修 士 論 文

音声中の感情表現に関連する物理量とその制御に関する研究

指導教官　赤木 正人 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

杉本 隆

2000 年 2 月 15 日

要 旨

音声中には言語情報の他にも個人性・感情といったパラ言語情報と呼ばれるものがある。よりよいマンマシンインターフェースを実現させるためには、このパラ言語情報に音声中のどのような特徴が関与するか知る必要がある。特に感情情報は、より人間らしく聞きやすい合成音声に欠かせない。音声中の物理量と感情情報との関係が明らかになれば、マンマシンインターフェースにおいて円滑なコミュニケーションがとれるようになるだけでなく、音声の本質の解明にとっても重要な知見をもたらすことが期待される。しかし、現在の合成音声は感情をうまく表現できていない。この一要因として、音声の生成機構に起因する物理量間の関係を、合成音では記述されていないことが考えられる。

本研究では音声中の物理量の変化が感情識別に与える影響を調べた。調べた結果と、音声の生成機構に起因する物理量間の関係より、感情制御のための物理量変換ルールを構築した。平静音声を、構築したルールによって感情制御し、感情知覚への影響を調べた。

目 次

図 目 次

表 目 次

第 1 章

序論

1.1 はじめに

人間は音声によるコミュニケーションの中で、言葉の意味だけでなく年齢・性別・個人性・感情・気分などの様々な情報を読み取っている。これらの情報は普通、表情や態度、声を組み合わせて表現されるが、例えば電話で話していても相手が誰で、どんな感情なのか分かるように、音声だけからでもある程度の情報伝達は可能である。すなわち、音声中には言語情報の他にも個人性・感情といったパラ言語情報 [1] と呼ばれるものがある。

現在、人間とコンピュータのインターフェースとして、音声を用いることが実用化されつつある。コンピュータと、音声についてコミュニケーションをとることを考えた場合、ただ単に言語情報の伝達だけを扱うのではなく、パラ言語情報の伝達が行なわれれば、より円滑なコミュニケーションの実現が期待される。特に、感情の伝達は人間同士において重要な役割を担っており、人間らしく聞きやすい合成音声には感情情報が必要不可欠だと考えられる。

感情情報に音声中のどのような特徴が関与するか知ることができれば、合成音品質の向上につながるものと期待される。また、マンマシンインターフェースにおいても貢献をもたらし、音声の本質の解明にとっても重要な知見をもたらすことが期待される。

1.2 本研究の背景

感情という心理量は文化的背景や状況的文脈、音声の個人性や音韻性といった要素を切り離して考えることができないために、工学的に扱うためにある種の制約を設ける必要がある。

音声中に含まれる感情情報の抽出および制御を可能にするためには、ある感情を表現している音声が、他の感情を表現している音声と区別できる独特の特徴を備えていなければならない。また、同じ感情を表現している音声同士では、その特徴も同等でなければならない。しかもそれが、広く普遍的に観察されることが必要である。

次に感情を表す物理量だが、従来の研究では、主に基本周波数・振幅・発話速度に注目して行なわれている。林は発声時間とピッチ曲線による感情識別・同定について研究しており、ピッチ曲線が感情情報の伝達に重要であると報告している [4]。北原らは感情を表現する韻律成分、およびその構造について検討を行ない、怒りの表現にとって時間構造、歓喜・悲哀の表現にはピッチ構造が各々重要であることを示した [2]。平賀らはピッチ・振幅の時系列変化パターンに着目して研究を行なった。その結果、怒り・歓喜は時間に対する変化率が大きく、悲哀は変化率が小さいと報告している。また、サンプルに十分な感情表現がなされていれば、その時系列変化パターンは個人差も比較的許容できる範囲に収まると述べている [3]。このことについては、Noad らも感情を含むことによって音声に生ずる物理的変動の傾向が、性別・経歴などに依らないと報告している [6]。これらはサンプルとして感情的自由度の高い言葉を選択し、場合によってはほぼ妥当な結果を得ている。

また、感情語や物理量の選択に対して依存性をなくすために、感情を含むことによって音声に生ずる物理的変動と、そこから知覚される感情との対応づけを試みた研究もある [5]。

しかし、これらの報告で作成された合成音声は、感情の種類によって結果にばらつきがあり、全ての感情に対しては表現できていないという欠点もみられる。この一要因として感情同士、または音声の生成機構に起因する物理量同士の関係が、合成音声で記述されていないことが考えられる。

1.3 本研究の目的

本研究では、平静音声とその他の感情を含んだ音声とを分析・比較し、感情情報の抽出を行なう。また、平静音声に対して感情制御を施し、感情知覚への影響を調べる。

音声中の感情情報を明らかにすることができれば、マンマシンインターフェース、合成音声の品質向上などに応用ができる。

本研究では、感情制御に用いるルールを、分析で得られた特徴のフィードバックだけでなく、音声の生成機構に起因する物理量同士の関係を取り入れることによって構築する。これによって、より人間らしい感情制御を目指す。

感情情報の抽出･感情制御には音声分析変換合成法 STRAIGHT[8] を用いる。STRAIGHT

で分解された物理量を、構築した制御ルールで変換し、合成する。これにより、積極的に感情情報を制御できるだけでなく聴取実験で感情知覚への影響を調べるのに十分な音質を保つことができるため、より正確な結果が期待される。

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、過去の感情情報の研究の現状と問題点を指摘し、本研究の目的を明らかにする。

第 2 章では、本研究で用いる音声データを採取する。また、採取したデータが、どのような感情を含んでいるのか分類するため、聴取実験を行なう。

第 3 章では、本研究で用いるパターン記述モデル STRAIGHT の構造を説明し、その有用性を示す。

第 4 章では、音声データを様々な物理量で分析し、各物理量における感情間の距離について検討を行なう。

第 5 章では、分析結果と、音声の生成機構に起因する物理量同士の関係により、感情制御のための物理量変換ルールを構築する。

第 6 章では、物理量変換ルールによって合成された音声を用いた聴取実験 (Close Test、Open Test) により物理量変換ルールの有効性、感情表現に対し重要な物理量について検討する。

第 7 章では、全体の考察を行ない、第 8 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

感情分類実験

2.1　目的

本章では、分析に用いるための音声データを採取する。本研究では、感情情報を音声データの分析によって獲得するため、この行程は非常に重要である。

また、本研究で扱う感情は、話し手ではなく聞き手側に存在するものを指す。よって、採取したデータが、どのような感情を含んでいるのか分類するため、聴取実験を行なう。

2.2　音声データの採取

分析に用いるための音声データの、発話者・文章・録音方法などを示す。

発話者

発話者は、男性 2 名 (m1、m2)、女性 3 名 (f1、f2、f3) である。いずれも 20 歳代である。f1、f2、f3 は演劇経験者である。m1、m2、f1、f2 はヴォーカルスクールに通っている。f1、f3 は声優養成所に通っている。

以上のような人選を行なったのは、演劇経験者などが一般人 (演劇経験なし) に比べ、音声によって感情状態の表現を行なう手法を的確に心得ているからである [3]。

文章

言語情報は感情に対して、かなりの影響を及ぼすことが想像できる。よって音声データに用いる文章は、言語情報に依存しないように感情的自由度の高いものにしなければなら

ない。また、本研究では基本周波数を扱うため、声帯の振動を伴う母音または有声子音で構成されたものが望ましい。以上の制約に従って、音声データは「いいじゃない」という言葉を用いた。

録音方法

様々な文脈で「いいじゃない」が発声される対話文を用意した (表 2.1)。発話者には B の部分を、文脈を考慮した上で自由に感情を込めて発声してもらった。A の部分は他の話者が読んだ。また、それ以外に喜・怒・哀・楽・平静を意識して発声してもらった。合計 167 サンプルの音声を採取した (表 2.2)。

録音は防音室にて行なった。対話全体を防音室の外の DAT レコーダに入力し、標本化周波数 48kHz で録音した。これを標本化周波数 20kHz にダウンサンプリングして、特定音声部分を切り出し、ワークステーション (WS) に保存し音声データとした。録音に使用した機器を表 2.3に示す。

表 2.1: 「いいじゃない」を含んだ対話文の例

	(飲み会で盛り上がっているときに)
A	そろそろ帰らなきゃ。
B	いいじゃない、まだ。じゃ、あと 30 分。
A	いや、明日、朝早いし…。
B	え〜、じゃ、あと 10 分。ね、いいじゃない。
	(友人が新しい時計をしているのを見て)
B	うぁ、いいじゃない。この時計。
A	そう？昨日買ったんだ。やっぱり、いい感じでしょ。
B	うん。いいじゃない。似合ってるよ。
A	いいでしょ。いいでしょ。
B	だから、いいじゃないって言ってるでしょ。
	(ちょっとしたミスをしたときに)
A	こんな失敗、小学生でもしないぞ。
B	いいじゃない(余計なお世話だ)。
A	もういい。お前にはもう頼まない。
B	いいじゃない(そんなぁ…)。

表 2.2: 発話者の条件と採取したサンプル数

発話者	m1	m2	f1	f2	f3
演劇経験者			○	○	○
ヴォーカルスクール	○	○	○	○	
声優養成所			○		○
採取したサンプル数	28	28	55	28	28

表 2.3: 録音に使用した機器

機器	メーカー、機種
マイクロフォン	SONY C-536P
DAT レコーダ	SONY TCD-D10 PRO 2
マイクロフォンアンプ	SONY AC-148F

2.3 聴取実験

本研究では、話し手側の抱いた感情ではなく聞き手側に存在する感情を基準とする。従って、採取した音声データが、どのような感情を含んでいるのか分類するために、聴取実験を行なった。

女性話者 1 名 (f1) の音声 (「いいじゃない」) を呈示し、その音声が平静・歓喜・怒り・悲哀のどの感情のものかを強制判断させた。また、平静・歓喜・怒り・悲哀のどれでもないと感じた場合は、どのような感情なのか、あるいはどの感情とどの感情の組合わせなのかを回答させた。

練習 4 サンプル、本番 55 サンプルをランダムに呈示した。被験者は大学院生 11 名であり、その全てが正常聴力を有する。

被験者は防音室内でヘッドフォンにより受聴した。受聴はモノラルの両耳受聴である。被験者には聞き直しを許し、パーソナルコンピュータ (PC) を用いて回答させた。音声データは防音室の外に設置された WS 内に保存されており、被験者の応答に応じて呈示される。聴取実験システムの全体図を図 2.1に、使用した機器を表 2.4に示す。

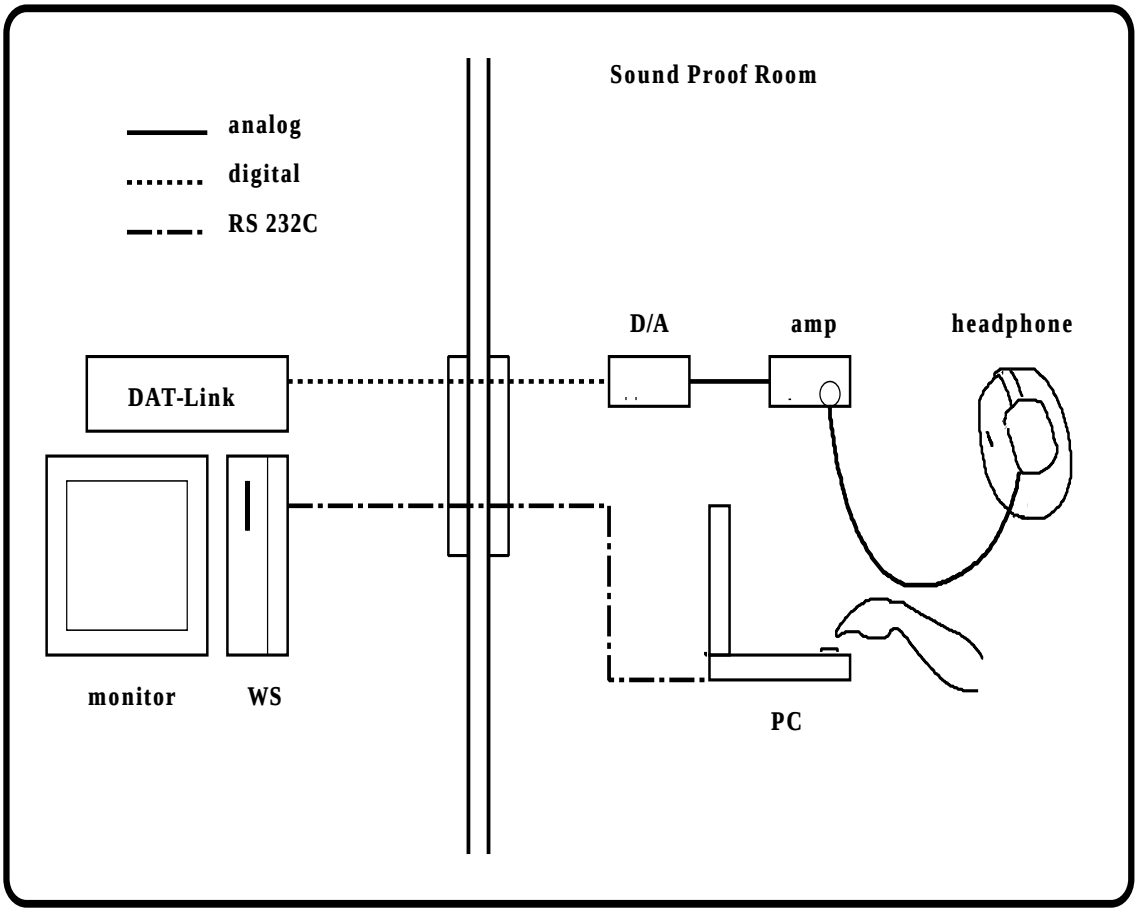

図 2.1: 実験システム

表 2.4: 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	SENNHEISER HDA 200
ヘッドホンアンプ	SANSUI AUα-907MR
WS	Sun S-4/IX
PC	Macintosh PowerBook Duo

