修 士 論 文

文音声中に含まれる
個人性情報の知覚に関する研究

指導教官　赤木 正人 助教授

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

鈴木 教郎

1999 年 2 月 15 日

目 次

図 目 次

表 目 次

第 1 章

序論

1.1 はじめに

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 text-to-speech において、音声合成の高品質化のために個人性をどのように付加し制御するかが問題となる。

このためにはまず音声中の個人性について明らかにすることが重要である。音声情報の中から個人性情報を抽出することができれば、多様な合成音声を生成できる音声合成システムの構築や、話者認識、ある話者の音声を徐々に他の話者の音声へと変化させると言う音声モーフィングなどマンマシンインターフェース技術にとって大変有益である。

さらに声質制御機能を備え、様々な合成音声を生成できる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、多話者間での音声翻訳システムにおける話者識別のためにも重要である。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。特に声質変換など合成音声の話者性制御を精度よく行うためには、音声の持つ個人性情報を把握することが非常に重要であり、そこで得られた知見は、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。

1.2 研究の背景

音声の個人性は、それぞれ声帯特性と声道特性に対する基本周波数とスペクトル包絡の双方に含まれる。また、生まれ育った環境によって話し方が違い、その違いも個人性のひとつである。話し方の違いは、アクセントや拍子など音声の時間情報にあらわれ、話者の特徴の時間的変化に着目した研究が多い。

これまでの個人性に関する研究は、大きく分けると話者の特徴の平均 (静的成分) を取り扱ったスペクトル包絡や平均基本周波数に関する研究と、話者の特徴の時間的変化 (動的成分) すなわちスペクトルの動きや基本周波数の時間変化に関する研究の 2 つに分類できる。

話者の特徴の静的成分に関する研究として、伊藤ら [1] は、基本周波数、スペクトル、音素持続時間が知覚に及ぼす影響を報告し、スペクトル包絡が個人性の知覚に与える寄与が大きいことを示した。また、北村ら [3] は単母音のスペクトル包絡成分に着目した個人性の分析を行っており、個人性情報は高域により多く含まれ、1,740Hz 付近に存在する peak 以上の帯域を利用して話者変換が可能であると述べている。

一方、話者の特徴の動的成分に関する研究として、基本周波数の時間変化の研究では、家永 [2] らが、単語音声により基本周波数の時間変化パターンを対象に研究を行い、基本周波数の時間変化パターンに個人性が多く含まれることを示している。また、スペクトル (ホルマント) の動きに関する研究は、粕谷ら [4] がホルマントの軌跡の時間的変化について研究を行った。その研究では、ホルマントの軌跡の時間的変化 (動的成分) よりもホルマントの平均 (静的成分) の方が個人性情報を多く含むことを示している。また、北村ら [5] が、連続母音を使ってスペクトル遷移パターン (動的成分) が話者識別に与える影響は小さいと報告している。

しかし、これらの報告では、スペクトルやホルマントの動き、基本周波数の時間変化について述べられているが、両方統合された変化に関しては検討されていないのが現状である。

1.3 本研究の目的

本研究では、話者知覚に与える物理量をあきらかにする。

音声中に含まれる話者特徴すなわち個人性をあきらかにすることができれば、話者認識や音声認識、音声合成、text-to-speech などの様々な音声処理技術に応用ができる。

また、前節の報告では、スペクトルやホルマントの動き、基本周波数の変化に関して検討は行われているが、両方統合された変化に関して検討されていない。スペクトル、基本周波数、これらの変化を同時に制御して、話者知覚での関連性を調べる必要がある。

そこで、本研究では、基本周波数、スペクトル包絡の双方の時間的変化を表す物理量を求め、基本周波数の時間変化、スペクトルの動きについて総合的に取り扱い、個人性について検討する。スペクトル、基本周波数を同時に扱うために本研究では、STRAIGHT[6] と S^2BEL-TD[7] を用いて文音声を分解する。そして、各パラメータの分析や聴取実験で

パラメータを話者間で入れ替えることで話者知覚への影響を調べ、それぞれの寄与、関与を調べる。

1.4　本論文の構成

本論文の構成を以下に示す。

第 1 章では、過去の個人性の研究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、STRAIGHT と S^2BEL-TD を用いて文音声を分解し、３つの要素 (スペクトル、基本周波数、その変化) を抽出するモデルの構築を行っていく。

第 3 章では、文音声を対象に、3 つ（スペクトル、基本周波数、その変化）の要素を分析する。

第 4 章では、聴取実験を通じて 3 つの要素と話者知覚の影響を調べていく。

第 5 章では、全体の考察を行い、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

時間変化パターン記述モデル

2.1 目的

本研究では、個人性関係物理量のうちスペクトル、基本周波数、その変化を総合的に取り扱う。そのため本節ではスペクトル、基本周波数、その変化を抽出するモデルの構築を行っていく。

2.2 時間変化パターンの記述モデルの概要

本論文では、時間変化を考慮した合成音声を作成する為に、以下の図 2.1 のように音声合成系 STRAIGHT[6] とテンポラルデコンポジション S^2BEL-TD[7] を使用し、時間変化パターンを記述するモデルを構築する。

本章では、図 2.1 の各過程について簡単に説明する。

2.3 STRAIGHT

2.3.1 はじめに

本論文では、時間変化を考慮して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為には、音声から個人性情報を抽出す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高品質な合成音声を作成できる STRAIGHT(Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram)[6] を採用する。

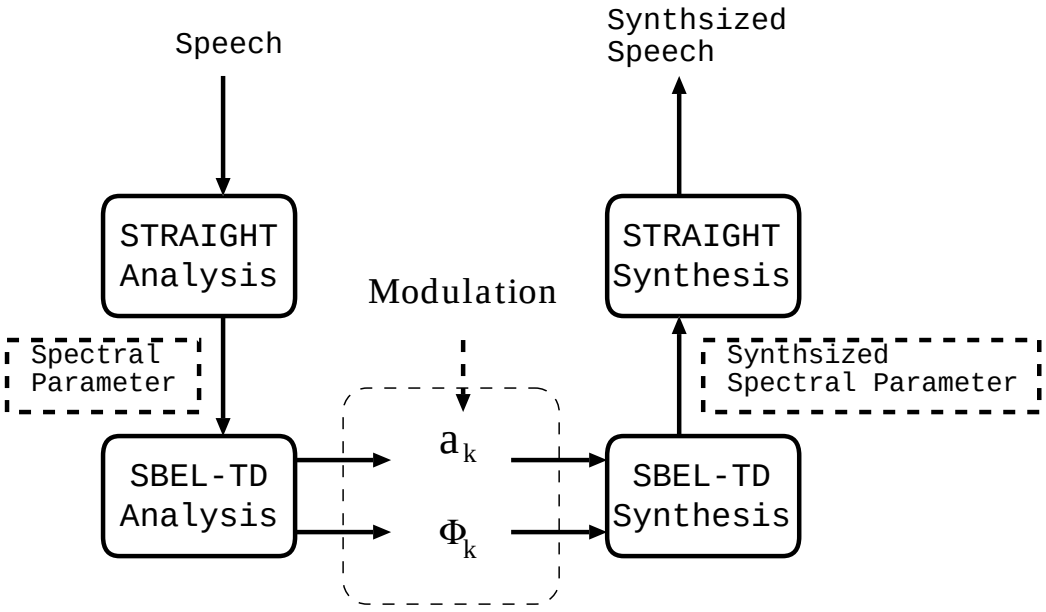

図 2.1: スペクトル記述モデルの構成

2.3.2 STRAIGHT の構造

STRAIGHT は、模式図 2.2 に示すように 3 つの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無い時間周波数表現を抽出する方法である。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おこなうことが中心的なアイデアである。

SPIKES は、合成に用いる駆動音源の位相 (正確には郡遅延) 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郡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。

TEMPO は、「基本波らしさ」という概念を導入することで、基本周波数の推定とその抽出精度の推定とを同時に行うことを可能にした方法である。

以下では本研究で取り扱う個人性関連物理量 (スペクトル、基本周波数) を抽出する STRAIGHT-core、TEMPO について簡単に説明を行う。

2.3.3 STRAIGHT-core

STRAIGHT-core の重要なアイデアは、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

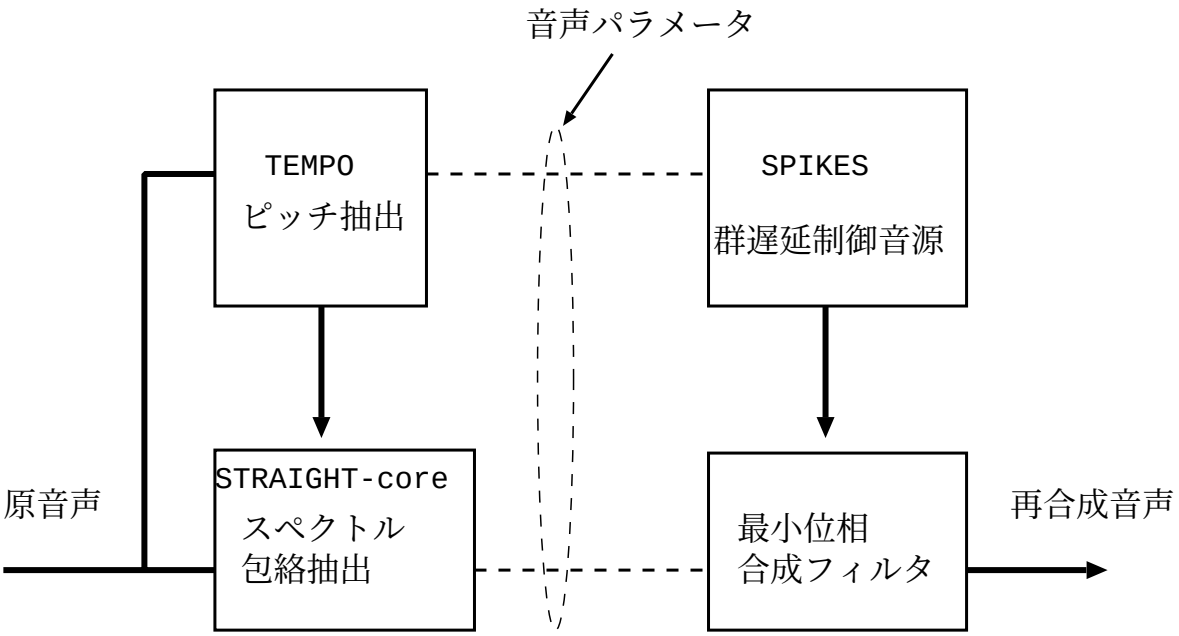

図 2.2: STRAIGHT の構成

割を担うものであると解釈するところにある。この解釈の下で、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のが、STRAIGHT-core の原理である。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 TEMPO 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桁違いに小さな誤差を有することが示されたのである。

