HOKKAIDO UNIVERSITY

Title	音響的類似単語の雑音ロバストフレーズ音声認識に関する研究
Author(s)	宮﨑, 善行
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第15559号
Issue Date	2023-03-23
DOI	https://doi.org/10.14943/doctoral.k15559
Doc URL	https://hdl.handle.net/2115/89874
Type	doctoral thesis
File Information	Yoshiyuki_Miyazaki.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP



博士論文

音響的類似単語の雑音ロバストフレーズ音声認識に関する研究

A Study on Robust Phrase Speech Recognition for Noisy Acoustically Similar Words

北海道大学大学院情報科学研究科

メディアネットワーク専攻　情報通信ネットワーク研究室

宮﨑　善行

令和 4 年 12 月

論文要旨

第 1 章では研究の背景と目的について述べる．音声認識は，有用なヒューマンインターフェースであり，音声入力インターフェースとして機器の音声制御，ハンズフリー操作などの実現に役立つ．近年，音声認識を搭載した機器が商品化され，日常的に音声制御の可能な製品が開発されつつある．一般的に，大語彙音声認識が可能で，カーナビ等において地名等の固有名詞を認識できるとして高い有効性を示すシステムでは,雑音環境下での認識率を向上させるのが難しいために,音響処理段階では耐環境順応特性を向上させる手法が導入され，現時点でもその性能向上が研究されている．

本研究にて提案するフレーズ音声認識方式は，音響的な特徴をできる限り音声分析により見つけ，その特徴に基づいて認識結果を推定する．この場合，雑音環境下で判別の困難な認識語とは，音響的な発音が似ていて認識の難しい単語と考えられる．一般的に，SNR（signal-to-noise ratio）が低くなるにつれて，認識精度は低くなるが，各々の単語について，一様に低下するわけではない．低 SNR 環境でも認識精度が高い単語がある一方，音響的類似単語が登録単語辞書に存在する場合，SNR がそれほど低くない環境であっても認識精度が劣化する場合がある．本研究では,上記のような認識精度が劣化する場合の認識率改善手法を提案する．

第 2 章では雑音ロバスト音声認識システムの構成要素を紹介する．具体的には音響分析として，音声生成の仕組み，音声分析，学習，認識と共に，雑音ロバスト音声認識について述べる．

第 3 章では，本論文で「難認識語」と呼ぶ認識が難しい単語に対する規定について述べ，難認識語の認識精度を改善する鍵となる音響的差異区間について説明する．雑音環境下における雑音ロバスト音声認識の単語別の認識率に関するヒストグラムにおいて，交差検証法に基づいて認識率の評価を行った．

SNR が低下するにつれて，他の単語よりも認識率が著しく低下する単語が存在する．例えば,「財布」と「ライス」のように音響的類似単語がそれらに相当する．さらに,共通の音素列を含む単語もそれらに含まれるような特徴を持つ単語を「難認識語」と呼ぶ．そして，音響的類似性を定量的に評価するために，DTW（dynamic time warping）距離を適用する．DTW 距離による評価において，局所的な差異が存在する区間を音響的差異区間とし，音響的差異区間を抽出して認識することができれば，認識率の向上が期待できる．

本研究では,発話音声の中から音響的に類似しない音響的差異区間を切り出し，認識することにより，音響的類似単語の認識精度の改善を試みる．音響的差異区間の切り出しは，ビタビアルゴリズムにより，自動的に行うことができる．したがって，その区間が，単語のどの位置にあっても，そして，破裂音でなくても適用できるという利点がある．

第 4 章では，認識精度が最も高いと期待される音響的差異区間を決定する方法について説明し，音響的差異区間の決定を行うと共に，難認識語の認識を行う．単語を発声した教師音声を分析して得られる音声特徴量列について，単語の HMM（隠れマルコフモデル，hidden Markov model）に対する対数尤度をビタビアルゴリズムで計算すると，ビタビ経路を求めることができる．このビタビ経路を利用して，単語の音声から，単語の HMM の始端状態番号から終端状態番号までに対応する区間に存在する音声を切り出し，音響的差異区間候補と呼ぶ．

音響的差異区間候補の音声を教師音声とし，バウム・ウェルチアルゴリズム（Baum-Welch algorithm）により，音響的差異区間候補 HMM のパラメータを求める．難認識語，あるいは，その誤認時頻出語を発声した音声を最も高い精度で認識すると期待される音響的差異区間をその候補の中から決定する．音響的差異区間候補 HMM による難認識語の認識精度は，最大値が，従来手法による難認識語

の認識精度を超えていれば認識精度の改善を期待できる．

第 5 章では，全話者 48 名を 8 名ずつの 6 グループに分けて，1 グループ 8 名を認識話者，残りの 5 グループ 40 名を教師話者とし，交差検証法により従来手法と提案手法の認識率を評価した．提案手法の認識率は，従来手法の認識率以上の結果となった．誤認時頻出語とペアになる難認識語の認識精度に対する改善度のほうが上回っている．尚，Speech babble，および，White noise 雑音環境下についても同様の結果が得られた．

目 次

論文要旨 i

第 1 章 序論 1

第 2 章 雑音ロバスト音声認識 6

2.1 音声認識とは …… 6

2.2 音響分析 …… 7

2.2.1 音声生成の仕組み …… 7

2.2.1.1 母音生成の仕組み …… 8

2.2.1.2 喉頭原音のスペクトル …… 8

2.2.1.3 声道特性 …… 10

2.2.1.4 放射特性 …… 11

2.2.1.5 音声認識にとって重要な情報 …… 13

2.2.2 音声分析 …… 13

2.2.2.1 フレーム化 …… 14

2.2.2.2 高域強調 …… 14

2.2.2.3 窓処理 …… 16

2.2.2.4 離散フーリエ変換 …… 16

2.2.2.5 絶対値演算 …… 18

2.2.2.6 メルスケール変換 …… 18

2.2.2.7 対数演算 …… 20

2.2.2.8 離散コサイン変換 …… 20

2.2.2.9 対数エネルギー …… 22

2.2.2.10 デルタ特徴量 …… 22

2.2.3 学習 …… 23
2.2.3.1 音響モデル …… 26
2.2.3.2 ベイズの定理の適用 …… 27
2.2.3.3 初期 HMM 生成 …… 28
2.2.3.4 Forward アルゴリズム …… 33
2.2.3.5 バウム・ウェルチアルゴリズム …… 33
2.2.4 認識 …… 34
2.3 雑音ロバスト音声認識 …… 35

第 3 章 音響的類似単語とその認識困難さ 39
3.1 難認識語 …… 39
3.2 音響的差異区間 …… 49

第 4 章 音響的類似単語の雑音ロバストフレーズ音声認識 50
4.1 音響的差異区間候補の抽出 …… 50
4.2 音響的差異区間候補の音声分析 …… 52
4.3 音響的差異区間候補の学習 …… 53
4.4 音響的差異区間の決定 …… 53
4.5 難認識語認識 …… 59

第 5 章 評価 62
5.1 実験結果 …… 62
5.2 単語認識率 …… 69

第 6 章 結論 70

謝辞 71

参考文献 72

著者による発表論文 76

著作権の表示 77

図 目 次

2.1 母音生成の仕組み …… 8
2.2 Rosenberg 波 …… 9
2.3 喉頭原音の振幅スペクトル …… 9
2.4 「あ」の声道特性の振幅スペクトル …… 10
2.5 声道を通過した音声の振幅スペクトル …… 11
2.6 放射特性の振幅スペクトル …… 12
2.7 放射された音声の振幅スペクトル …… 12
2.8 音声分析処理 …… 13
2.9 フレーム化 …… 14
2.10 フレームデータ …… 15
2.11 高域強調データと窓係数 …… 16
2.12 スペクトル実部 …… 17
2.13 スペクトル虚部 …… 17
2.14 メルスペクトルと振幅スペクトル …… 18
2.15 メルスケール …… 19
2.16 対数メルスペクトルと対数振幅スペクトル …… 19
2.17 MFCC …… 21
2.18 低次 MFCC 時系列データ …… 21
2.19 デルタ特徴量 …… 23
2.20 「上」の 1 次 MFCC 時系列データ …… 24

2.21 「上」の 5 次 MFCC 時系列データ …… 25
2.22 「下」の 1 次 MFCC 時系列データ …… 25
2.23 「下」 の 5 次 MFCC 時系列データ …… 26
2.24 「上」の音声特徴量列のヒストグラム …… 29
2.25 「下」の音声特徴量列のヒストグラム …… 30
2.26 「上」の初期 HMM …… 31
2.27 「下」の初期 HMM …… 32
2.28 雑音ロバスト音声認識のフローチャート …… 35
2.29 音声分析のフローチャート …… 36

3.1 雑音ロバスト音声認識の単語別認識率のヒストグラム …… 40
3.2 「財布」と「ライス」の音声波形 …… 42
3.3 「財布」と「ライス」のスペクトログラム …… 43

4.1 音響的差異区間抽出 …… 51
4.2 音響的差異区間候補に対する認識率改善マップ …… 56
4.3 難認識語認識フローチャート …… 61

表 目 次

2.1 区間別統計データ …… 31
2.2 対数確率密度 …… 33
2.3 雑音ロバスト音声認識の音声分析条件 …… 37

3.1 単語一覧（抜粋） …… 41
3.2 DTW 距離 …… 44
3.3 難認識語と易認識語の認識率(%) …… 45
3.4 難認識語の単語別認識率(%) …… 46
3.5 混同行列 …… 47
3.6 難認識語と誤認時頻出語のペア …… 49

4.1 差異区間音声分析条件 …… 52

5.1 難認識語の単語別認識率(%) …… 64
5.2 誤認時頻出語の単語別認識率(%) …… 67
5.3 混同行列の比較 …… 68
5.4 全単語の認識率(%) …… 69

第 1 章

序論

研究背景

音声認識[1]-[3]は，有用なヒューマンインターフェースであり，音声入力インターフェースとして機器の音声制御，ハンズフリー操作などの実現に役立つ．近年，音声認識を搭載したスマートフォン，タブレット端末，掃除ロボット等が商品化され，日常的に音声制御の可能な製品が開発されつつある．

一般的に，スマートフォン等に搭載されているクラウド型の大規模サーバーで処理される音声認識は，連続音声認識方式[4]-[14]が採用されている．一方では，機器制御等を行う音声認識には，より耐雑音性能を向上させた特定単語等を認識するようなフレーズ音声認識方式[15]-[21]が用いられている．

連続音声認識方式は，大語彙音声認識が可能で，カーナビ等において地名等の固有名詞を認識できるとして高い有効性を示すシステムであるが，雑音環境下での認識率を向上させるのが難しいために,音響処理段階では,ノイズキャンセラーや，高性能な超指向性マイクを使用するなどの対策を講じることや，認識処理段階で，ミッシングフィーチャー理論[8]-[10]，[13]や，ディープラーニング手法[14]による耐環境順応特性を向上させる手法が導入され，現時点でもその性能向上が研究されている．これらの手法は，基本となる認識処理が大語彙対応のため，

雑音環境下において認識性能を上げようとした場合，近距離マイクが多くの場合の認識条件となり，さらに認識トリガとなるスイッチを押してから音声認識を開始することが考慮されている．

一方で，フレーズ音声認識方式は，認識対象フレーズ数が数百程度と，認識するためのフレーズ数に制限があるが，制限された数の認識対象のため，雑音環境下でのロバスト音声認識が比較的容易に実現でき，認識トリガを用いない場合でも，マイクから 2〜3 m 離れたところからの音声認識が可能となっている．実用性を考慮する場合には，雑音環境下で，かつ，マイクから離れたところからでも高認識率を維持する状況（例えば，SNR 10 dB の環境でも，90%以上の音声認識率）があるため，このようなシステムが必要とされている．

株式会社レイトロンでは，北海道大学大学院情報科学研究科の宮永研究室で研究・開発された雑音ロバスト音声認識アルゴリズムを搭載した音声認識エンジン「VoiceMagic」のチップ化を行い，SNR10dB の環境で 90%以上の音声認識率を実現した．

雑音ロバスト音声認識エンジンはコミュニケーションロボットへ実装し，東京ビッグサイトや幕張メッセなどの展示会において，向かいのブースでプレゼンテーションを行っている様な実環境の強雑音環境下で，1m以上離れた場所からの音声認識を実演している．

また，雑音ロバスト音声認識システムのコンシューマー向け商品として，音声認識コミュニケーションロボット「Chapit」（チャピット）を開発し，一般ユーザー向けにインターネットの販売サイトにて販売している．高齢者向けとしても，厚生労働省の「介護ロボットの開発・実証・普及のプラットフォーム事業」において認定され，高齢者施設向けとして提供している．

さらに，音声認識よるリモコン操作に特化した音声認識リモコンとして，1 つの音声認識フレーズで 20 種類のリモコンコードを出力可能な音声認識スマートコントローラ「LisPee360」（リスピーサンロクマル）を開発し，同じくインターネットの販売サイトにて提供している．例えば,「ただいま」というフレーズを認識することで，ダイニングやリビングの照明，エアコンやテレビの電源を入れる

といった操作が可能となる．

ビジネス向け製品としては，「Chapit」や「LisPee360」と同様に音声認識エンジン「VoiceMagic」として商標登録を行い，組込み型のモジュールとして提供している．シリーズとして「VoiceMagicStandard」と共に，MEMS マイク一体型のオールインワン製品として「VoiceMagicUSB」を販売している．用途としては，工場などのタッチパネル操作を行う生産現場で，手が汚れていてパネルを触れないなどの状況において，音声認識モジュールを組込むだけで声の操作のユーザーインターフェイスを搭載することを可能としている．雑音ロバスト音声認識エンジンはチップで構成されているのでオフラインでの操作が可能で，高速で低消費電力の性能を実現している．

その一方で，連続音声認識方式に基づく耐雑音単語認識方式[8]-[14]や，雑音ロバスト音声認識[15]-[21]を適用しても雑音環境下で判別の困難な認識語が多数存在する．例えば，文献[10]では，対象数を限定した認識（「0」から「9」の数字認識）の場合，ミッシングフィーチャー理論により認識性能を向上させることができることを示した．一方，逆に汎化性が低下し，大語彙認識では従来システムに比べて性能が低下すると説明している．認識対象を限定することで，性能を向上させることは古くから考えられているが，どのような制限が適しているかは，導入している手法の特性に依存しており，普遍的な考え方を見つけるのは容易ではない．

本研究の目的

本研究にて提案するフレーズ音声認識方式は，音響的な特徴をできる限り音声分析により見つけ，その特徴に基づいて認識結果を推定する．この場合，雑音環境下で判別の困難な認識語とは，音響的な発音が似ていて認識の難しい単語と考えられる[22]，[23]．これは，音響的類似単語であり，例えば，母音の配列が同じで，一箇所だけ異なる子音を含んでいる単語などとなる．これに対する高精度認

識手法は，ミッシングフィーチャー理論でも検討されているが，対象フレーズが数百〜数千程度の場合，高い認識性能を実現することが難しい．

一般的に，SNR が低くなるにつれて，認識精度は低くなるが，各々の単語について，一様に低下するわけではない．低 SNR 環境でも認識精度が高い単語がある一方，音響的類似単語が登録単語辞書に存在する場合，SNR がそれほど低くない環境であっても認識精度が劣化する場合がある．本論文では，上記のような認識精度が劣化する場合の認識率改善手法を提案する．提案手法では，雑音環境下で高い認識精度を得られない単語のペアに対して，雑音ロバスト音声認識を行った後に，両単語の中で，音響的に類似していない区間を抽出する．その後，その区間のみを，さらに詳細に特徴抽出を行って再度雑音ロバスト音声認識をすることにより，雑音環境下での音声認識率が向上できることを示す．

本論文の構成

本論文は 6 章で構成される．

2 章では，雑音ロバスト音声認識システムの構成要素を紹介する．具体的には音響分析として，音声生成の仕組み，音声分析，学習，認識と共に，雑音ロバスト音声認識について述べる．

3 章では，本研究で「難認識語」と呼ぶ，認識が難しい単語に対する規定について述べ，次に，難認識語の認識精度を改善する鍵となる音響的差異区間について説明する．

4 章では，音響的類似単語の雑音ロバストフレーズ音声認識手法を提案する．具体的には，音響的差異区間の決定を行った後に，難認識語の認識を行う．

音響的差異区間の決定では,まず音響的差異区間候補の抽出を行う必要がある．抽出した音響的差異区間候補の音声分析をし，HMM 学習した後に，音響的差異区間の決定を行う．難認識語認識は，従来手法である雑音ロバスト音声認識と音響的差異区間に対する音声認識を組み合わせた手法で行い，難認識語と誤認時頻

出語のペアである場合，音響的差異区間の音声を切り出して，音響的差異区間 HMM で認識を行い，ペアでない場合，従来手法と同じく，尤度最大の HMM の単語を認識結果として出力する．

5 章では，全話者を 6 グループに分けて，認識話者と教師話者として，交差検証法により従来手法と提案手法の認識率を評価し，提案手法の平均認識率は，すべての場合で従来手法よりも効果的であったことを説明する．

最後に 6 章にて，本論文の成果をまとめ，今後の研究課題について述べる．

第 2 章

雑音ロバスト音声認識

2.1 音声認識とは

音声認識とは，音声から話者が話した単語や文を推定することである．最近，音声認識機能を応用した製品やシステムを目にする機会が増えてきている．スマートフォンの音声検索システムや音声リモコン，または，会話ができるロボットなどである．音声認識によって，音声で情報や命令を入力できるようになる．音声入力には，手軽に使える入力手段であるという利点がある．それから，入力するときに手足を使う必要がないという利点もあり，例えば，運転しながらでも使うことができる．

音声認識の方式では，話者を制限するかどうかによって分類される．事前に音声を学習させたユーザーだけが使用できる音声認識システムがあるが，このように使用する話者を制限する音声認識のことを特定話者音声認識と呼び，事前に音声を学習させなくても使用できる音声認識システムのことを不特定話者音声認識と呼ぶ．

さらに，音声認識方式は，入力する音声の単位によっても分類される．単語やフレーズを連続させずに発声した音声を認識する方式を孤立単語音声認識と呼び，読み上げられた文章や会話などのように連続的に発声された音声を認識する方式を連続音声認識と呼ぶ．本研究では，孤立単語音声認識について説明する．

音声認識処理では，入力された音声信号から，音声特徴量と呼ばれるパターンの時系列データを抽出して，パターン認識を行う．まず，音声生成の仕組みでは，音声信号からどのような音声特徴量を抽出すればよいのかについて判断し，次に，音声分析では，音声特徴量を抽出する方法について説明し，学習では，参照パターンに相当する HMM を生成する方法について紹介し，認識では，認識結果について説明する．

2.2 音響分析

2.2.1 音声生成の仕組み

ヒトの聴覚器官は音声波から周波数帯ごとに音の大きさを感知し，神経インパルスと呼ばれる電気的信号に変換し，それを脳に伝達する．この過程で各聴覚器官がどのように働いているかについて，多くのことが解明されている．

一方，脳は伝達された神経インパルスを手がかりにこれまでの学習で獲得した記憶や知識に基づいて発声された単語や文を推定する．このとき，脳のどの部位が活動しているかについては解明されつつあるようであるが，どのように働いているかについては解明されていない．

脳に神経インパルスとして伝達される周波数帯ごとの音の大きさに関する情報には，発話の大きさや音程などの音声認識にとっては重要ではない情報が含まれている．また，中国語のように声調がある言語では，音声認識にとって音程は重要な情報である．その場合，音程の情報としてピッチ周波数を使用することが多い．また，日本語でも雨と飴のようにアクセントが異なる同音異義語を区別する場合も音程は有用な情報となる．したがって，脳で音声認識にとって重要な情報を抽出していると想像されるが，その仕組みは解明されておらず，音声生成の仕組みから音声認識にとって重要な情報の正体を探ることにする．

