周波数帯域パワーとフォルマント周波数を用いた母音認識の初歩的研究

数理･情報システム学科
計算機科学講座
田中研究室
S013037　齋藤　寿樹

目次

1. 序論

現在、情報技術の分野において音声認識は音声ディクテーションシステムやカーナビゲーションシステムなどとして利用されている。音声認識を利用したこれらのシステムは話者による周波数の違いや残響音などにより正確な認識ができないことがある。また、声は話者によって感情や体調などの変化によって、同一話者であっても常に一定の音声が出るとは限らない。これらのことが音声認識の精度の低下の原因となっている[1]。

母音は声帯の振動による周期的なパルスで駆動して、声道で共振を起こさせて生成する。母音は子音に比べて通常長い継続時間長を持ち、スペクトルも比較的明確である。よって、母音は通常容易にかつ確実に認識されることができるので、人間と機械の両者にとって音声認識では重要な役割を果たす[2]。

音声認識を行うためには、言語情報に対応する特徴量のみを音声波形から抽出しなければならない。現在知られている特徴量として、周波数帯域パワー、フォルマント周波数、ピッチ周波数、メル周波数などがあげられる[3]。今回の研究において、理論的にわかりやすい周波数帯域パワーと、母音の認識や生成などによく用いられるフォルマント周波数の両者の特徴量をとり、母音音声の判別を行った。

2. 本研究の流れ

本研究で行った流れを図 2-1 で示す。

まず、教師データとなる音声を録音し、その音声ファイルを Matlab で読み込ませる。そして、そのデータを直流電圧のノイズの除去を行い、音声区間を検出する。音声を切り出してから、特徴量を抽出し、その特徴量の平均と分散を計算する。その教師データを元に、次に比較データを教師データと同様に特徴量を抽出し、教師データの平均と分散から入力された比較データが『あ』～『お』のどの音声であるかを分類した。

特徴量としては、周波数帯域パワーとフォルマント周波数を用いた。周波数帯域パワーではスペクトル分析を行い、またフォルマント周波数では LPC 分析により抽出を行った。

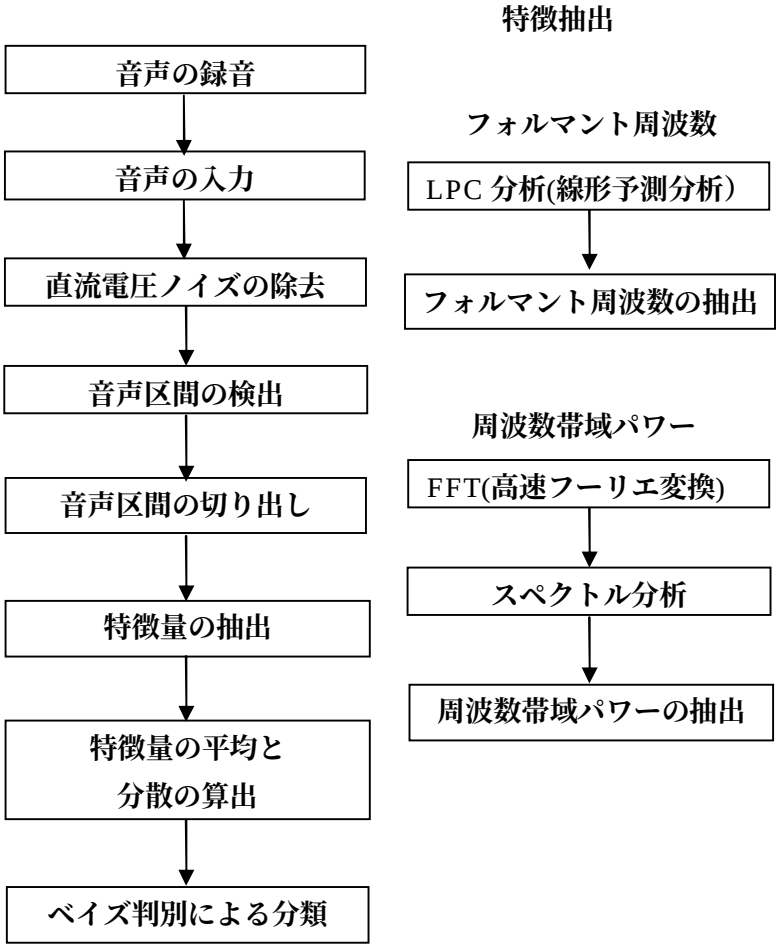


図 2-1：研究の流れ図

3. 使用した音声ファイル

3.1　音声の録音

Windows 標準のサウンドレコーダーを用いて、男性 6 人の声を wav ファイルとして録音した。声のとり方は 6 人全員に『あ』ならば『あ』を 20 回間隔をおいて録音した。これを教師データ用と比較データ用の 2 セット『あ』～『お』を録音した。サンプリング周波数はサウンドレコーダーでデフォルトの 22.050kHz において録音する。その録音した例を図 3-1 に示す。

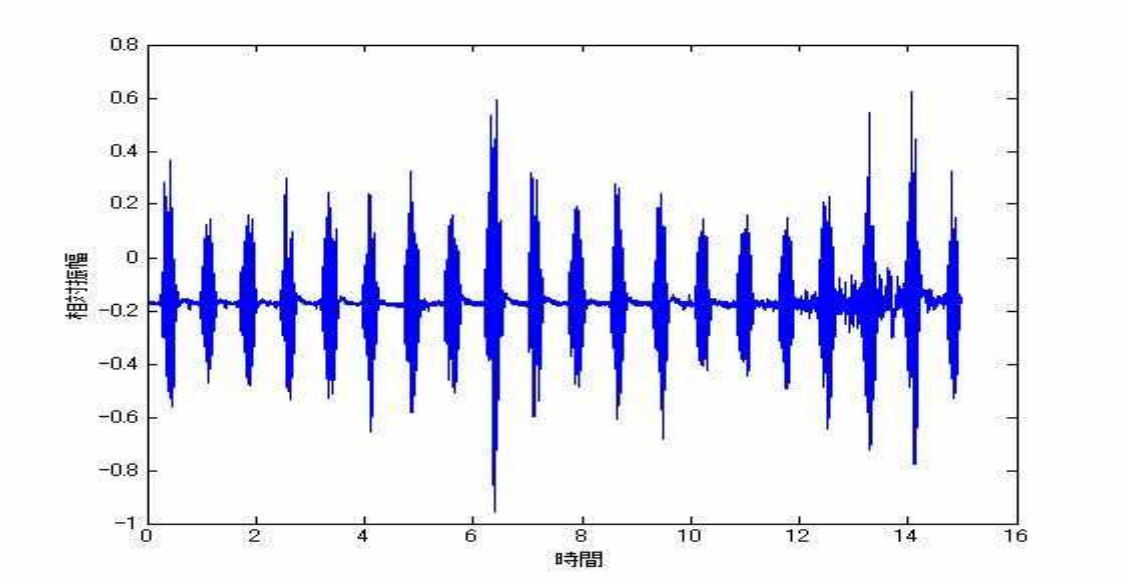


図 3-1：音声ファイルの波形図

3.2　ノイズの除去

音声処理を行うために、図 3-1 の音声区間を検出する必要がある。しかし、最初に録音機材の影響により、直流成分のノイズが存在している。これは無音区間であっても、振幅が 0 にならず、マイナスに傾いている。これはマイクなどの録音機材が電気的な影響を受けてしまっているために現れる。このようなノイズがある場合、音声処理を行う際に影響を及ぼす[1]。そこで、無音ファイルを作成し、そのときのノイズの平均をとり、録音した音声ファイルのデータからノイズの平均値を引いて、ノイズを除去した。ノイズを除去した図を図 3-2 に示す。二つの波形図の赤い線付近を見ると、上の図では無音区間が約-0.17 であったが、下の図では無音区間が約 0 に近い値になっていることがわかる。