2.4　実験結果と考察

実験結果を図 2.2に示す。
図は、縦軸が識別率[1]、横軸がサンプル番号[2]である。上から、平静・歓喜・怒り・悲哀の結果である。

図から分かるように、識別率 100%のサンプルが、平静・怒りについて 2 個、悲哀について 3 個存在する。歓喜については識別率 100%のサンプルが無いが、識別率 80%以上をその感情音声だと考えると 5 個存在し、同様に平静について 9 個・怒りについて 4 個・悲哀について 8 個存在する。
上記の識別率およびサンプル数は、感情情報の抽出を目的とした分析にとって十分だと考えられる。

[1]本論文では、あるサンプルについて、同一の感情を表現していると回答した被験者の割合を識別率と呼ぶことにする。

[2]平静に対する識別率が高い順に並べてある。

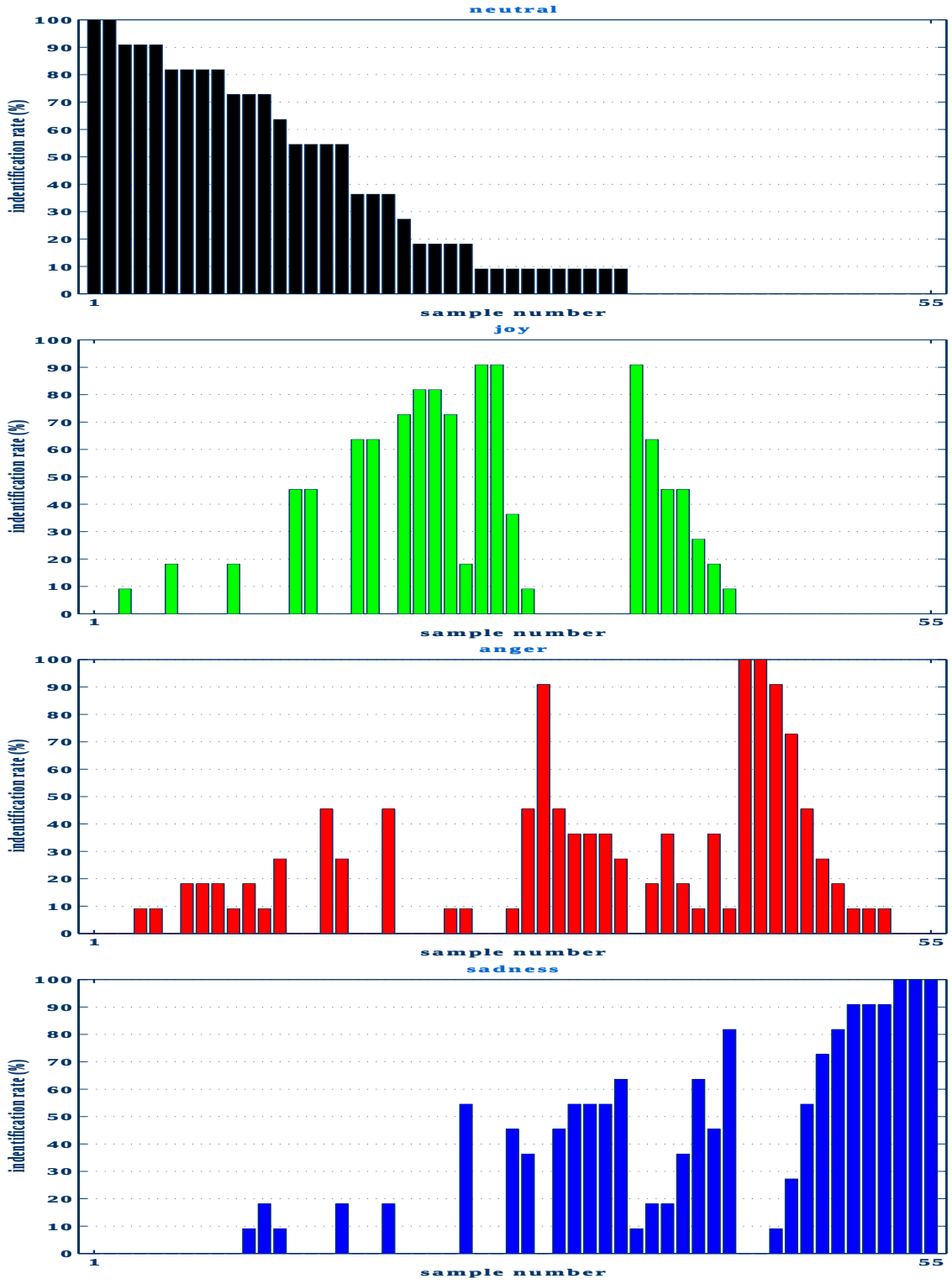

図 2.2: 感情分類実験の結果

第 3 章

パターン記述モデル

3.1 目的

本研究では、感情情報に関する物理量としてパワー・基本周波数・発話時間・スペクトルを取り扱う。そのため、これら時間変化パターンの情報を抽出するモデルを構築しなければならない。また本研究では、時間変化を考慮し、抽出した物理量の変換および合成音声の作成を行なう。よって時間変化パターン記述モデルとして音声分析変換合成系を用いる必要がある。このための音声分析変換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 using Adaptive Interpolation of weiGHTed spectrogram)[8] を採用する。

本章では、その構造を説明し、本研究での有用性を示す。

3.2 STRAIGHT

STRAIGHT は STRAIGHT-core、SPIKES、TEMPO2 の 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行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[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。同一のパワースペクトルであっても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。

[1]正確には群遅延である。

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。

3.2.1 STRAIGHT-core

STRAIGHT-core の重要なアイデアは、有声音に見られるほぼ周期的な励振を、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。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 次のカーディナル B-スプライン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、周期性の影響を雑音や誤差に強い形で選択的に除去することを狙っている。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。

$$s(t) = \sum_{k \in N} \alpha_k(t) \sin\left(\int_{t_0}^{t} k \left(\omega_0(\tau) + \omega_k(\tau) \right) d\tau + \phi_k \right) \tag{3.1}$$

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

実効的な長さが 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ガボール型時間窓は、そのような窓の一例である。

$$\omega_G(t) = \exp\left(-\pi\left(\frac{t}{\eta^{t_0}}\right)^2\right)\exp\left(\frac{j2\pi t}{t_0}\right) \tag{3.2}$$

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。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 pi 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持つようになる。

$$\omega_c(t) = \omega(t)\sin\left(\pi\frac{t}{\tau_0}\right) \tag{3.3}$$

時間方向に伸長した時間窓 $(\eta > 1)$ で得られたスペクトル $P_o(\omega, t)$ 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。

$$P_r(\omega, t) = \sqrt{P_o^2(\omega, t) + \xi(\eta)P_c^2(\omega, t)} \tag{3.4}$$

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伸ばすだけ $(\eta > 1.3)$ 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 次のカーディナル B-スプライン基底関数 $h_t(\omega)$ を周波数方向の平滑化関数とする。

$$h_t(\omega) = 1 - \left|\frac{\omega}{\omega_0(t)}\right| \tag{3.5}$$

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。

$$S(\omega,t)=\sqrt{g^{-1}\left(\int_D h_t(\lambda,t)g(|P_r(\omega-\lambda,t)|^2)d\lambda\right)} \tag{3.6}$$

ここで D は、平滑化関数の定義域を表す。式 3.6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、過剰な平滑化が行なわれてしまうためである。最適平滑化関数は、スプライン関数の性質を利用すると、窓関数の周波数表現と 2 次のカーディナル B-スプライン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見なしたときの逆フィルタの応答を計算することで求めることができる。

3.2.2 SPIKES

SPIKES のアイデアは、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になった。

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。

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

$$y(t)=\sum_{t_i\in Q}\frac{1}{\sqrt{G(f_0(t_i))}}v_{ti}(t-T(t_i)) \tag{3.7}$$

$$v_{ti}(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} V(\omega, t_i)\Phi(\omega) \exp\left(j\omega(t)\right) d\omega \tag{3.8}$$
$$where\ T(T_i) = \sum_{t_k \in Q, k<i} \frac{1}{G(f_0(t_k))}$$

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。

ここで、$V(\omega, t_i)$ は、$A(), u(), r()$ をそれぞれ振幅、周波数、時間軸の変換としたとき、変換された振幅スペクトル $A(S(u(\omega), r(t)), u(\omega), r(t))$ 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。q は、ケフレンシーを表す。

$$V(\omega, t) = \exp\left(\frac{1}{\sqrt{2\pi}} \int_0^{\infty} h_t(q) \exp\left(j\omega q\right) dq\right) \tag{3.9}$$
$$h_t(q) = \begin{cases} 0 & (q < 0) \\ c_t(0) & (q = 0) \\ 2c_t(q) & (q > 0) \end{cases} \tag{3.10}$$
$$and \qquad c_t(q) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(j\omega q\right) \log A d\omega \tag{3.11}$$

オールパスフィルタの設計

オールパスフィルタの最初の成分 $\Phi_1(\omega)$ は、次式で表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。

$$\Phi_1(\omega) = \exp\left(j\omega f_s T_d\right) \tag{3.12}$$

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、正規化角周波数が 2π のときの値が、n を任意の整数とすると $2n\pi$ であるとの拘束条件を満たすように、次式を用いる。

$$\Phi_1(\omega) = \exp\left(j(\pi f_s T_d + p(\omega))\right) \tag{3.13}$$

$$p(\omega) = \begin{cases} \frac{2\pi a}{1+\exp((\omega+\pi)/\omega_\omega)} & \omega \geq 0 \\ \frac{2\pi a}{1+\exp((\omega-\pi)/\omega_\omega)} & \omega < 0 \end{cases} \tag{3.14}$$

$$a = \lceil f_s T_d \rceil - f_s T_d \tag{3.15}$$

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。

3.2.3 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 3.2) とカーディナル B-スプライン基底関数 (式 3.5) 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 $\omega_c(t;\lambda)$ へ等しく写像される点 (不動点) の集合 $\Lambda(t)$ は次のように定義される。

$$\Lambda(t) = \{\lambda | \omega_c(t;\lambda) = \lambda, \omega_c(t;\lambda-\epsilon) - (\lambda-\epsilon) > \omega_c(t;\lambda+\epsilon) - (\lambda+\epsilon)\} \tag{3.16}$$

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。図 3.1に、音声「いいじゃない」の基本周波数推定の結果を示す。図は、上から不動点のマップ、C/N 比、推定された基本周波数を表している。横軸は時間である。

