修 士 論 文

文音声中の基本周波数の時間変化に含まれる個人性に関する研究

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

大野 宏

1998 年 2 月 13 日

目 次

図 目 次

表 目 次

第 1 章

緒言

今日実用化が望まれている音声を用いたマンマシンインタフェース技術には音声認識、規則音声合成、話者認識等があるが、その実用化のためには音声に含まれる個人性をどのように扱うかが問題となる。そのため個人性に関する研究はかねてより行なわれてきた。

過去の研究から音声における個人性は、音源特性である基本周波数や声帯波形、声道特性であるスペクトル包絡やホルマント周波数等に含まれるとされている。これらをさらに分類すると以下のようになる。

- 声帯特性
    1. 平均ピッチ周波数
    2. ピッチ周波数の時間変化パターン
    3. ピッチ周波数の揺らぎ
    4. 正門体積波形
- 声道特性
    1. ホルマント周波数の値
    2. スペクトル包絡の時間変化パターン
    3. スペクトル包絡の形と傾斜
    4. 平均スペクトル包絡特性

これらの個人性に関する研究としては様々な研究がなされ、スペクトル包絡や平均基本周波数といった静的な特徴量により多く個人性が含まれているといった報告もなされているが、まだ最も個人性を含む特徴はどれかということは明らかにされてはいない [1, 2]。

しかし、近年基本周波数が音韻の認識に影響を与えることや基本周波数の時間変化のような音声の動的変化に個人性が多く含まれることが注目され、静的な情報だけでなく、スペクトル包絡、基本周波数、時間構造を総合的に考えた音響特徴距離と個人性知覚との関係などが報告されている [3]。また、基本周波数の時間変化 (以下基本周波数パターン) を対象とした個人性に関する研究も幾つか行なわれており、過去に 3 モーラ単語の基本周波数パターンの個人性について分析検討もなされている [4]。だが、前述のマンマシンインタフェースを実用化するためには単語だけでなく文音声に含まれる個人性情報についても明らかにしておかなければならない。

そこで、本論文では文音声中での基本周波数パターンに着目し、そこに含まれる個人性情報について調べる。ここでいう個人性とは、同じ話者の発話が同じ話者のものだとわかることであると定義する。手法としては、基本周波数パターンをパラメータ化するため基本周波数パターン記述モデルとして生理学的見地からも理論が検証されている藤崎モデルを用い、各パラメータに現れる個人差について分析を行なう。次にスペクトル包絡を他話者のものに変換した音声 (以下スペクトル包絡変換音声) による聴取実験によって、基本周波数パターンに個人性情報が存在することを示す。最後に、基本周波数パターンのパラメータを他話者と変更した音声 (以下基本周波数変形合成音声) により個人性知覚への影響を調べ、単語音声の場合との比較検討を行なう。

第 2 章

音声分析合成系

2.1 はじめに

本論文では基本周波数を変形した合成音の作成を行なう為に、次のような音声合成系を使用した。本章では、この音声合成系の各過程について簡単に説明する。

2.2 STRAIGHT

聴取実験に用いる音声は、基本周波数パターン以外の情報が各話者間で同一で、基本周波数パターンの操作が可能である必要がある。そのような合成音声を作成するためには、何らかの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 Adaaptive Interpolation of weiGHTed spectrogram) [5] を採用する。

STRAIGHT は 、時間周波数表現を求める STRAIGHT-core、駆動音源の位相を操作する SPIKES、基本周波数を求める TEMPO、の 3 つの要素から構成される音声分析変換合成法であり、原音声に匹敵する自然な音声の合成が可能である。

本論文で用いた合成音は、全て STRAIGHT により作成した。基本周波数変形合成音の作成は、変形した基本周波数パターンを用いて音声の合成を行なう。

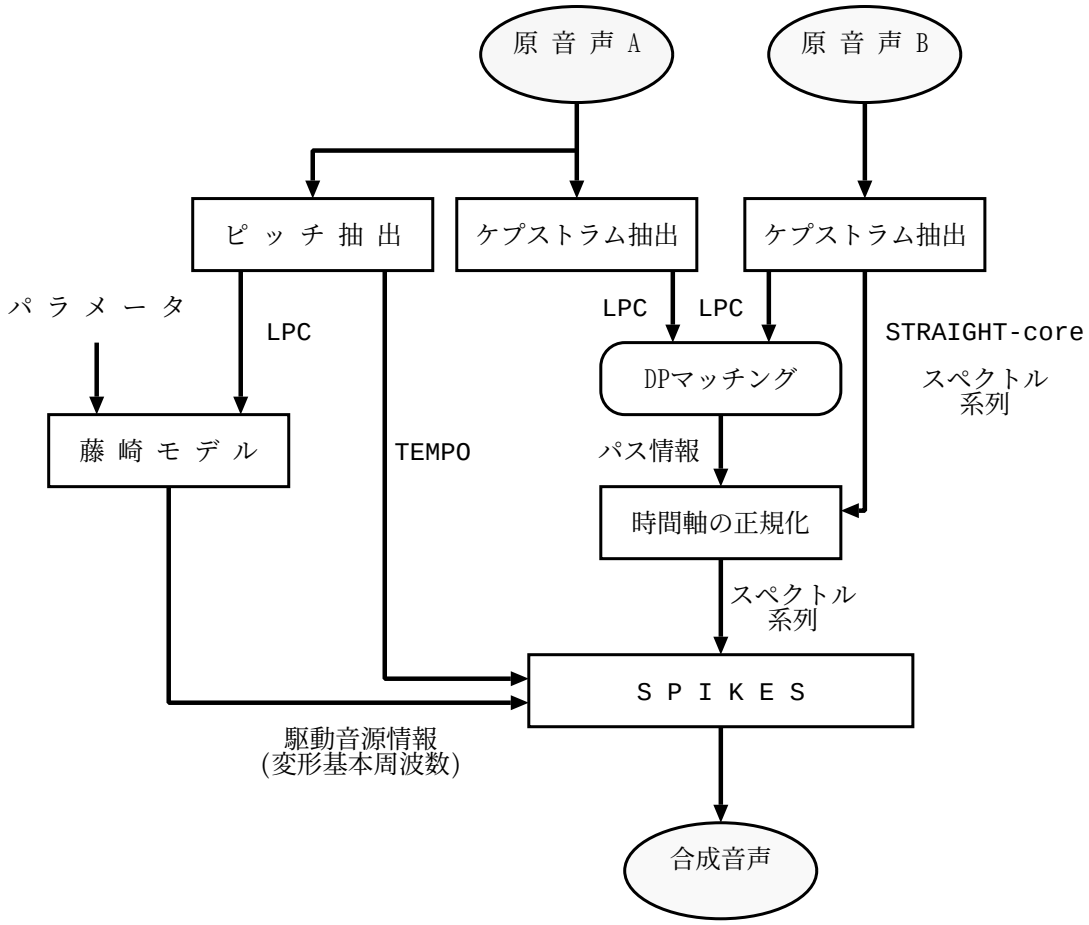


図 2.1: 音声合成系

2.3 藤崎モデル

日本語はいわゆるピッチアクセントの言語でありピッチパターンの様相が言語認知に強い影響を与えていることは良く知られている [9]。故に日本においては古くからピッチパターン生成のモデルの研究が行なわれてきた。

本論文では、基本周波数パターンの操作を行なうために基本周波数記述モデルとして藤崎モデル [6, 7] を採用する。

ピッチ周波数 F_0 の時間変化に現れる個人性をいくつかのパラメータとして表すために時間変化パターンをモデル化する必要がある。ピッチパターン生成モデルには点ピッチモデル [8] などのモデルもあるが、本研究では下記の理由により藤崎モデルを採用した。

- ピッチパターンをフレーズ成分とアクセント成分と呼ばれる２つの要素に分割しており、一般的な認識と整合性がよい。
- 生理学的見地から理論が検証されている。
- 日本語の規則音声合成で一般的に使用されている。
- ピッチ周波数の時間変化パターンの記述に要するパラメータの数が少ない。

藤崎モデルそれ自身は規則合成音の高品質化を目的として作成され、研究されてきたが、ピッチ周波数の時間変化に含まれる個人性のうち、大きな変化に関する個人性については十分表現可能であると考える。

藤崎モデルでは ピッチパターンの生成について以下のような仮定を行なっている。

1. フレーズコマンドはインパルスの並びで表現され、フレーズ成分は入力に対する臨界制動２次の線形システムの応答で表される。
2. アクセントコマンドはステップ関数の並びで表現され、アクセント成分は入力に対する臨界制動２次の線形システムの応答で表される。
3. フレーズおよびアクセント成分は対数軸上で重ね合わされ F_0 の時間変化パターンを記述する。

このような仮定に基づいた システムの出力は以下のように定義される。

$$\begin{aligned}\ln F_0 = \ln F_{\min} &+ \sum_{i=1}^{I} A_{\mathrm{p}i} G_{\mathrm{p}i}(t - T_{0i}) \\ &+ \sum_{j=1}^{J} A_{\mathrm{a}j}\{G_{\mathrm{a}j}(t - T_{1j}) - G_{\mathrm{a}j}(t - T_{2j})\},\end{aligned} \tag{2.1}$$

$$G_{\mathrm{p}i}(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t) & (t \geq 0), \\ 0 & (t < 0), \end{cases} \tag{2.2}$$

$$G_{\mathrm{a}j}(t) = \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t),\ \theta_j] & \\ & (t \geq 0), \\ 0 & (t < 0), \end{cases} \tag{2.3}$$

ここで $G_{\mathrm{p}i}(t)$ と $G_{\mathrm{a}j}(t)$ は、それぞれフレーズ制御機構のインパルス応答とアクセント制御機構のステップ応答である。式中のパラメータを以下に示す。

$F_{\min}$: 基底周波数
I: フレーズコマンド回数 $(i = 0, 1, \cdots I-1)$
J: アクセントコマンド回数 $(j = 0, 1, \cdots J-1)$
$A_{\mathrm{p}i}$: i 番目のフレーズコマンドの大きさ
$A_{\mathrm{a}j}$: j 番目のアクセントコマンドの振幅
T_{0i}: i 番目のフレーズコマンドの生成時刻
T_{1j}: j 番目のアクセントコマンドの開始時刻
T_{2j}: j 番目のアクセントコマンドの終了時刻
α_i: i 番目のフレーズ成分の固有角振動数
β_j: j 番目のアクセント成分の固有角振動数
θ_j: j 番目のアクセント成分の天井値

フレーズコマンド、アクセントコマンドの回数は通常視察により決定されるが、発話終了時のピッチパターンの急激な降下に対応する為、負の値のフレーズコマンドが文の終端で 1 つ余分に使われる。この終端のフレーズコマンドの大きさの絶対値はそれまでのすべてのフレーズコマンドの大きさの和に等しい。

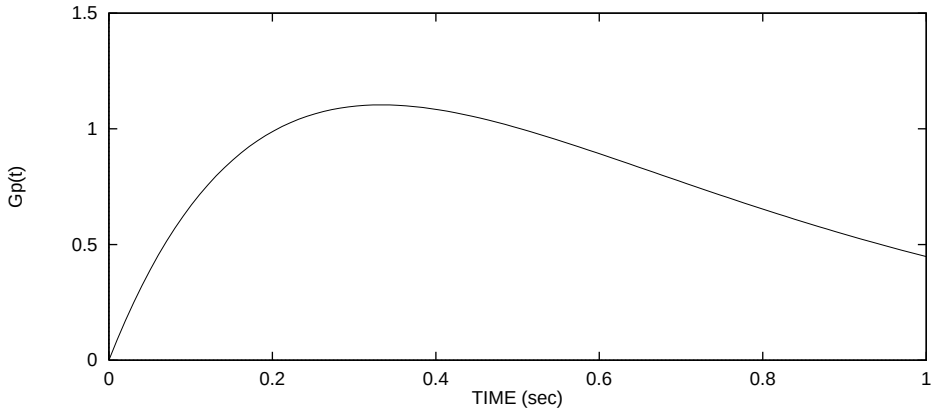
以下、図 2.2 にフレーズ成分のインパルス応答とアクセント成分のステップ応答を示す。また、藤崎モデルによるピッチパターンの生成過程を図 2.3 に示す。

