JAIST Repository
https://dspace.jaist.ac.jp/

Title	音声のスペクトル包絡における個人性に関する研究
Author(s)	北村，達也
Citation	
Issue Date	1997-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/838
Rights	
Description	Supervisor:赤木 正人，情報科学研究科，博士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

博 士 論 文

音声のスペクトル包絡におけ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

北村 達也

1997 年 1 月 16 日

Copyright ©1997 by Tatsuya Kitamura

目 次

1 序論 **1**

1.1 はじめに 2

1.2 用語の定義 3

1.3 研究の背景 4

1.3.1 個人性を表す物理量に関する研究 4

1.3.2 話者認識に適した特徴パラメータに関する研究 7

1.3.3 声質変換に関する研究 8

1.4 本研究の目的 8

1.5 本論文の構成 9

2 単母音のスペクトル包絡における個人性の存在確認 **12**

2.1 まえがき 13

2.2 実験 2-1 既知話者の音声データを用いた聴取実験 13

2.2.1 目的 13

2.2.2 実験条件 14

2.2.3 実験結果と考察 17

2.3 実験 2-2 未知話者の音声データを用いた聴取実験 20

2.3.1 目的 20

2.3.2 実験条件 20

2.3.3 実験結果と考察 22

2.4 むすび 23

3 個人性が顕著に現れるスペクトル包絡の帯域の検討 **25**

3.1 まえがき 26
3.2 スペクトル包絡の分析 27
3.2.1 目的 27
3.2.2 音声データとスペクトル包絡の計算方法 27
3.2.3 話者間の分散と音韻間の分散 28
3.2.4 分析結果と考察 28
3.3 実験 3-1 帯域毎の個人性の分布 30
3.3.1 目的 30
3.3.2 実験条件 30
3.3.3 実験結果と考察 34
3.4 実験 3-2 個人性と音韻の特徴が顕著に現れる帯域の調査 41
3.4.1 目的 41
3.4.2 実験条件 41
3.4.3 実験結果と考察 44
3.5 実験 3-3 個人性が顕著に現れる帯域の調査 47
3.5.1 目的 47
3.5.2 実験条件 48
3.5.3 予備実験の結果 50
3.5.4 スペクトル包絡の低域と高域の音韻が等しい場合 51
3.5.5 スペクトル包絡の低域と高域の音韻が異なる場合 53
3.6 むすび 54

4 話者識別に寄与する物理量の検討 55
4.1 まえがき 56
4.2 実験 4-1 スペクトル包絡の微細構造と話者識別の関係 57
4.2.1 目的 57
4.2.2 実験条件 57
4.2.3 実験結果と考察 59
4.3 実験 4-2 スペクトル包絡のピークとディップが話者識別に与える影響の検討 60
4.3.1 目的 60
4.3.2 実験条件 62

4.3.3 実験結果と考察 66
4.4 むすび 68

5 スペクトルピークに着目した個人性が顕著に現れる帯域の検討 69
5.1 まえがき 70
5.2 実験 5-1 F3 を含む帯域と個人性の関係 71
5.2.1 目的 71
5.2.2 実験条件 71
5.2.3 実験結果と考察 74
5.3 実験 5-2 20 ERB rate 付近のピーク以上の帯域と個人性の関係 76
5.3.1 目的 76
5.3.2 実験条件 78
5.3.3 実験結果と考察 79
5.4 むすび 83

6 連続音声中の母音における個人性に関する検討 84
6.1 まえがき 85
6.2 実験 6 連続音声中の母音のスペクトル包絡と基本周波数における個人性の検討 85
6.2.1 目的 85
6.2.2 実験条件 86
6.2.3 実験結果と考察 89
6.3 むすび 91

7 単純類似度法による話者認識に適した帯域の推定 92
7.1 まえがき 93
7.2 実験方法 93
7.2.1 音声データ 93
7.2.2 標準パターンとテストパターンの作成方法 94
7.2.3 認識方法 94
7.2.4 AD 値 94
7.3 実験 7-1 3 帯域の比較 96

7.3.1 目的 . . . 96
7.3.2 使用データ . . . 96
7.3.3 実験結果と考察 . . . 96
7.4 実験 7-2 話者認識に適した帯域の調査 . . . 99
7.4.1 目的 . . . 99
7.4.2 使用データ . . . 99
7.4.3 実験結果と考察 . . . 99
7.5 実験 7-3 男女各 10 名の音声データを用いた実験 . . . 100
7.5.1 目的 . . . 100
7.5.2 使用データ . . . 100
7.5.3 実験結果と考察 . . . 100
7.6 実験 7-4 標準パターンとテストパターンの音韻が異なる場合 . . . 102
7.6.1 目的 . . . 102
7.6.2 使用データ . . . 106
7.6.3 実験結果と考察 . . . 106
7.7 むすび . . . 107

8 全体考察 **110**

9 結論 **114**
9.1 本論文で明らかにされたことの要約 . . . 115
9.2 今後の課題 . . . 115

参考文献 **119**

付録 聴取実験に用いた回答用紙 **127**

第 1 章

序論

1.1 はじめに

音声には話者の特徴すなわち個人性が含まれており、人間はこれを話者の識別のみならず音声の認識にも利用している。人間は、個人性を利用して話者に適応することにより話者毎に特徴の異なる音声を正確に認識することができる [加藤 88]。

最近、Bregman による聴覚の情景分析 (Auditory Scene Analysis) の概念が注目を集めているが [Bregman 90]、複数の話者が同時に話している状況で特定の話者の音声を聞き分ける能力 (カクテルパーティ効果) にも個人性が重要な役割を果たしている。また、音声から話者の年齢や、場合によっては人格特徴や職業まで正確に判定できるという報告もある [Knapp 72], [鈴木 85]。

このように、個人性は音声によるコミュニケーションにおいて非常に重要な意味を持っている。それにもかかわらず、個人性知覚過程すなわち人間が音声中のどのような物理量をどのような処理によって個人性として知覚するのか、ということに関してほとんど明らかになっていないのが現状である。

個人性知覚に利用される物理量を明らかにすることができれば、人間の個人性知覚過程の解明に近づくばかりでなく、この物理量をさまざまな音声処理技術に応用することが可能である。例えば、これを正規化することにより、不特定話者を対象にした音声認識の話者に対する頑健性が向上することが期待できる。この物理量を利用して話者認識を行うことも可能である。また、合成音声に個人性を付加すれば、より人間らしく聞きやすい音声が合成できるようになる。ある話者の音声を別の話者のものに変換する声質変換の技術にも応用できる。そこで、本研究では「人間が話者識別に利用している物理量が個人性を表す重要な物理量である」という作業仮説をおき、音声中の物理量の変化が話者識別に与える影響を調べ、その関係から個人性を表す物理量を求める。

音声の個人性は、それぞれ声帯特性と声道特性に対応する基本周波数とスペクトル包絡の双方に含まれる。スペクトル包絡には音韻の特徴も含まれており、音声の認識に重要な意味を持つ。そのため、個人性を表す物理量を求め音韻の特徴との関係を明らかにすることができれば、上述した音声処理技術に与える効果が大きい。よって、本研究ではスペクトル包絡における個人性について検討する。

1.2 用語の定義

本研究で用いる用語を定義する。

個人性 (Speaker individuality) とは、音声に含まれる話者の特徴である。個人性には、誰が話しているのかに関する情報や、「通る声」、「澄んだ声」、「甲高い声」などの**声質** (Voice quality) に関する情報などが含まれる。人間が音声から個人性を抽出することを**個人性知覚**と呼ぶ。

話者識別 (Speaker identification) とは、人間が個人性を利用して話者を特定することである。言い換えれば、複数存在する話者のカテゴリの中から、1 つのカテゴリを選択する行為である。本研究では、便宜上コンピュータにより話者識別を行うことを**話者認識** (Speaker recognition) と呼んで区別する。話者認識や音声認識に利用される物理量のことを**特徴パラメータ**と呼ぶ。また、信号処理の技術を用いてある話者の音声を別の話者のものに変換することを**声質変換** (Voice quality control) と呼ぶ。

個人性知覚に関する研究において用いられる音声データは、**既知話者**のものまたは**未知話者**のものに分けられる。既知話者とは被験者が日常の生活もしくは事前の学習によってその個人性を記憶し、話者識別ができる話者のことをいう。一方、未知話者とは被験者がその個人性を記憶しておらず、話者識別できない話者のことである。

音韻 (Phoneme) は、音声の単位の 1 つであり、音素と同義である。日本語の場合、/a/、/i/、/u/、/e/、/o/を**母音** (Vowel) と呼び、それ以外を**子音** (Consonant) と呼ぶ。音声中の音韻に対応する特徴を**音韻性** (Phonetic quality) と呼ぶ。人間が音韻性を利用して音韻を特定することを**音韻識別** (Phoneme identification) と呼ぶ。

単母音 (Isolated vowel) とは単独発声された母音である。**連続音声** (Continuous speech) は文章や単語のように複数の音韻が連続して発声された音声である。連続音声が母音からのみ成る場合を特に**連続母音**と呼ぶ。

以下では、話者識別に関する研究に用いられる聴取実験の方法について説明する。

Naming 法 (Naming method) とは、被験者に既知話者の音声データから作成した刺激音を呈示し、話者の名前を回答させる方法である。Naming 法には、被験者に誰の声かわからない刺激音を棄却 (reject) することを許すものと、棄却を許さず必ず誰かの話者を回答させる (**強制判断**) ものがある。**一対比較法** (Method of paired comparison) とは、被験者に 2 つの刺激音を連続して呈示し、それらの話者の声質を比較させる方法である。**ABX 法** (ABX method) とは、被験者に 3 つの刺激音 A、B、X を連続して呈示し、刺激音 X の話者が刺激音 A と B のどちらの話者かを回答させる方法である。

1.3 研究の背景

1.3.1 個人性を表す物理量に関する研究

従来より音声中の種々の物理量と個人性知覚の関係を調べる研究が行われてきているが、その数はあまり多くない。これらの研究の多くは、音声分析合成系を利用して音声中の物理量を変化させた刺激音を用いて聴取実験を行い、知覚との対応を調査するという方法を採っている。

伊藤らは、既知話者の連続音声と単母音を対象にして、Naming 法と ABX 法によりスペクトル包絡と基本周波数に関する種々の物理量の変化と個人性知覚の関係について調査した。その結果、スペクトル包絡と基本周波数とテンポの中ではスペクトル包絡が最も重要であると報告した。また、スペクトル包絡の復元精度も個人性知覚に影響を与え、復元精度が低いときには基本周波数の情報が利用されることを示した [伊藤 82]。

Furui らは電話帯域 (0〜4 kHz) の連続音声を対象にして、種々の物理量と個人性知覚との関係を調べた。そして、スペクトル包絡に定常的に現れる個人性を捉えていると考えられる、時間平滑スペクトル包絡の 2.5〜3.5 kHz の帯域の距離と心理的距離の相関が高いことを示した [Furui 85a]。

橋本らは、既知話者と未知話者の連続音声を対象にして、ABX 法によりスペクトルと基本周波数と音韻継続時間の 3 要因の個人性知覚への寄与率を求めた。そして、個人性知覚には基本周波数の寄与が最も大きいことを報告した。また、既知話者と未知話者では寄与率が異なり、既知話者の場合に未知話者よりもスペクトルの寄与率が高くなる傾向があると報告した [橋本 95], [橋本 96]。

阿部は、既知話者の連続音声を対象にして一対比較法により、スペクトルと基本周波数の寄与度を求め、これらは個人性の決定要因として同程度に重要であると報告した。また、個人性を決定する周波数帯域を ABX 法により調べ、2360 Hz 以下の低域のスペクトルの差が重要であると報告した [阿部 95]。さらに、これらの知見をもとに「ある話者の音声を徐々に他の話者の音声へ変化させる」**音声モーフィング** (Speech morphing) の実装も行われている [Abe 96]。

桑原らは、5 母音のみを含む無意味音声を対象として、声道特性と個人性知覚の関係を Naming 法により調べた。実験にはフォルマントの周波数とバンド幅を独立に制御した刺

激音を用いた。その結果、F4 以上よりも F3 までのフォルマントにより多くの個人性が含まれ、特に F3 が最も重要であると報告した [桑原 86]。

Kasuya らは、連続母音を対象にして、フォルマントの周波数とバンド幅と音源情報の個人性知覚への寄与を調べた。その結果、音源情報は個人性知覚にあまり寄与せず、F1 と F2 の動的特性が重要であると報告した。さらに、これらのフォルマントを操作することにより声質変換が可能であることを示した [Kasuya 96]。

上記の阿部の研究において個人性を決定する帯域であるとした 2360 Hz 以下の帯域は、日本語の母音において F1 と F2、母音によっては F3 が現れる帯域である。従って、連続音声において個人性知覚に寄与する帯域に関しては、上記の阿部、Kasuya et al.、桑原ら 3 者の結果はほぼ一致しているといえる。

以上の研究の他、基本周波数に関しては平均基本周波数 [桑原 93]、基本周波数のゆらぎ [鈴木 85]、基本周波数の動的特性 [Akagi 97] と個人性知覚の関係に関する研究が行われている。また、フォルマントの動的特性 [X. Yang 96] や話速 (Speaking rate)[Francis 96] と個人性知覚に関する研究も行われている。さらに、話者識別に有効な音韻は話者により異なるという報告もなされている [松井 93]。

これらの研究は、いずれも物理量と個人性知覚に関する重要な知見を示しているものの、個人性知覚過程の解明やモデル化には至っていない。また、個人性を表す物理量に関して音声生成系の個人差と対応させる必要がある。近年の MRI (Magnetic Resonance Image) などの計測技術の進歩により、音声生成系に関する新しい知見が得られている。個人性知覚に関する研究はこれらの研究と連携していく必要がある。

1.3.2 話者認識に適した特徴パラメータに関する研究

話者認識の重要な課題として、話者認識に有効な特徴パラメータの特定がある。話者認識に利用する特徴パラメータに求められる性質としては、

1. 発話内容に依存しないこと
2. 話者内では変動が少なく話者間では変動が大きいこと
3. 発声時期差による変動が少ないこと

が挙げられ、このような特徴パラメータに関する検討が行われている。

Furui は、ケプストラムの動的特徴の話者認識への利用が有効であることを明らかにした。この特徴パラメータはデルタケプストラムと呼ばれ、話者認識ばかりでなく音声認識にも有効であることが確認されている [Furui 86a]。

早川らは、DTW (Dynamic Time Wrapnig) による話者認識に音声の高域に含まれる個人性の利用が有効であることを示した。また、発声時期差による変動の問題に関しても、高域は長期間にわたって安定であることを示した [早川 95]。また、早川らは基本周波数のゆらぎに対応する線形予測残差スペクトルの帯域内強度差の利用が有効であることも報告している [早川 96]。この他、韻律情報の利用が Yegnanarayana らなどにより提案されている [Yegnanarayana 96]。

先にも述べたように、話者認識に有効な特徴パラメータの開発は重要な課題である。しかし、このテーマに関して十分な検討が行われてきたとは言い難い。個人性知覚の研究による知見にもとづいた新たな特徴パラメータの提案が期待される [古井 86b]。

1.3.3 声質変換に関する研究

近年の音声合成に関する技術の進歩に伴い、声質変換に関する数多くの研究が行われている [Kuwabara 95]。これらの研究は、パラメトリック (parametoric) な手法によるもの [小坂 95], [Mizuno 95], [Kasuya 96] とノンパラメトリック (non-parametric) な手法によるもの [Abe 90], [松本 94], [Iwahashi 95], [阿部 95] に大別できる。

これらの研究の多くはスペクトルや韻律情報を変換する手法自体に主眼が置かれている。しかし、話者認識の特徴パラメータと同様に、個人性知覚の研究による知見を利用し、どの物理量を変換させるのが効果的なのかを調べる研究 [阿部 95],[Kasuya 96] も必要である。

1.4 本研究の目的

以上のことを背景として、本研究では人間の個人性知覚過程を明らかにするために、個人性を表す物理量に関する検討を行う。その際、「人間が話者識別に利用している物理量が個人性を表す重要な物理量である」という作業仮説をおき、音声中の物理量の変化が話者識別に与える影響を調べ、その関係から個人性を表す物理量を求める。そして、その物理量を話者認識、声質変換に応用することを試みる。

音声の個人性は、子音よりも母音により多く現れると言われている。よって、本研究では母音の個人性に関して検討する。また、1.3.1節でみたように、個人性は様々な物理量に含まれているが、それらの中でも特にスペクトル包絡における個人性に関して検討する。なぜなら、スペクトル包絡には音韻性に関する情報も含まれているため、個人性と音韻性との関係を明らかにすることができれば、話者認識や音声認識や音声合成などの技術に与える効果が大きいからである。

1.3.1節で挙げた研究では、連続音声を対象にしていた。連続音声のスペクトル包絡には、

時間的に静的な成分に含まれる個人性と、動的な成分に含まれる個人性が混在していると考えられる。これら双方とも個人性知覚に重要な意味を持っているが、静的な成分に含まれる個人性は話者の先天的な声道形状に依存しているため、その話者の個人性の基礎になっていると考えられる。そこで、本研究では音声の静的な成分のスペクトル包絡において個人性を表す物理量について検討する。そして、その物理量を声質変換や話者認識に応用することを試み、さらにその物理量の性質から個人性知覚過程を推定する。

本研究では、はじめに単母音を対象にして個人性を表す物理量を求める。スペクトル包絡に個人性が存在することを確認したうえで、個人性が顕著に現れる帯域、その帯域中で個人性を表す物理量を聴取実験により求める。そして、その物理量を用いて声質変換が可能であることを示す。次に、単母音を対象にして得られた結果をもとに連続音声中の母音の個人性について検討を行なう。最後に、聴取実験により求めた物理量を単純類似度法による話者認識の特徴パラメータにすることで高い弁別能力が得られることを示す。

1.5 本論文の構成

本論文の構成を以下にに示す。(図 1.1参照)

第 1 章では、本論文が対象としている研究分野の現状と問題点を指摘し、本論文の目的を明らかにする。

第 2 章では、単母音のスペクトル包絡に個人性が存在することを確認するための聴取実験を行う。人間が既知話者でも未知話者でも単母音のスペクトル包絡の情報により話者識別が可能であることを確認するため、既知、未知 2 種類の話者による音声データを用いて聴取実験を行う。

第 3 章では、単母音を対象にしてスペクトル包絡において個人性が顕著に現れる帯域を

調査する。まず、周波数軸上でスペクトル包絡の話者間の分散と音韻間の分散を計算する。話者間や音韻間の差が顕著に現れる帯域には、それぞれ話者識別や音韻識別に重要な情報が存在すると考えられるからである。次に、聴取実験によりスペクトル包絡の変形と話者識別の定量的な関係を求める。そして、スペクトル包絡における個人性は高域に顕著に現れることを示す。

第 4 章では、スペクトル包絡の高域のどの物理量が話者識別に寄与するのか、すなわちどの物理量に個人性が顕著に現れるのかについて検討する。特に、スペクトル包絡の微細構造とスペクトル包絡高域のピークとディップの話者識別への寄与について検討する。そして、話者識別にはピークの寄与が大きいことを示す。この結果を受けて、第 5 章ではスペクトルピークに着目し個人性が顕著に現れる帯域を検討する。F3 を含む帯域、20 ERB rate 付近のピークを含む帯域の 2 帯域に関して調査を行い、個人性は後者に顕著に現れることを明らかにする。

第 5 章までは、単母音を対象に検討を行うが、第 6 章では連続音声中の母音のスペクトル包絡における個人性ついて、またスペクトル包絡と基本周波数の話者識別に対する役割について調べる聴取実験を行う。

第 7 章では、聴取実験による知見を単純類似度法 [飯島 89] による話者認識に適用することを試みる。話者認識に適したスペクトル包絡の帯域を求め、高い弁別性能が得られる帯域が人間の話者識別に重要な持つ帯域と一致するか否かを調べる。

第 8 章では、全体の考察を行い、第 9 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。

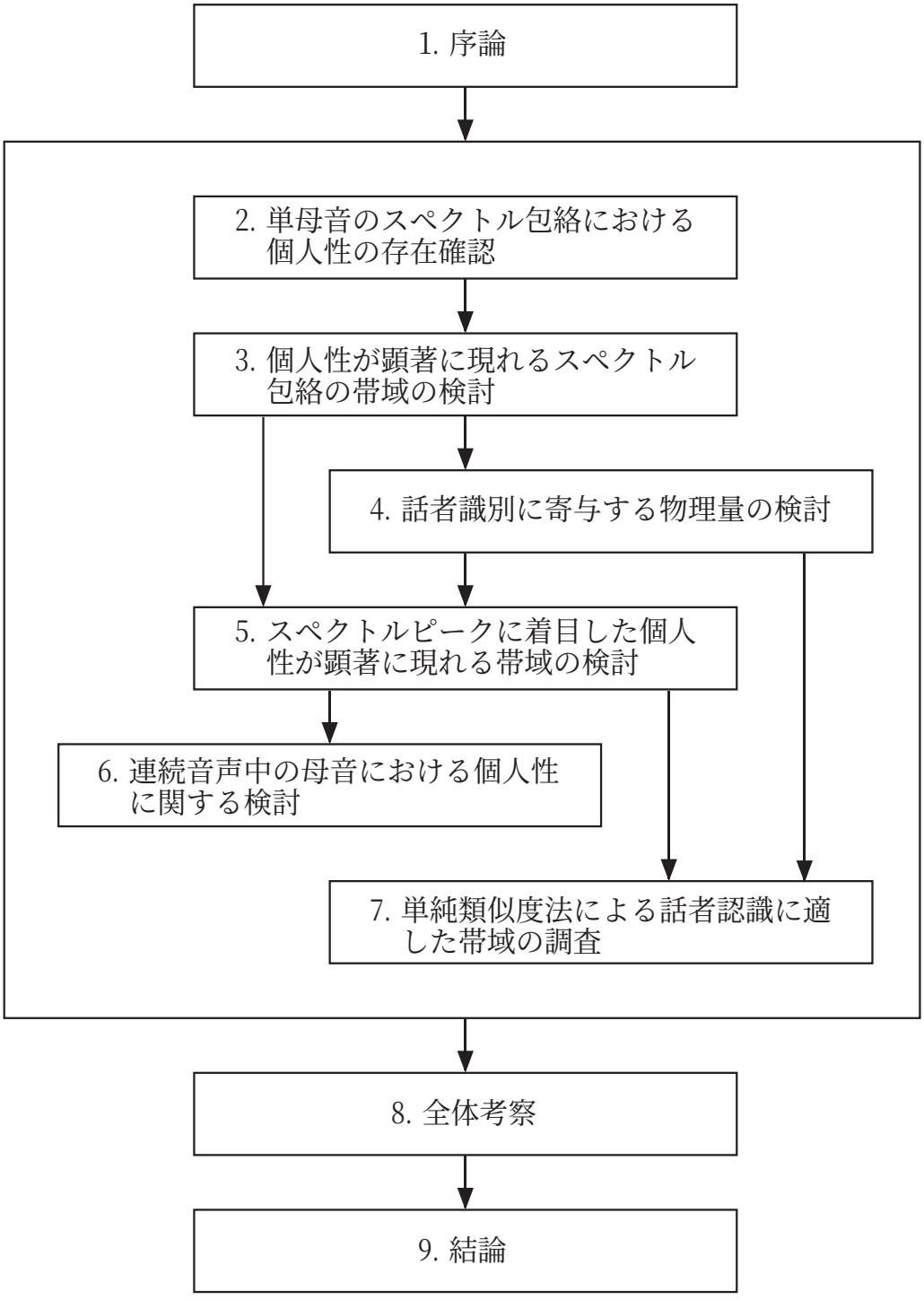
1. 序論
2. 単母音のスペクトル包絡における個人性の存在確認
3. 個人性が顕著に現れるスペクトル包絡の帯域の検討
4. 話者識別に寄与する物理量の検討
5. スペクトルピークに着目した個人性が顕著に現れる帯域の検討
6. 連続音声中の母音における個人性に関する検討
7. 単純類似度法による話者認識に適した帯域の調査
8. 全体考察
9. 結論

図 1.1: 各章の相互関係

第 2 章

単母音のスペクトル包絡における個人性の存在確認

2.1 まえがき

本研究では母音のスペクトル包絡における個人性に関する検討を行うが、その前提として母音のスペクトル包絡に個人性が存在することを確認する必要がある。そこで、単母音を用いてこのことを確認するための聴取実験を行う。

本章では、人間がスペクトル包絡の情報のみで話者識別が可能であることを示すことにより、スペクトル包絡に個人性が存在することを確認する。その際、既知話者と未知話者の音声データを用い、どちらの場合でもスペクトル包絡に存在する個人性を知覚することが可能であることを示す。

既知話者の音声データを用いる聴取実験は Naming 法により行う (実験 2-1)。一方、未知話者の音声データを用いる聴取実験では、被験者に 2 つの刺激音の声質の近さを一対比較法により 5 段階で評価させる (実験 2-2)。

両実験とも、刺激音の作成にはスペクトル包絡と基本周波数を独立に制御でき、かつ高品質の合成音声が得られる **LMA (Log Magnitude Approximation) 分析合成系** [今井 78], [今井 79] を利用する。そして、話者間でスペクトル包絡のみが異なる刺激音を用いて聴取実験を行い、単母音のスペクトル包絡に個人性が存在することを示す。

2.2 実験 2-1 既知話者の音声データを用いた聴取実験

2.2.1 目的

実験 2-1 では、既知話者の音声データを用いて Naming 法により聴取実験を行う。そして、スペクトル包絡の情報のみで被験者が話者を識別できることを示すことにより、単母音のスペクトル包絡に個人性が存在することを確認する。

2.2.2 実験条件

音声データ

音声データは、基本周波数が 125 Hz 前後である 23〜37 歳の男性 5 名による 5 母音である。スペクトル包絡における個人性に関する実験を行うためには、話者毎の基本周波数の違いが話者識別に与える影響を極力抑える必要がある。そこで、録音の際 125 Hz の純音をスピーカから 1 s 呈示し、その後に声の高さを純音に合わせて発声するよう話者に指示した。これにより音声データの基本周波数をほぼ 125 Hz にそろえることができた。

録音は騒音レベル 22.7 dB(A) の防音室にて行った。マイクロフォンからの距離を約 15 cm に保って発声させた音声を防音室の外の DAT レコーダに入力し、標本化周波数 48 kHz で録音した。この音声を標本化周波数 20 kHz にダウンサンプリングして WS (Workstation) に保存し音声データとした。録音に使用した機器を表 2.1に示す。