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。

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

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す。また、$\alpha_k(t)$ は、それぞれの成分の強さを表し、ϕ_kは、k番目の高次調波成分の初期位相を表す。この信号の短時間 Fourier 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となる。

時間方向の位相干渉の効果の軽減

実行的な長さが 1 基本周期以上でサイドローブが十分に減衰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タである。

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

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ゼロとなる部分が出現する。このゼロ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ゼロとなるのは、調波と調波の中間の周波数で上の調波成分と下の調波成分の位相が逆になる部分である。したがって、ある窓 $w(t)$ に対して上下の調波の位相を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は、下の窓によるスペクトルがゼロの部分で最大値を持つようになる。

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

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 $P_r(\omega, t)$ が求められる。

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

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、$P_r(\omega, t)$ 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。

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

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

前節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。

2.3.4 TEMPO

TEMPO では、式 (2.1) の基本波成分の瞬時周波数として基本周波数を定義する。「基本周波数が分からなければ基本波を選ぶことができない。」という問題を、基本周波数を利用せずに求めることができる「基本波らしさ」という尺度を導入することにより回避するというアイデアが TEMPO の鍵となっている。

基本波らしさ

複数の調波成分から構成される複合音を定 Q フィルタバンクを用いて分析したときに、出力の中に基本波成分のみが含まれているフィルタを選択する方法を考える。各々のフィルタの遮断特性は、高域では急峻で低域ではなだらかであるとする。このような条件の下では、基本波のみを含んでいるフィルタ出力の瞬時周波数の変動と瞬時振幅の変動は最小となる。そこで、この性質を利用して、「基本波らしさ」M_cを次のように定義する。

$$\begin{aligned} M_c \;=\; & -\log\left[\int_\Omega \left(\frac{d|D|}{du} - \mu_{AM}\right)^2\right] \\ & -\log\left[\int_\Omega \left(\frac{d^2 \arg(D)}{du^2} - \mu_{FM}\right)^2 du\right] \\ & +\log\left[\int_\Omega |D|^2 du\right] + \log\Omega(\tau_0) + \log\tau_0 \end{aligned} \tag{2.7}$$

$$\mu_{AM} = \frac{1}{\Omega}\int_{\Omega}\left(\frac{d|D|}{du}\right) \tag{2.8}$$

$$\mu_{FM} = \frac{1}{\Omega}\int_{\Omega}\left(\frac{d^2\arg(D)}{du^2}\right) \tag{2.9}$$

ここで、Dは、以下で定義される wavelet 変換である。用いている関数 $g_{AG}(t)$ は、一例であり、前述の条件を満たす広い範囲の関数が利用できる。

$$D(t,\tau_0) = |\tau_0|^{\frac{1}{2}}\int_{-\infty}^{\infty} s(t)g_{AG}\left(\frac{\bar{t}-u}{\tau_0}\right)du \tag{2.10}$$

$$g_{AG}(t) = e^{-\pi(\frac{t}{\eta})^2}e^{-j2\pi t}(1+j\sin\pi t) \tag{2.11}$$

なお、式を必要以上に複雑にしないため、式 (2.7) などでは積分をΩという積分区間で行われるものと略記している。実際には Gauss 型の窓をかけて重み付きの計算を行っている。

基本周波数の抽出

基本周波数は、このようにして計算される「基本波らしさ」が最大となるチャンネルの瞬時周波数として求められる。実装に置いては、選択されるチャンネルの切り替え時点で基本周波数の時系列が 1 次の導関数まで連続になるように、隣接するチャンネルの出力 cos の加重の下で合成している。

STRAIGHT を用いる理由

STRAIGHT を用いる理由は、文音声から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 TEMPO により計算される。さらに高品質な合成音声を生成できることから STRAIGHT を採用した。

本論文で用いた合成音は、全て STRAIGHT により作成した。また、TEMPO で得られた基本周波数の異常値が見られるときは修正を行った。合成音の作成は、変形した各パラメータを用いて音声の合成を行う。

なお、分析合成に用いた分析条件は、すべて、表 2.1の分析合成条件を用いた。

2.4 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 を通じてスペクトルパラメータから時間変化パターンとスペクトルの安定する位置に対する

表 2.1: STRAIGHT 分析合成系の条件

分析窓長	40ms
分析シフト幅	1ms
FFT 長	1024

スペクトルに分解する。そのため、S^{2}BEL-TD で分解するスペクトルパラメータとして、線形補間性に優れている Line Spectral Frequncies(LSF) を用いる。

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のよいパラメータとしても、LSF が報告されている [7] 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。

STRAIGHT で得たスペクトルからスペクトルパラメータ (LSF) に変換する方法は以下の通りにおこなった。

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 $X[k],\quad 0 \leq k \leq N-1$ を用いてパワースペクトル $S[k]$ を計算する。

$$S[k] = |X[k]|^2,\quad 0 \leq k \leq N-1 \tag{2.12}$$

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によって相関関数を求めれば

$$R[n] = \frac{1}{N}\sum_{k=0}^{N-1} S[k]e^{j\frac{2\pi kn}{N}} \tag{2.13}$$

となる。今、この相関関数を有する過程 $\mathbf{x}(t)$ が all-pole 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 $\{a_l^L\}, l = 1, 2, \cdots, L,\ 0 < L < N/2$ として、

$$P_L = R[0] - \sum_{l=1}^{L} a_l^L R[l] \tag{2.14}$$

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 $\{a_l^L\}, l = 1, 2, \cdots, L$ を決める。このときのフィルタ係数 $\{a_l^L\}, l = 1, 2, \cdots, L$ は LPC の予測係数と一致する。また、STRAIGHT で得られるパワースペクトルを

$$S(\omega) = \frac{P_L}{|1 - a_1^L e^{-j\omega T} - a_2^L e^{-j2\omega T} \cdots - a_L^L e^{-jL\omega T}|^2}$$

のように表すこととなる。

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

$$\begin{cases} A_L(Z) = 1 - \sum_{l=1}^{L} a_l^L Z^{-l}) \\ B_L(Z) = Z^{-(L+1)} A_L(Z^{-1}) \end{cases} \tag{2.15}$$

これを用いれば、LSF を計算できる。

LSF への計算は式 (2.15) より、

$$\begin{aligned} P(z) &= A_L(z) - B_L(x) \\ Q(z) &= A_L(z) + B_L(x) \end{aligned} \tag{2.16}$$

となる。$P(z)$ は反対称な係数をもつ $(p+1)$ 次の多項式、$Q(z)$ は対称な係数をもつ $(p+1)$ 次の多項式である。式 (2.15) から、L を偶数と仮定すると、$P(z)$ と $Q(z)$ は次のように因数分解される。

$$\begin{aligned} P(z) &= (1 - z^{-1})\Pi_{i=2,4,\cdots,l}(1 - 2z^{-1}\cos\omega_i + z^{-2}) \\ Q(z) &= (1 + z^{-1})\Pi_{i=1,3,\cdots,l-1}(1 - 2z^{-1}\cos\omega_i + z^{-2}) \end{aligned} \tag{2.17}$$

ただし、ω_iは次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 と呼ぶ。そして LSF 上での補間が可能となる。LSF 係数の次数は個人性情報を有する 30 次を用いた。

2.5　S^{2}BEL-TD

2.5.1　はじめに

時間変化を記述するためにスペクトルの時間変化パターンをモデル化する必要がある。そこで下記に示す理由により S^{2}BEL-TD[7] を採用した。

- S^{2}BEL-TD はスペクトルパラメータをスペクトルの時間変化パターンとイベント位置におけるスペクトル情報に置き換えることができる。

ここで、イベント位置とは、スペクトル変化の安定点である。

簡単に S^{2}BEL-TD について説明する。

2.5.2 S^{2}BEL-TD の構造

S^{2}BEL-TD それ自身はスペクトルパラメータをコーディングし、データ量を削減する目的として作成され研究されてきた。

S^{2}BEL-TD では、スペクトルパラメータからスペクトルの時間変化パターンとイベント位置に対するスペクトルを得る為に、以下のような計算を行っている。

1. 初期イベントターゲットの決定

$$\mathbf{A}^{(0)} = \left[a_k^{(0)}\right]_{1\leq k\leq K} \tag{2.18}$$

2. 初期イベントファンクションの決定

$$\mathbf{\Phi}^{(0)} = \left[\phi_k(n)^{(0)}\right]_{1\leq k\leq K, 1\leq n\leq N} \tag{2.19}$$

3. イベントターゲットとイベントファンクションの繰り返し計算

$$(\mathbf{A}^{(0)}, \mathbf{\Phi}^{(0)}) \rightarrow (\mathbf{A}^{(1)}, \mathbf{\Phi}^{(1)}) \rightarrow \cdots (\mathbf{A}^{(S)}, \mathbf{\Phi}^{(S)}) \tag{2.20}$$

このような過程に基づいたシステムの出力となるスペクトルパラメータは以下のようになる。

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \quad 1 \leq n \leq N \tag{2.21}$$

ここで、$\mathbf{a}_k$と$\phi_k(n)$ は、それぞれk番目のイベントダーゲットとイベントファンクションである。これにより、スペクトルパラメータは、イベント位置のスペクトル $\mathbf{a}_k$を重みとした時間変化パターン$\phi_k(n)$ の線形和として表される。

2.5.3 S^{2}BEL-TD を用いる理由

イベントターゲット、イベントファンクションはスペクトル変化の安定点にあらわれる。イベントターゲットとイベントファンクションの模式図を図 2.3 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 a_kとして与えられ、これがイベントターゲット (スペクトルターゲット) になる。イベントファンクション$\phi_k(n)$ の意味するところは、スペクトル変化の安定点 kから次の安定点 $k+1$ 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である。