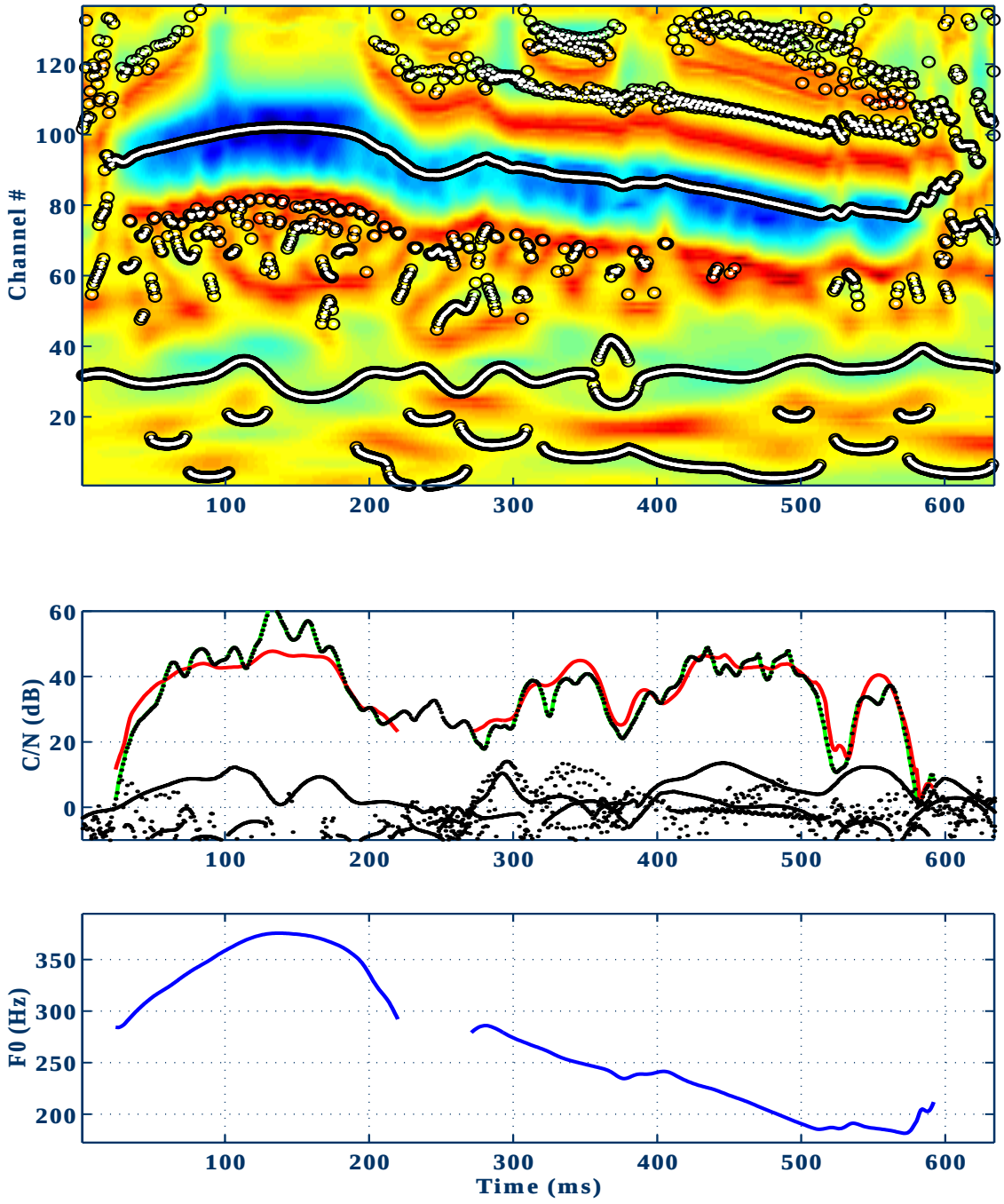

図 3.1: TEMPO2 による基本周波数推定結果

3.3 まとめ

STRAIGHT を用いる理由は、パワー・基本周波数・発話時間・スペクトルといった時間変化する特徴を抽出・変換・合成するためである。

基本周波数は TEMPO2 により、パワー・発話時間・スペクトルは STRAIGHT-core により計算される。計算された物理量は、感情情報を抽出するのに十分な正確さを持っている。また、SPIKES により、物理量を変換した後の合成で、高品質な合成音声を生成できる。

本研究では、STRAIGHT によって得られた物理量を分析した。

また、その物理量を変換し STRAIGHT で合成することで合成音を作成した。

なお、本研究での STRAIGHT による分析条件は、表 3.1の通りである。

表 3.1: STRAIGHT の分析条件

分析窓長	40 (ms)
分析シフト幅	1 (ms)
FFT 長	1024 (point)
基本周波数検索範囲	20 - 1000 (Hz)

第 4 章

分析

4.1 目的

音声データについて、時間変化パターンを STRAIGHT により抽出し、抽出された各パラメータに表れる感情間の距離について分析、検討を行なう。

4.2 過去の研究の分析結果

音声データを分析する前に、過去の研究の主要な分析結果を紹介する。

基本周波数についての分析結果

怒り・歓喜の基本周波数は類似した形状を有しており、平静に比べ高い [6]。また、基本周波数のダイナミックレンジ・変化率[1]も大きい。悲哀の基本周波数は平静に比べ低く、ダイナミックレンジ・変化率も小さい [2, 3]。

基本周波数の構造は、歓喜・悲哀の感情表現において重要な役割を担っている [2]。

振幅についての分析結果

怒りの振幅は平静に比べ大きく、悲哀は小さい [7]。歓喜は平静と同じパターンを有する [2]。

[1] フレーム毎の差分の絶対値を取り、平均を求めたものを変化率とする。

発話速度についての分析結果

怒りは平静に比べ発話速度が速く、悲哀は遅い。歓喜は平静からの時間伸縮は殆んど見られない [3]。

発話速度は、怒りの感情表現において重要な役割を担っている [2]。また、歓喜の感情表現においては、あまり重要ではない [3]。

4.3 音声データの分析

前節をふまえ、感情識別実験で使用した音声データを分析する。

図 4.1に、時間変化パターンの抽出結果の例を示す。図は上から、原波形、基本周波数、パワー、スペクトルムーヴメント、スペクトログラムである。音声はスペクトルムーヴメントを手がかりに音韻毎に分けた。

分析した物理量を、以下に示す。

1. 全発声区間
   (a) 発話時間
   (b) 基本周波数の長時間平均
   (c) パワーの長時間平均
   (d) 基本周波数の変化率
   (e) パワーの変化率
2. 音韻毎 (語頭の /ii/ と語尾の /ai/ のみ)
   - 全発声区間と同様

また、本研究では、各感情音声の、平静音声からの比率について検討する。平静音声は、感情分類実験で、平静の認識率が 100%のサンプルを用いた。

以下に示す分析結果の図では、感情間の特徴を明らかにするために、各感情における高認識率のサンプル 3 個ずつについては、別の印でプロットしてある。(歓喜：○、怒り：＊、悲哀：×)

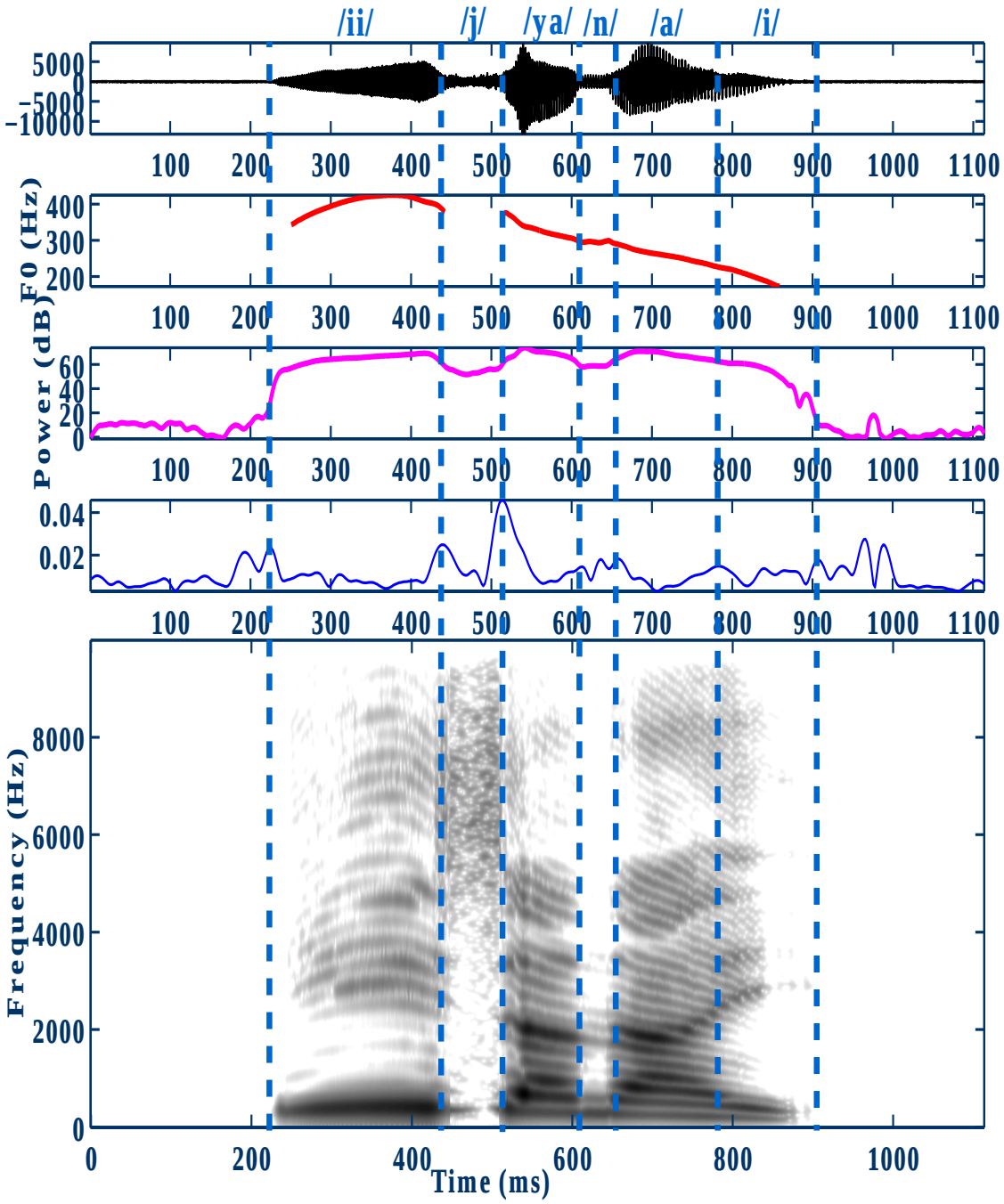

図 4.1: 時間変化パターンの抽出結果

全発声区間での基本周波数・パワーの長時間平均、発話時間

全発声区間での基本周波数・パワーの長時間平均、発話時間における平静からの比率について検討する。

平静音声を基準とした発話時間の伸縮率を $TC(\%)$、基本周波数・パワーの長時間平均の増減率を、それぞれ $FC(\%)$・$PC(\%)$ とする。分析結果を図 4.2に示す。

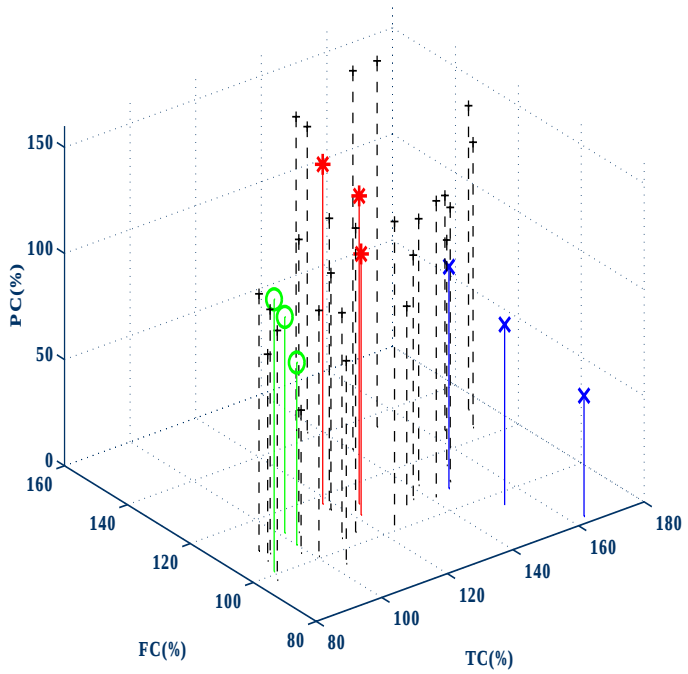

図 4.2: 分析結果 - 全発声区間での基本周波数・パワーの長時間平均、発話時間

図 4.2より、以下のことが分かる。

歓喜：TC、FC、PC 全てにおいてほぼ 100%、つまり平静と同じである。

怒り：$TC \simeq 120(\%)$、$FC \simeq 120(\%)$、$PC \simeq 140(\%)$ である。基本周波数・パワーの長時間平均が共に調査した感情中最大であり、この結果は過去の研究結果と合致する。しかし、発話時間が平静よりも長いという結果は、過去の研究結果と異なる。