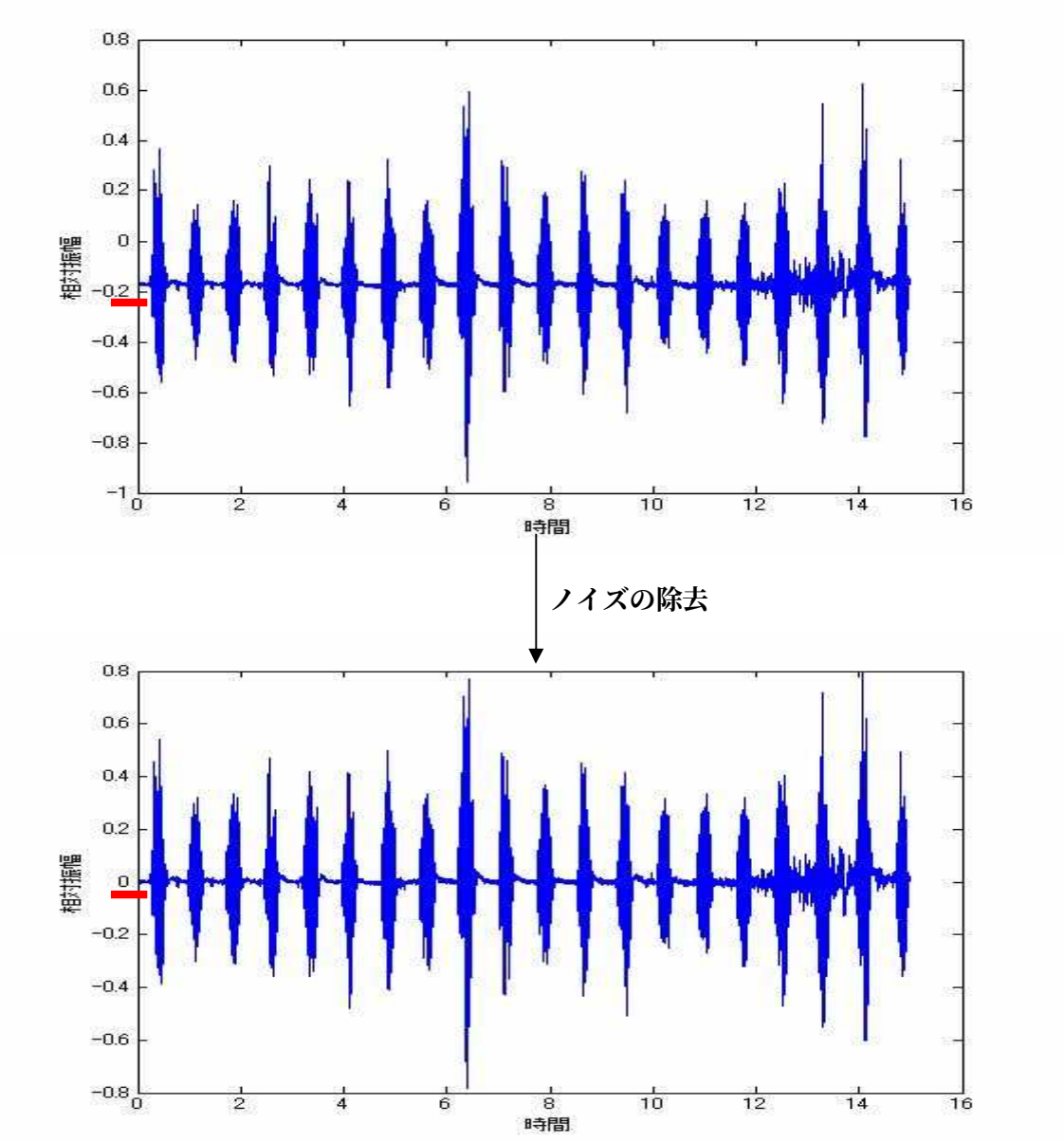


図 3-2：音声ファイルのノイズ除去

3.3　音声区間の検出

このノイズ除去を行った波形から、しきい値をある値に定め、区間のはじめとなる部分の検出を行えると考えた。しかし、『い』や『う』といった声は少し他の声よりも出しづらく、自然と音量が小さくなってしまう。また、元から声が小さい人やマイクとの距離などといった要因から、音声区間であってもしきい値を満たさない場合がある。それを解決するために、データの値に掛けて最大値が 1 となるような定数を求め、その定数倍をすべてのデータに掛け合わせて音量を増幅した。増幅した図を図 3-3 に示す。

増幅したデータとしきい値から、切り出し開始点を探し出す。そして、切り出し開始点からデータを 2048 ポイント抜き出して音声の切り出しを行う。

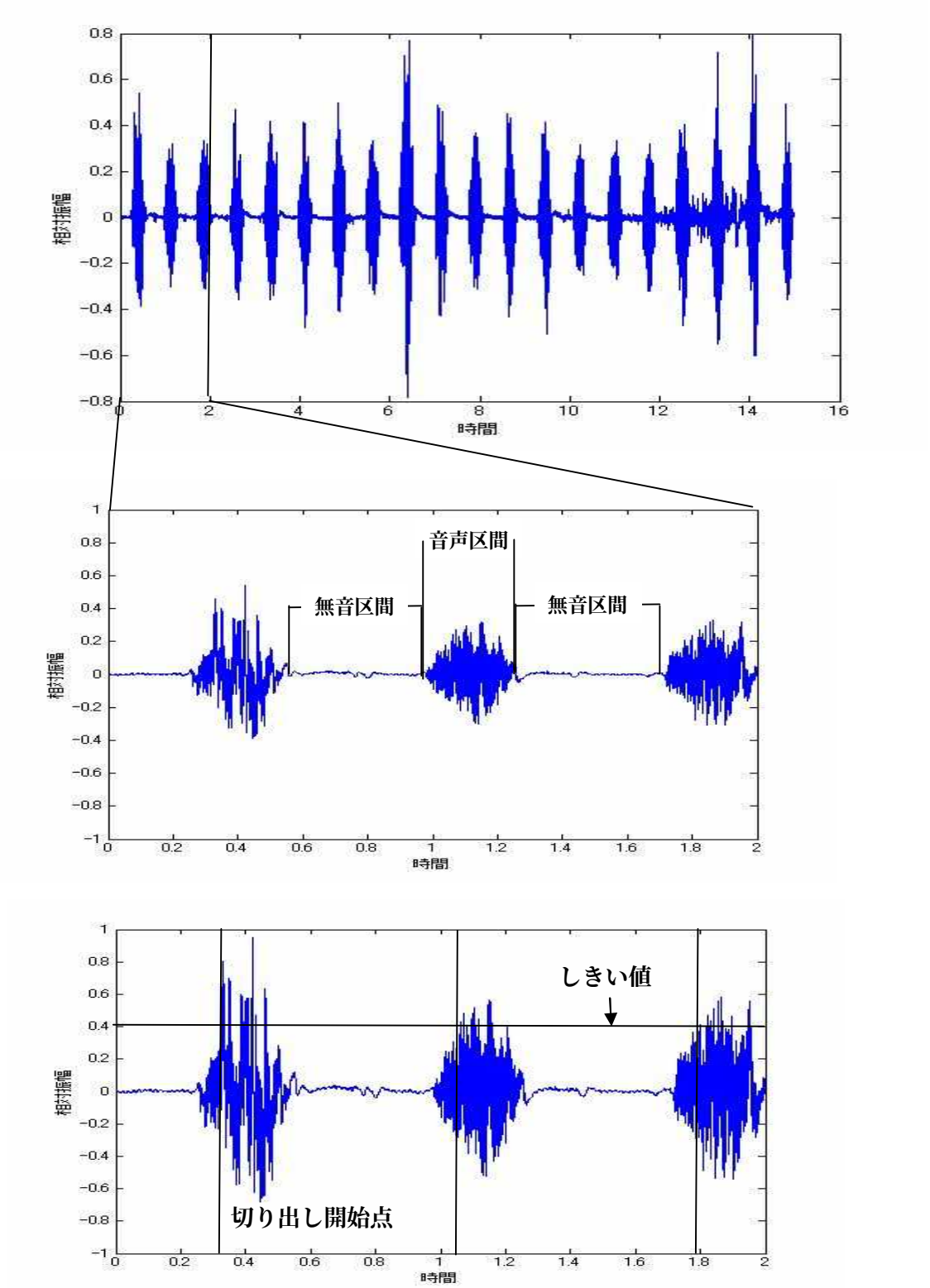


図 3-3：切り出し開始点の検出

4. 特徴抽出のための音声処理

切り出された音声から特徴量を抽出しなければならない。周波数帯域パワーではスペクトル分析を、フォルマント周波数では LPC 分析を行う。この章ではその分析手法について述べる。

4.1 高速フーリエ変換(FFT : Fast Fourier Transform)

離散フーリエ変換(DFT : Discrete Fourier Transform)は定義式として、

$$X_k = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn} \quad (k = 0,1,2,\cdots,N\text{-}1) \qquad (3・1)$$

が与えられる[2]。今回この離散フーリエ変換の計算量を減らした FFT を行い、$|X_k|^2$ を求めてスペクトル分析を行い、周波数ごとの音の大きさ(dB)を求める。

4.2 スペクトル分析

母音の認識を行う場合、5 母音はスペクトルが大きく異なることが一般的に知られている[2]。このことから、母音はスペクトルの形から判別されることが多い。このことから、本研究では音声のスペクトルから特徴量の抽出を行った。

短時間スペクトル分析は音声から連続する数十 ms 程度の時間長の信号区間を切り出し、切り出された信号に対して高速フーリエ変換を行い、その 2 乗値であるパワースペクトルが注目すべきスペクトル表現である。音声波形とそれに対応するパワースペクトルを図 4-1 に示す。

