなる。日本語の音節は、通常、子音と母音の組からなる。日本語では、外来語の表記も含めて約 110 個の音節が考えれるが、母音は/a/, /i/, /u/, /e/, /o/ の 5 種類である。母音は子音に比べて通常長い継続時間長を持ち、スペクトルも比較的明確である。よって、母音は通常容易にかつ確実に認識されることができるので、人間による音声認識でも機械による音声認識でも、重要な役割を果たしている。

特に、本研究ではより少量のサブセットデータによる音声モーフィングを行なう。そのため、本節では最小限の情報として、目標話者 B の単独発話 5 母音（/a/,/i/,/u/,/e/,/o/）をサブセットデータとして用いた音声モーフィングを試みる。以下、その手順を示す。

音声データ

録音条件は前説で説明した通りである。目標話者 B に関するサブセットデータとして採取した音声データは、目標話者 B による単独発話 5 母音/a/、/i/、/u/、/e/、/o/である。

モーフィング手法

1. 話者 B の単独発話 5 母音をそれぞれ STRAIGHT で分析、スペクトルパラメータ LSF に変換する。

[1]音声データにおける有声・無声区間や母音・子音区間などの時間情報が明記してあること

表 3.2: LPC ケプストラム距離

比較音声	LPC ケプストラム距離
音声 A と音声 X-1	1.7160
音声 B と音声 X-1	1.8187
音声 A と音声 B	1.9786

2. LSF を S^{2}BEL-TD で分析し、得られたイベントターゲットとイベント関数のうち、イベントターゲットを単独発話 5 母音それぞれの母音イベントとする。

3. 話者 A の元音声に関しても同様の分析を行ない、イベント抽出をする。得られたイベントにおいて、ラベルによる無声区間、子音区間、母音区間それぞれとの対応付けを行なう。

4. ラベル付けにより得られた話者 A の元音声の母音イベントを、話者 B の単独発話母音イベントと入れ替え、合成する。これにより、話者 A の音声の母音に相当する部分が話者 B の単独発話母音に変化し、話者 A から話者 B へのモーフィングが期待できる。図 3.1にその概要を示す。

以上のような操作により、目標話者 B に関する最小限の情報量である単独発話 5 母音イベントの操作による音声モーフィングが可能となる。表 3.2に 2 話者 A、B の音声（A、B）とモーフィング音声 X-1 の LPC ケプストラム距離を示す。

3.4 なまけ情報

連続音声中では、単音の物理的性質はその置かれた環境によって単独に発声された場合とは異なってくる現象、調音結合が生じている。この調音結合の影響により、単音が単独に発声されたときの声道の形を目標値とすれば、連続音声中では、目標値に十分に達しないで次の音に移る、いわゆる‘ なまけ ’の現象が現れる。すなわち、サブセットデータとして、さらに‘ なまけ ’情報を考慮することにより、より自然なモーフィング音声を生成することが期待できる。本節では、この‘ なまけ ’情報を付与する手法として、以下のものを考える。