表 2.1: 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 DTC-57ES
スピーカ	JBL control1 1 台
パワーアンプ	audio-technica AT-SA50

刺激音

聴取実験には以下の 5 種類の刺激音を用いた。

A. 原音声

B. LMA 分析合成音声

C. 平均基本周波数、基本周波数の変化の時間特性、音声波形の振幅を話者間で全て共通にし、スペクトル包絡には処理を加えない音声

D. C において、スペクトル包絡の時間順序をランダムに並べ替え、スペクトル包絡の時間情報を崩した音声

E. D において、スペクトル包絡の傾きを一定にした音声

刺激音 B は、LMA 分析合成系による処理が話者識別と音韻識別に与える影響を調べるためのものである。LMA 分析合成系は合成フィルタとして LMA フィルタを用いる。LMA フィルタの作成には 60 次の FFT ケプストラムを用いた。FFT ケプストラムを求める際のフレーム長は 51.2 ms、フレーム周期は 12.8 ms である。

刺激音 C、D、E はそれぞれの刺激音において変化させた物理量が話者識別と音韻識別に与える影響を調べるためのものである。刺激音 C は、LMA 合成系に図 2.1に示す駆動音源情報を入力することによって作成した。この駆動音源情報は、実際の音声における基本周波数の変化の時間特性を摸擬したものである。

刺激音 D は、時間順序をランダムに並べ替えた FFT ケプストラムから作成した LMA フィルタを用いて合成した。この刺激音において話者間で異なるのはスペクトル包絡のみである。刺激音 D で話者識別が可能か否かによって、単母音のスペクトル包絡に個人性が存在するか否かを判断する。

刺激音 E はスペクトル包絡の傾きが話者識別、音韻識別に与える影響を調べるためのものであり、FFT ケプストラムの 1 次と 2 次の値を音韻毎の平均値として作成した LMA フィルタを用いて合成した。FFT ケプストラムの 1 次と 2 次を正規化したのは、これらがスペクトル包絡の全体的な傾きを表しているからである。

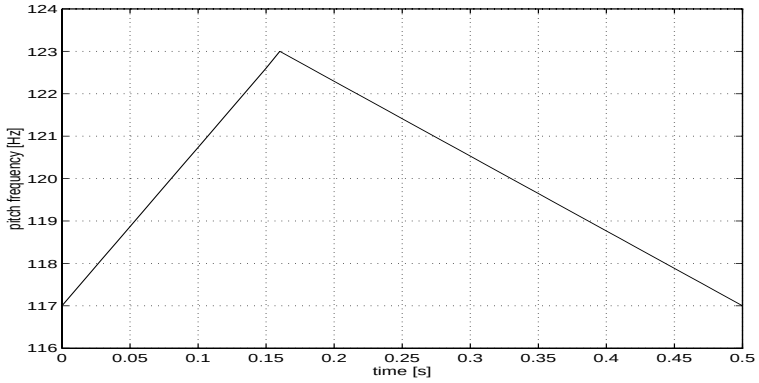

図 2.1: 刺激音 C、D、E の基本周波数の変化の時間特性

被験者

音声データの集録の対象とした話者と日頃接している者 12 名により、刺激音 D を用いて話者識別実験を行った。そして、その話者識別率が 50 %を越えた男性 7 名、女性 1 名の計 8 名を被験者とした。被験者は正常聴力を有する。

実験方法

実験では 5 話者の 5 母音を各 3 回、計 75 個をランダムに約 5 s 間隔で呈示した。被験者は防音室内でヘッドフォンにより受聴した。受聴は各被験者の聞きやすいレベルによる両耳受聴である。そして、回答用紙に書いてある話者と音韻を同時に選択する。ただし、判断不可能の場合に限り “X” と回答すること (棄却) を許した。実験に用いた回答用紙を付録に示す。聴取実験に使用した機器を表 2.2に示す。

表 2.2: 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 SR-λ pro.
パワーアンプ	STAX SRAM-1/MK-2 pro.
DAT プレーヤ	SONY DTC-57ES
DSP	M.I. Systems VMEDSP56K Engine

2.2.3 実験結果と考察

図 2.2に話者識別率と音韻識別率の平均値を示す。被験者が “X” と回答した場合には識別誤りをしたものとして識別率を求めている。刺激音 A の話者識別率と音韻識別率より、被験者は高い精度で原音声の話者と音韻を識別できることがわかる。

各刺激音の話者識別率の間に有意差があるか否かを有意水準 5 %で F 検定を行った [天野 91], [Grimm 93], [大村 95] ($F(1,14) = 4.01, p < .05$)。その結果、刺激音 B と C の間には有意差があるが ($(1,14)F = 17.74$)、刺激音 A と B ($F(1,14) = 0.46$)、刺激音 C と D ($F(1,14) = 0.15$)、刺激音 D と E ($F(1,14) = 2.56$) の間には有意差がないことがわかった。また、音韻識別率に関して検定を行ったところ、全ての刺激音の間に有意差がないことがわかった。

刺激音 A と B の間では話者識別率と音韻識別率のいずれについても有意差がないことから、LMA 分析合成音声は話者識別実験や音韻識別実験に用いるに十分な品質を有しているといえる。

次に、話者識別率に関する検定の結果より刺激音 C、D、E で変化させた物理量が話者識別に与える影響について考察する。まず、刺激音 B と C に有意差があることから平均基本周波数と基本周波数の変化の時間特性には個人性が含まれていることがわかる。これは従

来の報告と一致する [伊藤 82], [桑原 93]。また、刺激音 C と D に有意差がないことから、単母音のスペクトル包絡の時間順序には個人性が含まれているとはいえない。これは、単母音の場合、そのスペクトル包絡は時間的変動が少ないことに起因していると考えられる。加えて、刺激音 D と E に有意差がないことから、単母音のスペクトル包絡の傾きに個人性が含まれているとはいえないこともわかる。

刺激音 D において話者間で異なる物理量はスペクトル包絡のみである。この聴取実験のチャンスレベルが $\frac{1}{5} \times 100 = 20\%$ であることを考えると、74.2 %という刺激音 D の話者識別率はスペクトル包絡には十分に個人性が含まれていることを示しているといえる。

一方、音韻識別率の検定の結果から、平均基本周波数、基本周波数の変化の時間特性、音声波形の振幅、スペクトル包絡の時間順序、そしてスペクトル包絡の傾きのいずれも単母音の音韻識別には影響を与えないことがわかる。

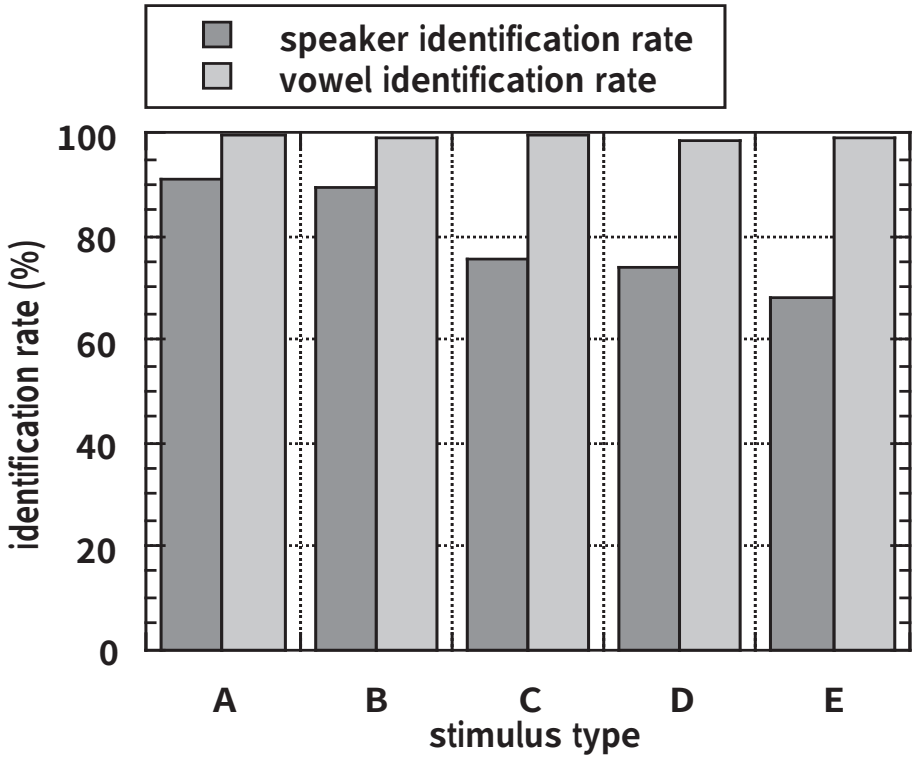

図 2.2: 話者識別率と音韻識別率の平均値 (%)

2.3 実験 2-2 未知話者の音声データを用いた聴取実験

2.3.1 目的

実験 2-2 では、スペクトル包絡に個人性が存在することを確認するために、未知話者の音声データを用いて聴取実験を行う。実験では、被験者に 2 音声の声質の近さを一対比較法により評価させる。そして、話者内の心理的類似度が話者間の心理的類似度よりも高くなることから、未知話者の場合でもスペクトル包絡の情報のみで個人性知覚が可能であることを示す。

2.3.2 実験条件

音声データ

音声データは、ATR の音声データベースの男性 5 名の 5 母音である (タスクコード SY)[武田 88]。この 5 名の話者は、音声データベースの男性 10 名の基本周波数の平均値 125.3 Hz に近い話者を選択した。話者は mnm、mtk、msh、mtm、mms の 5 名である。

刺激音

聴取実験に使用する刺激音は以下の 3 セットである。合成音声の作成には 60 次の FFT ケプストラムを用いた LMA 分析合成系を使用した。

S1. A ：原音声

B ：原音声

S2. A ：原音声

B ：LMA 分析合成音声

S3. A ：原音声

B ：平均基本周波数、基本周波数の変化の時間特性、音声波形の振幅を話者間で全て共通にし、スペクトル包絡の時間順序をランダムに並べ替えた音声

S1、S2 は被験者が原音声または LMA 分析合成音声によって個人性知覚が可能であることを確認するためのものである。S3 の B は、話者間でスペクトル包絡のみが異なる音声である。S3 によって、スペクトル包絡における情報のみで個人性知覚が可能か否かを調べる。

被験者

被験者は実験 2-1 と同じ 8 名であるが、この実験では被験者は音声データの話者の声を全く知らない。

実験方法

聴取実験では、図 2.3のように音声 A、B を 1 セットとして呈示する。1 つのセットは実験全体で 3 回呈示される。なお、音声 A と B の音韻は同じものにした。音声の長さは約 0.2 s であり、2 つの音声 A、B を約 1.0 s 間隔で呈示する。各セットの間隔は約 3.0 s であり、被験者はこの間に評価を行う。

被験者は、A と B の 2 つの音声の話者に関して「全く同じ」、「似ている」、「どちらともいえない」、「似ていない」、「全く異なる」の 5 段階で評価する。付録に、この実験に用いた回答用紙を示す。

実験結果の評価には以下の方法で求める心理的類似度を用いた。被験者によって与えられた 5 段階の評価にそれぞれ 4、3、2、1、0 の整数値を割り当てる。「全く同じ」を 4、「全く異なる」を 0 とする。$a_{ijv}(n)$ を話者 i と話者 j の母音 v の n 回目 $(n = 1, 2, 3)$ の評価であ

るとする ($a_{ijv}(n) = 1, \ldots, 4$)。話者 i と話者 jの間の心理的類似度 ps_{ij}は、この $a_{ijv}(n)$ を評価の最大値 4 で正規化し、n に関して加算平均を求めた値であり、式 2.1で求められる。この値が話者間より話者内で高くなれば、被験者が個人性知覚できたと考える。

$$ps_{ij} = \frac{\sum_v^5 \sum_n^3 a_{ijv}(n)}{3 \times 4} \times 100 \tag{2.1}$$

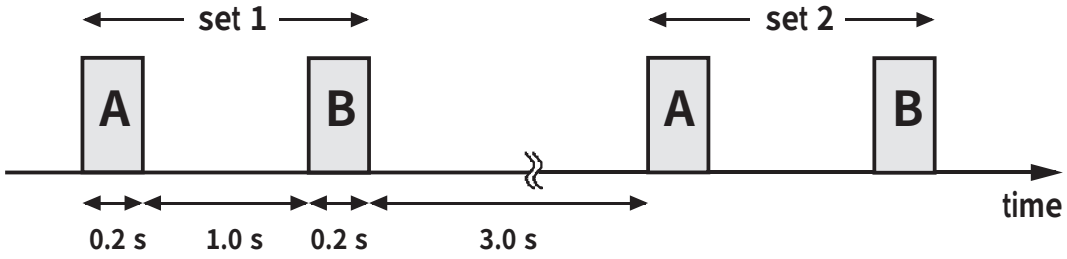

図 2.3: 音声の呈示タイミング

2.3.3 実験結果と考察

S1、S2、S3 の心理的類似度を表 2.3、表 2.4、表 2.5に示す。これらに関して、話者内の心理的類似度が話者間のものより有意に大きいかを有意水準 5 %の F 検定で検定した。その結果、全ての聴取実験に関して話者内の心理的類似度が話者間のものよりも有意に大きいことがわかった。このことは、人間が未知話者でもスペクトル包絡の静的成分の情報を用いて個人性知覚が可能であることを示している。以上の結果より、単母音のスペクトル包絡に個人性が存在することが確認された。

S2 の話者内の心理的類似度 (対角成分) は S1 のものよりも低くなっている。これは、LMA 分析合成系による処理の影響である。また、S3 の話者内の心理的類似度 (対角成分) は S2 のものより低い。これは、平均基本周波数、基本周波数の変化の時間特性、音声波形の振

幅に含まれる個人性が S3 の B の刺激音には含まれていなかったためであると考えられる。これは実験 2-1 の結果と矛盾しない。

2.4 むすび

本章では、単母音のスペクトル包絡に個人性が存在することを確認するための聴取実験を行った。そして、既知話者と未知話者のどちらでも人間がスペクトル包絡の情報を用いて個人性知覚できることから、スペクトル包絡に個人性が存在することを示した。

実験 2-1 では、既知話者の音声データを用いて聴取実験を行った。そして、話者間でスペクトル包絡のみが異なる刺激音に対する話者識別率から、スペクトル包絡に個人性が存在することが確認された。さらに、平均基本周波数と基本周波数の時間変化の時間特性にも個人性が含まれていることがわかった。これは従来の報告と一致する。

実験 2-2 では、未知話者の音声データを用いて聴取実験を行った。この聴取実験では、2 つの音声の声質の近さを被験者に 5 段階で評価させた。その結果、未知話者の場合でも被験者はスペクトル包絡から個人性を知覚できることが明らかになり、スペクトル包絡に個人性が存在することが確認された。

表 2.3: S1 の心理的類似度

	mms	mnm	msh	mtk	mtm
mms	98	30	41	32	42
mnm		98	37	19	43
msh			95	50	20
mtk				99	43
mtm					98

表 2.4: S2 の心理的類似度

	mms	mnm	msh	mtk	mtm
mms	84	50	51	36	37
mnm		72	46	28	55
msh			85	51	29
mtk				85	40
mtm					81

表 2.5: S3 の心理的類似度

	mms	mnm	msh	mtk	mtm
mms	59	18	25	19	19
mnm		59	22	11	26
msh			57	30	12
mtk				60	26
mtm					59

第 3 章

個人性が顕著に現れるスペクトル包絡の帯域の検討

3.1 まえがき

本章では、スペクトル包絡において個人性が顕著に現れる帯域を調査する。

Furui らは電話帯域 (0〜4 kHz) の単語音声を対象として個人性知覚と種々の物理量との関係を調べた。そして、時間平滑スペクトル包絡の 2.5 〜3.5 kHz の帯域の距離と心理的距離の相関が高いことを示した [Furui 85a]。しかし、この研究では単語全体のスペクトル包絡を平均しているため、音韻によるスペクトル包絡の違いが考慮されていない。この違いは個人性知覚に影響を与える可能性があるため、スペクトル包絡に関する個人性は音韻毎に調べる必要がある。また、Furui らの研究では心理的距離との相関が高いとした帯域を操作して知覚に対する影響を調べるということを行っていないが、個人性を表す物理量を明らかにするためにはこのような検討が必要である。

そこで、本章では単母音を対象にして、スペクトル包絡において個人性が顕著に現れる帯域を調べる。まず、周波数軸上でスペクトル包絡の話者間の分散と音韻間の分散を計算する。スペクトル包絡の形の上で個人差と音韻間の差が顕著に現れる帯域には、それぞれ話者識別と音韻識別に重要な情報が存在する可能性があるからである。

次に、3 つの聴取実験により個人性が顕著に現れる帯域を調査する。聴取実験には LMA 分析合成系を利用してスペクトル包絡の特定の帯域のみに変形を加えた刺激音を用いる。そして、スペクトル包絡の変形と話者識別の定量的な関係を求め、そこから帯域と個人性の関係を求める。

3.2 スペクトル包絡の分析

3.2.1 目的

スペクトル包絡の形の上で個人差と音韻間の差が顕著に現れる帯域には、それぞれ話者識別と音韻識別に重要な情報が存在する可能性がある。そこで、周波数軸上で話者間の分散と音韻間の分散を計算する。

3.2.2 音声データとスペクトル包絡の計算方法

音声データは ATR の音声データベースの男女各 10 名の話者による 5 母音である (タスクコード SY)[武田 88]。標本化周波数は 20 kHz である。

フレーム長 51.2 ms、フレーム周期 6.4 ms にて FFT ケプストラムを計算し、有声区間で平均した。そして、周波数軸上でスペクトル包絡のパワーを正規化するために、FFT ケプストラムの 0 次を一定にし、60 次までを用いてスペクトル包絡を求めた。さらに、人間の内耳の基底膜の周波数分析特性を考慮した表現に対応させるため、周波数軸を ERB rate[Glasberg 90] に変換した。

ERB rate は等価矩形帯域幅 (Equivalent Rectangular Bandwidth, ERB) を幅 1 として周波数軸を変形したものである。ERB と ERB rate はそれぞれ以下の式で求められる。

$$ERB = 24.7(4.37F + 1) \tag{3.1}$$

$$ERB\ rate = 21.4\log_{10}(4.37F + 1) \tag{3.2}$$

ここで、F は周波数 (kHz) である。式 3.2は基底膜上の周波数マッピングを近似的に求めている [Greenwood 90], [赤木 94]。

3.2.3 話者間の分散と音韻間の分散

話者 s $(s = 1, \ldots, 10)$ により発声された音韻 v $(v = 1, \ldots, 5)$ のスペクトル包絡を $E_{sv}(n)$ と表す。ここで n は ERB rate を表す。このとき、音韻 vの話者間のスペクトル包絡の分散は、

$$\sigma_v^2(n) = \frac{1}{9}\sum_{s=1}^{10}\{E_{sv}(n) - \mu_v(n)\}^2 \tag{3.3}$$

で与えられる。ここで、$\mu_v(n)$ は音韻 jのスペクトル包絡の平均であり、

$$\mu_v(n) = \frac{1}{10}\sum_{s=1}^{10}E_{sv}(n) \tag{3.4}$$

で与えられる。次に、話者 i の音韻間のスペクトル包絡の分散は

$$\sigma_s^2(n) = \frac{1}{4}\sum_{v=1}^{5}\{E_{sv}(n) - \mu_s(n)\}^2 \tag{3.5}$$

で与えられる。ここで、$\mu_s(n)$ は話者 i の音声のスペクトル包絡の平均であり、

$$\mu_s(n) = \frac{1}{5}\sum_{v=1}^{5}E_{sv}(n) \tag{3.6}$$

で与えられる。

3.2.4 分析結果と考察

話者間の分散を図 3.1に示す。上図は女性 10 名、下図は男性 10 名の分散である。この図から個人差は約 22 ERB rate (2212 Hz) 以上の帯域に顕著に現れることがわかる。音韻間の分散を図 3.2に示す。この図から音韻間の差は 12〜25 ERB rate (603〜3142 Hz) の帯域に顕著に現れることがわかる。

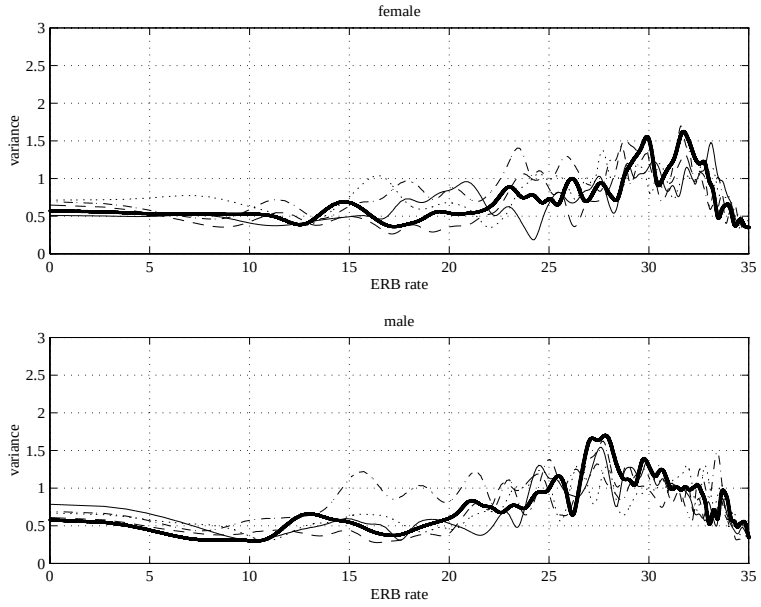

図 3.1: 話者間の分散 σ_j^2 (上：女性 10 名、下：男性 10 名)

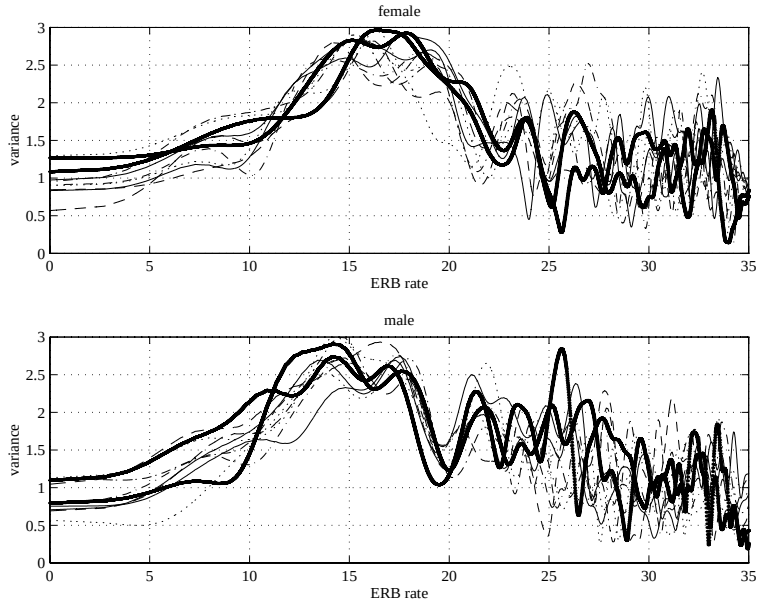

図 3.2: 音韻間の分散 σ_i^2 (上：女性 10 名、下：男性 10 名)

3.3　実験 3-1　帯域毎の個人性の分布

3.3.1　目的

本節では、聴取実験により個人性が顕著に現れる帯域を調べ、前節の結果との対応をみる。実験では、スペクトル包絡の 0〜10、10〜20、20〜30 ERB rate の 3 帯域のうちで個人性がどの帯域により多く現れるか調べる。ここで、スペクトル包絡を ERB rate 上で等間隔に分割するのは、人間の基底膜上の周波数の表現に対応させるためである。

3.3.2　実験条件

音声データ

ATR 音声データベースの男性話者 9 名 (mau, mht, mmy, mnm, msh, mtk, mtm, mtt, mxm) による標本化周波数 20 kHz の 5 母音 (タスクコード SY)。

刺激音

刺激音は LMA 分析合成系により合成した。LMA フィルタの作成に用いるケプストラムは、改良ケプストラム法 [今井 79] により求めた。分析条件はフレーム長 25.6 ms、フレーム周期 6.4 ms、加速係数 1.0、近似回数 3 である。刺激音 A、B、X には以下のものを用いた。

A : 話者 9 名中 1 名のスペクトル包絡を持つ音声

B : 話者間で加算平均したスペクトル包絡を持つ音声

X : 刺激音 B のスペクトル包絡の以下の 4 帯域を刺激音 A のもので置換した音声

X1. 全帯域 (刺激音 A と同じ)

X2. 0〜10 ERB rate (0〜442 Hz)

X3. 10〜20 ERB rate (442〜1740 Hz)

X4. 20〜30 ERB rate (1740〜5544 Hz)

図3.5に話者 mht の/a/のスペクトル包絡 (刺激音 A)、刺激音 B の/a/のスペクトル包絡、20〜30 ERB rate を話者 mht の/a/で置換したスペクトル包絡 (刺激音 X4) を示す。

刺激音 A は各話者の各音韻の有声区間を時間平均した 60 次までのケプストラム c_A から合成し、刺激音 B は c_A を音韻毎に話者間で加算平均したケプストラム c_B から合成した。刺激音 X は c_A と c_B を用いて合成した。刺激音 X2 を例に作成方法を説明する (図 3.3参照)。はじめに、c_A と c_B に 512 点 DFT をかけて対数スペクトラム s_A、s_B を得る。次に、s_B の 0〜10 ERB rate を s_A の 0〜10 ERB rate で置換する。置換した対数スペクトラムに 512 点 IDFT をかけ、再びケプストラムを得る。このケプストラムから LMA フィルタを作成し、合成音声を得る。

変形を加えた対数スペクトルには不連続点が生じることがある。しかし、この不連続点が合成音声のスペクトル包絡に現れることはほとんどない。なぜなら、対数スペクトル包絡上の不連続点はそれに IDFT をかけて得られるケプストラムの高次に影響を与えるが、LMA フィルタの作成には 60 次までの低次のケプストラムを用いるためである。

本節以降の聴取実験で用いるスペクトル包絡を変形した合成音声の作成も、これと同様の方法で行う。すなわち、ケプストラムを一旦対数スペクトラムに変換し、この領域で変形を加えた後、再びケプストラムに変換し、LMA フィルタを作成して合成音声を得る。

合成音声の平均基本周波数は、9 名の話者の基本周波数の平均である 130 Hz にした。基本周波数は図3.4に示す時間特性を持っている。刺激音の長さは 0.5 s で、振幅を正規化し、さらに刺激音の前後部 0.05 s を sin 関数で重み付けした。