生成過程の流れは以下のようになっている。

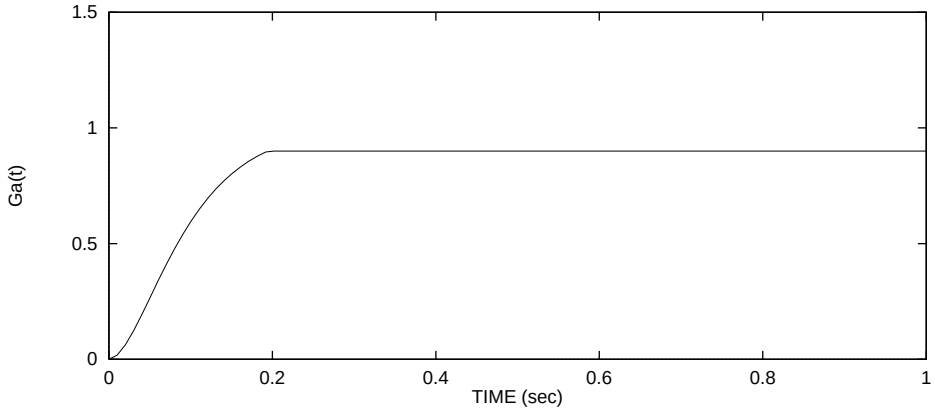
1. フレーズコマンドのインパルス信号列とアクセントコマンドのステップ関数列をそれぞれフレーズ生成機構、アクセント生成機構に入力する

2. それぞれに入力に応じて、フレーズ生成機構のインパルス応答としてフレーズ成分を、アクセント生成機構のステップ応答としてアクセント成分を生成する

3. 対数周波数軸上でフレーズ成分とアクセント成分を加算し、基底周波数 $\ln F_{\min}$ を足して、出力波形 $\ln F_0(t)$ を得る



(a) フレーズ成分 $G_p(t)$ （ $\alpha = 3.0\,\mathrm{s}^{-1}$）



(b) アクセント成分 $G_a(t)$ （$\beta = 20.0\,\mathrm{s}^{-1}$,$\theta = 0.9$）

図 2.2: フレーズ成分、アクセント成分の形状

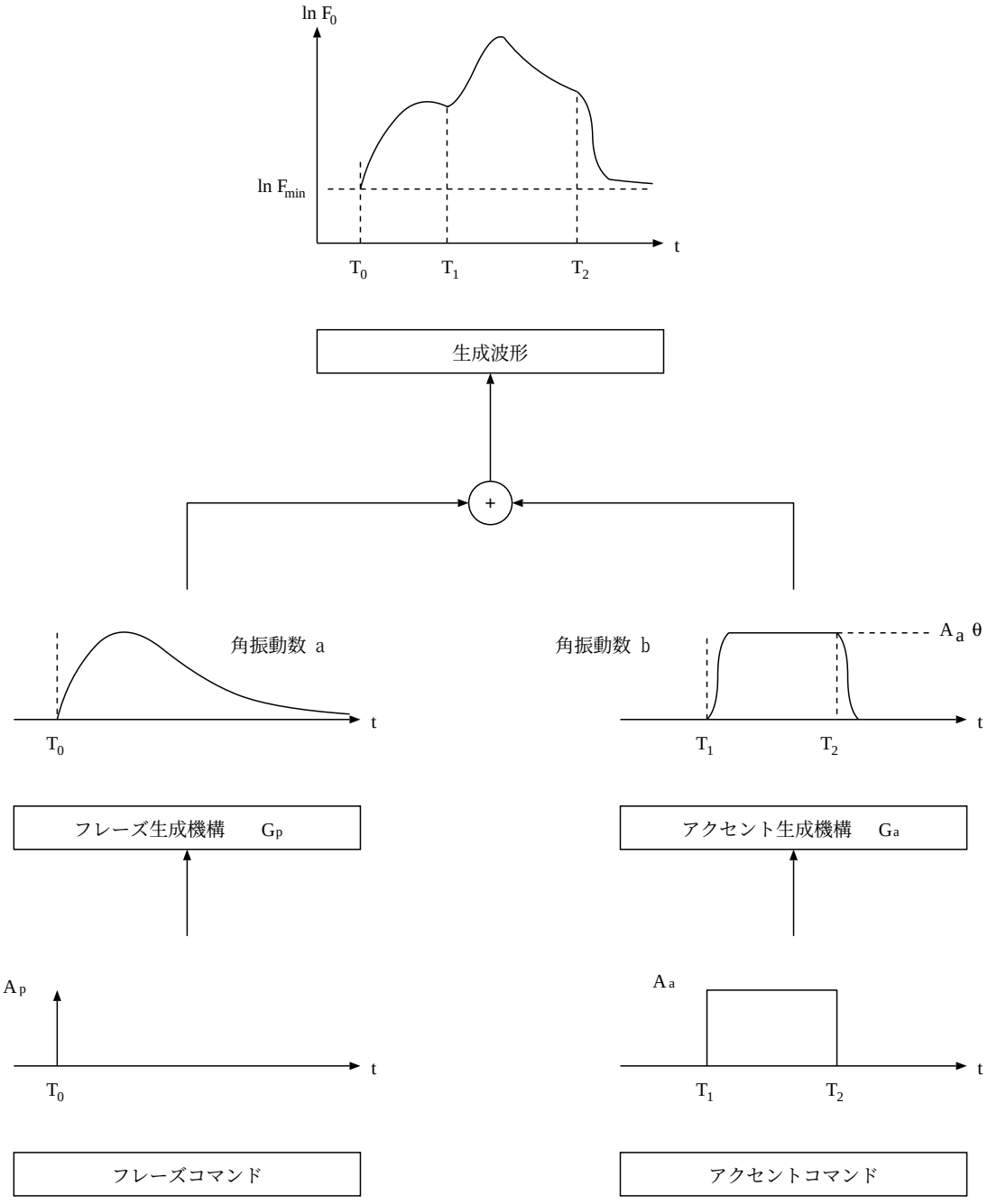


図 2.3: 藤崎モデルのピッチパターン生成過程

2.4 時間軸の正規化

時間軸の正規化は、原音声から LPC ケプストラムを計算し LPC ケプストラム距離尺度による DP マッチングを行ない DP パスを求め、この DP パスを STRAIGHT-core で求めたスペクトル系列に適用することにより行なう。信号 S_x、S_y の LPC ケプストラム距離尺度 $d(S_x, S_y)$ は次式で表される。c_iは i 次の LPC ケプストラムである。

$$d(S_x, S_y) = \sqrt{2\sum_{i=1}^{p}(c_i^x - c_i^y)^2} \tag{2.4}$$

また本論文では、DP マッチングの局所パス規制には図 2.4のものを使用した。

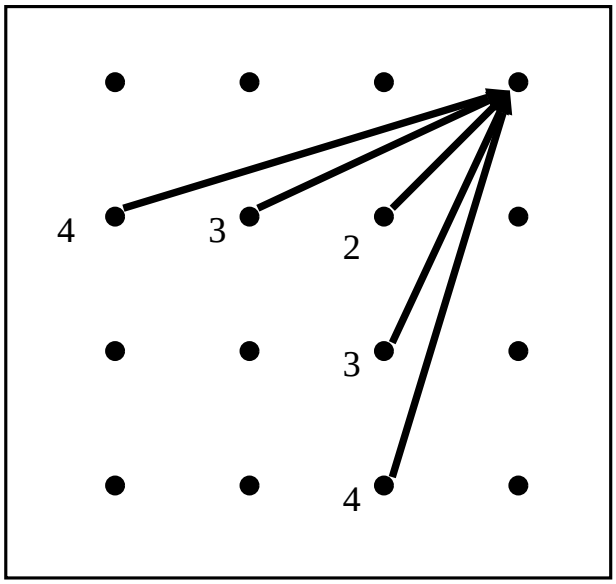


図 2.4: 局所パス規制

第 3 章

文音声の基本周波数の時間変化に現れる個人差の分析

3.1 目的

文音声について基本周波数パターンを抽出し藤崎モデルによる分析を行ない、抽出された各パラメータに現れる個人差について分析、検討する。

3.2 実験音声

分析に用いる音声には、男性の大学院生 8 名の発話による 5 文章各 10 サンプルの中から 5 話者を選び、話者毎に基本周波数パターンが似ている音声について 3 組ずつ選択した。録音条件を表 3.1 に示す。録音の際には、アクセント位置や後の時間軸の正規化を考慮し発話速度に関して極端にならないように指示を行なった。

表 3.1: 録音条件

マイクロフォン	SONY C-536P
DAT レコーダ	SONY TCD-DIO PRO2
サンプリング周波数	48 KHz

音声データに用いた文章は、基本周波数パターンを扱う研究であるため声帯振動を伴

表 3.2: LPC 分析 / 基本周波数抽出の仕様

サンプリング周波数	20 KHz
分析窓	Hanning 窓 （窓長 = 25.6msec）
フレーム周期	6.4 msec
LPC 次数	16 次

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録音を行なった 5 文章の中から「青い葵が青い屋根の上にある」を採用した。

3.3 分析方法

分析に使用する基本周波数パターンは、 LPC 分析の予測誤差信号の短時間自己相関関数のピークから求める。基本周波数抽出の仕様を表 3.2 に示す。

一般的に含まれる基本周波数の抽出誤りに関しては以下のように対処した。

- sonagram 等による正確な有声区間の割りだしによる適切な分析区間の決定
- 基本周波数抽出時の周波数範囲を変化させ抽出誤り位置を変化させることによる誤り部分の修正
- 上記の方法で修正できない部分の視察による基本周波数の修正

こうして得られた基本周波数パターンについて、藤崎モデルによる分析を行なう。藤崎モデルによる分析は、基本周波数パターンとモデルのパターン間で $I = 3,\ J = 5$ として analysis-by-Synthesis 法により平均自乗誤差を最小にすることにより行なった。分析に用いるパラメータの刻み幅は、T_0、T_1、T_2について 0.01[sec]、A_pと A_aについて 0.01 とした。また、F_{min}は基本周波数パターンとモデルの差の総和が 0 になるようにとり、α は 3.0[sec^{-1}]、β は 20.0[sec^{-1}]、θ は 0.9 で一定とした。

このように抽出されたパラメータのうち、話者間で個人差の大きいパラメータを調べるために F比による分析を行なう。

F 比は、カテゴリの数を n 、サンプル数を N、第 i カテゴリの第 j サンプルの特徴量を C_{ij} とすると、級間分散と級内分散の比

$$F = \frac{\sum_i^n \left(\bar{c}_i - \frac{1}{n}\sum_i^n \bar{c}_i\right)^2}{\frac{1}{N}\sum_i^n \sum_j^N (c_{ij} - \bar{c}_i)^2}, \quad \left(\bar{c}_i = \frac{1}{N}\sum_j^N c_{ij}\right) \tag{3.1}$$

で与えられるものであり、その値が大きいほどそのパラメータがカテゴリを分類する尺度として有用であることを示す。

3.4 分析結果

mkh,mmm,mnt,msh,myk の 5 話者について分析した結果を図 3.1から図 3.5に示す。

図 3.1は、基本周波数の変動の大きさと平均、基底周波数である。ここで、◯が抽出を行なったパラメータ値である基底周波数、線で結んであるのが各話者の平均基本周波数の値である。

