修士学位論文

音声による感情認識システムに関する研究

東北大学大学院工学研究科
電気・通信工学専攻　阿曽研究室

門谷 信愛希

目 次

第 1 章　序論 **1**

1.1　研究の背景 1

1.1.1　ディクテーションシステムの課題 1

1.1.2　エージェントシステムと音声インターフェース 3

1.2　研究の目的 4

1.3　論文の構成 5

第 2 章　感情の連続音声認識システムへの影響 **6**

2.1　はじめに 6

2.2　感情の分類 6

2.3　感情音声データベースの構築 7

2.4　実験の内容 9

2.5　実験条件 9

2.6　実験結果 10

2.6.1　音素正解精度 10

2.6.2　誤認識の分布 12

2.7　まとめ 13

第 3 章　音声による感情の判別 **16**

3.1　はじめに 16

3.2　感情の推定方法 16

3.3　判別分析とは 17

3.4　特徴量の選定 18

3.5　特徴量の抽出 18

3.5.1　基本周波数の抽出 18

3.5.2　その他の特徴量の抽出 20

3.6　実験内容 21

3.7　実験結果 21

3.7.1　話者ごとの場合 21

3.7.2 話者混合データの場合 21
3.7.3 従来手法との比較 26
3.8 聴取実験による評価 27
3.8.1 実験の方法 27
3.8.2 実験結果 28
3.8.3 考察 28
3.9 まとめ 29

第 4 章 感情判別に有効な特徴量 30
4.1 はじめに 30
4.2 特徴量の寄与割合 30
4.2.1 話者ごとの環境 31
4.2.2 話者混合環境 31
4.3 性別差補正の導入 32
4.3.1 性別による特徴量の相違 32
4.3.2 補正による判別率の向上 36
4.4 まとめ 36

第 5 章 結論 37
5.1 本研究の成果 37
5.2 今後の課題 37

謝辞 39

参考文献 40

研究業績 42

図 目 次

1.1 連続音声認識システムの概要 2
1.2 話者の心理状態に対応できるエージェント 3

2.1 大語彙連続音声認識エンジン JULIUS の構成 10
2.2 平均音素正解精度 (男性話者) 11
2.3 平均音素正解精度 (女性話者) 11

3.1 基本周波数軌跡と各パラメータの関係 19
3.2 基本周波数抽出の流れ 19
3.3 話者ごとの判別率 (男性話者) 22
3.4 話者ごとの判別率 (女性話者) 22
3.5 話者ごとの判別率の平均値 23
3.6 混合データの判別率 23
3.7 サンプルの分類個数と割合 (男性話者) 24
3.8 サンプルの分類個数と割合 (女性話者および全話者) 25
3.9 被験者ごとの評価値の分布割合 (話者 1) 29

4.1 話者ごとの標準化判別係数 (同程度型) 33
4.2 話者ごとの標準化判別係数 (多種突出型) 34
4.3 話者混合環境の標準化判別係数 35

表 目 次

2.1 会話文の内容 7
2.2 各音素の使用数 8
2.3 会話文「山田君，そこのドア閉めてくれる?」の想定場面例 8
2.4 データベースの録音環境 8
2.5 連続音声認識の実験条件 10
2.6 感情ごとの誤認識変動率上位 5 音素 12
2.7 感情ごとの誤認識変動率下位 5 音素 12
2.8 感情ごとの誤認識変動率 (男性話者) 14
2.9 感情ごとの誤認識変動率 (女性話者) 15

3.1 基本周波数の抽出パラメータ 20
3.2 従来手法との比較 26
3.3 聴取実験の方法 27
3.4 各話者への被験者の割り当て 27
3.5 話者ごとの相関係数 28

4.1 特徴量の平均値の性別による相違 32
4.2 補正前と補正後の判別率の比較 36

第1章

序論

1.1 研究の背景

近年の計算機性能の向上に伴い，これまでごく限定された用途に用いられていた音声認識技術が一般に浸透してきている．音声は元来人間のコミュニケーションの最も基本的な手段であるため，人間と計算機とのインターフェースとしての音声への期待は，計算機が発明された当初から存在した． もし音声をインターフェースとして用いることができるならば，初心者や高齢者には使い勝手の悪いキーボードやマウスなどの入力デバイスに頼らなくても計算機を操作できるようになる可能性がある．また，秘書ロボット等のエージェントシステムにおいては，人間対人間の間で行われるような，自然なコミュニケーションの能力を備えることが要求される．このような「人間らしさを備える」という観点から考えた場合，音声はエージェントシステムのインターフェースとして必須と言えるほど重要であると考えられる．

1950 年代からこれらの要求に応えるべく精力的に研究が行われた結果 [1]，現在では音声の文章への書き起こし (ディクテーション) を目的とするシステムで見れば，家庭向けの比較的廉価な計算機環境においても，実時間で実用に耐えうる精度を得ることが可能なシステム [5] が提供されるようになってきた．これらの聞き取りシステムの認識部を利用した計算機操作用のソフトウェアも開発されるなど，音声認識システムは着実に進歩している．

1.1.1 ディクテーションシステムの課題

ここで，ディクテーションを目的とした連続音声認識エンジンの概要を述べる．現在の連続音声認識は HMM(Hidden Markov Model) などの統計的な確率モデルを利用して認識を行う方法 [2] が主流である．処理の流れを図 1.1 に示す．入力された音声は，そのままで扱うことはデータ量の点から問題があるため，特徴量とよばれる低次のベクトル系列に変

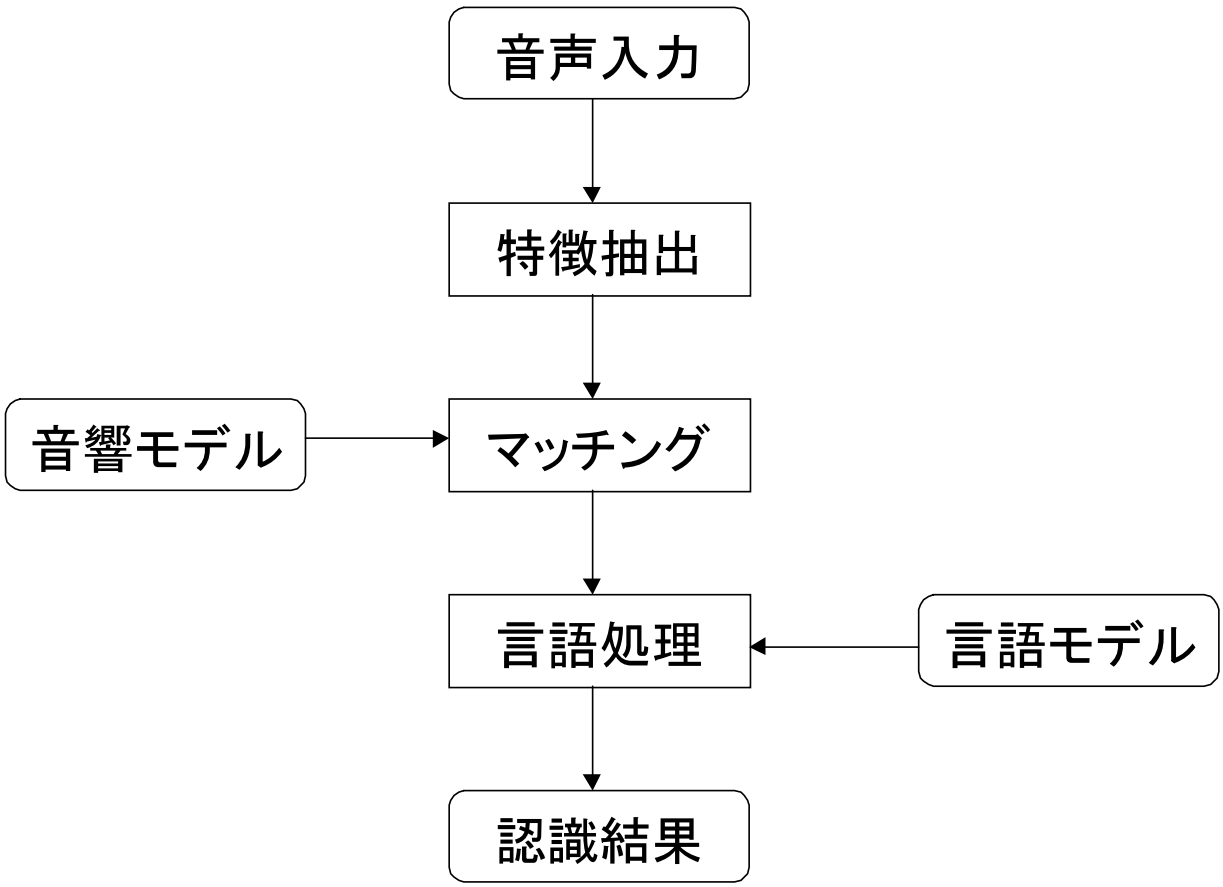


図 1.1: 連続音声認識システムの概要

換される．変換された特徴量は，大量の音声データを元に学習した音響モデルとマッチングされ，第 1 段階のスコアが出力される．音響モデルの構築には，はあらかじめ背景雑音の極めて少ない無響室等の理想的な環境でアナウンサーが発話した音声が用いられる．通常音響モデルは単語や文章といった長い単位ではなく，音響的言語的に意味のある最小の単位である音素ごとに作成される．音素の種類は分類の方法によってその数が異なるが，およそ 30 ~ 40 種類に分類されるのが普通である．次にマッチングによりスコアリングされた音素の系列を，言語モデルを用いて再評価する．言語モデルはデータ量と入手の容易さから，書き言葉である新聞記事データベース等を元に作成されることが多い．言語モデルは単語辞書と，隣接する単語の共起確率を数万単語についてモデル化したものからなり，後者には該当単語の一つ手前の単語を見る bi-gram と，二つ手前までを見る tri-gram がある．bi-gram を粗い言語モデル，tri-gram をより詳細な言語モデルとして，候補数を絞り込みながら二段階の評価を行うことで高速化を図るシステムも存在する．最終的に最もスコアの高い系列が認識結果として出力される．

このように現在の連続音声認識は，話者が一定の声質でしかも文章表現で用いられるような書き言葉調の音声が入力された場合に最適な出力を得られるようになっている．しか

しながら，実際の発話においては話者の音声は発話内容以外の要因に起因するさまざまな「ゆらぎ」が存在する．例えば話し言葉特有のくだけた表現，心理状態の変化に伴う声の強弱や周波数の変動などである．このような事前に学習した音声とは声質，強さなどが変化した音声が入力された場合，ディクテーションシステムの認識結果にどのような影響が出るのかはこれまであまり考慮されていなかったのが現状である．

1.1.2　エージェントシステムと音声インターフェース

現在の社会で用いられている音声を利用したシステムは，機器類の操作をガイドするもの，つまり機器からユーザへの情報伝達のみを行うものが主流である．例えば銀行の ATM や電車の切符販売機，電話によるチケット予約システムなどがこれにあたる．

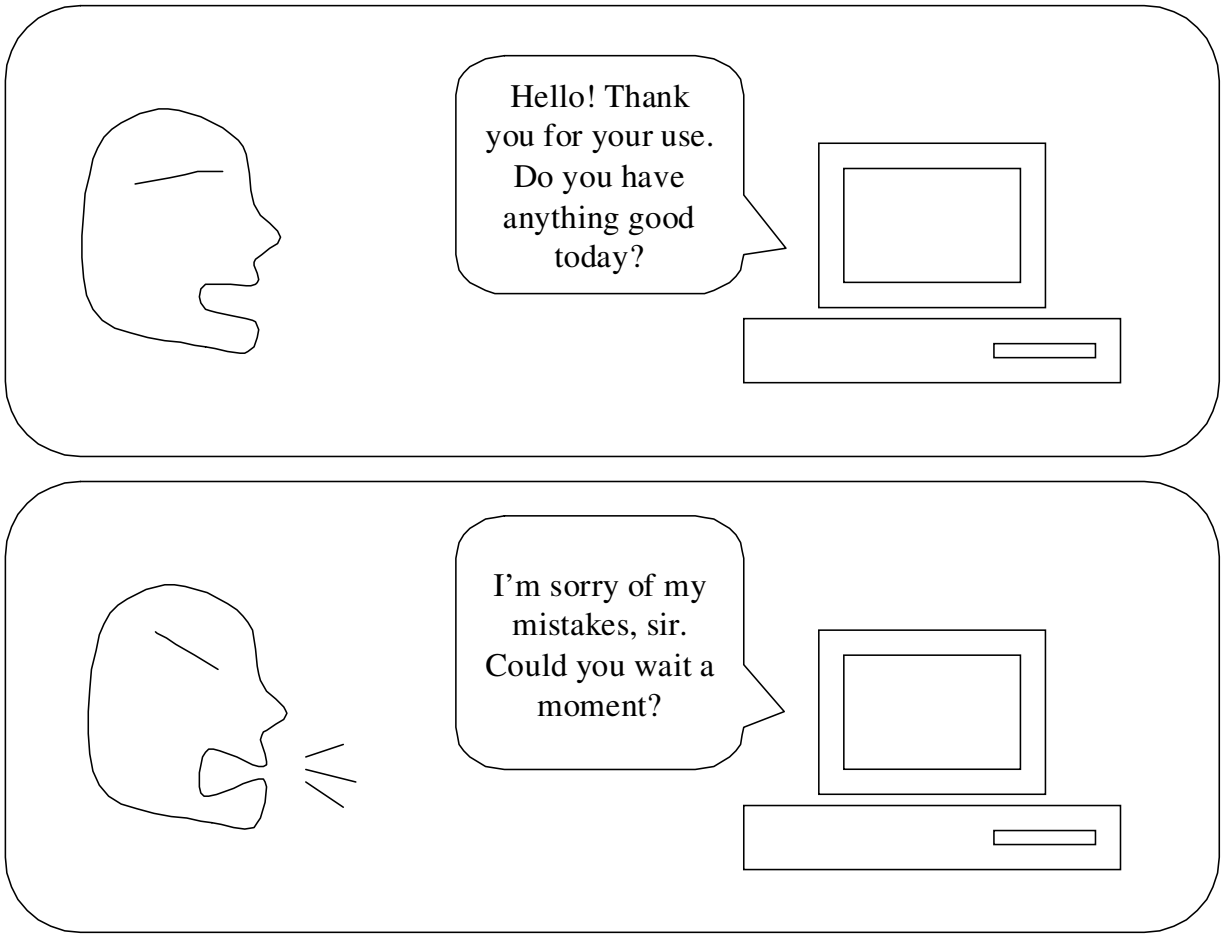


図 1.2: 話者の心理状態に対応できるエージェント

これらのシステムでは，情報を入力するためのデバイスは押しボタンなど別に存在しており，音声は情報伝達を確実にするための補助的機構としてしか用いられていない．これに対して，今後はユーザと機器とが音声によって双方向の情報伝達を行う機能を備えたシ