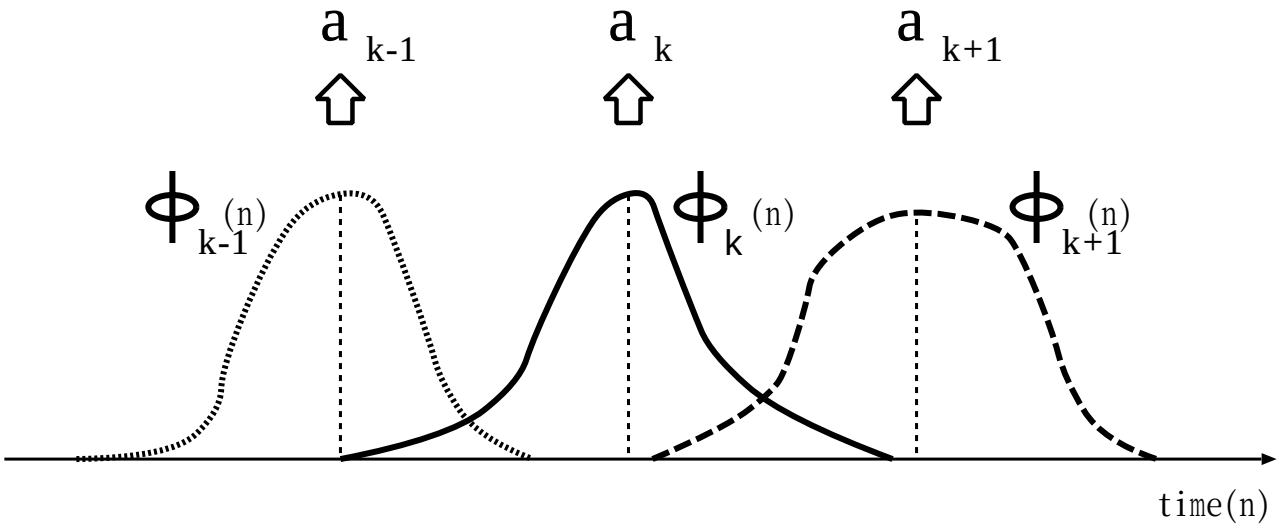

図 2.3: イベントターゲット $\mathbf{a}_k$とイベントファンクション$\phi_k(n)$

また、STRAIGHT で得られた基本周波数 $p(n)$、振幅成分 $g(n)$ はイベントファンクション$\phi_k(n)$ と基本周波数ターゲット、振幅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 2.4)。ただし、基本周波数、振幅成分は対数変換をおこなったものを用いた。

$$\hat{p}(n) = \sum_{k=1}^{K} p_k \phi_k(n), \quad 1 \leq n \leq N \tag{2.22}$$

$$\hat{g}(n) = \sum_{k=1}^{K} g_k \phi_k(n), \quad 1 \leq n \leq N \tag{2.23}$$

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り、$\hat{g}(n)$ と g_kは、それぞれ再現された振幅成分と k番目の振幅ターゲットである。

このように、イベントファンクションが、スペクトルや基本周波数、振幅の時間変化の構造を、またイベントターゲットがイベント位置における 3 つの要素 (スペクトル、基本周波数、振幅) を示すことができるため S^{2}BEL-TD を採用した。

2.6 位置情報数 (イベント数) の話者間での正規化

S^{2}BEL-TD で得たパラメータを話者間で入れ替えを行う為、話者間でのパラメータの対応付けが必要となってくる。パラメータはイベントの発生位置に出現するので、イベントの位置が重要になってくる。

このため、話者間でのパラメータの対応付けを行うためにひとつのルールを用いること

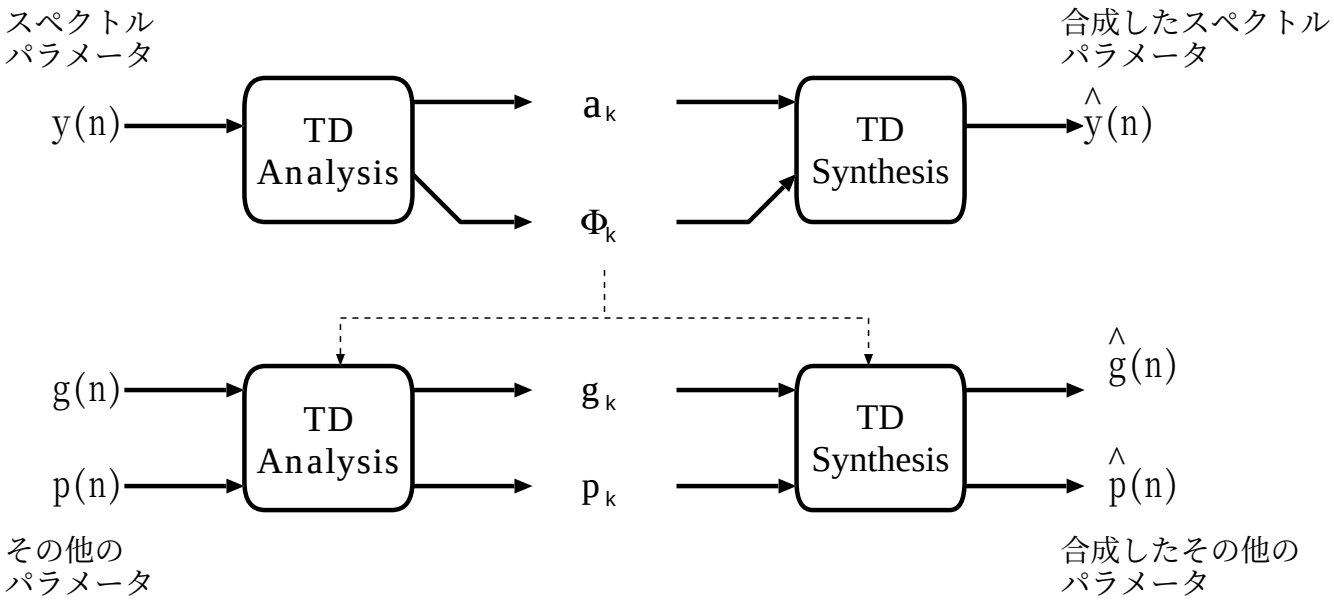

図 2.4: 基本周波数、振幅成分のコーディング

にする。そのルールは、1 つの音韻に対して 1 つのイベントをわりあてることである。このことにより、話者が違っても同じ文章ならば同じ音韻数を取ることができ、イベントの対応がつくことになる。

2.7 まとめ

本章では、時間変化パターンを記述できるモデルについて述べてきた。STRAIGHT では、スペクトル、基本周波数を STRAIGHT-core、TEMPO を用いて抽出した。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、振幅のパラメータおよびこれらの時間変化に分解することができた。

第 3 章

分析

3.1 目的

文音声についての時間変化パターンを S^2BEL-TD により抽出し、抽出された各パラメータに現れる話者間の物理的距離について分析、検討を行う。

3.2 分析する音声、パラメータ

音声データに用いた文章は、声帯振動の伴う母音または有声子音で構成した。分析に用いた音声データの文章には「いいえ、うえにある」を採用し、これを 10 回、普通とゆっくりの速さで発話するように指示した。発話者は 36 歳から 50 歳の男性の大学院教員 5 名である。5 人中 3 人は普通の速さ、2 人はゆっくりの速さのデータを用いた。

録音条件を表 3.1 に示す。録音は、防音室で行った。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)

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	48kHz

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。

分析に使用するパラメータは、スペクトルの時間変化パターン$\phi_k(n)$ とイベント位置のスペクトルパラメータ LSF(30 次)、基本周波数、振幅成分 (LSF 残余 P_L) である。

はじめに、イベント (音韻中心) 位置に置ける各パラメータの分析をおこなう。つぎの節で時間変化パターンの分析をおこなう。

3.3 イベント位置に置ける各パラメータの分析

3.3.1 分析方法

文音声データから抽出した各パラメータについて各話者間での距離を求め、それを元に多次元尺度構成法により 3 次元または 2 次元分布を求める。

スペクトルパラメータの話者間の物理的な距離 **CD** については、LSF を LPC 係数に、LPC 係数を LPC ケプストラムに変換し、イベント位置のケプストラム距離 ($\mathbf{cd}_k$) を求め、全イベント数 Kの平均として求めた。

$$\mathbf{cd}_k = D_b\sqrt{2\sum_{i=1}^{p}(c_{ik}^{(x)} - c_{ik}^{(y)})^2} \tag{3.1}$$

$$\mathbf{CD} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{cd}_k \tag{3.2}$$

ここで、p はケプストラム次数 (30 次)、$c_{ik}^{(x)}, c_{ik}^{(y)}$は k番目のイベントに対する話者 x と話者 y のケプストラム係数、D_bは距離尺度をデシベルに変換するための定数で、$D_b = 10/\ln 10$ である。

また、基本周波数、振幅も同様に話者間の物理的距離 **DIST_F0**、**DIST_振幅**を求める。

$$\mathbf{DIST\_F0} = D_b\sqrt{\sum_{i=1}^{K}(p_i^{(x)} - p_i^{(y)})^2} \tag{3.3}$$

$$\mathbf{DIST\_振幅} = D_b\sqrt{\sum_{i=1}^{K}(g_i^{(x)} - g_i^{(y)})^2} \tag{3.4}$$