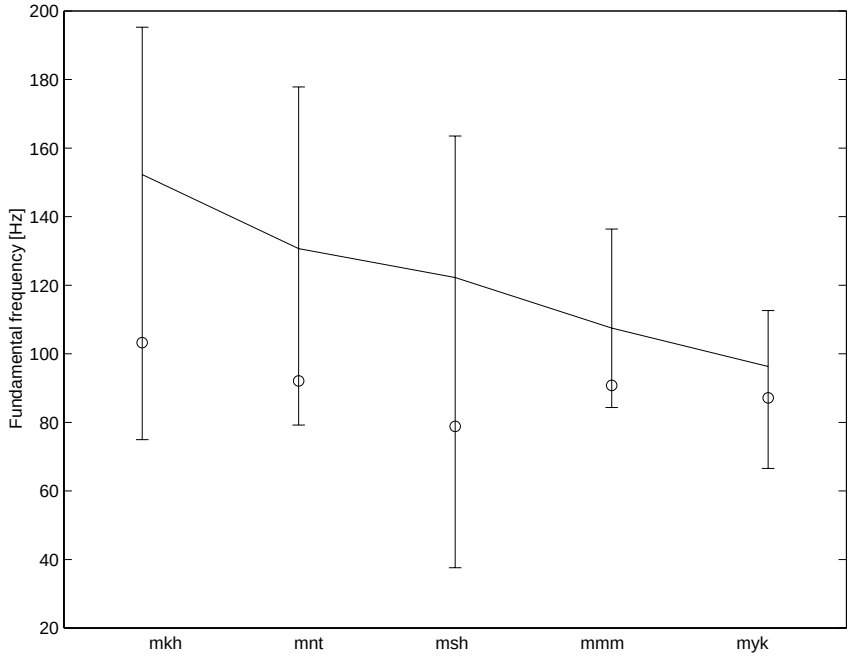


図 3.1: 各話者の基本周波数の変動と平均・基底周波数

図 3.2は各話者のフレーズ成分の値、図 3.3は各話者のアクセント成分の値である。ただし、話者はそれぞれ mkh:◯、mmm:×、mnt:＋、msh:□、myk:◇である。

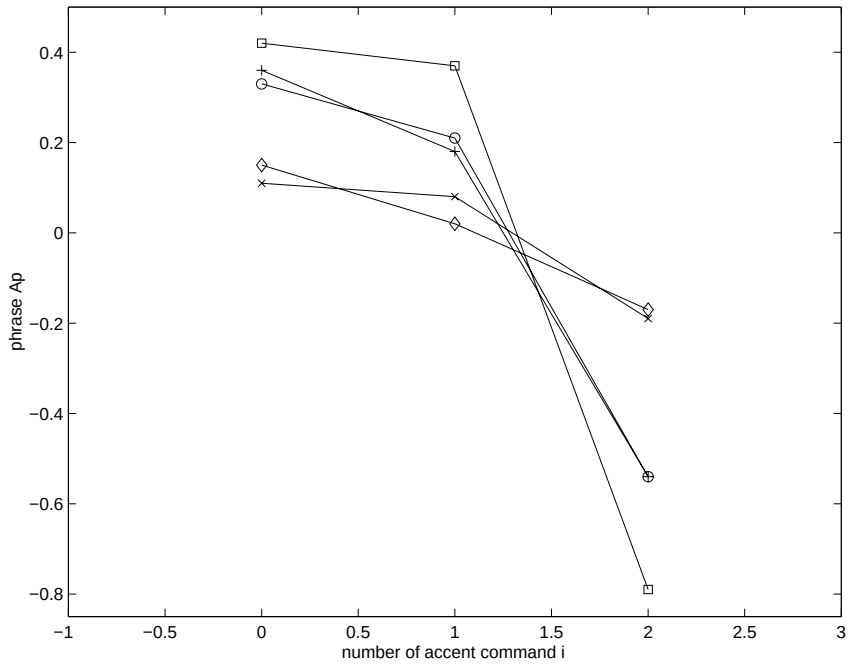


図 3.2: 各話者のフレーズ成分

図 3.4は、各話者の時間構造の分析結果である。ここで ΔT_{0i} 、 ΔT_{1j} 、 ΔT_{2j} はそれぞれフレーズコマンドの立上り時間 T_{00} からの相対時間である。

図 3.5は、F 比による分析結果である。

3.5 考察

図 3.1の分析結果から、基本周波数の変動の大きさは mkh と mmm、msh が大きく、mmm と myk が小さいことがわかる。また基底周波数は、mkh が他と比べて大きいことが分かる。

フレーズ成分は msh が他と比べて大きく、mkh と mnt、mmm と myk がそれぞれ近い値を持っていることがわかる。(図 3.2)

アクセント成分では、mnt が A_{a3}に、msh が A_{a2}に、myk が A_{a4}にそれぞれ特徴があり、mkh と mmm では mkh が全体的に大きいということがわかる。(図 3.3)

時間構造は msh が他と比べて極めて長く、文末のアクセント成分の継続時間を見ると、

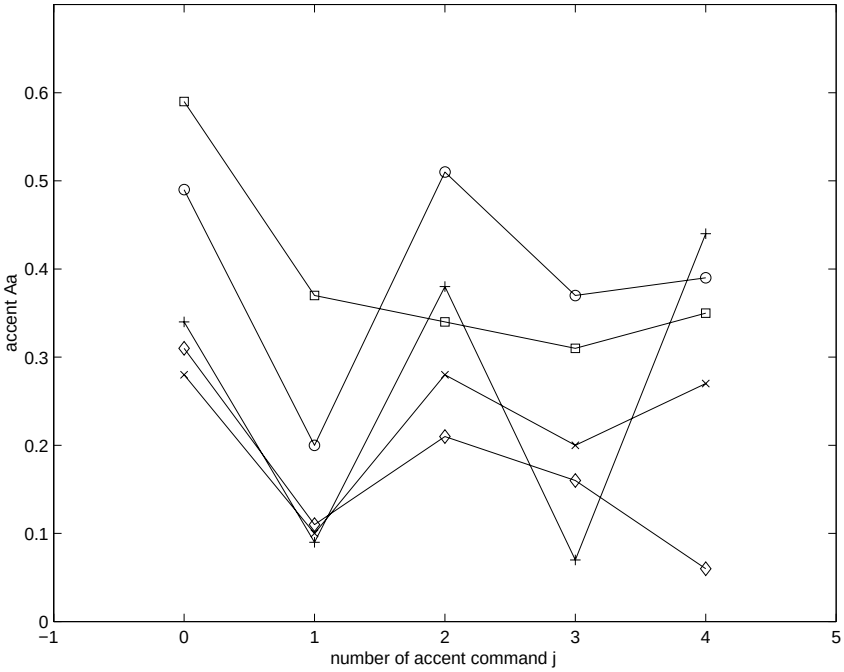


図 3.3: 各話者のアクセント成分

mmm と myk が短く、残りは長いことがわかる。(図 3.4)

F 比による分析の結果 (図 3.5) では「青い葵が青い屋根の上にある」という文章においては、F 比の値は文末の「上にある」にかかるアクセント成分 A_{a4}が 44.84 と最も大きく、続いて基底周波数 F_min の 32.38、立ち下がりのフレーズ成分 A_{p2}の 22.85 と続き、時間構造を示すΔTに関しては全て 10 以下の値となった。また、アクセント成分の立上りから立ち下がりまでのアクセントの継続時間、アクセント成分の立ち下がりから次のアクセント成分の立上りまでのアクセントの休止時間に関しても F 比による分析を行なったが、F 比の値はより小さくなった。以上の分析により、基底周波数とアクセント成分、フレーズ成分に個人差が大きく、時間構造にはあまり個人差がないという結果が得られた。

今回の分析に用いた話者と同一話者による別の文章 (「青いりんごの甘い匂いが匂う」) による F 比の分析結果は、図 7.2のような結果が得られた。この結果は図 3.5の結果とは完全には一致しない。このことより、文章によって個人差の現れやすい部分というのは異なるのだと思われる。文章中で実際に個人差の現れやすい部分がどこかを限定するには、より多くの文章についてより詳しい分析検討が必要である。

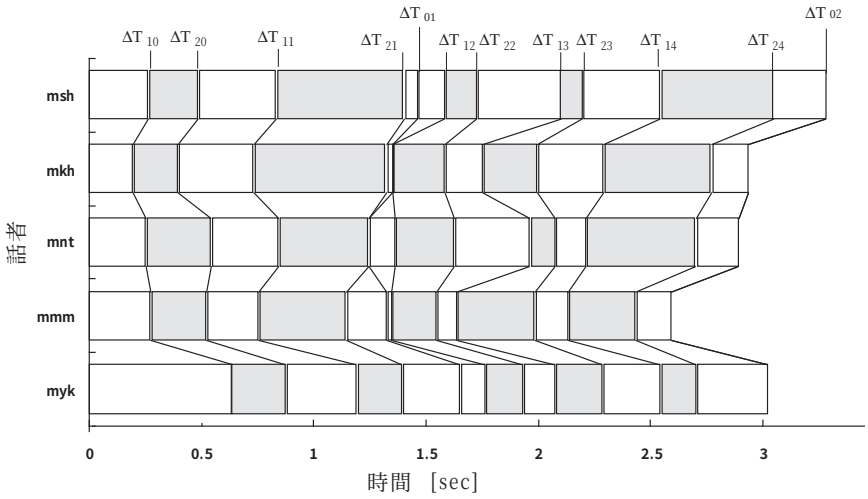


図 3.4: 各話者の時間構造

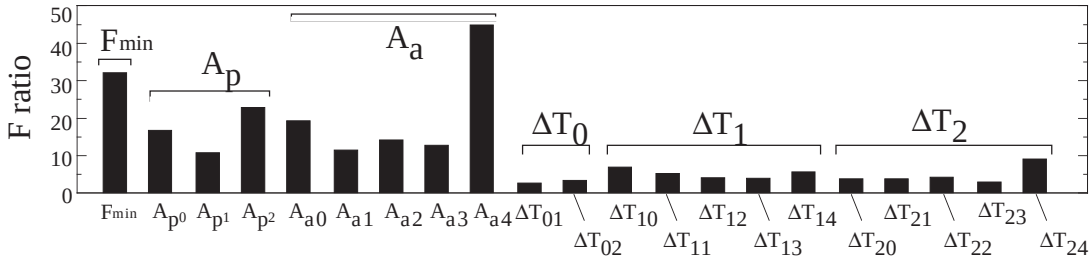


図 3.5: F 比による分析

以上のように、ここでは各パラメータの個人差について分析した。このことをふまえて、以下の章では聴取実験により基本周波数パターンの個人性情報について調べ、実際に個人性情報を多く含んでいるパラメータを明らかにする。

第 4 章

合成音声の個人性情報に関する聴覚実験

4.1 目的

ここでは、スペクトル包絡変換合成音声を用いた聴取実験により、被験者が基本周波数パターンの違いをどの程度聞き分けられるか調べる。また、基本周波数パターンに藤崎モデルにより近似したパターンを用いた場合の個人性知覚への影響を調べる。

4.2 実験条件

4.2.1 実験音声

実験音声としては、次の 3 種類の音声を用いる。

1. 原音声
2. スペクトル包絡変換音声 (基本周波数パターンは TEMPO により抽出したもの)
3. スペクトル包絡変換音声 (基本周波数パターンは藤崎モデルにより記述したもの)

実験に使用する原音声には、前節で分析を行なった 5 話者による各 3 サンプルの発話文音声 (「青い葵が青い屋根の上にある」) を採用した。

スペクトル包絡変換音声は、前述の STRAIGHT 音声分析合成系により作成した。その際用いるスペクトル包絡の情報は、録音してもらった 8 名の話者のうち刺激音に基本周波数パターンを用いた 5 名を除いた 3 話者より選択した。