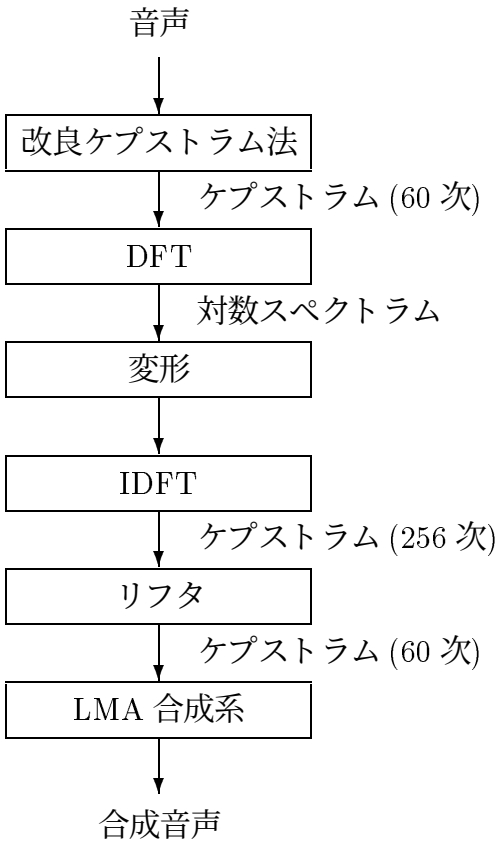

図 3.3: スペクトル包絡を変形した音声作成方法

この実験で用いる刺激音において、話者間で異なる物理量はスペクトル包絡のみである。よって、実験結果はスペクトル包絡の違いのみに起因する。

被験者

正常聴力を有する 23〜25 歳の男性 9 名、女性 1 名の計 10 名。これらの被験者は音声データの話者の声を知らない。

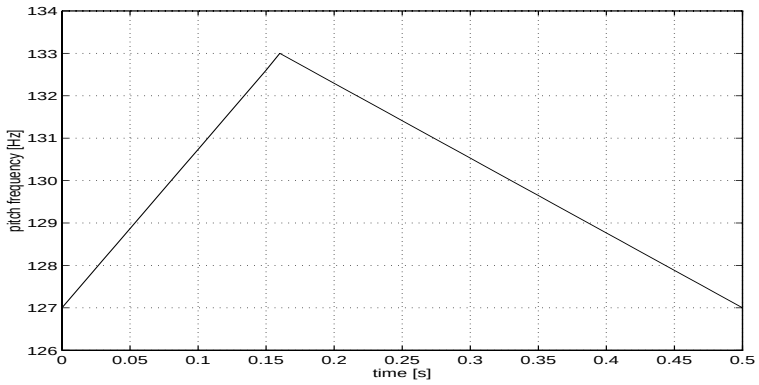

図 3.4: 刺激音の基本周波数の時間特性

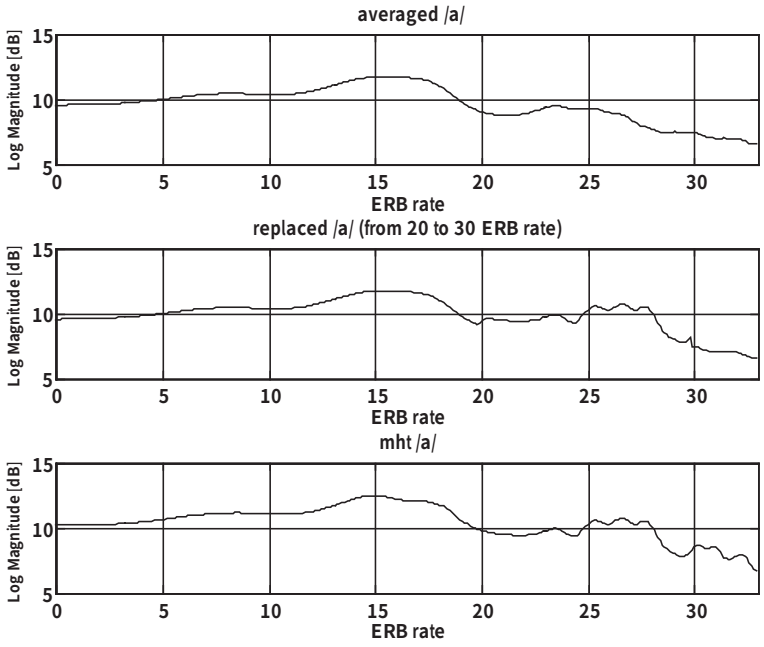

図 3.5: 上：話者間で加算平均した/a/のスペクトル包絡、中：20〜30 ERB rate を話者 mht の/a/で置換したスペクトル包絡、下：話者 mht の/a/のスペクトル包絡。

実験方法

ABX 法により行った。同じ音韻の刺激音 A、B、X を約 2 s の間隔で呈示し、刺激音 X の話者が A と B の話者のどちら似にているかを強制判断させた。継時効果を打ち消すために、BAX の順についても実験を行った。A、B、X の 3 つの刺激音の組を 1 刺激とし、1 刺激につき ABX、BAX を各 3 回、計 6 回ランダムに呈示した。

被験者は防音室内でヘッドフォンにより受聴した。受聴は各被験者の聞きやすいレベルによる両耳受聴である。被験者には聞き直しを許し、パーソナルコンピュータ (PC) を用いて回答させた。なお、実験中は PC の HDD (Hard Disk Drive) を停止させ RAM のみを使って稼働させることにより、HDD の回転音によるノイズは発生しないよう考慮してある [北村 96]。

刺激音は防音室の外に設置されたワークステーション (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。WS から出力された刺激音は D/A 変換され、さらに 8 kHz (33.3 ERB rate) の LPF を通過させることにより高域に発生するノイズを除去した [北村 96]。聴取実験システムの全体図を図 3.6に、使用した機器を表 3.1に示す。

3.3.3 実験結果と考察

被験者が刺激音 X の話者を刺激音 A の話者に似ていると回答した割合を図 3.7に示す。この値はスペクトル包絡の置換によって刺激音 X の個人性が代わった割合を表している。よって、以下ではこの割合を**変換率**と呼ぶ。変換率が高い帯域ほど個人性がより多く現れていることになる。置換した帯域が 0〜33.3 ERB rate (8 kHz) の場合はスペクトル包絡の全帯域を置換した場合に相当する。

図 3.7から、置換する帯域が高くなるに従い変換率が増加する傾向があることがわかる。

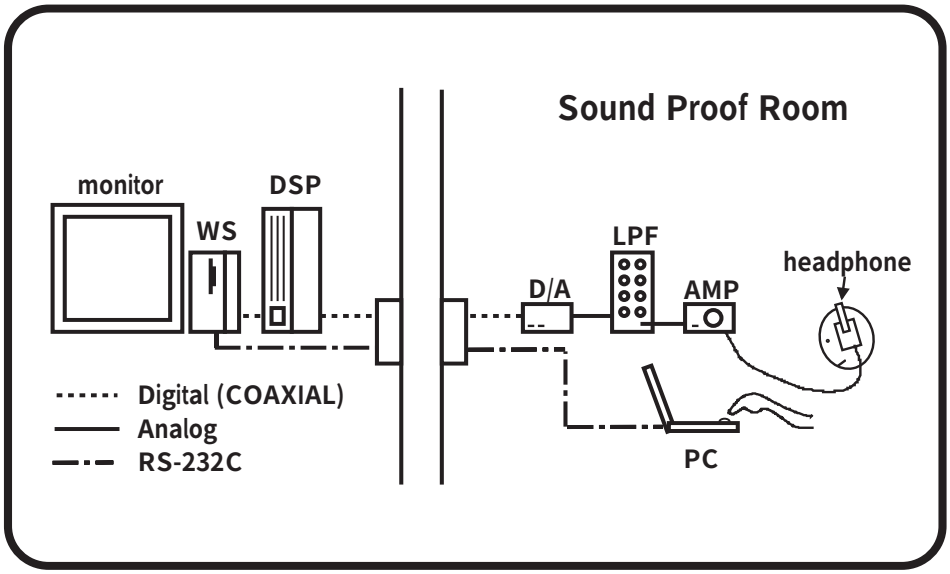

図 3.6: 聴取実験システムの全体図

表 3.1: 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 SR-λ pro.
アンプ	STAX SRAM-1/MK-2 pro.
LPF	NF P-86
D/A プロセッサ	STAX DAC-TALENT BD
DSP	VMEDSP56K Engine
WS	Sun S-4/IX
PC	Macintosh PowerBook Duo

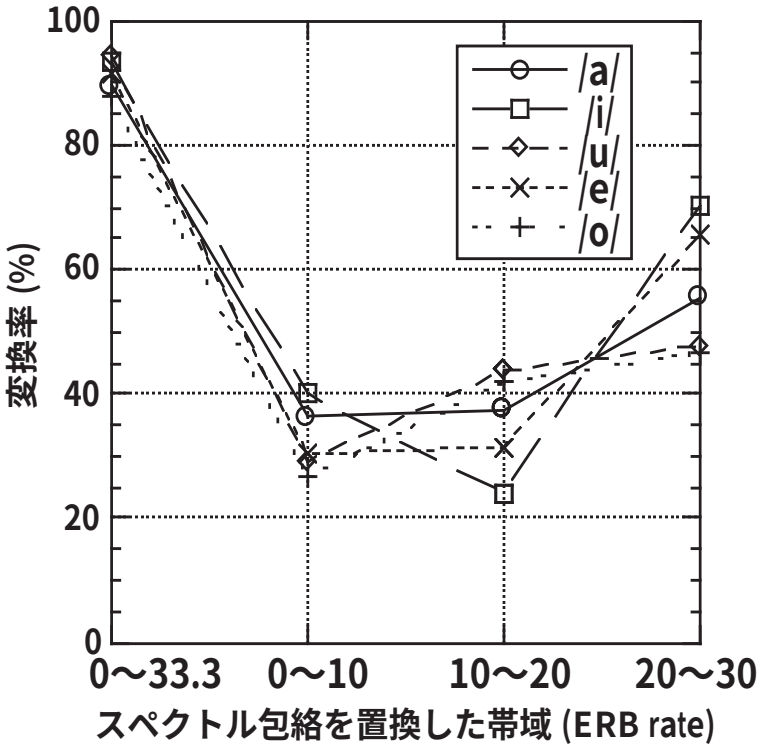

図 3.7: 変換率の平均値と標準偏差

このことは音声の個人性はスペクトル包絡の全帯域に現れるが、高域により多く現れることを示している。これは前節での分析結果と対応している。

また、刺激音 X4 の/a/の変換率を表 3.2に示す。この表からこれは話者により変換率が大きくばらついていることがわかる。このばらつきには、各話者のスペクトル包絡と話者間で加算平均したスペクトル包絡との距離が話者により異なることが関係している可能性がある。そこで、各話者のスペクトル包絡と話者間で加算平均したスペクトル包絡とのユークリッド距離を 0〜10、10〜20、20〜30 ERB rate の帯域毎に求め、話者識別率との関係を調べた。以後この距離を**スペクトル距離**と呼ぶ。話者識別率とスペクトル距離の関係の典型的な例として、図 3.8に話者 mht の/a/、図 3.9に話者 mnm の/i/の結果を示す。

表 3.2: 刺激音 X4 の/a/の変換率 (%)

話者	変換率	話者	変換率	話者	変換率
mau	63.3	mht	83.3	mmy	26.7
mnm	20.0	msh	66.7	mtk	88.7
mtk	86.7	mtt	83.3	mxm	10.0

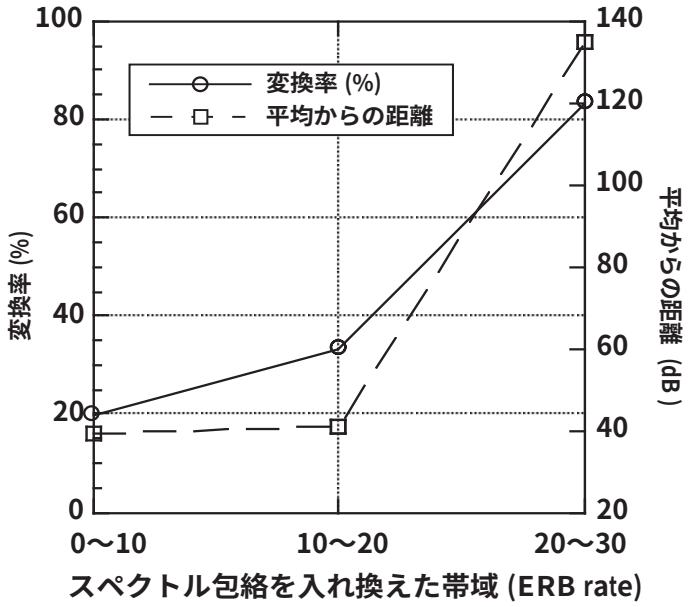

図 3.8: mht の/a/の変換率とスペクトル距離の関係

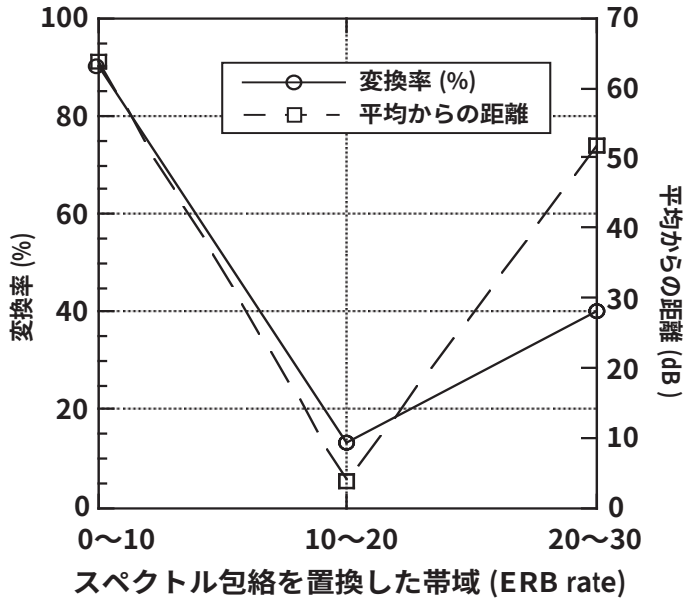

図 3.9: mnm の/i/の変換率とスペクトル距離の関係

これらの図から、話者識別率とスペクトル距離には相対的な対応関係があることがわかる。そこで、話者識別率が 60 %を越える話者と音韻の組合せと、平均値よりも大きいスペクトル距離を持つ話者と音韻の組合せの間の一致度を求めることにより、話者識別率とスペクトル距離の相関を求め、考察を行う。一致度の計算方法を以下に説明する。

1. 各話者の話者識別率に関して 60 %以上の場合には 1、60 %以下の場合には 0 とした行列を作成する。この実験で用いた話者は 9 名で、話者識別率はスペクトル包絡を置換した帯域によって 3 種類あるため、音韻毎に 9×3 行列が得られる。/a/の話者識別率に関する行列を表 3.3に示す。
2. スペクトル距離に関して音韻毎に平均値を求め、各話者の各帯域のスペクトル距離が平均値よりも大きい場合には 1、小さい場合には 0 とした 9×3 行列を作成する。スペクトル距離に関する行列を表 3.4に示す。
3. 音韻毎に話者識別率に関した行列とスペクトル距離に関した行列の要素毎に排他的論理積を求め、一致度とする。一致度の最大値は 27 であり、値が大きいほど話者識別率とスペクトル距離の間の相関が高いと考える。表 3.3と表 3.4の 2 つの行列から得られる一致度は 24 である。

以上の方法で求めた一致度を図 3.10に示す。この図から話者識別率とスペクトル距離の相関が高いことがわかる。このことから、スペクトル距離が大きい帯域ほど、話者変換の効果が大きいといえる。つまり、話者本来のスペクトル包絡に近づくほど、知覚上もその話者に近づくということである。このことはスペクトル距離が知覚上の距離と相関が高いことを示している。これはスペクトル包絡の高域に限らず、0〜10 ERB rate の低域にもいえることである。

しかし、ABX 法による聴取実験では被験者が音色の違いによる判断を行っていた可能性

表 3.3: /a/の話者識別率に関する行列

freq.	speaker								
band	mau	mht	mmy	mnm	msh	mtk	mtm	mtt	mxm
0〜10	0	0	0	0	1	0	0	1	0
10〜20	0	0	0	0	0	1	0	0	0
20〜30	1	1	0	0	1	1	1	1	0

表 3.4: /a/のスペクトル距離に関する行列

freq.	speaker								
band	mau	mht	mmy	mnm	msh	mtk	mtm	mtt	mxm
0〜10	0	0	0	0	1	0	1	0	0
10〜20	0	0	0	0	0	0	0	0	0
20〜30	1	1	0	0	1	1	1	1	0

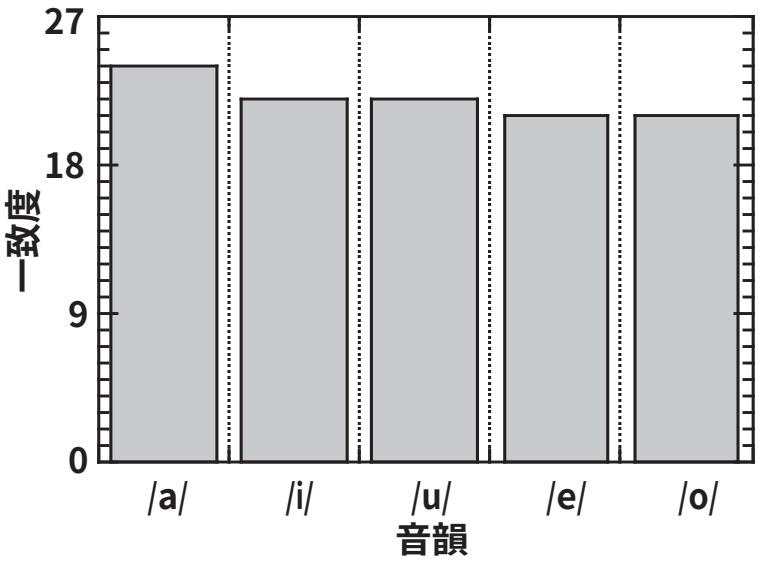

図 3.10: 話者識別率とスペクトル距離の一致度

もあり、結果が個人性のみの影響を反映したものかについては疑問が残る。そこで、以下の聴取実験ではNaming法により聴取実験を行う。

3.4 実験3-2 個人性と音韻の特徴が顕著に現れる帯域の調査

3.4.1 目的

3.2節において、スペクトル包絡の22 ERB rate以上の帯域で個人差が大きく現れ、12〜25 ERB rateの帯域に音韻間の差が大きく現れることが示された。そこで、22 ERB rateを境に個人性が顕著に現れる帯域と音韻の特徴が顕著に現れる帯域とを分割することができるという仮説をたてて、これを検証するためにNaming法による聴取実験を行った。

3.4.2 実験条件

音声データ

2.2節の実験2-1で用いたものと同じ、男性5名による5母音。

刺激音

刺激音は実験2-1の刺激音D、すなわち話者間でスペクトル包絡のみが異なるLMA分析合成音声の12〜22 ERB rateまたは22 ERB rate以上の帯域のスペクトル包絡を変形させた音声を用いる。これは、それぞれの帯域が話者識別と音韻識別にどのような影響を与えるのかを調べるためのものである。

スペクトル包絡の変形は、スペクトル包絡のピークとディップに関して行った。人間の聴覚にはスペクトルのピークが重要であるため、これに変形を加えることはその帯域の情報をこわすことに相当すると考えたからである。ここでピークとはスペクトル包絡におい

てその回帰直線より値の大きい部分を指し、ディップとは回帰直線より値の小さい部分を指す。

スペクトル包絡の変形は下記の 2 つの方法により行った。$E(n)$ を変形前のスペクトル包絡、$E'(n)$ を変形後のスペクトル包絡、$R(n)$ をスペクトル包絡の回帰直線とする。n は回帰直線である。そして、$n = N_1 \sim N_2$ ERB rate の帯域に変形を加えるとする。

1. スペクトル包絡を回帰直線で置換する方法

$$E'(n) = \begin{cases} R(n) & n = N_1 \sim N_2 \\ E(n) & otherwise \end{cases} \tag{3.7}$$

2. スペクトル包絡を回帰直線に対して反転させる方法

$$E'(n) = \begin{cases} R(n) - (E(n) - R(n)) & n = N_1 \sim N_2 \\ E(n) & otherwise \end{cases} \tag{3.8}$$

図 3.11に、LMA 分析合成音声、22 ERB rate 以上の帯域を回帰直線で置換したスペクトル包絡、22 ERB rate 以上の帯域を回帰直線に対して反転させたスペクトル包絡を示す。

上述の変形方法を用いて以下の 4 種類の音声を作成し、話者、音韻識別実験を行った。

A. 変形なし

B. 12〜22 ERB rate を回帰直線に対して反転させた音声

C. 22 ERB rate 以上を回帰直線に対して反転させた音声

D. 22 ERB rate 以上を回帰直線で置換した音声

被験者

実験 2-1 と同じ、音声データの集録の対象とした話者と日頃接している男性 7 名、女性 1 名の計 8 名。

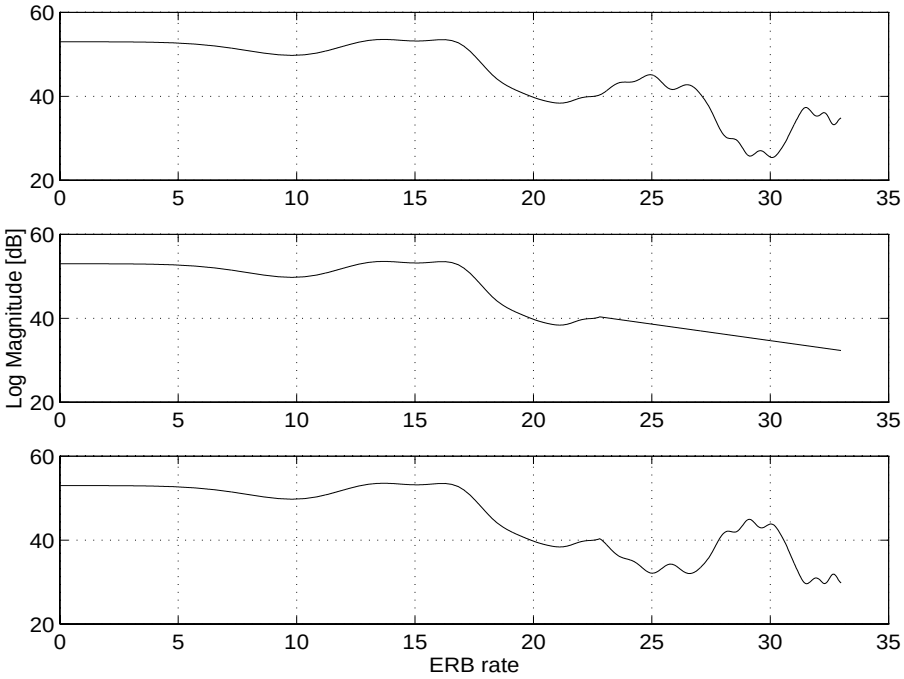

図 3.11: 22 ERB rate 以上の帯域を変形させたスペクトル包絡 (上：変形なし、中：回帰直線で置換、下：回帰直線に対して反転)

実験方法

実験 2-1 と同様に、被験者は防音室内でヘッドフォンにより受聴した。受聴は各被験者の聞きやすいレベルによる両耳受聴である。そして、回答用紙に書いてある話者と音韻を選択する。ただし、判断不可能の場合に限り “X” と回答することを許す。

3.4.3 実験結果と考察

図3.12に実験結果を示す。この実験結果に関して有意水準 5 % で F 検定を行った ($F(1,14) = 4.60, p < .05$)。はじめに、変形する帯域によって話者識別率に有意差があるか否かを検定した。その結果、刺激音 A と B の間 ($F = 25.1$)、刺激音 A と C の間 ($F(1,14) = 88.9$)、刺激音 B と C の間 ($F(1,14) = 11.8$) に有意差があることが明らかになった。

次に、変形する帯域によって音韻識別率に有意差があるか否かを検定した。その結果、刺激音 A と B の間 ($F(1,14) = 342.9$)、刺激音 B と C の間 ($F(1,14) = 223.9$) には有意差があるが、刺激音 A と C の間 ($F(1,14) = 4.51$) にはないことが明らかになった。

最後に、変形方法によって話者、音韻識別率に差があるか否かを検定した。その結果、刺激音 C と D の話者識別率には有意差があり ($F(1,14) = 14.3$)、音韻識別率には有意差がない ($F(1,14) = 2.9$) ことが明らかになった。

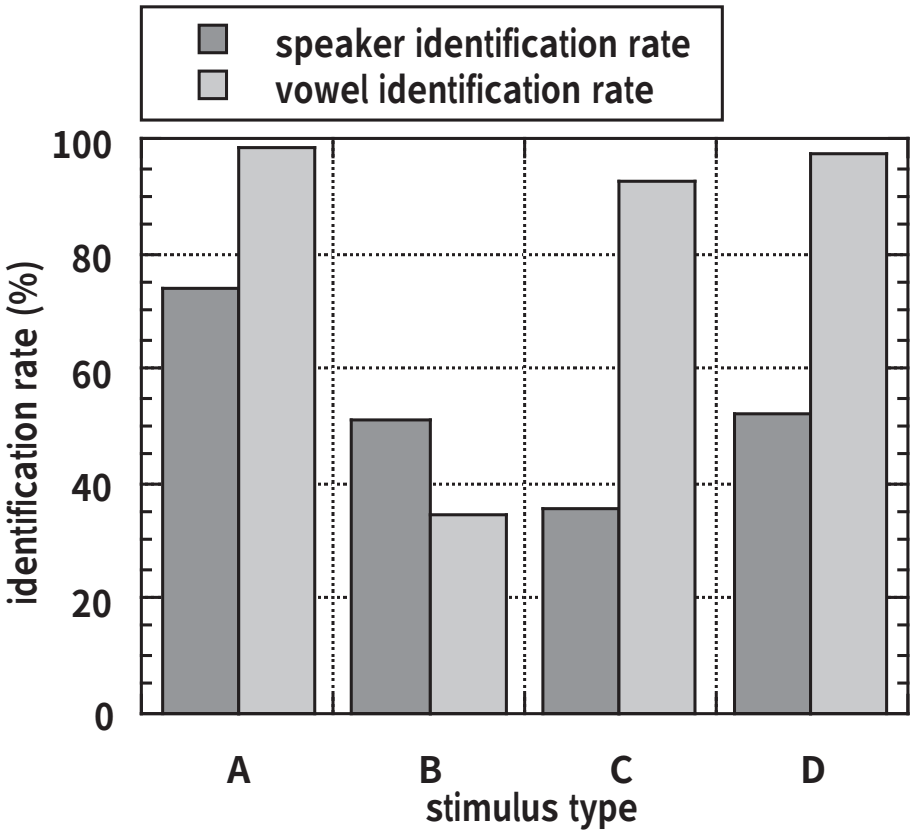

図 3.12: 話者識別率と音韻識別率の平均値

図 3.12と検定の結果から以下のことが結論される。

1. スペクトル包絡の 22 ERB rate 以上の帯域の変形は話者識別には影響を与えるが、音韻識別には影響を与えない。
2. 変形方法による影響の比較では、スペクトル包絡をその回帰直線に対して置換するよりも反転させる方が話者識別率が低くなる。
3. スペクトル包絡の 12〜22 ERB rate の帯域の変形は、音韻識別に大きな影響を与える。この帯域の変形は話者の識別にも影響を与えるが、22 ERB rate 以上の帯域の変形が与える影響よりは小さい。また、この帯域を変形させた音声の話者識別率は音韻識別率よりも有意に大きい。