悲哀：$TC \simeq 150(\%)$、$FC \simeq 90(\%)$、$PC \simeq 80(\%)$ である。発話時間が最長、基本周波数・パワーの長時間平均が共に最小であり、この結果は、過去の研究結果と合致する。

歓喜は平静と同じ傾向を示しており、怒りと悲哀は、パワー・基本周波数の長時間平均では互いに逆の傾向を示していることが分かる。怒りにおける発話時間の、過去の研究結

果との違いは、怒りという感情の中でも、Hot Anger と Cold Anger と呼ばれるものがあり、物理量の変化の仕方が違う [3] ためだと考えられる。

全発声区間での基本周波数・パワーの変化率

全発声区間での基本周波数・パワーの変化率における平静からの比率について検討する。

平静音声を基準とした基本周波数・パワーの変化率の増減率を、それぞれ$FVC(\%)$・$PVC(\%)$とする。分析結果を図 4.3に示す。

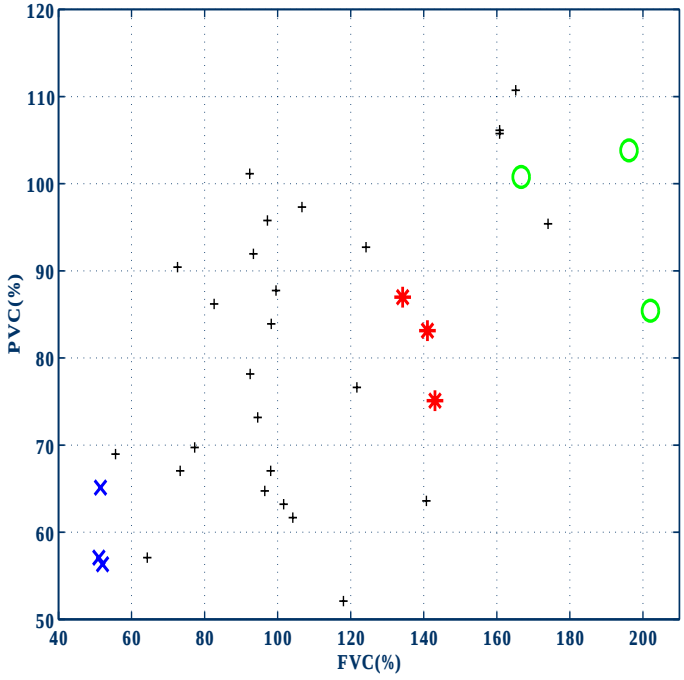

図 4.3: 分析結果 - 全発声区間での基本周波数・パワーの変化率

図 4.3より、以下のことが分かる。

歓喜：$FVC \simeq 200(\%)$、$PVC \simeq 100(\%)$ である。基本周波数の変化率は唯一、全発声区間で分析した物理量で平静と歓喜とが大きく異なるものである。

怒り：$FVC \simeq 140(\%)$、$PVC \simeq 80(\%)$ である。基本周波数の変化率が平静より大きいという結果は過去の研究結果と合致する。しかし、パワーの変化率が平静より小さいという結果は過去の研究結果と異なる。

悲哀：$FVC \simeq 50(\%)$、$PVC \simeq 60(\%)$ である。基本周波数・パワーの変化率が共に最小であり、この結果は、過去の研究結果と合致する。

基本周波数の変化率は歓喜の感情表現にとって重要であると考えられる。パワーの変化率は、同じ感情内でも値のばらつきが大きく、感情に対する重要度が低いと考えられる。

音韻/ii/、/ai/での基本周波数・パワーの長時間平均、発話時間

全発声区間と同様にして、音韻/ii/、/ai/について、基本周波数・パワーの長時間平均、発話時間における平静からの比率について検討する。

音韻/ii/についての分析結果を図 4.4に示す。

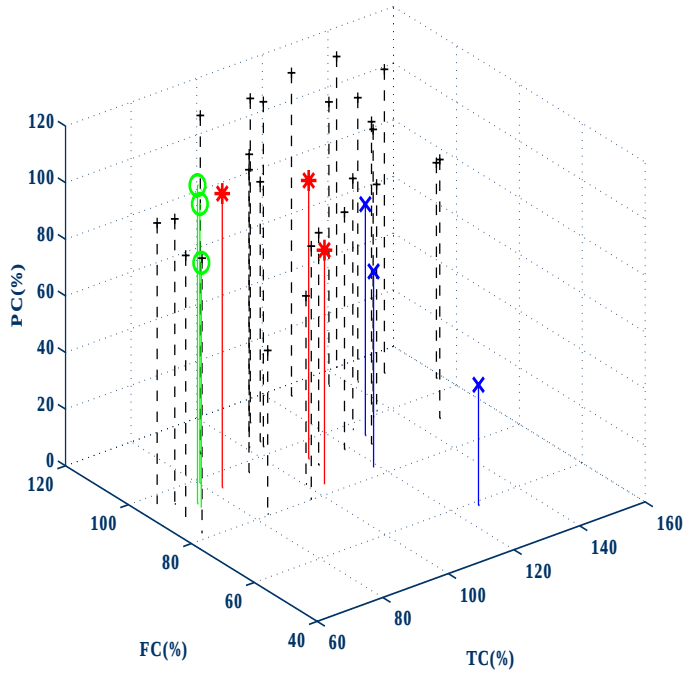

図 4.4: 分析結果 - 音韻/ii/での基本周波数・パワーの長時間平均、発話時間

図 4.4より、以下のことが分かる。

歓喜：$FC \simeq 100(\%)$、$PC \simeq 100(\%)$、$TC \simeq 80(\%)$ である。

怒り：$TC \simeq 100(\%)$、$PC \simeq 100(\%)$、$FC \simeq 90(\%)$ である。

悲哀：$TC \simeq 130(\%)$、$FC \simeq 80(\%)$、$PC \simeq 80(\%)$ である。発話時間が最長、基本周波数・パワーの長時間平均が共に最小であり、この結果は、全発声区間での分析結果と合致する。

先の発声区間全体での分析を考慮すると、これらは妥当な結果である。
音韻/ai/についての分析結果を図 4.5に示す。

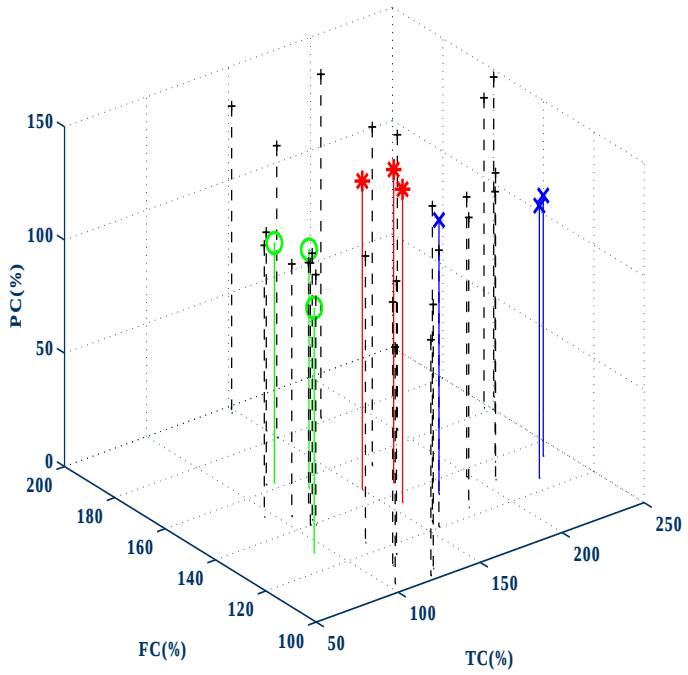

図 4.5: 分析結果 - 音韻/ai/での基本周波数・パワーの長時間平均、発話時間

図 4.5より、以下のことが分かる。

歓喜：$TC \simeq 100(\%)$、$PC \simeq 100(\%)$ であるが、$FC \simeq 140(\%)$ である。基本周波数が高いのは、全発声区間での分析結果と異なる傾向である。

怒り：$TC \simeq 150(\%)$、$FC \simeq 140(\%)$、$PC \simeq 120(\%)$ である。この結果は全発声区間での分析結果と同じ傾向である。

悲哀：$TC \simeq 200(\%)$、$FC \simeq 120(\%)$、$PC \simeq 110(\%)$ である。

歓喜の基本周波数が高いのは、全発声区間では見られなかった傾向である。語尾の基本周波数が高いことが、歓喜の感情表現にとって重要である可能性がある。

音韻/ii/、/ai/での基本周波数・パワーの変化率

全発声区間と同様にして、音韻/ii/、/ai/について、基本周波数・パワーの変化率における平静からの比率について検討する。

音韻/ii/についての分析結果を図 4.6に示す。

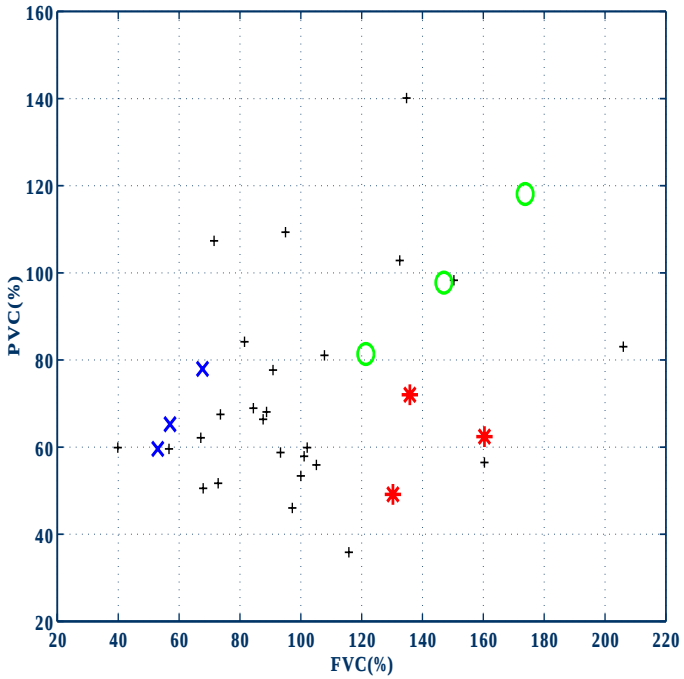

図 4.6: 分析結果 - 音韻/ii/でのパワー・基本周波数の変化率

図 4.6より、以下のことが分かる。

歓喜：$FVC \simeq 150(\%)$、$PVC \simeq 100(\%)$ である。この結果は全発声区間での分析結果と同じ傾向を持っている。

怒り：$FVC \simeq 140(\%)$、$PVC \simeq 60(\%)$ である。この結果は全発声区間での分析結果と同じ傾向を持っている。

悲哀：$FVC \simeq 50(\%)$、$PVC \simeq 70(\%)$ である。この結果は全発声区間での分析結果と同じ傾向を持っている。

各感情において、全発声区間の分析結果と同じ傾向が表れた。よって音韻/ii/でのパワー・基本周波数の変化率における特徴は、全発声区間での特徴に包含される。

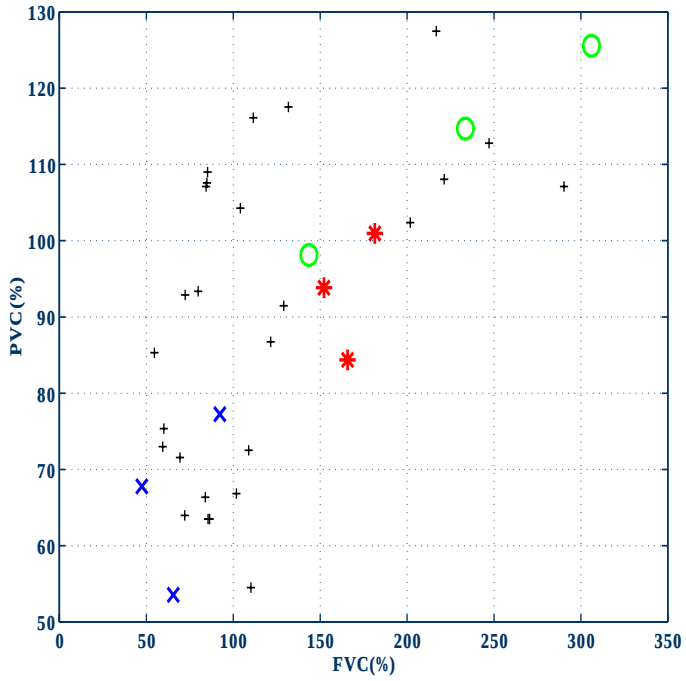

図 4.7: 分析結果 - 音韻/ai/でのパワー・基本周波数の変化率

音韻/ai/についての分析結果を図 4.7に示す。
図 4.7より、以下のことが分かる。

歓喜：$FVC \simeq 250(\%)$、$PVC \simeq 110(\%)$ である。この結果は全発声区間での分析結果と同じ傾向を持っている。但し、基本周波数の変化率の増大が顕著になっている。