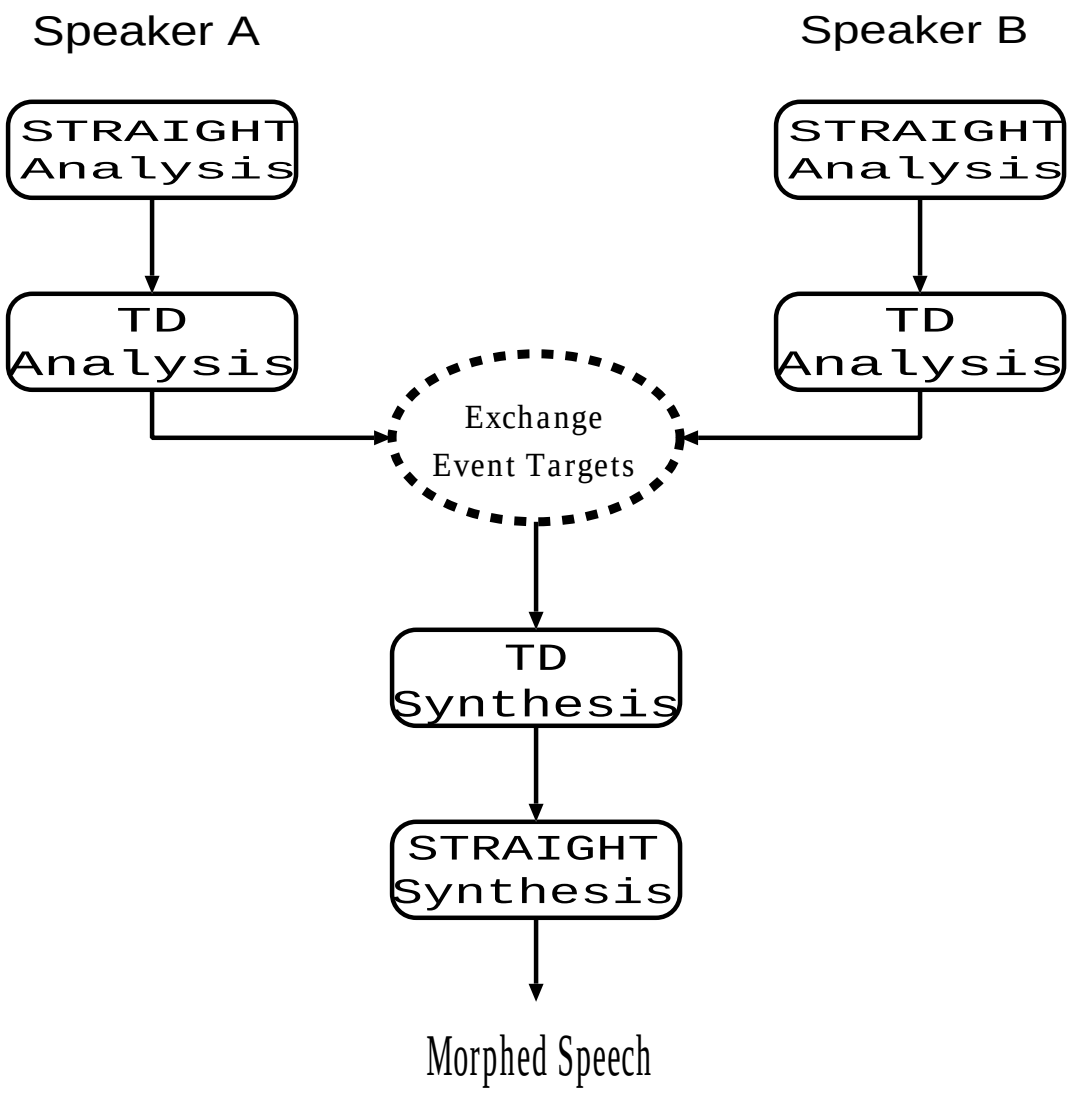

図 3.1: イベントターゲットの入れ替えによるモーフィング手法の概要

表 3.3: LPC ケプストラム距離

比較音声	LPC ケプストラム距離
音声 A と音声 X-1	1.7160
音声 B と音声 X-1	1.8187
音声 A と音声 X-2	1.2479
音声 B と音声 X-2	1.3153
音声 A と音声 B	1.9786

3.4.1 連続母音

目標話者 B のなまけ情報が付与されたサブセットデータのための母音イベントとして、3 連続母音における第 2 母音イベントによる音声モーフィングを試みた。

連続母音とは、母音のみで発声された音声である。なまけは、特に 3 連続母音における第 2 母音に現れると考えられる。すなわち、3 連続母音中では第 2 母音はその前後の母音である第 1 母音と第 3 母音の影響を受けている。よって、なまけ情報が付与された母音として、3 連続母音を用いた。

音声データ（3 連続母音）

録音条件は前説で説明した通りである。目標話者 B に関するサブセットデータとして採取した音声データは、目標話者 B による 3 連続母音/oie/、/ieu/である。これらは、モーフィグ対象音声データ「そびえる」に対応した 3 連続母音である。これらの第 2 母音/oie/、/ieu/を用いた。また、「そびえる」における第 1 母音/o/、第 4 母音/u/ に関しては、3 連続母音における/oie/、/ieu/を用いた。

得られた 3 連続母音に関するイベントのうち、ラベルとの対応によって得られる第 2 母音区間におけるイベント、すなわち第 2 母音イベントをモーフィングパラメータとして採用した。図 3.3に 2 話者 A、B の音声（A、B）と本手法によるモーフィング音声 X-2 の LPC ケプストラム距離を示す。

3.5 基本周波数

過去の研究から、音声における個人性は声帯特性と声道特性の 2 つに含まれると考えられているが、声帯特性である基本周波数に関する個人性情報は最も重要なパラメータの一つである。特に、基本周波数の時間的変化パターンには個人性が多く含まれるとの報告もある [9]。よって、本論文では、この基本周波数に関連した個人性情報をモーフィングパラメータとして採用した。

平均基本周波数

目標話者 B に関する最小限の基本周波数情報として、目標話者 B の平均基本周波数を採用した。手法としては、STRAIGHT により話者 A 、話者 B 双方の基本周波数を抽出し、得られた話者 A の基本周波数の平均を、話者 B の基本周波数の平均に合わせることを行なった。

基本周波数イベント

話者によって基本周波数の時間的変化パターンは異なるが、これはアクセントによる基本周波数の立ち上がりの違い、すなわちアクセントの強さを表わすことができる。この基本周波数の立ち上がりの違い、すなわちアクセントの強さの違いによって音色は変化する。

よって、これらの情報をサブセットデータとして取り入れるために、様々なアクセント型の音声データが必要である。そこで、同話者では単語におけるアクセントの強さには違いがないものと仮定し、代表的なアクセント型の音声データをサブセットデータとして用いる。

STRAIGHT で得られた基本周波数を S^2BEL-TD で分解し、その結果得られる基本周波数イベントをモーフィングパラメータとして採用した。手法は以下の通りである。

