NAIST-IS-DT0061004

博士論文

音声対話インタフェースのための
雑音に頑健な音声入力の研究

伊田 政樹

2004年2月6日

奈良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

本論文は奈良先端科学技術大学院大学情報科学研究科に
博士 (工学) 授与の要件として提出した博士論文である.

論文番号： NAIST-IS-DT0061004

提出者： 伊田 政樹

審査委員： 鹿野 清宏 教授
木戸出 正繼 教授
中村 哲 博士
猿渡 洋 助教授

提出日： 2004 年 2 月 6 日

音声対話インタフェースのための
雑音に頑健な音声入力の研究*

伊田 政樹

内容梗概

次世代のマンマシンインタフェースはどうあるべきか．簡単で誰もが使えること．高精度で誤りが少ないこと．高速に情報の入出力が行えること．人にとっても機械にとっても負荷が小さいこと．などが挙げられる．多様なマンマシンインタフェースの中で，本研究では音声対話インタフェースに注目する．

近年，音声認識の研究を行う上での環境が大きく進歩し，さまざまなアプリケーションで音声認識技術が一般ユーザに利用されるようになった．しかしながら，音声認識は技術としていまだ道半ばであると言わざるを得ない．未完成と考える最大の要因は，音声認識技術の利用形態にかかる制約が大きい点にある．

実環境下における音声認識システムの実用化について考えると，ハンズフリー音声入力インタフェースの実現と，それに伴う雑音混入に対してロバストな音声認識性能の実現が大きな鍵を握っている．本研究では，ハンズフリー音声入力インタフェースの実現のためマイクロホンアレーによる指向性マイクロホンとスペクトル減算による雑音除去を併用する．さらに音響モデルの雑音環境適応化により雑音混入による音声認識性能の低下を防ぐ．ここで，以下の三点に注目する．第一に，音響モデルを雑音環境適応化する際に必要な適応データ量の削減．第二に，時々刻々変動する雑音環境に対してロバストな音響モデルの構築．第三に，実際の音声対話アプリケーションにおいて利用可能な音声入力インタフェースの構築．以上の三点である．

* 奈良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 博士論文，NAIST-IS-DT0061004, 2004年2月6日.

第一の課題および第二の課題に対し雑音 GMM 適応化と SN 比別マルチパスモデルを用いた HMM 合成法を提案する．HMM 合成法は環境雑音のモデル化を行う際に適応データとして雑音データのみを用いる方法である．したがって適応化にユーザ負担を生じない長所がある．HMM 合成法の欠点は多量の適応データ量を必要とする点と雑音の変動に対して対応できない問題がある．提案法では，さまざまな雑音データを用いて準備した初期雑音 GMM の適応化することで適応データ量を削減できる．また，複数の SN 比に対応した適応化 HMM を作成し 1 つのモデルの中に並列に構築する．評価実験の結果，従来法に比べて適応データ量を 10 分の 1 に削減することができた．

第三の課題に対し，情報提供端末向けにハンズフリー音声入力インタフェースを開発した．音声対話が有効なアプリケーションとして券売機や情報提供端末を取り上げ，遅延和形マイクロホンアレ―処理とスペクトル減算による雑音除去を組み合わせた音声入力インタフェースを試作した．実環境下における認識性能評価実験の結果，216 単語の孤立単語認識で 91.6%の認識性能を達成し，十分な認識性能が得られることが確認できた．

以上のように，本研究では実環境下での音声認識システムの利用における課題について検討し，成果を得た．

キーワード

音声認識，HMM 合成法，マルチパスモデル，マイクロホンアレ―，スペクトル減算

Robust Speech Recognition in Real Environments for Spoken Dialog Systems*

Masaki Ida

Abstract

What are the key requirements for the next-generation human-machine interface? Some of the requirements may be ease of use, high accuracy, high speed, and low cost. While there are various approaches to create a human-machine interface, I focus on a spoken dialog interface in this study.

In recent years, a great progress has been made in speech recognition technology. Speech recognition technology has been used in various applications for general users. However, the current state of the technology is still immature, because there are a large number of constraints over the use of the technology.

For utilization of a speech recognition system under real environments, two major problems must be solved: One is the realization of a hands free speech input interface, and the other is robust recognition performance against noisy speech input. In this thesis, a microphone array serves as a directive microphone. Combined with spectral subtraction for noise reduction, a hands-free interface becomes feasible. The noise problem is addressed by adapting acoustic models to the emvironmental conditions.