第 1 の結論は、スペクトル包絡の 22 ERB rate 以上の帯域は話者識別に重要な意味を持ち、この帯域おける個人性は音韻識別と独立に取り扱うことが可能であることを意味している。また、第 2 の結論は話者識別にはスペクトル包絡のピークとディップの位置関係が重要な意味を持っていることを示唆している。さらに、第 3 の結論は人間は音韻識別ができない音声に対しても話者識別できることを示している。しかし、12〜22 ERB rate の帯域に変形を加えることによって、音韻識別のみならず話者識別にも影響を与えることから、スペクトル包絡における音韻性を個人性と独立に取り扱うことはできないことがわかる。

以上の点から、スペクトル包絡の 22 ERB rate 以上の帯域における個人性は音韻性と独立に制御できるが、スペクトル包絡の 12〜22 ERB rate における音韻性は個人性と独立には制御できないことが結論される。この原因として、人間の話者識別過程は音韻識別過程との何らかのかかわりがあるということが考えられるが、これに関する検討は今後の課題である。

Mokhtari らは/CVd/の音声データを対象にして、音声認識に利用する帯域の上限と認識

率の関係を調べている。ここで、C は/h, b, d, g, p, t, k/のうちのいずれかの子音、V は鼻音化母音でない母音である。彼らは、話者が 1 名の場合と話者が複数の場合について実験を行っている。その結果、話者が複数の場合、1780 Hz (20.2 ERB rate) 以上の帯域を利用すると認識率が低下するが、話者が 1 名の場合にはそのような低下はみられないことを報告している [Mokhtari 94]。

Mokhtari らの結果が示すように、従来の音声認識や話者認識でケプストラムを用いる場合には、周波数軸上で一様な重みを用いていたために、高域における個人差の影響を受けて認識精度が低下していたと考えられる。そこで、この聴取実験で得られた結果を利用して、高域の重みを小さくする処理を施せば、不特定話者音声認識の性能を向上させることができると考えられる。この考えは、正規型自然観測法理論を不特定話者の母音認識に応用した飯島らによる研究で適用され効果を上げている [飯島 97]。一方、話者認識の場合には、高域の重みを大きくする処理を加えることにより、認識性能が向上すると考えられる。

3.5 実験 3-3 個人性が顕著に現れる帯域の調査

3.5.1 目的

前節の結果から、スペクトル包絡の 22 ERB rate 以上の帯域には個人性が顕著に現れることが明らかになった。しかし、この帯域の境界である 22 ERB rate という値は 3.2節で求めた話者間の分散から恣意的に決定したものである。本節では、個人性が顕著に現れる帯域をより正確に求めることを目的として、スペクトル包絡の高域を別の話者のもので置換した音声を用いて聴取実験を行う。実験ではスペクトル包絡を置換する帯域の境界を変化させ、話者識別率との関係を求める。

3.5.2 実験条件

音声データ

音声データは普段の音声の基本周波数が 120 Hz に近い男性 3 名による 5 母音である。話者毎の基本周波数の違いが話者識別に与える影響を極力抑えるため、録音の際 120 Hz の純音をスピーカから 1 s 呈示し、その後に声の高さを純音に合わせて発声するよう話者に指示した。標本化周波数は 20kHz である。

刺激音

刺激音の合成には 60 次の FFT ケプストラムを用いて作成した LMA 分析合成系を使用した。FFT ケプストラムを求める際のフレーム長は 51.2 ms、フレーム周期は 12.8 ms である。

被験者がスペクトル包絡の違いのみでこの実験の話者を識別ができることを確認するため、以下の 3 種類の刺激音を用いて予備実験を行う。

1. 原音声
2. LMA 分析合成音声
3. 基本周波数と音声波形の振幅を話者間で共通にし、スペクトル包絡をランダムに並べ替え時間順序を崩した音声

個人性が顕著に現れる帯域を調べる実験に用いる刺激音は、予備実験の刺激音 3 のスペクトル包絡を、図 3.13に示すように、話者 A のスペクトル包絡の高域を話者 B のもので置換した音声である。

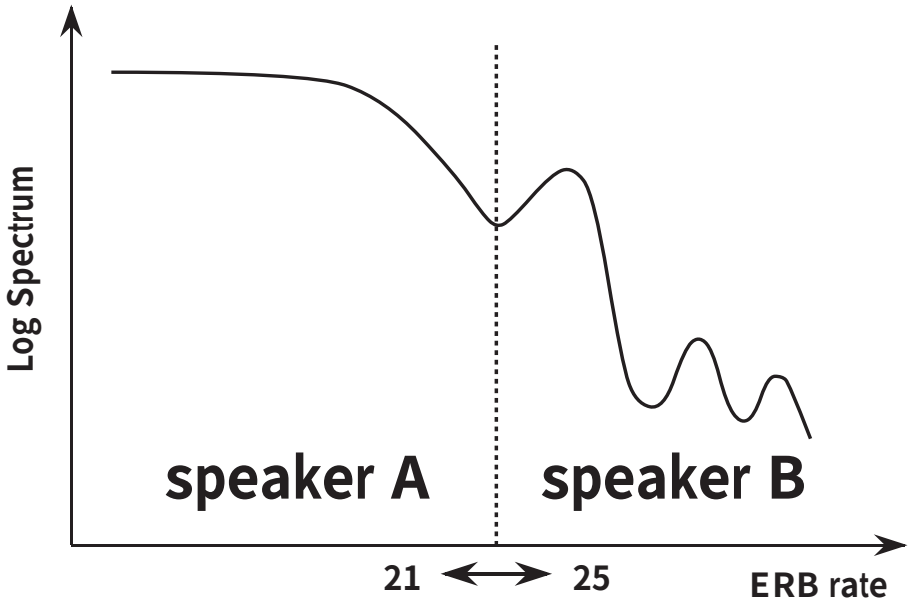

図 3.13: 話者 A のスペクトル包絡の高域を話者 B のものと置換したスペクトル包絡

被験者

正常聴力を有し、音声データの収集の対象とした話者と日頃接している 23〜25 歳の男性 4 名。実験 2-1 の被験者とは異なる。

実験方法

各刺激音をランダムに 3 回呈示する。データ長は約 0.5 s、呈示の間隔は約 6 s である。被験者は防音室内でヘッドフォンにより両耳受聴し、話者と音韻を回答する。ただし、判断不可能の場合には “X” と回答することを許した。

3.5.3 予備実験の結果

表 3.5に被験者間で平均した話者識別率と音韻識別率を示す。刺激音 3 において、話者による違いはスペクトル包絡のみである。この刺激音に対して 79.4 %の話者識別率が得られたことから、被験者はスペクトル包絡のみを手がかりにしてこの実験の話者を識別できることが確認された。

表 3.5: 予備実験の話者識別率と音韻識別率の平均値 (%)

刺激音	話者識別率	音韻識別率
1	94.4	98.9
2	87.2	98.3
3	79.4	99.4

3.5.4 スペクトル包絡の低域と高域の音韻が等しい場合

低域と高域の音韻を同じものとし、話者の全ての組合せで音声を作成した。そして、低域と高域の境界を 21〜25 ERB rate (1963〜3142 Hz) の範囲で 1 ERB 毎 5 段階に変化させ、話者識別と音韻識別に与える影響を調べる。境界を変化させる範囲は、3.2節のスペクトル包絡における話者間の分散の結果を参考に決定した。

置換を行う帯域の境界と話者識別率、音韻識別率の関係を図 3.14に示す。図中の speaker A、B はそれぞれ低域、高域の話者であると回答した割合であり、other はこれら以外の話者および判断不可能であると回答した割合である。また、0、35 ERB rate における話者識別率と音韻識別率は置換を行わない場合、すなわち、予備実験の刺激音 3 の話者識別率と音韻識別率である。

スペクトル包絡を置換する帯域の境界が 21〜23 ERB rate (1963〜2489 Hz) のとき、高域の話者であると回答された割合が低域の話者であると回答された割合よりも高いことから、スペクトル包絡における個人性は 21 ERB rate 以上の帯域に顕著に現れることがわかる。この結果は、スペクトル包絡の 21 ERB rate 以上の帯域を置換することにより声質変換の効果が得られることを意味している。

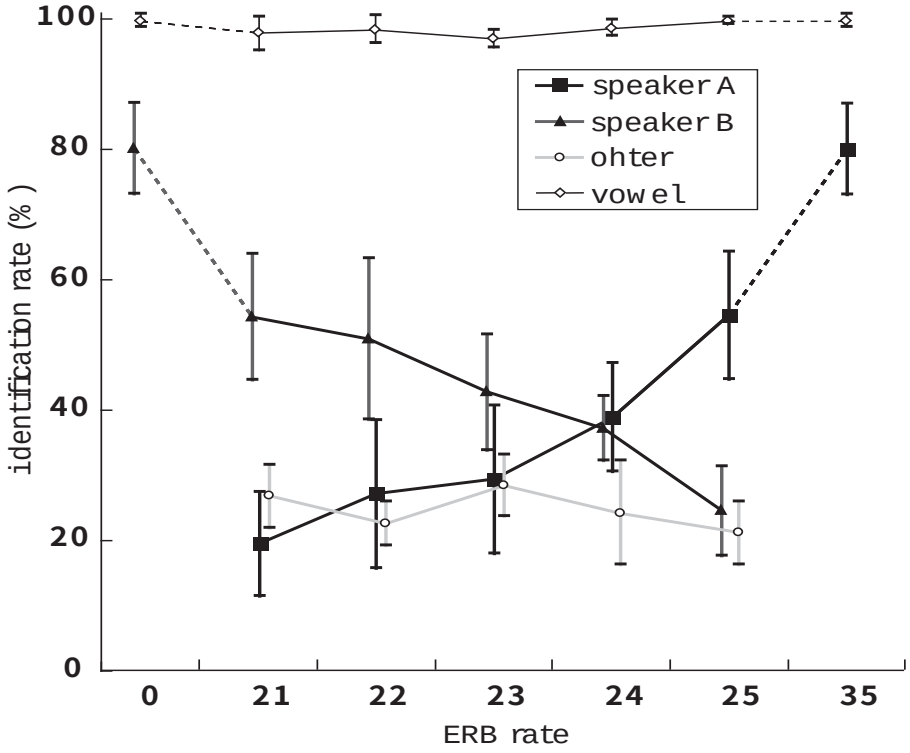

図 3.14: スペクトル包絡を置換する帯域の境界と話者識別率と音韻識別率の平均値と標準偏差の関係

3.5.5 スペクトル包絡の低域と高域の音韻が異なる場合

上述の結果は、スペクトル包絡の低域と高域の音韻を同じものにした場合のものである。この結果が高域の音韻と関係があるか否かを調べるために、高域を異なる音韻のもので置換した刺激音を作成し聴取実験を行った。3 名の話者の中から 1 名を低域の話者とし、残りの 2 名を高域の話者とした。低域と高域の境界は 22 ERB rate (2212Hz) に固定した。

実験の結果を表 3.6に示す。図 3.14の音韻識別率よりもこの表の音韻識別率が低いことから、低域と高域の音韻を異なるものにすると音韻識別が影響を受けることがわかる。これは、低域と高域のそれぞれの音韻性の影響を受けるためであると考えられる。

また、低域と高域の音韻を異なるものにすることにより、声質変換の効果もなくなることがわかる。この原因に考えられることとして、

1. スペクトル包絡の個人性は音韻によって異なる
2. 人間の音韻識別過程と話者識別過程には何らかの関係があり、音韻識別が困難になると話者識別も困難になる

という 2 つが考えられるが、この解明は今後の課題である。

表 3.6: スペクトル包絡の低域と高域の音韻が異なる場合の話者識別率と音韻識別率の平均値 (%)

話者識別率			音韻識別率
speaker A	speaker B	other	
26.9	38.7	34.5	84.2

3.6 むすび

本章では、スペクトル包絡における個人性の周波数軸上での分布を調査した。まず、周波数軸上におけるスペクトル包絡の話者間の分散と音韻間の分散を計算し、前者は22 ERB rate以上の帯域、後者は12〜25 ERB rateの帯域で値が大きくなることを明らかにした。

さらに、ABX法とNaming法による3つの聴取実験により、周波数軸上で個人性が顕著に現れる帯域を調査した。そして、スペクトル包絡の高域には個人性が顕著に現れることを示し、これを利用して話者変換が可能であることを示した。次章では、この帯域の中で話者識別に寄与する物理量の検討を行う。

第 4 章

話者識別に寄与する物理量の検討

4.1 まえがき

前章にて単母音のスペクトル包絡における個人性は高域に顕著に現れることを示した。本章では、その帯域の中のどの物理量が話者識別に寄与するのか、すなわちどの物理量に個人性が顕著に現れるのかについて検討する。ここでは特に、

1. スペクトル包絡の**微細構造** (実験 4-1)
2. スペクトル包絡高域の**ピーク**と**ディップ** (実験 4-2)

について検討する。

従来、話者識別にはどの程度細かいところまでのスペクトルの情報が必要なのか調べられたことがなかった。しかし、話者認識や声質変換に用いる特徴パラメータを特定するという観点からも、このことを明らかにすることが必要である。そこで、実験 4-1 では、LMA フィルタの作成に用いる FFT ケプストラムの次数と話者識別率の関係を求め、話者識別にはどの程度細いスペクトルの情報が必要なのかを調べる。

実験 4-2 では、スペクトル包絡の F3 以上の帯域におけるピークとディップが話者識別に与える影響について調査する。3.4節の実験 3-2 よりスペクトル包絡のピークとディップの位置関係が話者識別に重要な意味を持っていることが示唆された。また、ピークとディップはスペクトル包絡の全体的な形状を決定していることから、知覚的に重要であることが予想される。そこで、スペクトル包絡の F3 以上の帯域のピークまたはディップを除去した音声を刺激音として聴取実験を行い、これらが話者識別に与える影響について調査する。

4.2 実験 4-1 スペクトル包絡の微細構造と話者識別の関係

4.2.1 目的

本節では、スペクトル包絡のどの程度細かな形状が話者識別に必要なのかを調べる。LMA フィルタの作成に用いる FFT ケプストラムの次数を変化させて、スペクトル包絡の微細構造と話者識別の関係を調査する。

4.2.2 実験条件

音声データ

音声データは 2.2節の実験 2-1 で用いたものと同じ、男性 5 名による 5 母音である。

刺激音

刺激音は実験 2-1 の刺激音 D、すなわち話者間でスペクトル包絡のみが異なる LMA 分析合成音声を用いた。そして、LMA フィルタの作成に用いる FFT ケプストラムの次数を 10、15、20、25、30、40、50、60 次の 8 段階に変化させた。高い次数の FFT ケプストラムを用いるほど、原音声のスペクトルの細い形状を保存した音声になる。図 4.1にスペクトル包絡の作成に用いる FFT ケプストラムの次数とスペクトル包絡の形状の関係を示す。

被験者

被験者は実験 2-1 と同じ、音声データの集録の対象とした話者と日頃接している男性 7 名、女性 1 名の計 8 名である。

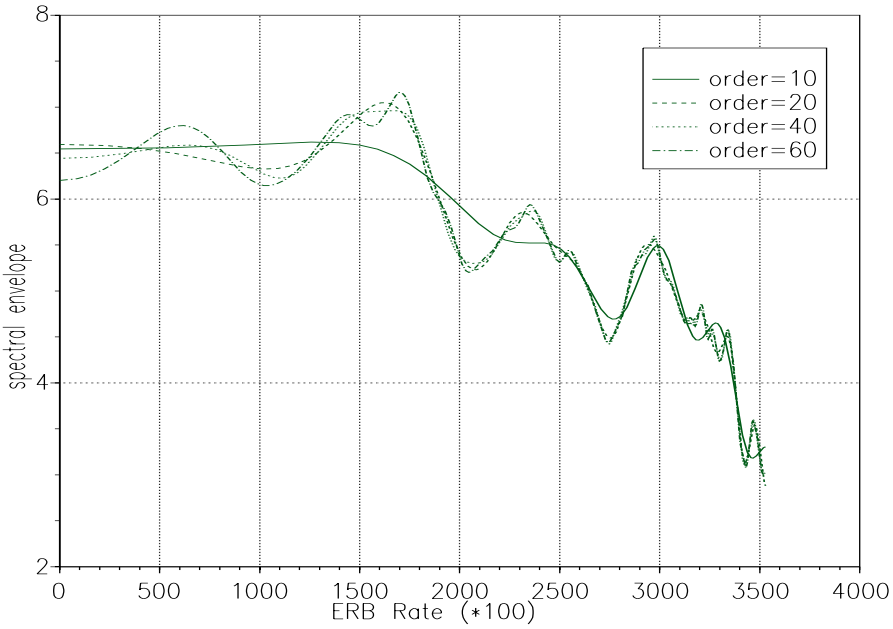

図 4.1: スペクトル包絡の作成に用いる FFT ケプストラムの次数とスペクトル包絡の形状の関係 (音声データ：ATR 音声データベースの話者 mau による/a/)

実験方法

実験では 5 話者の 5 母音を各 3 回ランダムに約 5 s 間隔で呈示した。被験者は防音室内でヘッドフォンにより受聴した。受聴は各被験者の聞きやすいレベルによる両耳受聴である。そして、回答用紙に書いてある話者と音韻を選択する。ただし、判断不可能の場合に限り “X” と回答すること (棄却) を許す。聴取実験に使用した回答用紙と機器は実験 2-1 と同じである。

4.2.3 実験結果と考察

図 4.2に LMA フィルタの作成に用いる FFT ケプストラムの次数と被験者間で平均した話者識別率、音韻識別率の関係を示す。話者識別率と音韻識別率に関して有意水準 5% で F 検定を行った ($F(1, 14) = 4.60, p < .05$)。結果を以下に述べる。

第 1 に、FFT ケプストラムの次数の違いによって話者識別率に有意差があるか否かを検定した。その結果、30 次と 25 次の間には有意差がないが ($F(1, 14) = 0.21$)、25 次と 20 次の間には有意差がある ($F(1, 14) = 6.03$) ことがわかった。

第 2 に、FFT ケプストラムの次数の違いによって音韻識別率に有意差があるか否かを検定した。その結果、30 次と 25 次の間 ($F(1, 14) = 0.43$)、25 次と 20 次の間 ($F(1, 14) = 4.23$) には有意差がないが、20 次と 15 次の間には有意差がある ($F(1, 14) = 10.51$) ことがわかった。

検定の結果から、FFT ケプストラムの次数を下げていくと、話者識別率は 25 次と 20 次の間に初めて差が現れ、音韻識別率は 20 次と 15 次の間に初めて差が現れることがわかる。これは、話者識別には音韻識別よりも細かいスペクトル包絡の情報が必要であることを意味している。また、25 次の FFT ケプストラムから作成するスペクトル包絡はスペクトルの全体的な形状を反映しているものであることから、この全体的な形状に個人性が現れて

いることがわかる。しかし、ほとんどの被験者が「次数を下げるに従って話者識別が困難になる」という内観報告をしていたことから、スペクトル包絡の微細構造にも個人性が現れていることがわかる。

従来、音声認識や話者認識にFFT ケプストラムを用いる場合には、次数の決定は経験的に行われていた。しかし、この実験結果から、標本化周波数が20 kHz の場合には音声認識では次数を20 次に設定すれば母音を認識できること、話者認識では次数を25 次以上に設定する必要があることが示された。これらの次数を標本化周波数が12kHz の場合に対応づけると、20kHz の20 次は12Hz の12 次、20kHz の25 次以上は12kHz の15 次以上に対応する。

4.3 実験4-2 スペクトル包絡のピークとディップが話者識別に与える影響の検討

4.3.1 目的

第3章の実験から、スペクトル包絡における個人性は高域に顕著に現れることが明らかになっている。また、ピークとディップはスペクトル包絡の全体的な形状を決定していることから、知覚的に重要であることが予想される。そこで、本節ではスペクトル包絡のF3 以上の帯域のピークとディップが話者識別に与える影響について調査する。F3 以上の帯域に変形を加えるのは音韻識別に影響を与えないためである。

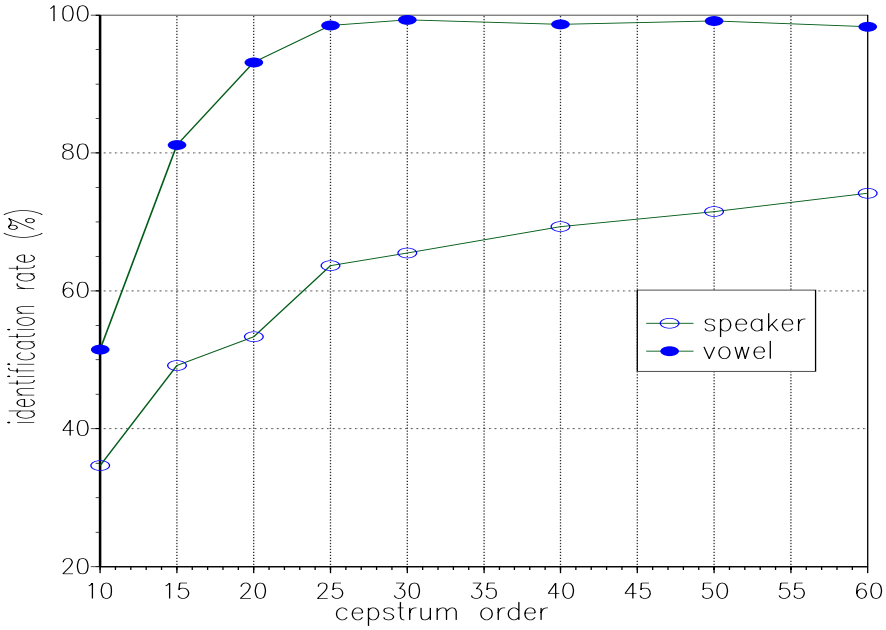

図 4.2: LMA フィルタの作成に用いる FFT ケプストラムの次数と話者識別率、音韻識別率の平均値

4.3.2 実験条件

音声データ

音声データは、基本周波数が 125 Hz 前後である 24〜26 歳の男性 5 名による 5 母音である。話者毎の基本周波数の違いが話者識別に与える影響を極力抑えるため、録音の際話者に 125 Hz の純音をヘッドフォンにより呈示し、それに声の高さを合わせるよう指示した。

録音は騒音レベル 22.7 dB(A) の防音室にて行った。マイクロフォンからの距離を約 15 cm に保って発声させた音声を防音室の外の DAT レコーダに入力し、標本化周波数 48 kHz で録音した。この音声を標本化周波数 20 kHz にダンサンプリングして WS に保存し、さらに定常部約 200 ms を切り出して音声データとした。録音に使用した機器を表 4.1に示す。

表 4.1: 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 レコーダー	SONY TDC-D10 PRO II
ヘッドフォン	STAX SR-λ pro.
ヘッドフォンアンプ	STAX SRAM-1/MK-2 pro.