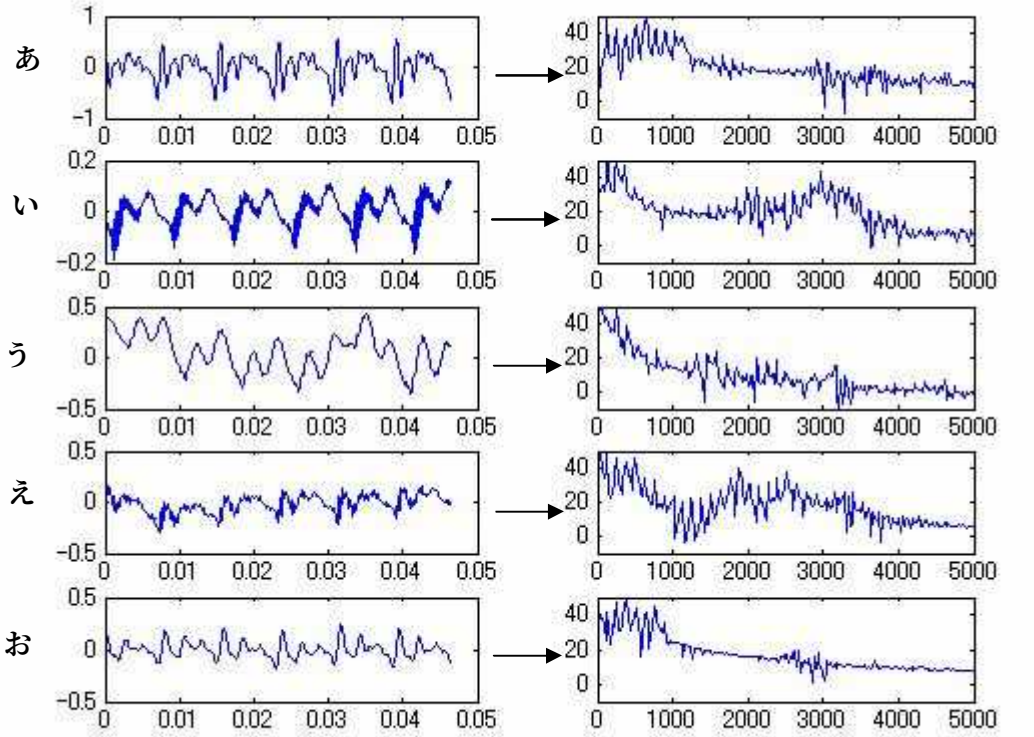


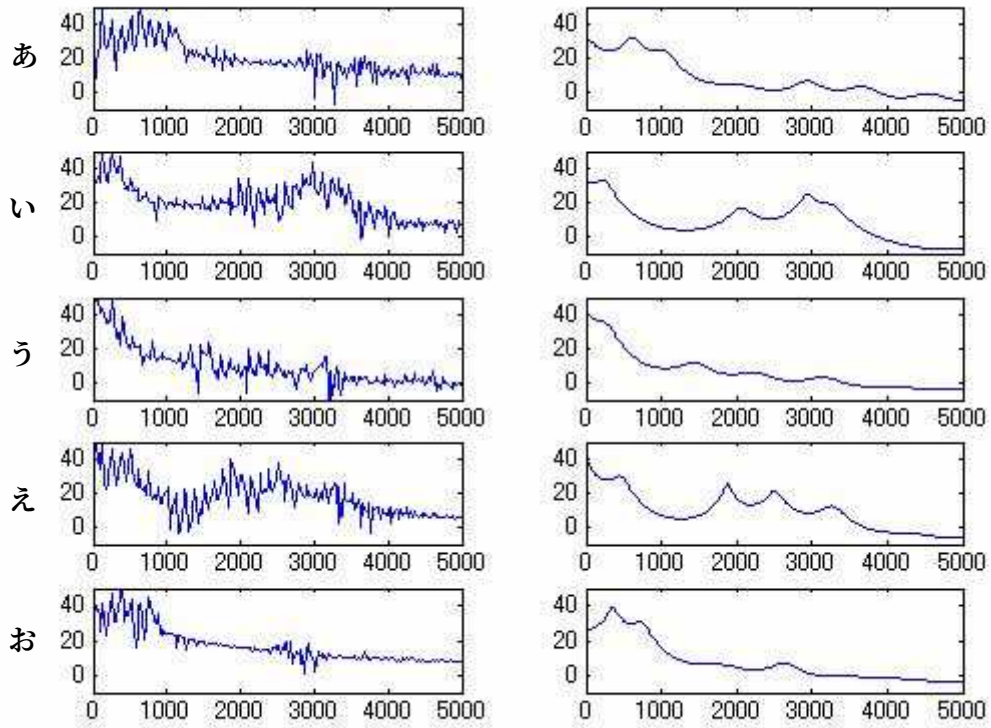
図 4-1：音声波形とそれに対応するパワースペクトル

4.3 線形予測分析(LPC : Liner Predictive Coding)

LPC 分析は音声のスペクトル包絡情報を効率よく表すことのできる分析手法である[3]。切り出された信号 x(n) から LPC 分析を行う場合、その信号の近似式としては、

$$\hat{x}(n) = -a(2)x(n-1) - a(3)x(n-2) - \ldots - a(p+1)x(n-p) \qquad (3・2)$$

で表されるとき、その誤差 $e(n) = x(n) - \hat{x}(n)$ が最小となる a(i) (i=0,1,…p)を求めることである。元のスペクトル波形とこの LPC 分析を行ったスペクトル包絡を表した図を以下に示す。



横軸：周波数[Hz] 縦軸：パワー[dB]　　横軸：周波数[Hz] 縦軸：パワー[dB]

図 4-2：パワースペクトルとそれに対応する LPC パワースペクトル

5. 特徴抽出

5.1 周波数帯域パワー

周波数帯域パワーはまず短時間スペクトル分析を行う。スペクトルからある周波数で間隔を分けて区切り、その周波数での大きさ(dB)の和によりひとつの特徴ベクトル $\boldsymbol{P}$=(P1，P2，P3)として求めた。周波数の分け方は

F1：500～1700[Hz]，　F2：1700～2600[Hz]，　F3：2600～3800[Hz]

としている。このように周波数帯域を分けているが、これは[1]から、このように分割した場合、よい結果が現れているため採用した。

信号 s(n)に FFT を行ったものを S(k)とすると、それぞれのベクトルは

$$P1=\sum_{k=500}^{1700}\left|S(k)\right|^2,\quad P2=\sum_{k=1700}^{2600}\left|S(k)\right|^2,\quad P3=\sum_{k=2600}^{3800}\left|S(k)\right|^2 \qquad (3・3)$$

と計算される。この 3 変量をひとつの特徴ベクトルとして、教師データの平均と分散を求め、分類を行った。

図 5-1 に『あ』と『い』のスペクトル波形を示す。この図において、F1 の領域では『あ』の振幅やパワーレベルが大きく、それに対し『い』は振幅やパワーレベルとも小さい。F2、F3 においても同様に波形の違いが見られる。また、このような波形の違いは他の母音にも見られる。