I propose a new acoustic model adaptation method with a priori noise GMM (Gaussian Mixture Model) and multi-SNR (Signal to Noise Ratio) models for

* Doctor's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT0061004, February 6, 2004.

HMM (Hidden Markov Model) composition-based model adaptation. The standard HMM composition method requires only noise data for adaptation, not burdening users with adapting process. However, there are a couple of issues with this method. First, it requires much data for adaptation. Second, it lacks robustness against changes in the noise environment. The proposed method demonstrated a reduction of the amount of noise data needed for adaptation by adapting the initial noise GMM with a variety of noise data in the database. The method also employed a multiple HMMs for several SNRs from which the best model is selected based on the acoustic likelihood to deal with unknown SNRs. Experimental results show that the amount of adaptation data is reduced to 10% compared with the conventional HMM composition method.

I construct a hands-free speech input interface for information kiosk terminals using the combination of a microphone array and spectral subtraction method. I implemented the interface into ticket vending machines and information kiosk terminals to illustrate the effectiveness of the interface. The experimental results in the real environments show 91.6% word recognition rate in the 216 word recognition task, confirming the applicability.

Keywords:

speech recognition, HMM composition, multipath model, microphone array, spectral subtraction

目 次

図 目 次

表 目 次

第1章

序論

1.1. 本研究の背景

近・現代と続いた工業社会は今その最終段階である情報化社会を迎えた．私たちの周囲の生活空間はさまざまな情報機器に囲まれている．もはやこれらの情報機器に接することなく日々生活することは不可能であるといっても過言ではない．これまでの情報機器の発展の歴史を振り返ると，高速化，小型化，多機能化という3つのキーワードでまとめることができるだろう．最も端的に現れたものが携帯電話である．回線の高速化で多くの情報を伝送できるようになったことで音質が向上した．端末の小型化によりまさに「携帯」するにふさわしい端末を実現した．多機能化により「電話」の領域を超えた魅力を創出した．情報機器の次の発展のために必要なキーワードは何か，について考える必要がある．

来たるべき社会のニーズを先取りし，予測する手法の一つに SINIC 理論がある．SINIC とは Seeds-Innovation Need-Impetus Cyclic evolution の略で，科学と技術と社会の間には円環的な関係があり，次の 2 つの方向から互いに刺激を与えあっているとする．ひとつの方向は，新しい科学が新しい技術を生み，それが社会へのインパクトとなって社会の変貌を促すというもの．もうひとつの方向は，逆に社会のニーズが新しい技術の開発を促し，それが新しい科学への期待となるというもの．この 2 つの方向が相関関係により，お互いが原因となり結果となって社会が発展していくという理論である [1]．SINIC 理論によれば,「物質文明を繁栄させた工業社会は，情報化で『人間らしく』という欲求をもたらし，物質文明と精神文明の不均衡が生じることから，その是正に向けて両者が融合していく」

とされ，情報化社会の次に来るべき社会を最適化社会と名付けられている．この情報化と人間らしさの関係を，本研究の対象である情報機器にあてはめて考えると以下のようになる．情報機器において情報化とは多機能化である．一方，人間らしさとはその機器の使いやすさである．一般に単機能の製品は操作が簡単で使いやすいと考えられている．多機能で使いやすい情報機器が実現できないか，という点が次世代の情報機器に求められる技術になろう．あらゆる分野に情報機器が存在する社会を考えたときに，情報機器の使いやすさを決定づけるマンマシンインタフェース技術に求められる役割が非常に大きくなる．なぜなら，判断・意思決定・創造といった処理には人間の介在が不可欠であり，人間と機械の最適なバランスが必要だからである．

最適化社会におけるマンマシンインタフェースはどうあるべきか．簡単で誰もが使えること．高精度で誤りが少ないこと，高速に情報の入出力が行えること．人にとっても機械にとっても負荷が小さいこと．などが挙げられる [2, 3]．多様なマンマシンインタフェースの中で，本研究では音声認識技術を中心とした音声対話インタフェースに注目する．音声入出力は，人間が生まれながらにして持っているコミュニケーションツールの一つである．したがって，年齢，経験を問わず，特別な訓練も特別な道具も必要としない．この利点が見込めることから，音声対話インタフェースの実用化に期待が寄せられている．

近年，音声認識の研究を行う上での環境が大きく進歩を遂げた．第一に HMM (Hidden Markov Model) による音声認識アルゴリズムが確立したこと．第二に計算機性能が大幅に向上したこと．第三に実用に近い音声データベースが整備されたこと．以上の三点により音声認識技術は飛躍的に進歩した．