ステムに対するニーズが高まると予想される．本論文ではこの機能を持つシステムをエージェントシステムと呼ぶ．エージェントは現在人間が行っている業務を肩代わりしたり，省力化することをその目的としていて，例えば介護ロボットや秘書ロボット，情報検索エージェント，カーナビゲーションシステムなどが考えられる．これらのシステムで利用される音声インターフェースには，不特定の話者に対応することや，公共の場所など背景雑音の多い環境に対する頑健性を持つことなども重要であるが「そのシステムが人間にどれだけ近いか」という点もまた極めて重要である．図1.2に例を示した．この例では，システムは話者の音声に気楽さや陽気さが感じられる場合と，怒気を含む場合とで最初の応答メッセージを変化させている．人間対人間のコミュニケーションにおいてはこのような適応処理は暗黙のうちに行われているので，エージェントがこの暗黙の了解を破るような応答をした場合に，情報伝達が不確実になるという問題が発生する．従って，ユーザの心理を推察して対応を変化させる柔軟性は，ユーザの使用感を向上させるために重要であると考えられる．

1.2　研究の目的

現在一般的に用いられている連続音声認識システムは，ゆらぎの少ない音声を対象に設計されているため，話者の心理状態の変化に起因する音声の変動に頑健でない可能性があることを前節で述べた．もしも無視できない影響があるならば，感情に応じて音響モデルを切り替えるなどの対策を行う必要がある．これらの影響を検証するためには，感情を含んだ音声によって認識実験を行うのが最も重要であると考えられるが，感情を含んだ音声データは収集が困難であり，一般に公開されている日本語のデータベースは存在していない．そのため，本研究ではまず感情音声を含んだデータベースを作成することから始める．次に，これらの音声を用いて連続音声認識システムに与える影響を認識実験により検証，考察する．

また，より人間らしさを持つエージェントシステムを構成するために重要な，音声を用いてユーザの感情推定を行う方法について検討する．感情推定をある程度の確度で行うことが出来れば，連続音声認識システムの音響モデル切り替え機構として利用することが可能であるし，エージェントシステムが対応を変更する際の判断基準を与えることができる．

この音声による感情推定には，発話内容(コンテキスト)から得られる情報を用いる方法，つまり実際に音声認識を行ってその結果を利用する方法と，韻律や振幅，強勢などの信号処理的方法によって行う方法とが考えられるが，前者は連続音声認識システムの前段としての用途には計算量の面からみて現実的ではない．従って，本研究では後者の方法により感情音声から感情評価パラメータを抽出して感情推定を行うこととする．また，推定にはどのようなパラメータが有効であるのかを感情判別実験により明らかにし，考察する．

1.3　論文の構成

本論文の構成は以下の通りである．

第 1 章 序論であり，研究の背景及び目的を述べた．

第 2 章 判別すべき感情の分類および構築する感情音声データベースについて述べる．次に，感情を含んだ音声が既存の大語彙連続音声認識エンジンに与える影響を調査，分析する．

第 3 章 感情を判別するための特徴量を提案し，判別実験と評価，分析を行う．また，従来手法との比較により提案手法の有効性を検証する．

第 4 章 3 章で得られた結果をもとに，感情の違いが特徴量のどの部分に反映されるかをさらに詳しく分析する．　また，判別率向上のための特徴量の補正方法について述べる．

第 5 章 結論である．

第2章

感情の連続音声認識システムへの影響

2.1 はじめに

序論では，既存の連続音声認識エンジンが話者の音声にゆらぎが少なく，平静の状態であることを想定して設計してあるため，感情が込められた音声に対する頑健性に疑問があることを述べた．そこで本章では実際に感情を含んだ音声を連続音声認識エンジンに入力して影響を調査する．まず，計算機上において感情をどのように分類し取り扱うかについて述べる．次に，一般に利用できる感情を含んだデータベースが存在しない現状を踏まえ，データベースの構築作業について述べる．そして，これらの感情を含んだ音声データを用いて連続音声認識実験を行い，感情が与える影響について調査する．最後に認識結果の分析により，誤認識の特徴を調査する．

2.2 感情の分類

計算機が感情を推定するには，そもそも人間がどのように感情表現を行うかを知る必要があり，これまでは主に心理学者らによって心理学的側面からさまざまな研究が行われてきた．これまでの計算機による感情推定を試みた研究おいても，認識するべき感情をどのように取り扱うかは非常に大きな問題であり，研究者によって対応が千差万別であるのが現状である．例えば，感情を含む言葉 46 種を主観的に抽出した列挙型の研究 [14] や，怒り，驚き，悲しみ，嫌悪，喜びなどの数感情にとどめる少数分類型の研究 [11][12][13][17]，また感情が「快-不快」「興奮-沈静」「緊張-弛緩」の 3 次元空間を形成するとして，単なる分類ではなく感情の強弱を得ようとした研究 [15] などがあり，取り扱いには大きな違いがある．

エージェントにより人間に近い能力を持たせるためには，強弱を伴う形で認識するのが好ましいと思われるが，そもそも人間がユーザの感情の程度を知って，どのように対応を

変化させるかを手順化することは非常に困難であると考えられ，現状ではエージェントの対応するべき感情は数通り程度にするのが現実的である.

これらのことから，本研究においては最も基本的な感情の分類であると考えた「怒り」，「悲しみ」，「喜び」，「標準」の 4 種類の感情を取り扱うことにする.

2.3　感情音声データベースの構築

2.2 節で決定した感情の分類に基づいて，感情を含んだ音声データを採取した.　音声データの内容は，特定の感情の影響を比較的受けにくいと思われる会話文 20 文を主観的に選定して構成した. 会話文の内容を表 2.1 に示す. また，これらの文章で使われている音素の分布を表 2.2 に示す.

表 2.1: 会話文の内容

No	内容
1	電車，いつ来るのかなあ
2	山田君，そこのドア閉めてくれる?
3	今日は，8 時に会社にいかないと
4	机の上を，片づけなさい
5	私は別に，問題はありませんよ
6	明日の予定，あいてます?
7	そんな自分勝手なこと，ダメだよ
8	マイクの音量を，上げて下さい
9	まな板を，洗っておいて
10	車の調子が，悪いんだ
11	今日は，雨が降ったね
12	勉強もっとしなさい
13	この色で，塗っていいのかい
14	あの時計，3 分進んでいるよ
15	昨日，みんなでカラオケに行ったの?
16	そんなに心配しなくても，大丈夫だよ
17	後で，手紙を書いておきます
18	辞書で調べれば，わかるよ
19	電話がくるのを，待っています
20	最近，天気がよく変わるね

表 2.2: 各音素の使用数

a	i	o	e	u	n	k	t	sp	N	m
78	50	44	37	34	30	30	24	20	18	17
d	r	s	sh	y	w	q	o:	g	b	j
16	15	14	11	8	8	7	6	6	6	4
ts	ky	p	ch	z	ry	h	f	a:		
3	3	2	2	1	1	1	1	1		

これらの文章を番号順に披験者に提示し，それぞれの話者が最もよくその感情を表現していると思われる方法で発話してもらった. また表 2.3 に示すような感情ごとに想定される場面の一例を示し，話者が感情移入しやすいようにした. また録音前の練習は自由に行ってよいこととした.

表 2.3: 会話文「山田君，そこのドア閉めてくれる?」の想定場面例

感情	想定される場面
怒り	「虫の居所の悪い上司のつもりで」
悲しみ	「憂鬱な会社に 8 時に行くなんて … 」
喜び	「仲の良い友達の家での会話」

表 2.4: データベースの録音環境

使用機器	DAT + ヘッドセット (独 Senheizer 社 HMD-410)
サンプリング周波数	16KHz，16bit/sample (DAT-Link によりダウンサンプリング)
録音環境	一般木造家屋 (絨毯張り)， 晴天時録音
話者数	13 名 (男性 7 名，女性 6 名)
データ数	1040 サンプル (1 感情 20 文・4 感情・13 名=1040)

録音時の環境を表 2.4 に示す. これまでの音声による感情認識の研究では，感情を含んだ大規模な音声データベースが存在しないことが大きな障害であった. 研究者は基本的に感情を含んだデータを単独で収集する必要があり，感情を含んだ音声データはせいぜい数名程度の話者による数文章，全体で数十文程度であることが多く，量的な面に疑問の余地を残すことがあった. 本研究においては，そのような問題を回避すべく広範にデータ収集を

行った結果，この分野では比較的大規模である 13 名の話者からなる全 1040 サンプルを採取することができた.

2.4　実験の内容

2.3 節のデータを用いて，感情ごとにどのような影響があるかを分析する. まず，話者ごとに連続音声認識を行い，感情によってどのような影響が現れるかを評価する.　評価は 1.1.1 節で述べた音声認識における最小の認識単位である音素を用いて行う.　式 (2.1) がその評価尺度である音素正解精度 (phoneme accuracy) である. 音素正解精度は音素の挿入，置換，脱落誤りを考慮した評価尺度で，音素認識率の評価尺度として一般的に用いられているものである.

$$phoneme\ accuracy = \frac{N - R - D - I}{N} \cdot 100 \quad [\%] \tag{2.1}$$

但し N は文中の全音素数で，R，D，I は順に置換，脱落，挿入誤りを示すものとする. 各誤り数の把握は入力文章の音素表記と連続音声認識エンジンから出力される音素系列を DP マッチングすることで行う. このようにして得られた結果によって感情ごとの影響度合いを調べる. また，誤認識がどのような音素に分布しているかの傾向も調査する.

2.5　実験条件

使用した連続音声認識エンジンは IPA の援助により京都大学にて開発された JULIUS[6][7][8] である. JULIUS を用いたのは，既存の読み上げ音声コーパスを用いた評価実験で 90%以上の認識率を得ることができる点 [9] と，ソースコードも含め全体が無償で提供されている点に注目したためである. システムの構成を図 2.1 に示す. JULIUS は 2 パスの認識器であり，第 1 パスで簡易な言語モデルである bi-gram を用いて候補を絞り込み，第 2 パスでより詳細な tri-gram を用いることで，認識の高精度化を図っている. 実験時のパラメータを表 2.5 に示す. 音響モデルは男性話者と女性話者で切り替えた. これ以外のオプションはすべてデフォルト値を採用した.