2.2.1.1 母音生成の仕組み

図 2.1 は，母音生成の仕組みを示した図である．日本語には「あ」・「い」・「う」・「え」・「お」の 5 種類の母音があるが，日本語に限らずあらゆる母音の音源は，図 2.1 に示すように，喉頭音源と呼ばれる音である．喉頭音源は，喉頭にある声帯を振動させて作られる．

次に，喉頭音源の音声波は，声道（喉頭から口唇）を通過するとき，音声波の各周波数帯の音量は，強められたり，弱められたりする．

最後に，音声波は，口唇から空中へ放射される．このとき，周波数が高くなるほどその周波数帯の音量が強められる．

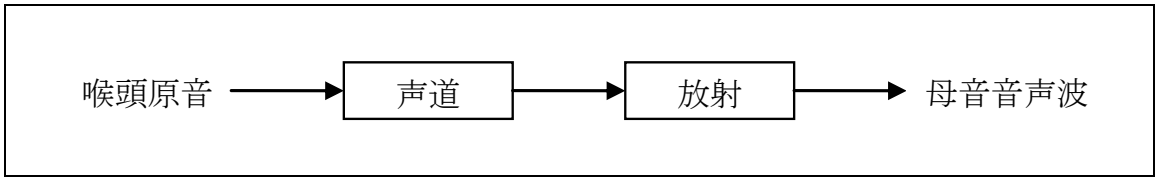


図 2.1 母音生成の仕組み

2.2.1.2 喉頭原音のスペクトル

喉頭原音は，ブザーのように聞こえる音で，基音とその倍音を成分に持つ周期的な音である．その音声波は，図 2.2 に示す Rosenberg（ローゼンバーグ）波に似ていて，図 2.3 に示すような振幅スペクトル（周波数ごとの振幅）を持っている．喉頭原音は，典型的には，周波数が 1 オクターブ高くなるごとにその周波数帯の音量（振幅の 2 乗）が約 12 dB（約 16 分の 1）減衰するような音である．

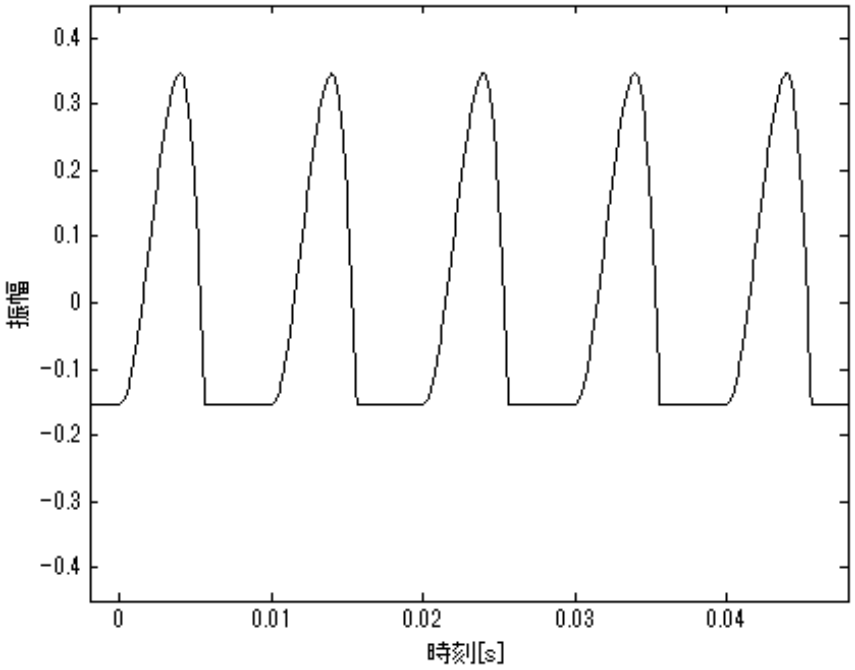


図 2.2 Rosenberg 波

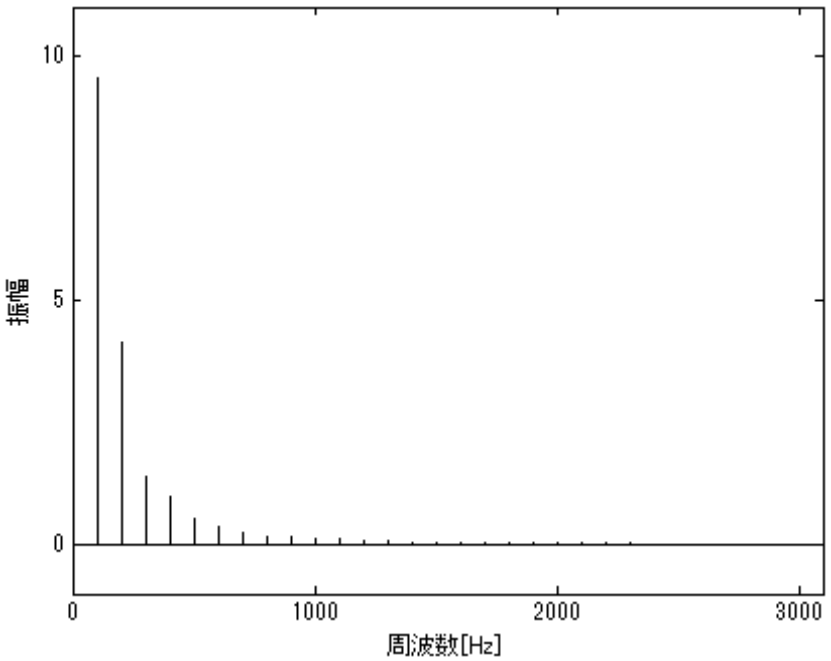


図 2.3 喉頭原音の振幅スペクトル

2.2.1.3 声道特性

喉頭原音は，声道を通過するときに各周波数帯の音量が強められたり，弱められたりする．つまり，声道は音楽プレーヤーのイコライザのような働きをしている．

各周波数帯の増幅率は，声道の形状，すなわち，口の開き具合や舌を盛り上げる位置などによって決まる．そして，声道の形状は発声する母音に依存する．例えば，「あ」を発声するときは口の開きを広くし，「い」を発声するときは口の開きを狭くして舌の先のほうを盛り上げ，そして，「う」を発声するときは口の開きを狭くして，今度は，舌の根元のほうを盛り上げる．つまり，声道の形状を整えることは，イコライザのレバーの位置を調整することに相当する．

図 2.4 に，ある話者が「あ」を発声したときの声道特性のスペクトル，つまり，声道における各周波数帯の増幅率を示す．図 2.5 に，この声道を通過した音声の振幅スペクトルを示す．イコライザのレバーの位置を図 2.4 の曲線を形成するように調整して，喉頭音源にイコライザをかけると，この声道を通過した音声を再現できる．

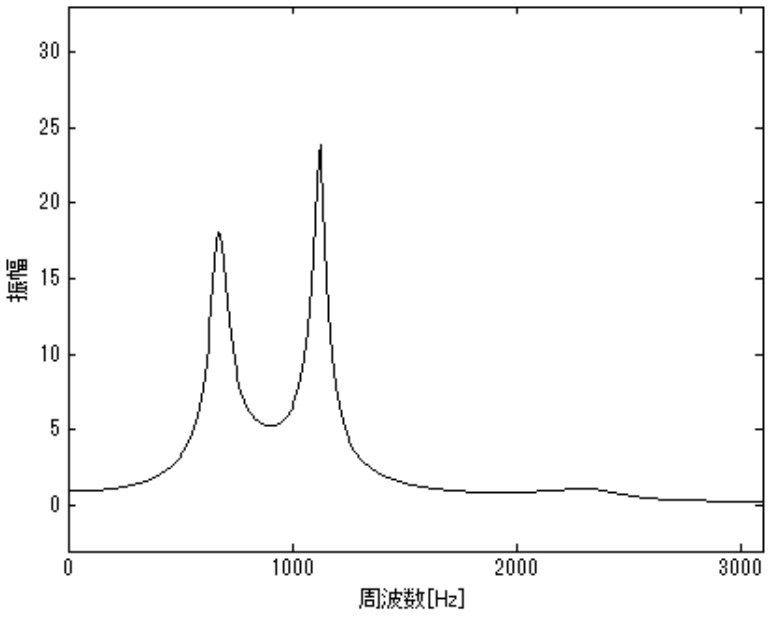


図 2.4 「あ」の声道特性の振幅スペクトル

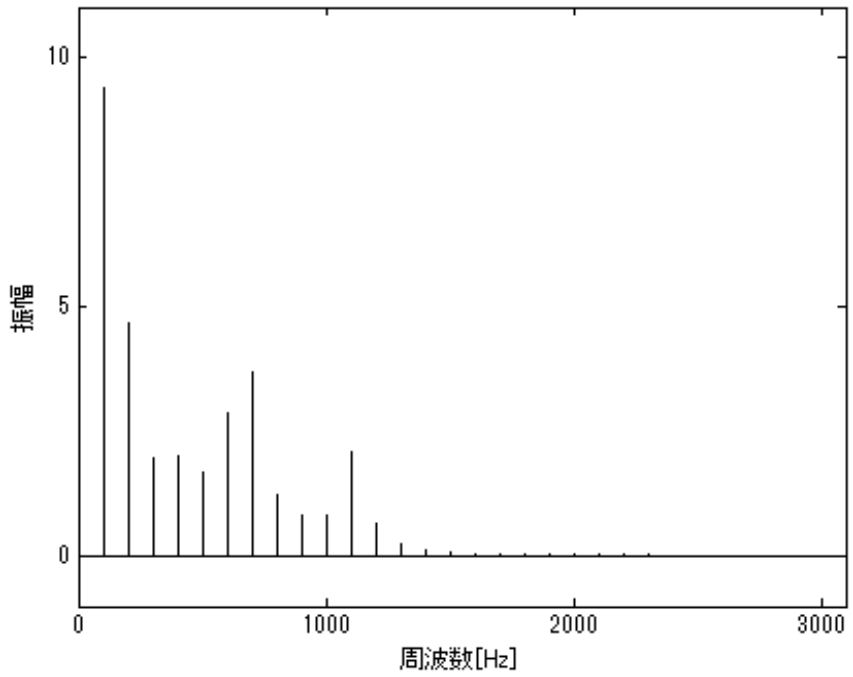


図 2.5 声道を通過した音声の振幅スペクトル

2.2.1.4 放射特性

声道を通過した音声波は，空中に放射されるときに，周波数が高くなるにつれて，その周波数帯の音量が強められる．図 2.6 に示すように，放射による各周波数帯の音量の増幅率は，周波数が 1 オクターブ高くなるごとに約 6 dB ずつ増加する．つまり，ここにもある種のイコライザが働いている．ただし，声道のときとは違って，このイコライザのレバーの位置は調整できない．

図 2.7 に放射された音声の振幅スペクトルを示す．声道を通過した音声に，図 2.6 の曲線を形成するようにレバーの位置を調整されたイコライザをかけると，放射された音声を再現できる．

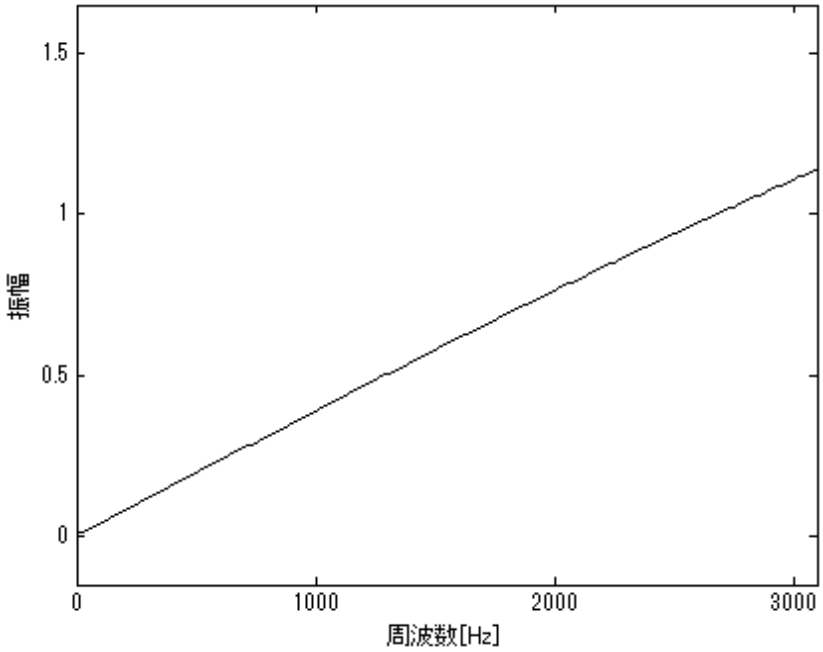


図 2.6 放射特性の振幅スペクトル

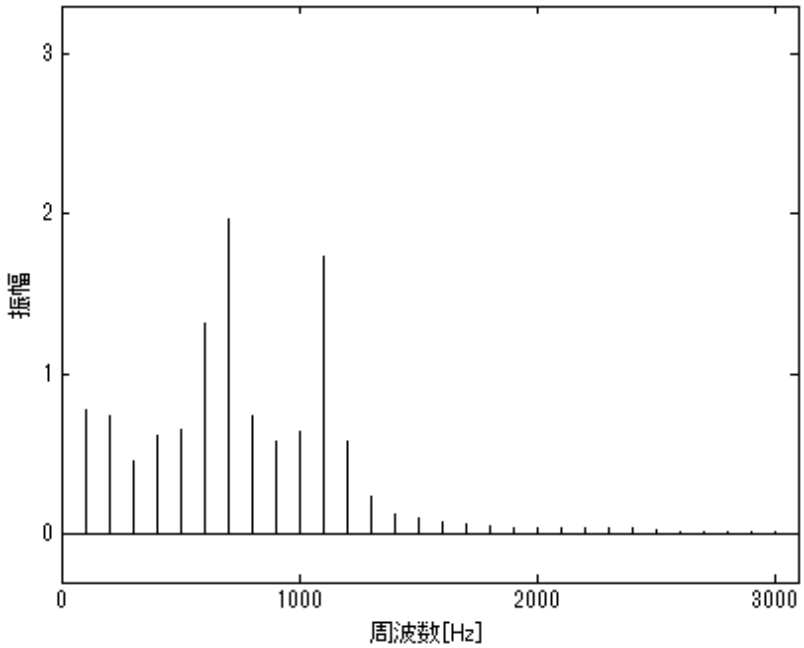


図 2.7 放射された音声の振幅スペクトル

2.2.1.5 音声認識にとって重要な情報

声道特性で説明したように，ある母音を発声するには，声道を母音ごとに定められた形状に整える必要がある．したがって，音声波から，声道においてどの周波数帯が強められたのか，あるいは，弱められたのかが分かれば，声道の形状，すなわち，どの母音が発声されたのかが分かるということである．しかも，同じ母音を発声するときは，誰もが声道をよく似た形状に整えることから，声道特性はある程度似通うことになる．つまり，音声認識にとって重要な情報は声道特性のスペクトルであるということになる．そして，このことは子音についても言える．

2.2.2 音声分析

音声分析とは，音声信号から音声特徴量と呼ばれる音声認識に有用なパターン情報を抽出することである．音声特徴量として，声道特性のスペクトルを使うことができそうであるが，音声認識では MFCC（メル周波数ケプストラム係数，Mel-frequency cepstral coefficients）を使うことが多い．MFCC は声道特性のスペクトルと同等の情報を持っていて，かつ，声道特性のスペクトルよりもデータの個数が少ない．図 2.8 に，音声信号から MFCC を抽出する一連の処理を示す．

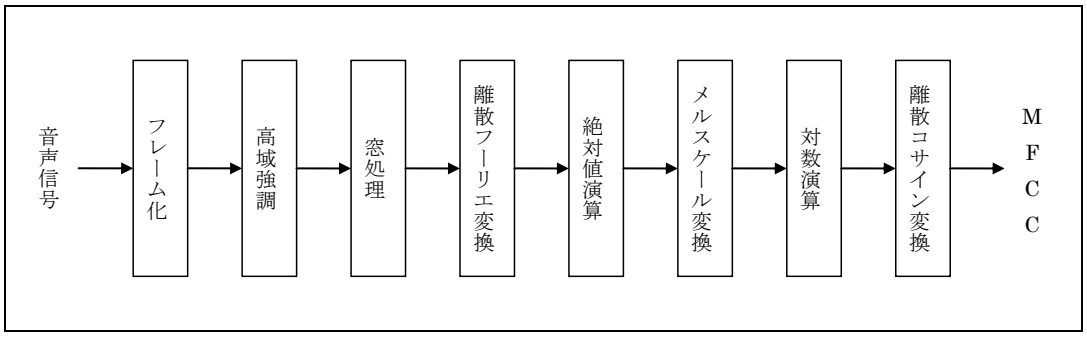


図 2.8 音声分析処理

2.2.2.1 フレーム化

図 2.9 に示すように，音声信号から長さ 20～30 ミリ秒の区間を切り出す．区間の切り出しは，その位置を約 10 ミリ秒ずつずらしながら行う．切り出した区間のことをフレームと呼び，図 2.10 に，図 2.9 の実線の枠（時刻 0.05～0.075 秒）で切り出したデータを示す．そして，図 2.8 に示す一連の処理を行って，MFCC をフレームごとに求める．

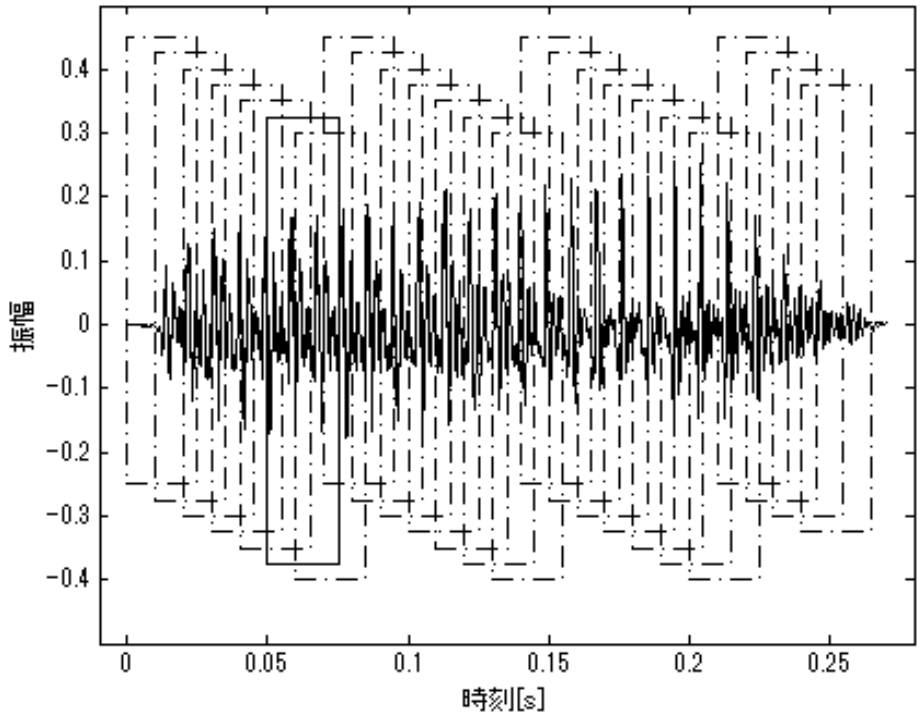


図 2.9 フレーム化

2.2.2.2 高域強調

入力音声のスペクトル包絡は，声道特性のスペクトル，喉頭原音のスペクトル包絡，放射特性のスペクトルを合成したものである．喉頭原音のスペクトル包絡は，周波数が 1 オクターブ高くなるごとに約 12 dB 減衰するような傾斜を持っている．放射特性のスペクトルは，周波数が 1 オクターブ高くなるごとに約 6 dB 増加するような傾斜を持っている．そこで，入力音声に対して，各周波数帯の音量