1. 話者 B の発話音声を STRAIGHT で分析し、スペクトルと基本周波数に分離する。
2. 得られた基本周波数をイベント関数を用いて S^2BEL-TD で分析する。これにより、イベント位置における基本周波数イベントが求まることになる。
3. 話者 A の元音声に関しても同様の分析を行ない、基本周波数イベントを抽出する。
4. 2 話者間でのアクセント位置における基本周波数イベントの入れ替えを行なう。この操作により、アクセントなどの基本周波数の動的変化による音声モーフィングが

期待できる。

ここでは、同話者では、単語におけるアクセントの強さには違いがないものと仮定している。以上のように、アクセントのある音韻における基本周波数イベントの入れ替え操作による音声モーフィングを試みる。これにより、基本周波数の動的変化による音声の個人性の違いが表わせるものと考えられる。

3.6 まとめ

本章では、音声モーフィングを行なうためのモーフィングパラメータを取り上げ、その内容を述べてきた。本論文では、サブセットデータとして特に単独発話母音イベント、3連続母音中の第 2 母音イベント、平均基本周波数、基本周波数イベントを選択し、これらの組み合わせによる音声モーフィングを行なうことにする。

第 4 章

聴取実験

4.1 目的

モーフィングを行なった音声が、どれだけ目標話者に近づいたかを確認するために、また、入れ替えたパラメータによるモーフィング音声への影響を調べるために、聴取実験を行う。

4.2 実験方法

音声データ

音声データは前説で説明した音声で、話者 A の音声データとして ATR 音声データベース男性話者 mms の「そびえる」を、話者 B の音声データとして「そびえる」、単独発話 5 母音（/a/、/i/、/u/、/e/、/o/）、3 連続母音（/oie/、/ieu/）および基本周波数を採用した。

刺激音

聴取実験には、前説で説明したモーフィング手法の組み合わせにより得られる以下 9 種類の刺激音を用いた。

表 4.1にモーフィング手法を、表 4.2は実験で用いたモーフィング音声の一覧を示す。

音声 a： 基本周波数イベントを話者 A から目標話者 B へと操作した音声（Fev）

音声 b： 平均基本周波数を話者 A から目標話者 B へと操作した音声（Fav）

表 4.1: モーフィング手法

操作	内容
Fev	話者 A の基本周波数イベント変化を目標話者 B に合わせる
Fav	話者 A の平均基本周波数を目標話者 B の平均基本周波数に合わせる
X-1	話者 A の母音イベントを話者 B の単独発話母音イベントに入れ替える
X-2	話者 A の母音イベントを話者 B の 3 連続母音中の第 2 母音イベントに入れ替える
X-3	話者 B の母音イベントを話者 B の 3 連続母音中の第 2 母音イベントに入れ替える

音声 c：話者 A の母音イベントと目標話者 B の単独発話母音イベントを入れ替え操作した音声（X-1）

音声 d：話者 A の母音イベントと目標話者 B の単独発話母音イベントを入れ替え、平均基本周波数を話者 A から目標話者 B へと操作した音声（Fav + X-1）

音声 e：話者 A の母音イベントと目標話者 B の単独発話母音イベントを入れ替え、基本周波数イベントを話者 A から目標話者 B へと操作した音声（Fev + X-1）

音声 f：話者 A の母音イベントと目標話者 B の 3 連続母音中の第 2 母音イベントを入れ替え操作した音声（X-2）

音声 g：話者 A の母音イベントと目標話者 B の 3 連続母音中の第 2 母音イベントを入れ替え、平均基本周波数を話者 A から目標話者 B へと操作した音声（Fav + X-2）

音声 h：話者 A の母音イベントと目標話者 B の 3 連続母音中の第 2 母音イベントを入れ替え、基本周波数イベントを話者 A から目標話者 B へと操作した音声（Fev + X-2）

音声 i：話者 B の母音イベントと話者 B の 3 連続母音中の第 2 母音イベントを入れ替え操作した音声（X-3）

被験者

被験者は正常聴力を有し、音声データのモーフィング目標話者 B と日頃接している 23 歳から 38 歳までの学生 10 名とした。

表 4.2: 実験で用いたモーフィング音声

音声	a	b	c	d	e	f	g	h	i
Fev	○				○			○	
Fav		○		○			○		
X-1			○	○	○				
X-2						○	○	○	
X-3									○

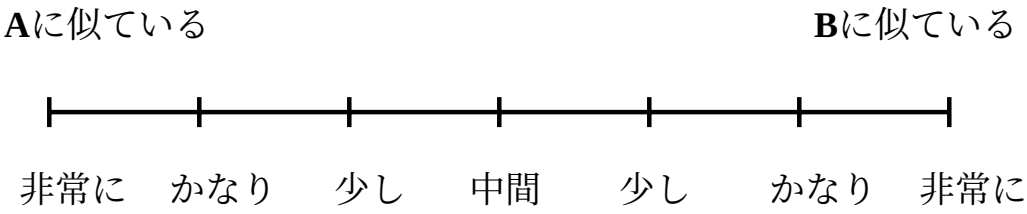

図 4.1: 評価表

実験方法

実験は ABX 法により行なった。刺激音 A、B、X を約 1 s の間隔で呈示し、刺激音 X が A と B の話者どちらに似ているかを「中間」を中心に、話者 A、B に「少し」、「かなり」、「非常に」の 7 段階の類似度で強制判断させた。図 4.2に音声の呈示タイミングを、図 4.1に評価表を示す。継時効果を打ち消すために、B A X の順についても実験を行なった。A、B、X の 3 つの刺激音の組を 1 刺激とし、1 刺激につき A B X、B A X を各 3 回、計 56 回をランダムに呈示した。