まず，音声認識アルゴリズムに関しては，1980 年代後半から主流となった HMM による確率・統計理論に基づいた方法が確立された [4]．これにより，音声認識技術は確固たるものとなった．アルゴリズムの確立による音声認識精度向上とともに，HTK [5]，Julius [6] といった研究用ツールが公開されたことで研究基盤が整備された．

データベースの面に関しては，1997 年に新聞記事読上げ連続音声コーパスが整備された [7]．約 60 時間に相当する音声データベースは 306 人の新聞記事読み上

げ音声で構成されている．このコーパスに基づいた音声認識システムと音響モデルは日本語の音声認識におけるベースラインシステムとして研究の礎を築いた．HMM による音声認識を用いる際には，データベースの整備は統計モデルの信頼度向上に直結する．現在，より自由度の高いタスクとして講演音声などを収録した日本語話し言葉コーパス (CSJ) の収集・整備が進んでいる [8, 9]．実環境下における音声認識を目的としたデータベース整備に着目すると，名古屋大学 CIAIR では，自動車運転中に機器を操作する状況で音声データを収集し，実環境での利用を想定した研究が進められている [10, 11]．また，雑音環境下における音声認識性能の国際的な比較評価タスクとして，DARPA による SPINE2 プロジェクト [12, 13] や ETSI STQ-AURORA DSR Working Group による AURORA2 プロジェクト [14, 15, 16] などが標準タスク，データベースとして利用されている．

計算機性能の向上は，もう一方で音声認識の実時間処理を可能にした．IBM の ViaVoice[17] や Dragon Speech[18] に代表されるパソコン向けの音声認識パッケージソフトウェアが市販されている．これらのソフトウェアはボイスコマンド機能やディクテーション機能によりマウスやキーボードを用いることなくコンピュータに情報を入力する手段を実現している．音声認識技術が実用化されている一例を挙げると，電話による受付・問い合わせ・予約などのコールセンタ業務も音声認識による電話音声自動応答システムが導入されている [19]．受付・問い合わせ・予約などの顧客と企業を結び付ける手段には音声による通話のほかに DTMF(プッシュトーン信号)，ファクシミリ，電子メール，WWW など，さまざまな手段が用意されている．しかしながら，エンドユーザの立場で見た場合，誰にでも使える，誰もが一度は使ったことがあるという簡単さと，即座に応答が返される安心感の 2 つの点で音声通話に大きな優位性が存在する．これに対し，サービスを提供する側の立場では，音声通話によるサービスは人件費の面で高コストであり，24 時間 365 日同等のサービスを維持・提供することは非常に難しい．これら両者のニーズが一致したことが，いち早く音声認識が応用される土壌となった．計算機性能の向上は，もう一方で音声認識システムを小型に，安価に提供することも可能にした．これにより，カーナビゲーションシステム，家庭用ゲーム機，玩具，家電などの多様な製品に音声認識技術が応用され，一般ユーザに利用されている．

しかしながら，音声認識は技術としていまだ道半ばであると言わざるを得ない．未完成であると考える最大の要因は，音声認識技術の利用形態にかかる制約が大きい点にある．たとえば，パソコンで利用するソフトウェアや，家電・カーナビなどに組み込んで利用する状況においては，そのシステムを利用する場所やユーザが限られていたり，利用できるサービスが限定されたものとなっている．また，電話自動応答のアプリケーションにおいても，入力音声が電話回線を経由した品質の音声に限定しているからこそ成立しているサービスであると言える．音声認識技術の実用化を考えた場合，応用に上記のような制約を加えることで人と機械の間のインタフェースとして成立しているのが現状であり，この制約がインタフェースとして幅広い普及を阻んでいる．