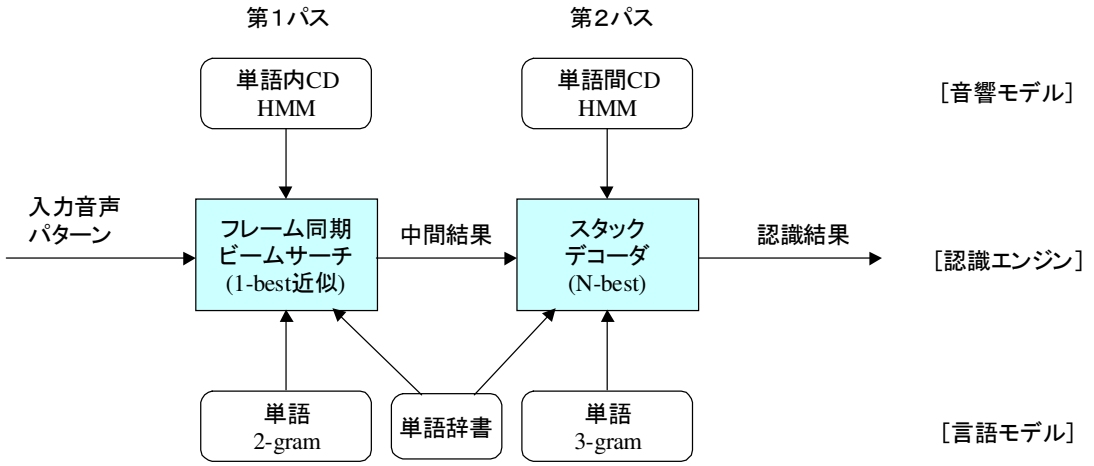


図 2.1: 大語彙連続音声認識エンジン JULIUS の構成

表 2.5: 連続音声認識の実験条件

使用システム	大語彙連続音声認識エンジン JULIUS-3.1p2
データ形式	MFCC(12 次元)
音響モデル	性別依存 triphone-HMM (ガウス分布混合数:16，状態打ち切り数:s3000)
言語モデル	20K 単語辞書+tri-gram (未知語率 1.32%)

2.6　実験結果

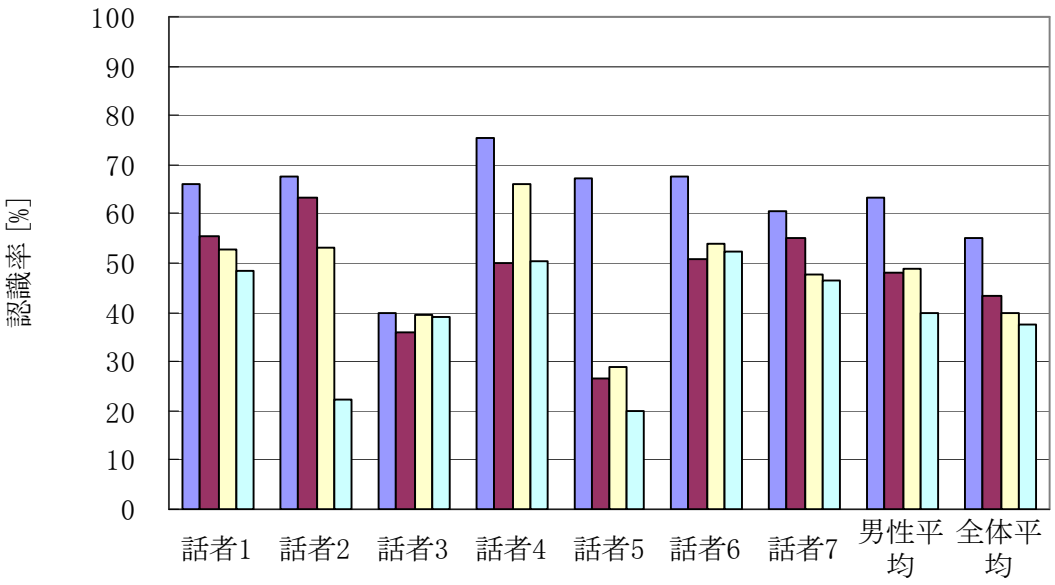
2.6.1　音素正解精度

男性話者の認識率を総合した結果を図 2.2 に，女性話者の認識率を図 2.3 に示す. 値はいずれも 20 文の平均値である.「全体平均」は全話者の平均認識率を表す.

男性話者では 60 ~ 70%程度の「標準」の認識率が得られているのに対し，その他の感情ではいずれも 10%から 20%，最大で 50%近い認識率の低下が見られた. 特に「喜び」の影響が大きいことがわかる. 男性話者の平均では標準 → 怒り → 悲しみ → 喜びの順で影響度が増している.

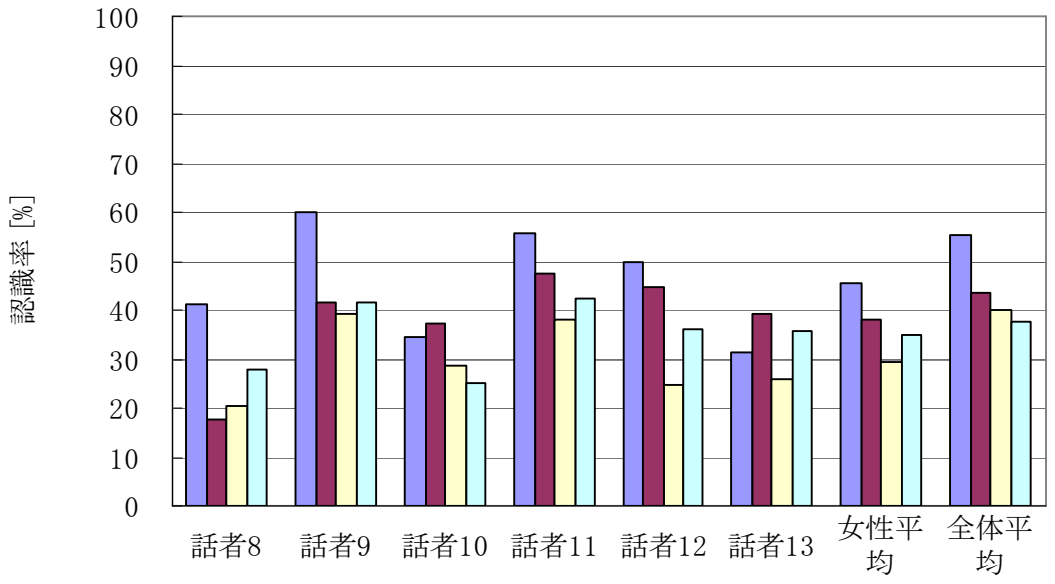
一方，女性話者では 50 ~ 60%程度の精度が得られているが，感情ごとの影響度合では，「悲しみ」の方が「喜び」よりも大きいことが判明した. 平均でも標準 → 怒り → 喜び → 悲しみの順に精度が低下している. これらの結果は，男女の感情表現に差があることを示していると考えられるが，一方で影響度合いは話者によって著しく異なっている.

そこで，感情によって有意差を伴って認識率への影響が生じるかどうかを検定によって



	話者1	話者2	話者3	話者4	話者5	話者6	話者7	男性平均	全体平均
標準	65.9	67.6	39.9	75.3	67.3	67.5	60.4	63.4	55.1
怒り	55.5	63.2	35.7	50.1	26.4	50.7	54.9	48.1	43.4
悲しみ	52.7	53.1	39.5	66.2	29.0	53.8	47.7	48.9	39.9
喜び	48.5	22.3	39.2	50.6	20.0	52.2	46.6	39.9	37.5

図 2.2: 平均音素正解精度 (男性話者)



	話者8	話者9	話者10	話者11	話者12	話者13	女性平均	全体平均
標準	41.2	60.0	34.5	55.7	49.9	31.3	45.4	55.1
怒り	17.6	41.6	37.1	47.5	44.7	39.4	38.0	43.4
悲しみ	20.5	39.2	28.6	37.9	24.8	25.9	29.5	39.9
喜び	27.8	41.6	25.3	42.5	36.1	35.6	34.8	37.5

図 2.3: 平均音素正解精度 (女性話者)

検証した. 図 2.2，図 2.3 のデータについて，感情間に平均認識率の有意差を仮定した検定を話者ごとに行った. その結果，有意水準 5%では話者 3 で仮定が棄却されたが，それ以外の 12 名の話者については仮定が採択された. このことから，平静音声に感情が含まれることで，有意差を伴う認識率の低下を起こすことが確認された. 従って，感情を含む音声に対して頑健な連続音声認識システムを構成するには，認識の前段階で感情推定によって音響モデルを切り替えるなどの対策を行う必要があることが分かった.

2.6.2　誤認識の分布

次に音声が感情に含まれることによって，それぞれの音素にどのような影響があるかを調べた. 2.6.1 節において行った実験で得られた文章ごとの認識誤りを分析し，音素ごとに式 (2.2) で表される変動率を定義し，これにより評価を行った. 変動率が大きいほど「標準」の感情から誤りが増えており，影響が大きいと見なすことができる.

$$\text{感情 } X \text{ の変動率} = \frac{\text{感情 } X \text{ における認識誤り数}}{\text{感情「標準」における認識誤り数}} \cdot 100 \quad [\%] \tag{2.2}$$

認識誤り数は挿入，置換，脱落の各誤りを合計したものであり，サンプル中に 10 回以上出現した音素のみを対象として調査した. また同率の場合は出現回数の多い音素を上位とした. 実際に変動率の値が大きい順に 1 位から 5 位までの音素を 1040 サンプル全体に対して調べたものが表 2.6 である. 同様に小さい順に調べた結果を表 2.7 に示す.

表 2.6: 感情ごとの誤認識変動率上位 5 音素

感情	1 位	2 位	3 位	4 位	5 位
怒り	sh 176%	N 155%	s 153%	k 150%	j 146%
悲しみ	sh 173%	k 168%	N 161%	r 148%	i 148%
喜び	k 176%	sh 173%	s 169%	o: 167%	N 165%

表 2.7: 感情ごとの誤認識変動率下位 5 音素

感情	1 位	2 位	3 位	4 位	5 位
怒り	ky 106%	g 112%	ts 113%	b 115%	r 115%
悲しみ	ky 94%	t 119%	b 123%	o: 127%	w 127%
喜び	b 176%	d 173%	g 169%	t 167%	e 133%

影響の大きい音素は，感情によって若干の違いがあるものの，摩擦音である sh，破裂音の中でも口蓋の動きによる k，鼻音である N の 3 つの音素が目立っている．このことは事前の「母音区間に感情表現が最も強く現れるのではないか」という予想とは異なったものとなった．　影響の小さいものでは，ky，b，t などが目立っている．

また，変動率を話者ごとに分析した結果を，男性話者については表 2.8，女性話者については表 2.9 に示す．　話者レベルにおいても，表 2.6 の全体の場合とほぼ同じような傾向が見られている．ほとんどの話者において，最も変動が大きい音素に N，sh，k，j などがあり，これらの音素が感情の変動を受けやすいことが判明した．

2.7　まとめ

本章ではまず，「標準」「怒り」「悲しみ」「喜び」の 4 つの感情からなる会話文のデータベースを構築した．次に音声に感情が付加されることによって連続音声認識エンジンにどのような影響があるのかを認識実験によって検証した．話者ごとに認識率の有意差を仮定した検定を行った結果，話者 3 以外のすべての話者で仮定が採択され，感情が付加されたことによって認識に影響を及ぼすことを確認した．また話者別では最大で怒りは 25.2%，悲しみは 25.1%，喜びは 47.3%の認識率の低下をもたらすことが分かった．また全体の平均では喜び (17.6%低下)，悲しみ (15.2%低下)，怒り (11.7%低下)，の順で影響度が大きいことが分かった．

音素ごとの誤り分布では sh，k，N の各音素が感情による影響が大きく，k，b，t の音素では影響が小さいことが分かった．

表 2.8: 感情ごとの誤認識変動率 (男性話者)

話者	感情	1 位	2 位	3 位	4 位	5 位
話者 1	怒り	s 300%	sh 300%	n 188%	u 167%	d 167%
	悲しみ	sh 300%	s 250%	r 200%	n 188%	u 178%
	喜び	sh 500%	s 350%	r 300%	o: 300%	n 238%
話者 2	怒り	s 233%	n 200%	w 200%	j 200%	m 140%
	悲しみ	s 300%	j 300%	w 250%	n 240%	u 211%
	喜び	n 500%	i 410%	w 400%	j 400%	q 300%
話者 3	怒り	k 155%	g 150%	j 150%	y 133%	q 133%
	悲しみ	k 155%	o: 150%	g 150%	j 250%	y 133%
	喜び	o: 150%	j 150%	g 125%	s 122%	y 117%
話者 4	怒り	k 650%	n 550%	q 500%	t 467%	e 460%
	悲しみ	k 400%	e 300%	n 300%	t 267%	r 250%
	喜び	k 750%	n 750%	t 400%	e 380%	u 380%
話者 5	怒り	n 386%	i 340%	k 300%	N 300%	q 300%
	悲しみ	e 300%	k 300%	r 300%	q 300%	i 260%
	喜び	k 500%	n 350%	q 350%	N 320%	a 318%
話者 6	怒り	k 500%	sh 400%	o: 300%	u 222%	n 200%
	悲しみ	k 500%	o: 400%	r 300%	sh 300%	w 250%
	喜び	k 500%	sh 500%	o: 500%	q 300%	w 250%
話者 7	怒り	o: 300%	w 167%	a 154%	k 160%	e 139%
	悲しみ	j 300%	q 200%	o: 200%	b 200%	N 183%
	喜び	o: 500%	k 240%	s 225%	q 200%	j 200%

表 2.9: 感情ごとの誤認識変動率 (女性話者)

話者	感情	1 位	2 位	3 位	4 位	5 位
話者 8	怒り	j 300%	s 217%	k 208%	i 206%	m 170%
	悲しみ	s 300%	j 300%	w 250%	n 240%	u 211%
	喜び	j 300%	k 192%	i 167%	s 167%	m 140%
話者 9	怒り	N 325%	o: 250%	sh 233%	s 220%	k 200%
	悲しみ	s 300%	j 300%	w 250%	n 240%	u 211%
	喜び	N 250%	o: 250%	w 233%	a 230%	s 220%
話者 10	怒り	g 400%	N 350%	sh 300%	y 300%	o 164%
	悲しみ	N 500%	g 400%	y 350%	sh 300%	o 179%
	喜び	N 700%	b 300%	sh 250%	y 250%	g 200%
話者 11	怒り	j 200%	sh 160%	w 150%	g 125%	q 120%
	悲しみ	w 175%	q 140%	r 130%	d 127%	m 123%
	喜び	sh 500%	s 350%	r 300%	o: 300%	n 238%
話者 12	怒り	N 200%	b 167%	s 143%	k 140%	o 126%
	悲しみ	N 300%	g 250%	k 200%	b 200%	t 183%
	喜び	N 275%	g 200%	o: 167%	b 167%	e 159%
話者 13	怒り	sh 175%	j 150%	m 107%	i 100%	N 100%
	悲しみ	sh 200%	j 150%	i 141%	N 125%	n 119%
	喜び	N 125%	sh 125%	y 117%	i 114%	a 100%