怒り：$FVC \simeq 150(\%)$、$PVC \simeq 90(\%)$ である。この結果は全発声区間での分析結果と同じ傾向を持っている。

悲哀：$FVC \simeq 70(\%)$、$PVC \simeq 70(\%)$ である。この結果は全発声区間での分析結果と同じ傾向を持っている。

各感情において、全発声区間の分析結果と同じ傾向が表れた。但し、歓喜の基本周波数の変化率の増大が顕著になっており、基本周波数の長時間平均の上昇と組み合わせて考えることができる。

4.4 まとめ

分析の結果、過去の研究の分析結果とほぼ同じ傾向が得られた。しかし、怒りの発話時間など若干異なる傾向も認められる。

全発声区間での分析結果を表 4.1に示す。

表 4.1: 分析結果

感情	TC(%)	FC(%)	PC(%)	FVC(%)	PVC(%)
歓喜	100	100	100	200	100
怒り	120	120	140	140	80
悲哀	150	90	80	50	60

パワーの変化率は、同じ感情内でも値のばらつきが大きく、感情に対する重要度が低いと考えられる。

また、音韻/ii/、/ai/でも分析したが、発声区間全体の分析結果と照合すると、類似している部分が多い。よって音韻毎での特徴は、全発声区間での特徴に包含されていると考えられる。

以上より、次章の物理量変換ルールでは、パワーの変化率および音韻毎での分析結果をフィードバックしない。

ただし、音韻/ai/での歓喜の基本周波数の長時間平均・変化率の上昇は、全発声区間での分析ではない新しい結果であるので、物理量変換ルールに組み込むことにする。

表 4.1を見ると、PC と FC に線形な関係が存在することが想像できる。この詳細については次章に示す。

第 5 章

感情制御のための物理量変換ルール

5.1　分析により導かれるルール

前章の分析結果をフィードバックさせることによって導かれる、感情制御のための物理量変換ルールは以下の通りである。

歓喜：発話時間・基本周波数の長時間平均・パワーの長時間平均は平静と同じ。
基本周波数の変化率を 2 倍に増加させる。
語尾のみ基本周波数を上げる。

怒り：発話時間を 1.2 倍に伸長させる。
基本周波数の長時間平均を 1.2 倍に増加させる。
パワーの長時間平均を 1.4 倍に増加させる。
基本周波数の変化率を 1.4 倍に増加させる。

悲哀：発話時間を 1.5 倍に伸長させる。
基本周波数の長時間平均を 0.9 倍に減少させる。
パワーの長時間平均を 0.8 倍に減少させる。
基本周波数の変化率を 0.5 倍に減少させる。

また、発話時間の伸縮に対して特別な処置を施す。

発話時間の伸縮

図 5.1から、どの感情でも子音部の発話時間がほぼ同じであることが分かる。

従って、本研究では、スペクトルの変動を手がかりにして、図 5.2のような時間軸変換マップを作成し、発話時間を伸縮した場合に子音部があまり伸縮されないようにした。図は、発話時間を 0.5 倍に縮小させる場合のマップである。

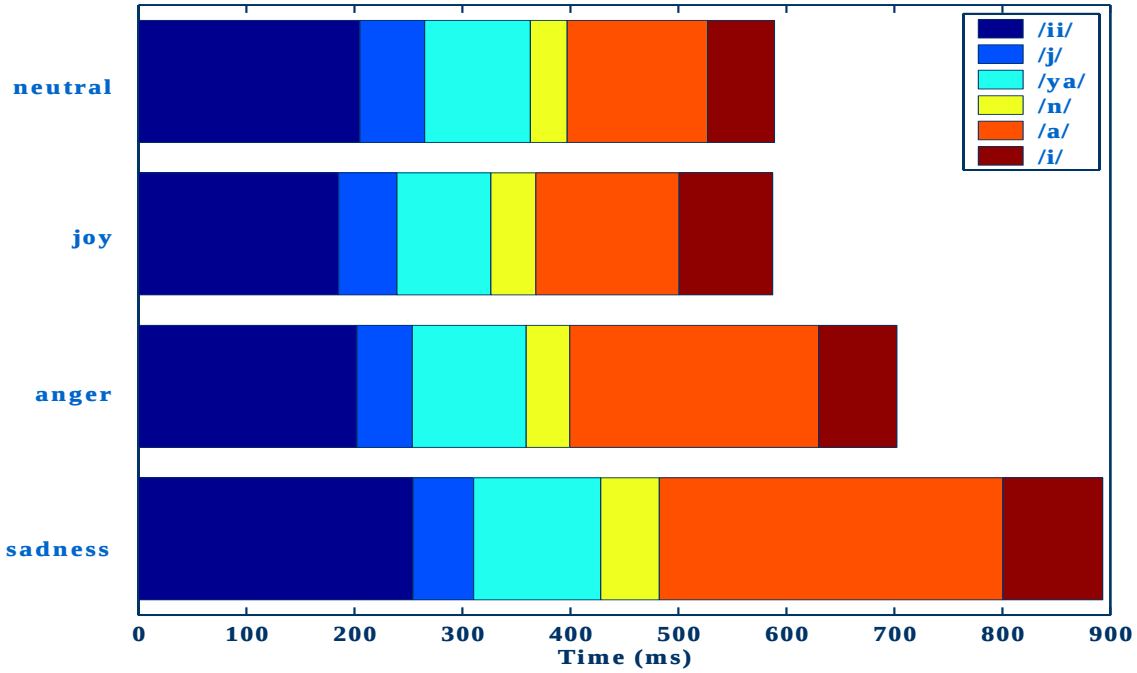

図 5.1: 音韻毎の発話時間

5.2 Lombard Effect

雑音の存在で発声者の発声様式が変化し、その結果音声波形が変形することを Lombard Effect[9] と呼ぶ。

これは、声量が大きくなることによって音声中の他の物理量も関連して変化する例である。本研究では、怒りの感情表現にて声量が大きくなることが分かっている。従って、怒りの感情制御に Lombard Effect のモデル化が大変有効だと考えられる。

Lombard Effect によって音声は次のような変化をすることが報告されている [10]。

- 基本周波数の上昇
  基本周波数は声量が大きくなることにより、通常の声より上昇する。

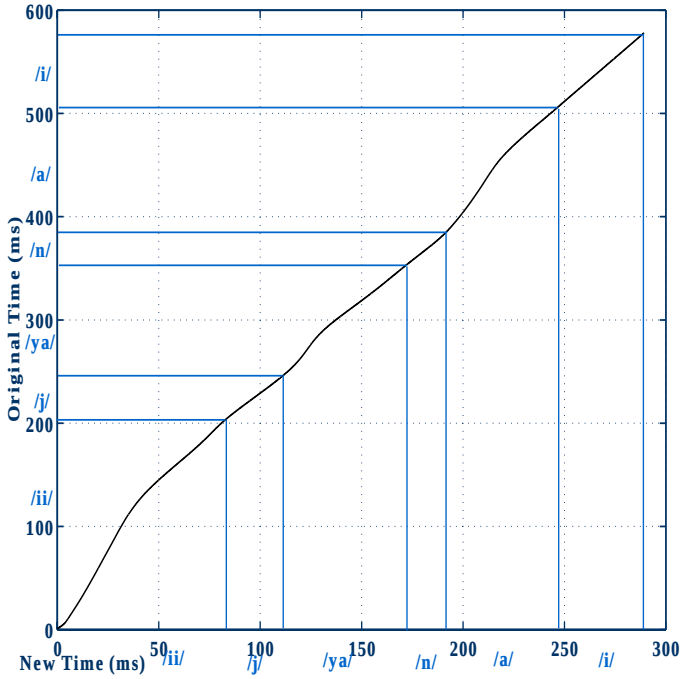

図 5.2: 発話時間の伸縮

- ホルマント周波数のシフト
  ホルマント周波数は Lombard Effect により、通常の声より 1.5kHz を境にして、それよりも高い周波数にあるものは低域に、低い周波数にあるものは高域に、120Hz 前後移動することが報告されている [11]。

- 高域スペクトル成分の増加
  Lombard Effect が起こると、高域スペクトル成分が顕著に増加する。

5.2.1 基本周波数の上昇

発声区間全体のパワーの長時間平均と、基本周波数の長時間平均との分析結果を図 5.3 に示す。

図 5.3から分かるように、パワーの長時間平均と、基本周波数の長時間平均との間には線形的な関係が成り立っている。この関係を、簡単な式で表す (式 5.1)。

$$FC = 0.5 \times PC + 50 \tag{5.1}$$

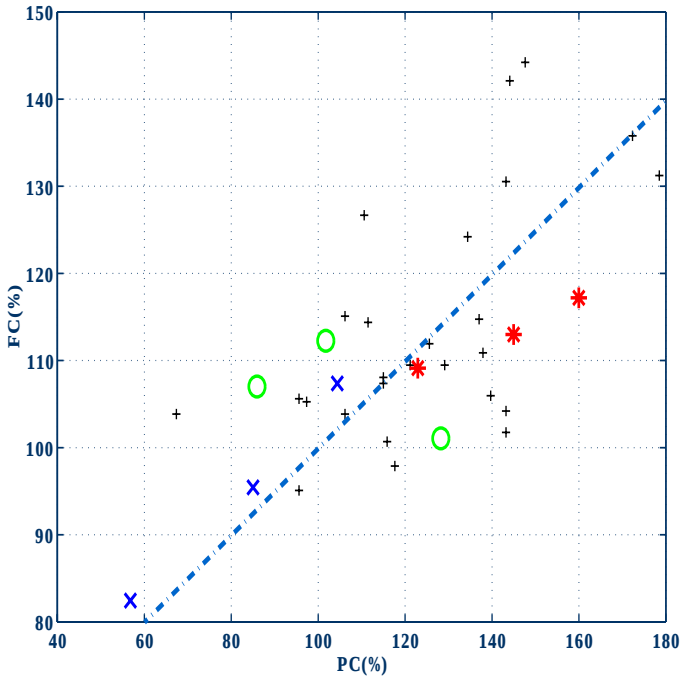

図 5.3: パワーと基本周波数の関係

分析により導かれるルールは、既に式 5.1と対応している。

5.2.2 ホルマント周波数のシフト

Lombard Effect によりホルマント周波数は、1.5kHz を境にして、それよりも高い周波数にあるものは低域に、低い周波数にあるものは高域に移動する。

従って、本研究では怒りの感情制御のとき、図 5.4のような周波数変換マップを作成し、ホルマント周波数を、1.5kHz 以下は上昇、1.5kHz 以上は下降させる。

5.2.3 高域スペクトル成分の増加

Lombard Effect により高域スペクトル成分が顕著に増加する。

従って、本研究では怒りの感情制御のとき、図 5.5のように高域[1]スペクトル成分を 5dB 増加させる。図は、左が高域スペクトル成分増加前、右が高域スペクトル成分増加後である。