の増幅率を 1 オクターブ上がるごとに約 6 dB ずつ増加させるような処理を行うと，理論上，喉頭原音と放射特性に由来する傾斜を打ち消すことができる．その結果，高域強調した入力音声のスペクトル包絡は，声道特性のスペクトルと一致することになる．高域強調と呼ばれるのは，この処理によって周波数が高い成分ほど強められるためである．図 2.11 に高域強調されたデータを示す．

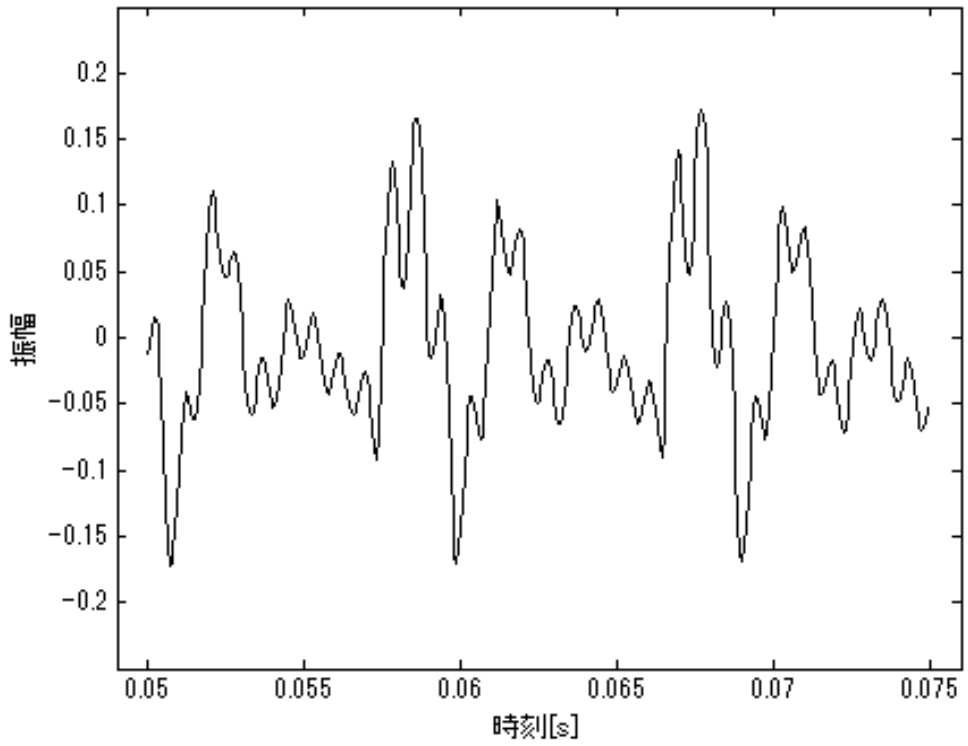


図 2.10 フレームデータ

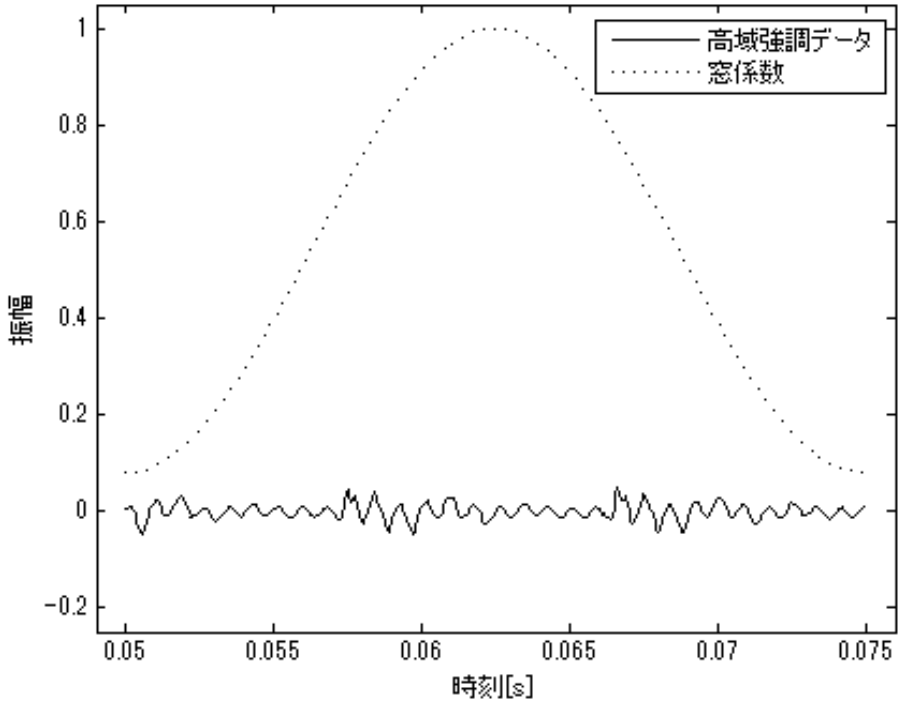


図 2.11 高域強調データと窓係数

2.2.2.3 窓処理

フレームの両端のデータには連続性がないため，このまま次のステップの離散フーリエ変換を行うと，音声信号のある周波数帯に小さなピークが存在するときに，そのピークが埋もれて見えにくくなる．そこで，フレームのデータ各々に窓係数（図 2.11）と呼ばれる重みを乗算して，データの不連続性を抑制する．窓係数として，ハミング窓やハン窓の係数を使うことが多い．

2.2.2.4 離散フーリエ変換

離散フーリエ変換を行って，スペクトルを求める（図 2.12，2.13）．スペクトルは複素数であり，周波数帯ごとの振幅と位相の情報を持っている．尚，計算は高速フーリエ変換で行う．

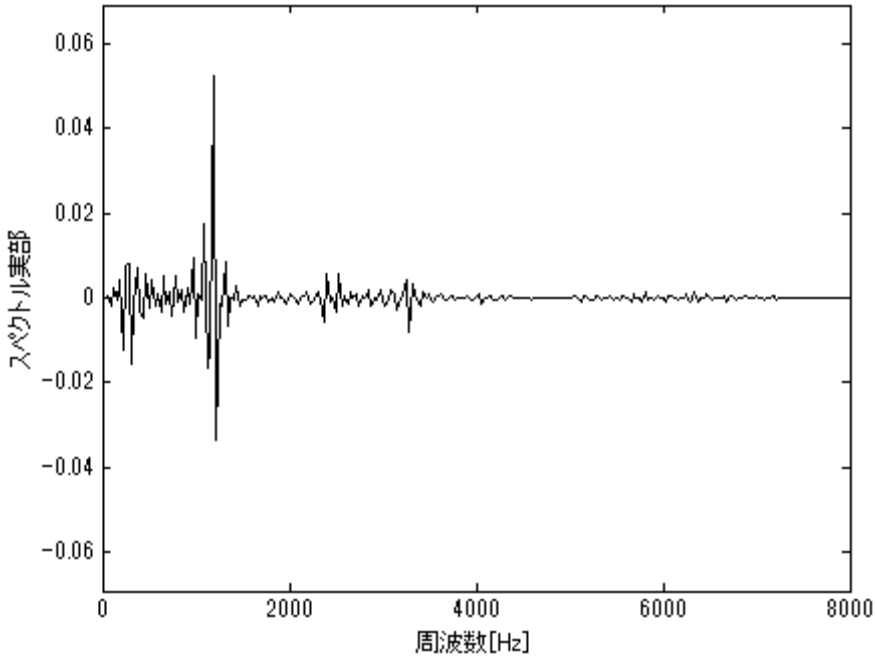


図 2.12 スペクトル実部

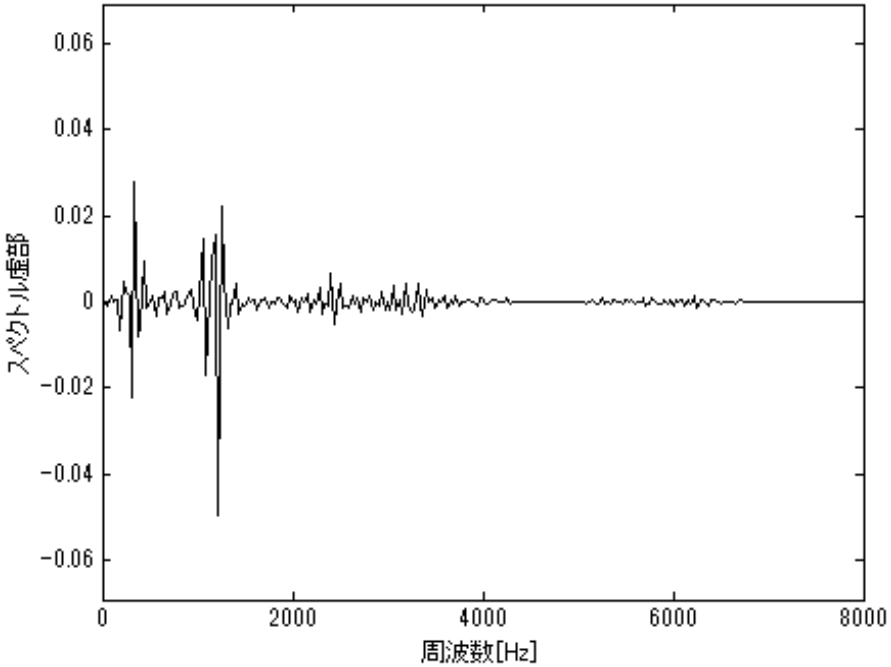


図 2.13 スペクトル虚部

2.2.2.5 絶対値演算

音声認識では周波数帯ごとの音量が重要であって，位相は重要ではないと考えられている．そこで，スペクトルの絶対値（実部と虚部の二乗和の平方根）である振幅スペクトル（図 2.14）を求める．

2.2.2.6 メルスケール変換

ヒトの聴覚器官は周波数帯ごとに音の大きさを感知していると説明したが，それらの周波数帯の幅は,周波数が高くなるにつれて広くなることが知られている．

このようなヒトの聴覚の特性を考慮して，周波数が高くなるほど目盛りが粗くなるようなスケールを使い，そのようなスケールとして，図 2.15 に示す mel（メル）単位に基づくスケールを使うことが多い．振幅スペクトルを mel 単位で一定間隔ごとにリサンプリングすると，メルスペクトルになる．このとき，近傍にある振幅スペクトルの加重平均操作を行うフィルタバンクを用いて計算する．図 2.14 に示すように，周波数が高くなるにつれて，近傍に存在する振幅スペクトルの個数が増加し，振幅スペクトルをより大まかに評価していることがわかる．

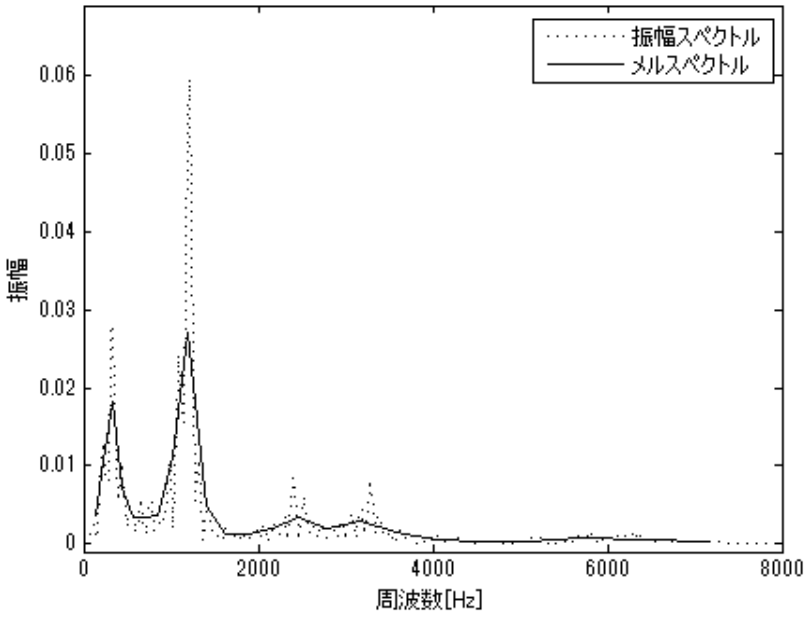


図 2.14 メルスペクトルと振幅スペクトル

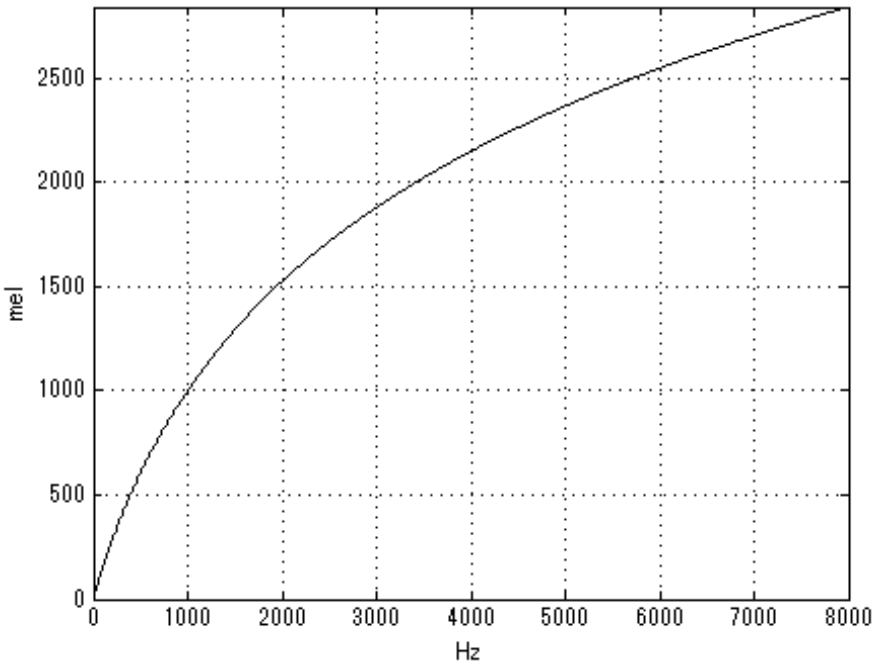


図 2.15 メルスケール

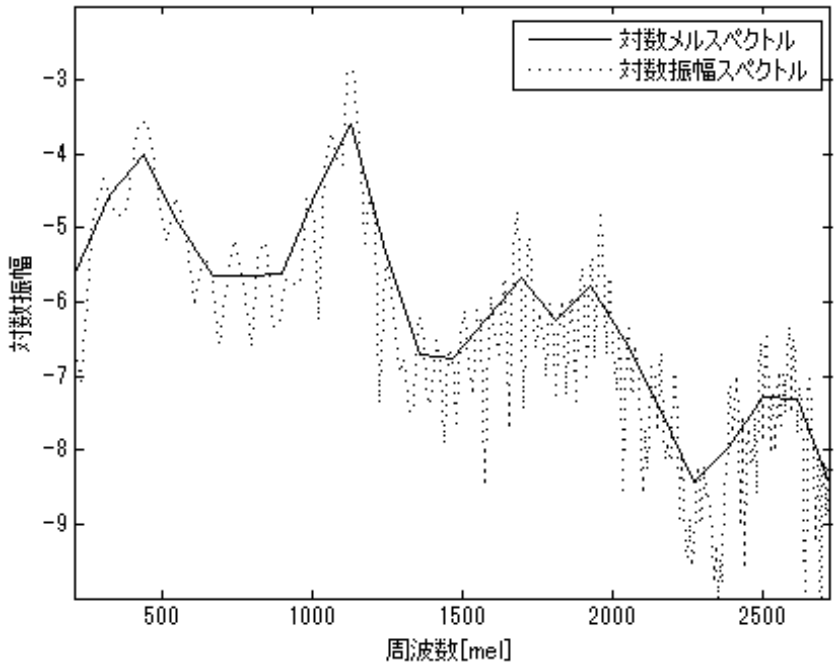


図 2.16 対数メルスペクトルと対数振幅スペクトル

2.2.2.7 対数演算

ヒトの聴覚に関する実験により，ヒトが感じる音の大きさは物理的な刺激量である音量の対数に比例すること，すなわち，（ヒトが感じる音の大きさ）＝（比例定数）×log（音量）が成り立つことが確かめられている．聴覚以外の感覚についても，感覚量が物理的な刺激量の対数に比例することが確かめられていて，これをフェヒナーの法則（Fechner's law）という．

このようなヒトの聴覚の特性を考慮して，メルスペクトルに対して対数演算を適用し，対数メルスペクトル（図 2.16）を求める．ヒトの聴覚器官で対数演算に相当する処理が行われているのは興味深いことだと思われる．

2.2.2.8 離散コサイン変換

対数メルスペクトルを離散コサイン変換して，MFCC（図 2.17）を求める．声道特性のスペクトルに関係があるのは低次の MFCC であるので，枠内の低次の MFCC だけを使う．音声認識では，1〜12 次 MFCC を使うことが多い．

パーセバルの定理から，2 つの低次の MFCC の差の二乗和は，それぞれの対数メルスペクトル包絡上のデータ列の差の二乗和に等しくなる．したがって，低次の MFCC の類似度を評価することは，メルスペクトル包絡の類似度を評価することに相当する．メルスペクトル包絡は声道特性のスペクトルの概形であることから，低次の MFCC が声道特性のスペクトルと同等の情報を持つことがわかる．

以上の処理をすべてのフレームに対して行うと，図 2.18 に示すように，MFCC の時系列データを得る（図 2.18 の枠内のデータは，図 2.17 の枠内のデータである）．図 2.18 の凡例のC_nはn次 MFCC のことであり，C_5〜C_{10}を割愛している．

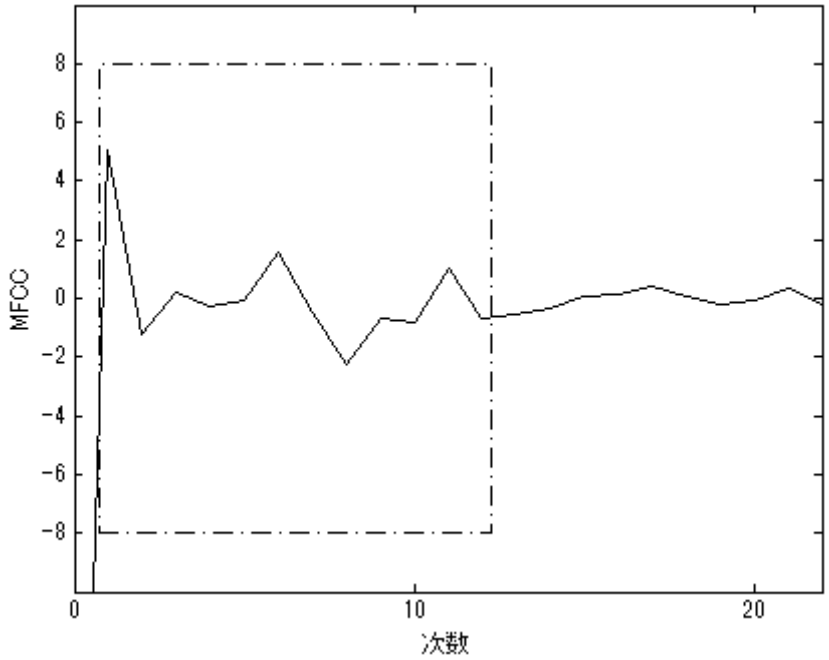


図 2.17 MFCC

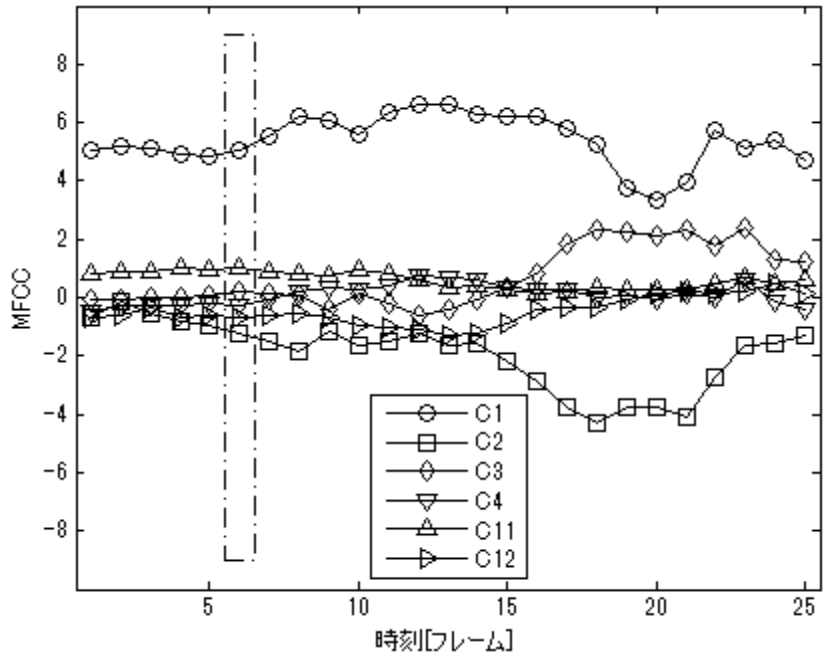


図 2.18 低次 MFCC 時系列データ

2.2.2.9 対数エネルギー

MFCC のほかに，音声認識でよく使われる音声特徴量として，対数エネルギーがある．エネルギーとは音量のことであり，切り出したフレーム内の全データ（データの値は振幅を表す）の二乗和を計算して求める．そして，その対数が対数エネルギーになる．