ここで、$p_i^{(x)}, g_i^{(x)}, p_i^{(y)}, g_i^{(y)}$はそれぞれ話者 x と話者 yの基本周波数、振幅の係数である。

次の節では、各パラメータの物理的な距離を用いて多次元尺度構成法による分析結果を示す。

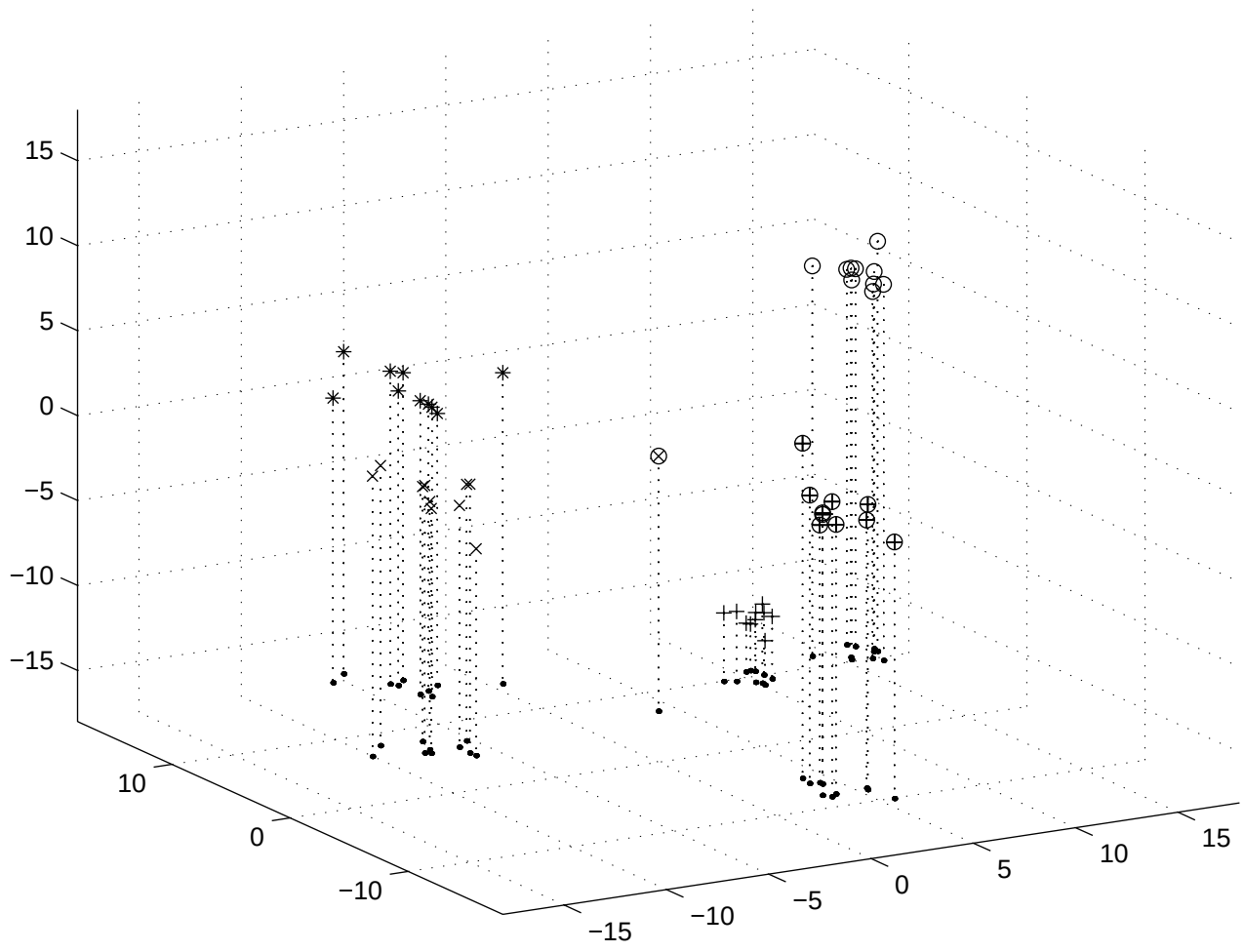

図 3.1: ケプストラム距離の付置図 (話者 a:○、話者 b:×、話者 c:+、話者 d:∗、話者 e:⊕、話者平均:⊗)

3.3.2 話者間のスペクトルパラメータの分析結果

イベント位置のケプストラムの距離を話者間で求め、多次元尺度構成法により 3 次元上に配置した (図 3.1)。また話者間でケプストラムパラメータを平均したものも図中に⊗としてプロットしている。

図 3.2 では、話者間の平均からの各話者のケプストラム距離を示している。平均からより近い話者は話者 c であり、より遠い話者は、話者 a、e である。また、話者 b、d は平均からの距離が同じくらいであり、方向も類似した方向にある。

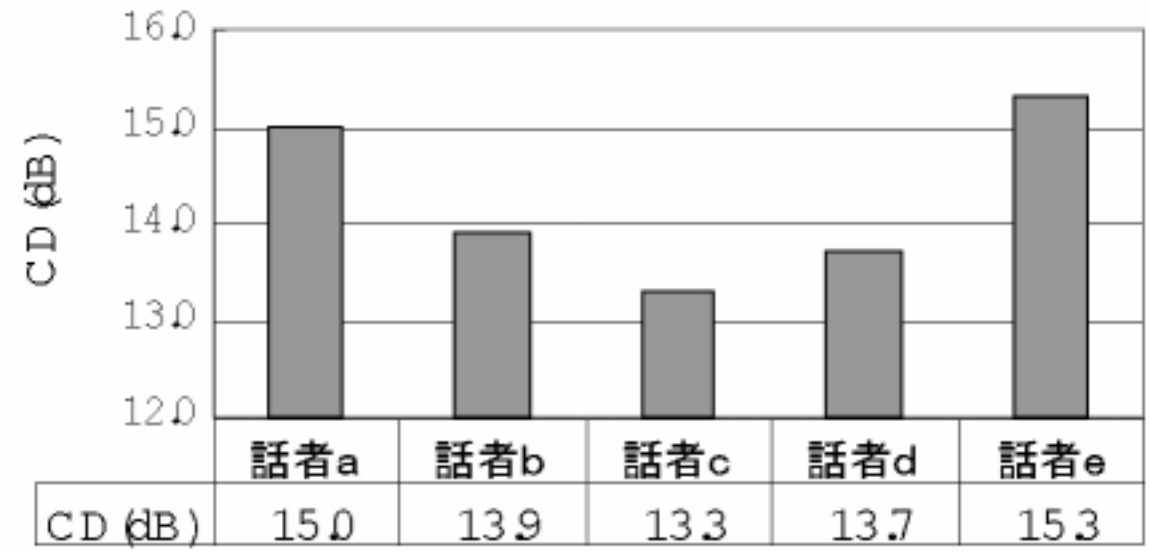

図 3.2: 平均から各話者の距離 (ケプストラム 30 次)

3.3.3 話者間の基本周波数パラメータの分析結果

ケプストラムと同様に基本周波数 F0 の距離を話者間で求め、それを元に多次元尺度構成法により 2 次元上に配置した (図 3.3)。また、図 3.4 には、平均からの各話者の距離を

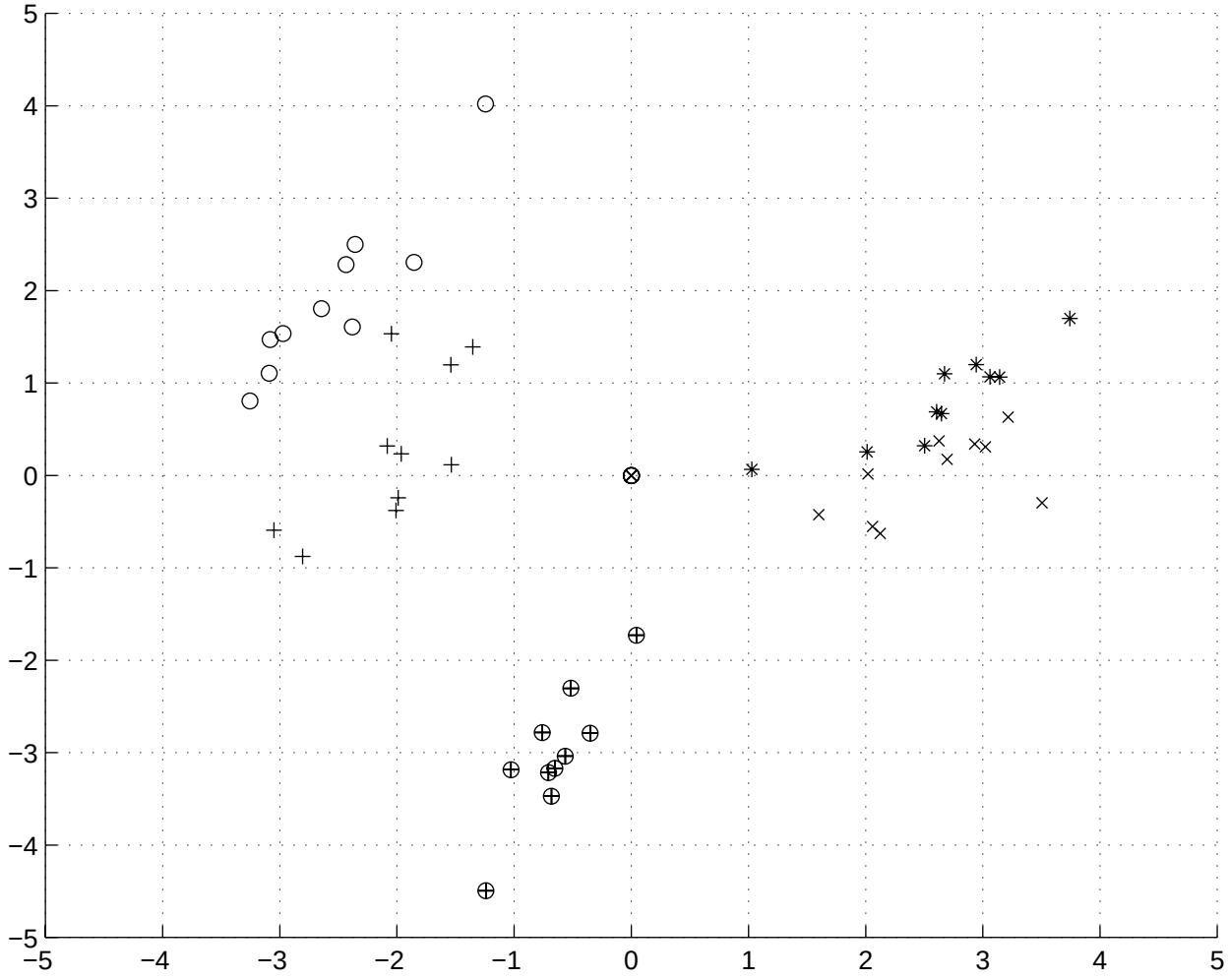

図 3.3: 基本周波数距離の付置図 (話者 a:○、話者 b:×、話者 c:+、話者 d:∗、話者 e:⊕、話者平均:⊗)

示している。

図 3.3 、図 3.4 から、平均から一番遠い話者は話者 a であり、話者 b、d は平均から同じ距離方向にあることがわかる。

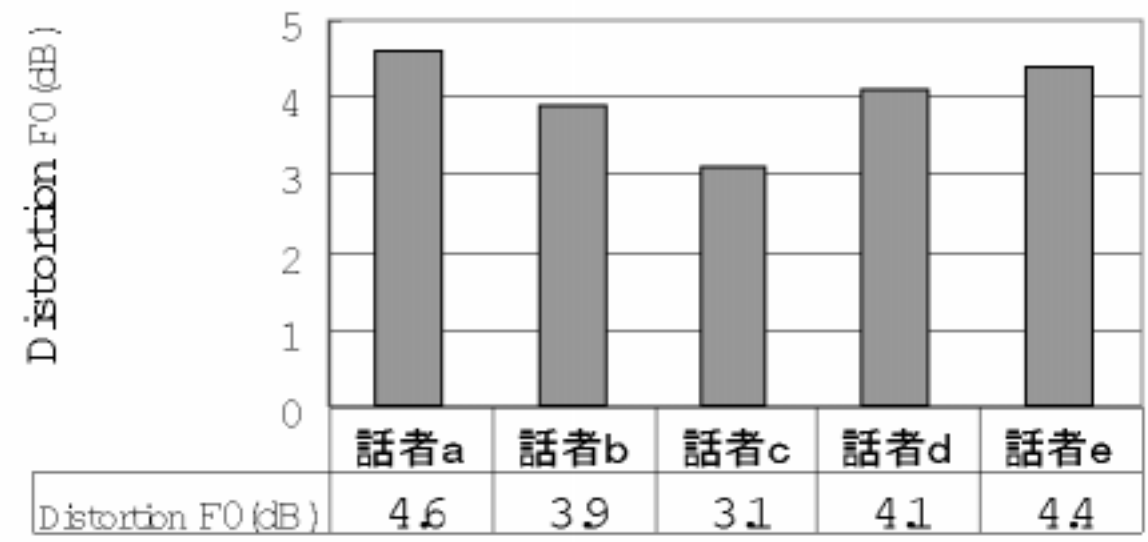

図 3.4: 平均からの各話者の距離 (基本周波数 F0)