表 4.1: 実験条件

話者	5 名
被験者	5 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 76 dB (A)

4.2.2 実験方法

聴取実験は、聴き直しを許さない環境で対比較により行なった。被験者には a と b の 2 つの音声を聴いてもらい、a の音声の話者と b の音声の話者が同じであるかどうかを-2(全く異なる) から 2(全く等しい) までの 5 段階評価で評価してもらった。この時、聞き取れなかった場合やわからなかった場合には 0 と評価してもらった。

前述の 3 つの音声について、聴取実験を行なった。被験者は実験に使用した音声の話者をよく知っている同じ講座内の大学院生 (男性) 5 名である。以降、本論文で行なわれた聴取実験は全て同じ 5 名の被験者により行なった。

4.3 実験結果

実験の結果を表 4.2、表 4.3に示す。これは呈示刺激音声 AB の組合せ

(1) A と B が全く同じ音声の場合

(2) A と B は全く同じではないが同じ話者による音声の場合

(3) A と B が同じ話者による音声の場合

(4) A と B が異なる話者による音声の場合

の 4 種類において、刺激音声に原音声、スペクトル包絡変換音声 (TEMPO)、スペクトル包絡変換音声 (藤崎モデル) を用いた場合の、被験者の答えた 5 段階評価の平均 (表 4.2) とその標準偏差 (表 4.3) である。

表 4.2: 実験結果 (平均)

音声サンプル	(1)	(2)	(3)	(4)
原音声	2.0	2.0	2.0	-2.0
TEMPO	1.973	1.707	1.796	-1.761
藤崎モデル	1.88	1.74	1.787	-1.751

表 4.3: 実験結果 (標準偏差)

音声サンプル	(1)	(2)	(3)	(4)
原音声	0	0	0	0
TEMPO	1.61	0.90	0.752	0.79
藤崎モデル	0.66	0.97	0.88	0.81

表 4.4 は実験に使用した 3 種類の合成音声間に対して、表 4.5 は上述の 4 つの呈示組合せ間に対して、それぞれ表 4.2、表 4.3の結果を用いて t 検定を行なった結果である。

図 4.1は、上述の 4 つの呈示組合せ間のうち同じ話者の音声 (3) と異なる話者の音声 (4) に対して、それぞれ表 4.2、表 4.3の結果を用いて正規分布を示したものである。ここで青色が (3)、赤色が (4) の分布を示す。またインパルスが原音声、点線が TEMPO で抽出した基本周波数パターンによる合成音声、実線が藤崎モデルで近似した基本周波数パターンによる合成音声である。

4.4 考察

表 4.4の結果より、スペクトル包絡を変換して基本周波数の情報に TEMPO で抽出した基本周波数パターンを用いた音声や藤崎モデルにより記述した基本周波数パターンを用いた音声を原音声と比較したとき、全く同じ音声を聞いた場合には有意水準 5 %で同じものであるといえたが、同じ話者の発話による音声や違う話者の音声を聞いた場合には原音声とスペクトル包絡変換音声は同じ母平均をもっているとはいえなかった。原音声の場合には完全に話者を判断できたことを考えると、スペクトル包絡を交換することによって個

表 4.4: 実験結果の t 統計量 (合成音声間)

音声サンプル	(1)	(2)	(3)	(4)
原音声,TEMPO	1.424	3.985	4.079	9.111
原音声, 藤崎モデル	1.585	3.299	3.654	9.199
TEMPO, 藤崎モデル	1.187	0.309	0.115	0.265

$t_{0.05} = 1.960, t_{0.01} = 2.576$

表 4.5: 実験結果の t 統計量 (呈示刺激音声 AB の組合せ間)

音声	(1) と (4)	(2) と (4)	(3) と (4)	(1) と (2)
TEMPO	41.024	48.932	61.221	2.534
藤崎モデル	37.722	47.394	57.52	1.131

$t_{0.05} = 1.960, t_{0.01} = 2.576$

人性情報の一部が失われ、知覚が影響を受けたと考えられる。

また、同一のスペクトル包絡を用いた TEMPO と藤崎モデルの合成音声間では同一とみなすことができるという結果が得られた。これより、基本周波数パターンに藤崎モデルで記述したパターンを用いても、個人性情報は殆ど失われないことが確かめられた。しかし同じ音声を聞いた場合には図 4.1 を見てもわかるように、藤崎モデルにより近似したパターンと TEMPO により抽出されたパターンとの間には何らかの違いがみてとれた。

また t 検定による 4 つの呈示組合せ間での分析では、表 4.5からわかるように、TEMPO によるスペクトル包絡変換音声、藤崎モデルによるスペクトル包絡変換音声は、同じ音声の場合と同じ話者の場合ではそれらが同じ母平均をもつということが有意水準 1 %で採択された。また、違う話者と同じ話者とを比較した場合に、それらが互いに全く異なるものであるという結果となった。

以上の結果より、被験者はスペクトル包絡が同じで基本周波数パターンだけが異なる音声を聞いて、基本周波数パターンから同じ音声のみでなく同じ話者の音声も充分聞き分けられることが確かめられた。これは、本論文での個人性の定義 (同じ話者の発話した音声について同じ話者による音声であることが判断できること) から、基本周波数パターンは十分に個人性情報を含んでいると考えることができる。また、藤崎モデルにより記述され

た基本周波数パターンは、TEMPO により抽出された基本周波数パターンとほぼ同程度の個人性情報を含んでおり、聴取実験で基本周波数パターンに藤崎モデルを用いても問題がないと考える。

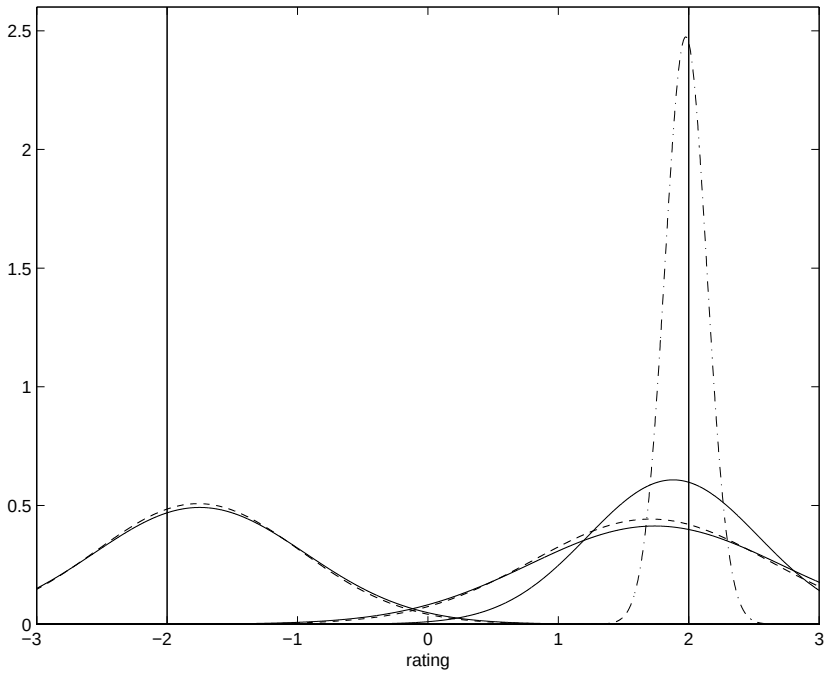


図 4.1: (3) と (4) の正規分布

第 5 章

基本周波数の時間変化の各パラメータの個人性知覚への影響に関する聴覚実験

5.1 目的

ここでは、基本周波数変形合成音声を用いた聴取実験により、被験者が知覚している基本周波数パターンの違いの中でどのパラメータがより多く影響を与えているかを明らかにする。

5.2 実験条件

5.2.1 実験音声

実験では、前節の藤崎モデルによるスペクトル包絡変換音声で話者 a の藤崎モデルのパラメータの一部を話者 b のものに変換した基本周波数変形合成音声を使用する。変換する藤崎モデルのパラメータとしては、以下のものを考える。

1. 基底周波数 F_{min}
2. フレーズ成分 A_{pi}
3. アクセント成分 A_{aj}
4. 時間構造 T_{0i}, T_{1j}, T_{2j}

表 5.1: パラメータの組み合わせ

組み合わせ	A	B	C	D	E	F	G	H
基底周波数	a	b	a	a	a	b	b	a
フレーズ	a	a	b	a	a	b	a	b
アクセント	a	a	a	b	a	a	b	b
時間構造	a	a	a	a	b	a	a	a

話者 a・b による藤崎モデルのパラメータの変換パターンには、表 5.1 のような 8 通りを考えることとする。

パラメータを変換する基本周波数パターンには、前節の実験で用いた各話者 3 サンプルの発話の中から 1 つずつをランダムに選択した。またスペクトル包絡には、前節のスペクトル包絡変換音声と同じものを使用した。

5.2.2 実験方法

実験は、聞き直しを許す環境で ABX 法による聴取実験を行なった。基本周波数に藤崎モデルにより近似した基本周波数パターンを用いたスペクトル包絡変換音声 a・b と藤崎モデルのパラメータの一部を a から b に入れ換えたパラメータ変換音声 x を聴いてもらい、x の音声が a と b どちらの話者の発話による音声と感じるかを強制判断してもらった。聴取回数は各組合せにつき 2 回行ない、順序効果を排除するため a・b の順序を半分ずつ入れ替えた。被験者は前節同様、男性 5 人である。

5.3 実験結果

実験結果を図 5.1示す。これは、話者 a の基本周波数パターンの一部を話者 b のものに変換した基本周波数変形音声 x を聴いて b の話者の音声に近いと答えた割合を、被験者毎に全ての話者の組合せの知覚率を平均した結果である。

この聴取実験の結果について視覚的に分りやすく表にしたものを、表 5.2 に示す。

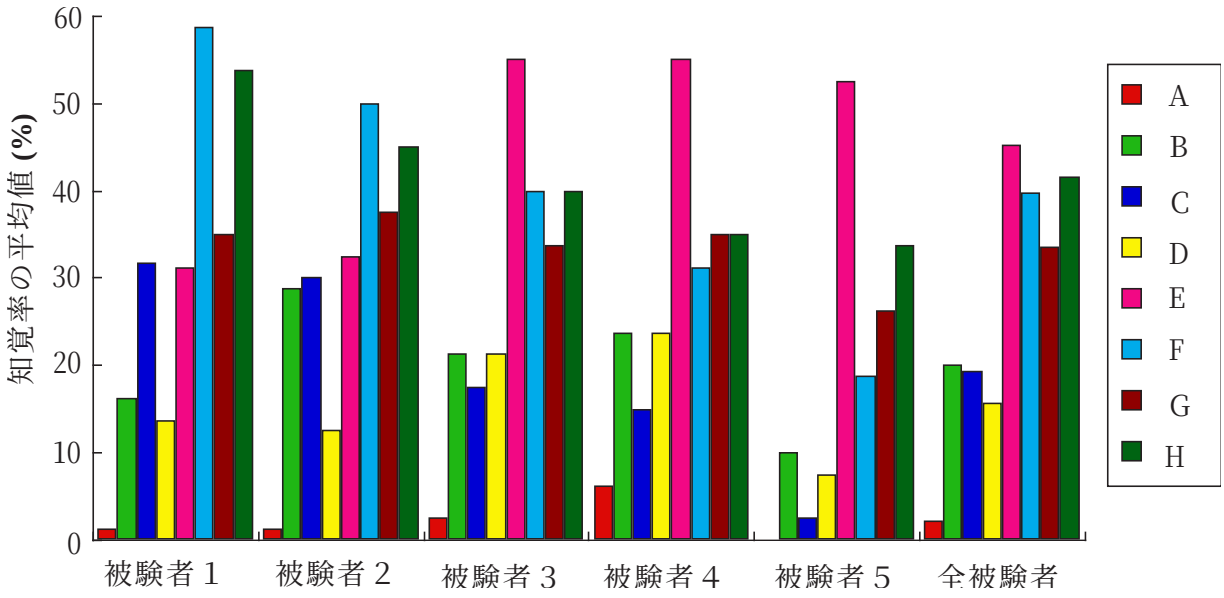


図 5.1: 各被験者のパラメータの組み合わせ毎の知覚率

ただし、×, △, ○, ◎はそれぞれ知覚率が 5 %未満,5〜20 %,20〜40 %,40 %以上を表す。

また、基底周波数だけを入れ替えた場合 (B) の被験者平均の各話者の組合せの知覚率についてそれぞれ図 5.2に示す。

図 5.3は話者間での基底周波数の差の平均との比を求めたもので、値が大きい物程その話者間では基底周波数に差があることを表している。

同様にフレーズ成分だけを入れ替えた場合 (C)、アクセント成分だけを入れ替えた場合 (D)、時間構造だけを入れ替えた場合 (E) の被験者平均の各話者の組合せの知覚率をそれぞれ図 5.4から 5.6に示す。