声帯が振動しているときとしていないときとで対数エネルギーの値が大きく異なるため，声帯を振動させて発声する有声音（母音,「が」や「ざ」の子音など）と声帯を振動させずに発声する無声音（「か」や「さ」の子音など）を区別するのに有用である．

ただし，対数エネルギーそのものを使わずに，次に説明するデルタ特徴量であるデルタ対数エネルギーやデルタデルタ特徴量であるデルタデルタ対数エネルギーを使うことが多い．

2.2.2.10 デルタ特徴量

「わ」と「うあ」という 2 つの音声を発声するときは，どちらも，最初は口をすぼめて舌の後ろを盛り上げる「う」の舌の構えをしてから,口を大きく開く「あ」の構えに移る．このような共通点があるために，それらの音声特徴量列は最初と最後が似通う．しかし，発声開始から「あ」に至るまでの時間については,「わ」のほうが「うあ」よりも短いため，音声特徴量が変化する量が異なる．したがって,音声特徴量が 1 フレームあたりに変化する量も音声認識にとって有用である．音声特徴量が1フレームあたりに変化する量のことをデルタ特徴量と呼ぶ.MFCC のデルタ特徴量のことをデルタ MFCC と呼ぶ．さらにデルタ特徴量の 1 フレームあたりの変化量であるデルタデルタ特徴量もよく使われる．図 2.19 に，ある話者が「下（した)」と発声した音声から抽出した 1 次 MFCC，および，そのデルタ特徴量とデルタデルタ特徴量を示す．

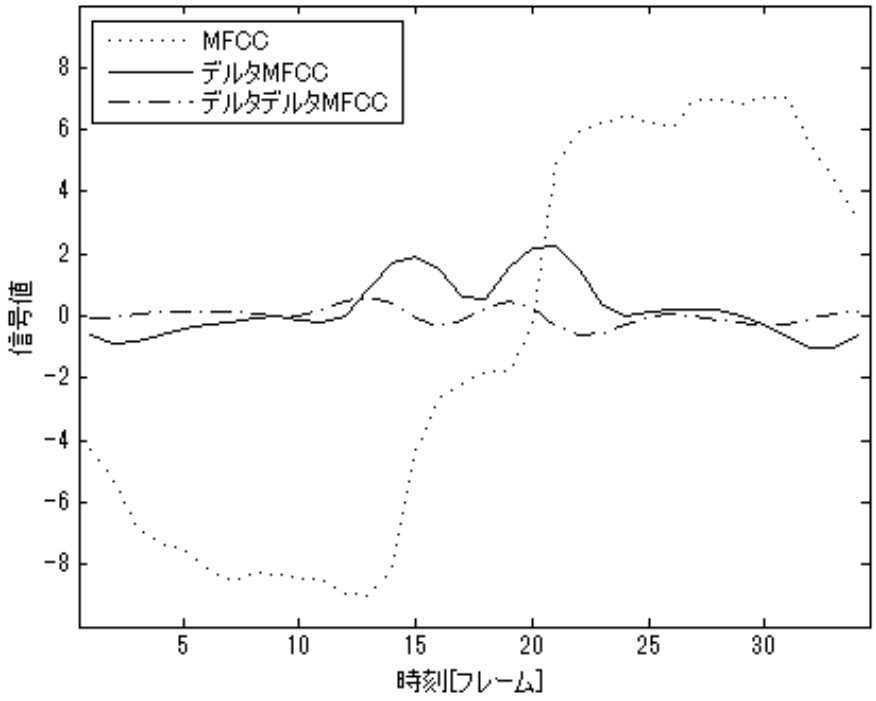


図 2.19 デルタ特徴量

2.2.3 学習

学習の手順であるが，まず，発声した単語が同一の音声信号を数十名以上の話者から収録し，それらの音声信号から音声特徴量の時系列データ（以降，音声特徴量列と呼ぶ）を抽出する．学習に使用する音声のことを教師音声，その話者を教師音声話者と呼び，教師音声から抽出した音声特徴量列を標準的な時系列パターンとして学習し，発声した単語の音響モデルを作る．

使用する音声特徴量であるが，1～12 次の MFCC，1～12 次のデルタ MFCC の 24 種類, あるいは, それにデルタ対数エネルギーを加えた 25 種類とすることや，1～12 次の MFCC，1～12 次のデルタ MFCC，1～12 次のデルタデルタ MFCC の 36 種類, それにデルタ対数エネルギーとデルタデルタ対数エネルギーを加えた 38 種類とすることが多い．

発声する単語が「上（うえ)」か「下（した)」のどちらかであったとして，入

力音声から抽出した音声特徴量列からどちらが発声されたかを推定することを考える.男性話者 6 名が発声した音声から抽出した音声特徴量列を図 2.20～図 2.23 に示す．教師音声話者は話者 B～F の 5 名とする．話者 A の音声を入力音声として，認識の節で使う．使用する音声特徴量は 1 次 MFCC C_1と 5 次 MFCC C_5の 2 種類とする.

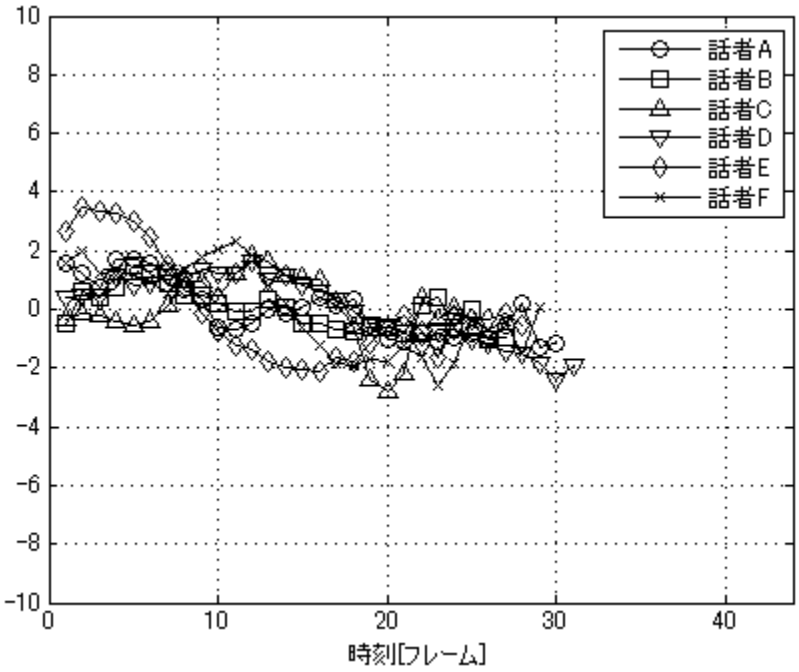


図 2.20 「上」の 1 次 MFCC 時系列データ

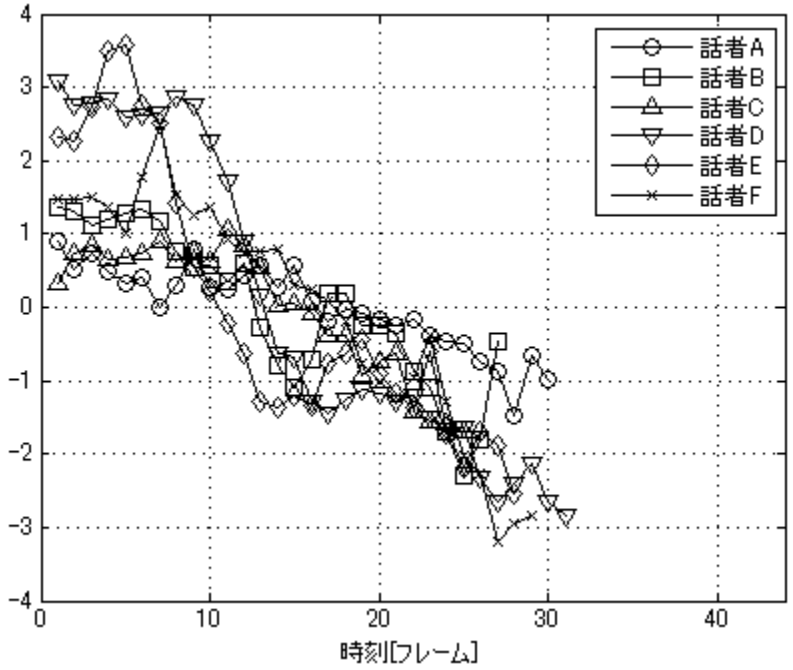


図 2.21 「上」の 5 次 MFCC 時系列データ

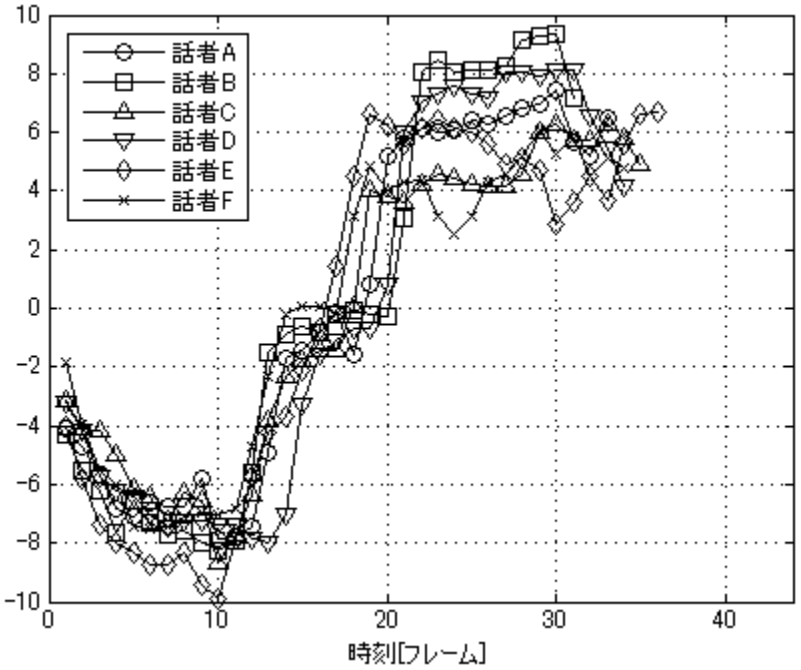


図 2.22 「下」の 1 次 MFCC 時系列データ

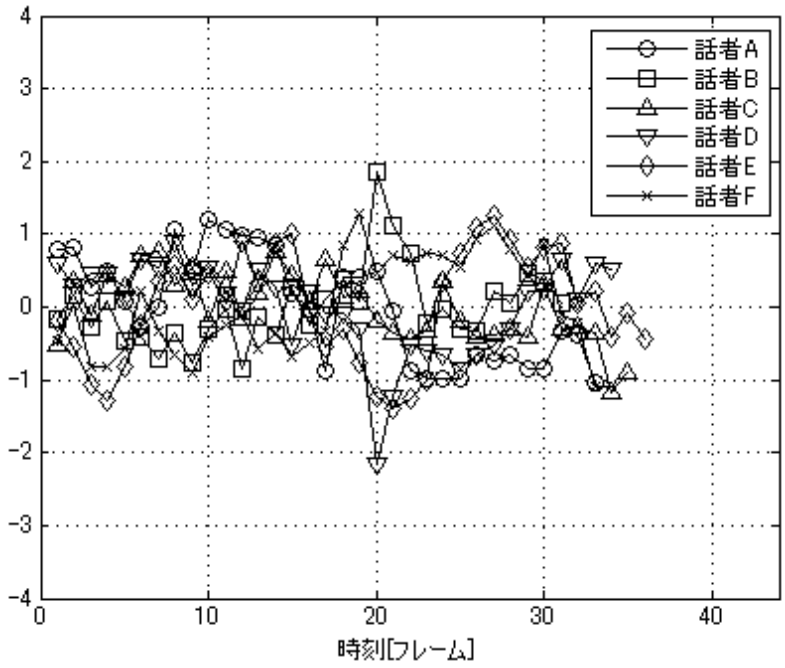


図 2.23 「下」の 5 次 MFCC 時系列データ

2.2.3.1 音響モデル

本稿で扱う音響モデルは，任意の音声特徴量列に対してそれが生成される確率を計算できる枠組みを与えるもので，音声特徴量列を入力するとその音声特徴量列が生成される確率を出力する関数のようなものである．

音声分析で見たように，同じ単語を発声すると類似した音声特徴量列が得ることができる音声特徴量を使っている．したがって，教師音声から抽出した音声特徴量列を学習することによって教師音声から抽出した音声特徴量列に類似した音声特徴量列ほど確率が高くなるような音響モデルを作成すると，教師音声話者が収録時以外に発声した音声から抽出した音声特徴量列が音響モデルから生成される確率は高くなる．さらに，多くの教師音声話者の教師音声を学習することにより，教師音声話者以外の話者，すなわち，不特定話者が同じ単語を発声した音声から抽出した音声特徴量列も生成される確率が高くなるような音響モデルとなる．肝心の音響モデルであるが，音声認識では，HMM を使用することが多い．

2.2.3.2 ベイズの定理の適用

入力音声から抽出した音声特徴量列から，発声された単語が「上」である確率P_1と「下」である確率P_2を計算できたと仮定する．そうすると，これら 2 つの確率 P_1と P_2を比較して，確率が高くなるほうの単語が発声されたと推定するのが合理的である．しかし，これらの確率を直接計算することは非常に困難で，現実的には，不可能である．そこで，ベイズの定理から間接的にこれら 2 つの確率を計算して比較する．

発声される単語が「上」(「下」) のときに抽出した音声特徴量列を得る確率をQ_1（Q_2)，発声される単語が「上」(「下」) である確率をR_1（R_2)，抽出した音声特徴量列を得る確率をSとすると，ベイズの定理から，$P_1 = Q_1R_1/S$ と $P_2 = Q_2R_2/S$ が成り立つ．

まず，各右辺の分母に現れるSであるが，これは，P_1とP_2の大小関係に影響を与えないので，S を計算する必要はない．次に，R_1とR_2であるが，孤立単語音声認識では，通常，どの単語も発声される確率は等しいとするので，R_1とR_2は同じ値である．したがって，P_1とP_2の大小関係は，Q_1とQ_2の大小関係から決まることになる．

また，連続音声認識では，一般的に，単語ごとに発声される確率が異なる．例えば，「きのう本を」の後に続く単語について考えると，「買った」が続く確率のほうが「割った」が続く確率よりも高いと考えるのが合理的である．そこで，言語モデルを導入し，単語ごとに発声される確率を計算できる枠組みを与える．そうすると，入力音声が「きのう本を買った」であったときに，「買った」と発声した部分について，雑音などの影響により，音響モデルで「割った」のほうが「買った」よりも確率が高いと誤って推定しても，言語モデルを加味したときに「買った」である確率のほうが「割った」である確率よりも高いと推定できるようになる．音声認識では，直前の$N-1$個の単語から次に発声される単語の確率を与える N-gram（N-グラム）と呼ばれる言語モデルを使用することが多い．

さて，Q_1（Q_2）であるが，これを厳密に求めるには，「上」(「下」) と発声した音声をすべて集めて，その音声から抽出した音声特徴量列について同じ数列とな

るものごとに数え上げる必要があり，現実的には不可能である．しかしながら，多くの教師音声話者が発声した教師音声から抽出した音声特徴量列を学習することにより，それらに類似した音声特徴量列ほど生成される確率が高くなるような音響モデルを作ることは可能で，そのような音響モデルからQ_1やQ_2の近似的な値を計算することができる．

2.2.3.3 初期 HMM 生成

HMM の作り方について説明する．まず，「上」の教師音声から抽出した音声特徴量列を，対応するフレームの音声波形を観察しながら，以下の 3 つの区間に分ける．

・区間 U1：発声開始から U2 の直前までの区間

・区間 U2：「う」から「え」に移り変わる区間

・区間 U3：U2 直後から発声終了までの区間

「下」の音声特徴量も，以下の 3 つの区間に分ける．

・区間 S1：「し」の子音がある区間

・区間 S2：S1 と S3 の間にある区間

・区間 S3：「た」の母音がある区間

この作業をすべての教師音声について行う．このような区間の分け方をしたが，実際には，音素ごとに区間を分け，さらに各区間をいくつかの区間に分ける．区間ごとに集計した結果をヒストグラムとして図 2.24 と図 2.25 に示す．

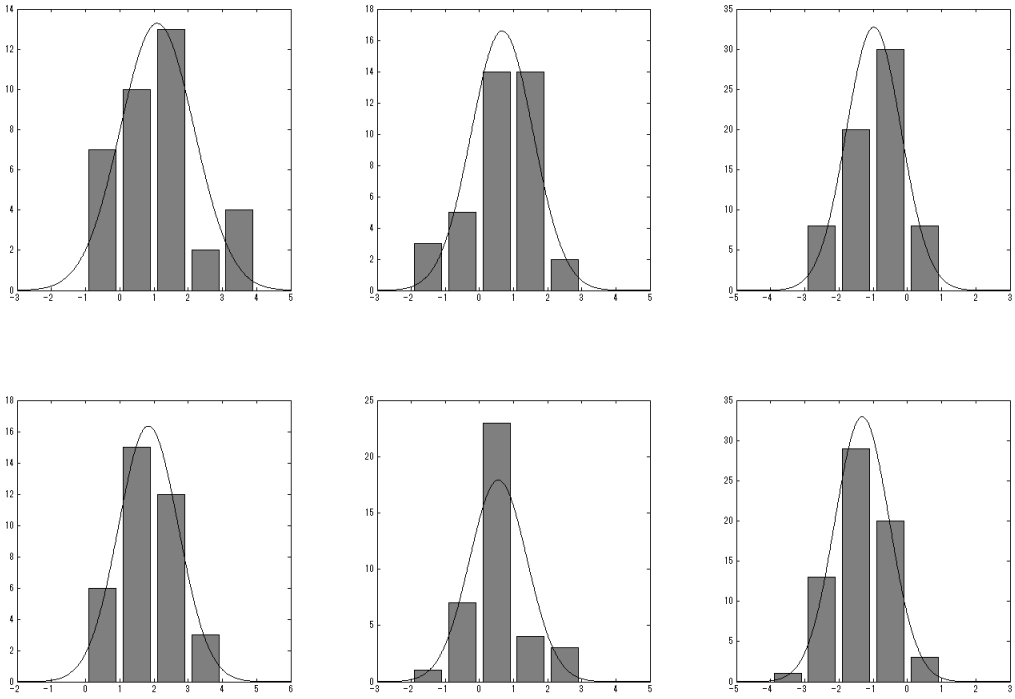


図 2.24 「上」の音声特徴量列のヒストグラム

図 2.24 の上段の左から順に区間 U1 の 1 次 MFCC，区間 U2 の 1 次 MFCC，区間 U3 の 1 次 MFCC，下段の左から順に区間 U1 の 5 次 MFCC，区間 U2 の 5 次 MFCC，区間 U3 の 5 次 MFCC である．