[1]本研究では 5kHz 以上を高域とした。

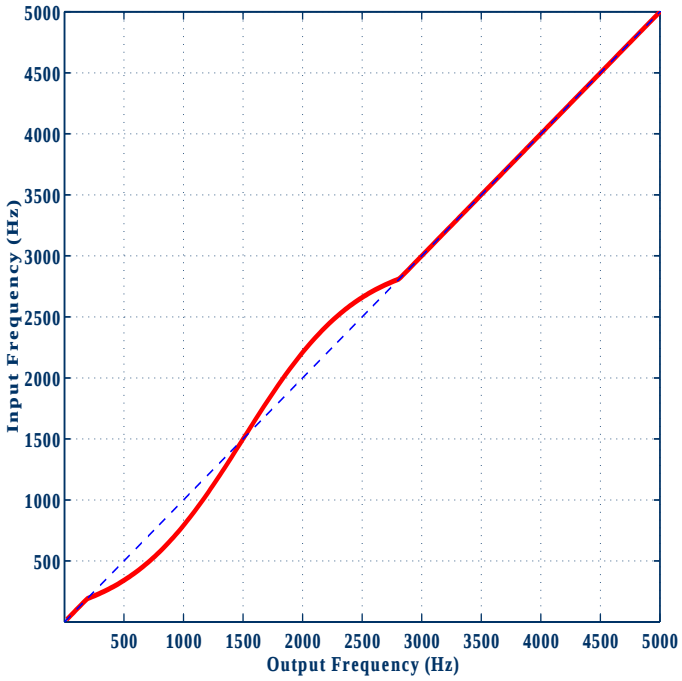

図 5.4: ホルマント周波数のシフト

5.3 まとめ

以上より構築される感情制御のための物理量変換ルールを以下に示す。

歓喜： 発話時間・基本周波数の長時間平均・パワーの長時間平均は平静と同じ。
基本周波数の変化率を 2 倍に増加させる。
語尾のみ基本周波数を上げる。

怒り： 発話時間を 1.2 倍に伸長させる (子音部伸縮なし)。
基本周波数の長時間平均を 1.2 倍に増加させる。
パワーの長時間平均を 1.4 倍に増加させる。
基本周波数の変化率を 1.4 倍に増加させる。
図 5.4によってホルマント周波数をシフトさせる。
高域スペクトル成分を 5dB 増加させる。

悲哀： 発話時間を 1.5 倍に伸長させる (子音部伸縮なし)。
基本周波数の長時間平均を 0.9 倍に減少させる。

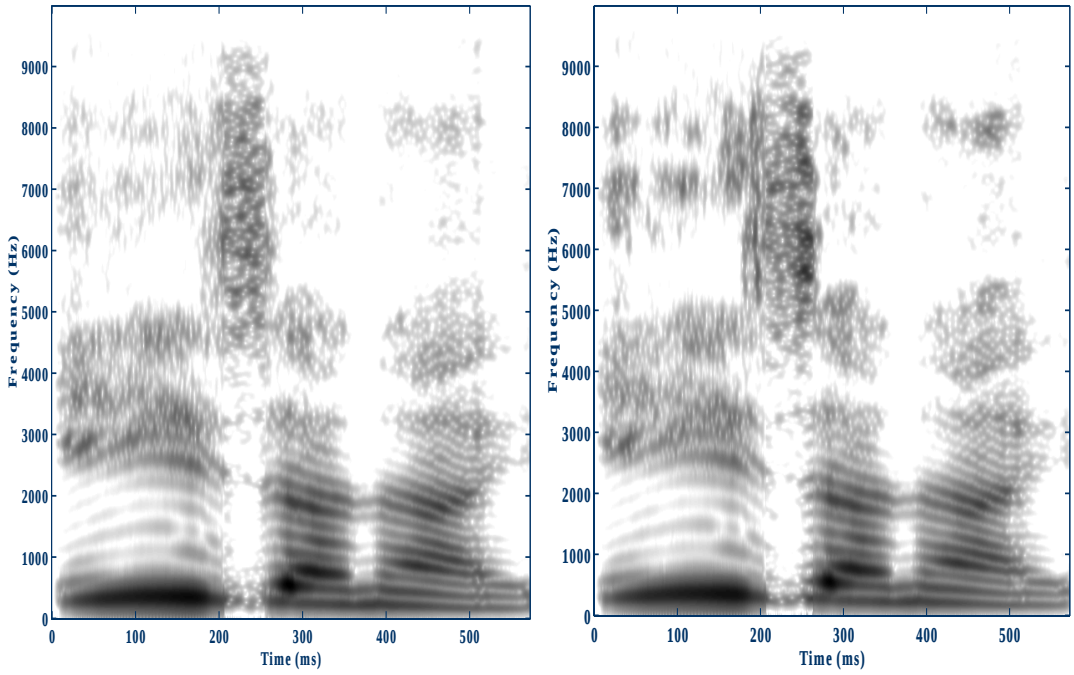

図 5.5: 高域スペクトル成分の増加

パワーの長時間平均を 0.8 倍に減少させる。
基本周波数の変化率を 0.5 倍に減少させる。

第 6 章

聴取実験

6.1 目的

感情制御した合成音声が、感情を含んでいるのか確認するために、聴取実験を行なった。

6.2 実験方法

実験に用いる合成音声の、物理量変換ルールを以下に示す。

neutral1： 平静音声あるいはデータベース上の音声を STRAIGHT で再合成する (物理量変換なし)。

joy1： 基本周波数の変化率を 2 倍に増加させる。

joy2： 語尾のみ基本周波数を上げる。

anger1： パワーの長時間平均を 1.4 倍に増加させる。

anger2： 基本周波数の長時間平均を 1.2 倍に増加させる。

anger3： 基本周波数の変化率を 1.4 倍に増加させる。

anger4s： 発話時間を 0.8 倍に短縮させる (子音部伸縮なし)。

anger4l： 発話時間を 1.2 倍に伸長させる (子音部伸縮なし)。

anger5： ホルマント周波数をシフトさせる。

anger6： 高域スペクトル成分を 5dB 増加させる。

sadness1： パワーの長時間平均を 0.8 倍に減少させる。

sadness2： 基本周波数の長時間平均を 0.9 倍に減少させる。

sadness3： 基本周波数の変化率を 0.5 倍に減少させる。

sadness4： 発話時間を 1.5 倍に伸長させる (子音部伸縮なし)。

怒りの感情の発話時間に対して、本研究と過去の研究で分析結果が異なっているため、伸長・短縮の両方を用意した。

実験では、合成音声を呈示し、その音声が平静・歓喜・怒り・悲哀のどの感情のものかを判断させた。

はじめに平静音声を呈示し、その後、平静音声に物理量変換ルールを適用して作成した音声刺激をランダムに呈示した。被験者は大学院生 10 名であり、その全てが正常聴力を有する。

被験者は防音室内でヘッドフォンにより受聴した。受聴はモノラルの両耳受聴である。被験者には聞き直しを許し、PC を用いて回答させた。音声データは防音室の外に設置された WS 内に保存されており、被験者の応答に応じて呈示される。使用した機器を表 6.1 に示す。

表 6.1: 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
ヘッドホンアンプ	STAX SRM-1/MK-2 P.P
WS	Sun S-4/IX
PC	Macintosh PowerBook Duo

6.3 Close Test

Close Test として、分析に用いた音声「いいじゃない」の平静音声に物理量変換ルールを適用して作成した音声刺激に、感情分類実験で高識別率を得たサンプルを加えて聴取実験を行なった。結果を図 6.1、表 6.2に示す。

表 6.2: 実験結果 - 「いいじゃない」の認識率

サンプル	平静 (%)	歓喜 (%)	怒り (%)	悲哀 (%)	適用したルール
1	**80**	0	5	15	平静の識別率 100%
2	**80**	0	0	20	neutral1
3	25	**50**	10	15	joy1
4	10	**90**	0	0	joy1、2
5	0	**100**	0	0	歓喜の識別率 91%
6	95	0	**0**	5	anger1
7	25	5	**25**	45	anger1、2
8	10	15	**75**	0	anger1、2、3
9	10	25	**65**	0	anger1、2、3、4s
10	15	30	**55**	0	anger1、2、3、4l
11	5	0	**95**	0	anger1、2、3、4s、5
12	0	25	**75**	0	anger1、2、3、4l、5
13	0	0	**100**	0	anger1、2、3、4s、5、6
14	0	0	**100**	0	anger1、2、3、4l、5、6
15	0	0	**100**	0	怒りの識別率 100%
16	90	0	0	**10**	sadness1
17	80	0	5	**15**	sadness1、2
18	25	0	0	**75**	sadness1、2、3
19	10	10	0	**80**	sadness1、2、3、4
20	0	0	0	**100**	悲哀の識別率 100%

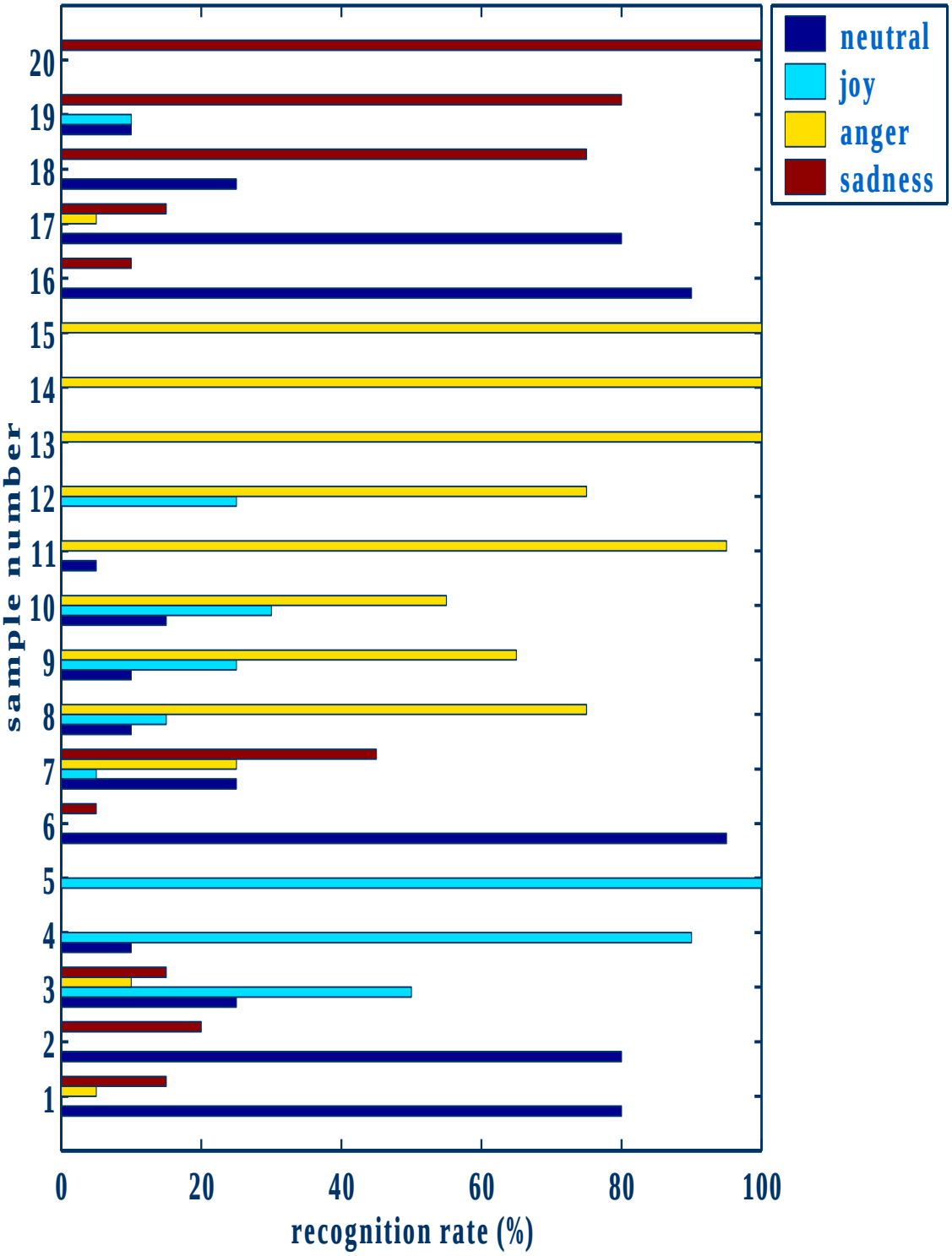

図 6.1: 実験結果 - 「いいじゃない」

図 6.1、表 6.2より、以下のことが分かる。

- サンプル番号 4(歓喜)、13 および 14(怒り)、19(悲哀) と、全ての感情において高い認識率を獲得している。
- 歓喜の感情表現においてルール joy1、joy2 は 2 つ共に有効・重要である。
- 怒りの感情制御のときに、物理量同士の関係を記述することで、過去の研究での変換 (サンプル番号 9 まで) よりも高い認識率が得られている。
- サンプル番号 18 で、急激に悲哀の認識率が上がっている。
- すべての感情において、今回作成した合成音は、高識別率を得たサンプルと同等もしくは近い認識率が得られている。