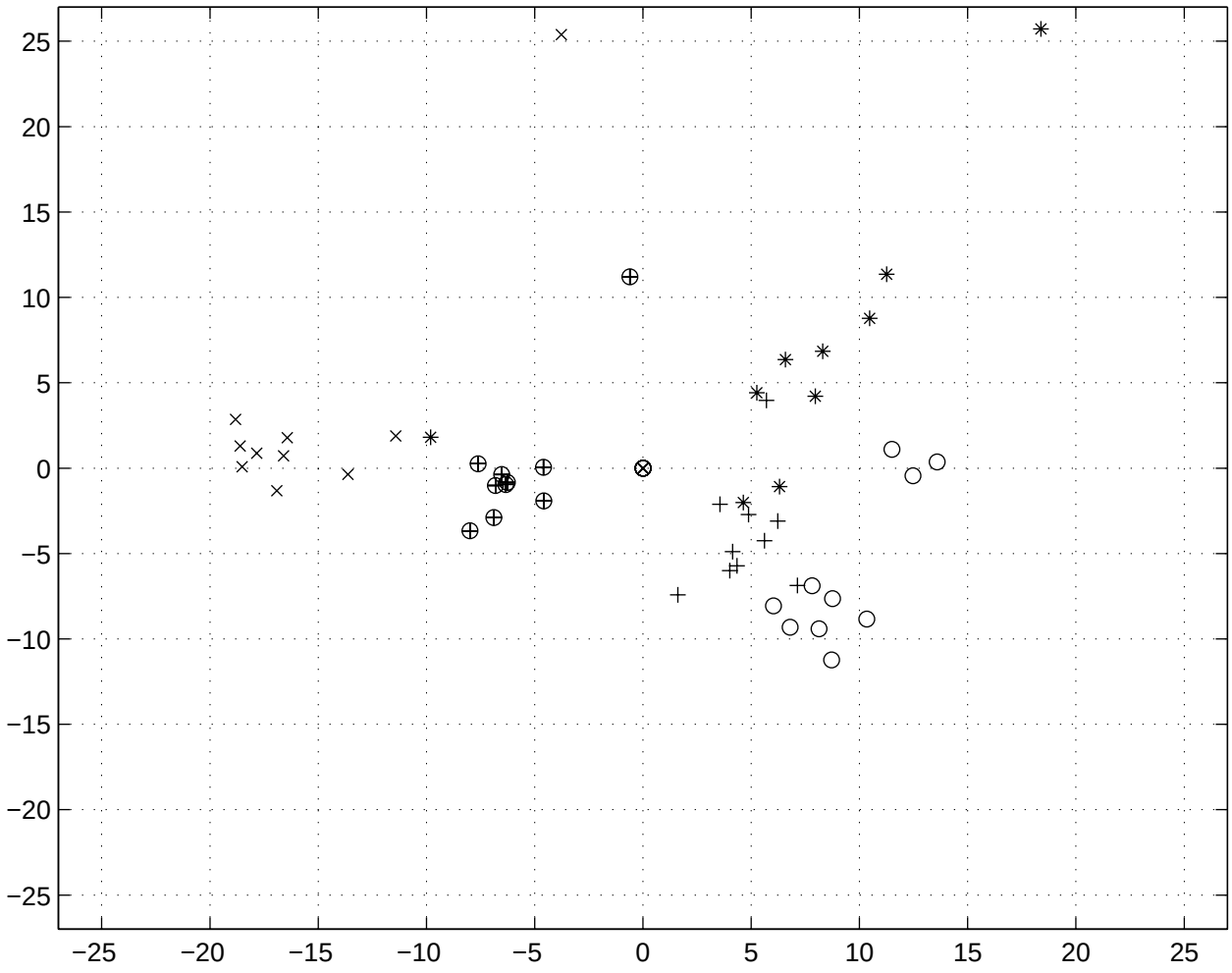

図 3.5: 振幅距離の付置図 (話者 a:○、話者 b:×、話者 c:+、話者 d:∗、話者 e:⊕、話者平均:⊗)

3.3.4 話者間の振幅パラメータの分析結果

図 3.5 に、振幅パラメータの距離から各話者の 2 次元分布を示してある。また、平均からの各話者の距離を図 3.6 に示してある。

図 3.5 、図 3.6 より平均から一番近い話者は話者 c である。

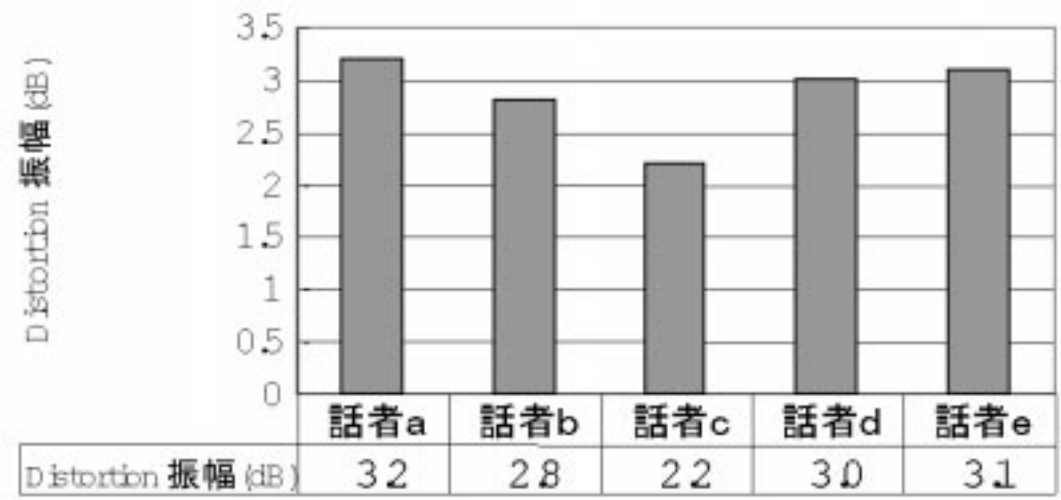

図 3.6: 平均からの各話者の距離 (振幅成分)

3.4 時間変化パターンの分析

3.4.1 分析方法

文音声「いいえ、うえにある」の各音韻の長さに個人差があるか検討をおこなう。

時間変化パターンの話者間の距離は、文音声中の各音韻長の話者間の距離 $\mathbf{DIST_t}$ を求め (式 (3.5))、それを元に多次元尺度構成法から 2 次元分布を求める。

$$\mathbf{DIST_t} = \sqrt{\sum_{i=1}^{K} (s_i^{(x)} - s_i^{(y)})^2} \tag{3.5}$$

ここで、$s_i^{(x)}, s_i^{(y)}$は話者 x、話者 y の i 番目の音韻長である。

また、後述する知覚実験の被験者から、文章の始まりの部分と文章の終わりで話者の差が大きいという内感報告があったことから、「いいえ」と「うえにある」に関して距離を求め 5 話者間の距離マトリックスから多次元尺度構成法により付置図を作成し、5 話者の位置関係を調べた。

3.4.2 分析結果

各話者で分析を行った結果を以下のように示す。

「いいえ、うえにある」を「いいえ」と「うえにある」で別々に多次元尺度構成法によりプロットした。

これらでは、「いいえ」だけの場合は各話者とも混在していることがわかり、「うえにあ

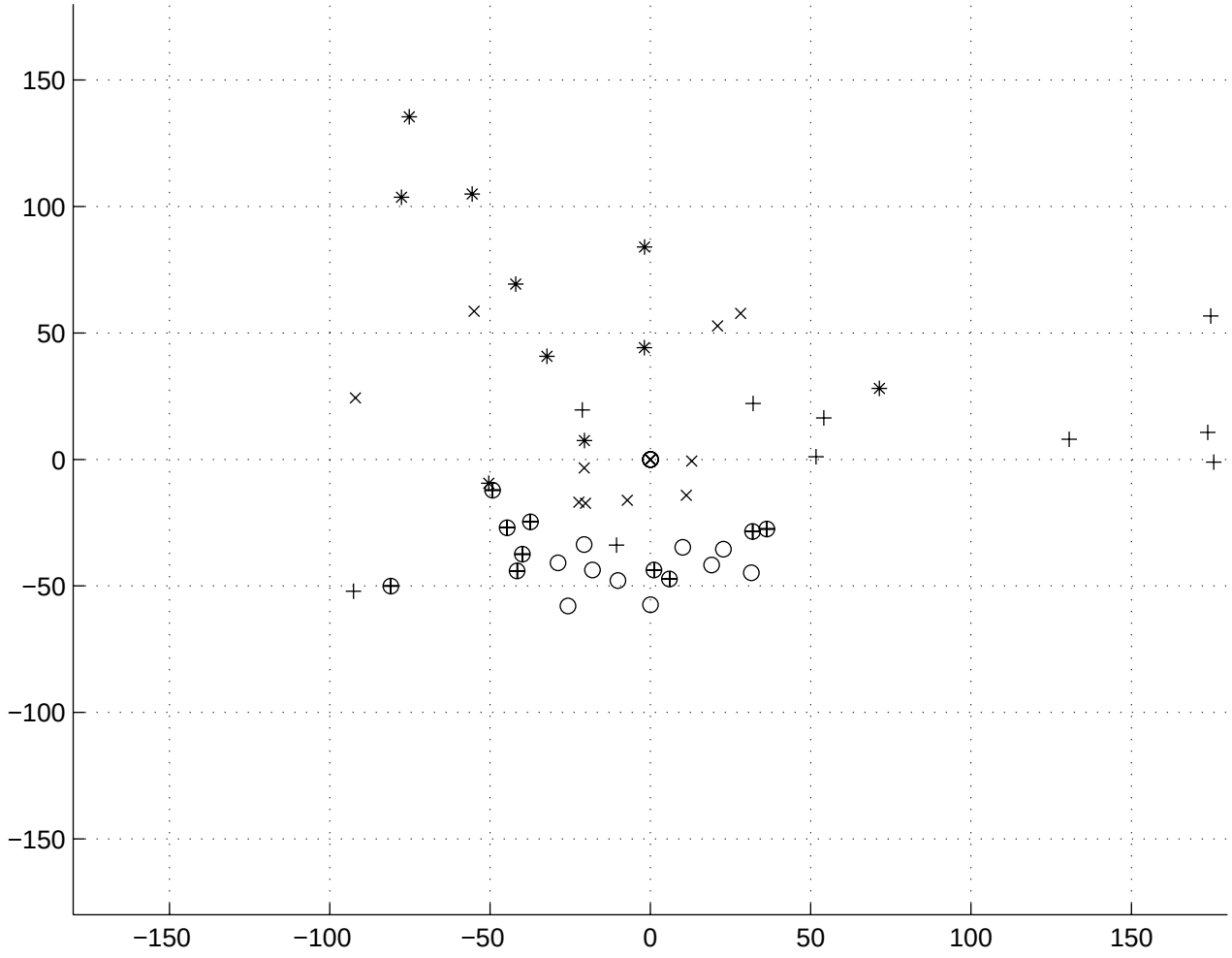

図 3.7: 音韻長距離の付置図「いいえ」(話者 a:○、話者 b:×、話者 c:＋、話者 d:∗、話者 e:⊕、話者平均:⊗)