市場ニーズ・社会ニーズの両面から実現を望まれながら，未だに音声入力インタフェースが広く普及していない分野の一例として社会システム機器について考える．社会システム機器とは，情報提供端末機やコンビニエンスストアに置かれるようなオンラインショッピング端末，鉄道の券売機や空港の自動チェックイン端末，金融機関の ATM，役所などの公共機関におかれる行政端末などである．ブロードバンドと分類される高速通信網の普及により，社会システムは各々のシステムで閉じた系におけるマンマシンインタフェースという位置づけではなくなった．社会システム機器は，その機器の後方につながれた通信回線を介して存在する膨大な情報に対するマンマシンインタフェースであることが必要となった．扱う情報量の増加に伴って機器操作は複雑さを増す．したがって，従来から主に用いられているタッチパネルやボタンでは操作が階層的になることを避けられず，目的の情報に至るまでのユーザ負担が大きくなる．このような実現方法では，ユーザにとってはサービスの低下であると捉えられてしまう．上記の複雑さの解消法として音声入力により目的の情報を直接選択できる方法の実現が望まれている．一方，社会システム機器には，音声入力の実現に対して社会的なニーズも存在する．なぜなら，これらの機器は誰もが同じように使えなければならないからである．近年，ユニバーサルデザインに関する意識があらゆるものに対して求められている [20]．特に公共機関においては，高齢者や身体障害者などの社会的弱者に対しても同等のサービスの提供を可能にすることが必須条件となってきている．

その実現手段の一つとして音声入力に注目が集まっている．

以上のような強力なニーズが存在していながら社会システム機器に音声認識が組み込まれていない理由の一つに，ユーザやシステム提供者が求める認識性能を得られていない点があげられる．すでに実現されている音声認識アプリケーションでは何らかの制約を加えることで所望の認識性能を達成しているのに対し，社会システム機器において同様の制約を加えることは難しく，所望の認識性能を達成できていない．この音声認識性能が低下する最大の原因は雑音混入の影響によるものであろう．社会システム機器における音声入力インタフェースとして，ヘッドセットマイクや電話の受話器などのように話者の身につけたり手で持ったりするものを使うことはできない．なぜなら，手で持つ，あるいは装着するという動作を伴うならば従来のインタフェースであるボタンやタッチパネルと比較した際にメリットが明確にならないからである．そこで，機器筐体に直接設置されたマイク素子で音声を収録することになる．機器筐体に設置したマイクで音声を収録した場合，話者の口 (音源) とマイク素子の距離が遠いために入力に混入する周囲の雑音の割合が大きくなり無視できなくなる．この雑音混入により認識性能の低下を招く．したがって，社会システム機器に音声認識を組み込むことを考えたとき，ハンズフリー音声入力インタフェースの実現とそれに伴う雑音混入に対してロバストな音声認識性能の実現が必須条件となる．

雑音混入に対してロバストな音声認識の実現方法として，以下の (1)，(2)，(3) の 3 つの段階でそれぞれに雑音対策を施すことができる．

(1) 音声入力インタフェースにおいては，音源から離れた位置のマイクロホンを用いて高い SN 比 (Signal to Noise Ratio) の音声収録を行うことを考える．高い SN 比の音声収録のためには，目的音声だけを収録し，背後の雑音を抑えることが必要で，そのために鋭い指向性のマイクロホンが必要になる．鋭い指向性を備えたマイクロホンの実現の方法として単一のマイクロホン素子で実現する方法と複数のマイクロホンの入力に信号処理を施すことで鋭い指向性を得る方法の 2 つがある．前者の方法は一般に素子そのものが大型であったり，マイクロホン素子のの実装に物理的制約があるなど，実用上の条件が厳しい．後者の方法はマイクロホンアレーと呼ばれ，さまざまな方法が提案され，研究が行われてきた [21, 22].

マイクロホンアレーの過去の研究事例として，遅延和形アレー [23, 24, 25]，遅延和形アレーの拡張として反射音を利用したマルチビームフォーミング [26, 27] がある．適応形アレーとしては Griffith-Jim 形の適応形アレー [28]，AMNOR 方式の適応形アレー [29] のほか，さまざまなフィルタ適応の手法が研究されている [30, 31, 32, 33]．ほかに，音声信号用マイクロホンと雑音受音用マイクロホンを用意して音声信号用マイクロホンに混入する雑音を推定し差し引く 2ch スペクトルサブトラクション法 [34, 35] などがある．

新情報処理開発機構が中心となって収集した先述の実環境音声・音響特性データベース [36] や名古屋大学統合音響情報研究拠点 (CIAIR)[10, 11]，AURORA データベース (後述)[14, 15, 16] など研究用データベースの整備にともない，近年急速に研究が進んでいる．