刺激音

刺激音は音声データから LMA 分析合成系を用いて合成した。刺激音の平均基本周波数は 125 Hz であり、図 4.3に示す時間特性を持つものである。これ以外の分析合成に関する条件は 3.3節の実験 3-1 と同じである。

ここで、$E_{sv}(n)$ を話者 s $(s = 1, \ldots, 5)$ により発声された音韻 v $(v = 1, \ldots, 5)$ の音声

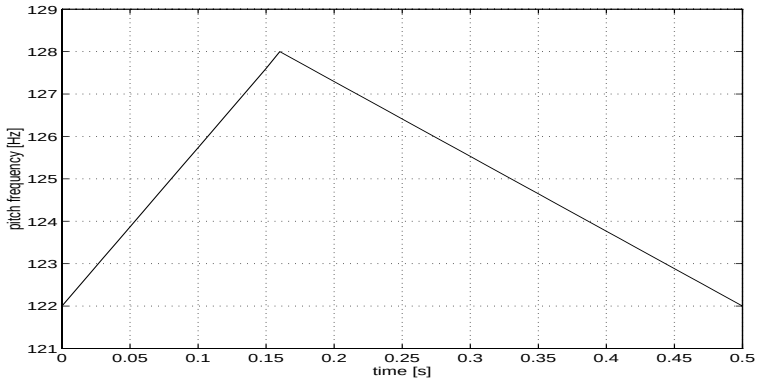

図 4.3: 刺激音の基本周波数の時間特性

データの n ERB rate におけるスペクトル包絡、$E'_{sv}(n)$ を刺激音のスペクトル包絡であるとする。さらに、$R(n)$ を F3 以上のスペクトル包絡の自己回帰直線であるとする。このとき、この実験で用いる 4 種類の刺激音 ORG、PEAK、DIP、REG は以下のように表される。なお、ここではスペクトル包絡の $R(n)$ より大きい部分をピーク、小さい部分をディップとする。

ORG LMA 分析合成音声

$$E'_{sv}(n) \quad = \quad E_{sv}(n) \tag{4.1}$$

PEAK F3 以上の帯域のディップを除去し、ピークを残した音声。スペクトル包絡において $R(n)$ より値の**小さい**部分 (ディップ) を $R(n)$ によって置換することによって作成する。

$$E'_{sv}(n) \quad = \quad \begin{cases} E_{sv}(n) & n < F3 \\ \max[E_{sv}(n), R(n)] & n \geq F3 \end{cases} \tag{4.2}$$

DIP F3 以上の帯域のピークを除去し、ディップを残した音声。スペクトル包絡において

$R(n)$ より値の**大きい**部分 (ピーク) を $R(n)$ によって置換することによって作成する。

$$E'_{sv}(n) = \begin{cases} E_{sv}(n) & n < F3 \\ \min[E_{sv}(n), R(n)] & n \geq F3 \end{cases} \tag{4.3}$$

REG F3 以上の帯域のピーク、ディップを除去した音声。F3 以上の帯域を $R(n)$ によって置換することによって作成する。

$$E'_{sv}(n) = \begin{cases} E_{sv}(n) & n < F3 \\ R(n) & n \geq F3 \end{cases} \tag{4.4}$$

/a/ の音声データをもとにした各刺激音のスペクトル包絡を図 4.4 に示す。全ての刺激音において F3 未満の帯域は各話者自身のスペクトル包絡を用いている。また、F3 は目視により決定した。なお、これらの刺激音の音韻性が保存されていることは実験前に確認してある。

被験者

正常聴力を有し、音声データの集録の対象とした話者と日頃接している 24〜29 歳の男性 6 名。前節までの被験者とは異なる。

実験方法

上述の 4 種類の刺激音をランダムに並べ変え、4 等分したものを 1 セッションとした。1 セッションは 125 個の刺激音から成っている。1 つの刺激音は 4 セッションのうちに 5 回現れる。被験者には防音室内でヘッドフォンにより受聴した。受聴は各被験者の聞き易いレベルによる両耳受聴である。被験者には聞き直しを許し、刺激音の話者を強制判断させた。回答は PC のディスプレイ上の話者の名前が書いてあるボタンをクリックすることにより行わせた [北村 96]。

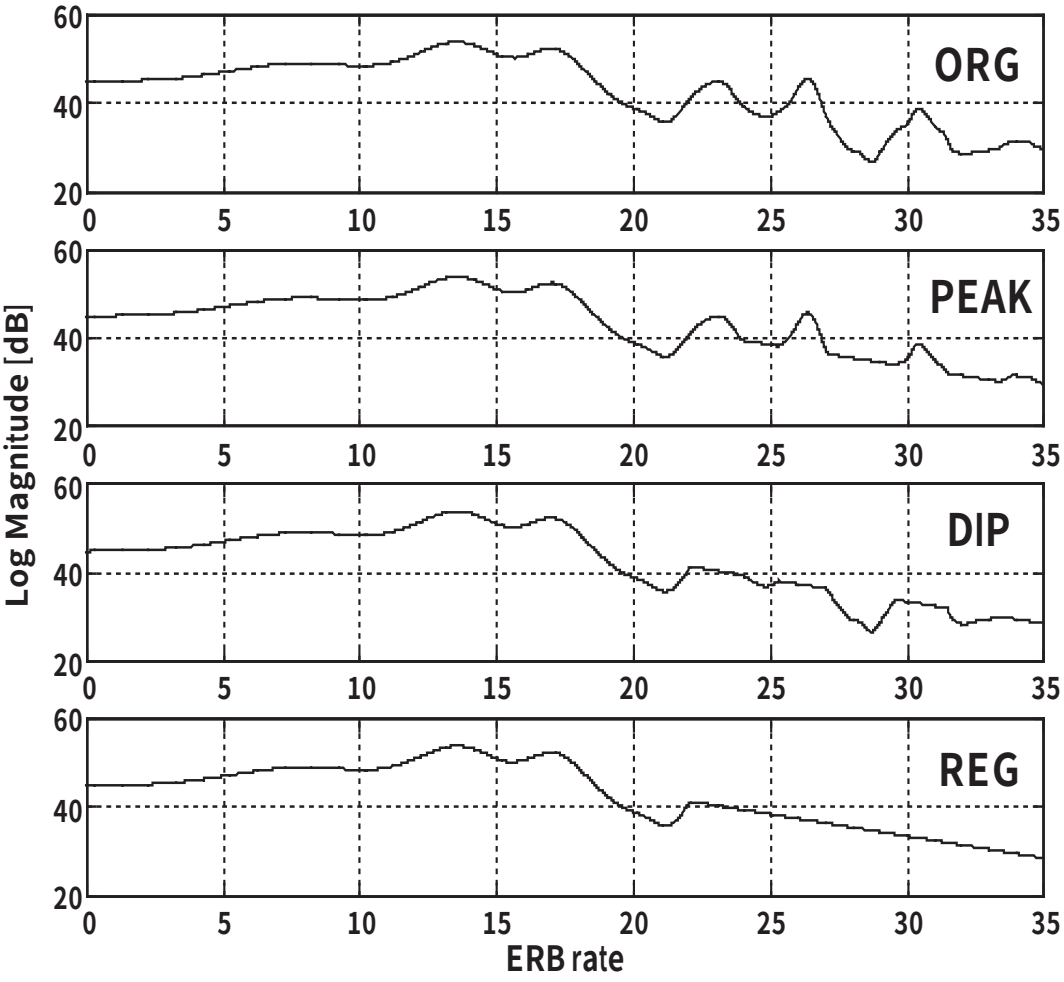

図 4.4: 実験 4-2 に用いた刺激音 ORG、PEAK、DIP、REG のスペクトル包絡 (1 段目：ORG、2 段目：PEAK、3 段目：DIP、4 段目：REG)

4.3.3 実験結果と考察

刺激音 REG の話者識別率の平均値と標準偏差を表 4.2に示す。標準偏差が比較的大きいことから、短時間のスペクトル包絡のみを利用して話者を識別する能力には個人差があることがわかる。そこで、この実験の結果は各被験者の刺激音 ORG の話者識別率から刺激音 PEAK、DIP、REG の話者識別率を減じた値により評価を行う。以下、この値を**減少値**と呼ぶ。減少値の単位はポイント (point) である。

被験者 s ($s = 1, \ldots, 6$) の刺激音 ORG の話者識別率を $I(s, \mathsf{ORG})$、刺激音 PEAK の話者識別率を $I(s, \mathsf{PEAK})$ とすると、被験者 s の刺激音 PEAK の減少値 $D(s, \mathsf{PEAK})$ は、

$$D(s, \mathsf{PEAK}) \quad = \quad I(s, \mathsf{ORG}) - I(s, \mathsf{PEAK}) \tag{4.5}$$

となり、図 4.5に示す被験者で平均した減少値 $D(\mathsf{PEAK})$ は、

$$D(\mathsf{PEAK}) \quad = \quad \frac{\sum_{i=1}^{6} I(s, \mathsf{PEAK})}{6} \tag{4.6}$$

で求められる。

これらの結果について有意水準 5% の分散分析を行ったところ ($F(1, 58) = 4.01, p < .05$)、刺激音 PEAK の減少値と刺激音 DIP の減少値の間 ($F(1, 58) = 8.45$)、刺激音 DIP の減少値と刺激音 REG の減少値の間 ($F(1, 58) = 7.27$) に有意差があった。これより、各刺激音の話者識別率が PEAK、DIP、REG の順で低くなっていくことがわかる。

この結果から、F3 以上の帯域におけるピークやディップは話者識別に寄与し、特にピークが重要であることがわかる。これは、3.4節の実験 3-2 において、スペクトル包絡のピークとディップの位置関係が話者識別に重要な意味を持っていることが示唆されたことや、人間の聴覚ではスペクトルのピークが重要であるという従来からの知見とも矛盾しない。

この実験では F3 以上のスペクトル包絡の自己回帰直線よりも大きい成分をピーク、小さ

表 4.2: 刺激音 ORG に対する話者識別率の平均値と標準偏差

	/a/	/i/	/u/	/e/	/o/
話者識別率 (%)	91.3	92.0	83.3	94.0	67.3
標準偏差	6.3	7.7	11.2	11.7	11.2

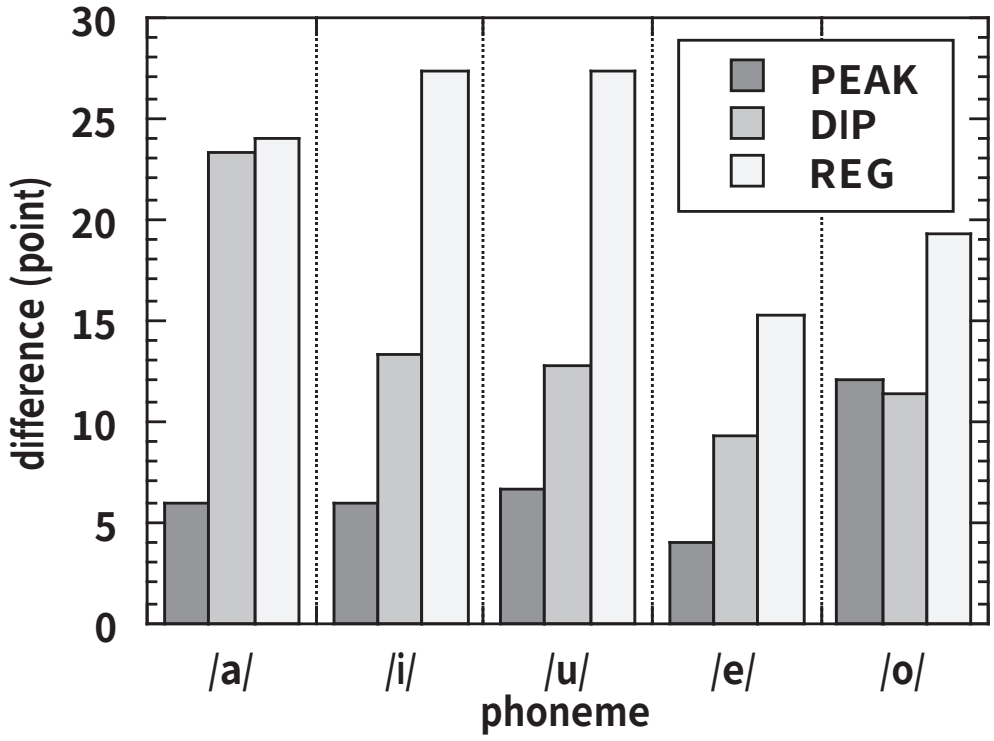

図 4.5: 刺激音 PEAK、DIP、REG の減少値

い成分をディップとしたため、ピークとディップの操作はスペクトル包絡の大まかな形状の操作に等しい。従って、この実験の結果はスペクトル包絡の大まかな形状にも個人性が現れているとした実験 4-1 の結論を支持するものである。

また、刺激音 DIP は F3 以上の帯域におけるピークの上部を除去したものと見なすことができることから、この実験の結果はピークの周波数やパワー、またピークとディップのパワー差も話者識別に寄与していることを示唆している。

4.4 むすび

本章ではスペクトル包絡において話者識別に寄与する物理量に関する検討を行った。

実験 4-1 では、スペクトル包絡の微細構造と話者識別、音韻識別の関係について検討した。聴取実験の結果、話者識別には音韻識別よりも細かいスペクトル包絡の情報が必要であることが明らかになった。具体的には、標本化周波数 20 kHz の音声データを用いる場合、音韻識別には 20 次までの FFT ケプストラムを用いれば十分であり、話者識別には 25 次以上が必要である。

実験 4-2 では、スペクトル包絡の F3 以上の帯域におけるピークとディップの話者識別への寄与について検討した。聴取実験の結果、ピークとディップは共に話者識別に寄与すること、ディップよりもピークのほうが話者識別への寄与がより大きいことが明らかになった。また、ピークの周波数やパワー、ピークとディップのパワー差も話者識別に寄与していることが示唆された。

第 5 章

スペクトルピークに着目した個人性が顕著に現れる帯域の検討

5.1 まえがき

第3章の実験より、単母音のスペクトル包絡における個人性は高域に顕著に現れることが示された。また、第4章において、話者識別にはスペクトル包絡のピークが重要な意味を持っていることが明らかになった。そこで、本章ではスペクトルピークに着目し再び個人性が顕著に現れる帯域を調査する。特に第3フォルマント (F3) と 20 ERB rate 付近のピークに着目し、F3 以上の帯域と 20 ERB rate 付近のピーク以上の帯域に着目して調査を行う。ここで、「F3 以上の帯域」とは F3 を含みそれ以上の帯域を指し、「20 ERB rate 付近のピーク以上の帯域」は 20 ERB rate 付近のピークを含みそれ以上の帯域を指す。以下、簡単のためこの表記を用いる。

日本語の母音の F3 は約 2〜4 kHz (約 21〜27 ERB rate) の範囲に現れ [三浦 80]、前章までに個人性が顕著に現れるとした帯域に含まれる。そこで、まず個人性が顕著に含まれる帯域が F3 を含むか否かを調査する。そして、この結果をもとに 20 ERB rate 付近のピークを含む帯域に関して検討を行う。

5.2　実験 5-1　F3 を含む帯域と個人性の関係

5.2.1　目的

本節では個人性が顕著に現れる帯域に F3 が含まれるか否かを調べる。

5.2.2　実験条件

音声データ

音声データは、基本周波数が 120 Hz に近い男性 3 名による 5 母音である。各話者の基本周波数の違いが話者識別に与える影響を抑えるため、録音の際話者に 120 Hz の純音に合わせて発声させた。標本化周波数は 20 kHz である。

刺激音

聴取実験に用いる刺激音の作成方法を以下に述べる。この実験は、個人性が顕著に現れる帯域に F3 が含まれるか否かを明らかにすることが目的である。そこで、スペクトル包絡の低域は話者間で加算平均したものとし、F3 以上の帯域を話者自身のものとした音声と F4 以上の帯域を話者自身のものとした音声を刺激音とする。

$E_{sv}(k)$ を話者 s $(s = 1, \ldots, S)$ によって発声された音韻 v $(v = 1, \ldots, V)$ の k $(k = 1, \ldots, K)$ 番目のフレームのスペクトル包絡であるとする。$E_{sv}(k)$ をフレームに関して加算平均したものを

$$E_{sv} = \frac{1}{K} \sum_{k=1}^{K} E_{sv}(k) \tag{5.1}$$

とする。E_{sv} を話者に関して加算平均したものを

$$E_{v} = \frac{1}{S} \sum_{s=1}^{S} E_{sv} \tag{5.2}$$

とする。そして、E_vのF3以上の帯域 (図5.1 (a)) またはF4以上の帯域 (図5.1 (b)) をE_{sv}と置換する。なお、F3とF4は目視により決定した。

刺激音は以下の5種類である。

a. 原音声

b. LMA分析合成音声

c. 平均基本周波数、基本周波数の変化の時間特性、音声波形の振幅を話者間で全て共通にし、スペクトル包絡の時間順序をランダムに並べ替えた音声。基本周波数は図2.1と同じものを用いた。

d. cの処理に加え、スペクトル包絡のF3以上の帯域を置換した音声 (図5.1 (a))。この音声には話者自身のF3が含まれる。

e. cの処理に加え、スペクトル包絡のF4以上の帯域を置換した音声 (図5.1 (b))。この音声には話者自身のF3が含まれない。

刺激音aは、被験者が原音声によって話者識別できることを確認するためのものである。刺激音bはLMA分析合成音声の品質を調べるためのものである。刺激音cでは話者により異なる物理量はスペクトル包絡のみである。この刺激音によって、被験者がスペクトル包絡の情報のみでどの程度の精度の話者識別が可能かを調べる。刺激音d、eは個人性がF3以上の帯域に含まれるか否かを調べるためのものである。

LMAフィルタの作成には60次のFFTケプストラムを用いた。データ長は約0.5 sである。

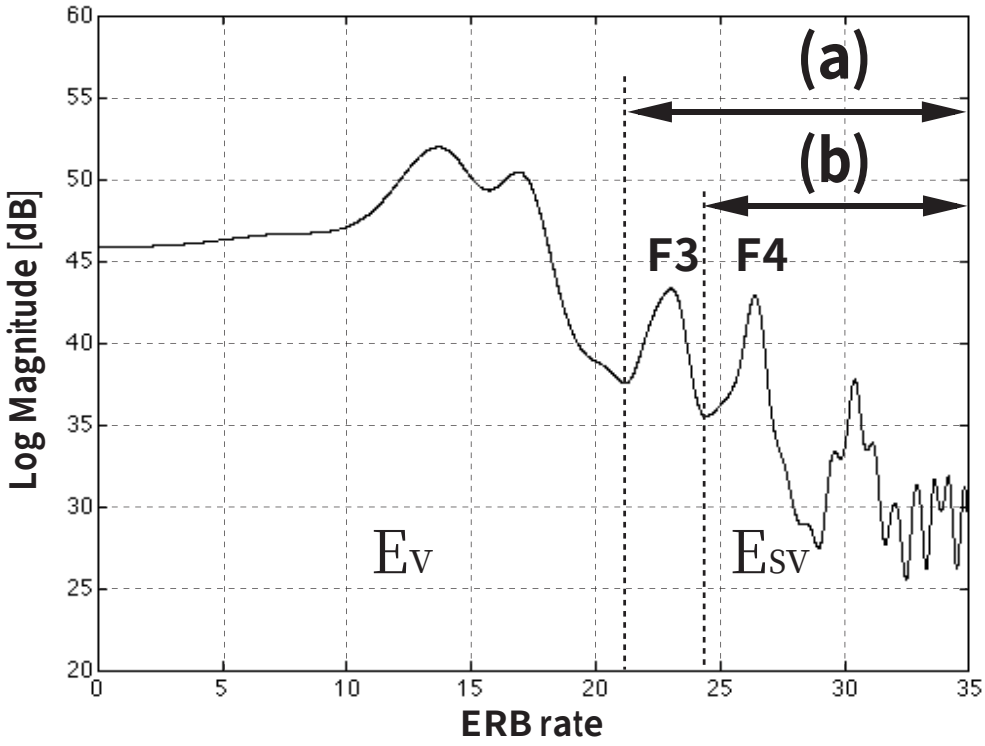

図 5.1: E_vの高域をE_{sv}と置換した/a/のスペクトル包絡。スペクトル包絡の (a) F3 以上の帯域または (b) F4 以上の帯域を置換する。

被験者

正常聴力を有し、音声データの集録の対象とした話者と日頃接している 23〜25 歳の男性 8 名。前章までの聴取実験の被験者とは異なる。

実験方法

各刺激音を 3 回ずつ呈示する。呈示の間隔は約 6 s である。被験者は防音室内でヘッドフォンにより両耳受聴し、話者と音韻を回答する。ただし、判断不可能の場合には “X” と回答することを許す。実験に用いた回答用紙を付録に示す。

5.2.3 実験結果と考察

各刺激音の音韻識別率は 99 %以上あり、全ての刺激音間に有意差がない ($F(4, 40) = 0.60$)。ここで、$F(4, 40) = 2.61, p < .05$ である。このことは、刺激音 b、c、d、e に加えた処理は音韻識別に影響を与えないことを示している。

各刺激音の話者識別率の平均値を図 5.2に示す。被験者が “X” と回答した場合には識別誤りをしたものとして識別率を求めている。これらの結果に関して有意水準 5 %の F 検定を行なったところ以下のことがわかった。なお、$F(1, 16) = 4.49, p < .05$ である。

/a/と/u/と/o/の話者識別率に関しては、刺激音 c と d の差は小さいが (/a/:$F(1, 16) = 9.67$, /u/:$F(1, 16) = 1.56$, /o/:$F(1, 16) = 3.66$)、刺激音 d と e の差は大きい (/a/:$F(1, 16) = 25.27$, /u/:$F(1, 16) = 3.45$, /o/:$F(1, 16) = 95.01$)。このことは、/a/と/u/と/o/に関しては F3 以上の帯域が話者識別に重要な意味をもつ、つまり個人性が顕著に現れることを示している。

一方、/i/と/e/の話者識別率に関しては、刺激音 c と d の間に有意差があるが (/i/:$F(1, 16) = 13.51$, /e/:$F(1, 16) = 43.40$)、刺激音 d と e の間にはない (/i/:$F(1, 16) = 0.03$, /e/:$F(1, 16) =$

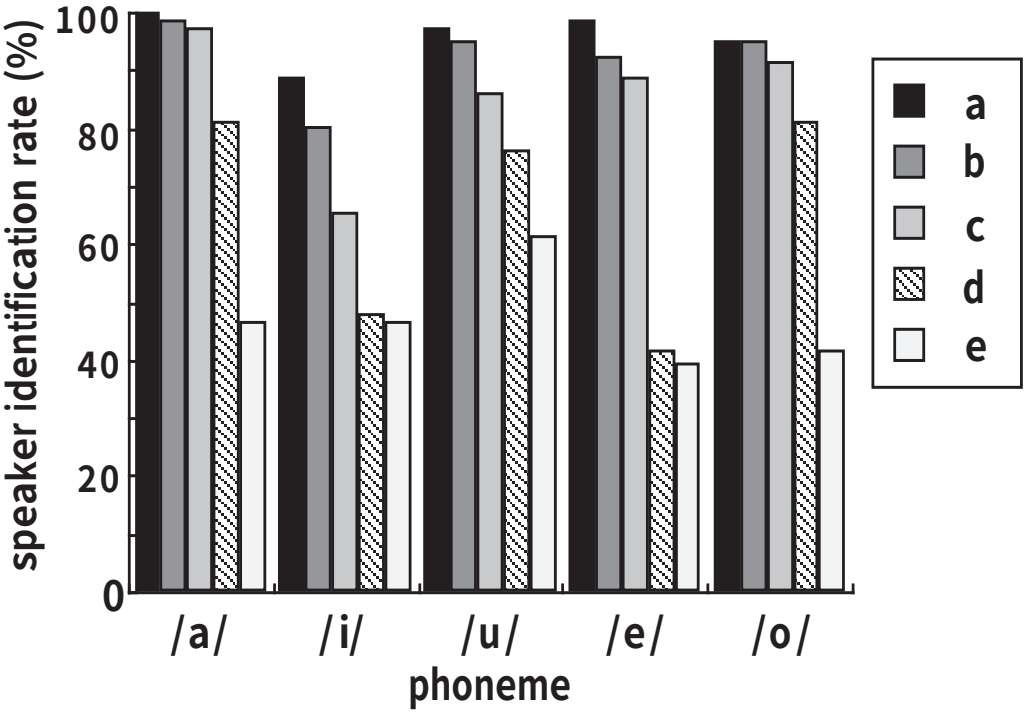

図 5.2: 各刺激音の話者識別率の平均値

0.15)。これは、/i/と/e/に関してはF3以上の帯域に個人性が顕著に現れるとはいえないことを示している。/i/と/e/の個人性が顕著に現れる帯域に関しては、次節にて改めて検討する。

5.3 実験5-2 20 ERB rate付近のピーク以上の帯域と個人性の関係

5.3.1 目的

前節で、/a/と/u/と/o/ではF3以上の帯域に個人性が顕著に現れるが、/i/と/e/では顕著に現れるとはいえないことがわかった。この結果と、/a/と/u/と/o/ではF3が現れる20 ERB rate (1740 Hz) 付近の帯域に/i/と/e/ではF2が現れることを併せて考えると、スペクトル包絡における個人性は音韻にかかわりなくこの20 ERB rate付近のピーク以上の帯域に顕著に現れることが推察される。

本節ではこの推察を確認するため、この帯域における情報で話者識別が可能か否かを調べる。さらに、スペクトル包絡の表現を簡略化し、制御を容易にすることを目的として、この帯域のピークを三角形で近似することを試みる。

以下、「20 ERB rate付近のピーク以上の帯域」を下線付きの「高域」で表し、「高域未満の帯域」を「低域」で表す。図5.3に/a/と/i/と/u/の20 ERB rate付近に存在するピークと高域と低域の範囲を図示する。「20 ERB rate付近に存在するピーク」は、/a/と/u/と/o/ではF3に相当し、/i/と/e/ではF2に相当する。

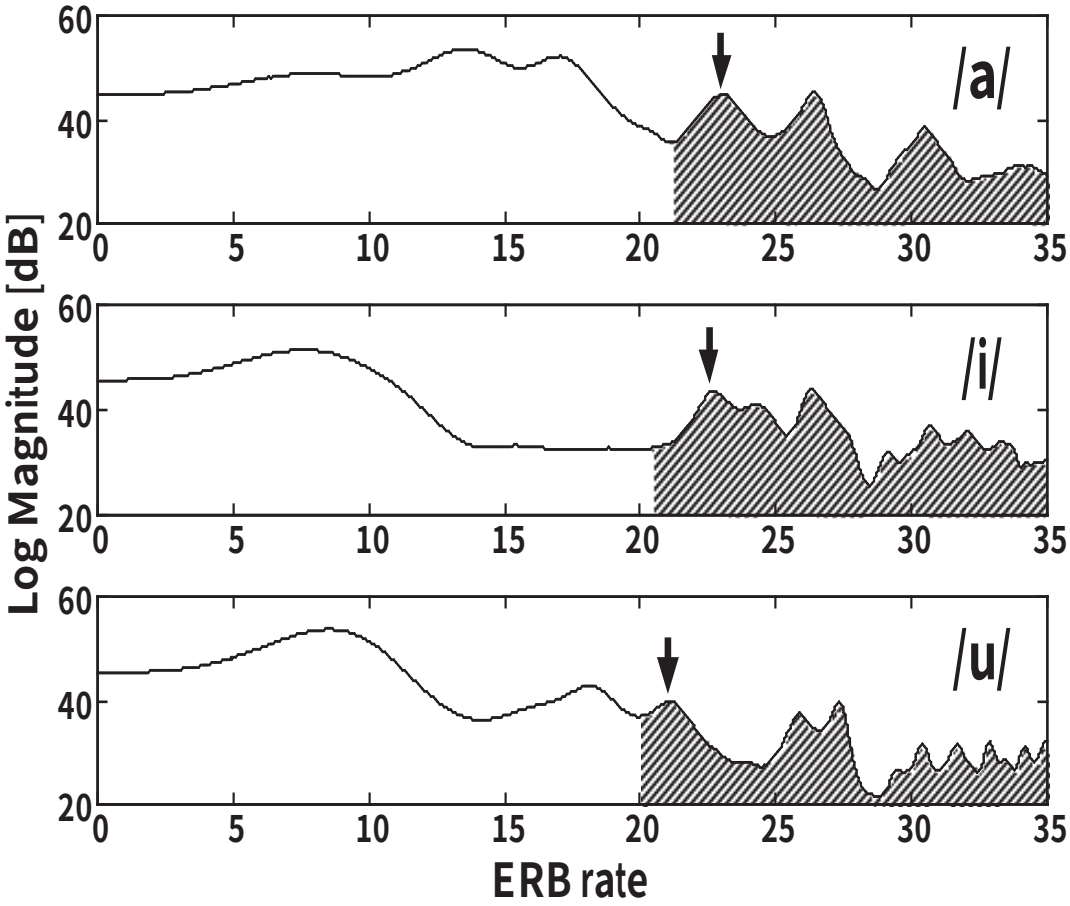

図 5.3: 20 ERB rate 付近に存在するピーク (矢印) と高域 (斜線) と低域 (白ヌキ) の範囲

5.3.2 実験条件

音声データ

4.3節の実験 4-2 と同じ、男性 5 名による 5 母音の定常部約 200 ms。

刺激音

刺激音は音声データから LMA 分析合成系を用いて合成した。刺激音の平均基本周波数は 125 Hz である。これ以外の分析合成に関する条件は 3.3節の実験 3-1 と同じである。本実験に用いた刺激音は以下の 2 種類である。これらの刺激音の音韻性が保存されていることは実験前に確認してある。