る」だけの場合は話者 d が各話者よりも離れて、話者 b と話者 c が比較的離れて、話者 b と e が同じ位置にいるということがわかる。

3.5 まとめ

以上で述べたように物理パラメータについて各話者がどのように分布しているかを調査した。

ケプストラム距離の付置図 (図 3.1) では、各話者から離れている話者が話者 a、e、比較的同位置にあるのは話者 b、d であった。また、平均にもっとも近かった話者は話者 c

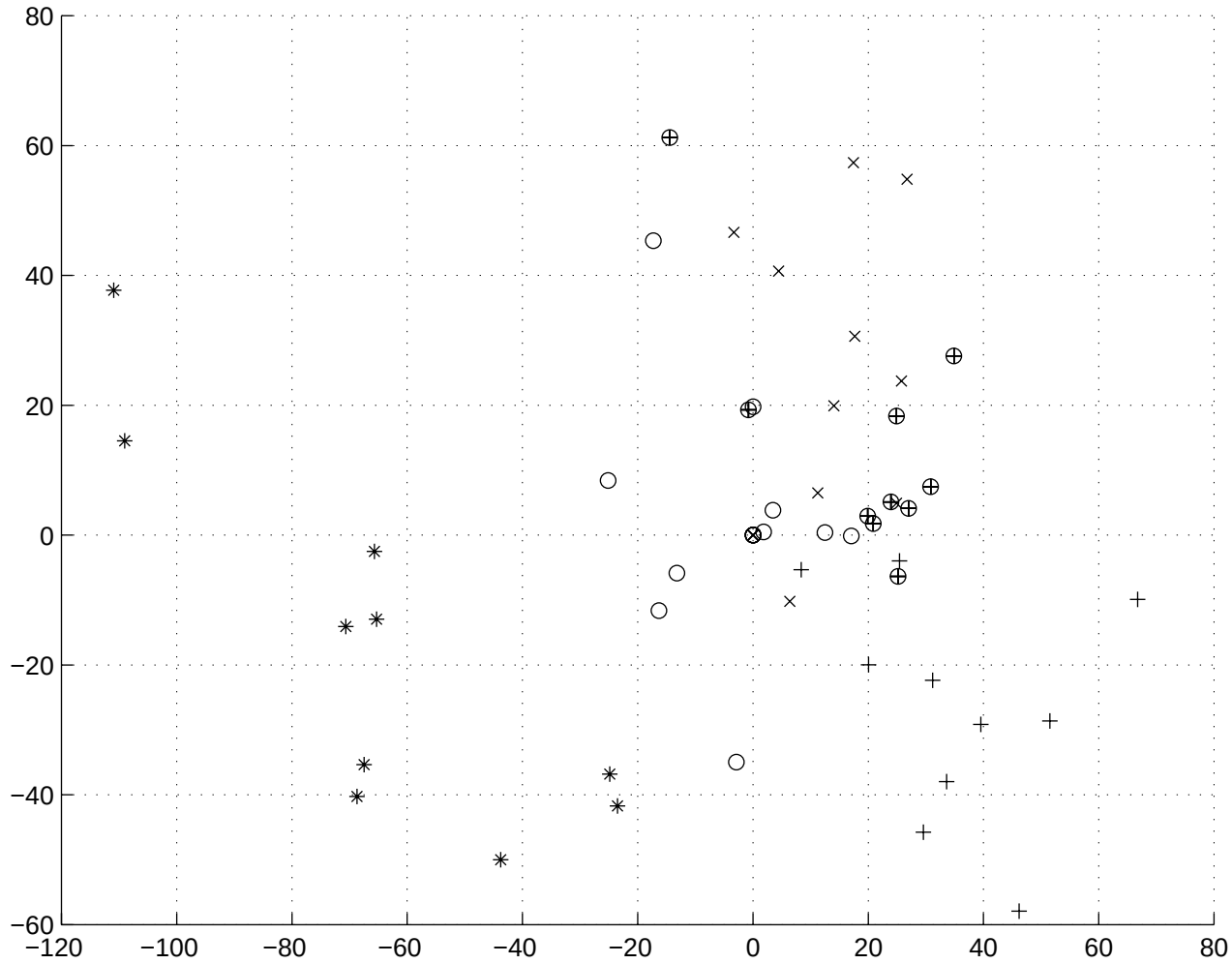

図 3.8: 音韻長距離の付置図「うえにある」(話者 a:○、話者 b:×、話者 c:＋、話者 d:∗、話者 e:⊕、話者平均:⊗)

である。

基本周波数距離の付置図 (図 3.1) では、平均からもっとも離れている話者は話者 a であり、比較的同位置にあるのは話者 b、d であった。

振幅距離の付置図 (図 3.1) では、平均からもっとも近かった話者は話者 c である。

音韻長の距離「いいえ」の付置図では話者が混在しているおり、あまり個人差が見られなかった。また、「うえにある」の距離を求めた付置図では、各話者から離れているのは話者 d であり、話者 b、c も離れてある。

第 4 章

知覚実験

4.1 目的

本章では、文音声中の 3 要素 (スペクトル、基本周波数、振幅) とこれらの時間変化が話者知覚に与える影響について聴取実験を行い調査する。

はじめに S^2BEL-TD で時間変化まで求めた合成音声に個人性が存在しているかを確認し、次に 3 要素 (スペクトル、基本周波数、振幅) に個人性が多く含まれているかを確認する。さらに各パラメータを話者間で入れ替えを行い、時間変化と 3 要素が話者知覚に影響を与えるかを確認していく。

以上のことをふまえて 3 つの実験をおこなう。

実験 1: S^2BEL-TD で時間変化まで求めた合成音声に個人性が存在しているかを確認する

実験 2: どのパラメータに個人性が多く含まれているを確認する。

実験 3: 時間変化を含めた 3 つの要素 (スペクトル、基本周波数) がどのように話者知覚に影響を与えるかを確認する。

4.2 実験 1：時間変化を考慮した合成音の個人性の存在確認

4.2.1 目的

実験 1 では、時間変化パターンを考慮した合成音声に個人性が存在することを確認する。

4.2.2 実験方法

音声データ

音声データは、前節で説明した録音条件で録音した音声で、36〜50 歳の本学の男性教官 5 名による文音声「いいえ、うえにある」を採用した。全部で 10 回録音したその中から、3 回分を用いた。

刺激音

聴取実験には以下の 4 種類の刺激音を用いた。

1-A. 原音声

1-B. STRAIGHT 分析合成音声

1-C. STRAIGHT で得られたスペクトルをスペクトルパラメータ (LSF30 次) まで分解し合成した合成音声

1-D. 1-C において、スペクトルパラメータを S^2BEL-TD を用いて時間構造、イベント位置に対するスペクトルパラメータや基本周波数などに分解し合成した合成音声

被験者

被験者は正常聴力を有し、音声データの収録の対象とした話者と日頃接している 22 歳から 36 歳の男性学生 10 名とした。

実験方法　実験は Naming 法により行った。一刺激につき 1 セットとし、計 4 セット行った。刺激音 1-A は、一話者につき 6 回、計 30 回でランダムに呈示した。刺激音 1-B〜D については、一話者につき 9 回、計 45 回をランダムに呈示した。刺激音は 8kHz の LPF により高域に発生するノイズを除去した。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。被験者には聞き直しを許し回答させた。回答は、PC のディスプレイ以上の話者の名前が書いてあるボタンをクリックすることにより行わせた。実験条件を表 4.1 に示す。

4.2.3 結果と考察

実験結果は、話者知覚できた割合を知覚率として図 4.1 に示す。

刺激音 1-B だけ知覚率が 99.8%と低くなった (他は 100%) が、刺激音 1-A と刺激音 1-B の話者知覚率に有意差があるか否かを有意水準 5%で F 検定を行ったところ、有意差が認

表 4.1: 実験条件 (実験 1)

話者	5 名
被験者	10 名
ヘッドフォン	SENNHEISER HDA 200 (両耳受聴)
ヘッドフォンアンプ	SANSUI AUα-907MR

められなかった (図 4.1)。これから刺激音 1-D は話者聴取実験に用いるのに十分な品質を有しているといえる。

4.3 実験 2：各物理量が話者知覚に与える影響の検討

4.3.1 目的

この実験では、スペクトル、基本周波数、振幅のうちどのパラメータが話者知覚に大きく影響するかを調べる。時間変化パターンについては個人性があるという前提でそのまま使用する。また、個人性に関係の少ないパラメータがわかれば、今後そのパラメータを考慮しないで実験を進めることが可能となり、被験者の負担を軽減することもできる。