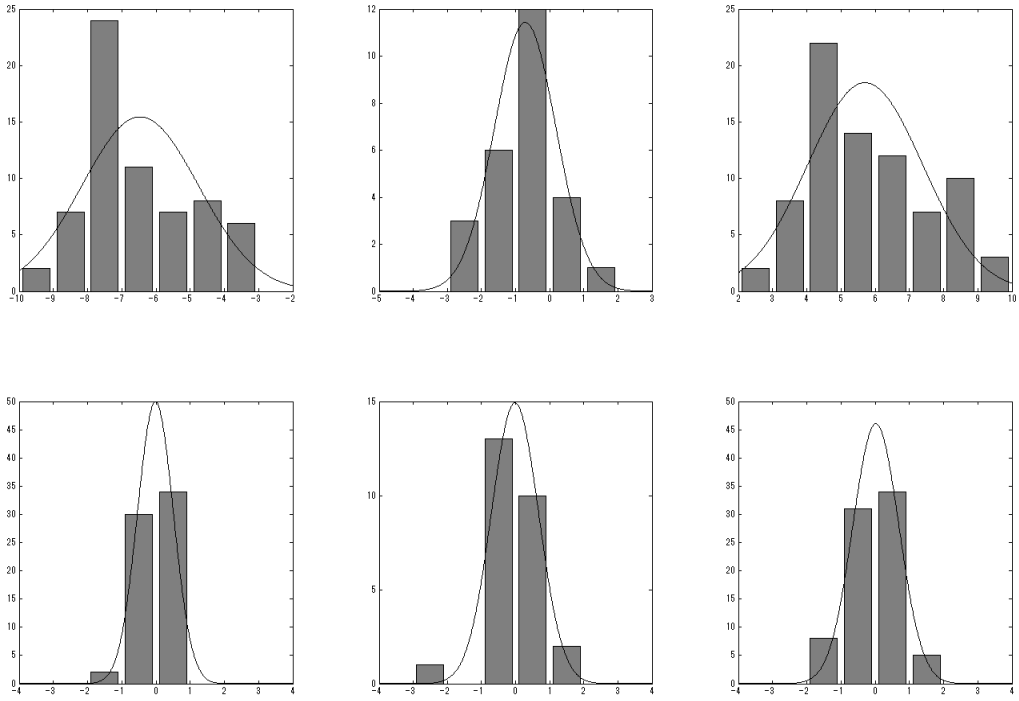


図 2.25 「下」の音声特徴量列のヒストグラム

図 2.25 の上段の左から順に区間 S1 の 1 次 MFCC，区間 S2 の 1 次 MFCC，区間 S3 の 1 次 MFCC，下段の左から順に区間 S1 の 5 次 MFCC，区間 S2 の 5 次 MFCC，区間 S3 の 5 次 MFCC である．

次に，区間ごとに統計処理を行い，その区間の長さの平均，音声特徴量の平均と分散（平均からのばらつきの度合いを表す量）を計算する．その結果を表 2.1 に示す．図 2.24 と図 2.25 に平均と分散から決まる正規分布曲線を重ねて描く．

表 2.1 区間別統計データ

区間	平均の長さ	C_1の平均	C_1の分散	C_5の平均	C_5の分散
U1	7.2	1.0954	1.1682	1.8381	0.7707
U2	7.6	0.6673	0.8339	0.5541	0.7157
U3	13.2	–0.9817	0.6464	–1.3216	0.6382
S1	13.2	–6.4624	2.9039	–0.0163	0.2775
S2	5.2	–0.7175	0.8232	–0.0176	0.4831
S3	15.6	5.7073	2.8354	0.0197	0.4557

表 2.1 のデータから，図 2.26 と図 2.27 に示すような初期 HMM を作成できる．

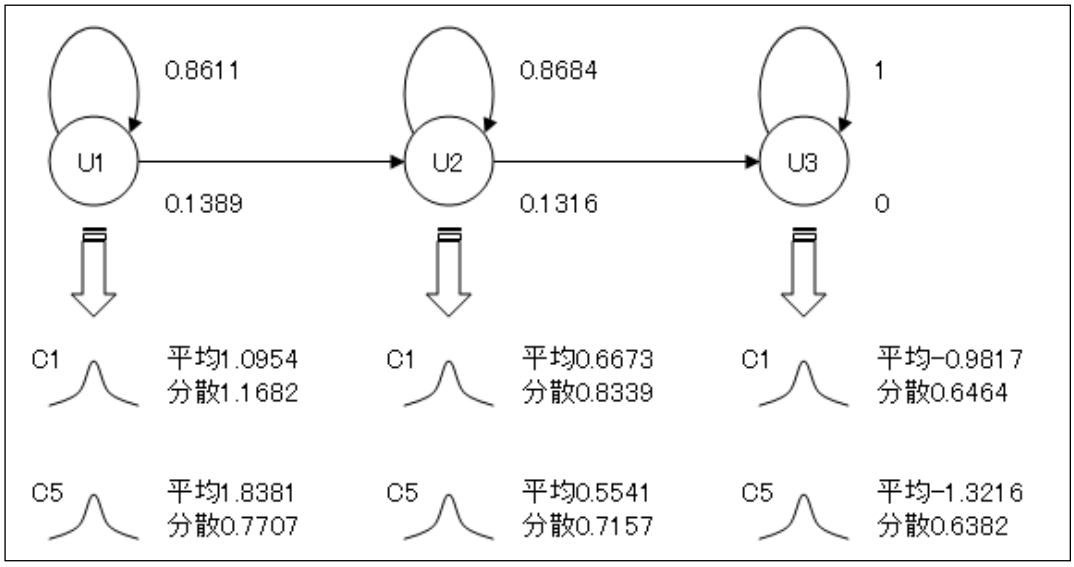


図 2.26 「上」の初期 HMM

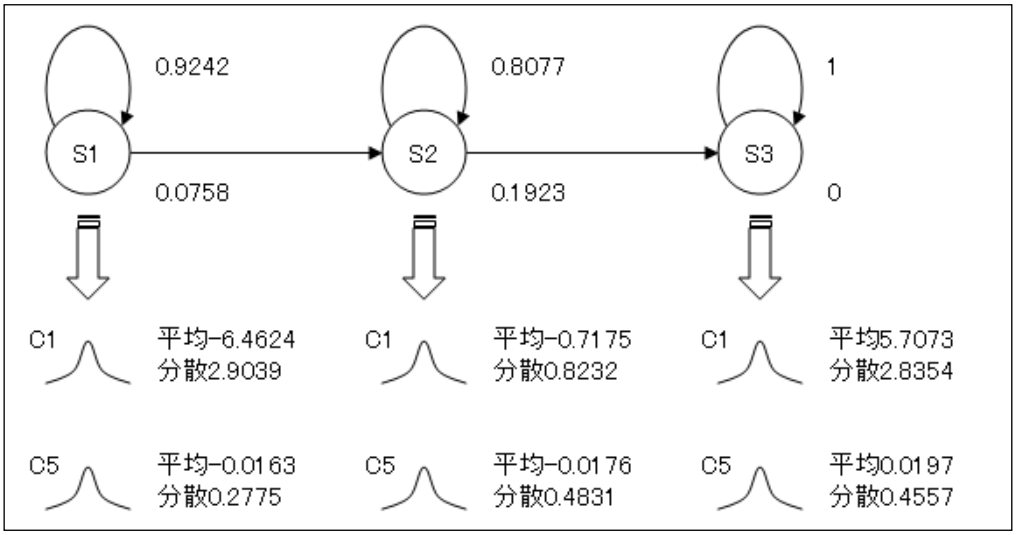


図 2.27 「下」の初期 HMM

図 2.26 と図 2.27 の円形のオブジェクトは，それぞれ表 2.1 の区間に対応している．この円形のオブジェクトをノードと呼ぶ．ノード同士を結ぶ矢印は，区間から区間への遷移，つまり，区間の切り替えを意味する．前のノードに戻る矢印がないのは，区間の間に時間的順序の制約があるためである．例えば,「上」を発声するときは，先に「う」を発声してからその後に「え」を発声するが，その順序を逆転して発声することはない．矢印の傍に記載されている数値は遷移確率と呼ばれ，そのうち，自身を結ぶ矢印の傍に記載された数値（自己遷移確率）が大きいほど，その区間の長さの平均が大きくなる．その値は，表 2.1 の区間の平均の長さの逆数を 1 から引くことで計算できる．例えば，ノード U1 の自己遷移確率は，$1 - 1/7.2 = 0.8611$で計算できる．ノード U1 からノード U2 への遷移確率は，$1 - 0.8611 = 0.1389$として計算する．ノード U3 については，他のノードへ遷移できないので，自己遷移確率を強制的に 1 にしている．

ノードから下向きに出ているブロック矢印は，各区間のC_1とC_5の値が記載されている平均と分散をもつ正規分布に従うことを示している．これらの平均や分散は，表 2.1 のデータである．

2.2.3.4 Forward アルゴリズム

Forward アルゴリズムによって, HMM から音声特徴量列が生成される確率密度を計算することができる.「上」の初期 HMM から話者 B～F が発声した「上」から抽出した音声特徴量列が生成される対数確率密度（確率密度の自然対数）を計算すると，表 2.2 のようになる.

表 2.2 対数確率密度

話者	初期 HMM	最終 HMM
B	–62.27	–58.43
C	–63.03	–57.57
D	–76.50	–78.02
E	–76.01	–72.18
F	–75.62	–74.13
総和	–353.43	–340.33

2.2.3.5 バウム・ウェルチアルゴリズム

バウム・ウェルチアルゴリズム（Baum–Welch algorithm）によって，初期 HMM の音響モデルパラメータ（遷移確率や音声特徴量の平均および分散）を初期値として, 話者 B～F の音声特徴量列が生成される確率密度をすべて掛け合わせた積, すなわち，総積（総乗）が大きくなるように音響モデルパラメータを逐次改善する. バウム・ウェルチアルゴリズムの中では, HMM を使って区間の再分割をし，再分割した区間ごとに音声特徴量の平均と分散，区間の平均の長さを計算して音響モデルパラメータを更新し，更新された HMM を使って区間の再分割をするという繰り返しがなされる．バウム・ウェルチアルゴリズムを適用した最終結果の HMM から, 話者 B～F の音声特徴量列が生成される対数確率密度とそれらの総和（確率密度の総積の対数に等しい）を表 2.2 に示す．バウム・ウェルチアルゴリ

ズムにより，対数確率密度の総和，すなわち，確率密度の総積が大きくなったことから, HMM が改善されたことが分かる．この最終結果の HMM を認識で使う．

2.2.4 認識

音声認識を行うには, まず, 入力音声を音声分析して音声特徴量列を抽出する．次に，その音声特徴量列が各 HMM から生成される確率密度を求める．最後に，確率密度が最も高い HMM に対応する単語を認識結果とする．確率密度を求めるときは，Forward アルゴリズムよりも計算量が少ないビタビアルゴリズムによって，近似的な値を求めることが多い．

ビタビアルゴリズムによって，話者 A の「下」の音声特徴量列が各 HMM から生成される対数確率密度を求めると,「上」の HMM については−814.7692,「下」の HMM については−97.5872となる．したがって,「下」の HMM から生成される確率密度のほうが高いので，認識結果は「下」となる．

2.3 雑音ロバスト音声認識

従来手法となる雑音ロバスト音声認識[15]-[21]は，主に HMM に基づく音声認識であり，そのフローチャートを図 2.28 に示す.

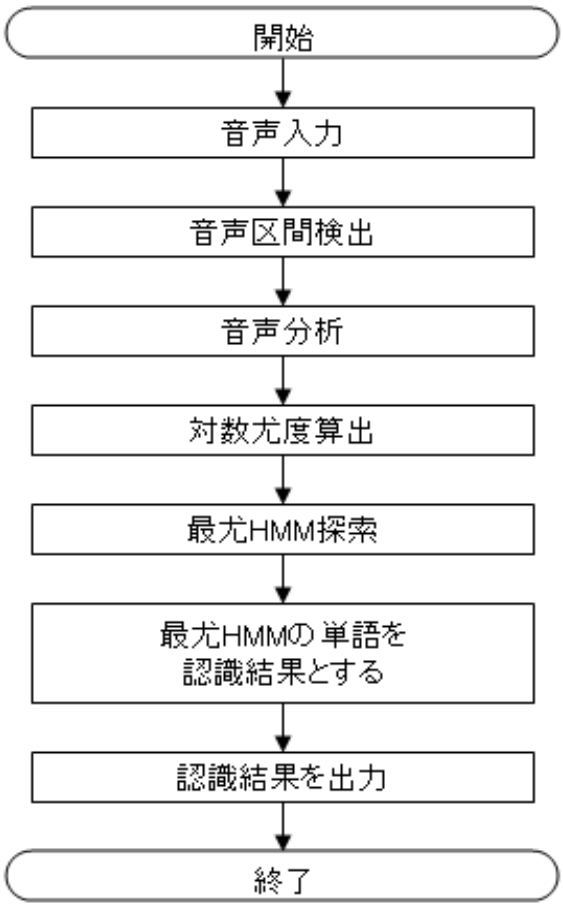


図 2.28 雑音ロバスト音声認識のフローチャート

雑音ロバスト音声認識では，入力音声に対して音声区間検出を行い，検出された音声信号 $s_i\ (i = 0, 1, \ldots, N-1)$ を図 2.29 に示すフローチャートに従って，表 2.3 に示す条件で分析し，音声特徴量ベクトル列 $\boldsymbol{X} = [\boldsymbol{x}_0\ \boldsymbol{x}_1\ \ldots\ \boldsymbol{x}_{T-1}]$ を求め

る．ここで$\boldsymbol{x}_t$は，t番目の音声フレーム $\boldsymbol{s}_t = [s_{128t}\ s_{128t+1}\ \cdots\ s_{128t+255}]$ から計算された 38 次元音声特徴量ベクトルである．また，T は，$\big((N-256)\big)/128+1$以下の最大の整数である．尚，図 2.29 の RSF (running spectral filtering)では，184 次 FIR フィルタによる変調周波数 2〜8 Hz 帯域のバンドパスフィルタリングを行った [16]．尚，RSF ではなく，CMS (cepstral mean subtraction) [18]，あるいは，RASTA [20]を適用することも考えられる．

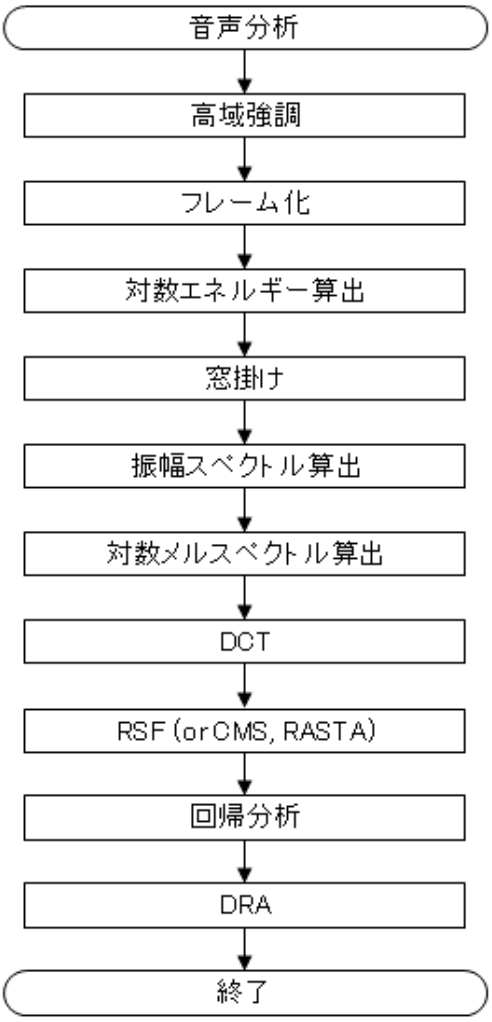


図 2.29 音声分析のフローチャート

表 2.3 雑音ロバスト音声認識の音声分析条件

標本周波数	11025 Hz
高域強調係数	0.97
窓係数	ハン窓
フレーム長	23.2 ms (256 points)
フレームシフト長	11.6 ms (128 points)
音声特徴量	12 次元 MFCC 12 次元 ΔMFCC 12 次元 ΔΔMFCC Δ 対数エネルギー ΔΔ 対数エネルギー
雑音ロバスト処理	Running spectral filter [16] Dynamic range adjustment [17]

次に，ビタビアルゴリズムにより，各単語 w の HMM に対する $\boldsymbol{X}$ の対数尤度 $\ln P(\boldsymbol{X}|\boldsymbol{\theta}_w)$ を求める．ここで，$\boldsymbol{\theta}_w$ は，バウム・ウェルチアルゴリズムによって求めた単語 w の HMM パラメータである．

本論文において，雑音ロバスト音声認識で扱う HMM は，32 状態 1 混合の単語モデルであり，マルチコンディション学習によって生成される．マルチコンディション学習とは，同じ学習音声でも複数の雑音環境下における音声を想定して，総合的に学習するもので，クリーン音声，Speech babble 雑音音声，White noise 雑音音声，および残響効果付加音声の 4 種類を使用した．Speech babble，White noise 等，本論文では，NOISEX-92 データベース[24]の雑音データを使用した．

最後に，対数尤度が最大となる HMM を探索し，認識結果w_1として出力する．こ

こで，w_1 は $w_1 = \underset{w}{\operatorname{argmax}} \ln P(X|\theta_w)$ により求める．

第 3 章

音響的類似単語とその認識困難さ

本章では，まず，本論文で「難認識語」と呼ぶ認識が難しい単語に対する規定について述べ，次に，難認識語の認識精度を改善する鍵となる音響的差異区間について説明する．

3.1 難認識語

図 3.1 は，SNR 10 dB，15 dB，および，20 dB の Factory floor noise 1 雑音環境下における雑音ロバスト音声認識の単語別の認識率に関するヒストグラムである．ここで，ヒストグラムを構成するバーは，右端から順に，認識率が 99%を超える単語の個数を高さとするバー，認識率が 98%を超え，かつ，99%以下となる単語の個数を高さとするバー，認識率が 97%を超え，かつ，98%以下となる単語の個数を高さとするバー，等々である．認識率の評価は，5 章で説明するように，交差検証法に基づいて行った．

本論文で使用した音声データベースは，18 歳以上の女性 48 名が防音室で 95 単語を 3 回ずつ発話した音声をデータベース化したものである．発話した全 95 単語の中から抜粋した単語の一覧を表 3.1 に示す．公開されている既存の音声データベースを使用しなかったのは，音響的に類似した単語を多く含む，話者 50 名程

度以上の孤立発話音声データベースを必要としたためである．

図 3.1 より，SNR が低下するにつれて，他の単語よりも認識率が著しく低下する単語が存在することがわかる．例えば，「財布」と「ライス」のように音響的類似単語がそれらに相当する．さらに，共通の音素列を含む単語もそれらに含まれる．このような特徴を持つ単語を「難認識語」と呼ぶことにする．また，難認識語以外の単語を「易認識語」と呼ぶことにする．