被験者は防音室内でヘッドホンにより受聴した。受聴は各被験者の聞きやすいレベルによる両耳受聴である。

刺激音は 防音室の外に設置されたワークステーション（WS）内に保存されており、WS から出力された刺激音は D/A 変換される。聴取実験システムの全体図を図 4.3に、使用した機器を表 4.3に示す。

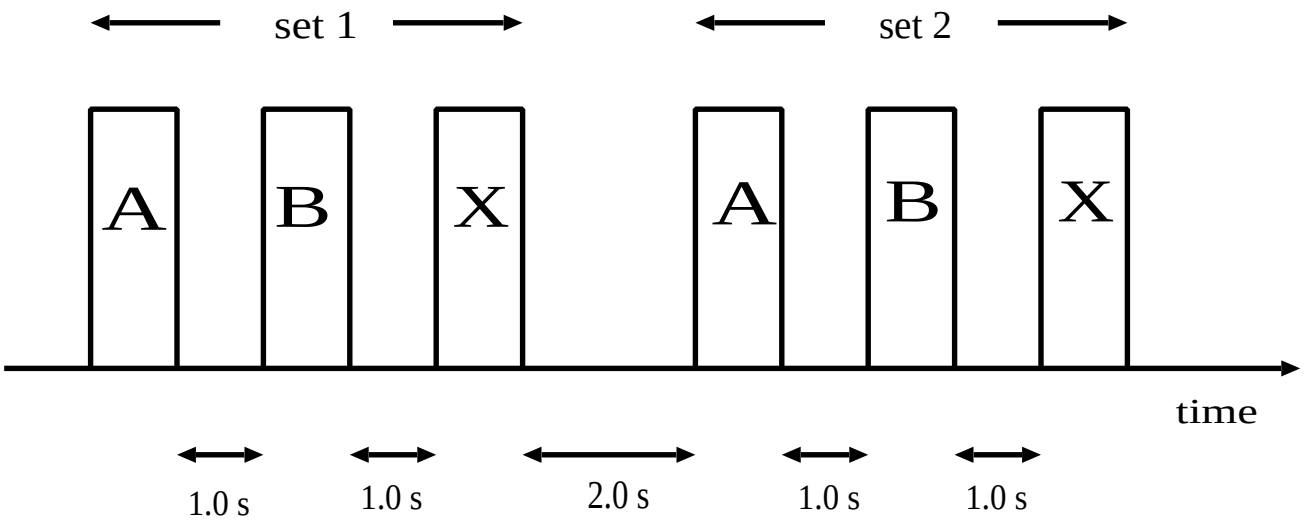

図 4.2: 音声の呈示タイミング

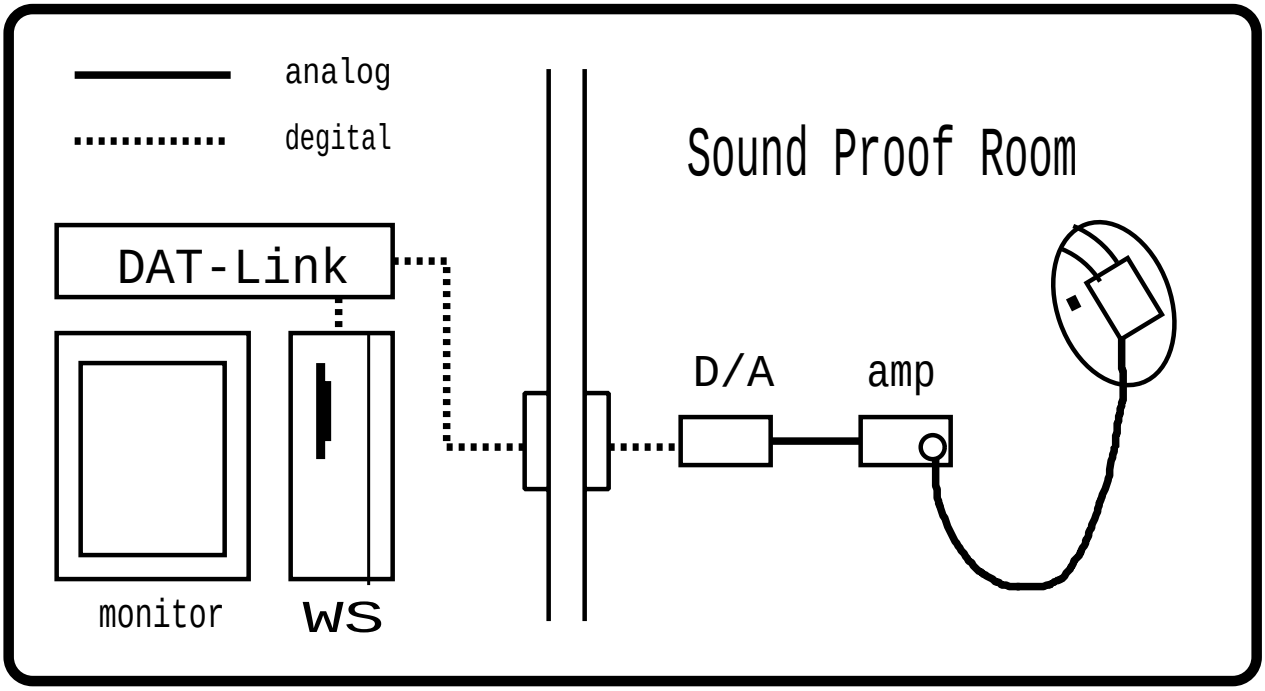

図 4.3: 聴取実験システムの全体図

表 4.3: 聴取実験に使用した機器

機器	メーカー、機種
ヘッドフォン	STAX Lambda Nova
ヘッドフォンアンプ	STAX SRM-1/MK-2 P.P
WS	Sun S-4/IX

4.3 実験結果と考察

実験結果は図 4.4のようになった。−3 に近づくほど話者 A に、3 に近づくほど目標話者 B に似ていることを表している。

得られた 9 種類のモーフィング音声を比較するために、これらを統計的な検定に基づいて考察する。まず、9 種類のモーフィング音声における等分散性を調べる。図 4.4から、音声 c、d、e、f、g、h においてはその分散値はほぼ等しいとみなすことができる。よって、特に音声 a と b、b と c、h と i の 3 組を有意水準 5 %での F 検定を行なった。その結果、a と b には有意差が認められなかったが、b と c、h と i の 2 組はいずれも有意差が認められた（表 4.4）。よって、得られたモーフィング音声は大別すると以下のように分類分けできる。

グループ F： 基本周波数のみ操作された音声（a,b）

グループ V： 母音イベントが操作された音声（c, d, e, f, g, h）

グループ B’： 目標話者 B の母音イベント・基本周波数イベント以外の情報が存在する音声（i）

基本周波数操作と母音イベント操作の違いによる影響

グループ F の結果から、基本周波数操作（Fav,Fev）のみのモーフィング音声はわずかながらも効果が表れている。このことから、基本周波数にも個人性が存在することが確認できたといえる。

グループ F とグループ V（音声 c、f）の結果から、基本周波数操作（Fav,Fev）よりも母音イベント操作（X-1,X-2）がより音声モーフィングに効果があることがわかる。このことは、個人性知覚は基本周波数よりも母音スペクトルによる影響が大きいと考えられる。