6.4 Open Test

Open Test として「いいじゃない」以外の音声として ATR 音声データベースより、女性アナウンサー fsu の

- 「けっこうです」
- 「そうですか」
- 「もちろん発表のときも日本語でよろしいのですね」

について同様に、物理量変換ルールを適用し、刺激音声を作成し聴取実験を行なった。

6.4.1 単語

「けっこうです」の場合についての結果を図 6.2、表 6.3に示す。
「そうですか」の場合についての結果を図 6.3、表 6.4に示す。
図 6.2、表 6.3、図 6.3、表 6.4より、以下のことが分かる。

- サンプル番号 16 で、急激に悲哀の認識率が上がっている。これは Close Test と同じ傾向である。

表 6.3: 実験結果 - 「けっこうです」の認識率

サンプル	平静 (%)	歓喜 (%)	怒り (%)	悲哀 (%)	適用したルール
1	**100**	0	0	0	データベース上の音声
2	**90**	0	10	0	neutral1
3	40	**0**	60	0	joy1
4	30	**70**	0	0	joy1、2
5	80	0	**20**	0	anger1
6	50	30	**10**	10	anger1、2
7	50	10	**40**	0	anger1、2、3
8	20	30	**50**	0	anger1、2、3、4s
9	10	10	**80**	0	anger1、2、3、4l
10	0	0	**100**	0	anger1、2、3、4s、5
11	0	20	**80**	0	anger1、2、3、4l、5
12	0	10	**90**	0	anger1、2、3、4s、5、6
13	10	10	**80**	0	anger1、2、3、4l、5、6
14	90	10	0	**0**	sadness1
15	70	0	0	**30**	sadness1、2
16	10	0	20	**70**	sadness1、2、3
17	30	0	20	**50**	sadness1、2、3、4

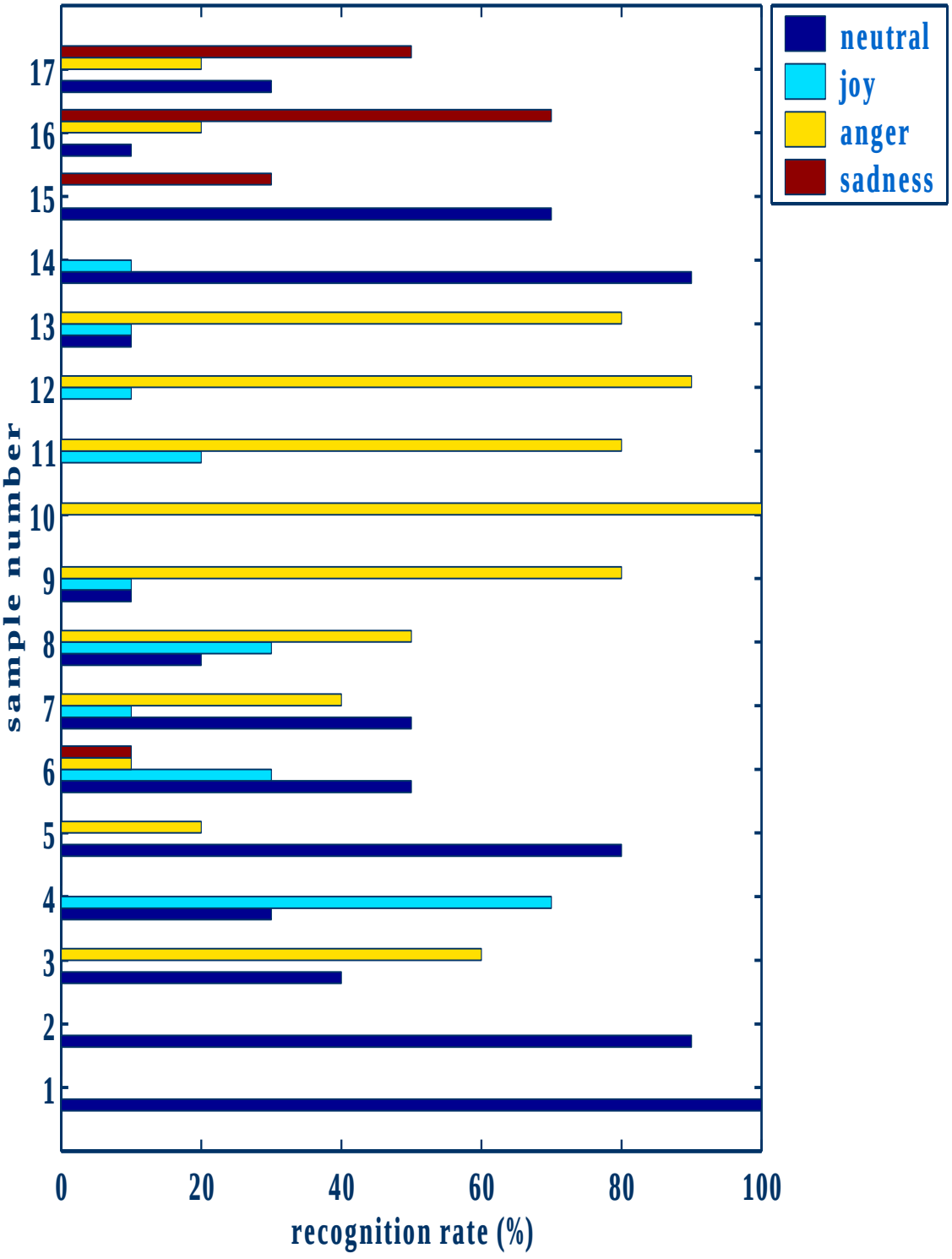

図 6.2: 実験結果 - 「けっこうです」

表 6.4: 実験結果 - 「そうですか」の認識率

サンプル	平静 (%)	歓喜 (%)	怒り (%)	悲哀 (%)	適用したルール
1	**86**	0	0	14	データベース上の音声
2	**50**	0	0	50	neutral1
3	18	**27**	55	0	joy1
4	0	**9**	91	0	joy1、2
5	82	0	**0**	18	anger1
6	41	27	**0**	32	anger1、2
7	32	45	**23**	0	anger1、2、3
8	27	23	**50**	0	anger1、2、3、4s
9	0	64	**36**	0	anger1、2、3、4l
10	36	23	**41**	0	anger1、2、3、4s、5
11	14	59	**27**	0	anger1、2、3、4l、5
12	18	27	**55**	0	anger1、2、3、4s、5、6
13	27	32	**36**	5	anger1、2、3、4l、5、6
14	68	0	5	**27**	sadness1
15	64	0	9	**27**	sadness1、2
16	0	0	5	**95**	sadness1、2、3
17	5	0	0	**95**	sadness1、2、3、4

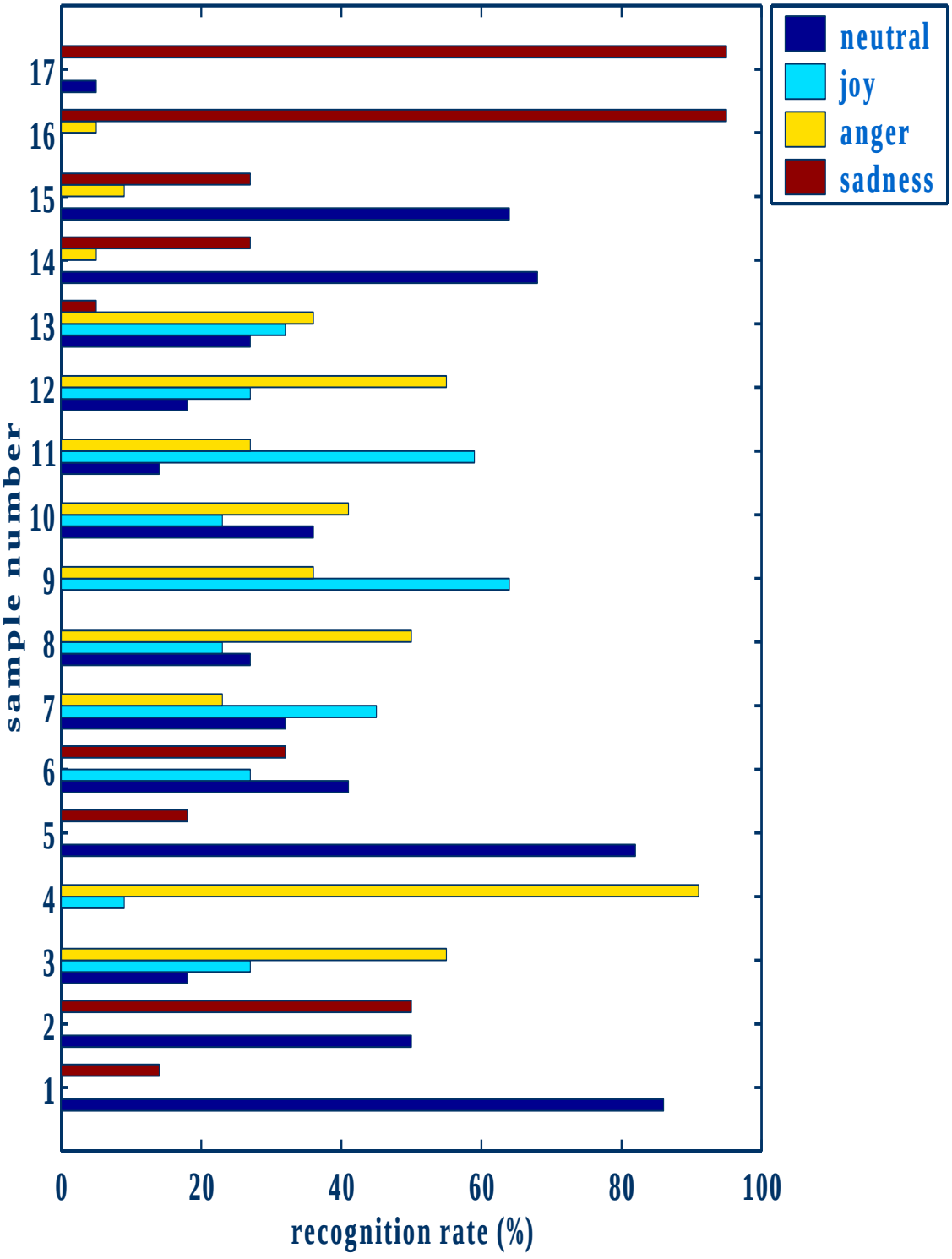

図 6.3: 実験結果 - 「そうですか」

- 全ての感情において高い認識率が得られているが、Open Test より若干低い。これは、日常生活の中で「けっこうです」という文章で明確な感情を感じることが殆んどないためである。また、「そうですか」の歓喜の感情制御をしたサンプルが、怒りと認識されているのは、語尾の基本周波数を上昇させることが、文章によっては嫌悪に聞こえるからだと考えられる [3]。さらに、怒りの感情制御をしたサンプルの認識率が低い要因として、基本周波数上昇に伴い合成音におけるスペクトル歪が生じ、音質が劣化したことが挙げられる。

6.4.2 文章

音声データを単語ではなく、文章でも実験する。物理量変換ルールは、文節毎に適用する。文節に分けると、以下のようになる。
もちろん / 発表の / ときも / 日本語で / よろしいのですね
「もちろん発表のときも日本語でよろしいのですね」の場合についての結果を図 6.4、表 6.5に示す。
図 6.4、表 6.5より、以下のことが分かる。

- 歓喜・悲哀については今までの実験の傾向とほぼ同じであり、怒りの認識率についても、認識率が低いながらもルールを適用するほどに高くなっている。これらより、今回作成したルールが、単語だけでなく文章にも通用することが分かる。

- 怒りの認識率が全体的に低い原因として、文脈に依ってしまったことが考えられる。

6.5 実験結果についての考察

Close Test では全ての感情において高い認識率が得られた。しかし Open Test では認識率が低い感情も存在した。これは言語情報に影響を受けたためだと考えられる。
それぞれの感情表現に用いたルールによる認識率の向上を図 6.5に示す。ここで図中の番号とルールは以下のようになっている。

1：(歓喜) 基本周波数の変化率を 2 倍に増加させる。

2：(歓喜) 語尾のみ基本周波数を上げる。

3：(怒り) パワーの長時間平均を 1.4 倍に増加させる。

表 6.5: 実験結果 - 「もちろん発表のときも日本語でよろしいのですね」の認識率

サンプル	平静 (%)	歓喜 (%)	怒り (%)	悲哀 (%)	適用したルール
1	**88**	0	0	12	データベース上の音声
2	**75**	0	0	25	neutral1
3	19	**54**	25	0	joy1
4	12	**76**	12	0	joy1、2
5	100	0	**0**	0	anger1
6	69	25	**0**	6	anger1、2
7	25	75	**0**	0	anger1、2、3
8	12	88	**0**	0	anger1、2、3、4s
9	63	31	**6**	0	anger1、2、3、4l
10	12	63	**25**	0	anger1、2、3、4s、5
11	31	38	**31**	0	anger1、2、3、4l、5
12	25	31	**44**	0	anger1、2、3、4s、5、6
13	63	12	**25**	0	anger1、2、3、4l、5、6
14	81	0	0	**19**	sadness1
15	50	0	6	**44**	sadness1、2
16	0	0	0	**100**	sadness1、2、3
17	0	0	0	**100**	sadness1、2、3、4