表 5.2: 実験結果 (パラメータの組み合わせ毎の知覚率)

組み合わせ	A	B	C	D	E	F	G	H
被験者 1	×	△	○	△	○	◎	○	◎
被験者 2	×	○	○	△	○	◎	○	◎
被験者 3	×	○	△	△	◎	◎	○	◎
被験者 4	△	○	△	△	◎	○	○	○
被験者 5	×	△	×	△	◎	△	○	○
合計	×	○	△	△	◎	○	○	◎

5.4 考察

実験の結果、なにもパラメータを変更しない場合 (A) には、被験者は殆ど正しい話者 (a) を選択することができた。間違いが 0 %にならなかった原因としては、前節同様スペクトル包絡を変換したことによる影響が考えられる。

基底周波数を変更した場合 (B) には、被験者 5 名全員が影響を受け、うち 3 名 (被験者 2,3,4) は特に強く影響を受けた。話者の組合せとしては、

- アクセント成分と時間構造にあまり差がない。
- 基底周波数の差が、フレーズ成分の差に比べて著しく大きい。

という場合に強く影響を受け、以上の条件のうちどれかが欠けた場合にはほとんど影響を受けなかった。

フレーズ成分の全てを入れ替えた場合 (C) には被験者 5 名中 2 名 (被験者 1,2) は強く影響を受けたが、1 名 (被験者 5) は全く影響を受けなかった。この場合の影響を受けた話者の組合せとしては、

- アクセント成分と時間構造にあまり差がない。
- フレーズ成分の差が、基底周波数の差に比べて著しく大きい。

という場合であり、条件のうちどれかが欠けた場合にはほとんど影響を受けなかった。

アクセント成分の全てを入れ替えた場合 (D) には被験者全員がある程度影響を受けた。この場合の影響を受けた話者の組合せとしては、

- アクセント成分 A_{a0}〜A_{a4}のうち大きく異なるものが存在する。
- 平均基本周波数や時間構造に極端な差がない。

という場合で、条件のうちどれかが欠けた場合にはほとんど影響を受けなかった。

時間構造の場合 (E) には被験者全員が強く影響を受け、うち 2 名 (被験者 4,5) は A から H までの組合せの中で最も影響を受けた。話者の中に 1 名他話者と比べ発話長が長い話者 (msh) がおり、被験者全員がその話者に対し強く影響を受けた。

実験の結果、パラメータを話者 a から話者 b のものへ単独で変更した場合には、個人性知覚への影響は時間構造が最も大きく、 次に基底周波数 > フレーズ成分 > アクセント成分 の順に大きかった。また、ある話者の組合せで被験者全員の評価が強く影響を受けたものは時間構造と基底周波数の場合のみであり、フレーズ成分に関しては被験者によるバラツキが大きく、なかには全く影響を受けない被験者もいた。また、アクセント成分の場合には影響自体が小さく、影響を受ける話者の組合せでも b であると知覚した割合は 50 %程度であった。

このことから、話者を知覚する上で時間構造と基底周波数が大きなウエイトを占めており、フレーズ成分やアクセント成分は前者に比べウエイトが小さいのではないかと思われる。

次に 2 つのパラメータを組み合わせた場合について考察する。

基底周波数とフレーズ成分を変換した場合 (F) には、4 名が強く影響を受けた。影響の少なかった 1 名の被験者は先にフレーズ成分に影響を受けなかった被験者である。

次に基底周波数とアクセントを変換した場合 (G) には、被験者全員が影響を受けた。

フレーズ成分とアクセント成分を変換した場合 (H) には、被験者全員が影響を受け、時間構造 (E) の影響を大きく受けた 2 名 (被験者 4,5) は比較的弱く、フレーズ成分 (C) の影響を大きく受けた 2 名 (被験者 1,2) は強く影響を受けた。

過去に 3 モーラ単語の基本周波数パターンにおいては、次のような報告がなされている [4]。

- 単語音声では藤崎モデルのパラメータのうち、時間構造に個人性情報は少なく、基底周波数とフレーズ成分、アクセント成分の 3 つのパラメータに個人性が多く含まれており、この 3 つを制御する事により話者変換が可能である。

- 基底周波数にも個人性は含まれるが、それのみでは基本周波数パターンの個人性においては支配的であるとはいえず、基本周波数の動的変化 (変化の大きさ) に多くの個人性が含まれている。

以上のことを踏まえて実験結果をみると、被験者は大きく分けると基底周波数やフレーズ成分に強く影響を受けるグループ (被験者 1,2) と時間構造に強く受けるグループ (被験者 4,5) に分かれ、被験者 3 はその中間に位置しているといえる。すなわち、被験者は 2 名 (被験者 1,2) が基本周波数の高さを、2 名 (被験者 4,5) がアクセント成分の立上り立ち下がりのタイミングを含めた基本周波数の動きを重視しており、残りの 1 名 (被験者 3) は高さとタイミングを総合的に聴いているものと思われる。これは、単語音声では基本周波数の高さが被験者の知覚に影響を及ぼしていたのに対し、文音声では基本周波数の高さも重要であるものの、時間構造の重要性が増し基本周波数の動きが知覚に強く影響しているものと考えられる。

また、話者 a から b にあるパラメータを入れ替えた場合と b から a に入れ替えた場合では、必ずしも同じ認識率とはならなかった。これは入れ替えるパラメータの差の大きさ自体は等しいが、話者を識別する際に話者がそのパラメータの他に個人性情報を多く含むパラメータを持っているかどうかによって、その判断が変わってくることが原因であると思われる。そのため、パラメータの組合せの中に他話者と比べて極めて大きな差を持ったパラメータが存在する時、そのパラメータを含んだ話者に知覚は強く影響される。この結果は、音響特徴の差が個人性の知覚に反映されるという報告とも一致している [3]。

B・C・D の結果から、4 つのパラメータのうち時間構造とどれか 2 つのパラメータを変換することにより、残りの 1 つのパラメータが話者を決定付けるほど知覚を左右する大きな個人性情報を含んでいるのでなければ、80 %以上被験者の知覚を変化させることが出来ることがわかった。このことから、極端に大きな差を持ったパラメータが存在しない場合には時間構造を含めた 3 つのパラメータを制御する事により話者変換は可能であると考えられる。また、極端に大きな差を持ったパラメータが存在する場合には、そのパラメータと時間構造を含めた 3 つのパラメータを制御する事によりより強く話者の判断を変化させる事ができると考えられる。これは単語の場合と比べ、文音声においては時間構造が非常に重要である事を示している。

基本周波数パターンの時間構造の F 比による解析では大きな個人差が見られなかったが、実際の個人性知覚において時間構造には個人性が多く含まれるという結果が得られ、

アクセント成分とフレーズ成分は時間構造や基底周波数と比べて知覚に与える影響が小さく個人性情報が少ないと考えられる。

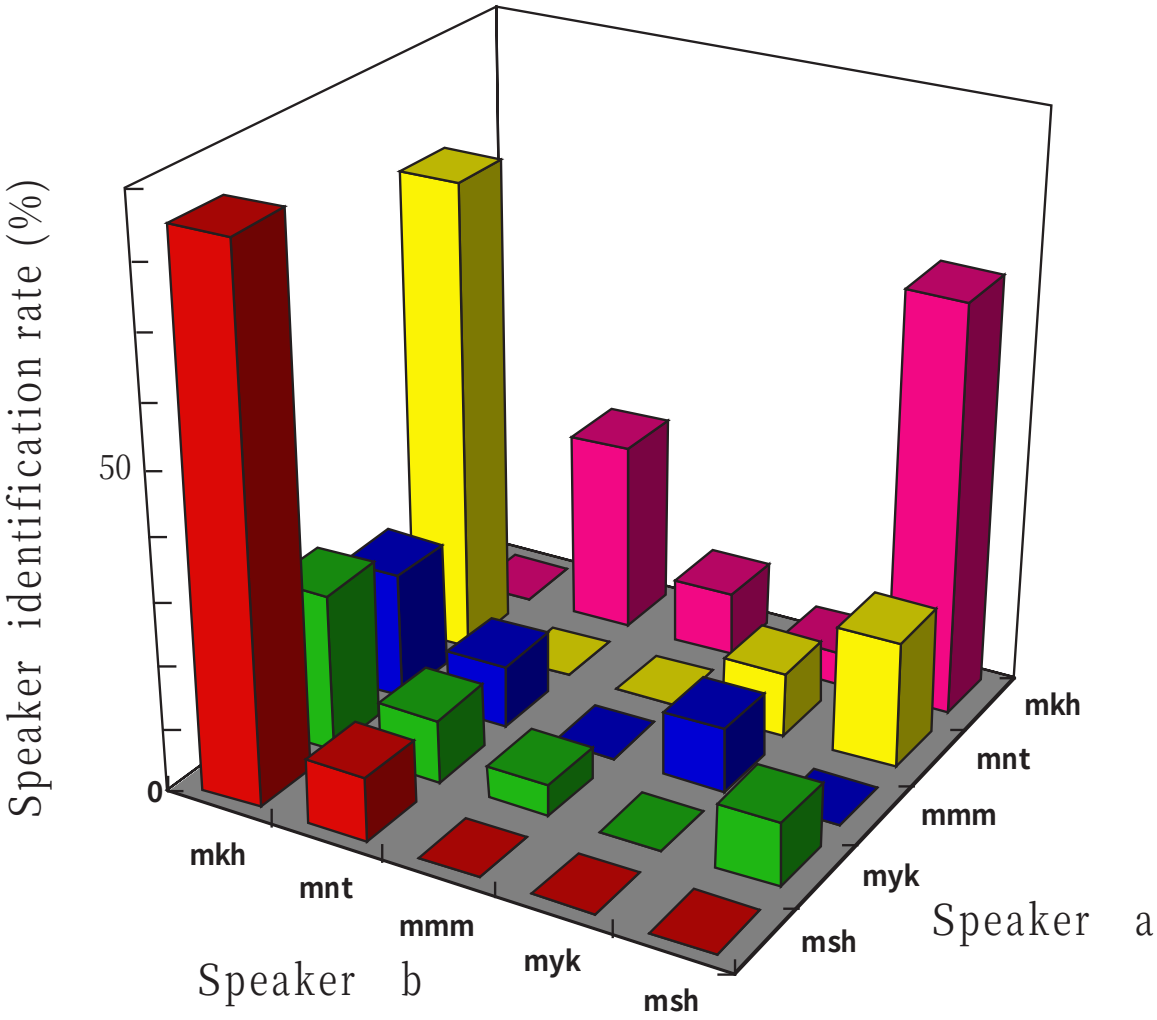


図 5.2: 基底周波数変換時の話者知覚率 (被験者平均)