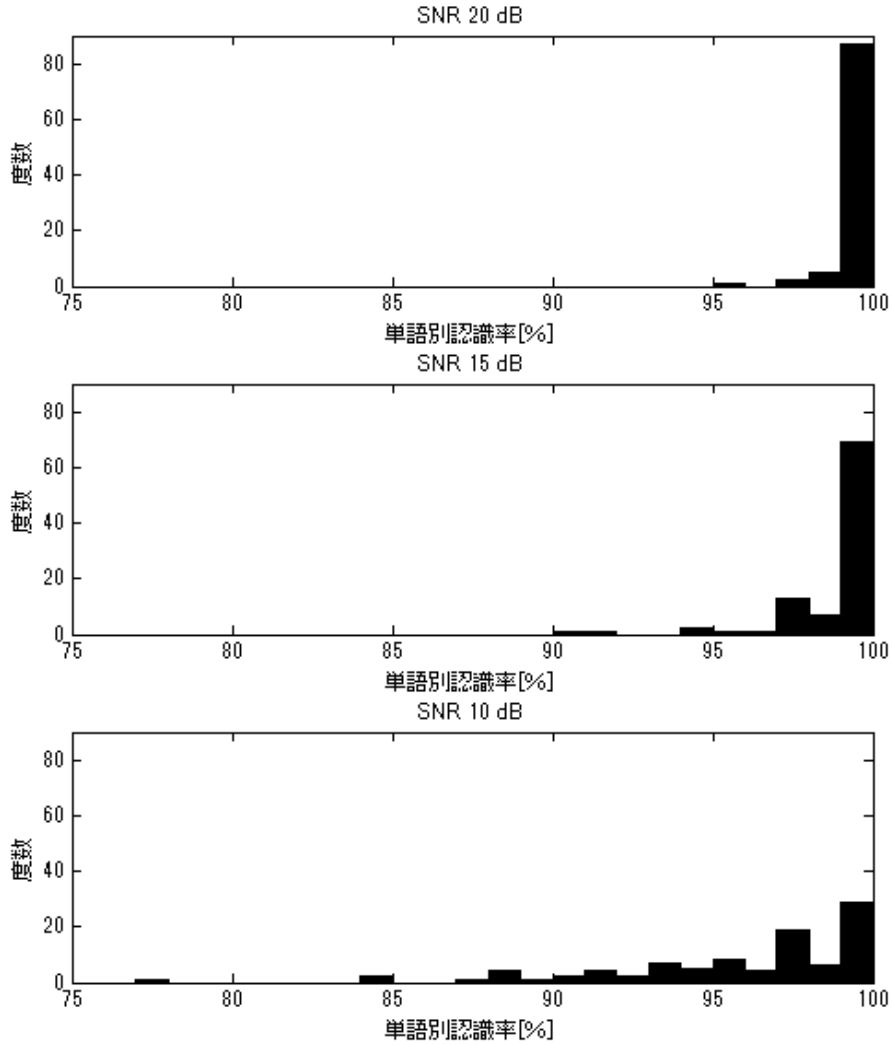


図 3.1 雑音ロバスト音声認識の単語別認識率のヒストグラム

ここで,「財布」と「ライス」を例に，単語間の音響的類似性について述べる．図 3.2 は，ある話者が「財布」と発話した 2 個の音声 S1，S2，および,「ライス」と発話した音声 R に，SNR 30 dB になるように White noise を加えて作成した雑音音声の波形である．尚，White noise を加える前に，各音声の平均パワーを揃えている．図 3.3 (a)は，各雑音音声の対数メルスペクトルのスペクトログラムである．図 3.3 (a)より，「財布」と「ライス」には，/i/と/u/の間（時刻 0.2～0.5 s）に鮮明な差異が存在することがわかる．発話開始部分にも/s/と/r/の差異が見えると期待されるが，時間が短いため，その差異は見えにくい．

表 3.1 単語一覧（抜粋）

番号	単語	番号	単語
2	トカゲ	25	ハザードランプ消（け）す
3	消化（しょうか）	27	ヘッドライト
4	文化（ぶんか）	30	上（うえ）にカーソル
5	定価（ていか）	31	フォグランプ
7	魚（さかな）	32	フォグランプ消（け）す
8	財布（さいふ）	37	ドアロック
12	畳（たたみ）	39	右（みぎ）にカーソル
13	ライス	40	左（ひだり）にカーソル
16	ランチ	43	ドアミラーオープン
18	河原（かわら）	49	ドアオープン
19	扉（とびら）	58	モニタオープン
24	ハザードランプ	63	AM ラジオ
—	—	64	FM ラジオ

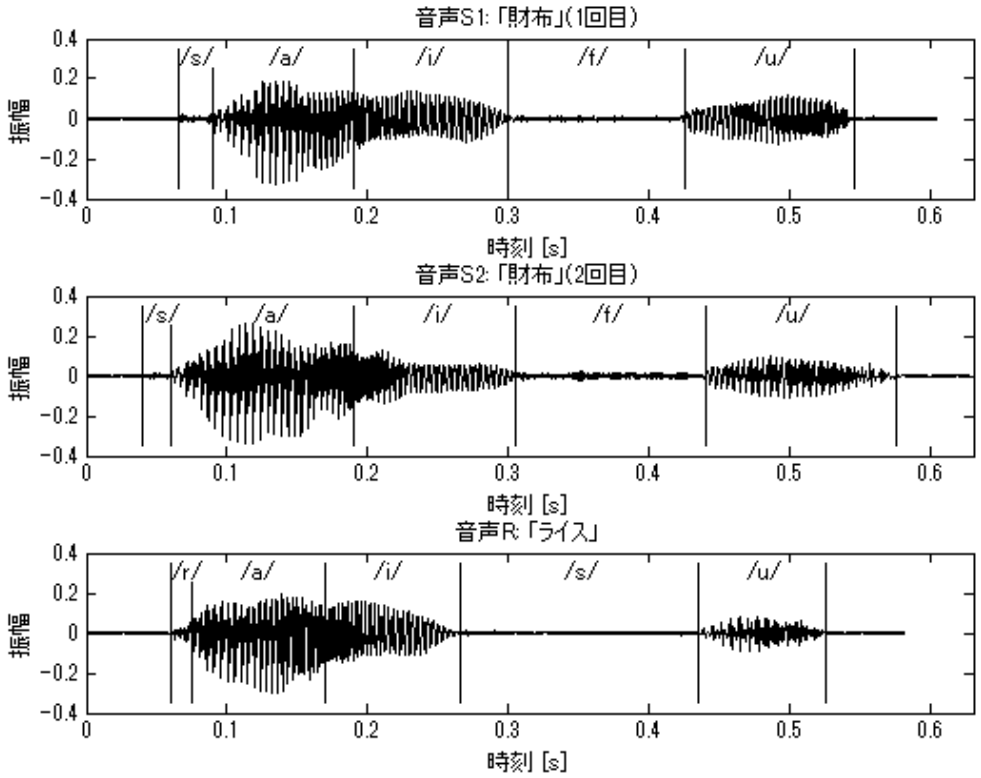
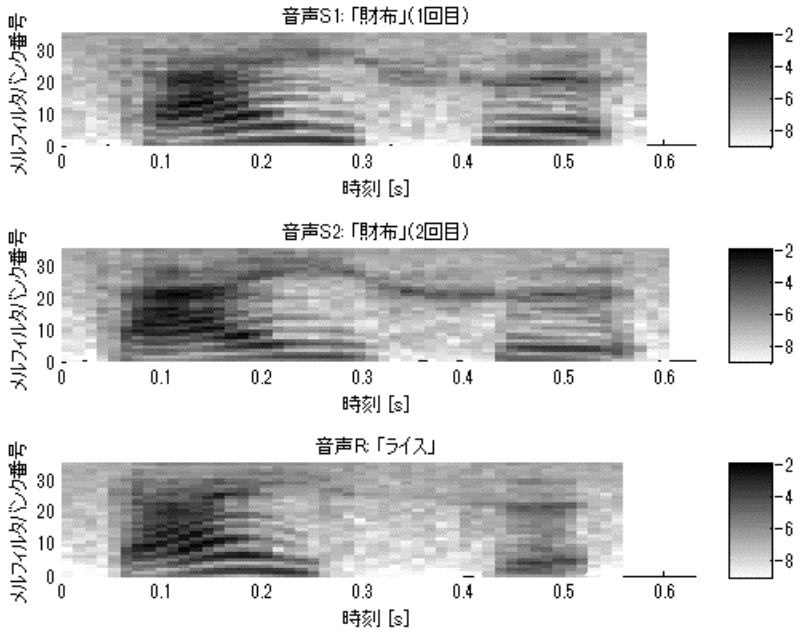


図 3.2 「財布」と「ライス」の音声波形

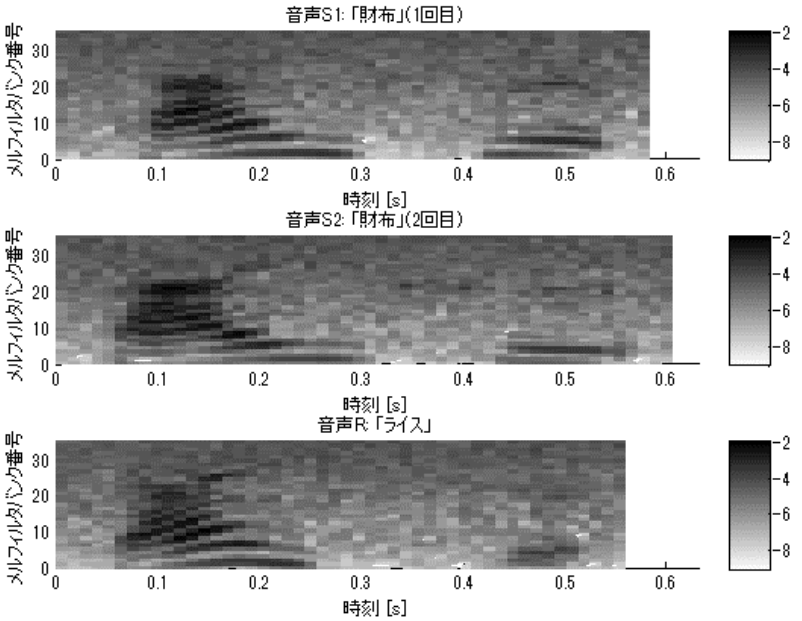
次に，同一の発話音声に，SNR 10 dB になるように White noise を加えた雑音音声の対数メルスペクトルのスペクトログラムを図 3.3 (b)に示す．図 3.3 (b)では，/i/と/u/の間の差異が図 3.3 (a)よりも不鮮明である．

音響的類似性を定量的に評価するために，DTW（dynamic time warping）距離を適用する[25]．表 3.2 に，音声 S1 を照合元データ，音声 S2，および，音声 R を照合先データとする非対称 DTW 距離を示す．

表 3.2 より，SNR 20 dB，および，10 dB では，音声 R の DTW 距離は，音声 S2 のそれよりも小さい，すなわち，音声 S1 のスペクトログラムに対し，音声 R のスペクトログラムは，音声 S2 のそれよりも類似性が大きい．



(a) SNR 30 dB



(b) SNR 10 dB

図 3.3 「財布」と「ライス」のスペクトログラム

しかし，DTW 距離による評価は，単語全体についての評価である．実際には，図 3.3 (b)の 0.2～0.3 s，および，0.4～0.5 s の間に見られるように，局所的な差異が存在する．その局所的な差異が存在する区間―音響的差異区間―を抽出して認識することができれば，認識率の向上が期待できる．

表 3.2 DTW 距離

	照合先データ	
SNR	S2	R
30 dB	305.94	309.95
20 dB	246.19	219.64
10 dB	189.37	172.63

本論文では，SNR 10 dB Factory floor noise 1 雑音環境下において，雑音ロバスト音声認識の認識率が 90%未満の単語を難認識語として扱う．ここで，90%という値に特別な意味は無く，要求される認識精度に応じて他の値を設定してもよい．表 3.3 に，難認識語と易認識語の認識率を示し，表 3.4 に，難認識語の単語別認識率を示す．尚，混合数を 8 まで増やして追加実験を行ったが，同様の結果が得られた．

表 3.3 難認識語と易認識語の認識率(%)

Noise name	SNR	易認識語	難認識語
Factory floor noise 1	10 dB	97.08	86.33
	15 dB	99.39	95.76
	20 dB	99.77	99.15
Speech babble	10 dB	97.03	88.17
	15 dB	99.53	96.84
	20 dB	99.84	99.07
White noise	10 dB	97.63	87.64
	15 dB	99.22	95.29
	20 dB	99.68	97.84

表 3.4 には，RSF の代わりに CMS を適用した場合に認識率が 90%未満となる単語も示している．CMS を適用した場合に認識率が 90%未満になる単語の個数は 14 個であるが，RSF を適用した場合のそれは，5 個少ない 9 個である．また，CMS を適用した場合の全 95 単語の平均認識率は 94.08%であるが, RSF を適用した場合のそれは約 1.2%高い 96.01%である．したがって，以降，本論文で使用した音声データベースに対して認識精度が高い RSF を適用した雑音ロバスト音声認識を採用する．

表 3.4 難認識語の単語別認識率(%)

難認識語(番号)	RSF	CMS
ハザードランプ消す(25)	77.08	68.06
財布(8)	84.03	80.56
ライス(13)	84.72	86.81
畳(12)	87.50	84.03
フォグランプ消す(32)	88.10	81.82
消化(3)	88.19	86.11
上にカーソル(30)	88.89	97.22
モニタオープン(58)	88.89	81.25
ランチ(16)	89.58	86.81
河原(18)	91.58	82.52
扉(19)	92.36	89.58
ヘッドライト(27)	95.83	88.19
ドアロック(37)	97.92	89.58
右にカーソル(39)	95.14	84.03
AM ラジオ(63)	91.67	89.58

次に，難認識語を認識したときに，誤ってどの単語を認識結果としているかを確認するために，入力に対する認識結果の個数を要素とする行列である混同行列を作成する．混同行列は 95 次正方行列であるが，難認識語に関連する部分行列を，表 3.5 に示す．番号は，表 3.1 の番号に対応する．

表 3.5 より，難認識語を誤って認識した場合，ある特定の単語を認識結果とする傾向があることがわかる．これらの単語を誤認時頻出語と呼ぶことにするが，

本論文では, 難認識語の認識を誤った時に, 3 回以上出現する認識結果の単語を, 誤認時頻出語と定めることにする. 3 回以上と定めたのは, 4 章で述べる音響的差異区間の決定に大量の計算をする必要があるという時間的な制約のためである. 表 3.6 は，表 3.5 から難認識語のそれぞれについて誤認時頻出語を探して作成した表である．誤認時頻出語は，難認識語と音響的に類似していることが多い．例えば,「財布」と「ライス」のように母音の並びが同一である場合があり,「ヘッドライト」と「ヘッドライト消す」のように共通部分の割合が多い場合もある.

文献[22]によると，難認識語であっても，パワーとスペクトルエントロピーに着目して発話開始直後の破裂音を切り出して認識することにより，単語の先頭に相異なる破裂音を持つ音響的類似単語の認識精度を改善できることが示されている.

表 3.5 混同行列

(a) 混同行列の部分 1

		認識結果							
	番号	2	7	12	3	4	5	19	他
入力単語	2	140	0	0	0	0	0	0	4
	7	0	143	0	0	0	0	0	1
	12	5	4	126	1	0	0	0	8
	3	0	0	0	127	5	5	3	4
	4	0	0	0	0	142	0	0	2
	5	0	0	0	0	6	135	0	3
	19	0	0	0	0	2	1	133	8

表 3.5 混同行列

(b) 混同行列の部分 2

		認識結果							
	番号	8	13	16	24	25	31	32	他
入力単語	8	121	17	2	0	0	0	0	4
	13	17	122	0	0	0	0	0	5
	16	1	7	129	0	0	0	0	7
	24	0	0	0	141	3	0	0	0
	25	0	0	0	33	111	0	0	0
	31	0	0	0	1	0	139	2	2
	32	0	0	0	0	4	12	126	1

(c) 混同行列の部分 3

		認識結果						
	番号	30	39	40	43	49	58	他
入力単語	30	128	11	4	0	0	0	1
	39	2	137	3	0	0	0	2
	40	4	0	136	0	0	0	4
	43	0	0	0	138	0	1	5
	49	0	0	0	0	135	0	9
	58	0	0	0	4	7	128	5

表 3.6 難認識語と誤認時頻出語のペア

難認識語(番号)	誤認時頻出語(番号)
ハザードランプ消す(25)	ハザードランプ(24)
財布(8)	ライス(13)
ライス(13)	財布(8)
畳(12)	トカゲ(2) 魚(7)
フォグランプ消す(32)	フォグランプ(31) ハザードランプ消す(25)
消化(3)	文化(4) 定価(5) 扉(19)
上にカーソル(30)	右にカーソル(39) 左にカーソル(40)
モニタオープン(58)	ドアミラーオープン(43) ドアオープン(49)
ランチ(16)	ライス(13)

3.2 音響的差異区間

本論文では，発話音声の中から音響的に類似しない区間を切り出し，その区間を認識するという手法により,音響的類似単語の認識精度の改善を試みる.また，この区間のことを音響的差異区間と呼ぶことにする．音響的差異区間の切り出しは，ビタビアルゴリズムにより，自動的に行うことができる．したがって，その区間が，単語のどの位置にあっても，そして，破裂音でなくても適用できるという利点がある.