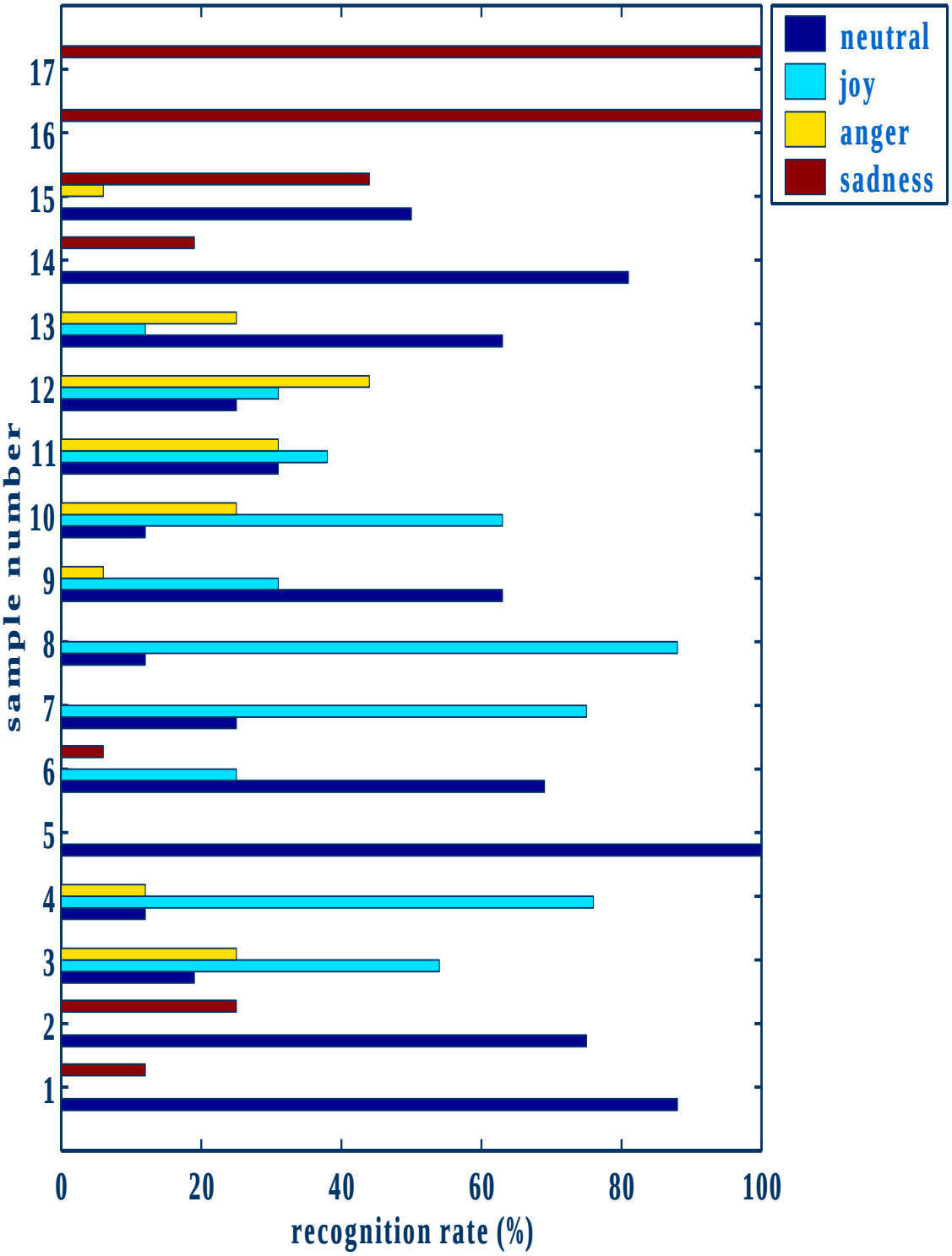

図 6.4: 実験結果 - 「もちろん発表のときも日本語でよろしいのですね」

4：(怒り) 基本周波数の長時間平均を 1.2 倍に増加させる。

5：(怒り) 基本周波数の変化率を 1.4 倍に増加させる。

6：(怒り) 発話時間を 0.8 倍に短縮させる (子音部伸縮なし)。

7：(怒り) 発話時間を 1.2 倍に伸長させる (子音部伸縮なし)。

8：(怒り) ホルマント周波数をシフトさせる。

9：(怒り) 高域スペクトル成分を 5dB 増加させる。

10：(悲哀) パワーの長時間平均を 0.8 倍に減少させる。

11：(悲哀) 基本周波数の長時間平均を 0.9 倍に減少させる。

12：(悲哀) 基本周波数の変化率を 0.5 倍に減少させる。

13：(悲哀) 発話時間を 1.5 倍に伸長させる (子音部伸縮なし)。

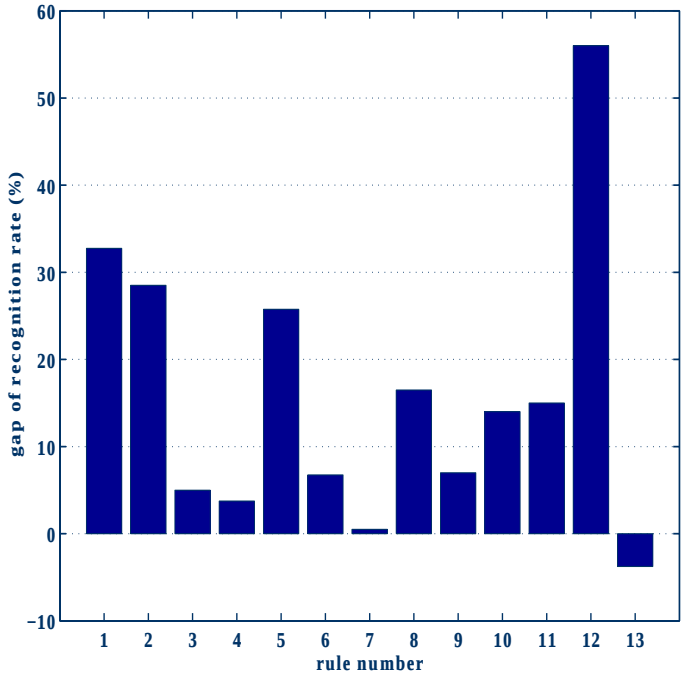

図 6.5: 物理量変換ルールによる認識率の向上

図 6.5より、以下のことが分かる。

- 歓喜の感情表現において、基本周波数の変化率の増加、語尾の基本周波数上昇は 2 つとも重要、有効なルールである。
- 怒りの感情表現において、基本周波数の変化率の増加、ホルマント周波数のシフトが特に有効であった。ホルマント周波数のシフト、高域スペクトル成分の増加によって認識率は確実に向上しており、怒りの感情制御において Lombard Efeect のモデル化が有効であったことが分かる。逆に発話時間の伸縮は、あまり重要ではない。
- 悲哀の感情表現において、基本周波数の変化率の減少が特に有効であった。発話時間の伸長によって認識率が低下しているのは、ルール適用以前に既にかなり高い認識率を獲得しているからである。

第 7 章

全体の考察

本研究は、音声中の感情表現に関連する物理量として、発話速度・基本周波数・パワーを扱った。

さらに Lombard Effect の利用によってパワーと基本周波数・スペクトルを関連付け、分析結果との組合わせにより、感情制御のための物理量変換ルールを構築した。

本章では研究で得られた結果の考察を行なう。

感情と物理量の関係

平静音声と、歓喜・怒り・悲哀の音声との間の物理量との関係は、矛盾点も皆無というわけではないが、過去の研究とほぼ合致した。

本研究の分析結果と、過去の研究の分析結果との違いは、怒りの感情表現における発話時間のみである。

Open Test では、幾らか認識率が低下する感情もあったが、多くの共通項を見いだすことができた。

各感情表現において、基本周波数の変化率が強く関わっており、特に悲哀については、重要な役割を担っていることが分かった。また、怒りと歓喜とは、基本周波数の構造が似ているため、これらは言語情報に影響を受けやすい。

怒りの感情制御において、ホルマント周波数のシフト、高域スペクトル成分の増加が有効であることが分かった。

本研究で扱ったサンプルでは、発話時間の伸縮は怒りの感情の認識率に影響を及ぼさなかった。

物理量同士の関係

分析により、パワーと基本周波数との線形的な関係が発見された。

Lombard Effect を利用し、パワーの増大によるホルマント周波数のシフトおよび高域スペクトルの増加を物理量変換ルールに組み入れた。これらは怒りの感情制御において重要であり、認識率の向上が得られた。

第 8 章

結論

8.1　本論文で明らかにされたことの要約

本研究では音声中の物理量の変化が感情識別に与える影響を調べた。その結果、過去の研究結果とほぼ合致した。調べた結果と、音声の生成機構に起因する物理量間の関係より、感情制御のための制御ルールを構築した。その中で、パワーと、基本周波数・ホルマントのシフト・高域スペクトルの増加を結び付けた。平静音声を、制御ルールによって感情制御し、感情知覚への影響を調べた。

その結果、各感情表現において、基本周波数の変化率が関わっており、特に悲哀については、重要な役割を担っていることが分かった。怒りの感情表現に対し、Lombard Effect を利用して、物理量同士を関連づけたところ、認識率の向上が得られた。

8.2　今後の課題

感情のとらえ方の個人差

怒りにおける発話時間において、本研究は過去の研究と異なる結果が出た。

これは、怒りの感情表現にて発話時間の伸縮が重要ではないという考え方もできるが、怒りという感情の中でも、Hot Anger と Cold Anger と呼ばれるものがあり、物理量の変化の仕方が違うためだとも考えられ、それらを分離する必要がある。また、本研究では感情を平静・歓喜・怒り・悲哀だけに絞ったため、感情のとらえ方の個人差が大きく影響する。

文脈依存性の問題

今回の実験では、Close Test では全ての感情において高い認識率が得られたが、Open Test に対しては感情によって認識率が低下した。

これは言語情報が感情に対して影響を及ぼしているからだと考えられる。本研究で構築した制御ルールにも言語情報に依るものがあるように思われるため、様々な文章で分析をしていかなければならない。

謝辞

日頃御指導いただき、貴重な御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] 前川喜久雄: “音声によるパラ言語情報の伝達:言語学の立場から”, 音学講論, 1-3-10, pp.381-384, Sep.1997.

[2] 北原義典, 東倉洋一: “音声の韻律情報と感情表現”, 信学技報, SP88-158, Mar.1989.

[3] 平賀裕, 斎藤善行, 森島繁生, 原島博: “音声に含まれる感情情報抽出の一検討”, 信学技報, HC93-66, pp.1-8, Jan.1994.

[4] 林康子: “感動詞「ええ」におけるピッチ曲線と感情認知”, 信学技報, H98-61, Jul.1998.

[5] T.Moriyama, H.Saito and S.Ozawa: “Evaluation of the relationship between emotional concepts and emotional parameters on speech”, Proc.of ICASSP, Vol.2, pp.1431-1434, Apr.1997.

[6] J.E.H.Noad, S.P.Whiteside and P.D. Green: “A MACROSCOPIC ANALYSYS OF AN EMOTIONAL SPEECH CORPUS”, Proc.of Eurospeech, pp.517-520, 1997.

[7] 小池和仁, 斎藤博昭, 中西正和: “感情音声の合成”, 信学技報, SP98-107, Dec.1998.

[8] 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, 1-2-1, pp.189-192, Sep.1997.

[9] 吉田勝, 小畑秀文: “ロンバート効果を考慮した低品質単語認識の一手法”, 音学講論, 1-Q-21, pp.183-184, Oct.1994.

[10] D.B.Pisoni, R.H.Bernacki, H.C.Nusbaum and M.Yuchtman: “Some acoustic-phonetic correlates of speech produced in noise”, Proc.of ICASSP, pp.1581-1584, Apr.1986.

[11] B.J.Stanton: “Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions”, Proc.of ICASSP, pp.331-334, Apr.1988.