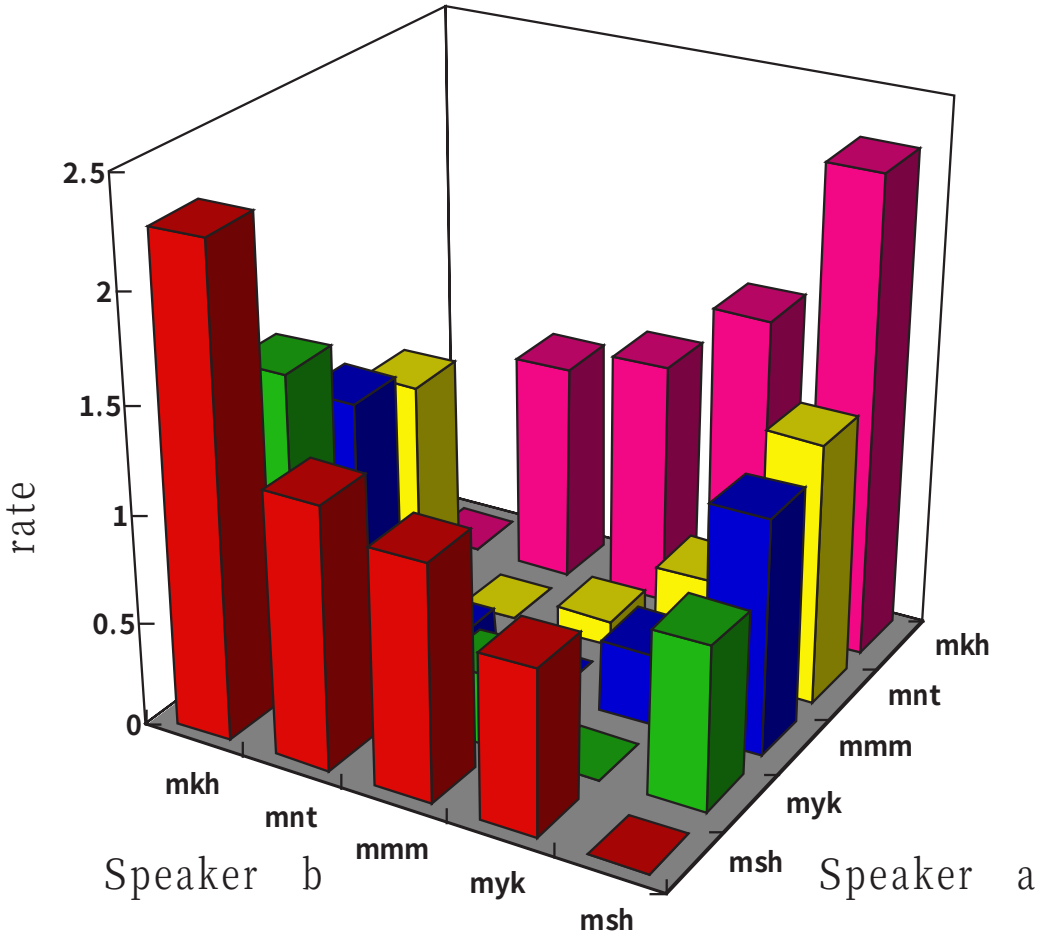


図 5.3: 話者間での基底周波数の差

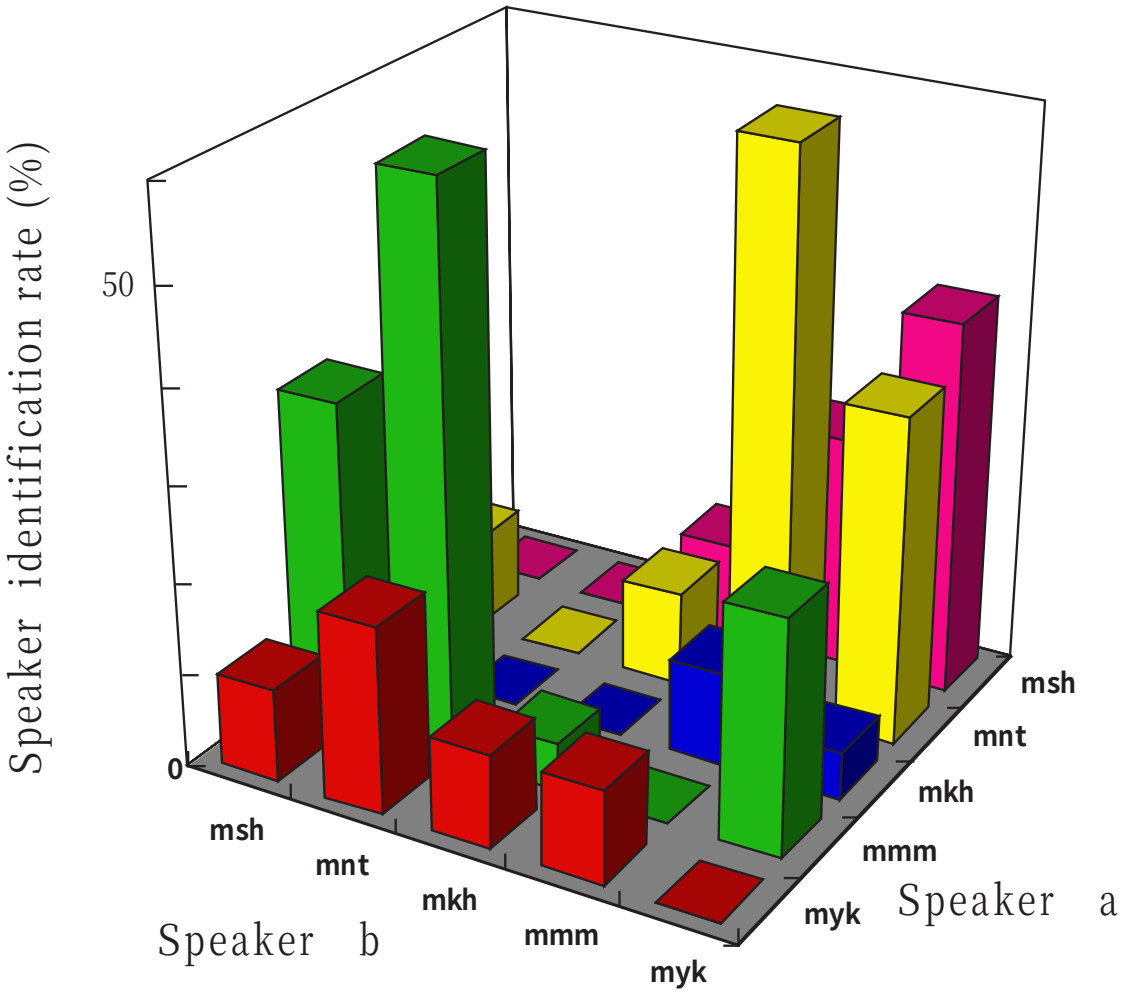


図 5.4: フレーズ成分変換時の話者知覚率 (被験者平均)

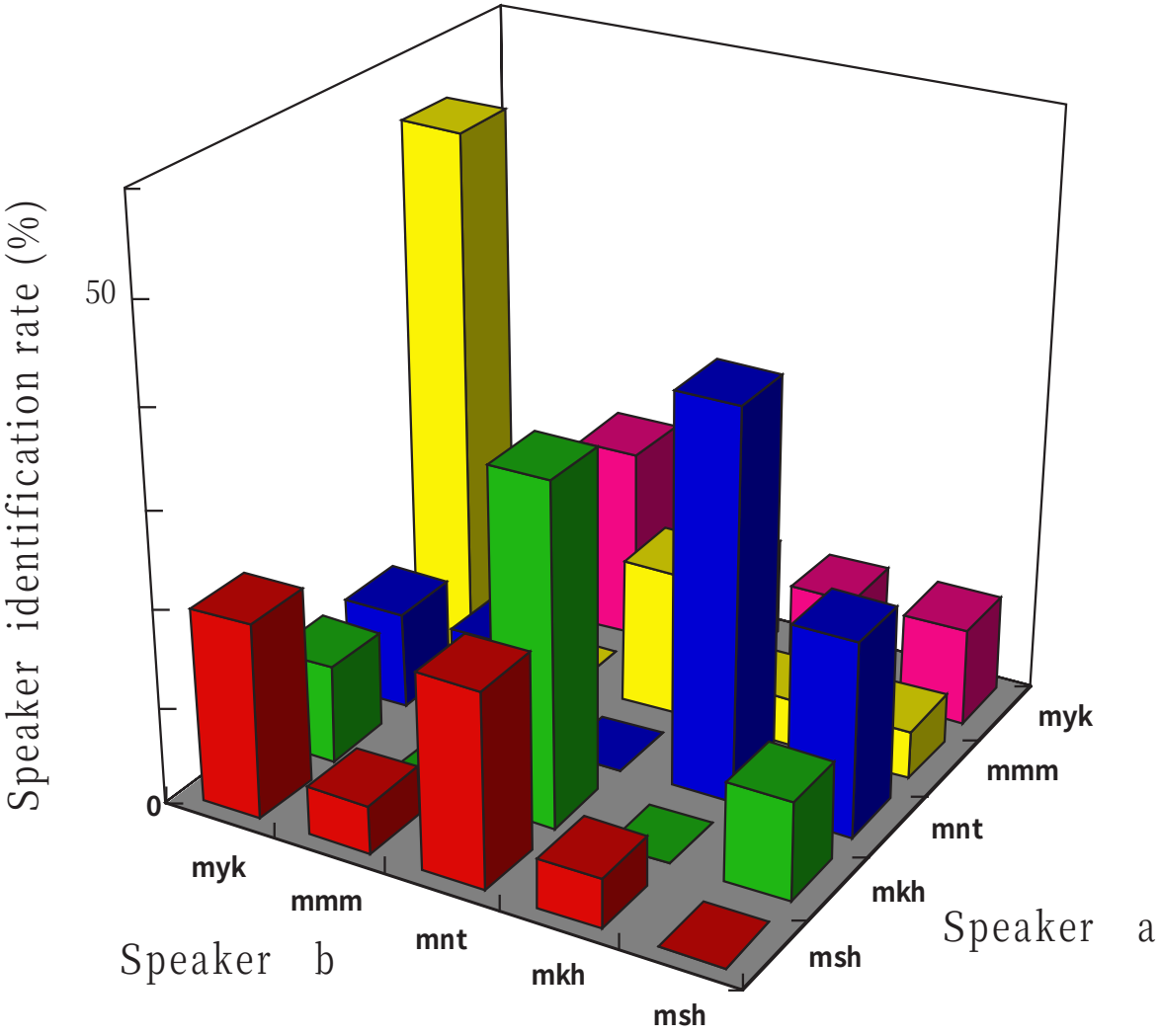


図 5.5: アクセント成分変換時の話者知覚率 (被験者平均)