第 4 章

音響的類似単語の雑音ロバストフレーズ音声認識

本章では，認識精度が最も高いと期待される音響的差異区間を決定する方法について説明する．

4.1 音響的差異区間候補の抽出

単語wを発声したr番目の教師音声$s_i^{(w,r)}$を分析して得られる音声特徴量列$X^{(w,r)}$について，単語vの HMM に対する対数尤度をビタビアルゴリズムで計算すると，ビタビ経路$q^{(w,v,r)}$を求めることができる．

$$\boldsymbol{q}^{(w,v,r)} = \underset{\boldsymbol{q}'}{\operatorname{argmax}} \ln P\left(\boldsymbol{X}^{(w,r)}, \boldsymbol{q}' \middle| \boldsymbol{\theta}_v\right) \tag{4.1}$$

このビタビ経路を利用して，図 4.1 に示すように，単語wの音声から，単語vの HMM の始端状態番号nから終端状態番号mまでに対応する区間に存在する音声を切り出す．この区間を，音響的差異区間候補と呼ぶことにする．

$q_t^{(w,v,r)} = n$を満たす最小のフレーム番号tを$b^{(w,v,n,r)}$，$q_t^{(w,v,r)} = m$ を満たす最大のフレーム番号tを$e^{(w,v,n,r)}$とすると，サンプリング番号$128b^{(w,v,n,r)}$から$\left(128e^{(w,v,m,r)} + 255\right)$までの音声が切り出される．したがって，切り出された音声は

$$s'^{(w,\boldsymbol{V},r)}_{i} = s^{(w,r)}_{i+128b^{(w,v,n,r)}} \tag{4.2}$$

となる．ここで，$0 \le i \le 128\left(e^{(w,v,m,r)} - b^{(w,v,n,r)}\right) + 255$ であり，$\boldsymbol{V}$は，単語v，始端状態番号n，終端状態番号mを要素とするベクトルである．

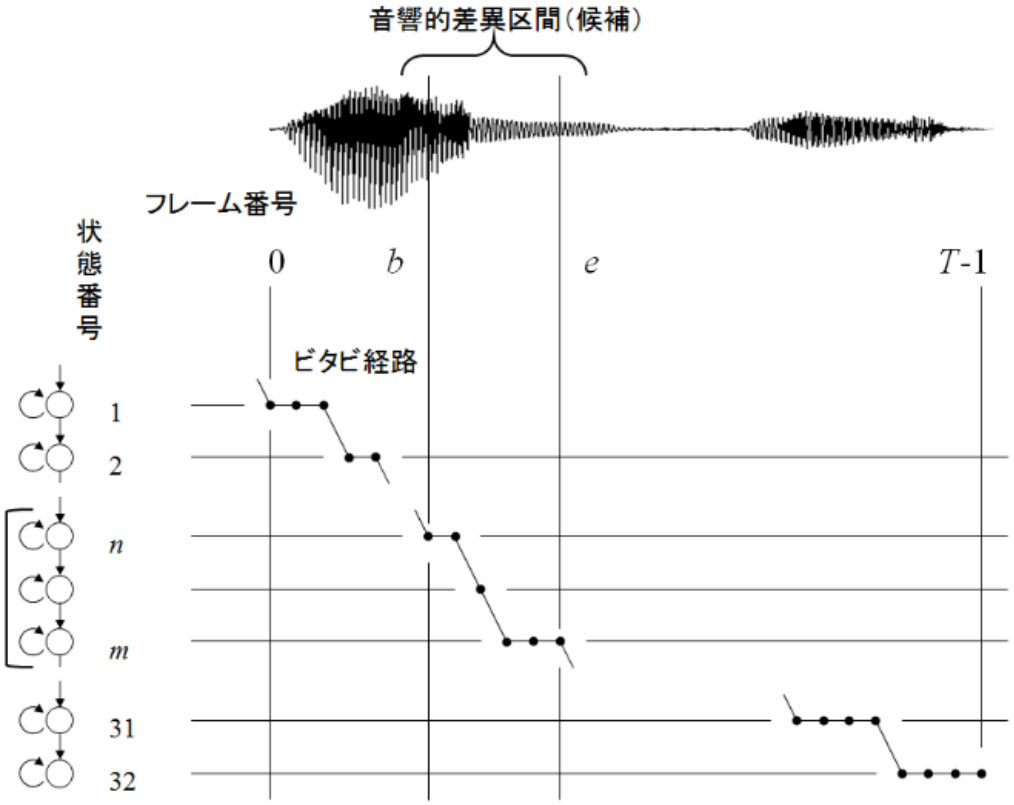


図 4.1 音響的差異区間抽出

4.2 音響的差異区間候補の音声分析

難認識語と誤認時頻出語の母音の並びが同一の場合，音響的に類似しない区間には子音が存在する．子音は母音よりも継続時間が短く，非定常的であることが多いため，音響的差異区間候補の音声分析におけるフレーム長とフレームシフト長を，それぞれ，従来手法の半分の長さとする．表 4.1 に，雑音ロバスト差異区間音声認識における音声分析の条件を従来手法のそれと異なる項目について示す．

雑音ロバスト音声認識では，RSF により乗法性雑音を除去するが，このフィルタの次数は 184 であるから，フィルタ長は時間に換算すると約 2 秒間である．多くの音響的差異区間候補の時間的長さはこれよりも短くなるため，音響的差異区間候補の音声に RSF を適用するのは適切ではない．そこで，雑音ロバスト処理において，RSF ではなく，CMS を適用する．

表 4.1 差異区間音声分析条件

フレーム長	11.6 ms　(128 points)
フレームシフト長	5.8 ms　(64 points)
雑音ロバスト処理	Cepstral mean subtraction Dynamic range adjustment

音響的差異区間候補の音声信号 $s'^{(w,V,r)}_i$ を図 2.29 に示すフローチャートに従って，表 4.1 に示す条件で分析すると，音声特徴量ベクトル列 $\boldsymbol{X}'^{(w,V,r)} = \left[\boldsymbol{x}'^{(w,V,r)}_0\ \boldsymbol{x}'^{(w,V,r)}_1\ \cdots\ \boldsymbol{x}'^{(w,V,r)}_{T'^{(w,V,r)}-1}\right]$ が得られる．ここで，$\boldsymbol{x}'^{(w,V,r)}_t$は，$t$番目の音声フレーム$s'^{(w,V,r)}_t = \left[s'^{(w,V,r)}_{64t}\ s'^{(w,V,r)}_{64t+1}\ \cdots\ s'^{(w,V,r)}_{64t+127}\right]$ から計算された 38 次元音声特徴量ベクトルである．また，$T'^{(w,V,r)}$は，$\left(e^{(w,v,m,r)} - b^{(w,v,n,r)} + 1 - 128\right)/64 + 1$ 以下の最大の整数である．

4.3 音響的差異区間候補の学習

音響的差異区間候補の音声 $s_i'^{(w,\boldsymbol{V},r)}$ を教師音声とし，バウム・ウェルチアルゴリズムにより，音響的差異区間候補 HMM のパラメータ $\boldsymbol{\lambda}_w^{(\boldsymbol{V})}$ を求める．音響的差異区間候補 HMM の状態数は，切り出す範囲に対応する単語 HMM の状態数の 2 倍とし，混合数は 1 とした．

4.4 音響的差異区間の決定

難認識語d，あるいは，その誤認時頻出語fを発声した音声を最も高い精度で認識すると期待される音響的差異区間をその候補の中から決定する．

単語d，および，fの 2 つの単語 HMM を用いて，従来手法により，単語w_1を発声した N_{w_1} 個の音声を認識する．ここでw_1は，dあるいはf のいずれかである．そして，認識結果がw_2 である音声の個数を N_{w_1,w_2} とする．ここでw_2は，dあるいはf のいずれかである．認識に使用する音声は，単語 HMM の教師話者音声に SNR 10 dB Speech babble，または，SNR 10 dB White noise を重畳した音声とした．

さらに，認識結果がd の場合に，HMM パラメータが $\boldsymbol{\lambda}_d^{(\boldsymbol{U})}$，$\boldsymbol{\lambda}_f^{(\boldsymbol{U})}$ の 2 つの音響的差異区間候補 HMM を用いて，再度認識した結果がw_3となる音声の個数を$N_{w_1,d,w_3}^{(\boldsymbol{U})}$とする．ここで$w_3$は，$d$あるいは$f$ のいずれかである．また，$\boldsymbol{U}=[u,n_u,m_u]$ の各要素について，単語u は，dあるいはf のいずれかであり，状態番号n_u，m_u は，$1 \le n_u \le m_u \le 32$を満たす．認識結果がf の場合も同様に，HMM パラメータが $\boldsymbol{\lambda}_d^{(\boldsymbol{V})}$，$\boldsymbol{\lambda}_f^{(\boldsymbol{V})}$の 2 つの音響的差異区間候補 HMM を用いて，再度認識した結果がw_3となる音声の個数を $N_{w_1,f,w_3}^{(\boldsymbol{V})}$ とする．ここで，$\boldsymbol{V}=[v,n_v,m_v]$ の各要素について，単語vは，dあるいはf のいずれかであり，状態番号n_v，m_vは，$1 \le n_v \le m_v \le 32$を満たす．

音響的差異区間候補 HMM による難認識語dの認識精度は，

$\left(N_{d,d,d}^{(\boldsymbol{U})}+N_{d,f,d}^{(\boldsymbol{V})}\right)/N_d$ であるから，これを最大にする$\boldsymbol{U}$，$\boldsymbol{V}$を探す．そして，最大値が，従来手法による難認識語$\boldsymbol{d}$の認識精度$N_{d,d}/N_d$を超えていれば認識精度の改善を期待できる．したがって，$\boldsymbol{U}$，$\boldsymbol{V}$は，次の制約条件を満たす必要がある．

$$N_{d,d,d}^{(\boldsymbol{U})}+N_{d,f,d}^{(\boldsymbol{V})}>N_{d,d} \tag{4.3}$$

最大化の結果，誤認時頻出語$\boldsymbol{f}$の認識精度が低下しないという次の制約条件も課す．

$$N_{f,d,f}^{(\boldsymbol{U})}+N_{f,f,f}^{(\boldsymbol{V})}\geq N_{f,f} \tag{4.4}$$

(4.4)式の左辺，および，右辺をN_f で除算すると，それぞれ，音響的差異区間候補 HMM，および，単語 HMM による誤認時頻出語$\boldsymbol{f}$の認識精度になる．

ところで，従来手法による認識結果が$\boldsymbol{d}$ の場合に，認識精度が改善する音響的差異区間候補 HMM が存在しない場合がある．このような場合，次の制約条件(4.5), (4.6)の下，$N_{d,f,d}^{(\boldsymbol{V})}$ を最大化する$\boldsymbol{V}$ を探す．

$$N_{d,f,d}^{(\boldsymbol{V})}>0 \tag{4.5}$$

$$N_{f,f,f}^{(\boldsymbol{V})}\geq N_{f,f} \tag{4.6}$$

以上の制約条件を満たす $\boldsymbol{U}$，$\boldsymbol{V}$ が存在しない場合，音響的差異区間 HMM による認識を行わずに，単語 HMM による認識結果を採用する．

認識精度を最大化する $\boldsymbol{U}$，$\boldsymbol{V}$ が複数ある場合，以下に述べる基準により選択した．まず，切り出す範囲に関する状態数 (m_u-n_u+1)，(m_v-n_v+1) が最少のものを選択した．そのような $\boldsymbol{U}$，$\boldsymbol{V}$ が複数ある場合は，始端状態番号 n_u，n_v がよ

り小さいもの，さらには，単語番号 u，v がより小さいものを選択した．

尚，「財布」と「ライス」のペアのように，単語 f が難認識語，単語 d がその誤認時頻出語でもある場合，$d_1 = d$，$d_2 = f$ として，次の制約条件(4.7)，(4.8)

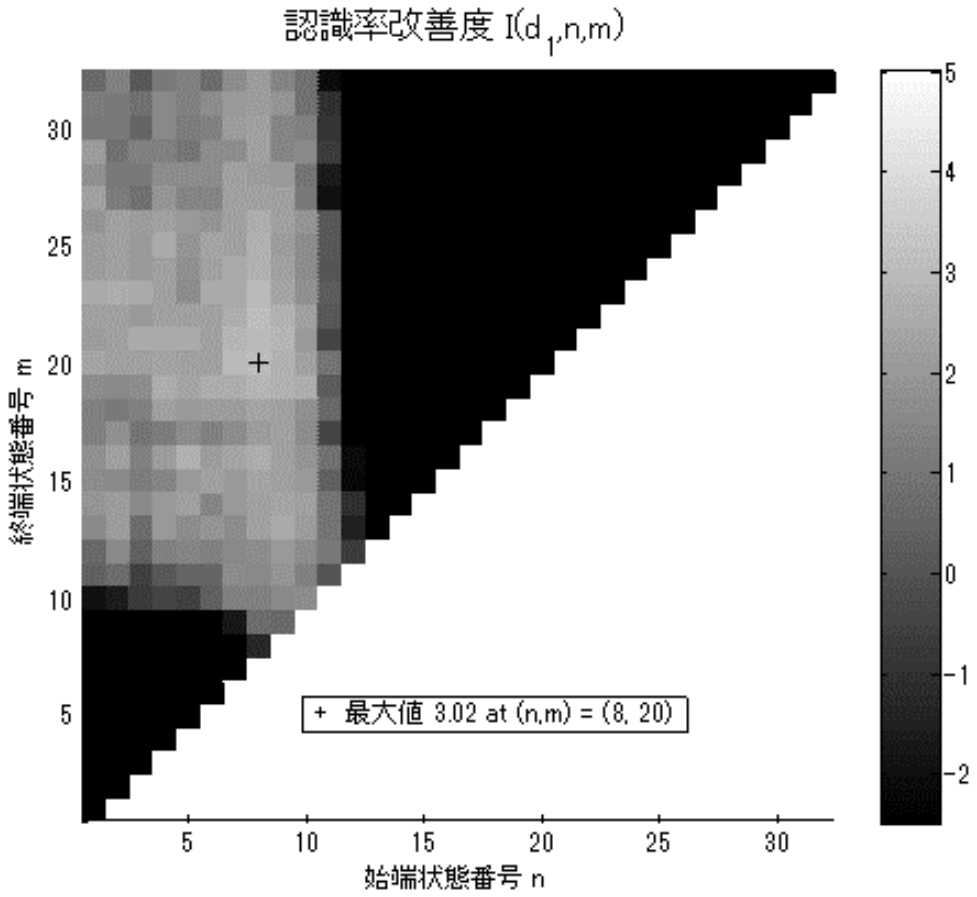
$$N^{(\boldsymbol{U})}_{d_1,d_1,d_1} + N^{(\boldsymbol{V})}_{d_1,d_2,d_1} \geq N_{d_1,d_1} \tag{4.7}$$

$$N^{(\boldsymbol{V})}_{d_2,d_2,d_2} + N^{(\boldsymbol{U})}_{d_2,d_1,d_2} \geq N_{d_2,d_2} \tag{4.8}$$

の下，次の(4.9)式を最大にする $\boldsymbol{U}$，$\boldsymbol{V}$ を探す．

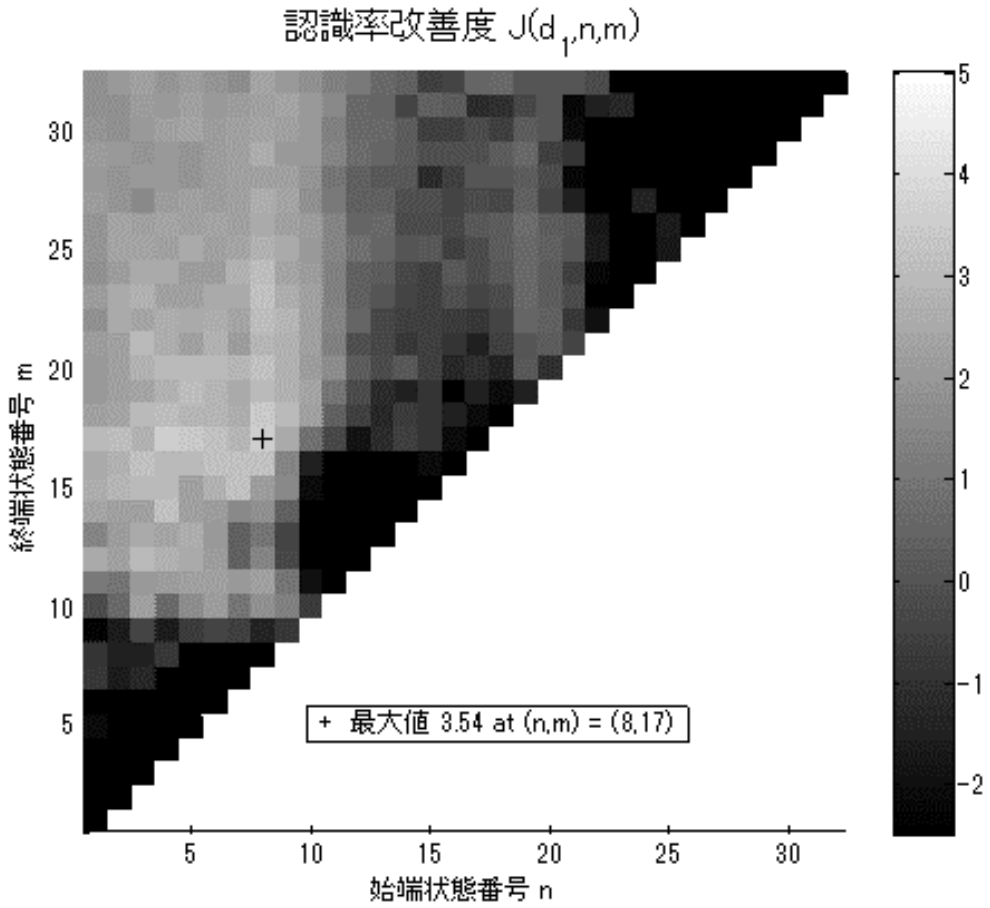
$$N^{(\boldsymbol{U})}_{d_1,d_1,d_1} + N^{(\boldsymbol{U})}_{d_2,d_1,d_2} + N^{(\boldsymbol{V})}_{d_1,d_2,d_1} + N^{(\boldsymbol{V})}_{d_2,d_2,d_2} \tag{4.9}$$

図 4.2 は, 交差検証のために生成した 6 個のデータセットの中の 1 個において, 単語 d_1 = 「財布」，d_2 = 「ライス」の教師話者音声に対する音響的差異区間候補 HMM による認識率から従来手法による認識率を減算して得られる認識率改善度（単位はポイント）をプロットした図である．



(a) 改善事例 1

図 4.2 音響的差異区間候補に対する認識率改善マップ



(b) 改善事例 2

図 4.2 音響的差異区間候補に対する認識率改善マップ

図 4.2 (a)にプロットした認識率改善度 $I(d_1,n,m)$ は，雑音ロバスト音声認識の結果が「財布」のときにのみ，d_1 =「財布」の単語 HMM の始端状態番号 n から終端状態番号 m までに対応付けられた音声を音響的差異区間候補 HMM による認識を行った場合の認識率改善度である．

$$I(d_1,n,m)=\frac{N^{(\boldsymbol{U})}_{d_1,d_1,d_1}+N^{(\boldsymbol{U})}_{d_2,d_1,d_2}-N_{d_1,d_1}}{N_{d_1}+N_{d_2}} \tag{4.10}$$

ここで，$\boldsymbol{U}=[d_1,n,m]$である．$I(d_1,n,m)$は，図 4.2 (a)に＋印で示しているが，$(n,m)=(8,20)$ のとき，最大となり，その値は 3.02 である．また，雑音ロバスト音声認識の結果が「財布」のときにのみ，d_2 = 「ライス」の単語 HMM の始端状態番号nから終端状態番号mまでに対応付けられた音声を音響的差異区間候補 HMM による認識を行った場合の認識率改善度 $I(d_2,n,m)$ は，$(n,m)=(4,17)$ のとき，最大となり，その値は 3.02 である．したがって，式(4.9)を最大にする $\boldsymbol{U}$ は，$\boldsymbol{U}=[d_1,8,20]$である．

図 4.2 (b)にプロットした認識率改善度 $J(d_1,n,m)$ は，雑音ロバスト音声認識の結果が「ライス」のときにのみ，d_1 =「財布」の単語 HMM の始端状態番号nから終端状態番号 m までに対応付けられた音声を音響的差異区間候補 HMM による認識を行った場合の認識率改善度である．

$$J(d_1,n,m)=\frac{N^{(\boldsymbol{V})}_{d_1,d_2,d_1}+N^{(\boldsymbol{V})}_{d_2,d_2,d_2}-N_{d_2,d_2}}{N_{d_1}+N_{d_2}} \tag{4.11}$$

ここで，$\boldsymbol{V}=[d_1,n,m]$ である．$J(d_1,n,m)$ は，図 4.2 (b)に＋印で示しているが，$(n,m)=(8,17)$ のとき，最大となり，その値 3.54 である．また，雑音ロバスト音

声認識の結果が「ライス」のときにのみ，d_2=「ライス」の単語 HMM の始端状態番号nから終端状態番号mまでに対応付けられた音声を音響的差異区間候補 HMM による認識を行った場合の認識率改善度 $J(d_2, n, m)$ は，$(n, m) = (7,17)$ のとき，最大となり，その値は 3.33 となる．したがって，式(4.9)を最大にする $\boldsymbol{V}$ は，$\boldsymbol{V} = [d_1, 8,17]$ である．

4.5 難認識語認識

提案手法である難認識語認識について説明する．図 4.3 に難認識語認識のフローチャートを示す．

難認識語認識は，従来手法である雑音ロバスト音声認識と音響的差異区間に対する音声認識（雑音ロバスト差異区間音声認識）を組み合わせた手法である．難認識語認識の手順は，最尤 HMM 探索までは，従来手法と同一であり，その認識結果を w_1 とする．

$$w_1 = \underset{w\prime}{\operatorname{argmax}} \ln P(\boldsymbol{X}|\boldsymbol{\theta}_{w\prime}) \tag{4.12}$$

次に，尤度 2 位の HMM を探索し，その結果を w_2 とする．

$$w_2 = \underset{w\prime \neq w_1}{\operatorname{argmax}} \ln P(\boldsymbol{X}|\boldsymbol{\theta}_{w\prime}) \tag{4.13}$$

w_1とw_2 が，表 3.6 に示した難認識語と誤認時頻出語のペアである場合，4 章で決定した音響的差異区間の音声を切り出して，音響的差異区間 HMM で認識を行い，その結果wを出力する．

$$w = \underset{w' \in \{w_1, w_2\}}{\operatorname{argmax}} \ln P(\boldsymbol{X}' | \boldsymbol{\lambda}_{w'}) \tag{4.14}$$