4.3.2 実験方法

音声データ

4.3 節の実験 1 で用いたものと同じ、各話者 3 データを用いた。

刺激音

刺激音は以下の合成音声を用いた。パラメータの平均は全データの平均をおこなった。

2-A. 実験 1 の刺激音 1-D と同じもの。

2-B. 実験 1 の刺激音 1-D で、話者間で振幅パラメータを平均したもの。

2-C. 実験 1 の刺激音 1-D で、話者間で基本周波数パラメータを平均したもの。

2-D. 実験 1 の刺激音 1-D で、話者間でスペクトルパラメータ (LSF30 次) を平均したもの。

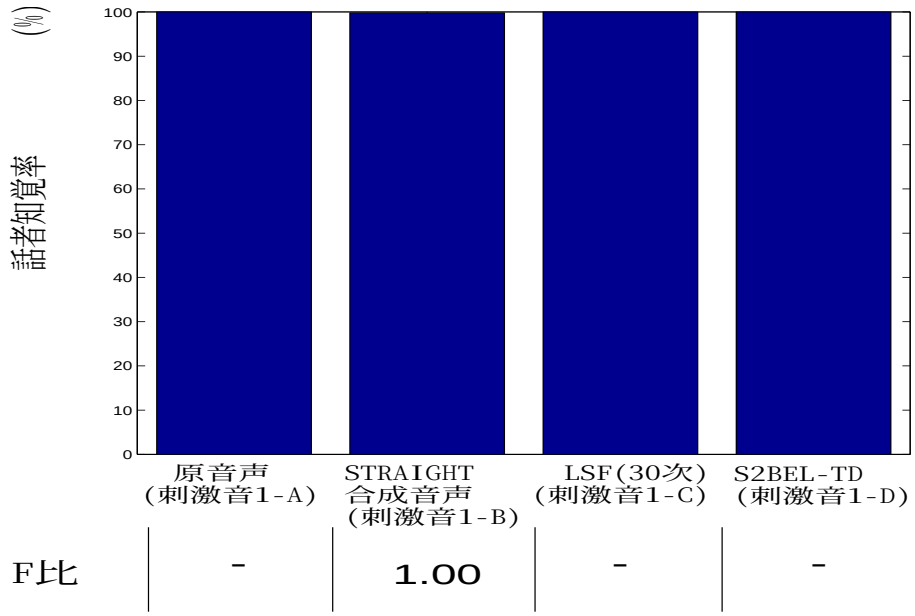

図 4.1: 話者識別率 {F(1,18:0.05)=4.41}

被験者

実験 1 と同じ、音声データの収録の対象とした話者と日頃接してたことのある男性 10 名。

実験方法　実験 1 と同様に、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。

実験は、5 話者のうちのどの話者かということを被験者に尋ねる Naming 法で行った。一話者につき 9 回、計 45 回をランダムに呈示した。被験者には聞き直しを許し、刺激音の話者を強勢判断させた。回答は、PC のディスプレイ以上の話者の名前が書いてあるボタンをクリックすることにより行わせた。実験条件を表 4.2 に示す。

4.3.3　結果と考察

実験結果は、図 4.2 のようになった。平均を行わないパラメータの話者と知覚した割合は、刺激音 2-A で 100%、刺激音 2-B で 99.6%、刺激音 2-C で 95.8%、刺激音 2-D で 79.6%である。

刺激音 2-A と各刺激音の話者知覚率の間に有意差があるか否かを有意水準 5%で F 検定を行った (表 4.3)。その結果、刺激音 2-A と 2-B の間には有意な差がなく、刺激音 2-A と 2-C、刺激音 2-A と 2-D には有意な差があることがわかった。

表 4.2: 実験条件 (実験 2)

話者	5 名
被験者	10 名
ヘッドフォン	SENNHEISER HDA 200 (両耳受聴)
ヘッドフォンアンプ	SANSUI AUα-907MR

表 4.3: F 検定

	振幅平均合成音声 (刺激音 2-B)	基本周波数平均合成音声 (刺激音 2-C)	スペクトル平均合成音声 (刺激音 2-D)
刺激音 2-A との F 比	1.00	29.8	43.4
刺激音 2-B との F 比	-	17.8	40.7

F(1,18:0.05)=4.41,F(1,18:0.01)=8.28

刺激音 2-D の知覚率が一番低いことがわかる。このとき元の話者以外だと知覚した率 20.4%のうち、77.2%が話者 c と回答した。スペクトルの分析 (図 3) の結果でも話者 c が平均から最も近いことから、このような結果が得られたと思われる。

また、刺激音 2-B と刺激音 2-C、2-D の話者知覚率の間に有意な差があるか否かを有意水準 5%で検定をおこなった (表 4.3)。その結果、刺激音 2-B と刺激音 2-C、刺激音 2-B と 2-D の話者知覚率の間に有意な差があることがわかった。この結果より、振幅パラメータが話者知覚に影響がないということがわかった。今後の実験では、振幅パラメータを話者間で平均したものを用いる。

4.4 実験 3：時間変化が与える影響と話者知覚の検討

4.4.1 目的

実験 3 は、時間変化パターンを含めた 3 つの要素 (時間変化パターン、スペクトル、基本周波数) が話者知覚にどのように影響を及ぼすかを調べる。そのために、3 つの要素を話者間で入れ替えを行い、聴取実験を行う。

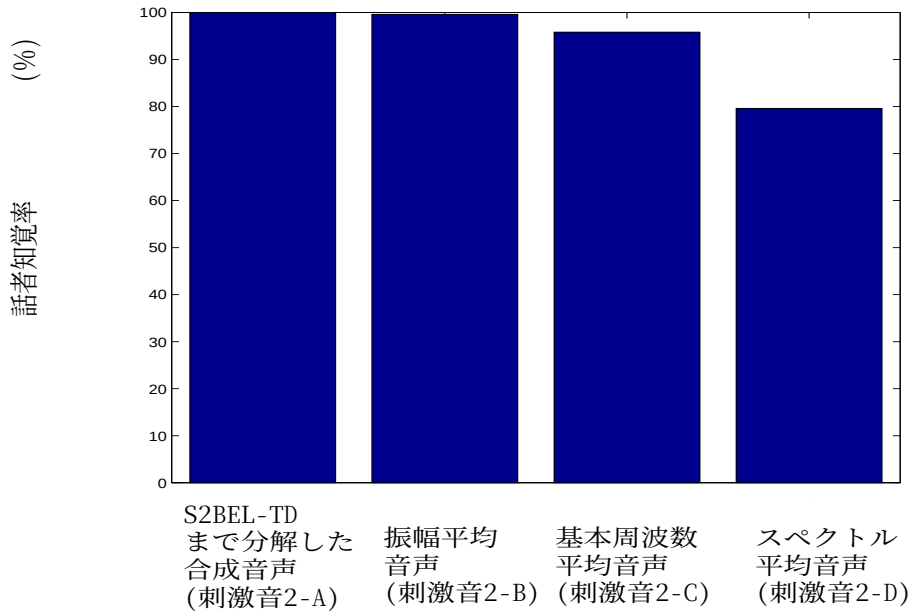

図 4.2: 話者知覚率

4.4.2 実験方法

音声データ

4.3 節の実験 1 で用いたものと同じ、男性教官 5 名による文音声。

刺激音

刺激音は、時間変化、スペクトルパラメータ (LSF)、基本周波数 (F0) の 3 つを話者間で入れ替えを行った合成音声を用いた。刺激音は 3 種類に分けることができる。括弧 () 中の数字は実験一セット当りの各刺激音の種類である。

3-A. 3 つのパラメータがすべて同じ話者のもの (5 通り)

3-B. 2 つのパラメータが同一話者のもの (60 通り)

3-C. 3 つのパラメータ全てが異なる話者のもの (60 通り)

被験者

他の実験と同じ男性学生 10 名である。

表 4.4: 実験条件 (実験 3)

話者	5 名
被験者	10 名
ヘッドフォン	SENNHEISER HDA 200 (両耳受聴)
ヘッドフォンアンプ	SANSUI AUα-907MR
受聴レベル	約 75〜80 dB (A)

実験方法　各話者の文音声の 1 データのパラメータ 3 種類を話者間で入れ替えを行い、一セットの合計を 125 回とし刺激音をランダムに被験者に呈示した。また、刺激音呈示レベルは約 75〜80dB(A) の範囲で、両耳にモノラルで呈示した。実験条件を表 4.1 に示す。

実験は前回と同様 Naming 法で行った。被験者には聞き直しを許し、刺激音の話者を強勢判断させた。回答は、PC のディスプレイ以上の話者の名前が書いてあるボタンをクリックすることにより行わせた。

実験は計 4 回行った。1 回につき各話者の文音声の 1 データを用いて行った。1 回目と 4 回目は同じ音声データを用いた。2 回目、3 回目については、各々音声データの種類を 1 回目と違うものを用いた。また、はじめの一回は分散が大きいとして除いた。したがって、実験結果は 3 回の実験の結果を示す。

4.4.3　実験結果

結果 1：全て同じ話者の場合　刺激音 3-A の知覚率は、3 回とも 100%であった (表 4.5)。ここで、知覚率は、被験者がその話者であると答えた割合を示してある。

結果 2：2 つが同じパラメータの場合　被験者が 2 つのパラメータの話者であると答えた割合を知覚率として、刺激音 3-B の知覚率を表 4.5 に示す。

LSF と時間変化、F0 と時間変化の組の回答率の内訳は、次の表 4.6 、表 4.7 のようになる。表 4.6 、表 4.7 では、2 つが同じ話者のパラメータ (左側の話者 a〜e) の刺激を被験者がどの話者 (上の話者 a〜e) に回答したかを回答率としてあらわしてある。other は 3 つの成分で構成する以外の話者と答えた割合である。

表 4.5: 話者知覚率 (結果 1,2)

	LSF	F0	時間変化	知覚率 (%)
結果 1	○	○	○	100
結果 2	○	○	other	98.8
	○	other	○	83.8
	other	○	○	20.3

表 4.6: 回答の内訳 (時間変化と LSF が同 一話者)

	話者					
	a	b	c	d	e	other
話者 a	95.0*	0	0	1.7	1.7	1.7
話者 b	0	78.3*	5.8	6.7	0.8	8.3
話者 c	0	3.3	84.2*	1.7	0	10.8
話者 d	0	5.0	5.8	67.5*	0.8	20.8
話者 e	0	0	1.7	1.7	94.2*	2.5

結果 3：すべて違うパラメータの場合　表 4.8 に結果を示す。知覚率は、そのパラメータの話者であると知覚した割合を示したものである。また、表 4.8 のその他の知覚率は、3 つの要素を構成する以外の話者と知覚した割合を示してある。また、表 4.9 には、刺激音 3-C を回答した内訳を示す。左側の話者 a〜e は LSF の話者を示してあり、上側の話者 a〜e は被験者が回答した話者を示し、other は 3 つの要素を構成する以外の話者と答えた割合である。

4.4.4　考察

考察 1：全て同じ場合　3 つとも同じ話者の要素を用いた場合、被験者は正確に知覚している。

考察 2：2 つが同じパラメータの場合　LSF と F0 が同一話者のものがいちばん話者知覚の割合が高く、個人性に関与しているといえる。また、LSF と時間変化、F0 と時間変化

表 4.7: 回答の内訳 (時間変化と F0 が同一話者)