A. 以下のスペクトル包絡を持つ合成音声

低域 … 話者間で平均したスペクトル包絡

高域 … 回帰直線より**小さい**成分を回帰直線によって置換

B. 刺激音 A において高域のピークを三角形で近似した合成音声

3.3節では 0〜10 ERB rate (442 Hz) の帯域におけるパワーの違いも話者識別に寄与することも明らかになった。低域を話者間で平均する際にはこの点を考慮し、5 名の話者を 0〜10 ERB rate に大きなパワーを持つ 2 名とそれ以外の 3 名のグループに分けた。そして、スペクトル包絡の 0〜10 ERB rate の帯域はそれぞれのグループ内で加算平均したスペクトル包絡により置換し、10 ERB rate 以上の帯域は 5 名間で加算平均したスペクトル包絡により置換した。

スペクトル包絡のピークを三角形で近似する方法を図 5.4に示す。ここでピークとはスペクトル包絡の高域においてその回帰直線よりも大きい値を持つ部分のことである。まず、

スペクトル包絡の回帰直線を引く。次に、三角形の頂点となるピークの頂点を目視により決定する。最後に、その頂点とスペクトル包絡と回帰直線の交点を直線で結ぶ。これによりピークの周波数とパワーとバンド幅が大まかに近似される。この実験で用いた音声データに対しては、この方法で決定される三角形の個数が 4 個以内におさまった。図 5.5に 1 名の話者の /a/ の音声データをもとにした刺激音 A と B のスペクトル包絡を示す。

被験者

4.3節の実験 4-2 と同じ男性 6 名。

実験方法

上述の刺激音をそれぞれ 1 セッションとして実験を行った。呈示順序はランダムであり、1 つの刺激音は 5 回呈示される。呈示条件や回答方法は 4.3節の聴取実験と同じである。

5.3.3 実験結果と考察

この実験の結果も 4.3節と同様に各被験者の LMA 分析合成音声 (ORG) の話者識別率から刺激音 A、B の話者識別率を減じた値 (減少値) により評価を行う。被験者間で平均した減少値を図 5.6に示す。

刺激音 A と 4.3節の刺激音 PEAK(F3 以上の帯域において回帰直線より大きい値を持つ部分を回帰直線により置換した音声) の減少値について有意水準 5% の分散分析を行ったところ、有意差は見られなかった ($F = 3.76, F(1, 58) = 4.01, p < .05$)。刺激音 PEAK と A における大きな違いは低域を話者間で平均しているか否かである。低域を話者間で平均したことによる影響がないことから、話者識別には高域がより重要であることがわかる。

また、実験に用いた刺激音の音韻性が保存されていることから、話者に関して加算平均

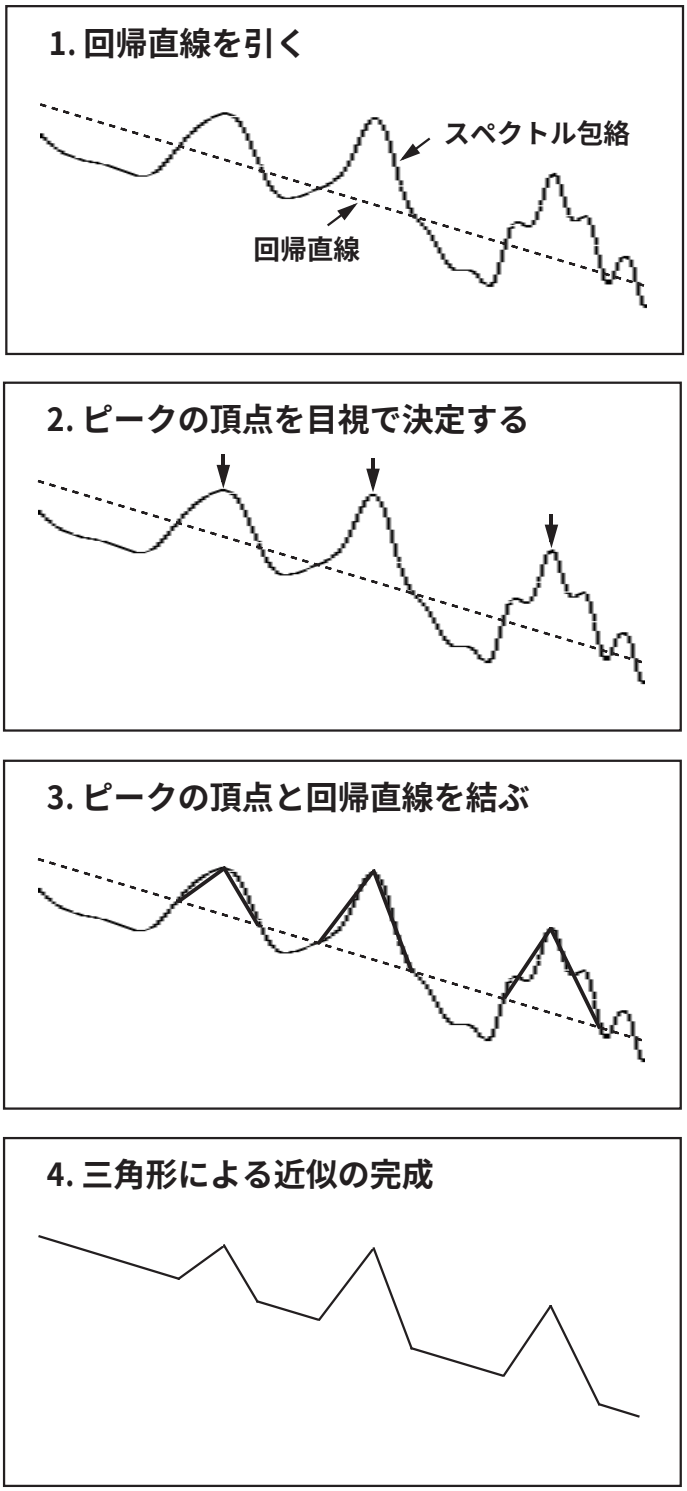

図 5.4: ピークを三角形で近似する方法

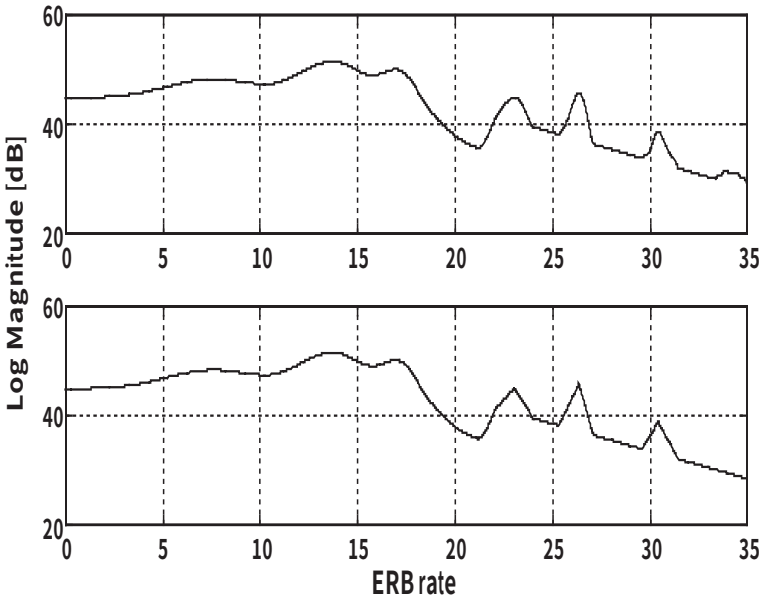

図 5.5: 刺激音 A と B のスペクトル包絡 (上：刺激音 A、下：刺激音 B)

したスペクトル包絡の高域を変換対象の話者のもので置換することにより、音韻識別に影響を与えずに声質変換ができることがわかる。

次に、刺激音 A について音韻により減少値に有意差があるか否かを調べたところ ($F(4, 25) = 2.76, p < .05$)、音韻間には有意差が見られなかった ($F = 0.91$)。従って、単母音のスペクトル包絡における個人性は音韻によらず 20 ERB rate 付近に存在のピーク以上の帯域 (高域) に顕著に現れることがわかる。高域は、Furui らが個人性知覚の心理的距離との相関が高いとした時間平滑スペクトル包絡の 2.5〜3.5 kHz の帯域を含む帯域である [Furui 85a]。よって、この実験の結果は彼らの結果を支持するものであるといえる。

一方、刺激音 A の減少値と刺激音 B の減少値の間には有意差が見られなかった ($F = 1.26$)。これはスペクトル包絡の表現を簡略化できる可能性を示唆している。また、三角形による近似はピークの周波数とパワーとバンド幅の情報を大まかに近似していると考えられるので、話者識別にはこれらの情報が重要であることを示唆しているといえる。

しかし、ほとんどの被験者から刺激音 B に対する話者識別の困難さを指摘された。これは、ピークを 1 つの三角形により近似することは個人性の劣化を引き起こすことを示している。しかし、1 つのピークを複数の三角形により近似することにより個人性の劣化を抑えつつ個人性情報の表現を簡略化することができる可能性が高く、音声合成等への応用が期待できる。

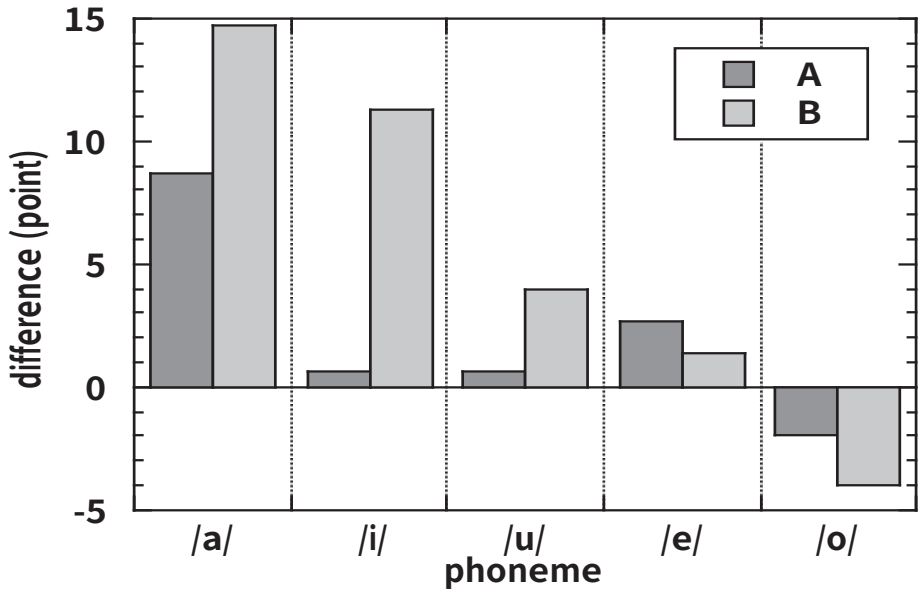

図 5.6: 実験 5-2 の減少値

5.4 むすび

本章ではスペクトルピークに着目し個人性が顕著に現れる帯域を調査した。

実験 5-1 では、F3 以上の帯域に個人性が顕著に現れるか否かを検討した。その結果、/a/と/u/と/o/に関しては顕著に現れることを示した。実験 5-2 では単母音のスペクトル包絡において話者識別に寄与する成分に関する検討を行った。その結果、単母音の話者識別にはスペクトル包絡の 20 ERB rate 付近のピーク以上の帯域 (高域) が重要な意味を持つことを示した。これは、人間が話者を識別する際にはこの帯域から個人性を抽出していることを示唆するものである。

また、この実験の結果から、高域を置換することにより声質変換が可能であることが示された。従来の声質変換に関する研究では、スペクトル包絡のどの部分を人間が話者識別に利用しているかを考慮せずに、スペクトル包絡全体の形状を変換対象の話者のものに近づけようとする手法が主流である [Kuwabara 95]。本章の結果は、人間が話者識別に利用している物理量を明らかにし、しかもそれが声質変換に利用できることを示した点において意義のあるものである。

第 6 章

連続音声中の母音における個人性に関する検討

6.1 まえがき

前章までは、単母音を対象にしてスペクトル包絡における個人性に関する検討を行った。そして、個人性が高域に顕著に現れることを明らかにした。この結果が、単母音のみでいえるのか、それとも連続音声中の母音でもいえるのかについて明らかにする必要がある。また、前章まではスペクトル包絡のみの個人性を対象にしてきたが、基本周波数の個人性との関係を明らかにする必要がある。以上の点から本章では、

1. スペクトル包絡における個人性も高域に顕著に現れるのか否か
2. スペクトル包絡における個人性と基本周波数における個人性との関係

を調べるための聴取実験を行う。

聴取実験には、スペクトル包絡に変形を加えた 3 種類のスペクトル包絡と 4 話者の基本周波数の全ての組合せで作成した刺激音を用いる。そして、スペクトル包絡と基本周波数の変化が話者識別に与える影響を定量的に調べる。

6.2 実験 6　連続音声中の母音のスペクトル包絡と基本周波数における個人性の検討

6.2.1 目的

本節では連続音声中の母音における個人性に関して、1) スペクトル包絡における個人性も高域に顕著に現れるのか否か、2) スペクトル包絡における個人性と基本周波数における個人性との関係を調べるための聴取実験を行う。

6.2.2 実験条件

音声データ

音声データは男性 4 名による/a/と/i/と/o/の 3 母音であり、各話者の各音韻につき 1 つずつ用意した。これらの母音は、「白い雲が青い屋根の上に浮かんでいる」という連続音声中の「青い」の部分から、サウンドスペクトログラムを参考に切り出したものである。音声データの長さは 50〜125 ms である。音声の際には話者に対して発声の仕方に関する指示は与えていない。

刺激音

刺激音は音声データから LMA 分析合成系を用いて合成した。ケプストラムは改良ケプストラム法により求めた。フレーム長は 25.6 ms、フレーム周期は 6.4 ms、加速係数は 1.0、近似回数は 3 である。求めたケプストラムをフレーム間で平均し、その 60 次までを用いて LMA フィルタを作成した。

刺激音の長さは 500 ms である。振幅は正規化し、さらに刺激音の立ち上がりと立ち下がりを滑らかにするため、前後 50 ms の部分を sin 関数によって重み付けした。

刺激音は表 6.1に示すスペクトル包絡を持つ ORG、LOW、HIGH の 3 種類を用意する。これらの刺激音は連続音声中の母音のスペクトル包絡における個人性も、単母音と同様に 20 ERB rate 付近のピークを含む高域に現れるか否かを調べるために用いる。表中の「平均」はスペクトル包絡を話者間で加算平均したスペクトル包絡を意味する。例えば、刺激音 LOW はある話者のスペクトル包絡の高域を話者間で加算平均したスペクトル包絡で置換したものである。

話者 YNZ の/a/のスペクトル包絡 (刺激音 ORG)、話者間で加算平均した/a/のスペクト

表 6.1: 各刺激音のスペクトル包絡

刺激音	低域	高域
ORG	本人	本人
LOW	本人	平均
HIGH	平均	本人

ル包絡、話者 YNZ の/a/の刺激音 LOW のスペクトル包絡、話者 YNZ の/a/の刺激音 HIGH のスペクトル包絡を図 6.1に示す。

刺激音は表 6.2に示す 4 話者の基本周波数の時間方向の平均値を持つ音源で駆動される。これは、話者識別におけるスペクトル包絡と基本周波数の役割を調べるためである。駆動音源における基本周波数は図 2.1と同様の時間特性を持つものである。

スペクトル包絡と基本周波数に関する処理は音韻毎に行う。従って、1 つの音韻につき、スペクトル包絡に加える操作 (3 種類)、スペクトル包絡 (4 話者)、基本周波数 (4 話者) の全ての組合せによる 48($3 \times 4 \times 4$) 種類の刺激音、3 音韻で 144 種類の刺激音を作成した。

表 6.2: 音声データの基本周波数の時間方向の平均値 (Hz)

	話者 AOK	話者 YNZ	話者 HYS	話者 KWM
/a/	95.7	104.5	132.0	128.0
/i/	127.0	111.9	178.2	193.2
/o/	142.6	123.4	186.8	165.9

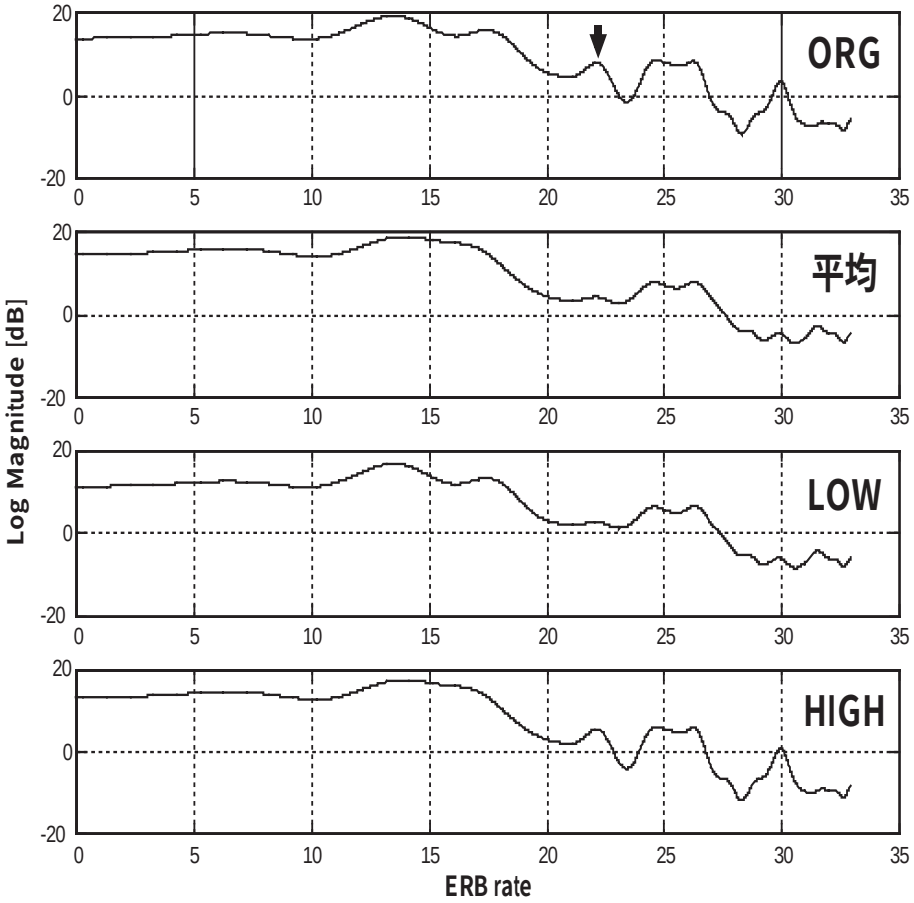

図 6.1: 変形を加えたスペクトル包絡。1 段目：話者 YNZ の/a/のスペクトル包絡 (刺激音 ORG)。矢印は「20 ERB rate 付近のピーク」を指している。2 段目：話者間で加算平均したスペクトル包絡。3 段目：話者 YNZ の/a/の刺激音 LOW のスペクトル包絡。4 段目：話者 YNZ の/a/の刺激音 HIGH のスペクトル包絡。

被験者

音声データの集録の対象とした話者と日頃接しており、正常聴力を有する 24〜31 歳の男性 7 名、女性 1 名の計 8 名。

実験方法

1 つの刺激音は 6 セッションのうちに 5 回現れる。上述の刺激音をランダムに並べ替え、6 等分したものを 1 セッションとした。1 セッションは 120 個の刺激音から成っている。

被験者は防音室内でヘッドフォンにより受聴した。受聴は各被験者の聴き易いレベルによる両耳受聴である。被験者には聴き直しを許し、刺激音の話者を強制判断させ PC を用いて回答させた。

6.2.3 実験結果と考察

被験者の中に刺激音 ORG においてスペクトル包絡と基本周波数の話者が等しい場合の話者識別率の平均値が 40 %に満たない者が 2 名いた。この被験者は母音の LMA 分析合成音声での話者識別が困難であるものとみなし、以下ではこの被験者の結果を除外した 6 名分の結果について議論する。

基本周波数とスペクトル包絡の話者が等しい場合

スペクトル包絡と基本周波数の話者が等しい場合の話者識別率を図 6.2 に示す。この結果に関して有意水準 5 %の F 検定を行った ($F(1,34) = 4.13, p < .05$)。その結果、刺激音 ORG と HIGH の話者識別率の間 ($F(1,34) = 0.51$) には有意差がないが、刺激音 ORG と LOW の話者識別率の間 ($F(1,34) = 52.13$) と刺激音 HIGH と LOW の話者識別率の間 ($F(1,34) = 38.00$) には有意差があることがわかった。このことから、連続音声中の母音と

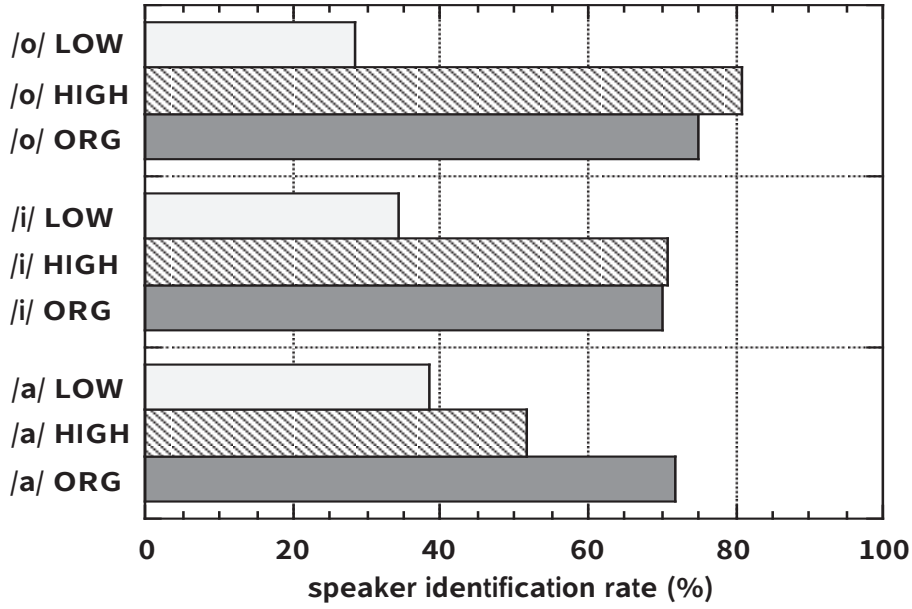

図 6.2: スペクトル包絡と基本周波数の話者が等しい場合の話者識別率の平均値

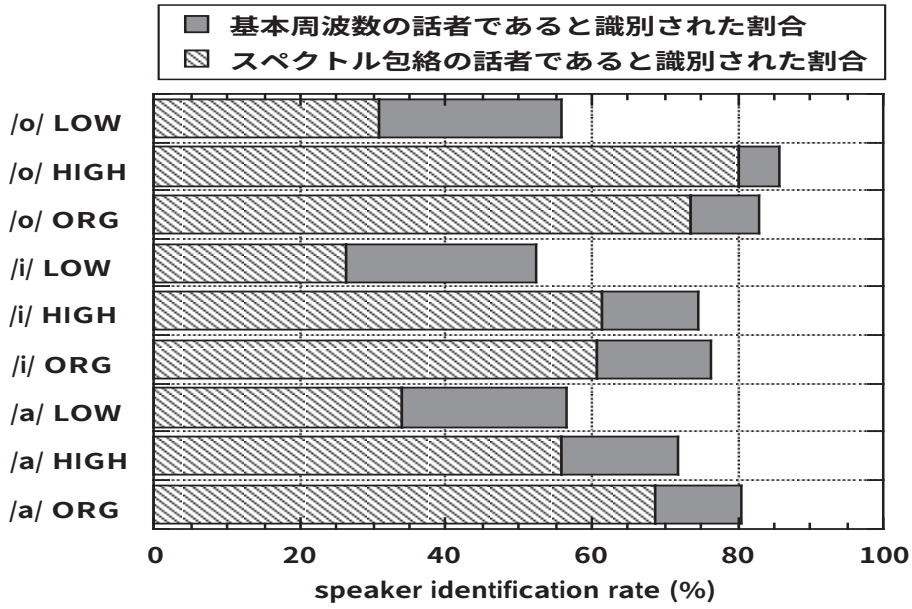

図 6.3: スペクトル包絡と基本周波数の話者が異なる場合の話者識別率の平均値

単母音とではスペクトル包絡の構造に違いがあるにもかかわらず、連続音声中の母音のスペクトル包絡における個人性は単母音と同様に高域に顕著に現れることがわかる。

本研究の音声データは、連続音声ではなくそこから切り出した母音である。そのため、連続音声におけるスペクトル包絡と基本周波数の時間特性は考慮されていない。しかし、コンピュータによる音声認識や話者認識においては、一般にフレーム単位の処理が行われることを考えると、この実験で得られた結果を応用することが可能であると考えられる。不特定話者を対象にした音声認識では、高域の重みを小さくする処理を施すことにより、話者に対する頑健性が向上すると考えられる [Mokhtari 94], [飯島 97]。逆に、話者認識では高域の重みを大きくする処理を施すことにより、認識性能が向上すると考えられる [早川 95]。

基本周波数とスペクトル包絡の話者が異なる場合

スペクトル包絡と基本周波数の話者が異なる場合の話者識別率を図 6.3に示す。刺激音 ORG において、基本周波数の話者であると識別された割合よりもスペクトル包絡の話者であると識別された割合が有意に大きい ($F(1, 34) = 413.93$)。このことから、この研究の実験条件のもとでは基本周波数よりもスペクトル包絡のほうが話者識別に寄与することがわかる。

6.3 むすび

本章では連続音声中の母音における個人性について検討を行った。話者 4 名の音声データを用い、スペクトル包絡と基本周波数の話者識別に対する役割について調べる聴取実験を行った。その結果、1) 連続音声中の母音のスペクトル包絡における個人性は高域に顕著に現れる、2) 本章の実験条件では基本周波数よりもスペクトル包絡のほうが話者識別に寄与することが明らかになった。