第3章

音声による感情の判別

3.1 はじめに

第2章では，感情によって音声が変化した結果，連続音声認識エンジンに有意差を伴う影響を与えることを述べた．これらの影響は，現在一般的に用いられている音声認識エンジンの音響モデルが平静音声を想定して設計されているために生ずるものであり，誤認識を減らすためには認識の前段階で感情ごとに音響モデルを切り替えるなどの対策を行う必要があることを示した．また序論でも述べたように，ユーザの心理状態によって対応を変化させる自然な対話機構を実現するためには，音声を用いて感情を推定する機構が不可欠である．本章では，これらの要求を満たすための音声による感情判別の方法を検討する．また実際に判別実験を行って提案する手法の有効性を確認する．

3.2 感情の推定方法

音声による感情推定は発話内容(コンテキスト)から得られる情報を用いる方法，つまり実際に音声認識を行ってその結果を利用する方法と，韻律や振幅，強勢などの信号処理的方法によって行う方法が考えられるが，前者は大語彙連続音声認識システムの前段としての用途には計算量の面からみて困難であることをすでに述べた．従って実際に感情推定を行うには，後者である感情音声から特徴量を抽出して感情推定を行う方法を用いることが現実的である．

抽出された特徴量を用いて感情推定を行うには，ニューラルネットを用いた方法や判別分析を用いる方法などが考えられる．しかし，ニューラルネットの構築には，一定量以上のサンプルが必要であることが難点として存在する．従って本研究においては判別分析を用いて特徴量の基本的な性能を調査することとした．

3.3　判別分析とは

判別分析とは，いくつかの変数に基づいて，各データがどの群に所属するかを判定する統計手法である．以下，判別分析について簡潔に述べる．

各群につき p 個のサンプルがあるとする．k 個の群についての母集団の平均を $\boldsymbol{\mu}_j = (\mu_{1j}, \mu_{2j}, \cdots, \mu_{pj})'$ 但し $j = 1, 2, \cdots, k$，観測値を $\mathbf{X} = (X_1, X_2, \cdots, X_p)'$ とする．また，各群の分散共分散行列を $\sum_j$，その逆行列を $\sum_j^{-1}$ とするとき，観測値 X の各群までのマハラノビス距離 d_j^2 は

$$d_j^2 = (\boldsymbol{X} - \boldsymbol{\mu}_j)' \sum\nolimits_j^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_j) \tag{3.1}$$

によって定義される．それぞれのサンプルは，このマハラノビス距離が最も小さい群に属すると判別するのである．ここで，もし各群の分散共分散行列が等しい，すなわち $\sum_1 = \sum_2 = \cdots = \sum_k = \sum$ を仮定することができるならば，式 (3.1) は，

$$\begin{aligned} d_j^2 &= (\boldsymbol{X} - \boldsymbol{\mu}_j)' \sum\nolimits^{-1} (\boldsymbol{X} - \boldsymbol{\mu}_j) \\ &= \boldsymbol{X}' \sum\nolimits^{-1} \boldsymbol{X} - 2\boldsymbol{X}' \sum\nolimits^{-1} \boldsymbol{\mu}_j + \boldsymbol{\mu}_j' \sum\nolimits^{-1} \boldsymbol{\mu}_j \end{aligned} \tag{3.2}$$

と書き換えることができる。第 1 項は各群に共通，第 3 項は各群ごとに異なる定数 (これを c_j とする) である．各ケースごとに異なるのは第 2 項のみであるから，マハラノビス距離の算出には

$$2\boldsymbol{X}' \sum\nolimits^{-1} \boldsymbol{\mu}_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p \tag{3.3}$$

を計算すればよいことになる．この場合の係数 $a_{1j}, a_{2j}, \cdots, a_{pj}$ は $\sum^{-1}$ の要素を σ^{is} とすれば，

$$a_{ij} = 2(\sigma^{i1}\mu_{1j} + \sigma^{i2}\mu_{2j} + \cdots + \sigma^{ip}\mu_{pj}), \quad i = 1, 2, \cdots, p \tag{3.4}$$

により求めることができる．式 (3.2) の第 1 項は無視できるので，式 (3.5) の値が最も小さい群に属すると判定すればよい．

$$f_j = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{pj}X_p + c_j \tag{3.5}$$

式 (3.5) を分類関数と呼ぶ．また，k 個の群の中から，任意の 2 群の組み合わせに対して，式 (3.6) で表される ${}_kC_2$ 個の判別関数を定義しておくこともできる．例えば，第 1 群と第 2 群の判別関数は，

$$\begin{aligned} Z_{1\cdot 2} &= d_1^2 - d_2^2 \\ &= (a_{11} - a_{12})X_1 + (a_{11} - a_{22})X_2 + \cdots + (a_{p1} - a_{p2})X_p + (c_1 - c_2) \end{aligned} \tag{3.6}$$

となる．a_{ij} は判別係数と呼ばれ，変数 X_i の判別への寄与割合を与える重要な指標である．

3.4　特徴量の選定

判別分析で用いる特徴量は，感情ごとの相違がよく反映されるものであることが重要であり，これまでにも様々な特徴量が検討されてきた．これまでの研究で音声の感情表現は，基本周波数やフォルマント周波数などの韻律面の構造，文の強勢やアクセントなどの振幅構造，そして文の長さなどの時間構造の三つの側面に現れると考えられ，従来の研究 [11][12][13] においてもその有効性がある程度確認されている．本研究ではこれらを考慮して基本周波数をベースとした特徴量を用いている重永の研究 [13] に着目した．重永の研究においては，

$F0_{max}$: 発話中の基本周波数 (F0) の最大値 [Hz]
A_{max} : 発話中の振幅の最大値
T : 発話の開始から終了までの時間長 [sec]
$F0_{init}$: 発話の開始直後の基本周波数 [Hz]
$F0_{range}$: 発話中の最大基本周波数 − 最小基本周波数 [Hz]

の 5 種の特徴量が基礎量として用いられていたが，本研究ではこれらに加えて，

T_{zero} : 発話中の非発声部分 (ポース) の累積時間長 [sec]
$F0_{avg}$: 発話中の基本周波数の平均値 [Hz]
$F0_{diff}$: 発話中の基本周波数の隣接フレーム間差分の平均値 [Hz]
A_{avg} : 発話中の振幅の平均値 [Hz]

の 4 種の統計量を加え，全部で 9 次元のパラメータを構成した．基本周波数軌跡上でこのうち 5 つを図示したものが図 3.1 である．

以下，提案パラメータを採用した理由について述べる．T_{zero} は事前に行った基本周波数波形の視察によって，特に悲しみの場合に単語間のポーズ (無音区間) が長くなる傾向が見られ，判別に寄与すると考えたためである．また $F0_{diff}$ については怒りや喜びの感情で基本周波数の変化が他にくらべて急峻になっている点に注目したためである．また $F0_{avg}$ については $F0_{max}$ や $F0_{range}$ と相関があり，A_{avg} も同様の相関性を A_{max} との間で持つことが予想されるが，判別においては平均値を加えることが性能向上につながるとされているため，補助的な意味合いで導入した．

3.5　特徴量の抽出

3.5.1　基本周波数の抽出

基本周波数の抽出には，発話全体の連続性を考慮した DP マッチングによる方法[10] を使用した．この手法は発話全体の情報が必要であるため実時間処理には向かないが，その

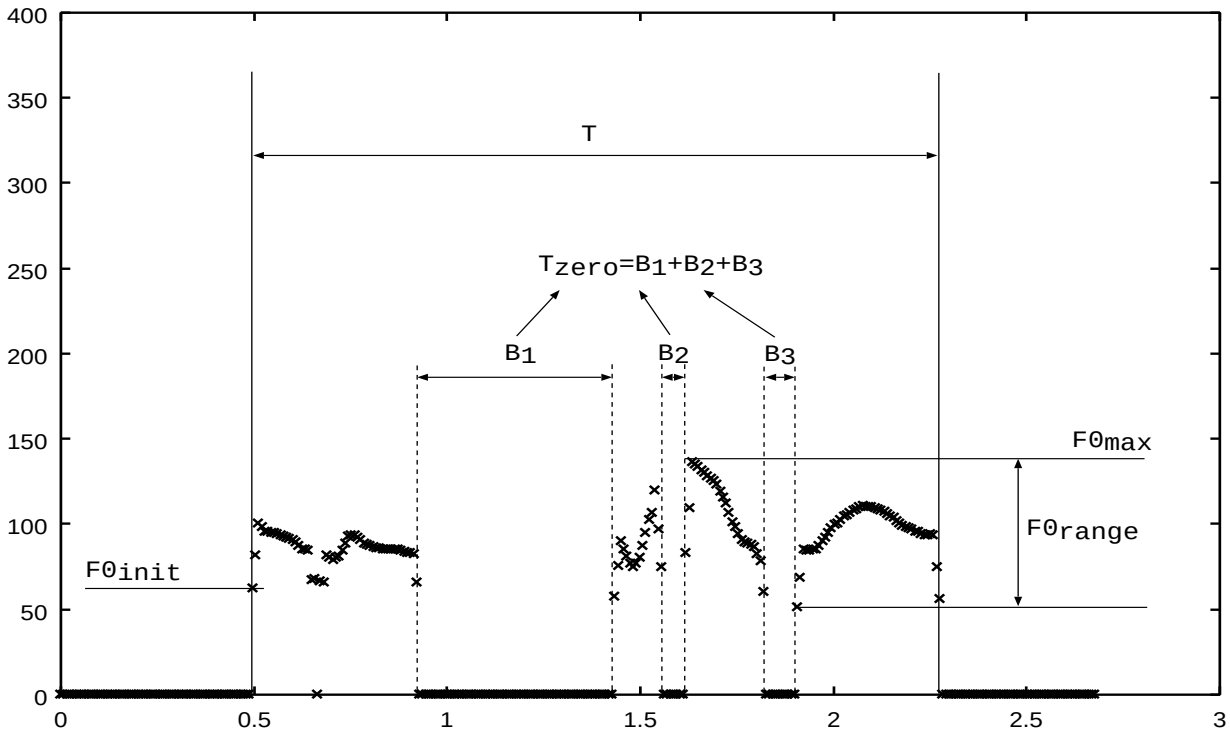


図 3.1: 基本周波数軌跡と各パラメータの関係

代わりに正確な基本周波数抽出が可能である，という利点がある．

以下，この手法を簡潔に説明する．図 3.2 に処理の流れを示す．音声データは DFT によって周波数領域のデータに一旦変換された後，対数を取って逆 DFT される．この操作によって，音声データはケプストラムとよばれる時間領域のデータになる．このケプストラムから，ピーク値の大きい順にある一定数を選び，このピークを DP によって滑らかにつなぐのがこの基本周波数抽出法の基本的な考え方である．DP のコスト評価は以下のようにして行う．

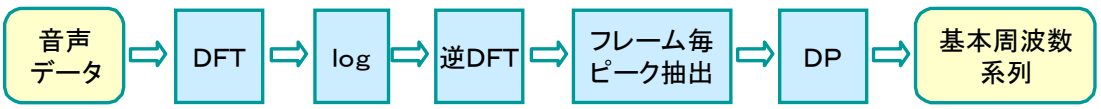


図 3.2: 基本周波数抽出の流れ

発話の i フレーム目におけるケプストラムのうち，パワーで正規化した値の大きい順に k 番目のピークの周波数を $P_{k,i}$，その値を $PH_{k,i}$ として，隣接した時刻 $i-2, i-1, i$ における第 n, m, k 番目のピークを結んだとき，コスト評価関数 C は式 (3.7) のように定義される．

$$C(i,k,m,n) = \begin{cases} -w_1 \cdot (PH_{k,2} + PH_{m,1}) + w_2 \cdot |P_{k,2} - P_{m,1}| & (i = 2) \\ -w_1 \cdot PH_{k,i} + w_2 \cdot |P_{k,i} - P_{m,i-1}| & \\ \quad +w_3 \cdot |P_{k,i} - 2 \cdot P_{m,i-1} + P_{n,i-2}| & (i > 2) \end{cases} \tag{3.7}$$

$P_{k,i}$ は基本周波数の抽出範囲に限定して選択する．また w_1 はピークの大きい候補を選ぶための係数であり，w_2，w_3 は順に 1 フレーム前，2 フレーム前までの連続性を保つための係数である．式 (3.7) を全フレームの全ての可能な経路について計算して総和を取り，最小となる経路を基本周波数軌跡とする．実際には 1 フレーム前までを考慮した漸化式である式 (3.8) を逐次的に計算する．

$$S(i,k) = \min_{m}[S(i-1,m) + C(i,k,m,n^{'})] \tag{3.8}$$

$n^{'}$ は各 m で $S(i-1,m)$ の最小値を与える n を意味する．式 (3.8) を $i = 2$ から $i = N$(N は総フレーム数) まで計算し，スコア最小の経路をバックトレースすることで最適系列が出力される．

基本周波数抽出時に使用したパラメータは表 3.1 の通りである．

表 3.1: 基本周波数の抽出パラメータ

DFT 次元数	1024 次元
フレームシフト	8ms(128 フレーム)
F0 抽出範囲	50 ~ 400Hz
C の重み係数	$w1 = 6000, w2 = 1.5, w3 = 1$