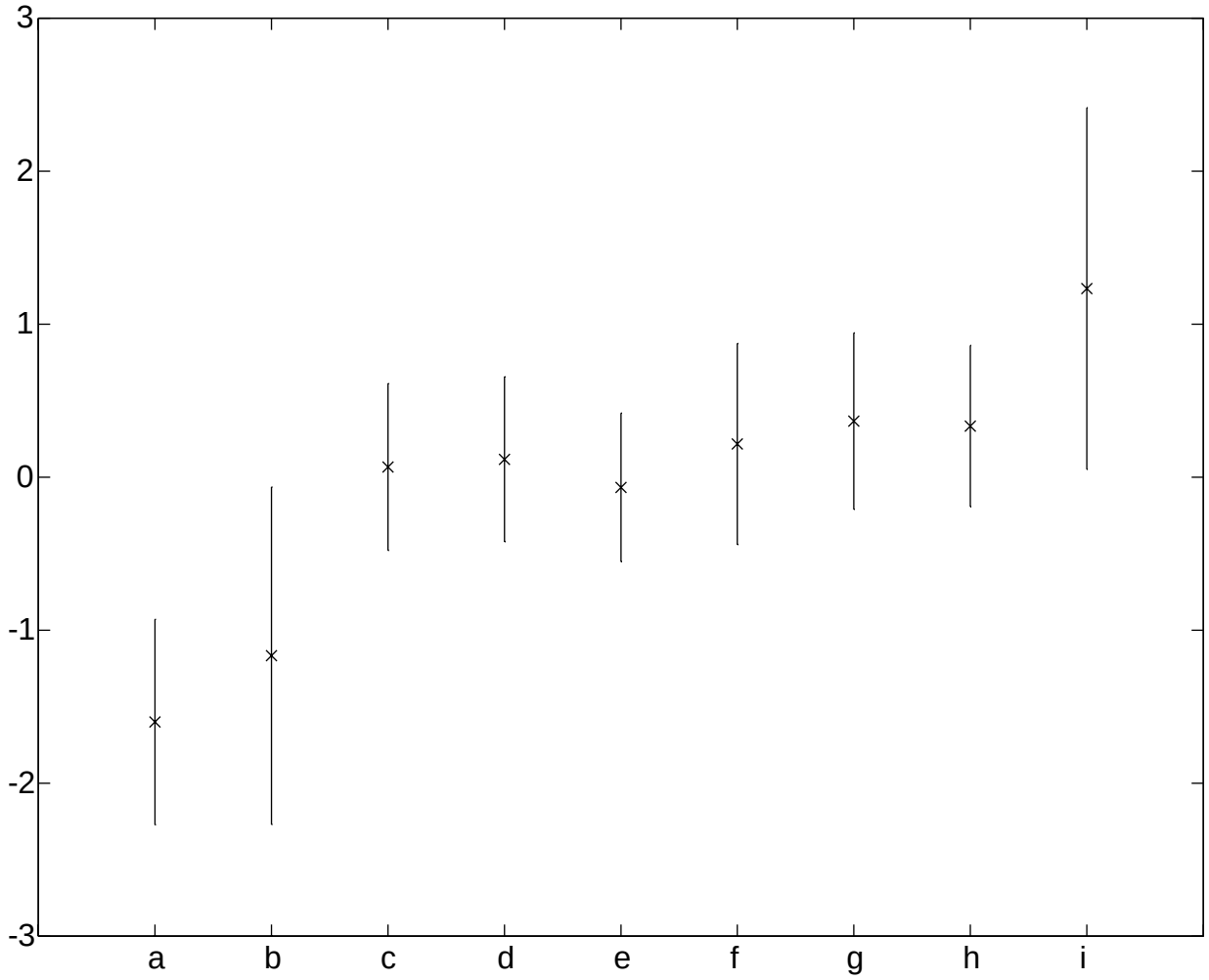

図 4.4: モーフィング音声「そびえる」の付置関係とその平均、標準偏差

表 4.4: F 検定

比較音声	F 比
a と b	2,43
b と c	4.19
h と i	5.15

F(9,9:0.05)=3.18

表 4.5: T 検定

	b	d	e	f	g	h
a	1.17	-	-	-	-	-
c	-	0.21	0.61	0.87	1.20	1.08
d	-	-	0.83	0.04	1.00	0.88
e	-	-	-	1.12	1.83	1.74
f	-	-	-	-	0.54	0.41
g	-	-	-	-	-	0.16

T(9,9:0.05)=1.73

基本周波数操作の違いによる影響

次に、基本周波数操作（Fav と Fev）の違いによるモーフィング音声を比較する。各グループ内でのモーフィング音声に対し、有意水準 5 %で T 検定を行なった。結果を図 4.5 に示す。

基本周波数操作（Fav と Fev）の違いによるモーフィング音声の組み合わせは a と b、d と e、g と h の 3 組である。これら 3 組は、いずれも有意な差はなかった。しかし、音声の個人性は主に基本周波数で知覚される話者とスペクトルによって知覚される話者があり、話者に依存する [10]。これらの理由で、基本周波数操作（Fav と Fev）の違いによるモーフィング音声はその知覚に差が表れにくかったものと考えられる。

母音イベント操作の違いによる影響

母音イベント操作（X-1 と X-2）の違いによるモーフィング音声の組み合わせは c と f、d と g、e と h の 3 組である。これら 3 組のうち、c と f、d と g の 2 組は有意な差はみられなかったが、e と h の組み合わせには有意な差が認められた。

このように、母音イベント操作の違いにより有意な差がはみられなかったが、全体的に操作 X-2（3 連続母音中の第 2 母音イベント）による音声モーフィングは操作 X-1（単独発話 5 母音イベント）によるものと比べるとその平均値も上がっている。さらに、操作 X-2 によるものは操作 X-1 によるものに比べて明らかに自然でなめらかなモーフィング音声が得ることができた。これらのことから、操作 X-2 による音声モーフィングはより効果のある手法だと考えられる。

その他のモーフィングパラメータによる影響

さらに、グループ V とグループ B'（操作 X-3）を比較すると、各平均値の距離は小さいとは言えない。また、グループ V のモーフィング音声は話者 A でも目標話者 B でもない、「中間」の音声という結果になっている。これから、より目標話者 B へと音声モーフィングを行なうためには、基本周波数と母音スペクトル以外に、さらに別のモーフィングパラメータが必要となることを表している。例えば、子音スペクトル成分を表す子音イベントや、スペクトルパラメータの時間変化を表すイベント関数などのパラメータが考えられる。

4.3.1 まとめ

聴取実験を通じて以下のことが明らかになった。

- 話者には依存はするが、基本周波数操作のみでも音声モーフィングは可能である。
- 母音イベント操作による音声モーフィングは十分に効果がある。特に、単独発話母音イベントよりもなまけ情報が付与された 3 連続母音中の第 2 母音イベントによる音声モーフィングがより効果がある。
- 基本周波数と母音スペクトル以外にも音声モーフィングに効果のあるモーフィングパラメータが存在する。
- 効果的な音声モーフィングを行なうためには、少なくともサブセットデータとして基本周波数と母音スペクトルを盛り込むことは必要条件である。