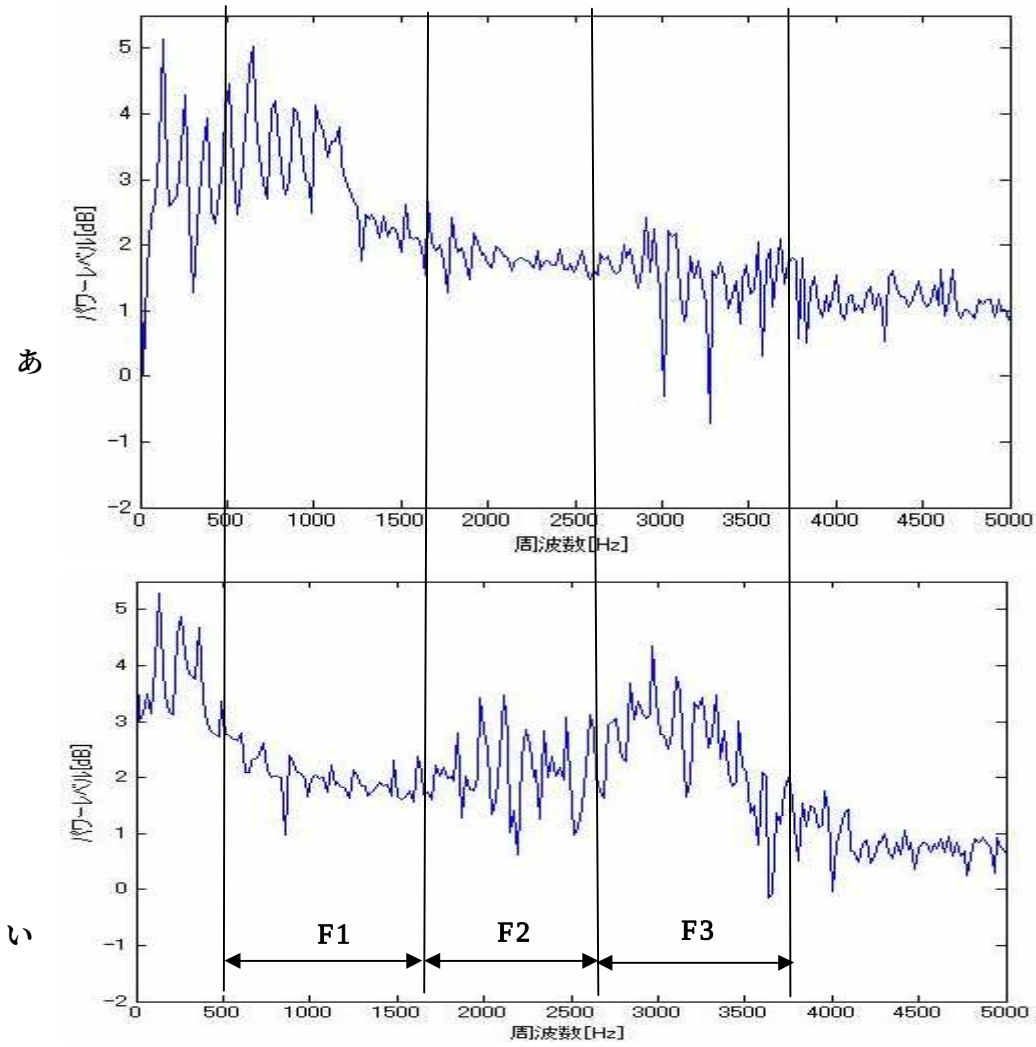


図 5-1：周波数帯域パワーの領域表示

5.2　フォルマント周波数

まず、切り出された信号に対して、LPC 分析により伝達関数を規定するパラメータである線形予測係数を決定し、信号のスペクトルの概形を決定する。LPC 分析を行った包絡に沿った曲線を図 5-2 に示す。そして、包絡に沿った曲線がいくつかの周波数でピークを持つ。そのピークはフォルマントと呼ばれ、低いものから第 1 フォルマント、第 2 フォルマント、…と呼ばれる。本研究では、一般的に『う』の声が第 4 フォルマントまで特徴を示すものとなっているとされているため[2]、第 4 フォルマントまで特徴ベクトルとして求めている。

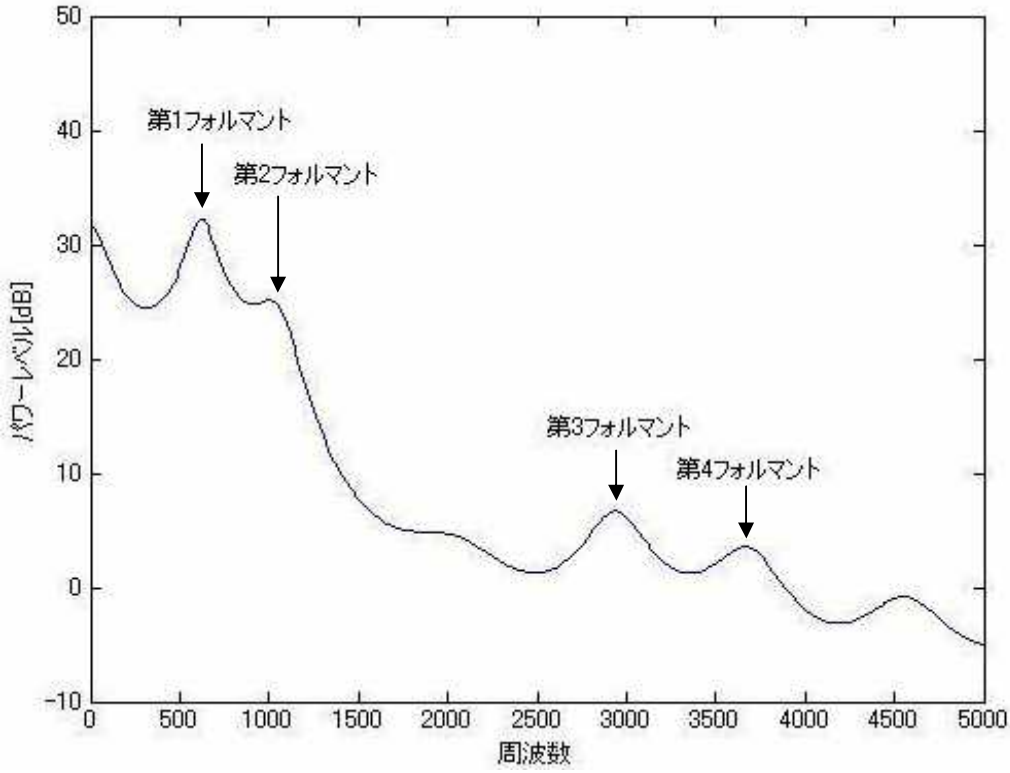


図 5-2：フォルマント周波数

5.3　その他の特徴量

本研究では、周波数帯域パワーとフォルマント周波数の 2 種類の特徴量での分類を行った。序論において、現在知られている今回行った 2 種類の特徴量以外の特徴量としてピッチ周波数やメル周波数などをあげた。

ピッチ周波数は、聴覚の上では、音の高さ、すなわち、いわゆるピッチに対応し、またピッチ周波数の緩やかな変化は、いわゆる抑揚となっている。したがって、ピッチ周波数の違いは、男女声の音色を区別したり、あるいは個人個人の音色の違いを聞き分けたりするために利用できることが考えられる。

メル周波数は、人間の聴覚の「周波数に対する非線形な特性」を考慮したものである。

6. 分類手法について

6.1　ベイズ判別法

今回抽出された特徴量はすべて正規分布に従うと考えられる。よって、多次元正規分布の確率密度関数は

$$P(x \mid c) = \frac{1}{(2\pi)^{M/2}\left|\Sigma_c\right|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_c)^T \Sigma^{-1}(x-\mu_c)\right\} \qquad (4・1)$$

と定義されている[5]。あるクラス c の平均 μ_c と分散共分散行列 Σ_c を求め、M は特徴ベクトル x の次元である。また、事後確率 $P(c \mid x)$ は

$$P(c \mid x) = \frac{P(x \mid c)\cdot P(c)}{\sum_{c=1}^{K} P(x \mid c)\cdot P(c)} \qquad (4・2)$$

と定義される。これは事前確率 $P(c)$ が既知である必要がある。この事後確率 $P(c \mid x)$ をクラス c の判別関数とし、それが最大となる c を判別結果とする判別方式のことをベイズ判別と呼ばれる。計算式において、右辺の分母は c によって変化することはない。よって、この識別は右辺の分子が最大であるとき、事後確率が c に関して最大であることと等価である。本研究では、式(4・2)の分子の値を求め、その値が最大であるか否かによって声が『あ』～『お』のどの声に分類されるか判別した。

7. 実験

7.1 実験環境

PC

CPU : Pentium4　2.40GHz

RAM : 1.00GB

Microphone :

ソフトウェア

Windows XP Professional SP1

Mathworks Matlab 7.0

Matlab Signal Processing Toolbox

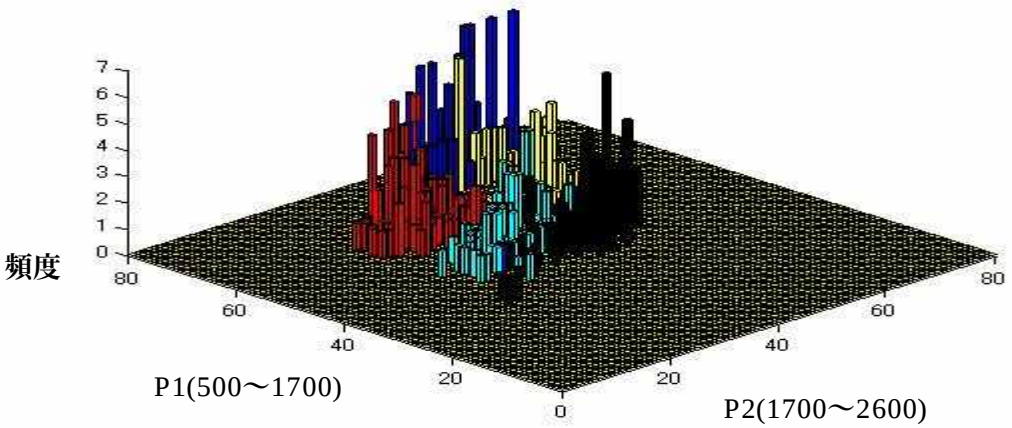
7.2 Matlab について

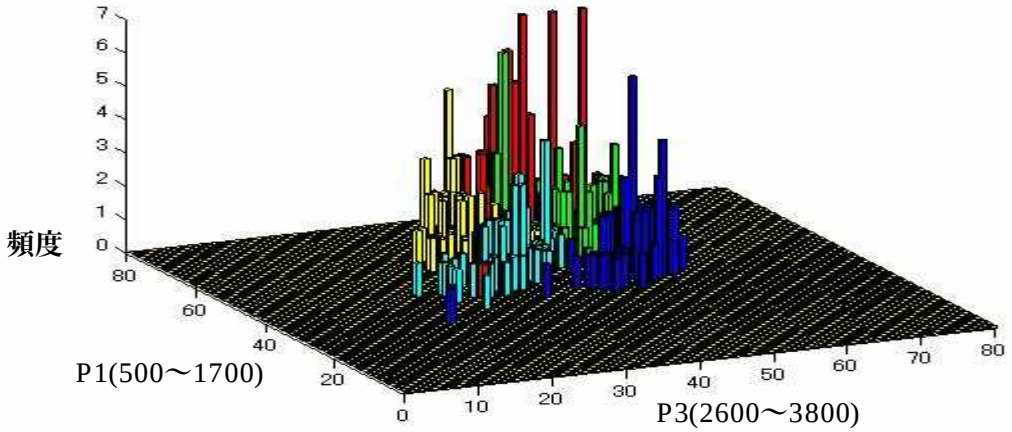
本研究では、Matlab を用いてプログラムを作成し、実験を行った。Matlab は Matrix Laboratory の意味を表している。その基本的な要素に次元を必要としない配列を持っているため、多くの技術計算の問題、特に行列とベクトルの形式を用いた問題を解くことを容易にしたツールである。計算、可視化、プログラミングなどを利用しやすい環境で統合している。M-ファイルとしてプログラムを作成する。