ペアでない場合，従来手法と同じく，尤度最大の HMM の単語w_1を認識結果として出力する．

尚，図 4.3 には示していないが，w_1 と w_2 が難認識語と誤認時頻出語のペアであったとしても，認識精度を改善する音響的差異区間が見つからなかった場合，尤度最大の HMM の単語w_1を認識結果として出力する．

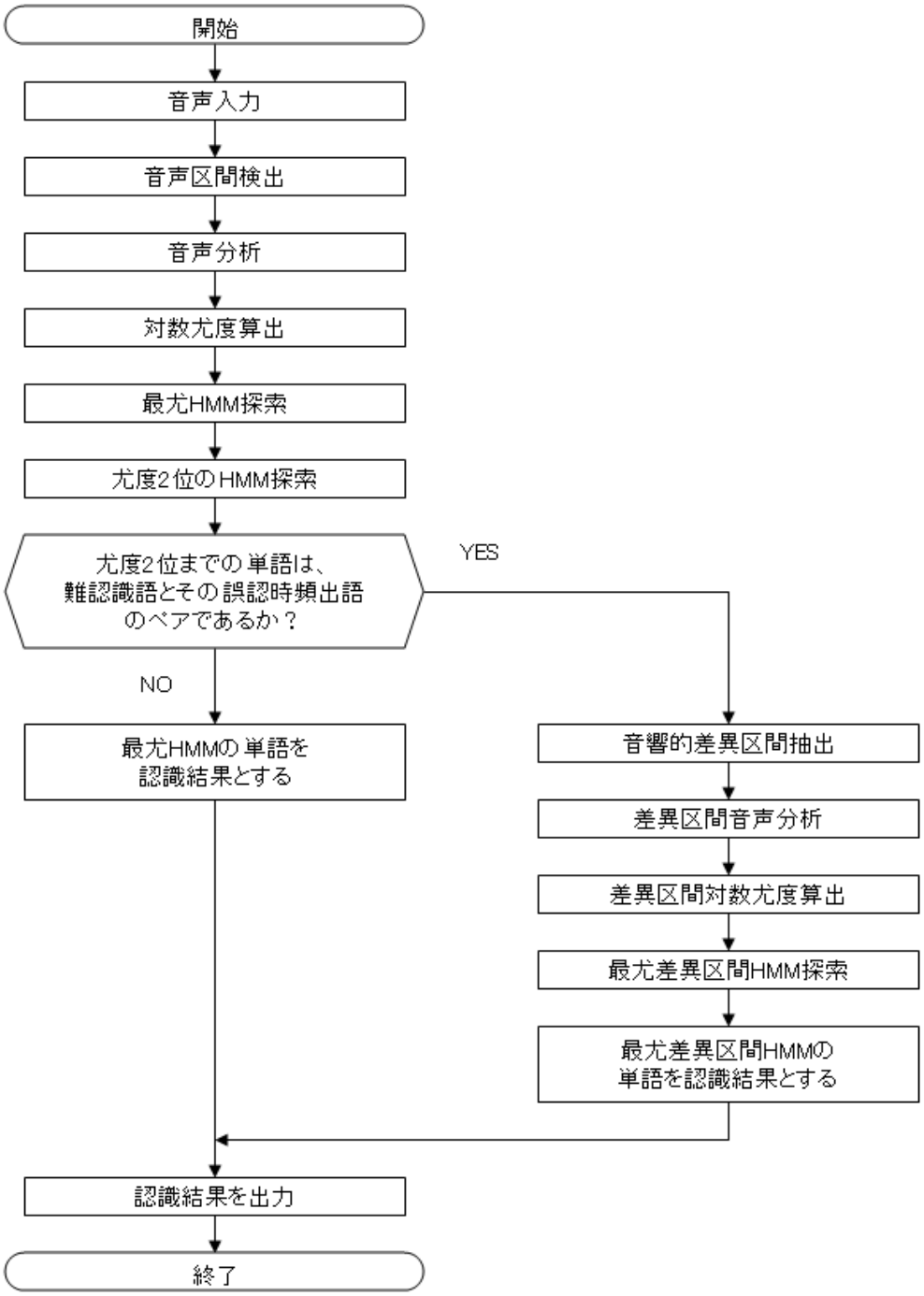


図 4.3 難認識語認識フローチャート

第 5 章

評価

5.1 実験結果

全話者 48 名を 8 名ずつの 6 グループに分けて，1 グループ 8 名を認識話者，残りの 5 グループ 40 名を教師話者とし，交差検証法により従来手法と提案手法の認識率を評価した．

表 5.2 は，難認識語の認識率を示した表である．提案手法の平均認識率は，すべての場合で従来手法のそれ以上であり，SNR 10 dB の雑音環境下において，Factory floor noise 1，Speech babble，White noise について，それぞれ，86.33%から 90.43%，88.17%から 92.43%，87.64%から 94.06%に向上した．単語別の認識率について述べると，表 5.2 に示すように，SNR 15 dB，および，10 dB の Speech babble 雑音環境下における「財布(8)」を除いて，提案手法の認識率は，従来手法のそれ以上である．尚，認識率が低下した場合でも，「財布」とペアになる「ライス(13)」の認識率に対する改善度は，「財布」のそれに対する改悪度を大幅に上回る．

表 5.2 は，誤認時頻出語（ただし，難認識語でもある「財布」，「ライス」，「ハザードランプ消す(25)」を除く）の Factory floor noise 1 雑音環境における認識率

を示した表である．提案手法の平均認識率は，SNR 15 dB，および，20 dB において, 従来手法のそれと同じであったが, SNR 10 dB においては, 約 0.5%低下した．しかしながら，難認識語の認識率に対する改善度のほうが大幅に上回る．誤認時頻出語の単語別の認識率について述べると，表 5.2 に示すように，SNR 15 dB における「フォグランプ(31)」，および，SNR 10 dB における「ハザードランプ(24)」と「フォグランプ」を除いて，提案手法の認識率は，従来手法のそれ以上である．表 5.3 に，認識率が低下した誤認時頻出語について，提案手法，および，従来手法における混同行列の部分行列を示す．尚，従来手法における部分行列の要素を括弧付きの数値で示した．表 5.3 に示すように，これらの誤認時頻出語とペアになる難認識語の認識精度に対する改善度のほうが上回っている．尚, Speech babble, および，White noise 雑音環境下についても同様の結果が得られた．

表 5.1 難認識語の単語別認識率(%)

(a) Factory floor noise 1 雑音環境

番号	Conventional			Proposed		
	10 dB	15 dB	20 dB	10 dB	15 dB	20 dB
3	88.19	97.92	100.00	90.97	99.31	100.00
8	84.03	90.28	97.22	86.11	93.06	97.92
12	87.50	97.92	99.31	91.67	98.61	100.00
13	84.72	95.83	98.61	90.97	98.61	100.00
16	89.58	97.92	100.00	90.28	98.61	100.00
25	77.08	91.67	99.31	86.81	97.22	100.00
30	88.89	94.44	98.61	90.28	95.14	98.61
32	88.10	97.22	99.31	97.19	100.00	100.00
58	88.89	98.61	100.00	89.58	99.31	100.00
Ave	86.33	95.76	99.15	90.43	97.76	99.61

表 5.1 難認識語の単語別認識率(%)

(b) Speech babble 雑音環境

番号	Conventional			Proposed		
	10 dB	15 dB	20 dB	10 dB	15 dB	20 dB
3	95.83	100.00	100.00	96.53	100.00	100.00
8	92.36	97.92	98.61	91.67	96.53	98.61
12	95.14	98.61	100.00	95.83	98.61	100.00
13	78.47	93.06	97.22	84.72	100.00	100.00
16	93.75	100.00	100.00	93.75	100.00	100.00
25	70.14	88.89	100.00	88.89	97.92	100.00
30	84.03	94.44	96.53	85.42	94.44	96.53
32	86.59	98.61	99.31	96.47	99.31	100.00
58	97.22	100.00	100.00	98.61	100.00	100.00
Ave	88.17	96.84	99.07	92.43	98.53	99.46

表 5.1 難認識語の単語別認識率(%)

(c) White noise 雑音環境

番号	Conventional			Proposed		
	10 dB	15 dB	20 dB	10 dB	15 dB	20 dB
3	72.22	90.28	94.44	90.28	96.53	97.92
8	83.33	89.58	94.44	90.97	95.14	95.83
12	98.61	99.31	99.31	100.00	100.00	100.00
13	88.19	95.83	98.61	91.67	98.61	99.31
16	95.83	98.61	99.31	97.22	98.61	99.31
25	77.78	93.06	97.92	93.75	99.31	100.00
30	85.42	93.75	97.22	85.42	94.44	97.22
32	88.77	97.89	99.31	97.89	100.00	100.00
58	98.61	99.31	100.00	99.31	99.31	100.00
Ave	87.64	95.29	97.84	94.06	97.99	98.84

表 5.2 誤認時頻出語の単語別認識率(%)

番号	Conventional			Proposed		
	10 dB	15 dB	20 dB	10 dB	15 dB	20 dB
2	97.22	99.31	100.00	97.22	99.31	100.00
4	98.61	99.31	100.00	98.61	99.31	100.00
5	93.75	98.61	98.61	93.75	98.61	98.61
7	99.31	99.31	100.00	99.31	99.31	100.00
19	92.36	98.61	99.31	92.36	98.61	99.31
24	97.92	99.31	100.00	93.06	99.31	100.00
31	96.53	97.92	100.00	93.75	96.53	100.00
39	95.14	100.00	100.00	95.14	100.00	100.00
40	94.44	98.61	100.00	96.53	100.00	100.00
43	95.83	97.92	98.61	96.53	97.92	98.61
49	93.75	99.31	99.31	93.75	99.31	99.31
Ave	95.90	98.93	99.62	95.45	98.93	99.62

表 5.3 混同行列の比較

(a) SNR 15 dB の場合

		認識結果				
	番号	24	25	31	32	他
入力単語	24	143 (143)	1 (1)	0 (0)	0 (0)	0 (0)
	25	4 (12)	140 (132)	0 (0)	0 (0)	0 (0)
	31	2 (2)	0 (0)	139 (141)	3 (1)	0 (0)
	32	0 (0)	0 (1)	0 (3)	143 (139)	0 (0)

(b) SNR 10 dB の場合

		認識結果				
	番号	24	25	31	32	他
入力単語	24	134 (141)	10 (3)	0 (0)	0 (0)	0 (0)
	25	19 (33)	125 (111)	0 (0)	0 (0)	0 (0)
	31	1 (1)	0 (0)	135 (139)	6 (2)	2 (2)
	32	0 (0)	0 (4)	3 (12)	139 (126)	0 (0)

5.2 単語認識率

最後に，表 5.4 に，全 95 単語の認識率を示す．全単語の認識精度に対する改善度は小さいが，これは全単語に占める難認識語の割合が約 10 %と小さいためである．

表 5.4 全単語の認識率(%)

Noise name	SNR	Conventional	Proposed
Factory floor noise 1	10 dB	96.01	96.35
	15 dB	99.04	99.23
	20 dB	99.71	99.75
Speech babble	10 dB	96.13	96.50
	15 dB	99.24	99.43
	20 dB	99.77	99.80
White noise	10 dB	96.63	97.27
	15 dB	98.82	99.08
	20 dB	99.50	99.60

第 6 章

結論

提案手法により，SNR 10 dB 雑音環境下における難認識語の認識率を，Factory floor noise 1, Speech babble, White noise について, それぞれ, 86.33%から 90.43%, 88.17%から 92.43%, 87.64%から 94.06%に向上することができた．したがって, マルチコンディション学習に使用していない Factory floor noise 1 雑音環境下でも認識率を改善できている．

一方，表 5.2 に示したように，SNR 10 dB Factory floor noise 1 雑音環境における誤認時頻出語（ただし，難認識語を除く）の平均認識率が約 0.5%低下した．しかしながら, 難認識語の認識率に対する改善度のほうが大幅に上回っているため, 提案手法は有用であると考えているが，今後，解決すべき課題としたい．

謝辞

本研究の遂行にあたり，適切なる御指導，御助言，多大なるご援助を頂き，常に有益な討論をして頂いた，北海道大学大学院情報科学研究科メディアネットワーク専攻情報通信システム学講座情報通信ネットワーク研究室宮永喜一名誉教授と筒井弘准教授に深く感謝し御礼申し上げます.

また，常に有益で適切な御討論をして頂き，貴重な御意見を頂いた株式会社レイトロンの研究員諸氏に厚く御礼申し上げます.

参考文献

[1] S.B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. on Speech and Signal Processing, pp.357-366, 1980.

[2] Sadaoki Furui, “Speaker-Independent isolated word recognition using dynamic features of speech spectrum,” IEEE Trans. on Acoust., Speech, and Signal Process., vol.ASSP-34, no.1, pp.52-59, Feb. 1986.

[3] Rabiner L.R., “A Tutorial on Hidden Markov Models and Se-lected Applications in Speech Recognition,” Proceedings of the IEEE, vol.77, no.2, Feb. 1989.

[4] K.M. Knill et al., “Use of Gaussian Selection in Large Vocabulary Continuous Speech Recognition Using HMMs,” in Proc. ICSLP 96, pp.470-473, 1996.

[5] Xuedong Huang, “Spoken Language Processing,” Prentice Hall, 2001, ISBN 0-13-022616-5.

[6] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, “Sphinx-4: A flexible open source framework for speech recognition,” Tech. Rep., Sun Microsystems Inc., 2004.

[7] A. Lee and T. Kawahara, “Recent Development of Open-Source Speech Recognition Engine Julius,” in Proc. Asia-Pacfic Signal and Information Processing Association Annual Summit and Conference, pp.131-137, Oct. 2009.

[8] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, “Robust automatic speech recognition with missing and unreliable acoustic data,” Speech Communication, vol. 34, no. 3, pp. 267-285, June 2001.

[9] B. Raj, M.L. Seltzer, and R.M. Stern, “Reconstruction of missing features for robust speech recognition,” Speech Communication, vol. 43, no. 4, pp. 275-296, Sept. 2004.

[10] S. Srinivasan, N. Roman, and D.L. Wang, “Binary and ration time-frequency masks for robust speech recognition,” Speech Communication, vol. 48, no. 11, pp. 1486-1501, Nov. 2006.

[11] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, “Automatic speech recognition and speech variability: A review,” Speech Communication, vol. 49, no. 10-11, pp. 763-786, 2007.

[12] L. Mangu, E. Brill, and A. Stolcke, “Finding consensus in speech recognition: Word error minimization and other applications of confusion networks,” Computer Speech and Language, vol. 14, no. 4, pp. 373-400, Oct. 2000.

[13] C. Cerisara, S. Demange, and J. Haton, “On noise masking for automatic missing data speech recognition: A survey and discussion,” Comput. Speech Lang., vol. 21, no. 3, pp. 443-457, July 2007.

[14] M.L. Seltzer, D. Yu, and Y. Wang, “An investigation of deep nueral networks for noise robust speech recognition,” in Proc. ICASSP, 2013.

[15] N. Hayasaka, Y. Miyanaga, and N. Hataoka, “Running spectrum filter for robust speech recognition,” IEICE Technical Report, CAS2003-6, VLD2003-16, DSP2003-36, 2003.

[16] Noboru Hayasa, Yoshikazu Miyanaga and Naoya Wada, “Running spectrum filtering in speech recognition,” SCIS Signal Processing and Communications with Soft Computing, Oct 2002.

[17] N. Wada, S. Yoshizawa and Y. Miyanaga, “A consideration about robust speech feature extraction for isolated word speech recognition,” in Proc. International Symposium on Intelligent Signal Processing and Communication Systems 2003, Vol.1, pp.478-483, December 2003

[18] B.S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” J. Acoust. Soc. Am., Vol.55, pp.1304-1312, 1974

[19] M.G. Rahim and B.H. Juang, “Signal bias removal for robust telephone based speech recognition in adverse environments,” in Proc. ICASSP-94, pp.I-445-I-448 April 1994.

[20] H. Hermansky and N. Morgan, “RASTA processing of speech,” IEEE Trans. Speech and Audio Process, vol.2, pp578-579, Oct 1994.

[21] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, “On the importance of various modulation frequencies for speech recognition,” in Proc. European Conference on Speech Communication and Technology, Vol.3, pp.1079-1082 Sep 1997.

[22] Yusuke Hashimoto, Wataru Takahashi, and Yoshikazu Miyanaga, “Robust Speech

Recognition for Plosive Sounds by Extracting Missing Features," in Proc. International Conference on Embedded Systems and Intelligent Technology (ICESIT), Vol.1, pp.137-142, Jan. 2013.

[23] 渡邊 真紀, 宮永 喜一,"線形予測理論を用いた類似発音単語識別のためのロバスト音声認識," 電子情報通信学会, 信学技報, Vol.113, No.78, pp.59-64, SIS2013-12, Jun. 2013.

[24] http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

[25] H. Sakoe, and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. on Acoust., Speech, and Signal Process., vol.ASSP-26, no.1, pp.43-49, Feb. 1978.

[26] レイ・D・ケント, チャールズ・リード著, 荒井 隆行, 菅原 勉 監訳, 音声の音響分析, 海文堂出版

[27] 古井 貞熙, 新音響・音声工学, 近代科学社

[28] 荒木 雅弘, フリーソフトでつくる音声認識システム, 森北出版

[29] 宮﨑 善行, 荒金 康人, 宮永 喜一, "音響的類似単語の雑音ロバストフレーズ音声認識," Journal of Signal Processing, 19 巻, 5 号, pp. 195-207, Sep. 2015. [doi.org/10.2299/jsp.19.195]

[30] 宮崎 善行, 荒金 康人, 宮永 喜一, "雑音環境下における音響的類似単語の認識について," 電子情報通信学会, 信学技報, vol. 113, no. 467, SIS2013-57, pp. 11-16, 2014 年 3 月.

著者による発表論文

1.宮﨑 善行, 荒金 康人, 宮永 喜一, “音響的類似単語の雑音ロバストフレーズ音声認識,” Journal of Signal Processing, 19 巻, 5 号, pp. 195-207, Sep. 2015. [doi.org/10.2299/jsp.19.195]

2.Yoshiyuki Miyazaki, Yasuhito Arakane, Yoshikazu Miyanaga, “Noise robust phrase speech recognition for similar words,” in Proc. International Symposium on Multimedia and Communication Technology (ISMAC), pp. 135-138, Aug. 2017.

3.Yoshiyuki Miyazaki, Yoshikazu Miyanaga, “New development and enterprise of robust speech recognition systems,” in Proc. International Workshop on Information Communication Technology, Aug. 2010.

4. 宮崎 善行, 荒金 康人, 宮永 喜一, “雑音環境下における音響的類似単語の認識について,” 電子情報通信学会, 信学技報, vol. 113, no. 467, SIS2013-57, pp. 11-16, 2014 年 3 月.

著作権の表示

本学位論文は著者による以下の論文・技術報告に基づき，その文章・図・表を含みます.

- 宮﨑 善行, 荒金 康人, 宮永 喜一, “音響的類似単語の雑音ロバストフレーズ音声認識,” Journal of Signal Processing, 19 巻, 5 号, pp. 195-207, Sep. 2015. [doi.org/10.2299/jsp.19.195]
  Copyright © 2015 Research Institute of Signal Processing Japan
- 宮崎 善行, 荒金 康人, 宮永 喜一, “雑音環境下における音響的類似単語の認識について,” 電子情報通信学会, 信学技報, vol. 113, no. 467, SIS2013-57, pp. 11-16, 2014 年 3 月.
  Copyright © 2014 IEICE