3.5.2　その他の特徴量の抽出

時間構造の抽出に用いる発声部分の検出には，振幅の絶対値がしきい値を下回ったフレームを非発声部分 (ポーズ) とすることで行った．しきい値は予備実験を行って，振幅波形の目視による分割と最も符合するよう経験的に定めた．

3.6　実験内容

3.5 節の方法により抽出された特徴量を用いて感情の判別分析 [18][19][21] を行った．評価の方法は各サンプルごとに該当する感情を示すラベルを付加して判別分析を行い，判別された感情と与えたラベルが一致すれば正しく判別されたものと見なし，そうでなければ誤判別とする．実験は以下のような，混合の有無と混合の仕方の違いによる 4 つの場合のそれぞれについて行った．

1. 話者ごとの 80 サンプルのみを別々に用いる (話者ごとの環境)
2. 男性話者の 560 サンプルを混合する (男性話者混合環境)
3. 女性話者の 480 サンプルを混合する (女性話者混合環境)
4. すべての話者の 1040 サンプルを混合する (全話者混合環境)

混合環境については重永の提案するパラメータのみで判別実験を行った場合と，本研究で提案する 4 つのパラメータを加えた場合とで比較し，提案したパラメータの有効性を検証した．

3.7　実験結果

3.7.1　話者ごとの場合

話者ごとの 4 感情計 80 文で閉じた環境にあるデータを用いて話者ごとに判別分析を行った結果を図 3.3 と図 3.4 に，その平均値を図 3.5 に示す．

話者によって判別率に若干のばらつきが見られるものの 80 ~ 95%程度，平均でも 90%近い判別率が得られた．また感情ごとの全話者平均値を見ると，標準 → 怒り → 悲しみ → 喜びの順で判別率が低下する傾向が見られた．

3.7.2　話者混合データの場合

次に 3 通りの方法でデータを混合した場合についての結果を図 3.6 に示す．

実験の結果，平均判別率は男性話者においては 64.82%，女性話者では 68.92% であった．また全体を混合した場合では 58.08%であった．結果を個別に見ると，男性話者の「怒り」の判別率が他と比較して特に高い値が得られている．これは，男性話者が総じて怒りの感情表現を，他の感情とは明らかに異なる方法で行っていることを示している．また，対照的に喜びの判別率は低い値であり，男性の喜びの表現方法が他の感情とあまり違わないためであるとも考えられる．

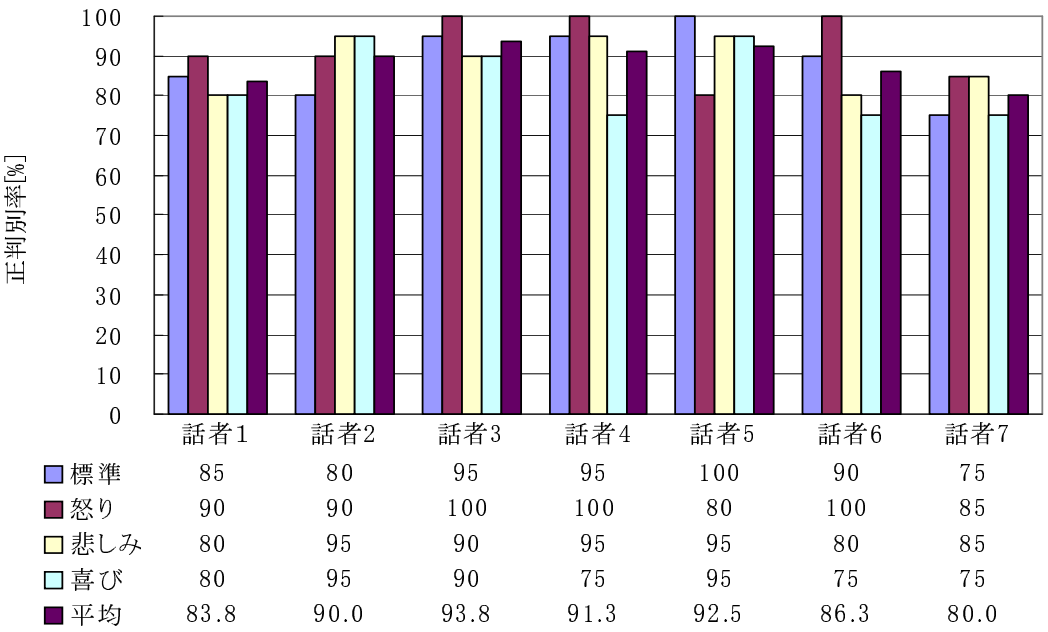


図 3.3: 話者ごとの判別率 (男性話者)

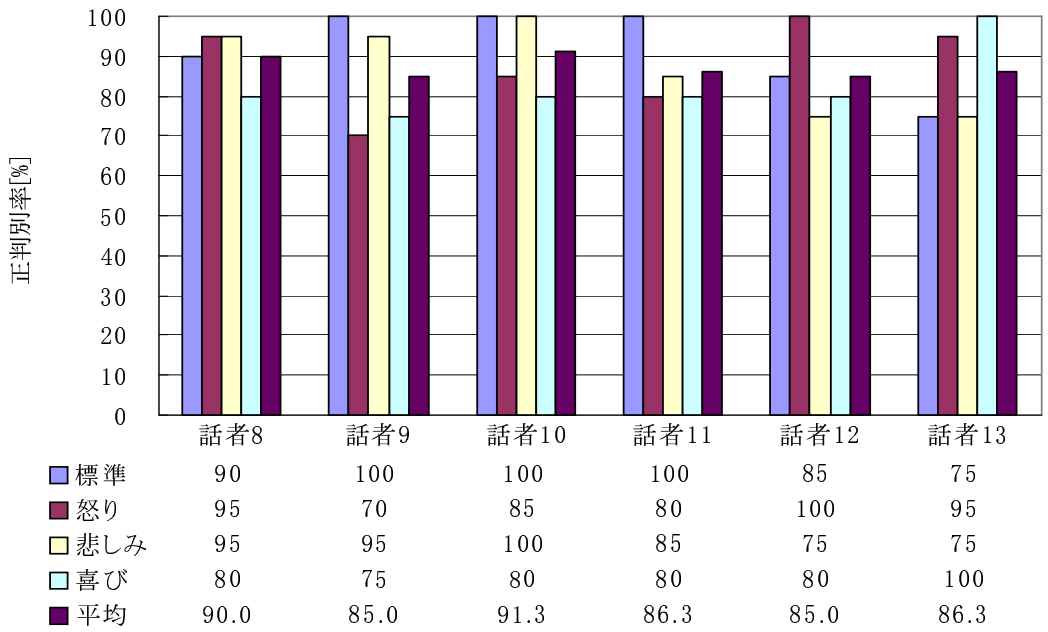


図 3.4: 話者ごとの判別率 (女性話者)

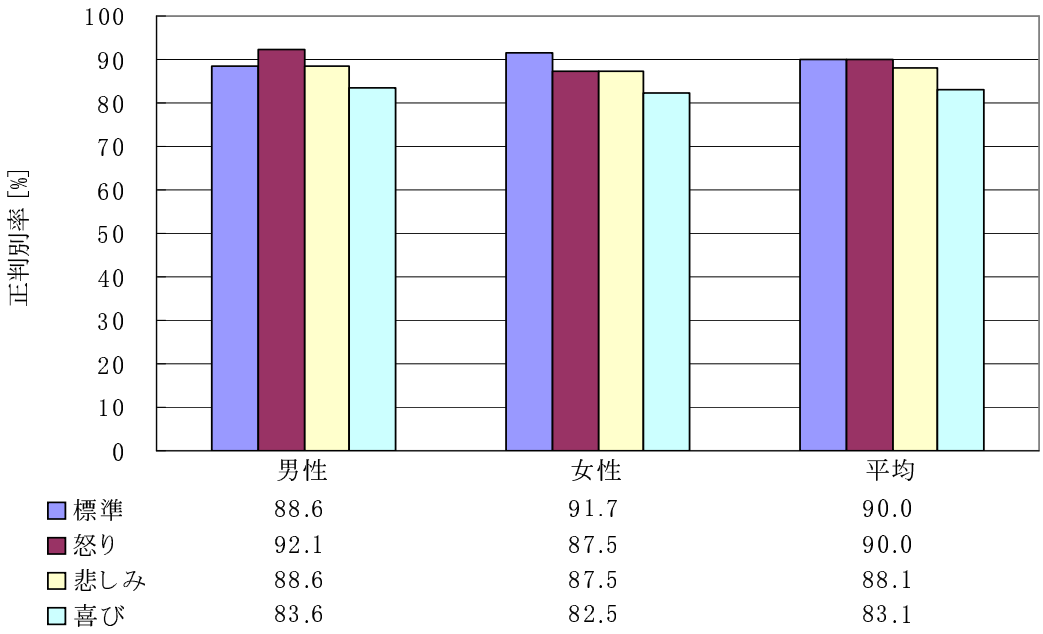


図 3.5: 話者ごとの判別率の平均値

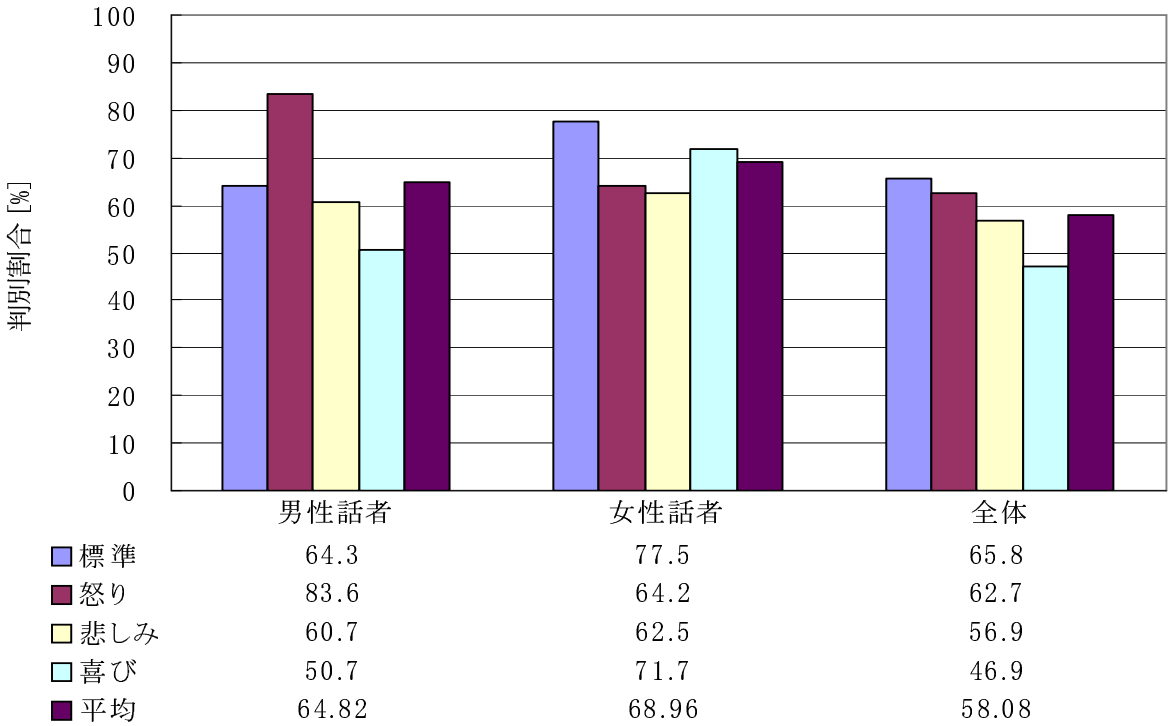


図 3.6: 混合データの判別率

一方女性話者については，標準と喜びの判別率が高く，怒りと悲しみの感情間では認識率にあまり差がみられなかった．しかしながら，女性話者では判別率は全体的に男性話者よりも高く，女性の感情表現が豊かであることの表れともいえる．

最後に全話者を統合した場合であるが，概ね両者の低い方の値に揃えたような値に近い結果が得られた．男性話者の怒りにあった特徴的な高い判別率や，女性話者の男性話者よりかなり高い喜びの判別率は失われており，データを混合することでこれら性別に依存した感情表現が埋もれてしまっていることが見て取れる．このことを確認するため，各サンプルが分類された個数を感情ごとに調べた．結果を図 3.7 と図 3.8 に示す．