(2) 雑音除去においては，代表的なものとしてスペクトル減算 (Spectral Subtraction)[37] と CMN (ケプストラム平均正規化：Cepstral Mean Normarization)[38, 39] があり，これらを応用した手法について種々提案されている．たとえば庄境らは音声/非音声判定誤りの影響を排除した Continuous-SS と乗法性歪み補正係数を詳細に求めることで補正性能を向上させた E-CMN [40] を提案している．北岡らは時間軸平滑化を行った SS[41] を提案している．Chen らはより長い分析窓を用いることで雑音の推定精度を向上させている [42]．藤本らはカルマンフィルタを用いた雑音除去法 [43] を提案している．また，雑音により欠落した情報を回復する方法として Missing Feature Theory に基づいた方法 [44, 45] や欠落した情報を別の特徴量により補う Multi-Stream HMM や Multi-Band HMM[46] についても検討が進められている．

(3) 音響モデルとの照合に関しては，音響モデルを雑音が混入する環境に適応化する技術の研究が種々行われている．音響モデルの適応化には従来から MAP[47, 48]，が用いられてきた．MAP 推定法は初期モデルに対して，観測された音声データの特徴量系列とその音声の音素列情報を用いて事後確率を最大にするようにモデルのパラメータを学習する．これに対し MLLR 法はモデルパラメータのアフィン変換により観測信号の空間に対応づける方法である．一般に適応データ量が多い場合に MAP 推定法が用いられ，オンライン適応には MLLR 法が用いられる．

両者を組み合わせて用いる方法 [49] も提案されている．近年になって，ヤコビ適応を応用した方法 [50] や十分統計量に基づいた方法 [51] の研究が進められている．また，音響モデルと照合する際に雑音混入に頑健な特徴量を用いる方法も検討されている．たとえば，Δ 特徴量 [52] や動的ケプストラム [53]，セグメント統計量を用いる方法 [54, 55]，RASTA-PLP 分析によるパラメータ [56] などがある．

1.2. 本研究の目的

前節で述べた通り，実環境下で音声認識システムを実用化することを考えたとき，ハンズフリー音声入力インタフェースの実現と，それに伴う雑音混入に対してロバストな音声認識性能の実現が必須条件である．本研究では上記の実現のため，環境雑音混入に対する音響モデルの適応化と実システムで利用可能なハンズフリー音声入力インタフェースの構築に着目する．

音響モデル適応化における課題は 2 つある．第一の課題は，音響モデルを適応化する際に必要な適応データ量を削減することである．実システムで運用することを考えた場合，音響モデル適応化のために取得する適応データの量は少ないほうが望ましい．なぜなら，適応データ取得を含む適応化に要する時間，ユーザは待たされることになるからである．第二の課題は時々刻々変化する雑音環境に対してロバストな音響モデルを構築することである．実環境下では周囲の環境雑音は時々刻々変化する．また，発声毎にユーザが発する音量が一定でないことやユーザとマイクの距離が一定でないことから，発声毎に SN 比も変動する．これらの変動は音響モデルの適応化で対処することが困難であるため，雑音環境の変動に対してロバストな音響モデルが必要となる．

ハンズフリー音声入力インタフェースの構築においては，第三の課題として実際のアプリケーションで利用可能な音声入力インタフェースの構築が挙げられる．社会システム機器に搭載することを考えたとき，手で持つ，あるいは装着するという動作を伴わない形態でなければ音声インタフェースの利点を明確にならない．さらに，社会システム機器に搭載する際には，機器筐体の大きさや設置場所，機器に搭載された他の部品との位置関係などの物理的制約が大きく影響する．これ

らの制約に加えて，設置場所の音響特性や利用時の状況を考慮して音声入力インタフェースの設計を行わねばならない．

第一の課題に対し，本研究では HMM 合成法に着目した．音響モデルの適応化手法は大きく二つの方法に分類できる．一方は適応データとして雑音の混入した音声データを用いる方法で，MLLR や MAP 推定法，あるいは十分統計量を利用した環境適応化手法などがある．もう一方は適応データに雑音データのみを用いる方法で，HMM 合成法やヤコビ適応法がある．前者は雑音の混入した音声データを取得する際にユーザに負担を生じる欠点があるのに対し，後者ではその負担がない．後者の方法について考えると，ヤコビ適応法は雑音環境の変動を線形近似で表現することで短時間での環境適応を可能にしている．しかし環境雑音が大きく変動した場合，近似誤差の影響を無視できなくなる問題点を持っている．これに対し，HMM 合成法は事前に環境雑音のモデル化を実雑音データに基づいて行う．したがって，より実雑音に対して忠実なモデル化が可能である．また，雑音データのみを用いるのでユーザ負担も小さいという利点を持っている．しかしながら，環境雑音のモデル化に統計的手法を用いているために多量の実雑音データを必要とする欠点がある．この欠点は雑音環境が時々刻々変化するような環境の場合，変化するたびに環境雑音のモデル化を行う必要があるため，非常に困難である．そこで，初期雑音 GMM を少量の実雑音データで現場の雑音環境に適応化することで，従来法の HMM 合成法に比べて適応化に用いる雑音データ量と計算量を削減する．