MATLAB の特色は、特定分野の解決策としてのツールボックス群があることである。今回利用した Signal Processing Toolbox では信号処理を扱う上で、大変よく使われる関数が M-ファイルの関数として登録されている。

7.3 教師データの作成

6 人の音声ファイルのひとつの音声ファイルから 15 個の音声を 2048 ポイントずつ取り出す。切り出された音声でそれぞれ処理を行い、特徴量を計算し、その特徴量の平均と分散共分散行列を作成する。平均と分散共分散行列を求めるのは、ベイズ判別にかけるためである。今回特徴量として抽出した周波数帯域パワーとフォルマント周波数の 2 次元のヒストグラムを図 7-1 に示す。





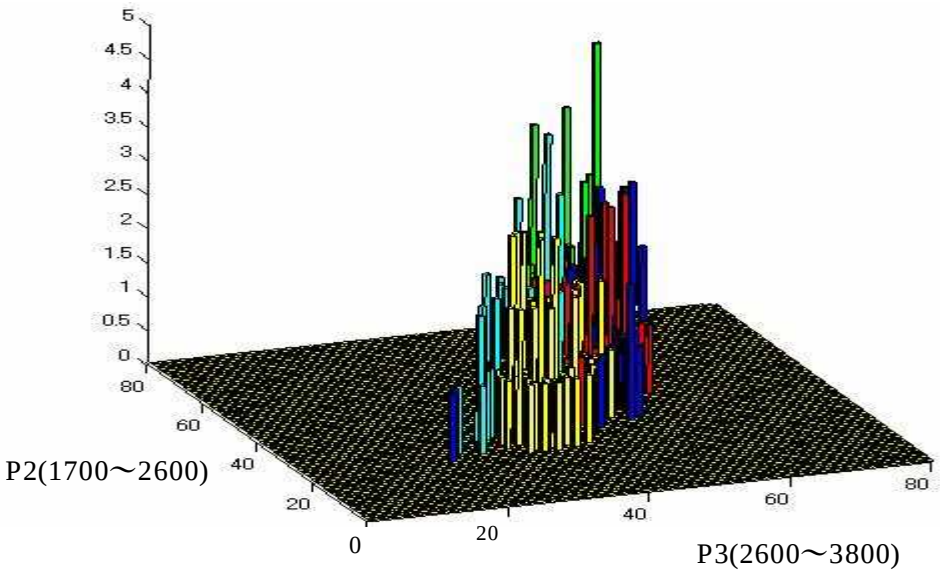


図 7-1：ヒストグラム(周波数帯域パワー)

図 7-1 のヒストグラムから、周波数帯域パワーが正規分布に従うことが見てとれる。この値から平均と分散を求め、その値を以下の表に示す。

表 7-1　周波数帯域パワー(平均)

	あ	い	う	え	お
P1(500～1700)	61.87192	43.57114	48.74451	57.81723	59.89041921
P2(1700～2600)	39.98619	46.01394	39.42635	51.1526	30.49964499
P3(2600～3800)	46.43378	50.97281	38.1217	49.57875	37.57610539

表 7-2　周波数帯域パワー(分散共分散行列)

あ

6.259034	1.633867	3.449736
1.633867	17.65661	12.61077
3.449736	12.61077	28.42607

い

7.866782	7.5354	4.964263
7.5354	39.67328	19.47299
4.964263	19.47299	30.47402

う

22.36154	23.11829	23.33349
23.11829	46.34095	33.65148
23.33349	33.65148	36.06944

え

13.04614	0.648588	-2.11829
0.648588	15.55112	18.7635
-2.11829	18.7635	27.71303

お

9.006189	4.527465	1.861255
4.527465	12.78709	12.14324
1.861255	12.14324	23.19746

これらの表を作成したデータの三次元散布図を図 7-2 に示す。散布図の記号の意味を下に記す。

＋：あ　　○：い　　＊：う　　×：え　　□：お

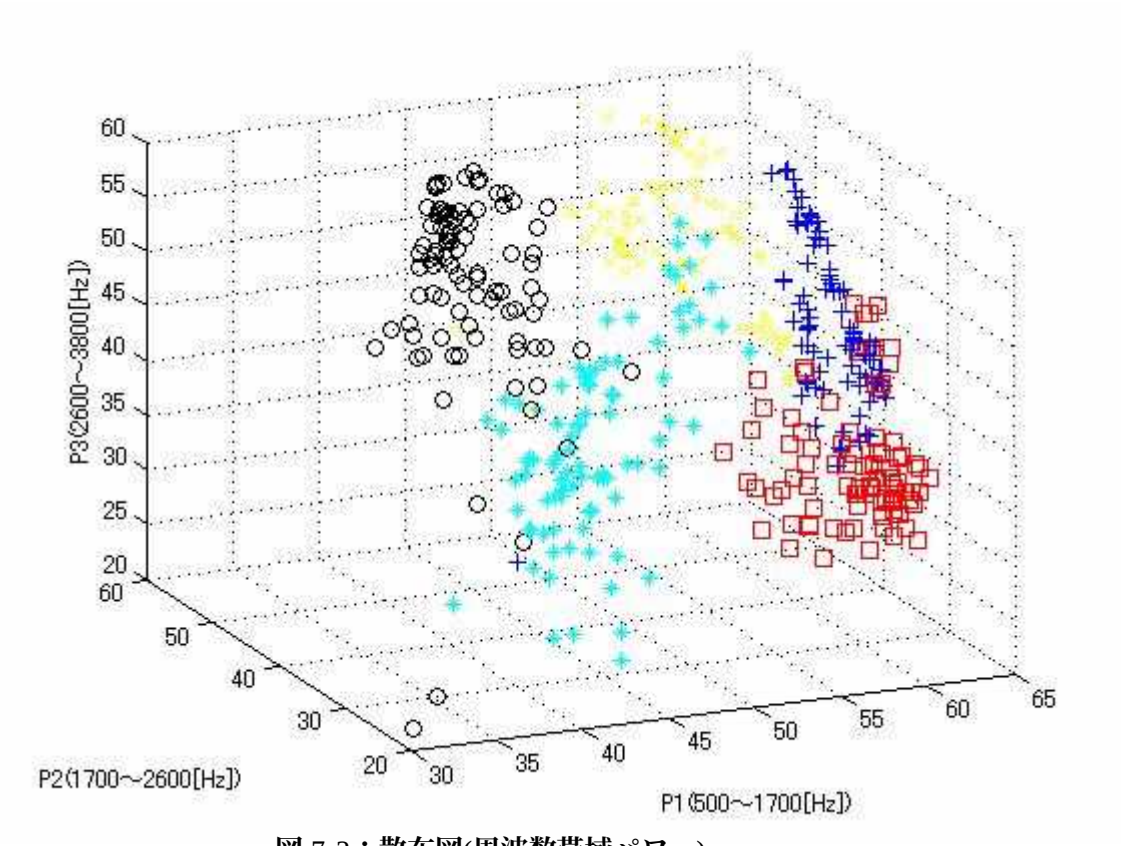
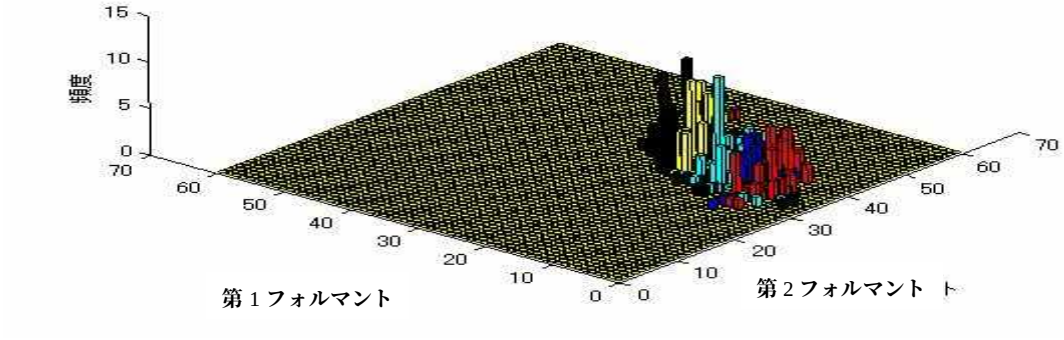
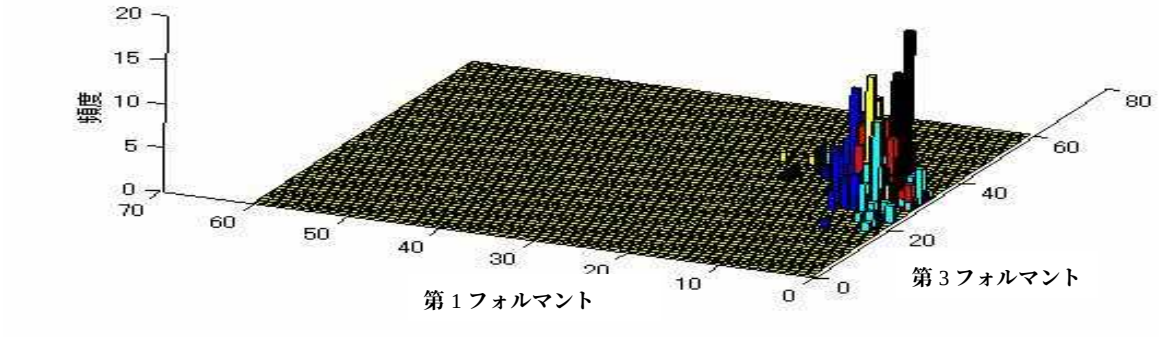