第 5 章

全体の考察

本研究では、音声モーフィングを行なうためのモーフィングパラメータとして基本周波数と母音スペクトルを取り扱い、モーフィング音声とモーフィングパラメータとの寄与、関与を調べてきた。

本節では、本研究で得られた結果の考察を行ない、効果的な音声モーフィングを行なうためにサブセットデータとして盛り込むべきモーフィングパラメータを検討する。

3 連続母音イベント

聴取実験を通じて、モーフィング音声に最も影響を与えたモーフィングパラメータは母音イベントであった。

また、単独発話母音イベントを用いたモーフィング音声よりも 3 連続母音イベントを用いたモーフィング音声が、目標話者に近い音声だと知覚されたが、それは歴然とした差ではなかった。

本実験ではその評価尺度を個人性・話者性としている。すなわち、話者 A、B どちらに似ているかという判断であり、モーフィング音声そのもののなめらかさや自然性などは評価対象には入っていない。単独発話イベント操作によるモーフィング音声と 3 連続母音イベント操作によるモーフィング音声は、個人性・話者性としては大きな変化が見られなかったが、音声そのものの自然性は明らかに後者がより自然な音声である。これらのことからも、‘ なまけ ’情報はより滑らかな音声モーフィングを行なうための重要なパラメータであるといえる。

さらに、S^2BEL-TD より得られる母音イベント、すなわち母音スペクトルは静的成分であり、声道情報の特性が含まれている。北村 [11] ら、鈴木 [12] らは静的スペクトルがもっとも話者知覚へ与える影響が大きいと報告しているが、本研究の結果から、さらに声道情

報の動的成分の一つでもある調音結合による‘ なまけ ’情報は個人性以上に、自然性に関与する重要なパラメータであることがいえる。

ただし、本研究のモーフィング手法である母音イベントの入れ替えによる音声モーフィングでは、S^{2}BEL-TD で合成をする際、パラメータが平均化されてしまう。よって、得られた音声スペクトルは高域成分が崩れてしまい、結果得られるモーフィング音声はこもった音質になってしまう。さらに、S^{2}BEL-TD でのスペクトルパラメータの時間分解において、子音成分と母音成分との分解能は完全ではなく、これによる音質低下も考えられる。より正確な音質評価を行なうためには、高域成分の補正や、S^{2}BEL-TD の改善も行なう必要がある。

基本周波数

本研究では、音声モーフィングに用いた音声データは単語「そびえる」であった。これは、アクセントなど基本周波数の動的変化が大きく表れにくいものである。そのために、基本周波数イベントの入れ替え操作による時間変化パターンによる音色の変化が表れにくかったものと考えられる。

さらに、音声の個人性の違いとして、その声質がスペクトルの変化に表れやすい話者と基本周波数の変化に表れやすい話者とがある [10]。本研究で用いた話者における音声モーフィングでは、基本周波数の変化による声質の変化は、スペクトルの変化によるものほど顕著には表れなかったことは樋口 [10] らの報告から示唆できる。

しかし、基本周波数操作のみのモーフィング音声でもわずかながらにも話者知覚へ影響を与えたという実験結果は、音声の基本周波数の時間変化パターンには個人性が含まれるという家永 [9] らの報告を支持する。

その他のパラメータ

その他のモーフィングパラメータとして子音イベントが挙げられる。子音イベントに関しては、予備実験として母音イベントと同様の入れ替え操作を行ない、その結果得られるモーフィング音声を評価した。しかし、話者 B の子音イベント操作によるモーフィング音声は、操作の前後で声質には変化がほとんど表れなかったため、聴取実験の対象音声から除外した。

サブセットデータ

以上のことを考慮した結果、サブセットデータで効果的な音声モーフィングを行なうにあたって、サブセットとして盛り込むべきモーフィングパラメータを母音イベントと平均基本周波数および基本周波数イベントとすることは必要条件であることがいえる。これらの条件を満たす音声データを得るためには、目標話者に関する

1. 主に 3 連続母音を中心に構成された音声
   実験結果から、個人性を表わすだけならば単独発話母音を用いるだけでも十分ではあるが、なめらかで自然な音声を得るためには 3 連続母音が必要となってくる。よって、フルサイズデータにおけるあらゆる音声データを自然性を保ちつつモーフィングするためには、サブセットデータとして多くの 3 連続母音の組み合わせが必要となる。

2. アクセント情報が含まれた音声
   アクセントの強さというものは同話者ではほぼ等しいものと考えられる。よって、サブセットデータとして多くの代表的なアクセント型の音声データが必要となる。

すなわち、目標話者に関する『3 連続母音の組み合わせを中心に構成され、かつ多くのアクセント情報が盛り込まれた文章』を採取することができれば、音声モーフィングは可能であるといえる。