	話者					
	a	b	c	d	e	other
話者 a	0.8*	12.5	25.0	5.8	24.2	31.7
話者 b	23.3	29.2*	14.2	7.5	24.2	1.7
話者 c	25.0	12.5	25.0*	9.2	22.5	5.8
話者 d	23.3	3.3	8.3	40.0*	22.5	2.5
話者 e	23.3	19.2	22.5	20.8	6.7*	7.5

表 4.8: 話者知覚率 (結果 3)

回数 (60 回)	知覚率 (%)
LSF	75.1
F0	11.1
時間変化	7.0
その他	6.8

の組合せを比べると LSF と時間変化の組が、より話者知覚されていることから LSF と時間変化の組の方が個人性に関与しているといえる。

時間変化と LSF の回答の内訳を示した表 4.6 を見ると、時間変化と LSF がある話者の時、その話者だと知覚した割合が高いことがわかる。特に話者 a、e では 90%以上の回答率を示している。一方、話者 b や話者 c、話者 d では若干知覚の割合が低い。このことから話者 a、e では、時間変化と LSF がより個人性に関与しており、また話者 b、c、d では、他の話者 b、c、d の F0 の要素に知覚がうつっていることから、LSF と時間変化の個人性の関与がより薄いことがわかる。

また、F0 と時間変化の回答の内訳を示した表 4.7 を見ると、F0 と時間変化が話者 a、e のとき、その話者と知覚した割合が少なく、他の話者の LSF に知覚がうつっている。このことから話者 a、e では、F0 と時間変化がほとんど個人性に関与していないといえる。

また、F0 と時間変化が話者 b や話者 c、話者 d では、その話者だと知覚した割合が比較的高い。話者 d に至っては 40%も知覚されている。これから、話者 b、c、d は、F0 と時間変化が若干個人性に関与しているといえる。

表 4.9: 回答の内訳 (全てが違う話者の場合)

	話者					
	a	b	c	d	e	other
話者 a	95.6*	0	0	2.8	1.1	0
話者 b	0	62.5*	21.1	13.9	2.5	7.5
話者 c	6.1	6.9	76.9*	9.7	0	5.0
話者 d	0	28.6	25.6	44.7*	1.1	19.7
話者 e	0	1.1	2.8	0	95.6*	1.7

考察 3：すべて違うパラメータの場合

表 4.8 では、3 つの要素のうち LSF の知覚率が高く、LSF が話者知覚に与える影響がより強い。3 つとも違う話者のパラメータを用いた合成音声でスペクトルパラメータの知覚に対する内訳の結果 (表 4.9) から、話者 a、e では、その話者と知覚した割合が高い。このことから、話者 a、e では、LSF がより個人性に関与していることがわかる。一方、話者 b、d では、その話者と知覚した割合が低く、半数近くがその他の要素 (F0 や時間変化) に知覚がうつっている。

また、表 4.8 より時間変化の知覚率が低いことから話者知覚にあまり影響を及ぼしていないと言える。

4.5 まとめ

3 つの聴取実験を通じて以下のようなことが明らかになった。

すべての話者についていえることは、次のようになった。

- 時間変化を考慮した合成音声に個人性が存在すること (実験 1)
- 時間変化以外の 3 つの要素 (スペクトル、基本周波数、振幅) のうち、個人性の関与しないものは振幅であること (実験 2)
- LSF が最も個人性に関与する (実験 2、実験 3 の結果 3)
- 時間変化は話者知覚にあまり影響を与えない (実験 3 の結果 3)

知覚に関して 5 話者は、大まかにわけて 2 パターンにわかれる (実験 3 の結果 2、3 より)。

- LSF(LSF と時間) のみで知覚できる話者
- F0 と時間変化の影響を受けやすい話者

第 5 章

全体の考察

本研究では文音声の個人性関係物理量のうちスペクトル、基本周波数、その変化を総合的に取り扱い、それぞれの寄与、関与を調べてきた。

本章では本研究で得られた結果の考察を行い、過去の研究との関係を調べる。

先天性の個人性情報

聴取実験の全体を通じて、3 つの要素 (スペクトル、基本周波数、時間変化) のうち、最も話者知覚に影響を与えたのはスペクトルであった (実験 2、実験 3 の結果 2、3 の結果より)。S^2BEL-TD より得られたイベント位置のスペクトルは静的成分であり、先天性の情報すなわち声道形状の特性が含まれている。このことから声道特性が話者知覚に与える影響が大きいと思われる。既知話者の条件のもと、静的スペクトルが一番話者知覚へ与える影響が大きいという結果は、伊藤 [1] ら、北村 [5] ら、橋本ら [8] の報告と一致する。

さらに、イベント位置のスペクトル、基本周波数の要素が同一話者のものであれば、どの話者でもその話者と知覚された (実験 3 の結果 2)。スペクトル、基本周波数の静的成分には声道特性と声帯特性が含まれており、先天性の個人情報が大きく話者知覚に影響すると考えられる。逆に言うと、時間変化には個人性があまり関与されていないと言える。聴取実験では、時間変化を話者間で入れ替えをおこなった。時間変化が違う話者と入れ替えられれば、音韻持続時間も入れ替わる。この結果は、音韻持続時間は話者知覚に大きな影響を与えないことを示している。これらは、伊藤ら [1]、北村ら [5]、橋本ら [8] の報告と一致する。

後天性の個人性情報

スペクトルの時間変化、基本周波数の時間変化は後天性の情報が含まれている。

聴取実験の結果 (実験 3 の結果 2、3) から、5 話者が『スペクトルと時間変化で知覚できる話者』と『基本周波数と時間変化の影響を受けやすい話者』に分かれた。結果から、前者は話者 a、e であり、後者は話者 b、c、d である。ここで、スペクトルの分析結果 (図 3.1) を見てみると、各話者から離れている話者は話者 a、e である。また、話者 b、c、d は、比較的似通った位置にある。このことをふまえると、『スペクトルの距離が遠い (スペクトルが違っている) 話者では、スペクトルと時間変化 (またはスペクトルのみ) を cue』とし、『スペクトルの距離が近い (スペクトルが似通っている) 話者では、基本周波数と時間構造 (または基本周波数のみ) を cue』として、話者知覚していることが考えられる。これらの結果は ABX 法を使って実験をおこなった橋本ら [8] の結果と類似するものとなった。

3 要素 (スペクトル、基本周波数、その変化) の個人性情報

この 3 つの要素を総合的に取り扱い、本研究は行われてきた。3 つの要素を総合的に取り扱うことでどんな意味を持つかを本研究で得られた結果をもとに、明らかにしていく。

話者間で 3 つの要素の入れ替えをおこなった聴取実験では、スペクトルが話者知覚に与える影響が大きいという結果 (実験 2、3) をもたらした。また、スペクトルが似通った話者ならば、基本周波数が話者知覚に影響を与えるという結果 (実験 3 の結果 2、3) があきらかになった。しかし、その時間変化が話者知覚に与える影響は見られなかった (実験 3 の結果 2)。

このことから、3 つの要素を総合的に取り扱った場合、『話者知覚の影響を与えるために十分なパラメータは、2 つの要素 (スペクトル、基本周波数) である』ことを明らかにした。

第 6 章

結論

6.1　本論文であきらかになったことの要約

本論文では、話者知覚に与える物理量をあきらかにする為に基本周波数、スペクトル包絡の双方の時間的変化を表す物理量を求め、基本周波数の時間変化、スペクトルの動きについて総合的に取り扱い、個人性に関する検討を行った。音声中の物理量の入れ替えを行うことで話者知覚に与える影響を聴取実験により調べ、その関係から個人性を表す物理量を求めた。

その結果、『スペクトルが違っている話者では、スペクトルを cue』とし、『スペクトルが似通っている話者では、基本周波数を cue』として、話者知覚していることを示した。

また、以上のことをふまえ、総合的に 3 つの要素を用いることであきらかになったことは、『話者知覚に影響を与える為に十分な要素は、スペクトルと基本周波数である』ことがわかり、その時間変化が話者知覚にあまり影響を与ないことがわかった。

6.2　今後の課題

以下に今後の課題を列挙する。

- 話者の性別の問題
  本研究の Naming 法による聴取実験で用いた音声データの話者はいずれも男性のみであった。そのため、実験結果が話者セットに依存している可能性は否めない。今後、大規模な話者セットによる聴取実験を行い、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。

- 文音声の数の問題
  本研究で対象になった文音声は有声子音と母性子音で構成された 1 文章の 10 セットを用いた場合であった。そのため音声データセット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セットによる聴取実験を行い、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

日頃ご指導いただき、貴重なご助言をいただきました赤木正人助教授、岩城護助手をはじめとする本学の教官の皆様、熱心にご討論いただいた赤木研究室をはじめとする本学の学生、OB の皆様に感謝いたします。また、ご多忙の中、音声を録音させていただいた 5 名の教官の皆様、聴取実験に参加いただいた 10 名の学生の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下さった全ての皆様に厚く感謝いたします。

本研究は科学技術振興事業団 (CREST) の援助を受けて行われた。

参考文献

[1] 伊藤, 斉藤：“音声の音響的特徴パラメータが個人性に及ぼす影響”, 電学論，**J65-A**, 1，pp. 101-108 （1982）

[2] M. Akagi and T. Ienaga：“Speaker individualities in fundamental frequency contours and its control”, J. Acoust. Soc. Jpn. (E) **18**, 2 (1997)

[3] Tatsuya Kitamura、Masato Akagi：“Speaker individualities in speech spectral envelopes”, J. Acoust. Soc. Jpn(E), **16**, 5 (1995)

[4] Weizhong Zhu、Hideki Kasuya：“Roles of static and dynamic features of formant trajectories in the perception of talk indedivduality”, Eurospeech97, Rhodes, Greece. ISSN 1018-4074, pp.2195-2198

[5] 北村達也、赤木正人、北沢茂良：“スペクトル遷移パターンが個人性知覚に与える影響について” 聴覚研究会, H98-97 (1998)

[6] 河原英紀：“音声分析・変換・合成方法 STRAIGHT-TEMPO における相補的な時間窓の利用について”, 聴覚研究会, H97-47 (1997)

[7] A.C.R.Nandasena and M. Akagi “Spectral stability based event localizing temporal decomposition” Proc. ICASSP98, II, 957-960

[8] 橋本誠、北川敏、樋口宣男 “音声の個人性知覚に影響を及ぼす音響的特徴の定量的分析” 音響学会誌, 54 巻 3 号 (1998)

学会発表リスト

鈴木、赤木：“文音声中に含まれる個人性情報の知覚”，音声研究会（1999.3 発表）