図 7-2：散布図(周波数帯域パワー)

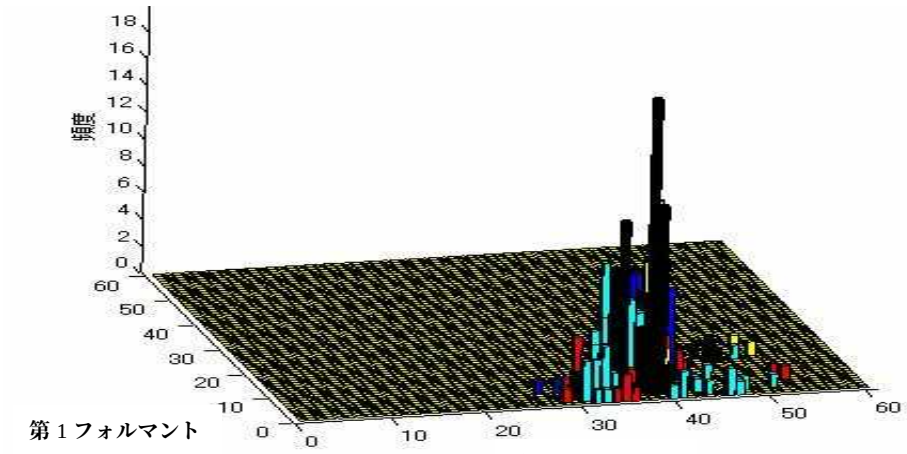
次にフォルマント周波数の 2 次元ヒストグラムを以下の図 7-３に示す。



第 1 フォルマント　　第 2 フォルマント



第 1 フォルマント　　第 3 フォルマント



第 1 フォルマント　　第 4 フォルマント

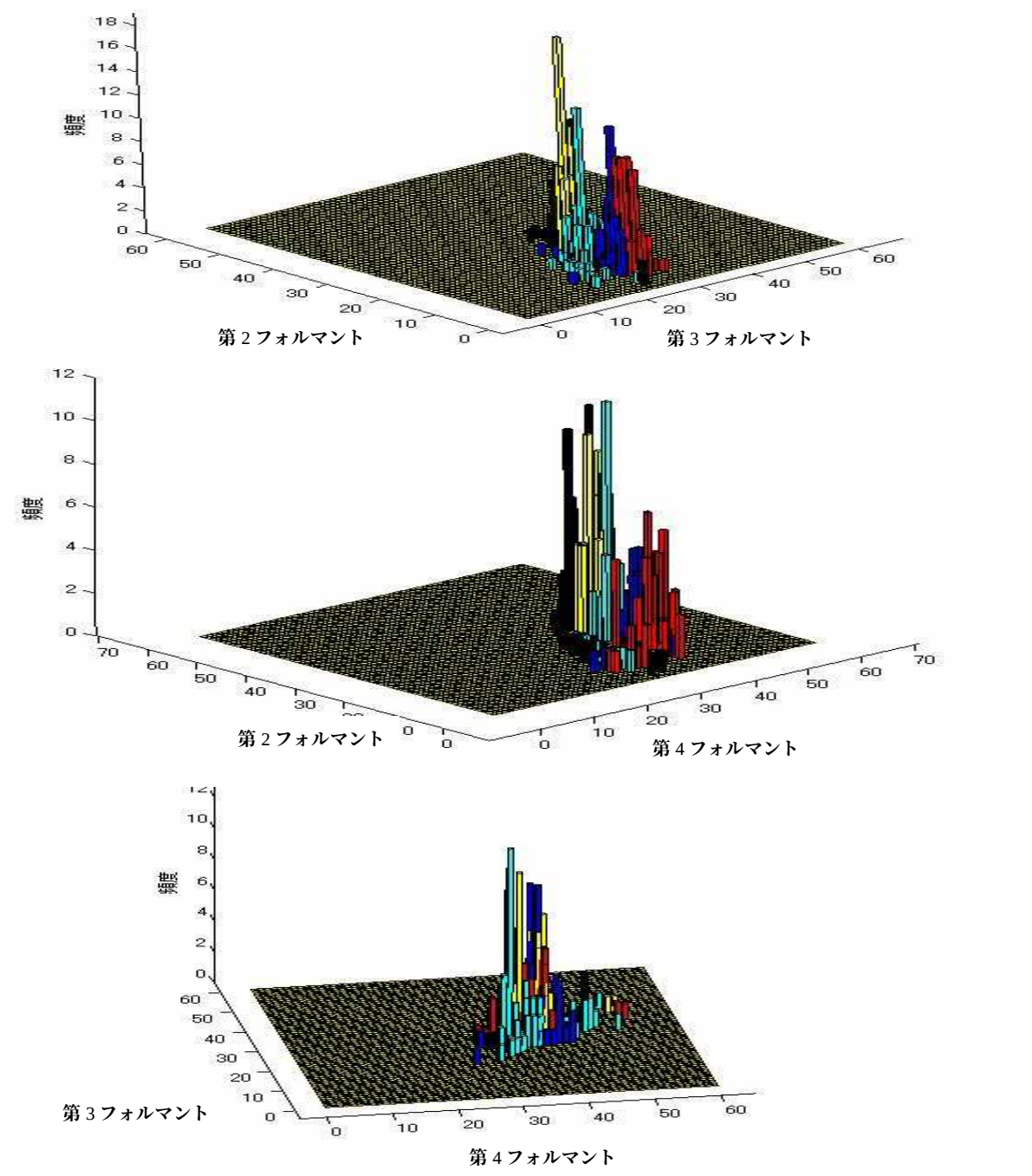


図 7-3：ヒストグラム(フォルマント周波数)

これらのヒストグラムから、フォルマント周波数が正規分布に従うことが見てとれる。この値から平均と分散を求め、その値を以下の表に示す。

表 7-3　フォルマント周波数(平均)

	あ	い	う	え	お
第 1 フォルマント	718.2567	338.3668	326.102	502.1525	496.5076
第 2 フォルマント	1145.933	2127.895	1435.103	1965.29	859.1324
第 3 フォルマント	2775.595	3036.409	2511.022	2803.668	2754.794
第 4 フォルマント	3758.458	3696.105	3400.181	3739.303	3476.7

表 7-4　フォルマント周波数(分散共分散行列)

あ

15670.74	6767.29	4067.857	7468.309
6767.29	60065.23	12063.95	30539.86
4067.857	12063.95	17958.41	15392.01
7468.309	30539.86	15392.01	88281.24

い

67445.43	26352.7	42340.44	60487.96
26352.7	271265.4	139560.5	119058.2
42340.44	139560.5	112563.6	103060.8
60487.96	119058.2	103060.8	136623.5

う

6653.701	7389.727	4775.589	14191.75
7389.727	57533.49	41131.21	71858
4775.589	41131.21	85229.66	70196.81
14191.75	71858	70196.81	167738.8

え

9099.147	3158.85	3370.907	1138.003
3158.85	23338.39	16702.51	16757.56
3370.907	16702.51	22471.56	16608.3
1138.003	16757.56	16608.3	37356.98

お

12981.08	10578.69	5394.052	6380.789
10578.69	23433.28	7373.52	14659.36
5394.052	7373.52	92184.87	55779.99
6380.789	14659.36	55779.99	134165.6