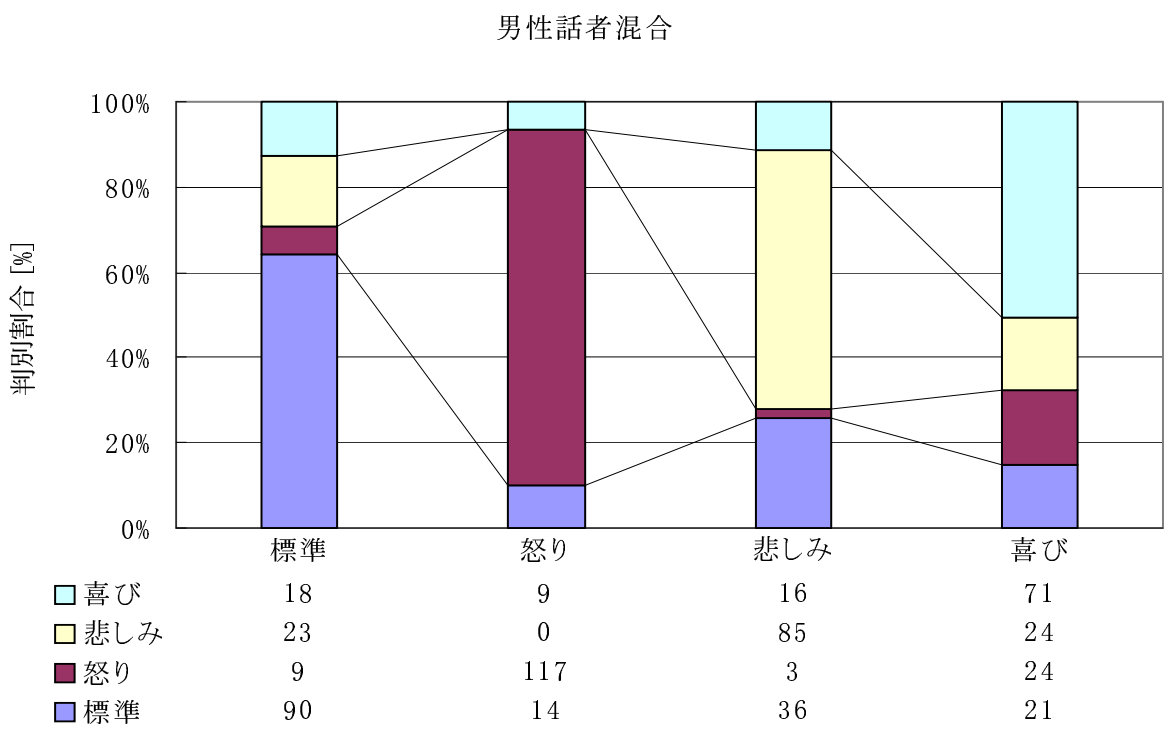


図 3.7: サンプルの分類個数と割合 (男性話者)

男性話者の場合は怒りが突出して高い判別率であったが，このことを裏付ける結果が得られた．また喜び以外の三感情では特定の一感情との判別誤りが多かったのに対し，喜びでは他の感情とまんべんなく誤りが起こっていることが分かった．女性話者ではこのような現象は見られなかった．

全体的に見て，話者本人が悲しみのつもりで発話しているのに，実際は怒りに判別されてしまうようないわば正反対の誤判別はごく少数であり，提案した特徴量でこの二感情間の判別は十分に行うことができると考えられる．しかしながら喜びについては怒りや標準との誤判別が多く，提案した特徴量では喜びに特有の感情表現を拾いきれていないことも明らかになった．

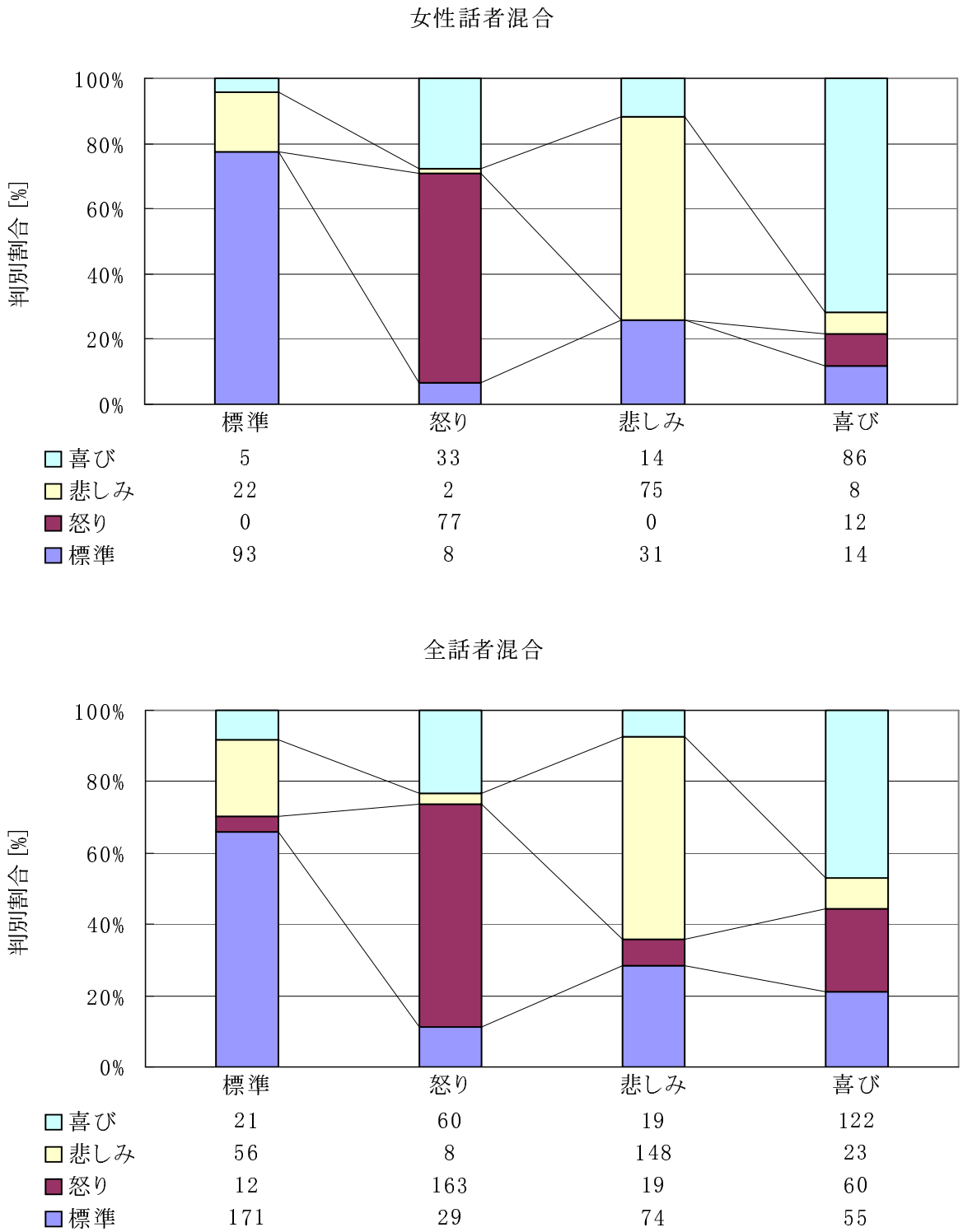


図 3.8: サンプルの分類個数と割合 (女性話者および全話者)

3.7.3　従来手法との比較

次に提案した 9 種の特徴量の有効性を検証するため，重永の研究において使用されていた特徴量のみを用いて判別分析を行った．この結果により提案手法と判別率の比較を行ったものが表 3.2 である．

表 3.2: 従来手法との比較

混合方法	感情	従来法	提案法
男性話者	標準	65.0	64.3
	怒り	80.7	83.6
	悲しみ	57.1	60.7
	喜び	42.1	50.7
	平均	61.25	64.82
女性話者	標準	69.2	77.5
	怒り	58.3	64.2
	悲しみ	60.8	62.5
	喜び	66.7	71.7
	平均	63.75	68.96
全体	標準	68.1	65.8
	怒り	61.9	62.7
	悲しみ	53.5	56.9
	喜び	46.5	46.9
	平均	57.50	**58.08**

青字は提案法が従来法を上回ったもの，赤字は下回ったものを示す．提案手法は男性話者混合の「標準」と，全体混合の「標準」以外のすべての場合で従来法を上回る判別率を得ることができた．特に最も重要な性能指標である全体混合の平均において 0.58%とわずかではあるが，提案法が上回る結果となった．このことから提案法は判別率の向上に有効であることが分かった．しかしながら男性話者，女性話者のそれぞれだけで混合した場合は 2 ~ 8%程度の判別率の上昇が見られるのに対し，全体を混合した場合では 0.4 ~ 3.4%とあまり向上が見られない．このことは，性別の異なる話者を混合することによって生ずる特徴量空間のひずみによる影響が関係しているものと思われる．

3.8　聴取実験による評価

前節までは，発声された音声には発声者の意図した感情が忠実に含まれている，ということを前提に分析を行ってきた．しかしながら発声者がある特定の感情を込めて発声したつもりであっても，それを聴いた人はかならずしもその意図通りに受け止める訳ではない．このことは実際の人間のコミュニケーションにおいても，感情の誤認識が頻繁に発生していることからも明らかである．そこで本節では，発声された感情を含む音声にどの程度その感情が含まれているかを聴取実験により評価付けし，判別分析により判別された結果との相関を調査する．

3.8.1　実験の方法

まず各サンプルにどの程度感情が含まれているかを人間の主観によって評価付けするため，聴取実験を行った．実験の方法を表 3.3 に示す．

表 3.3: 聴取実験の方法

被験者数	4 名 (いずれも男性,A ~ D とする)
被評価話者	話者 1 ~ 4(1 話者あたり 3 名の被験者が評価し平均化)
使用機器	計算機上の PCM 音源+ヘッドセット
評価方法	一話者分の 80 サンプルをランダムに提示して 5 段階評価 (1:全くない, 2:少しある, 3:普通, 4:かなりある, 5:すごくある)

評価付けには，各サンプルごとに怒り，悲しみ，喜びの各成分がどの程度含まれているかを評価する方法や，単にその感情の強さだけを評価する方法が考えられるが，前者の場合は被験者の負担が非常に大きいため，後者の方法により評価付けを行うこととした．実際の実験は話者 1 ~ 4 について行った．各話者への被験者の割り当てを表 3.4 に示す．

表 3.4: 各話者への被験者の割り当て

話者	割り当てた被験者
話者 1	被験者 A,B,C
話者 2	被験者 A,C,D
話者 3	被験者 B,C,D
話者 4	被験者 A,B,D

分析は聴取実験によって得られた各サンプルの平均評価値と，そのサンプルが属する感情の群からのマハラノビス距離をとの相関を調べることで行った．なお，距離の測定には 4.3 節で述べる補正を施した話者混合環境におけるサンプルを使用した．

3.8.2　実験結果

聴取実験により得られた相関係数と決定係数を表 3.5 に示す．

表 3.5: 話者ごとの相関係数

話者	相関係数	決定係数
話者 1	0.05168	0.00267
話者 2	0.09716	0.00944
話者 3	0.15866	0.02517
話者 4	0.14694	0.02159

相関係数はいずれも 0.16 以下と，極めて低くなっている．一般に相関係数が 0.7 ~ 0.9 程度であるとき相関があるとみなすことができるが，表 3.5 の結果はこの条件を満たしていない．従って相関はないと考えられる．

3.8.3　考察

聴取実験による評価と判別結果とに相関が見られなかった原因はいくつか考えられるが，被験者によって評価に大きな差が出たことが主な原因であると考えられる．図 3.9 は話者 1 のサンプルに対して 3 名の被験者により行われた聴取実験の評価値のヒストグラムである．被験者 A は評価値 4 を最頻値としており，平均評価値も 3.78 と比較的高い．被験者 B はほとんどの評価値が 3 か 4 に集中していて，平均値は 3.46 であった．一方，被験者 C は大半のサンプルが 3(普通) 以下であり，被験者 A,B とは大きく傾向が異なっている．

このように，同一条件で聴取実験を行ったにも関わらず，ランダムに感情を提示した場合の評価は被験者によって全く異なっているといってよい結果となった．このことは，実際の会話において人間が感情を判断する時には，平静の音声を知識として持っており，そこからの変化によって感情を認識しているため，知識なしに突然感情を込めた音声を与えられた場合に判断が分かれることを示しているとも考えられる．従ってこのような相関を調べるときは，まず標準を聴かせた後に感情を含んだ音声を聴いてもらい，どの程度変化したか，という評価を元に行う方法が妥当ではないかと考えられる．

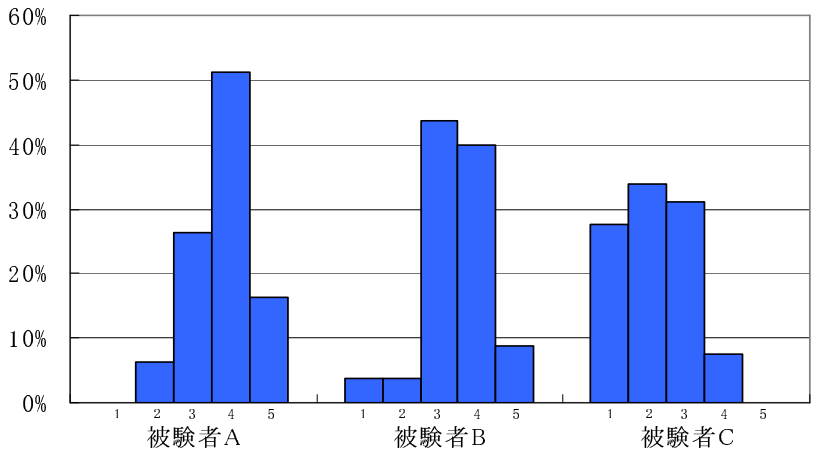


図 3.9: 被験者ごとの評価値の分布割合 (話者 1)