第 6 章

結論

6.1　本論文で明らかになったことの要約

本論文では、目標話者に関するサブセットデータを用いて音声モーフィングを行なった。用いたモーフィングパラメータがモーフィング音声に与える影響を検討し、効果的かつ自然性を保った音声モーフィングを行なうためにサブセットデータとして取り入れるべきパラメータを求めた。

その結果、3 連続母音イベント、平均基本周波数および基本周波数イベントは効果的なモーフィングを行なうためには重要なパラメータであることがわかった。すなわち、目標話者に関するこれらの情報が含まれた音声、『3 連続母音の組み合わせを中心に構成され、かつ多くのアクセント情報が盛り込まれた文章』をサブセットデータとして用いることが、サブセットデータでフルサイズデータの音声モーフィングを行なうための必用条件であることを示した。

得られたモーフィング音声は完全なものとはいかなかったが、サブセットデータで音声モーフィングを行なうための目安を明らかにした。

6.2　今後の課題

評価方法

本研究ではモーフィング音声の評価方法として ABX 法を用いたが、これはモーフィング音声を話者 A の音声と話者 B の音声で結ばれる直線上で比較したものである。そのため、モーフィング音声の個人性を詳細に評価できていない可能性もある。音声の個人性を

詳細に検討するためには、多次元での評価方法を行なう必要がある。

スペクトルの高域成分の補正

本研究での手法による音声モーフィングでは、高域成分が崩れてしまうことにより、得られたモーフィング音声はこもったような音になってしまった。これが原因で音質評価が十分に行なえていないということも考えられる。十分な音質評価を行なうためには、これらの高域成分を補正することが必要となってくる。

その他のモーフィングパラメータの影響

本研究ではサブセットデータに取り入れるモーフィングパラメータとして 3 連続母音と基本周波数を扱ったが、さらに効果のある音声モーフィングを行なうためには、その他のモーフィングパラメータによるモーフィングへの影響も検討しなければならない。特に、本研究では扱わなかったイベントターゲットの時間変化を表すイベント関数などによる影響も比較・検討する必要がある。

用いた音声データ

本研究で対象とした音声データは母音、有声・無声子音で構成された、男性話者によって発話された 1 単語のみであった。よって、得られた結果がこの音声データに依存している可能性が否めない。今後、数多くの音声データ、特に有声子音や拗音などが含まれた音声データ、さらには女性話者の音声データ用い、これらの音声モーフィングを行ない、聴取実験を通じて本研究で得られた結果が一般的なものか否かを検証する必要がある。

謝辞

日頃ご指導頂き、貴重なご助言をいただきました赤木 正人 教授をはじめとする本学の教官の皆様に感謝いたします。本研究を進める過程において、多大なアドバイスをくださり、熱心に御討論いただいた赤木研究室の皆様に感謝いたします。また、御多忙の中、音声を録音させていただいた皆様、聴取実験に参加いただいた皆様に感謝いたします。最後に、2 年間の研究生活を支えてくださった全ての皆様に厚く感謝いたします。

参考文献

[1] 阿部, “基本周波数とスペクトルの漸次変形による音声モーフィング,” 日本音響学会講演論文集, 2-1-8, pp.259-260, 1995.

[2] 小坂直敏, “Sinusoidal model を用いた母音の声質補間,” 日本音響学会講演論文集, 2-1-10, pp.263-264, 1995.

[3] 坂野秀樹, 武田一哉, 板倉文忠, “包絡と音源の独立操作による音声モーフィング,” 信学技報, SP96-6, May 1996.

[4] 土屋誠一郎, 陸金林, 鹿野清宏, “包絡と音源情報の補間による音声モーフィングの一検討,” 日本音響学会講演論文集, 3-7-7, pp.271-272, 1997.

[5] 大室仲, 板倉文忠, “積分スペクトル逆関数（IFIS）とその応用に関する検討,” 信学技報, SP89-72

[6] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, “Speaker interpolation for HMM-based speech synthesis system,” J. Acoust. Soc. Jpn. (E) 21, 4(2000)

[7] 河原英紀, “聴覚の情景分析と高品質音声分析変換合成法 STRAIGHT,” 音学講論, 1-Q-21, pp.183-184, Oct.1994.

[8] A.C.R.Nandasena and M.Akagi, “Spectral Stability Based Event Localizing Temporal Decomposition,” Proc.ICASSP98, II, 957-960

[9] 家永太郎, 赤木正人, “音声のピッチ周波数の時間変化パターンに含まれる個人性とその制御,” 信学技報, SP94-104, May, 1995.

[10] Norio Higuchi and Makoto Hashimoto, “ANALYSIS OF ACOUSTIC FEATURES AFFECTING SPEAKER IDENTIFICATION,” ESCA. EUROSPEECH’95.4^{th}, September 1995. ISSN 1018-4074

[11] 北村達也, 赤木正人, “音声のスペクトル包絡における個人性に関する研究,” JAIST 博士論文, January, 1997.

[12] 鈴木教郎, 赤木正人, “文音声中に含まれる個人性情報の知覚に関する研究,” JAIST 修士論文, Feburary, 1999.