周波数帯域パワー同様に、これらの教師データを作成した三次元散布図を図 7-4 に示す。記号の種類は図 7-2 と同様である。

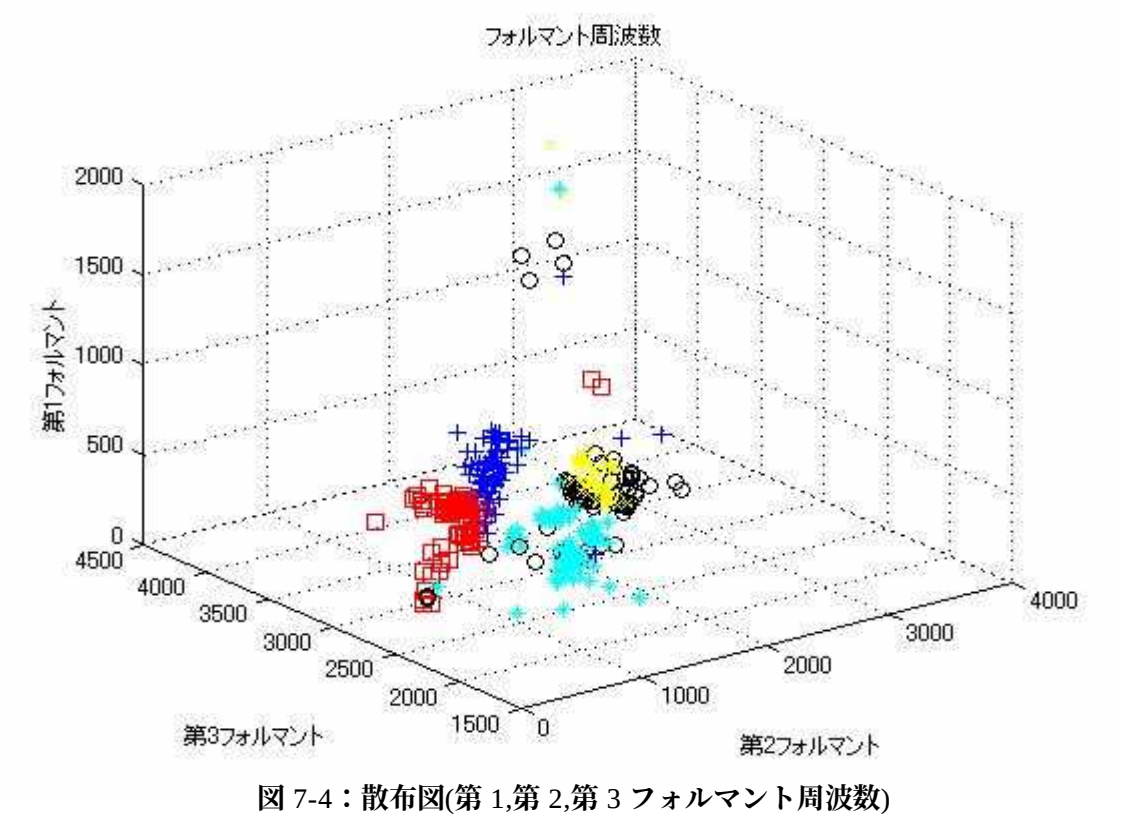


図 7-4：散布図(第 1,第 2,第 3 フォルマント周波数)

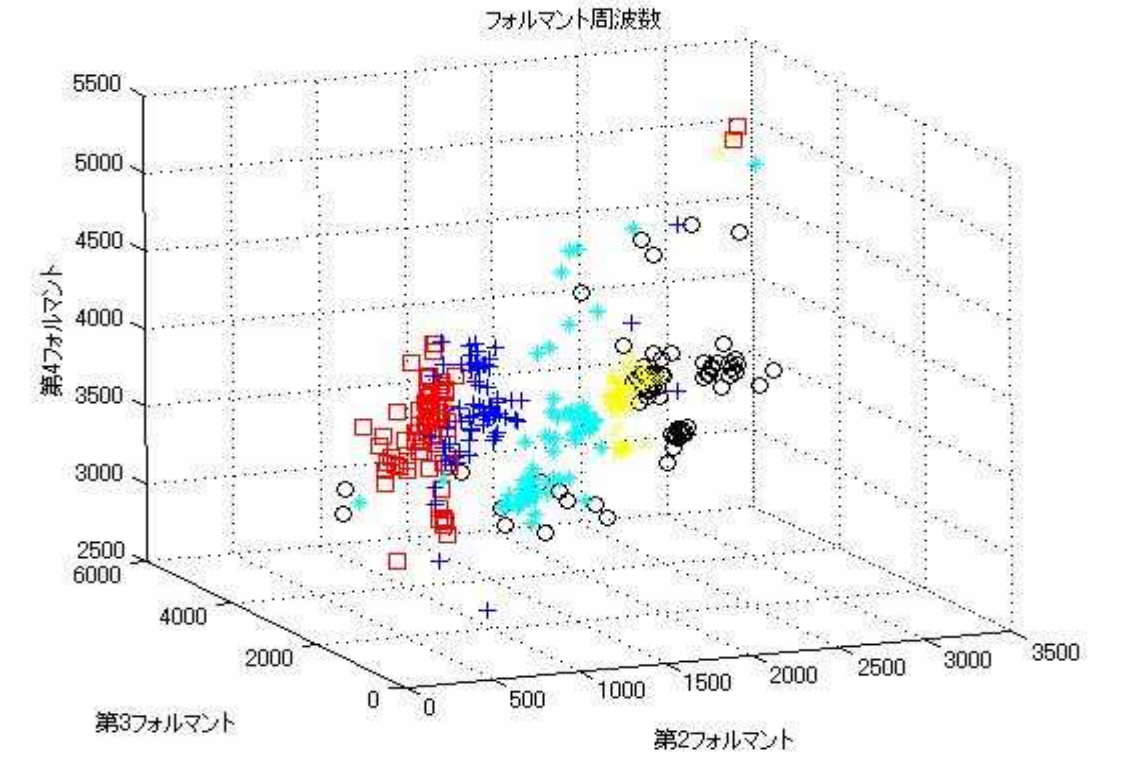


図 7-4：散布図(第 2,第 3,第 4 フォルマント周波数)

7.4　分類

録音した学習外データの切り出された音声を処理し、特徴量を算出する。そして、そのデータと教師データから入力された音声が『あ』～『お』のどの声になるのかをベイズ判別にかけて分類を行い、その分類が適切かどうかを確かめた。6 人の学習外音声データのひとつの音声データに対して 15 回(計 90 回)母音の分類を行う。2 次元でベイズ判別により求めた分類の境界線と学習外データの散布図を表示したときの周波数帯域パワー（図 7-4）とフォルマント周波数(図 7-5)の二つを作成した。

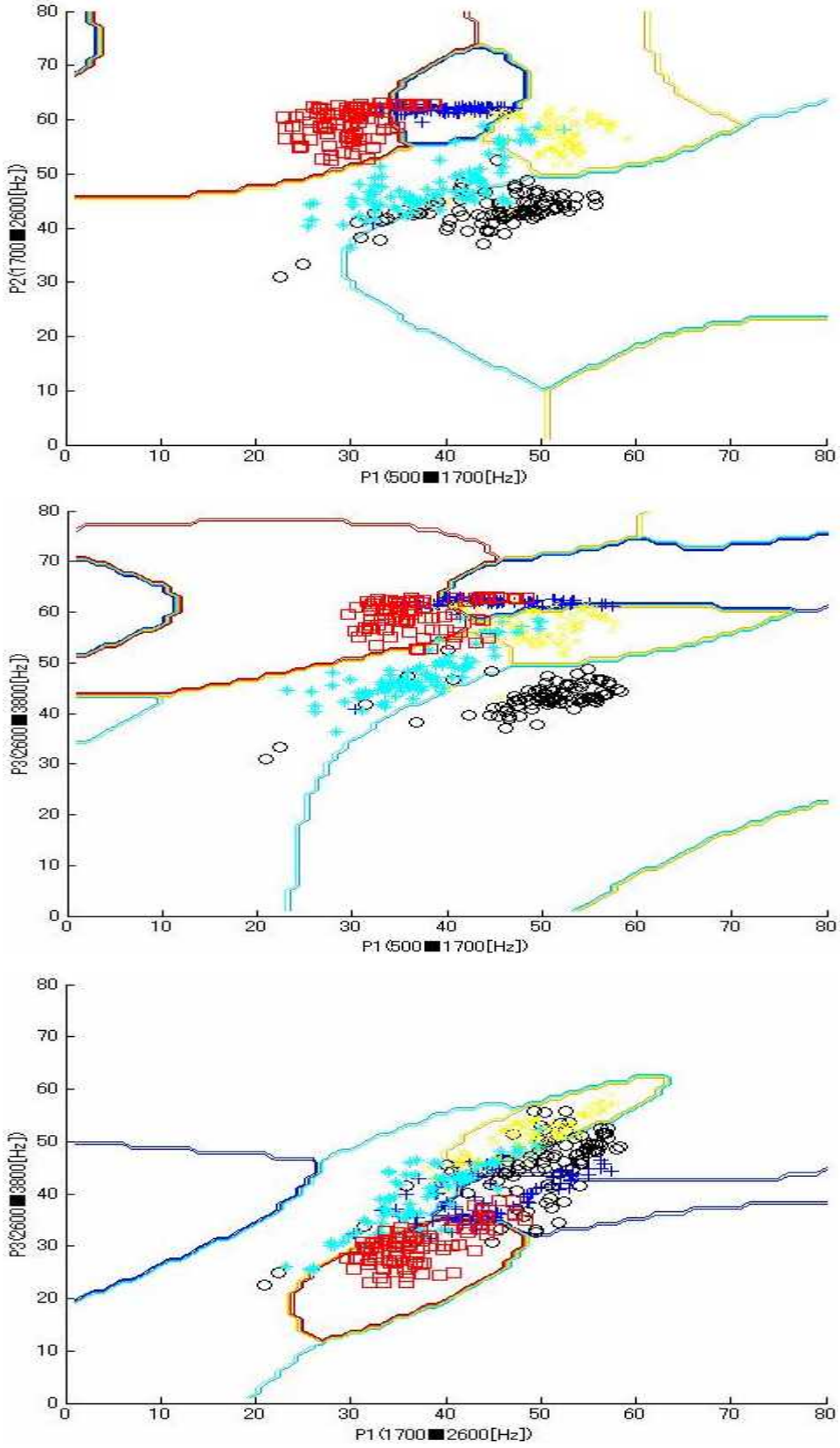


図 7-5：境界面(周波数帯域パワー)