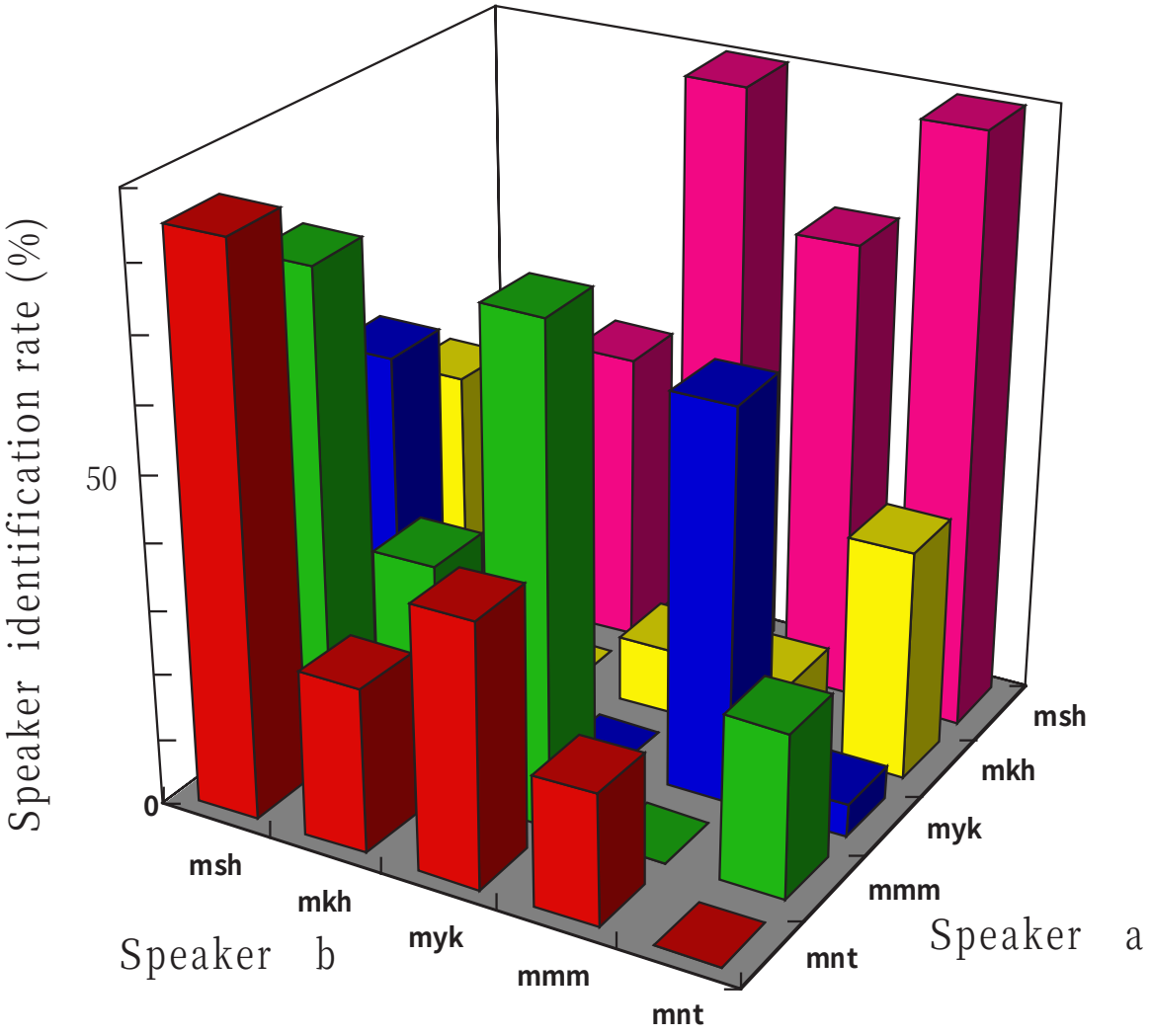


図 5.6: 時間構造変換時の話者知覚率 (被験者平均)

第 6 章

結言

本論文では、文音声の基本周波数パターンに含まれる個人性情報を調べるため、藤崎モデルを用いてパラメータの抽出を行ない、そこに現れる個人差について分析を行なった。また、文音声において基本周波数パターンが多くの個人性情報を含むことを聴取実験により確かめた。そして、実際に基本周波数パターンのパラメータを変更することにより、個人性知覚に及ぼす影響について調べ、検討を行なった。その結果、次のようなことが確認できた。

- 文音声においてスペクトル包絡を変換した音声でも話者が知覚でき、基本周波数パターンに個人性情報が多く含まれていることが確認できた。
- 単語の場合はあまり個人性情報を含んでいなかった時間構造が、文音声においては非常に多くの個人性情報を含んでおり、話者を知覚する上で重要な要素となっている。
- 被験者により各パラメータが個人性知覚に影響を与える大きさが異なり、被験者は主に基本周波数パターンの高さを重視するグループと基本周波数パターンのタイミングを重視するグループの 2 組に分けられる。
- 時間構造を含めた 3 つのパラメータを変換することにより、話者の知覚を変化させることが出来る。

今回、対象にした音声は「青い葵が青い屋根の上にある」という文音声のみで被験者も 5 名であった。今後はより多くの音声サンプル、より多くの被験者による同様の検討が必要である。

第 7 章

付録

7.1 藤崎モデルによるパラメータ抽出例

「青い葵が青い屋根の上にある」

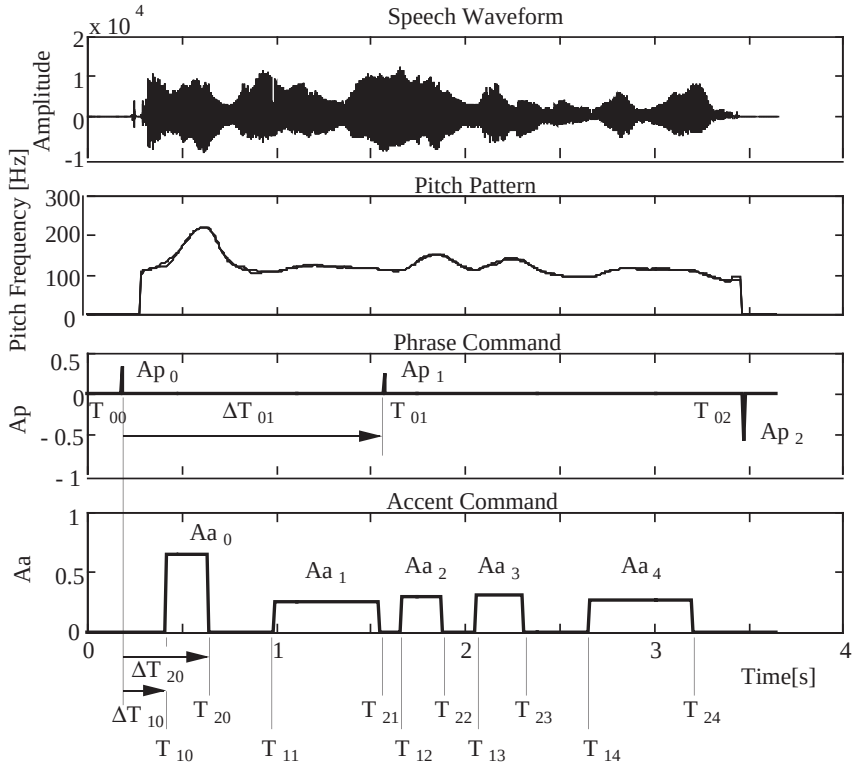


図 7.1: パラメータ抽出例 (msh03.ad)

7.2　F 比による分析結果

「青いりんごの甘い匂いが匂う」という文章について、「青い葵が青い屋根の上にある」で用いた 5 話者のパラメータ抽出 (I=3,J=5) を行ない、F 比による分析を行なった結果を図 7.2に示す。