第 7 章

単純類似度法による話者認識に適した帯域の推定

7.1 まえがき

本章では、単母音の話者認識に適したスペクトル包絡の帯域を調べるために、**単純類似度法** (Simple similarity) による話者認識実験を行う。その際、話者認識法の弁別性能の評価関数として AD (Averaged distance) 値を用いる。そして、その帯域が前章までの聴取実験の結果と一致するか否かを調べる。さらに、同様の実験を男女各 10 名の単母音を対象に行い、より一般的な評価を行う。また、単純類似度法による話者認識を行う場合、標準パターンの音韻をどのように設定するのが良いのかについても検討を行う。

単純類似度は 2 つのパターンの形状の近さを測る距離尺度であるので、高い弁別性能が得られる帯域には話者特有の形状が現れていることになる。その帯域が聴取実験の結果と一致すれば、人間の話者識別においてもその帯域の形状が手がかりになっている可能性がある。

7.2 実験方法

7.2.1 音声データ

基本周波数が 125 Hz 前後である 24〜26 歳の男性 5 名による 5 母音を用いた。話者毎の基本周波数の違いが話者認識に与える影響を抑えるため、録音の際に話者に 125 Hz の純音をヘッドフォンにより呈示し、それに声の高さを合わせるよう指示した。

録音した音声の定常部約 200 ms を切り出して音声データとした。サンプリング周波数は 20 kHz である。音声データは各話者の各音韻につき 5 個づつ作成した。これらの音声データは、4.3節と 5.3節にて聴取実験に用いたものである。

7.2.2 標準パターンとテストパターンの作成方法

標準パターンとテストパターンの作成方法を図 7.1に示す。テストパターンは、改良ケプストラム法により求めたケプストラム 60 次を有声区間で加算平均したものから作成したスペクトル包絡である。ケプストラムからスペクトル包絡に変換する際には、ケプストラムの 0 次を一定にした。分析条件はフレーム長 25.6ms、フレーム周期 6.4ms、加速係数 1.0、近似回数 3 である。

標準パターンは、テストパターン 5 個の作成に用いたケプストラムを加算平均したものから作成したスペクトル包絡である。標準パターンは各話者の音韻毎に作成し、標準パターンとテストパターンの音韻は同じものを用いる。

7.2.3 認識方法

単純類似度法 [飯島 89] を用いた。単純類似度法は標準パターンとの単純類似度が最大となるテストパターンの話者を認識結果とする。標準パターンを $r(\boldsymbol{n})$、テストパターンを $t(\boldsymbol{n})$ とするとき、単純類似度 $S[r,t]$ は以下の式で求められる。

$$S_0[r,t] = \frac{(r,t)^2}{||r||^2||t||^2} \tag{7.1}$$

7.2.4 AD 値

話者認識法の弁別性能の評価には飯島ら [飯島 97] により提案された AD(Averaged distance) 値を用いる。単純類似度を距離尺度としたとき、AD 値は以下の式で求められる。

$$AD = \frac{\sum_i^{N_{sp}} \sum_{j\neq i}^{N_{sp}} \sum_k^{N_{set}} (S_0[r_i,t_{ik}] - S_0[r_i,t_{jk}])}{N_{sp}(N_{sp}-1)N_{set}} \tag{7.2}$$

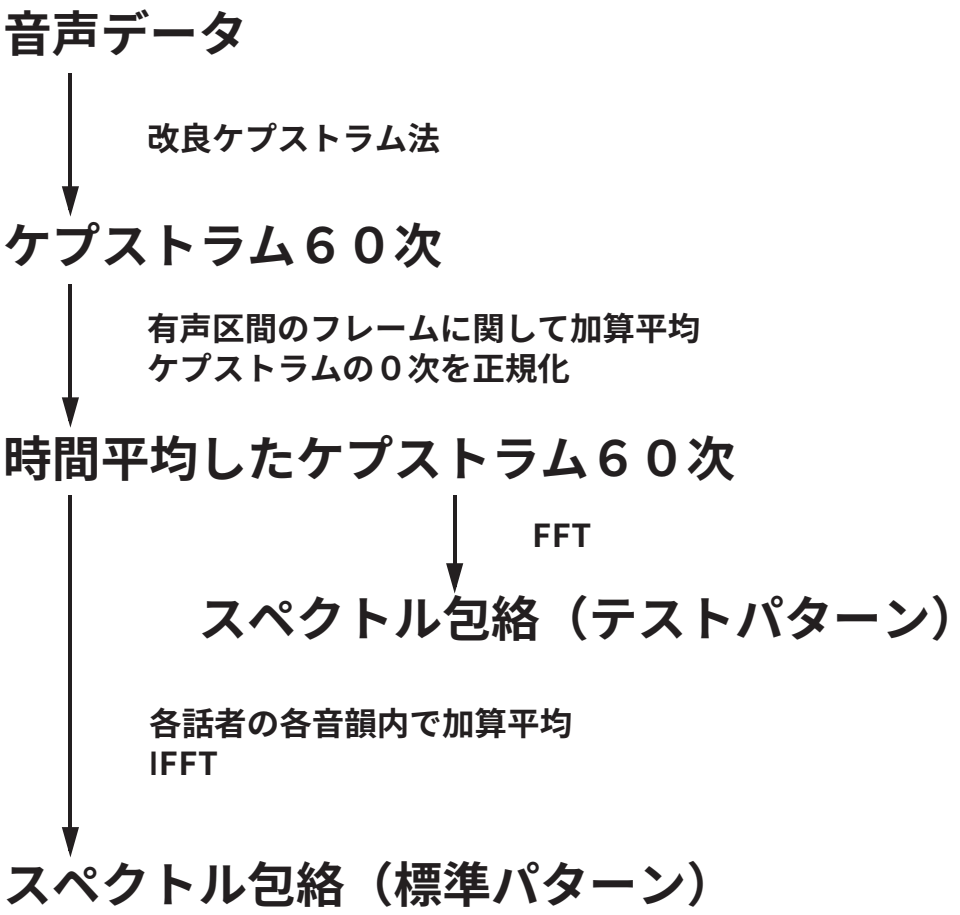

図 7.1: 標準パターンとテストパターンの作成方法

ここで、N_{sp} は話者数 (5 名)、N_{set} はテストパターンの数 (5 個)、r_i は話者 i の標準パターン ($i = 1, \ldots, N_{sp}$)、t_{ik} は話者 i の k 番目のテストパターンを表す。

AD 値は N_{sp} 人の話者における同話者内の単純類似度と異話者間の単純類似度の差の平均を求めている。つまり、AD 値が高いほど弁別性能の良い話者認識法であるといえる。また、単純類似度は 2 つのパターンの形状の近さを測るので、AD 値が高い帯域に話者特有の形状が現れていることを意味する。

7.3 実験 7-1 3 帯域の比較

7.3.1 目的

本節では聴取実験により個人性が顕著に現れることが明らかになった帯域が、単純類似度法による話者認識に適しているか否かを調べる話者認識実験を行う。

7.3.2 使用データ

話者認識に用いるスペクトル包絡の帯域として以下の 3 帯域を設定し、AD 値を求める。第 3 の帯域は個人性が顕著に現れるとされた帯域である。

1. 0〜33 ERB rate (0〜8000 Hz)
2. 0〜20 ERB rate (0〜1740 Hz)
3. 20〜33 ERB rate (1740〜8000 Hz)

7.3.3 実験結果と考察

表 7.1に上記の 3 帯域の AD 値を示す。この結果から、スペクトル包絡の 20〜33 ERB rate を用いた話者認識法が最も弁別性能が良いことがわかる。このことは、スペクトル包

絡のこの帯域は話者特有の形状を有しており、単純類似度法による話者認識に適していることを意味している。

早川ら [早川 95] はDTW (Dynamic Time Wraping) による話者認識の特徴パラメータとして用いる帯域と話者認識率との関係を調べ、高域の利用が有効であることを示している。本実験と彼らの研究では評価方法が異なるものの、スペクトル包絡の高域を用いることにより話者認識の性能が向上するという点では同じ結果が得られているといえる。

本実験ではどの帯域を用いても話者認識率が 100% であった。この結果は、スペクトル包絡の形の上では全ての帯域に個人差が存在し、これを話者認識に用いることが可能であることを示している。全ての話者認識率が 100% になったもう 1 つの理由として、変化の少ない母音の定常部を音声データとして用いたことが挙げられる。

認識率は話者認識や音声認識の分野でシステムの評価尺度として一般的に用いられている。そして、この値が大きいシステムは高性能であると評価され、この値を 100%に近づけることが目標とされている。しかし、認識率がシステムの性能を十分に表しているのかについては疑問が残る。

表 7.2に話者 AOK の/a/を標準パターンとした場合の各話者の/a/の単純類似度を示す。単純類似度はその定義より 0〜1 の値をとる。この表から、帯域の条件にかかわらず、正しく話者 AOK を認識していることがわかる。しかし、0〜33 ERB rate の場合の単純類似度は 0.890〜0.990、0〜20 ERB rate の場合は 0.910〜0.997 と変化範囲が狭く、ほとんど同じ形をしたスペクトル包絡のわずかな違いにより認識を行っている。これは、音声データの多少の変動により認識誤りが起きる可能性があることを意味している。一方、20〜33 ERB rate の場合の単純類似度は 0.162〜0.942 と変化範囲が広い。これは、違うものは違うものとして明確に区別できていることを示しており、音声データの多少の変動にも頑健

表 7.1: 3 帯域の AD 値の比較

freq. band	phoneme				
(ERB rate)	/a/	/i/	/u/	/e/	/o/
0 ∼ 33	0.077	0.110	0.068	0.103	0.046
0 ∼ 20	0.033	0.035	0.031	0.045	0.017
20 ∼ 33	0.510	0.639	0.579	0.351	0.514

表 7.2: 話者 AOK の/a/を標準パターンとした場合の単純類似度 (テストパターン: /a/)

freq. band	speaker				
(ERB rate)	AOK	IMD	KSG	UNK	YNZ
0 ∼ 33	0.990	0.890	0.869	0.903	0.925
0 ∼ 20	0.997	0.986	0.910	0.942	0.996
20 ∼ 33	0.942	0.162	0.571	0.616	0.419

であることが期待できる。

このような場合、スペクトル包絡の 0〜33 ERB rate、0〜20 ERB rate を用いた話者認識法よりも、20〜33 ERB rate を用いた話者認識法の方が高性能と言えるだろう。しかし、従来用いられてきた認識率という尺度では、このような性能を評価することができない。認識法の評価のためにこのような性能を評価できる尺度を用いる必要がある。AD 値はこの条件を満たす尺度の 1 つといえる。

7.4 実験 7-2 話者認識に適した帯域の調査

7.4.1 目的

前節ではスペクトル包絡の 20〜33 ERB rate を用いた話者認識法が最も弁別性能が良いことが明らかになった。本節では、この帯域の中で弁別性能が高い帯域をより詳細に求める。

7.4.2 使用データ

話者認識に用いるスペクトル包絡の帯域を 33 ERB rate から低域に広げた場合と帯域を 20 ERB rate から高域に広げた場合の AD 値を求める。

7.4.3 実験結果と考察

話者認識に用いるスペクトル包絡の帯域を 33 ERB rate から低域に広げた場合の AD 値を図 7.2に、20 ERB rate から高域に広げた場合の AD 値を図 7.3に示す。これらの図より、スペクトル包絡の帯域が 20〜28 ERB rate (1740〜4426 Hz)、30〜33 ERB rate (5544〜8000 Hz) における AD 値が高いことがわかる。これらの帯域には複数のスペクトルピー

クが存在しする。これらのピークは、話者識別に重要な意味を持っていることが前章までの聴取実験から明らかになっている。以上の点から、これらの帯域のピークは話者特有の形状を有し、その形状の違いは人間の話者識別の際にも重要な意味を持っている可能性があるといえる。

7.5 実験 7-3 男女各 10 名の音声データを用いた実験

7.5.1 目的

7.3、7.4節では男性話者 5 名の単母音を用いて実験を行っていた。しかし、話者数が少数のため得られた結果が話者セットに依存している可能性がある。また、これまで男性話者のみでしか実験を行っていなかったが、女性話者でも同様の結果が得られることを確認する必要がある。

7.5.2 使用データ

そこで、ATR 音声データベースの男女各 10 名による単母音 (タスクコード SY) のスペクトル包絡を用いて AD 値を求める。はじめに、0〜33 ERB rate、0〜20 ERB rate、20〜33 ERB rate の 3 帯域について AD 値を求めた。

7.5.3 実験結果と考察

男性話者 10 名、女性話者 10 名の単母音の AD 値を表 7.3、7.4 に示す。実験 7-1 と同様に、スペクトル包絡の高域のみを用いるほど AD 値が高いことがわかる。これはスペクトル包絡の高域の形状に個人性が顕著に現れていることを示唆する結果である。

高域に関してさらに詳細に AD 値を求めた。男性話者の話者認識に用いるスペクトル包

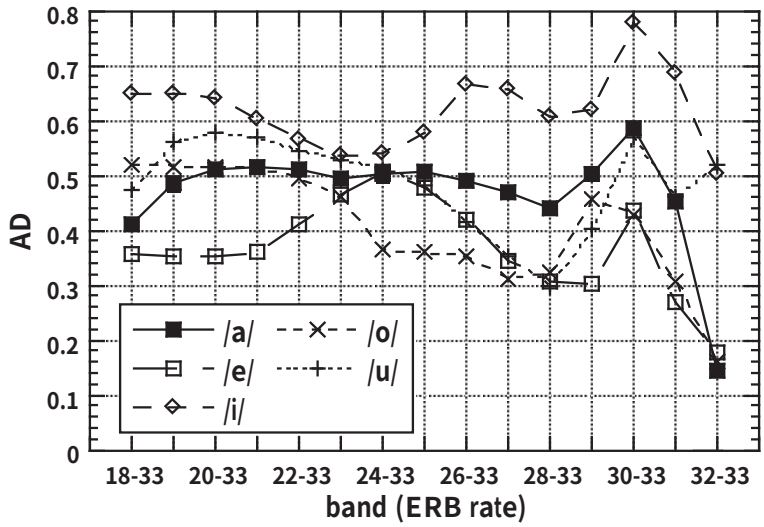

図 7.2: 話者認識に用いるスペクトル包絡の帯域を 33 ERB rate から低域に広げた場合の AD 値

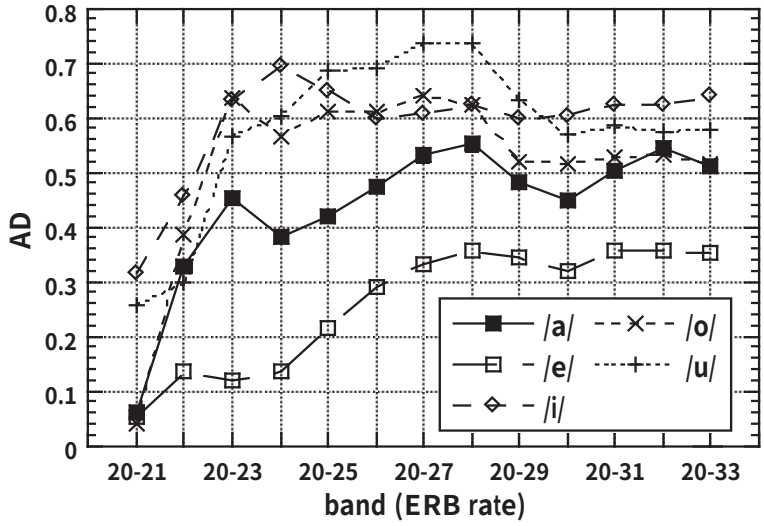

図 7.3: 話者認識に用いるスペクトル包絡の帯域を 20 ERB rate から高域に広げた場合の AD 値

絡の帯域を 33 ERB rate から低域に広げた場合の AD 値を図 7.4に、20 ERB rate から高域に広げた場合の AD 値を図 7.5に示す。同様に、女性話者の話者認識に用いるスペクトル包絡の帯域を 33 ERB rate から低域に広げた場合の AD 値を図 7.6に、20 ERB rate から高域に広げた場合の AD 値を図 7.7に示す。

図 7.5と 7.7においてグラフに右肩上がりの傾向が見られることから、スペクトル包絡の高域に個人性が顕著に現れていることがわかる。また、図 7.4、7.6 の 27〜33 ERB rate 付近で AD 値が高いことから、この付近のスペクトル包絡に話者特有の形状が現れていることがわかる。

以上の点から、単純類似度法による話者認識にはスペクトル包絡の高域を用いるのが適当であること、スペクトル包絡の高域には話者特有の形状が現れていることがより一般的に示された。

7.6 実験 7-4 標準パターンとテストパターンの音韻が異なる場合

7.6.1 目的

7.3、7.4章では標準パターンとテストパターンの音韻は同じものを用いていた。これがスペクトル包絡をパラメータとした単純類似度法による話者認識を行う際に必要な条件であるか否かを調べるために、本節では標準パターンとテストパターンの音韻が異なる場合について実験を行う。

表 7.3: 男性話者 10 名の単母音の AD 値

freq. band	phoneme				
(ERB rate)	/a/	/i/	/u/	/e/	/o/
0 ∼ 33	0.086	0.093	0.098	0.095	0.061
0 ∼ 20	0.025	0.015	0.035	0.024	0.017
20 ∼ 33	0.549	0.403	0.754	0.275	0.653

表 7.4: 女性話者 10 名の単母音の AD 値

freq. band	phoneme				
(ERB rate)	/a/	/i/	/u/	/e/	/o/
0 ∼ 33	0.100	0.137	0.126	0.123	0.082
0 ∼ 20	0.036	0.033	0.034	0.038	0.034
20 ∼ 33	0.440	0.553	0.847	0.397	0.743

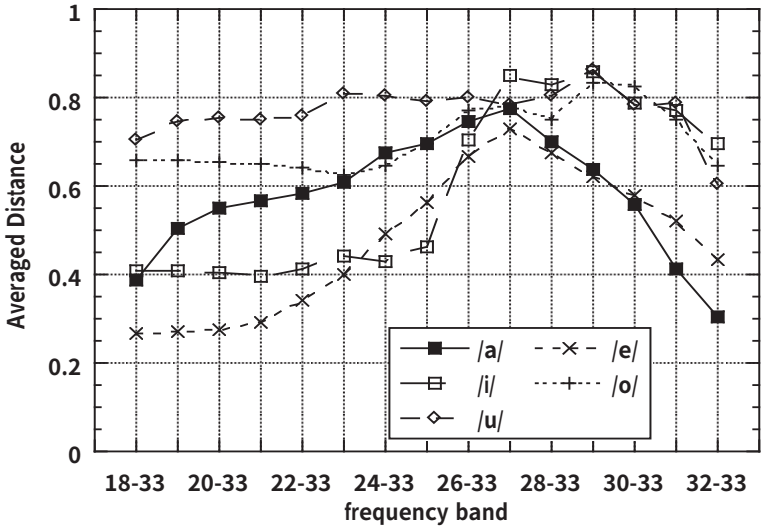

図 7.4: 男性話者の話者認識に用いるスペクトル包絡の帯域を 33 ERB rate から低域に広げた場合の AD 値

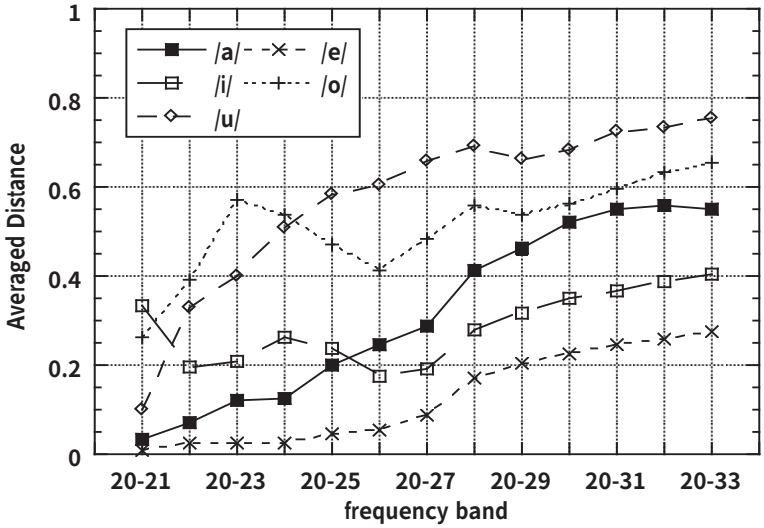

図 7.5: 男性話者の話者認識に用いるスペクトル包絡の帯域を 20 ERB rate から高域に広げた場合の AD 値

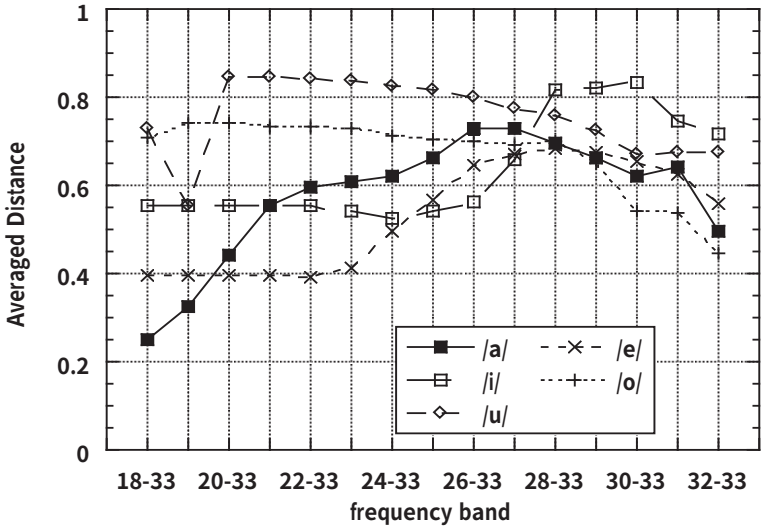

図 7.6: 女性話者の話者認識に用いるスペクトル包絡の帯域を 33 ERB rate から低域に広げた場合の AD 値

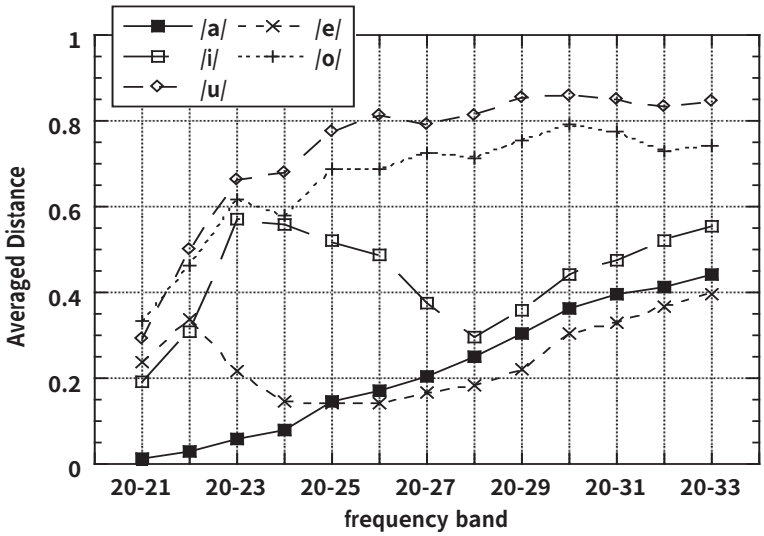

図 7.7: 女性話者の話者認識に用いるスペクトル包絡の帯域を 20 ERB rate から高域に広げた場合の AD 値

7.6.2 使用データ

音声データ、標準パターンとテストパターンは7.3、7.4章と同じものを用い、標準パターンとテストパターンの音韻を異なるものにしてAD値を求めた。AD値の計算に用いたスペクトル包絡の帯域は20〜33 ERB rateである。

7.6.3 実験結果と考察

標準パターンとして/a/と/i/を用いたときのAD値を表7.5に示す。標準パターンとテストパターンの音韻が異なる場合のAD値は同じ場合のものより小さく、その値には開きがあることがわかる。/a/と/i/以外の音韻を標準パターンとしたときも同様の結果が得られている。この結果から、スペクトル包絡の高域には音韻間に共通で話者に特有な形状の現れ方が小さいことがわかった。

さらに、話者に特有であるかを考えず、スペクトル包絡の高域に音韻間に共通の形状が現れているか否かのみを調べるために、音韻jの標準パターンと音韻k $(j \neq k)$のテストパターンの間の単純類似度の平均を式7.3で求めた。ここで、$S[r_{ij}, t_{ikl}]$は話者iにより発声された音韻jの標準パターンr_{ij}と話者iにより発声された音韻kのl番目のテストパターンt_{ikl}の単純類似度である。

$$\bar{S}_{0jk} = \frac{\sum_i^{N_{sp}} \sum_l^{N_{set}} S[r_{ij}, t_{ikl}]}{N_{sp} N_{set}} \tag{7.3}$$

標準パターンとして/a/と/i/を用いたときの単純類似度の平均を表7.6に示す。標準パターンとテストパターンの音韻が異なる場合の単純類似度が小さいことから、スペクトル包絡の高域には音韻間に共通の形状の現れ方が小さいことが示唆される。

以上の点から、スペクトル包絡をパラメータとして、単純類似度法により話者認識を行

う場合には標準パターンとテストパターンの音韻を同じものにする必要があることがわかる。また、スペクトル包絡の高域には音韻間に共通で話者に特有な形状が現れていないことが示唆される。

これらの結論は、3.5.5節の聴取実験の結果と対応している。この聴取実験では、スペクトル包絡の 22 ERB rate 以上の帯域を別の話者の別の音韻のものと置換した音声の話者を被験者に識別させた。その結果、スペクトル包絡の低域成分と高域成分の音韻が同じ場合 (3.5.4節) と異なり、スペクトル包絡の置換により話者変換の効果が得られないことが明らかになった。これも、本節の結果と同様に、スペクトル包絡に音韻間に共通した形状の現れ方が小さいことを示唆する結果である。

一方、音韻毎に異なるスペクトル包絡の微細構造の影響により、本節の分析方法では話者に特有で音韻間に共通の形状が見いだせなかったという可能性も残っている。本節で特徴パラメータとして用いたスペクトル包絡は 60 次のケプストラムから求めたものであり、これはスペクトルの微細な構造を有している。一方、4.2節のスペクトル包絡の微細構造と個人性の関係に関する聴取実験の結果が示すように、人間は話者識別の際にスペクトル包絡の全体的な形状も利用している。人間の場合、微細構造を持つスペクトル包絡からその全体的な形状を抽出し話者識別に利用することができると考えられるが、単純類似度の場合そのようなことはできない。従って、この大まかな形状に話者に特有で音韻間に共通した形状が現れている可能性もある。この点を明らかにすることは今後の課題である。