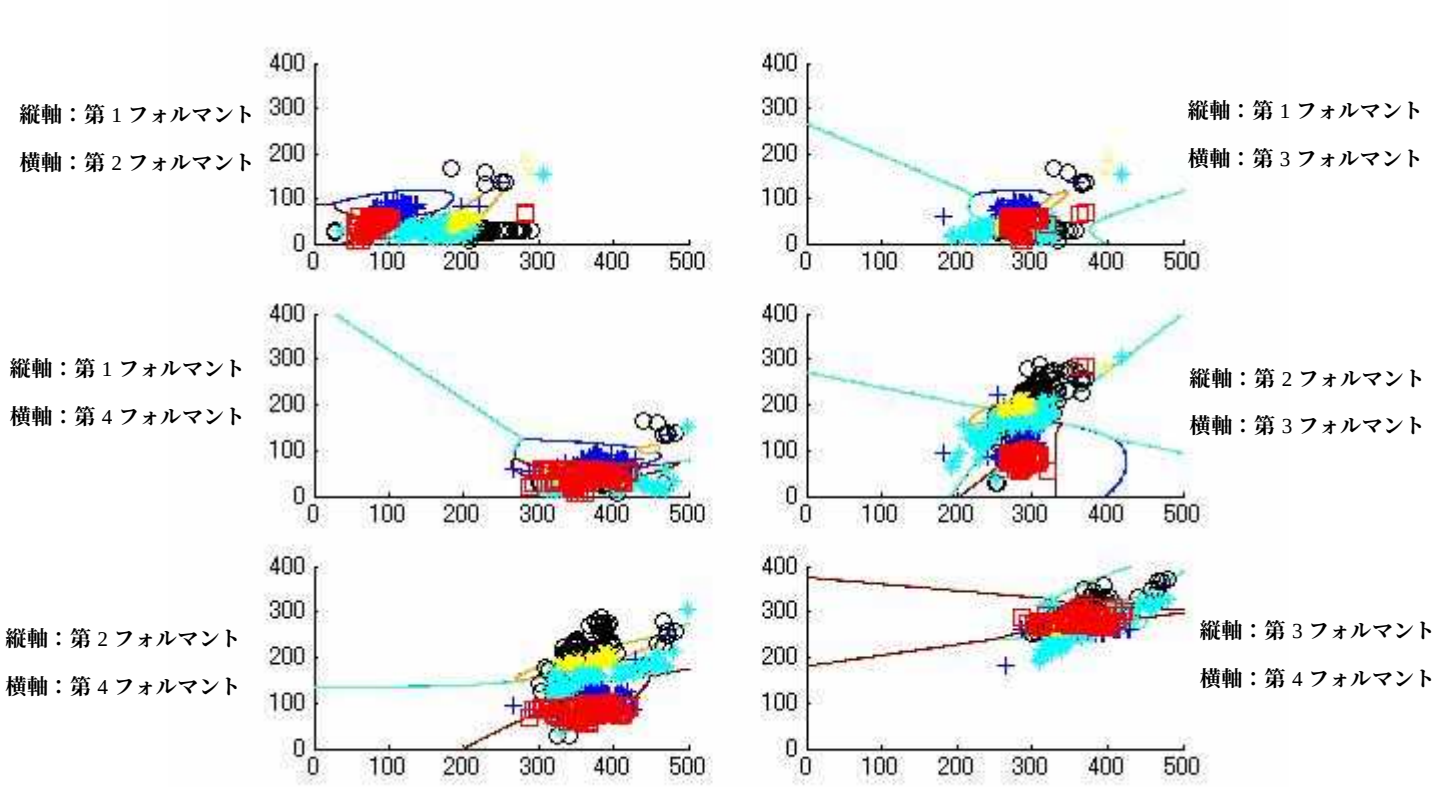


図 7-6：境界面(フォルマント周波数)

7.5　結果

周波数帯域パワーを分類した結果は表 6-5 に示す。この表において、本来の分類が一つの声に対して 90 個(比較データ数)存在し、その 90 個の声が『あ』～『お』のどの声に分類したかを示している。例えば、『あ』の声を『あ』と分類したのは 90 個の比較データのうち 72 個ということになり、『あ』の声を違う声と認識したのが 18 個となっている。『あ』を『あ』と認識した場合、その分類は成功しているのでこの場合の正当率は 72/90 となる。

表 7-5　エラーマトリクス(周波数帯域パワー)

		認識した分類					
		あ	い	う	え	お	正当率
本来の分類	あ	72	0	2	1	15	0.800
	い	0	87	2	1	0	0.967
	う	0	6	76	7	1	0.844
	え	1	0	3	86	0	0.956
	お	5	0	0	0	85	0.944
							0.902

この表から、『あ』を『お』に、また『お』を『あ』に誤認識することが多かった。また『う』が『い』や『え』に誤認識されることが多かった。

フォルマント周波数を分類した結果は以下の表に示す。

表 7-6　エラーマトリクス(フォルマント周波数)

		認識した分類					
		あ	い	う	え	お	正当率
本来の分類	あ	83	0	5	0	2	0.922
	い	0	80	8	0	2	0.889
	う	0	2	87	0	1	0.967
	え	0	1	3	86	0	0.956
	お	1	0	8	0	81	0.900
							0.927

この表から、『あ』→『う』、『い』→『う』、『お』→『う』に誤認識することが多かった。これは、周波数帯域パワーのときの誤りとは傾向が異なっている。

7.6　結論

周波数帯域パワーとフォルマント周波数の両者において、約 90%を超える確率で分類することができた。しかし、どちらにおいても誤りが存在している。両者の誤り方は図 7-5、図 7-6 の境界面と表 7-5、7-6 を見たとき、境界線で隣り合っている音声に誤りが多くなっていることがわかる。これは、周波数帯域パワーでは教師データがすでに『あ』『お』や『い』『う』『え』で近い値をとっているため、比較データにおいても誤認識していると考えられる。同様のことがフォルマント周波数でも言える。

本研究では、母音の判別を行ったが、実際に人が話をするときには子音含んだ言葉で行う。従って、母音だけでなく子音の判別も行えるようにする必要がある。また、本研究で用いた周波数帯域パワーとフォルマント周波数以外の特徴量での判別やそれらを組み合わせた特徴ベクトルなどを用いて判別し、認識率の向上を行うことを今後の課題とする。

謝辞

本研究において最後まで熱心な御指導をしていただきました田中章司郎教授には、心より御礼申し上げます。本研究室所属の大学院生である長田さん、喜代吉さんのさまざまな助言をしていただき感謝しております。同研究室の学部生の 6 名には声の提供、またさまざまな場面での助言をしていただきました。深く御礼申し上げます。なお、本論文、本研究で作成したプログラム及びデータ、ならびに関連する発表資料等の知的財産権を、本研究の指導教官である田中章司郎教授に譲渡致します。

参考文献・サイト

[1] 木下敦子、杉井大介
平成 14 年度卒業論文　周波数帯域パワーに基づいた音声認識の研究　東京電機大学

[2] 鹿野清宏、中村哲、伊勢史郎　共著
ディジタル信号処理シリーズ第 5 巻　音声・音情報のディジタル信号処理　昭晃堂

[3] 鹿野清宏、伊藤克亘、河原達也、武田一哉、山本幹雄　編著
IT Text 音声認識システム　オーム社

[4] 長尾真、宇津呂武仁、匂坂芳典、井口征士、片寄晴弘　執筆
マルチメディア情報学 4　文字と音の情報処理　岩波書店

[5] 英樹、津田宏治、村田昇
統計科学のフロンティア 6　パターン認識と学習の統計学～新しい概念と手法～

[6]http://www.hanlab.ee.kagu.tus.ac.jp/~norimatu/research5.html

[7]http://www.catnet.ne.jp/triceps/cdr/sample/cd002.pdf

[8tp://www.phon.ucl.ac.uk/courses/spsci/matlab/lect10.html