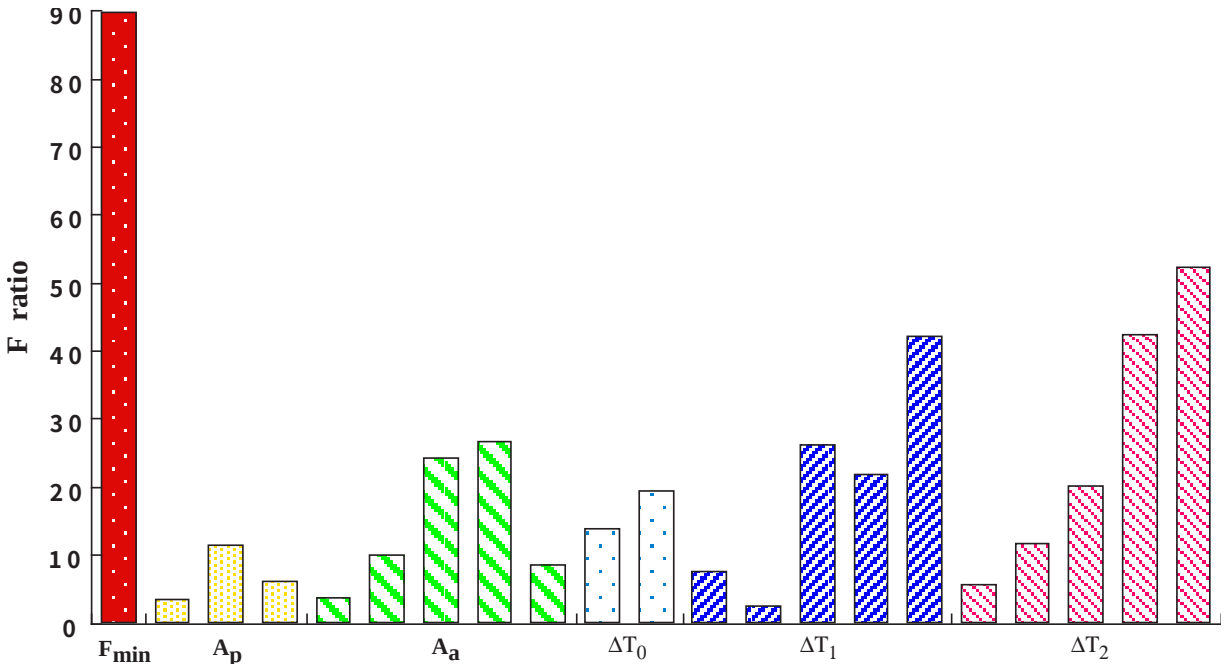
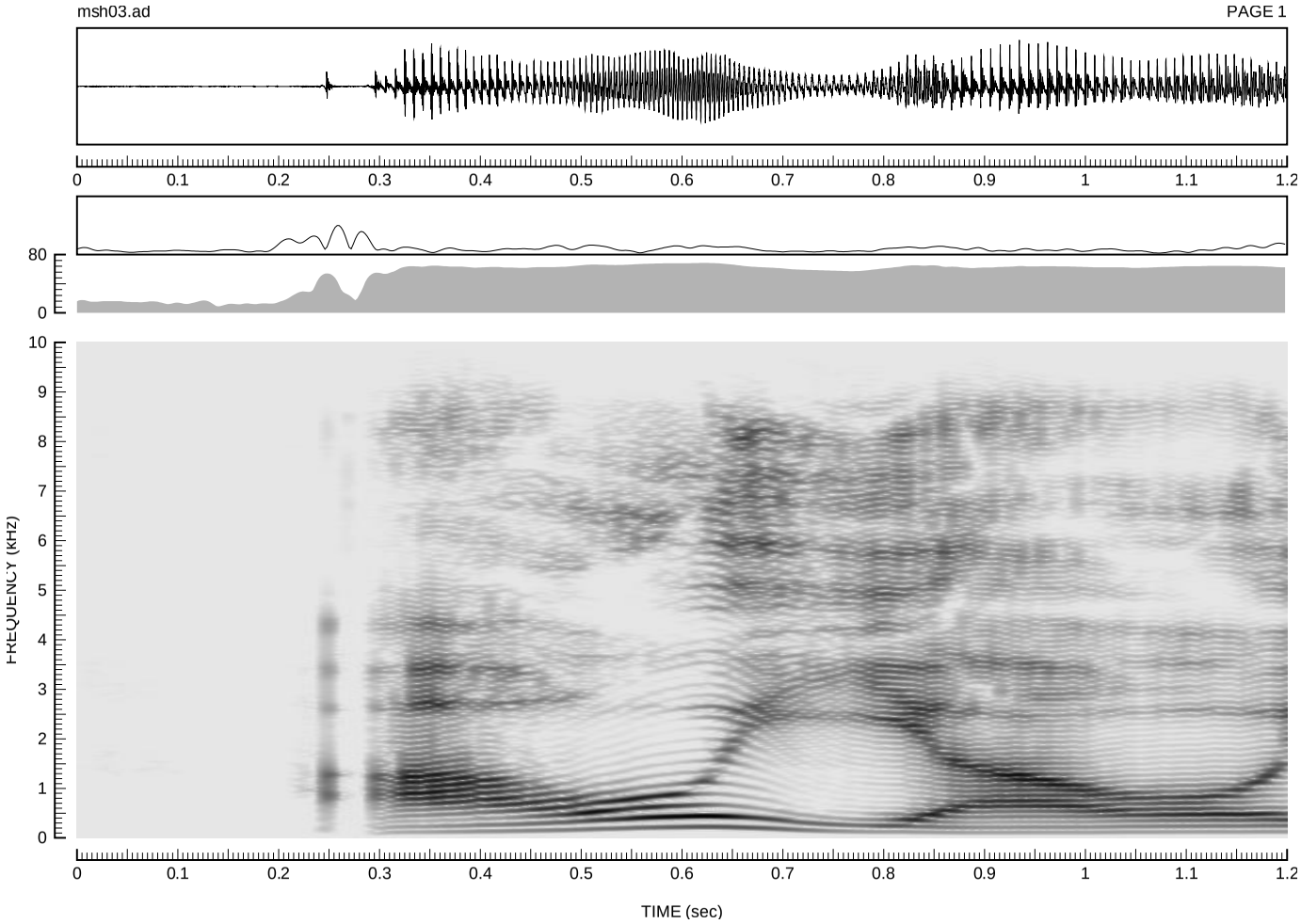
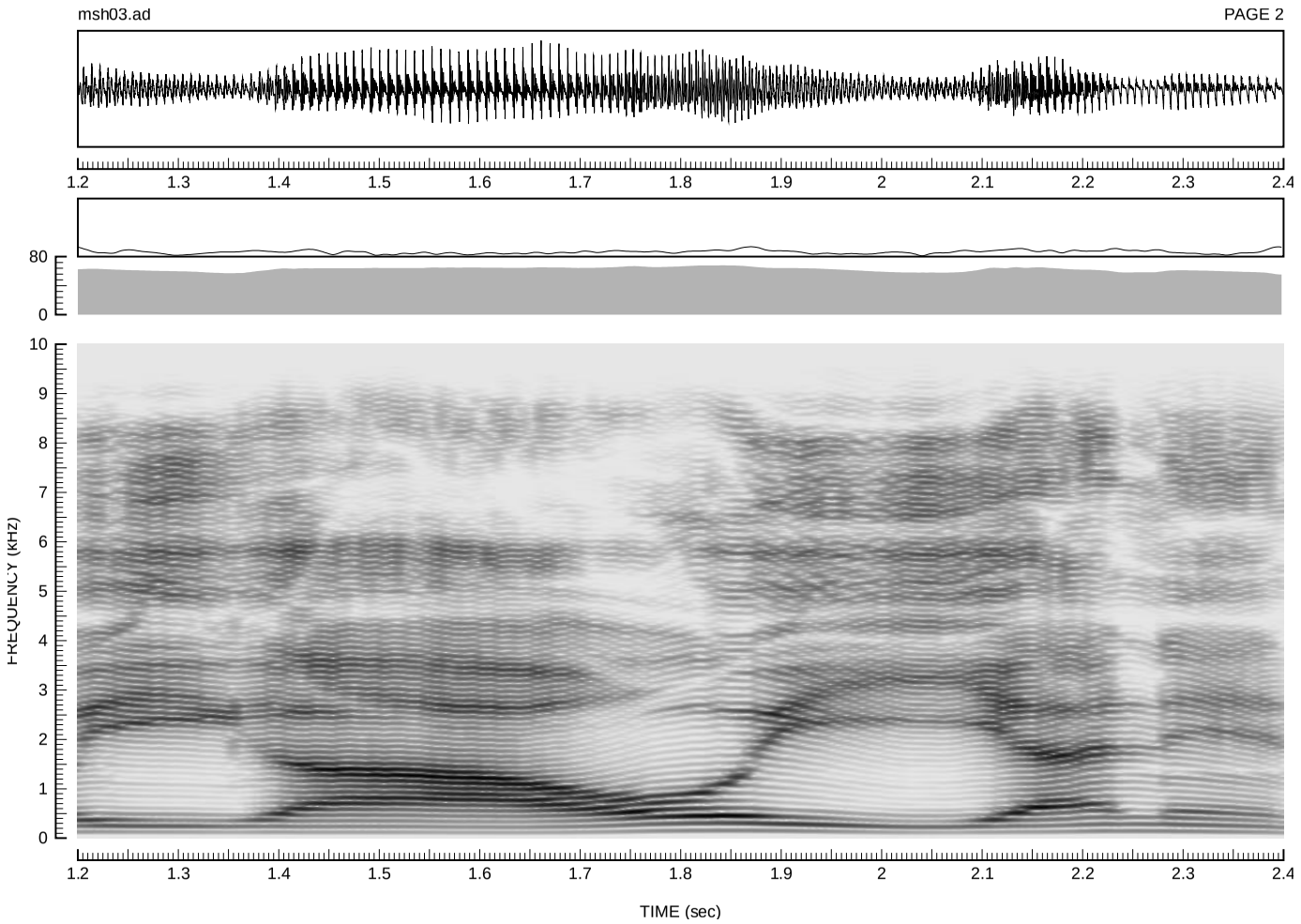
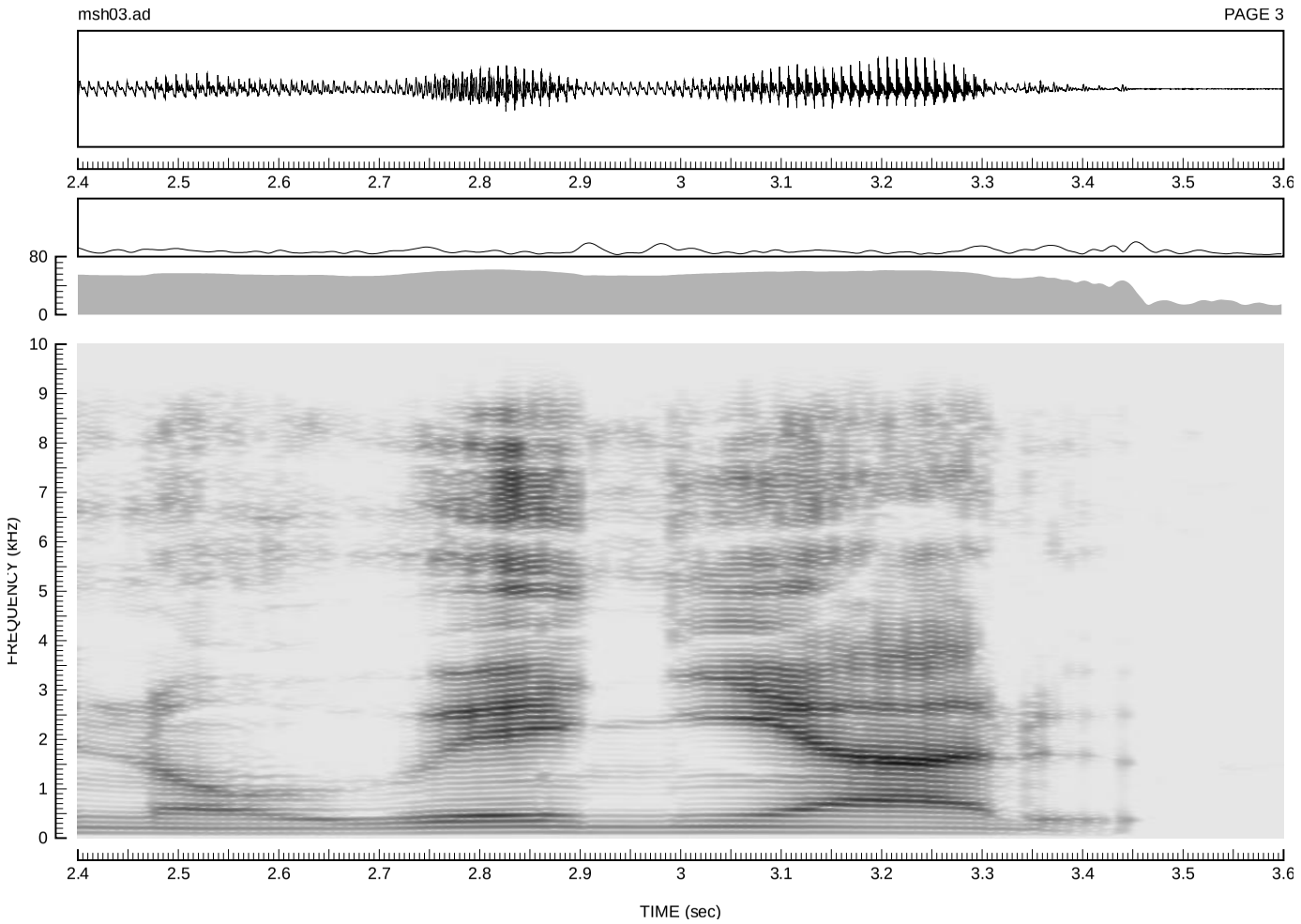


図 7.2: F 比による分析結果 2

7.3　ソナグラフ出力例


msh03.ad
PAGE 1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
80
0
10
9
8
7
6
5
4
3
2
1
0
FREQUENCY (kHz)
TIME (sec)


msh03.ad
PAGE 2
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2
2.1
2.2
2.3
2.4
80
0
10
9
8
7
6
5
4
3
2
1
0
FREQUENCY (kHz)
TIME (sec)


msh03.ad
PAGE 3
2.4
2.5
2.6
2.7
2.8
2.9
3
3.1
3.2
3.3
3.4
3.5
3.6
80
0
10
9
8
7
6
5
4
3
2
1
0
FREQUENCY (kHz)
TIME (sec)

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の皆様に感謝いたします。また、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下さった全ての皆様に厚く感謝いたします。

参考文献

[1] 桑原, 大串：“ホルマント周波数・バンド幅の独立制御と個人性判断”, 電学論，**J69-A**, 4，pp. 509-517 （1986）

[2] 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）

[3] 橋本, 樋口：“ Analysis of acoustic features affecting speaker identification”, Eurospeech'95, pp.435-438 (1995)

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

[5] 河原：“聴覚の情景分析と高品質音声分析変換合成法 STRAIGHT”，音響学会論文集, pp.189-192 （1997）

[6] 藤崎, 広瀬：“規則による音声合成”, 日本音響学会誌，**37**, 5, pp.204-209 （1981）

[7] H. Fujisaki and K. Hirose: “Analysis of voice fundamental frequency contours for declarative sentences of Japanese”, J. Acoust. Soc. Jpn. (E) **5**, 4 (1984)

[8] 濱上, 古村：“拡張点ピッチモデルによる韻率制御”, 音響論集 2-5-1,（1994 年 10 月）

[9] 北原，武田，市川，東倉：“音声言語認知における韻律の役割”，信学論（D），**J70-D**, 11，pp.2095-2101 （1987 年 11 月）

[10] 日本音響学会編：“音響用語辞典”，コロナ社,（1990 年）

[11] 武田一哉，匂坂芳典，片桐滋，阿部匡伸，桑原尚夫：“研究用日本語音声データベース利用解説書”，ATR Technical Report,（1988 年）

学会発表リスト

大野、赤木：“文音声中の基本周波数パターンに含まれる個人性の検討”，音声研究会（1998.3 発表）