7.7 むすび

本章では単母音を対象にして、単純類似度法による話者認識に適したスペクトル包絡の帯域を求めた。そして、AD 値により話者認識法の弁別性能を評価し、弁別性能の高い帯

表 7.5: 標準パターンとテストパターンの音韻が異なる場合の AD 値 (帯域: 20〜33 ERB rate)

phoneme of	phoneme of test pattern				
ref. pattern	/a/	/i/	/u/	/e/	/o/
/a/	0.511	0.184	0.089	0.179	0.174
/i/	0.182	0.639	0.156	0.126	0.106

表 7.6: 標準パターンとテストパターンの音韻が異なる場合の単純類似度 (帯域: 20〜33 ERB rate)

phoneme of	phoneme of test pattern				
ref. pattern	/a/	/i/	/u/	/e/	/o/
/a/	0.950	0.487	0.309	0.565	0.547
/i/	0.480	0.958	0.291	0.403	0.372

域が人間の話者識別において重要な意味を持つ帯域と一致するか否かを調べた。

その結果、スペクトル包絡の 20〜28 ERB rate、30〜33 ERB rate の帯域を話者認識に用いると、高い弁別性能が得られることが明らかになった。このことはスペクトル包絡の狭帯域のみで高性能の話者認識が実現できる可能性があることを意味している。また、これらの帯域には話者識別に重要な意味を持つスペクトルピークが存在する。このことから、これらのピークは話者特有の形状を有し、その形状の違いは人間の話者識別においても重要な意味を持っていることが示唆された。

さらに、男女各 10 名の音声データを用いて同様の実験を行い、この場合でもスペクトル包絡の高域を話者認識用いると高い弁別性能が得られることが明らかになった。これにより、単純類似度法による話者認識にはスペクトル包絡の高域を用いるのが適当であること、スペクトル包絡の高域には話者特有の形状が現れていることがより一般的に示された。

加えて、標準パターンの音韻についての検討を行い、単純類似度法により話者認識を行う場合には標準パターンとテストパターンの音韻を同じものにする必要があることを明らかにした。このことからスペクトル包絡の高域には音韻間に共通で話者に特有な形状の現れ方が小さいことがわかった。この点に関してはさらなる検討が必要である。

第 8 章

全体考察

本論文では人間の個人性知覚過程を明らかにするために、個人性に関する検討を行った。その際、「人間が話者識別に利用している物理量が個人性を表す重要な物理量である」という作業仮説をおき、音声中の物理量の変化が話者識別に与える影響を聴取実験により調べ、その関係から個人性を表す物理量を求めた。

まず、第 2 章にて単母音のスペクトル包絡に個人性が存在することを確認するための聴取実験を行った。聴取実験は既知話者、未知話者の 2 つの音声データに関して行い、どちらの場合でもスペクトル包絡のみの情報で個人性知覚が可能であることが確認された。同時に、平均基本周波数と基本周波数の時間特性にも個人性が含まれていることがわかった。これは、従来の報告と一致する結果である。

次に、第 3 章にて個人性が顕著に現れるスペクトル包絡の帯域を調査した。そして、スペクトル包絡の高域には個人性が顕著に現れることを示した。従来の音声認識や話者認識でケプストラムを用いる場合には、周波数軸上で一様な重みを用いていたために、高域における個人差の影響を受けて認識精度が低下していたと考えられる。この聴取実験で得られた結果を利用して、ケプストラムに高域の重みを小さくする処理を施すことにより、個人差の影響を抑え、不特定話者を対象とした音声認識の性能向上が可能になると考えられる。一方、話者認識にケプストラムを用いる場合には、高域の重みを大きくする処理を加え、個人性を強調することにより、認識性能が向上すると考えられる。この考え方は、第 6 章の単純類似度法による話者認識において実証されている。

また、スペクトル包絡高域を利用して声質変換が可能であることを示した。現在の声質変換に関する研究では、スペクトル包絡のどの部分を人間が話者識別に利用しているかを考慮せずに、スペクトル包絡全体の形状を変換したい話者のものに近づけようとする手法が主流である [Kuwabara 95]。本章の結果は、人間が話者識別に利用している部分を明らか

にし、しかもそれが話者変換に利用できることを示した点において意義のあるものである。

第 4 章では、スペクトル包絡の高域のどの物理量が話者識別に寄与するのかについて検討した。特に、スペクトル包絡の微細構造とスペクトル包絡高域のピークとディップの話者識別への寄与について検討した。

スペクトル包絡の微細構造の話者識別への寄与に関しては、話者識別には音韻識別よりも細かいスペクトル包絡の情報が必要であることを明らかにした。具体的には、標本化周波数 20 kHz の音声データを用いる場合、音韻識別には 20 次までの FFT ケプストラムを用いれば十分であり、話者識別には 25 次以上が必要である。

従来、音声認識や話者認識に FFT ケプストラムを用いる場合には、次数の決定は経験的に行われていた。しかし、この実験結果から、標本化周波数が 20 kHz の場合、音声認識では次数を 20 次に設定すれば話者間のばらつきによる影響を抑えて母音を認識することができること、話者認識では次数を 25 次以上に設定する必要があることが定量的に示された。これらの次数を標本化周波数が 12kHz の場合に対応づけると、20kHz の 20 次は 12Hz の 12 次、20kHz の 25 次以上は 12kHz の 15 次以上に対応する。

スペクトル包絡高域のピークとディップの話者識別への寄与に関しては、ピークとディップは共に話者識別に寄与すること、ディップよりもピークのほうが話者識別への寄与がより大きいことが明らかになった。これは、人間の聴覚ではスペクトルのピークが重要であるという従来からの知見とも矛盾しない。また、ピークの周波数やパワー、ピークとディップのパワー差も話者識別に寄与していることが示唆された。

第 5 章では、スペクトルピークに着目して個人性が顕著に現れる帯域を調査した。そして、単母音のスペクトル包絡における個人性は音韻によらず 20 ERB rate 付近のピークを含む高域 (高域) に顕著に現れることを示した。この帯域は、Furui らが個人性知覚の心理

的距離との相関が高いとした時間平滑スペクトル包絡の 2.5〜3.5 kHz の帯域を含む帯域である [Furui 85a]。よって、本研究の結果は Furui らの結果を支持するものであるといえる。

さらに、党らの結果 [Dang 96a], [Dang 96b] は本研究の結果を生成系から支持するものである可能性がある。党らによれば、喉頭部における声道の分岐である梨状窩 (pyriform fassa) は 2〜6 kHz の音声スペクトルに大きな影響を与える。梨状窩は声道内で相対的に不変な部分であるため、その音響特性は個人性の要因の 1 つである可能性があるとしている。この帯域は、本研究で高域と呼んだ帯域とほぼ一致している。これは、音声の個人性に対する梨状窩の影響を示唆するものである。

第 6 章では、連続音声中の母音を対象にし、話者識別におけるスペクトル包絡と基本周波数の役割について調査した。その結果、1) 連続音声中の母音のスペクトル包絡における個人性は高域に顕著に現れる、2) 本章の実験条件では基本周波数よりもスペクトル包絡のほうが話者識別に寄与することを明らかにした。

最後に第 7 章には、前章までの聴取実験による知見をふまえ、単純類似度法による話者認識に適した帯域を調査した。話者認識法の弁別性能の評価関数として AD 値を用い、AD 値の高い帯域が聴取実験の結果と一致するか否かを調べた。その結果、スペクトル包絡の 20〜28 ERB rate、30〜33 ERB rate の帯域を話者認識に用いると、高い弁別性能が得られることが明らかになった。このことはスペクトル包絡の狭帯域のみで高性能の話者認識が実現できる可能性があることを意味している。

また、これらの帯域には前章までの聴取実験により話者識別に重要な意味を持つことがわかっているスペクトルピークが存在する。このことから、これらのピークは話者特有の形状を有し、その形状の違いは人間の話者識別においても重要な意味を持っていることが示唆された。

第 9 章

結論

9.1 本論文で明らかにされたことの要約

本論文では人間の個人性知覚過程を明らかにするために、個人性に関する検討を行った。その際、「人間が話者識別に利用している物理量が個人性を表す重要な物理量である」という作業仮説をおき、音声中の物理量の変化が話者識別に与える影響を聴取実験により調べ、その関係から個人性を表す物理量を求めた。

その結果、単母音のスペクトル包絡における個人性は 20 ERB rate 付近に存在するピークを含む帯域 (高域) に顕著に現れることを示し、この帯域を利用して声質変換が可能であることを実証した。さらに、この帯域において、ピークとディップは共に話者識別に寄与すること、ディップよりもピークのほうが話者識別への寄与がより大きいことを示した。また、ピークの周波数やパワー、ピークとディップのパワー差も話者識別に寄与していることが示唆された。加えて、単純類似度法による話者認識に適した帯域についての検討を行い、スペクトル包絡の高域をパラメータとすることによって高い弁別性能が得られることを示した。

9.2 今後の課題

以下に、今後の課題を列挙する。

1. 話者数の問題

   本研究の Naming 法による聴取実験で用いた音声データの話者はいずれも 3〜5 名と少数である。そのうえ、話者は男性のみであった。そのため、実験結果が話者セットに依存している可能性は否めない。今後、大規模な話者セットによる聴取実験を行い、本研究で得られた結果が一般的なものか否かを検証する必要がある。

2. 連続音声における個人性

本研究で対象にしたのは母音定常部であったが、今後は連続音声における個人性に関する検討を行う必要がある。第 1 に、本研究で示した母音定常部のスペクトル包絡に関する結果が、連続音声にもあてはまるか否かを検証する必要がある。第 2 に、連続音声ではスペクトル包絡や基本周波数の時間特性が話者識別へ与える影響が大きくなることが予想されるため、この影響について調べる必要がある。この場合、時間特性をいかに表現するのかということも課題になる。

3. 音韻に共通した個人性

3.5.5節の聴取実験からスペクトル包絡の低域と高域を異なる音韻のものにすると話者識別が困難になるという結果が得られた。また、7.6節でもスペクトル包絡の高域に音韻間に共通の形状の現れ方が小さいことが示された。これらの結果は、音韻間に共通の個人性が存在しない可能性を示唆している。この場合、人間は音韻毎に個人性を学習し記憶するという非常に効率の悪い処理を行っていることになるが、このようなことは考えにくい。人間は、何らかの音韻間に共通した個人性を抽出して個人性知覚を行っていると考えるのが自然である。今後、この音韻間に共通した個人性の解明を行う必要がある。

4. 音声生成系との対応関係

本研究ではスペクトル包絡における個人性は高域に顕著に現れることを示したが、その結果を音声生成系と対応させる必要がある。すなわち、音声生成系のどの部分が高域の個人性を生み出すのかを明らかにする必要がある。これを明らかにできれば、上述した音韻間に共通の個人性の解明も期待できる。党らによる梨状窩の音響特性に関する報告は、スペクトル包絡における個人性と音声生成系の関係の研究に大きな示

唆を与えるものである [Dang 96a], [Dang 96b]。今後もさらなる研究が必要である。

5. 話者識別過程と音韻識別過程のかかわり合い

   3.4節の聴取実験から、スペクトル包絡の 12〜22 ERB rate に対する変形は音韻識別のみならず話者識別にも影響を与えることがわかった。これは、人間の話者識別過程は音韻識別過程との何らかの関係があることを示唆している可能性がある。

   音韻識別に話者適応が重要な役割を果たしていることを示す報告は多いが [加藤 88], [Magnuson 94]、話者識別に音韻識別がどのような意味を持っているのかについては現在のところ明らかになっていない。話者識別過程の解明、モデル化のためにも検討を行う必要があると考える。

謝辞

本研究を行うにあたり、北陸先端科学技術大学院大学 情報科学研究科の赤木正人助教授に熱心に御指導いただきましたことを深く感謝いたします。また、折に触れて御指導、御討論いただきました、北陸先端科学技術大学院大学 情報科学研究科の飯島泰蔵教授、岩城護助手に深く感謝いたします。さらに、本論文の草稿の段階から丁寧に御指導いただきました、金沢工業大学の垣田有紀教授に深く感謝いたします。その他、北陸先端科学技術大学院大学の開学以来の学生の皆様、特に赤木、飯島研究室の皆様に厚く御礼申し上げます。

なお、本研究の一部は文部省科学研究費補助金 (No. 07680388) 及び特別研究員奨励費 (No. 6157) によって行われたものであります。ここに感謝の意を表します。

最後に、常に励ましてくださった多くの皆様に心から感謝申し上げます。

参考文献

[Abe 90] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization”, J. Acoust. Soc. Jpn. (E), Vol. 11, No. 2 (1990)

[阿部 95] 阿部匡伸, “基本周波数とスペクトルの漸次変形による音声モーフィング”, 音響講論 (秋), pp. 259-260 (1995)

[Abe 96] M. Abe, “Speech morphing by gradually changing spectrum parameter and fundamental frequency”, Proc. of ICSLP 96 (1996)

[赤木 94] 赤木正人, “聴覚フィルタとそのモデル”, 信学誌, Vol. 77, No. 9, pp. 948-956 (1994)

[Akagi 97] M. Akagi and T. Ienaga, “Speaker individuality in fundamental frequency contours and its control”, J. Acoust. Soc. Jpn. (E) (in printing)

[天野 91] 天野成昭, “実験計画法と一対比較法”, 音響学会第 44 回技術講習会資料 (1991)

[Bregman 90] A. S. Bregman, “Auditory Scene Analysis The perceptual organization of sound”, MIT press (1990)

[党 95] 党建武, 本多清志, “母音発声時の音声スペクトルに対する梨状窩の影響”, 信学技報, SP95-10 (1995)

[Dang 96a] J. Dang and K. Honda,“An improved vocal tract model of vowel production implementing piriform presonance and transvelar nasal coupling”, Proc. of ICSLP 96 (1996)

[Dang 96b] J. Dang and K. Honda “Acoustic characteristics of the piriform fossa in models and humans”, J. Acoust. Soc. Am., Vol. 101, No. 1, pp. 456-465 (1996)

[Francis 96] A. L. Francis and H. C. Nusbaum, “Paying attention to speaking rate”, Proc. of ICSLP 96 (1996)

[古井 81] 古井貞熙, “話者認識”, 音響誌, Vol. 37, No. 5, pp. 234-238 (1981)

[Furui 85a] S. Furui and M. Akagi, “Perception of voice individuality and physical correlates”, 聴覚研資, H85-18 (1985)

[Furui 85b] 古井貞熙, “ディジタル信号処理”, 東海大学出版会 (1985)

[Furui 86a] S. Furui, “Research on individuality features in speech waves and automatic speaker recognition techniques”, Speech Commun., Vol. 5, No. 2, pp. 183-197 (1986)

[古井 86b] 古井貞熙, “音声知覚研究とその音声情報処理への応用”, 音響誌, Vol. 42, No. 12, pp. 953-958 (1986)

[Glasberg 90] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data”, Hearing Research, 47, pp. 103-138 (1990)

[Greenwood 90] D. Greenwood, “A cochlear frequency-position function for several species - 29 years later”, J. Acoust. Soc. Am., Vol. 87, No. 6, pp. 2592-2605 (1990)

[Grimm 93] Laurence G. Grimm, “Statistical applications for the behavioral sciences”, John Wiley & Sons, Inc. (1993)

[橋本 95] 橋本誠, 樋口宜男, “個人性の知覚に影響を及ぼす音響的特徴の分析”, 音響講論 (春), pp. 323-324 (1995)

[橋本 96] 橋本誠, 樋口宜男, “音声の個人性知覚における既知話者/未知話者の影響”, 音響講論 (秋), pp. 263-264 (1996)

[早川 95] 早川昭二, 板倉文忠, “音声の高域に含まれる個人性情報を用いた話者認識”, 音響誌, Vol. 51, No. 11, pp. 861-868 (1995)

[早川 96] 早川昭二, 板倉文忠, “線形予測誤差に含まれる個人性情報を用いた話者認識”, 信学技報, SP96-48 (1996)

[飯島 89] 飯島泰蔵, “パターン認識理論”, 森北出版 (1989)

[飯島 97] 飯島泰蔵, 岩城護, 北村義敬, “正規型自然観測法理論による単純類似度の多重化法 –不特定話者の母音認識への適用–”, 信学論 (印刷中)

[今井 78] 今井聖, 北村正, “対数振幅特性近似フィルタを用いた音声の分析合成系”, 信学論, Vol. J61-A, No. 6, pp. 527-534 (1978)

[今井 79] 今井聖, 阿部芳春, “改良ケプストラム法によるスペクトル包絡の抽出”, 信学論, Vol. J62-A, No. 4, pp. 217-223 (1979)

[今井 80] 今井聖 “対数振幅特性近似 (LMA) フィルタ”, 信学論, Vol. J63-A, No. 12, pp. 886-893 (1978)

[伊藤 82] 伊藤憲三, 斉藤収三, “音声の音響的パラメータが個人性の知覚に及ぼす影響”, 信学論, Vol. J65-A, No.1, pp. 101-108 (1982)

[Iwahashi 95] N. Iwahashi and Y. Sagisaka, “Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis fuction networks”, Speech Commun., Vol. 16, No. 2, pp. 139-151 (1995)

[Knapp 72] M. L. Knapp, “Nonverbal communication in human interaction”, Holt, Rinehart & Winston, Inc. (1972) (邦訳 牧野成一, 牧野泰子訳, “人間関係における非言語情報伝達”, 東海大学出版会)

[粕谷 93] 粕谷英樹, “声質に寄与する音響的特徴”, 音響講論 (秋), pp. 619-622 (1993)

[粕谷 95] 粕谷英樹, 楊長盛, “音源から見た音質”, 音響誌, Vol. 51, No. 11, pp. 869-875 (1995)

[Kasuya 96] H. Kasuya, W. Zhu, M. Matsuda, and C. S. Yang, “Voice quality conversion based on an ARX speech analysis-synthesis method and its application to the study of speaker individuality”, J. Acoust. Soc. Am., Vol. 100, No. 4, Pt. 2, p. 2600 (1996)

[加藤 88] 加藤和美, 筧一彦, “音声知覚における話者への適応性の検討”, 音響誌, Vol. 44, No. 2, pp. 180-186 (1988)

[北村 96] 北村達也, “聴取実験システムマニュアル”, JAIST Tech. Memo. IS-TM-96-0002M (1996)

[桑原 86] 桑原尚夫, 大串健吾, “ホルマント周波数・バンド幅の独立制御と個人性判断”, 信学論 Vol. J69-A, No. 4, pp. 509-517 (1986)

[桑原 93] 桑原尚夫, “個人性の音響的特徴とその制御”, 音響講論 (秋), pp. 615-618 (1993)

[Kuwabara 95] H. Kuwabara and Y. Sagisaka, “Acoustic characteristics of speaker individuality: Control and conversion”, Speech Commun., Vol. 16, No. 2, pp. 165-173 (1995) Q. Lin and C. Che, “Normalizing the vocal tract length for speaker independent speech recognition”, IEEE signal processing letteres, Vol. 2, No. 11 (1995)

[松本 94] 松本弘, 丸山靖史, 井上博夫, “教師あり/教師なしスペクトル写像による声質変換”, 音響誌, Vol. 50, No. 7, pp. 549-555 (1994)

[Mokhtari 94] P. Mokhtari and F. Clermont, “Contributions of selected spectral regions to vowel classification accuracy”, Proc. of ICSLP 94, pp. 1923-1926 (1994)

[中村 89] 中村哲, 鹿野清宏, “ファジイベクトル量子化を用いたスペクトログラムの正規化”, 音響誌, Vol. 45, No. 2 (1989)

[Osaka 94] N. Osaka, “An analysis of voice quality using sinusoidal model”, Proc. of ICSLP 94, pp. 1647-1650 (1994)

[小坂 95] 小坂直敏, “Sinusoidal model を用いた母音の声質補間”, 音響講論 (秋), pp. 263-264 (1995)

[大村 95] 大村平, “実験計画と分散分析のはなし”, 日科技連 (1984)

[音響用語辞典 88] 音響学会, “音響用語辞典”, コロナ社 (1988)

[Magnuson 94] J. S. Magnuson, R. A. Yamada, and H. C. Nusbaum, “Are representations used for talker identification available for talker normailzation ?”, ICSLP 94, pp. 1175-1178 (1994)

[松井 93] 松井知子, アーウィン・ポーラック, 古井貞熙, “連続音声中の音節による個人性知覚”, 音響講論 (秋), pp. 379-380 (1993)

[三浦 80] 三浦種敏, “新版 聴覚と音声”, 電子情報通信学会 (1980)

[Mizuno 95] H. Mizuno and M. Abe, “Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt”, Speech Commun., Vol. 16, No. 2, pp. 153-164 (1995)

[Nusbaum 92] H. C. Nusbaum and T. M. Morin, “Paying attention to differences among talkers”, in Speech perception, production, and linguistic structure, pp. 113-134 (1992)

[片桐 86] 片桐滋, 東倉洋一, 古井貞熙, “単音節知覚における時間情報の役割”, 音響誌, Vol. 42, No.2 pp. 97-105 (1986)

[鈴木 85] 鈴木誠史, “音声と話者の相関関係について”, 音響誌, Vol. 41, No. 12, pp. 895-890 (1985)

[Shikano 86] K. Shikano, K. F. Lee, and R. Reddy, “Speaker adaptation through vector quantization”, Proc. of ICASSP 86, pp. 2643-2646 (1986)

[武田 88] 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 Tech. Rep. TR-I-0028 (1988)

[東倉 90] 甘利俊一監修, 中川聖一, 鹿野清宏, 東倉洋一共著, “音声・聴覚と神経回路網モデル”, 4 章, オーム社 (1990)

[X. Yang 96] X. Yang, J. B. Millar, and I. Macleod, “On the sources of inter- & intra-speaker variability in the acoustic dynamics of speech”, Proc. of ICSLP 96 (1996)

[楊 95] 楊長盛, 粕谷英樹, “母音声道形状の個人性と正規化”, 信学技報, SP95-12 (1995)

[C. Yang 96] C. S. Yang and H. Kasuya, “Speaker individualities of vocal tract shapes of Japanese vowels measured by magnetic resonance images”, Proc. of ICSLP 96 (1996)

[Yegnanarayana 96] B. Yegnanarayana, S. P. Wagh, and S. Rajendran, “A speaker verification system using prosodic features”, Proc. of ICSLP 96 (1996)

本研究に関する研究業績

論文

[1] T. Kitamura, M. Akagi, “Speaker individualities in speech spectral envelopes”, J. Acoust. Soc. Jpn (E), Vol. 16, No. 5, pp. 283-289 (1995)

[2] 北村達也, 赤木正人, “単母音の話者識別に寄与するスペクトル包絡成分”, 音響誌, Vol. 53, No. 3, pp. 185-191 (1997)

国際会議

[1] T. Kitamura, M. Akagi, “Speaker individualities in speech spectral envelopes”, Proc. ICSLP 94, Vol. 3, pp. 1183-1186 (1994)

[2] T. Kitamura, M. Akagi, “Relationship between physical characteristics and speaker individualities in speech spectral envelopes”, ASA/ASJ Joint meeting 96 (1996)

[3] T. Kitamura, M. Akagi, “Speaker individualities in speech spectral envelopes”, EUROSPEECH 97 (準備中)

研究会

[1] 北村達也, 赤木正人, “音声のスペクトル包絡に含まれる個人性について”, 信学技報 SP93-146 (1994)

[2] 北村達也, 高木直子, 赤木正人, “個人性情報を含む周波数帯域について”, 信学技報 SP95-37 (1995)

[3] 北村達也, 赤木 正人, “話者識別に寄与するスペクトル包絡の成分について”, 信学技報 SP95-144 (1996)

[4] 北村達也, 赤木正人, “連続音声中の母音に含まれる個人性について”, 音響研資 H-98-98 (1996)

一般講演

[1] 北村達也, 赤木正人, “スペクトル包絡における個人情報に関する検討”, 音響講論 (春) 3-4-10 pp. 363-364 (1994)

[2] 北村達也, 赤木正人, “スペクトル包絡に含まれる個人性を利用した話者変換”, 音響講論 (秋) 1-9-17 pp. 439-440 (1994)

[3] 北村達也, 赤木正人, “スペクトル高域成分の変形と話者識別”, 音響講論 (春) 3-9-20 pp. 397-398 (1995)

[4] 北村達也, 高木直子, 赤木 正人, “スペクトル包絡と個人性判断の関係”, 音響講論 (秋) 3-3-10 pp. 399-400 (1995)

[5] 北村達也, 赤木正人, “話者識別に寄与するスペクトル包絡の成分について”, 音響講論 (春) 2-3-6 pp. 387-388 (1996)

[6] 北村達也, 赤木正人, “単純類似度法による話者識別に適した周波数帯域の検討”, 音響講論 (秋) (1996)

[7] 北村達也, 赤木正人, “連続音声中の母音の話者識別におけるスペクトル包絡と基本周波数の役割”, 音響講論 (春) (1997)

その他

[1] 北村達也, “聴取実験システムマニュアル”, JAIST Tech. Memo. IS-TM-96-0002M (1996)

その他の研究業績

論文

[1] 好田正紀, 北村達也, “離散分布型 HMM による単語音声認識におけるビタビ best-first サーチの検討”, 信学論 Vol. J77-D-II, No. 7, pp. 1187-1197 (1994)

研究会

[1] 好田正紀, 北村達也, “離散分布型 HMM による単語音声認識における Viterbi best-first サーチの検討”, 情報処理学会 東北支部研究会 (1992)

[2] 好田正紀, 北村達也, “離散分布型 HMM による単語音声認識における Viterbi best-first サーチの検討”, 信学技報 SP92-18 (1992)

[3] 北村達也, 相川清明, “Gammatone フィルタを用いた音声認識 – 時間周波数マスキングの効果–”, 信学技報 SP94-115 (1995)

[4] 寺朱美, 北村達也, 落水浩一郎, “WWW ブラウザを利用した日本語読解支援システム”, 日本語教育方法, Vol. 3, No. 1, pp. 10-11 (1996)

一般講演

[1] 好田正紀, 北村達也, “単語音声認識における Viterbi best-first サーチの検討”, 音響講論 (春) (1992)

[2] 寺朱美, 北村達也, 落水浩一郎, “WWW ブラウザを利用した日本語読解支援システム”, 日本科学教育学会, Vol. 20, pp. 103-104 (1996)

[3] 寺朱美, 北村達也, 落水浩一郎, “日本語読解支援システム dictlinker”, 日本語教育学会 (秋), pp. 43-48 (1996)

付録　聴取実験に用いた回答用紙