3.9　まとめ

本章では，判別分析を用いて感情を推定する方法について述べ，判別実験を行って提案手法の有効性を検証した．話者ごとに行った実験では平均で 90%程度の判別率を得ることができた．また男性話者，女性話者および全体のそれぞれを混合した場合では順に 64.82%，68.96%，58.08%の判別率を得た． また従来の研究による 5 次元の特徴量と提案した 9 次元の特徴量の判別率を比較した結果，全体を混合した場合で 0.58%の向上が得られた．これにより提案法で追加した特徴量が判別率の向上に有効であることが分かった．また, 判別した結果と人間の主観による感情評価との相関を調べたが相関は見られず，聴取実験の方法を再検討する必要があることが分かった．

第4章

感情判別に有効な特徴量

4.1 はじめに

前章では，感情を含む音声から特徴量を抽出し，実際に感情の判別を行う方法について述べた．また判別実験により，提案した特徴量が判別率の向上に寄与することを示した．しかしながら，判別率は性別や感情の種類によって著しく異なっており，ある話者に有効である特徴量が，他の話者に対しても普遍的に有効であるかは明らかではない．もしも有効性に話者による相違があるならば，話者ごとに正規化を行うなどの処理を行うことによって判別率の向上が期待される．そこで本章ではまず，話者ごとの特徴量の寄与割合を調査する．次に，特徴量の正規化について述べ，正規化を施したデータを用いて判別率分析を行い，正規化の有効性について検証する．

4.2 特徴量の寄与割合

感情を判別する際には，4.1 節で述べたように，判別に寄与する特徴量が話者の違いや感情の違いによって大きく異なることが予想される．怒りと悲しみの判別には振幅の最大値 (A_{max}) が有効であると考えられるし，悲しみは全体的に低い周波数で発声されることが予想されるので，$F0_{[max,range,init]}$ などが有効ではないかと考えられることなどがその例である．そこで本節では 3.3 節で述べた判別分析で感情間の判別に用いられる判別係数に注目し，特徴量ごとの判別への寄与割合を判定した．判別係数は標準化されているので，係数が大きい特徴量ほど判別に有効であるといえる．判別係数を 3.6 節で述べた「話者ごと」と，混合環境である「男性話者」,「女性話者」，そして「全体」の各分類ごとに調査した．

4.2.1　話者ごとの環境

話者ごとの環境では，判別係数の分布は大別して 2 つのグループに分類することができた．一つは，1 ~ 4 種類の特徴量が突出して高く，他の特徴量はいずれも同程度の値にとどまった「同程度型」のグループであり，話者 1,2,6,7,9,11,12,13 などがこのグループに該当した．これらの話者の判別係数を図 4.1 に示す．1 つのグラフには 4 種類の感情の中で任意の 2 つの感情を判別する ${}_4C_2 = 6$ 通りの標準化判別係数を示してあり，それぞれの列が 1 つの感情の組み合わせを示す．例えば「標– 怒」とある列には，標準と怒りのサンプルを判別するのに用いる係数を示してある．同程度型のグループでは，判別の決め手となる係数の大きい特徴量があることはあるが，それが 1 ~ 4 種類と少数にとどまっているのが特徴である．ほとんどの話者で $F0_{max}$ あるいは A_{max} が有効であり，それ以外の特徴量はいずれも低い．同程度型のグループでは，平均の判別率が 80%から最高でも 86%にとどまっている．

これに対して二つめのグループは，最初のグループと比較してまんべんなく 4 ~ 5 種類程度の特徴量が高くなっている「多種突出型」であるのが特徴で，話者 3,4,5,8,10 が該当した．これらの話者では同程度型よりも相対的に高い 90 ~ 95%の判別率を得ていた．特に同程度型ではほとんど判別に寄与していなかった T_{zero}，$F0_{range}$，A_{avg} などが有効に働いていることが注目される．

これらの結果から，本研究で用いた特徴量だけを見ても，話者によって感情表現を行う方法がかなり異なっていることが分かった．例えば同程度型のグループでは，感情の表現を主に振幅の最大値 (A_{max}) によって変化させている話者が多いのに対し，多種突出型の話者では，振幅に加えて基本周波数のダイナミックレンジ $(F0_{range})$ や，文節間の間合いの取り方 (T_{zero}) など複数の特徴を複合的に用いて感情を表現していることなどがその例である．

4.2.2　話者混合環境

男性，女性，全体の 3 通りの方法でデータを混合した場合の判別係数も同様に調査した．結果を図 4.3 に示す．男性話者においては振幅の最大値 (A_{max}) が大きく，標準– 怒りの判別などでこの傾向が目立っている．他にも $F0_{max}$ や $F0_{avg}$ などが判別に寄与していることが分かる．しかしながら，標準– 喜びの判別の場合には，どの特徴量も同程度の値を示している．このことは，男性話者の標準と喜びの判別は今回用いた特徴量が有効に働いていないか，あるいは男性話者が標準と喜びをあまり相違なく発話しているかのどちらかが原因であると考えられる．

一方女性話者においては，男性話者に比較して $F0_{max}$ の寄与率が高くなっているなど傾向が異なっている．実質的に $F0_{max}$ か A_{max} のみで判別が行われているとみなすこともできる．

最後に全体を混合した場合であるが，この場合は概ね男性と女性の平均を取ったような寄与割合になっていることが分かる．また全体的に係数が小さくなっており，混合することで特徴量空間にひずみが生じ，判別が難しくなっていることが考えられる．このことは男性話者だけでは 64.82%，女性話者だけでは 68.96%の判別率が得られたのに，混合データの場合では 58.08%の判別率しか得られないことからも見て取れる．

4.3　性別差補正の導入

4.3.1　性別による特徴量の相違

前節では，男性話者と女性話者を混合することで事前に予想した男女間の平均値である 67%程度よりも低い 58%程度の判別率しか得られず，混合による特徴量空間のひずみが考えられることを述べた．このことを検証するため，男性話者と女性話者の各特徴量の平均値を調べた．結果を表 4.1 に示す．

表 4.1: 特徴量の平均値の性別による相違

特徴量	単位	平均値		男女比
		男性話者	女性話者	(男性/女性)
$F0_{max}$	[Hz]	197.29	316.91	0.62255
A_{max}	-	19313	11500	1.67933
T	[sec]	1.809	2.087	0.86702
$F0_{init}$	[Hz]	75.19	120.68	0.62304
$F0_{range}$	[Hz]	140.38	245.31	0.57227
T_{zero}	[sec]	0.3778	0.4454	0.71349
$F0_{avg}$	[Hz]	82.27	138.65	0.59336
$F0_{diff}$	[Hz]	2.607	4.575	0.56976
A_{avg}	-	17.43	13.76	1.26609

一般に，男性の基本周波数は 50 ~ 200Hz，女性では 150 ~ 350Hz とされる [3] が，表 4.1 はそれを裏付けている．男性話者は振幅面では女性話者よりも大きいのに対し，他の特徴量では下回っており，最大で 2 倍近い相違が見られた．このことは，前節で述べた性別間の相違による特徴量のひずみを引き起こすものである．従って，判別率の向上にはこのようなひずみを何らかの形で補正する必要がある．

補正には様々な方法が考えられるが，本研究では男性話者と女性話者の特徴量の重心を

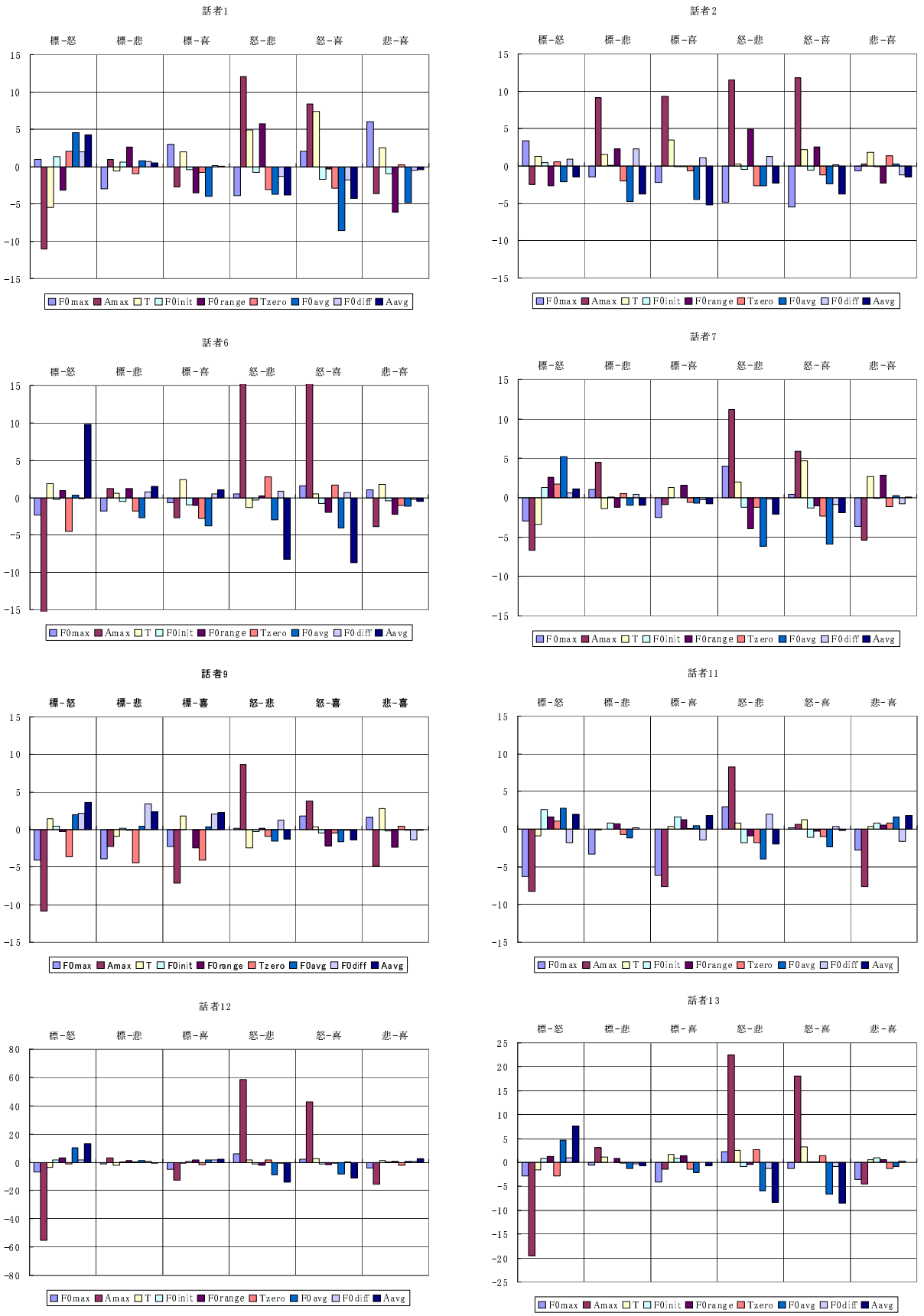


図 4.1: 話者ごとの標準化判別係数 (同程度型)

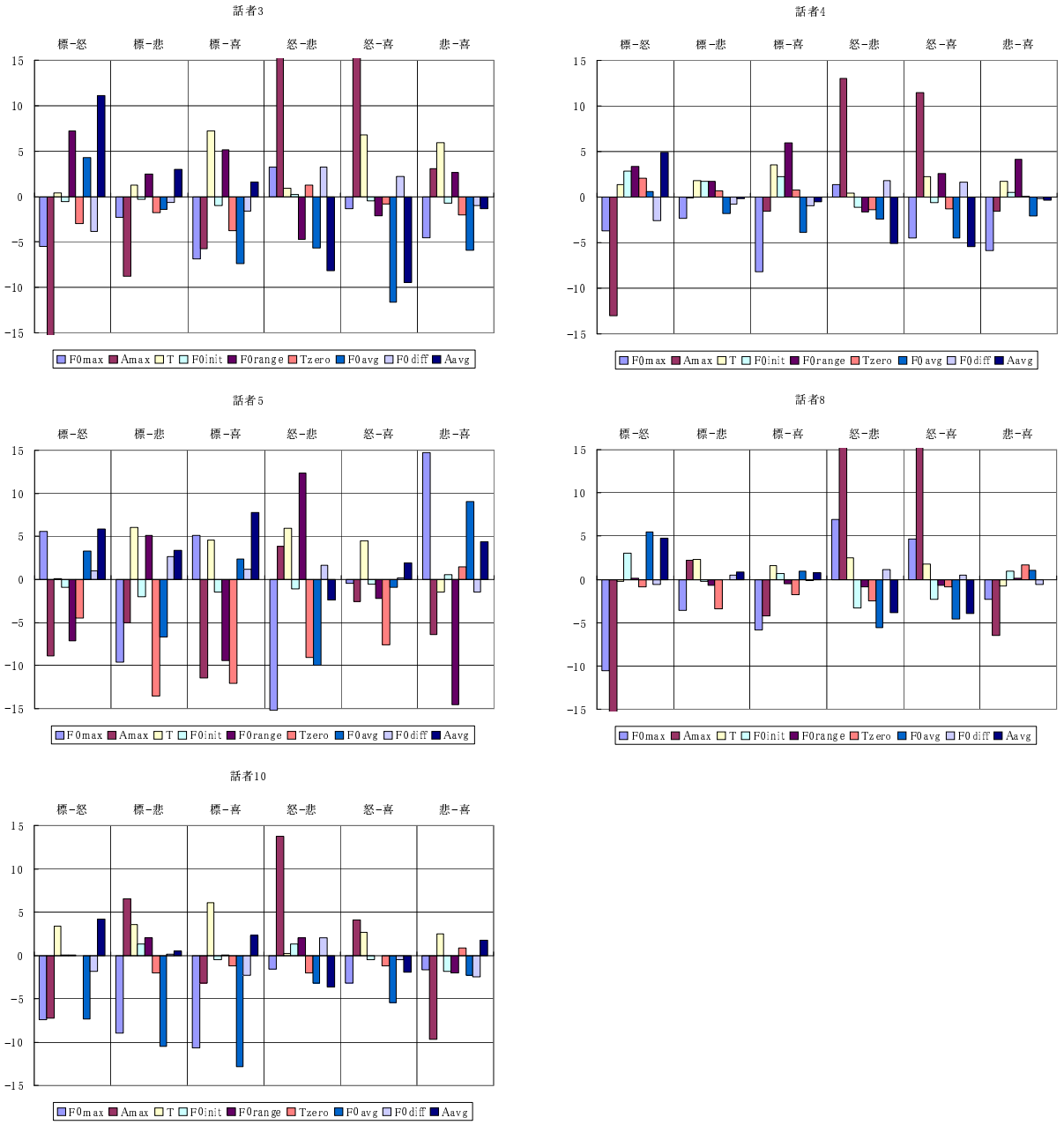


図 4.2: 話者ごとの標準化判別係数 (多種突出型)

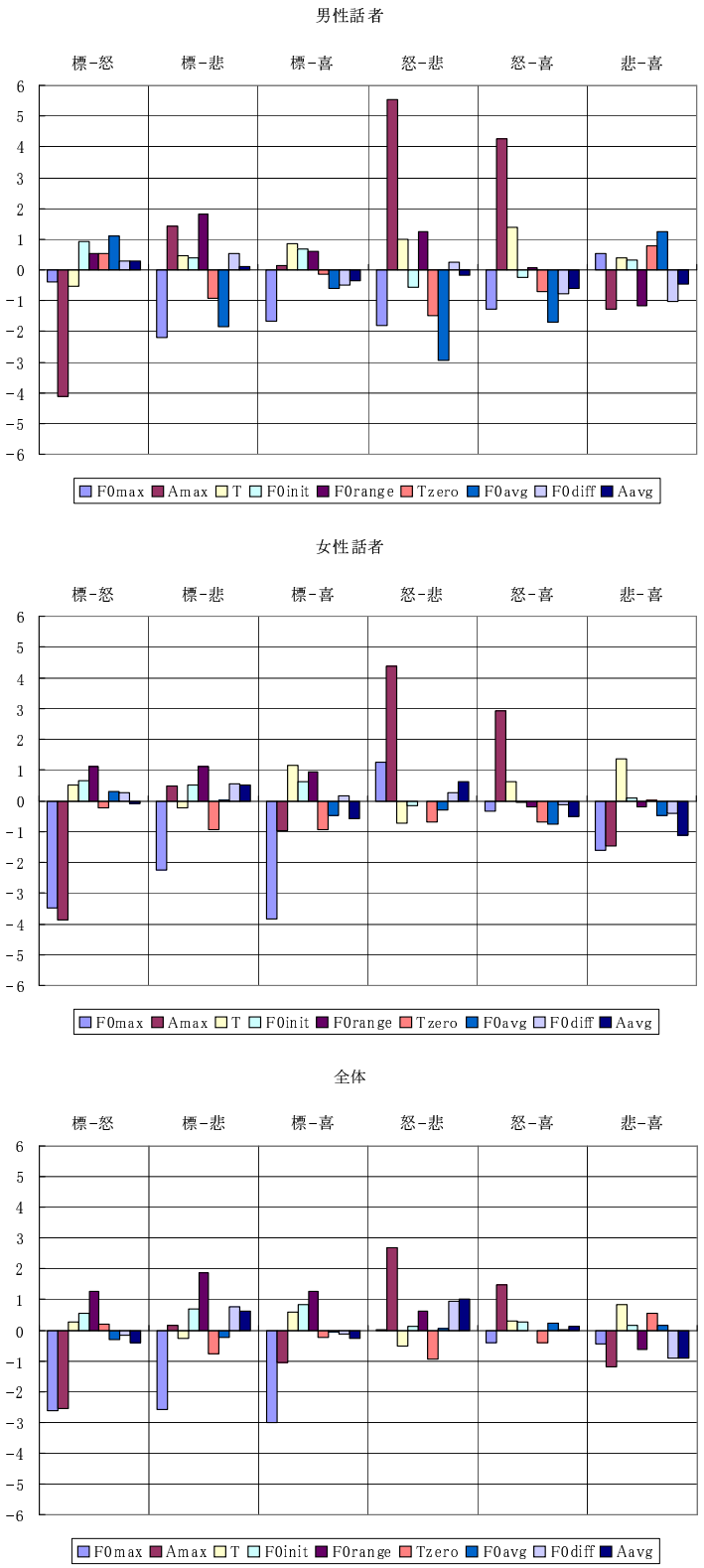


図 4.3: 話者混合環境の標準化判別係数

揃えることで，ひずみを低減させる方法を採用した．具体的には表 4.1 の男女比を女性話者のサンプルに乗ずることで行った．なお，この補正方法はあらかじめサンプルが男性か女性かどちらに属するかを知る必要があるが，男女の判断は基本周波数の平均値等の情報を用いて単純に行うことが可能であるため，適用が容易であるという利点がある．

4.3.2　補正による判別率の向上

性別差補正を施した場合の正判別率を表 4.2 に示す．結果は全体を混合した場合のデータについてのみ示してある．

表 4.2: 補正前と補正後の判別率の比較

混合方法	感情	補正前 [%]	補正後 [%]	改善度 [%]
全体	標準	65.8	72.7	6.9
	怒り	62.7	73.1	10.4
	悲しみ	56.9	60.4	3.5
	喜び	46.9	62.3	15.4
	平均	58.08	**67.12**	**9.04**

青字は補正により判別率が向上したことを表す．補正の結果，最低でも 3.4%の判別率の改善が得られた．特にもともと判別率の低かった喜びの感情では，15.4%と著しい改善が見られた．また，平均値でも約 9%の向上が見られ，この正規化手法の有効性が示された．

4.4　まとめ

本章では前章の判別実験の結果を受け，話者ごとの環境と男性話者，女性話者および全体の 3 通りの混合環境のそれぞれについて，特徴量の判別への寄与割合を標準化判別係数を用いて調査した．標準化判別係数の分布は話者ごとの環境では同程度型と多種突出型の 2 つのグループに分類することができ，ごく少数の特徴量のみが判別に寄与する前者よりも，まんべんなく多数の特徴量を用いて判別する後者のグループの方が高い判別率を得られることが分かった．

また，話者混合環境においては性別差によるひずみを吸収するための補正手法を導入した結果，すべての感情において 3.5 ~ 15.4%の判別率の改善が見られた．また平均値においても 9.04%の向上が見られ，提案した補正手法の有効性が確認された．

第5章

結論

5.1　本研究の成果

本研究の成果について述べる．

第2章では，これまで一般向けに利用できるデータベースが無かった感情音声を含んだ会話文のデータベースを構築した．また，このデータベースを用いて感情が大語彙連続音声認識エンジンに与える影響を連続音声認識実験により調査した．音素認識率を用いた評価の結果，感情の付加によって有意差を伴う認識率の低下がもたらされることを確認した．また，感情の付加による音素ごとの認識誤りの増加に一定の傾向があることを示した．

第3章では，感情音声から話者の感情を推定する方法について検討した．また基本周波数をベースとする9次元の特徴量を用いて判別分析を行い，提案した特徴量が各感情について2~8%程度，全体の平均でも0.58%の判別率の向上をもたらすことを示し，その有効性を確認した．

第4章では感情の判別にはどのような特徴量が有効であるのかを分析し，特徴量の性別によるひずみを補正する手法を提案し，各感情で3.5~15.4%，平均で9.04%と著しい判別性能の改善が得られることを示し，提案した正規化手法の有効性を確認した．

5.2　今後の課題

今後の課題としては以下のことが挙げられる．

第3章で行った判別実験では，話者ごとの環境では80~95% と実用上十分な判別率が得られているのに対し，話者混合環境では50~65%程度と不十分な判別率しか得られなかった．これは話者ごとの正規化を行っていないことが原因であると考えられる．第4章で正規化手法として性別差に着目したが，同様の相違は話者間にも存在すると考えられるため，

話者の相違をできるだけ相殺するような別の正規化手法について検討する必要がある．

また，今回は重永の研究から発展した 9 次元の特徴量を用いたが，他にも様々な特徴量が考えられる．判別率の向上には話者による相違に頑健な，別の特徴量を検討する必要があると考えられる．

また本研究では実際に判別システムを構築するところまでは行っておらず，システムを実装して多数の話者により評価を行うことも重要な点である．

謝辞

「主は私の羊飼い．私は，乏しいことがありません。
主は私を緑の牧場に伏させ，いこいの水のほとりに伴われます．」
詩篇 23 篇 1 ~ 2 節

本研究は愛する神様の助けなしには実現しませんでした．どうすればよいか分からずに悩んだときも，先が見えずに焦りがつのるときも，学会発表で緊張している最中も，あなたはいつも私と共にいて下さいました．最大の感謝をあなたに捧げます．

また，本研究を進めるにあたり，東北大学大学院工学研究科 阿曽弘具教授には，指導教官として多方面にわたり多大な御指導を頂きました．心から感謝します．

音声研究の分野においては東北大学大型計算機センター 牧野正三教授，鈴木基之助手，宇都宮大学工学部 森大毅助手より毎週の音声ゼミや学会等を通じて親身に御指導頂きました．心から感謝します.

本論文を纏めるにあたり，東北大学大電気通信研究所 矢野雅文教授に貴重なご意見を頂きました．深く感謝します．

感情音声の採取のために労して下さった，仙台福音自由教会の愛する兄弟姉妹の方々に心から感謝します．

最後に，本研究の元になるアイディアを与え，研究期間中も祈りをもって支えてくれた愛する両親と家族に感謝します．本研究に関係して下さったすべての方々の上に，神様の豊かな祝福がありますように．

参考文献

[1] Lawrence Rabiner,Biing-Hwang Juang, 古井貞熙,「音声認識の基礎 (上)(下)」, NTT アドバンステクノロジ株式会社 (1995).

[2] 中川聖一,「確率モデルによる音声認識」, 電子情報通信学会 (1988).

[3] 古井貞熙,「ディジタル音声処理」, 東海大学出版会 (1985).

[4] Lawrence Rabiner,R.W.Shafer, 鈴木久喜,「音声のディジタル信号処理」, コロナ社 (1983).

[5] 西村雅史, 伊東伸泰, 山崎一考, "単語を認識単位とした日本語の大語彙連続音声認識," 情処論, Vol.40, No.4, pp.1395-1403 (1999).

[6] 李晃伸, 河原達也, 堂下修司, "単語トレリスインデックスを用いた大語彙連続音声認エンジン JULIUS," 信学技報, SP98-3 (1998-04).

[7]「日本語ディクテーション基本ソフトウエアの開発」, http://winnie.kuis.kyoto-u.ac.jp/dictation/

[8]「大語彙連続音声認識デコーダ Julius」, http://winnie.kuis.kyoto-u.ac.jp/pub/julius/index.html

[9]「1999 年度 大語彙連続音声認識 評価実験結果」, http://winnie.kuis.kyoto-u.ac.jp/pub/julius/result99/

[10] 城風敏彦, 牧野正三, 城戸健一, "発話全体の連続性を考慮した基本周波数の抽出," 信学論 (A), Vol.J73-A, No.9, pp.1537-1539 (1990-09).

[11] 上床弘幸, 小林豊, 新美康永, "音声の感情表現の分析とモデル化," 信学技報, SP92-131 (1993-01).

[12] 川波弘道, 広瀬啓吉, “態度・感情音声における韻律的特徴の考察,” 信学技報, SP97-67 (1997-11).

[13] 重永実,“感情の判別分析からみた感情音声の特性 (III),” 信学技報,SP97-66 (1997-11).

[14] 森山剛, 斎藤英雄, 小沢慎治, “音声における感情表現語と感情表現パラメータの対応付け,” 信学技報, SP95-67 (1995-10).

[15] 浅井淑, 小沢慎治,“砂時計型ニューラルネットワークを利用した感情音声の分析・合成システム,” 信学技報, SP98-120 (1999-01).

[16] 重永実, 小川孝, 中尾光志,“単語音声による感情表現について,” 信学技報, SP95-15 (1995-05).

[17] 小池和仁, 斎藤博昭, 中西正和,“感情音声の合成,” 信学技報, NLC98-43,SP98-107 (1998-12).

[18] 有馬哲, 石村貞夫,「多変量解析のはなし」, 東京図書 (1987).

[19] 石村貞夫,「すぐわかる多変量解析」, 東京図書 (1987).

[20] 石村貞夫,「すぐわかる統計処理」, 東京図書 (1994).

[21] 青木繁伸,「Black-Box」,
http://aoki2.si.gunma-u.ac.jp/BlackBox/BlackBox.html

研究業績

学会発表等

1. 門谷信愛希, 阿曽弘具, 鈴木基之, 牧野正三，
   “音声に含まれる感情の判別に関する検討，”
   電子情報通信学会技術研究報告，NLC2000-34,SP2000-82 (2000-12)．