第二の課題に対して，雑音 GMM の適応化と SN 比別マルチパスモデルを用いた HMM 合成法による課題解決を試みる．実環境下においては，混入する雑音は時々刻々変化する．したがって雑音の変動に対してロバストな音響モデルが必要になる．ここで変動する雑音の問題を，雑音の種類が変動する問題と入力音声の SN 比が変動する問題に切り分けて考える．前者に関しては，先述の雑音 GMM の適応化を用いた HMM 合成法により解決できる．初期雑音 GMM を作成する際に多様な雑音環境に関する情報を与えることができる．後者に関しては，複数の SN 比に対応した適応化 HMM を作成し，1 つのモデルの中に並列に構築する手法を用いる．

第三の課題に対して，情報提供端末向けのハンズフリー音声入力インタフェースとして遅延和形マイクロホンアレーとスペクトル減算を組み合わせた構成を提案する．音声対話による操作が適当なシステムのひとつとして情報提供端末やコンビニエンスストアに置かれるようなオンラインショッピング端末，鉄道の券売機などの社会システム機器が挙げられる．これらの機器を介して提供されるサービスが多岐に拡がっていく傾向にあり，機器の高機能化とタスクの複雑化が年々進んでいる．それゆえ，複雑なタスクを簡単に操作することができる入力インタフェースとして音声認識を応用したインタフェースの導入を考える．情報提供端末機の大きさなどの物理的制約と設置環境の音響特性を考慮し，遅延和形マイクロホンアレーによる指向性受音とスペクトル減算による雑音除去を組み合わせた音声入力インタフェースを構成する．また，この音声入力インタフェースの有効性の評価を行うに際し，実騒音環境下において収録した評価データを用いた．

1.3. 本論文の構成

本論文は 5 章からなる．

2 章では，音声認識システムの概要について，HMM による音声認識の定式化について簡単に述べる．

3 章では，雑音環境下における音響モデル適応化の方法として HMM 合成法を取り上げる．雑音 GMM の適応化を用いた HMM 合成法を提案し，SN 比別マルチパスモデルを併用することで時々刻々変動する雑音に対してロバストな音響モデルを構築する．性能評価実験を日本語旅行対話タスクと英語連続数字タスクで行った．

4 章では，筆者らの開発した情報提供端末機向けハンズフリー音声入力インタフェースを紹介する．音声対話が有効なアプリケーションとして情報提供端末における音声対話を取り上げる．情報提供端末に組み込むことが可能な音声入力インタフェースの構成について述べ，次に構成要素である遅延和形マイクロホンアレー処理とスペクトル減算による雑音除去について述べる．性能評価実験は孤立単語音声認識実験により実施した．

最後に 5 章において結論と今後の課題についてまとめる.

第2章

実環境下における音声認識

2.1. はじめに

本章では実環境下において音声認識システムを構成する各要素について述べる．図 2.1 に概要を示す．まず，現在の音声認識技術において主流となっている HMM を用いた音声認識アルゴリズムについて述べる次に，実環境下で音声認識システムを利用するための雑音対策として本研究で検討を行った 3 つの方法について，その基本的な理論について述べる．第一に HMM 合成法による音響モデルの雑音環境適応化について．第二にマイクロホンアレー，特に遅延和アレー処理を用いた音声信号強調手法について．第三にスペクトル減算による雑音除去について述べる．

2.2. HMMによる音声のモデル化と音声認識への応用

音声認識システムを構築する際の問題点は，あいまい性を含んだ音声パターンを取り扱わねばならない点にある．あいまい性の要因はさまざまなものが挙げられる．音響的なあいまい性として，雑音の混入や音声伝達経路の音響特性の相違などの外的要因によるもの，口腔，のどなどの発声器官の個人差に基づく相違によるもの，発声前後の音響的文脈 (調音結合) によるもの，発声毎の発声速度や話し方の差異によるものなどがある．音声をモデル化する際には，以上のあいまい性を含むモデル化を行わねばならない．音声のあいまい性を考慮したモデル化の方法として，DP マッチング，マルチテンプレート法，HMM やニューラルネット，

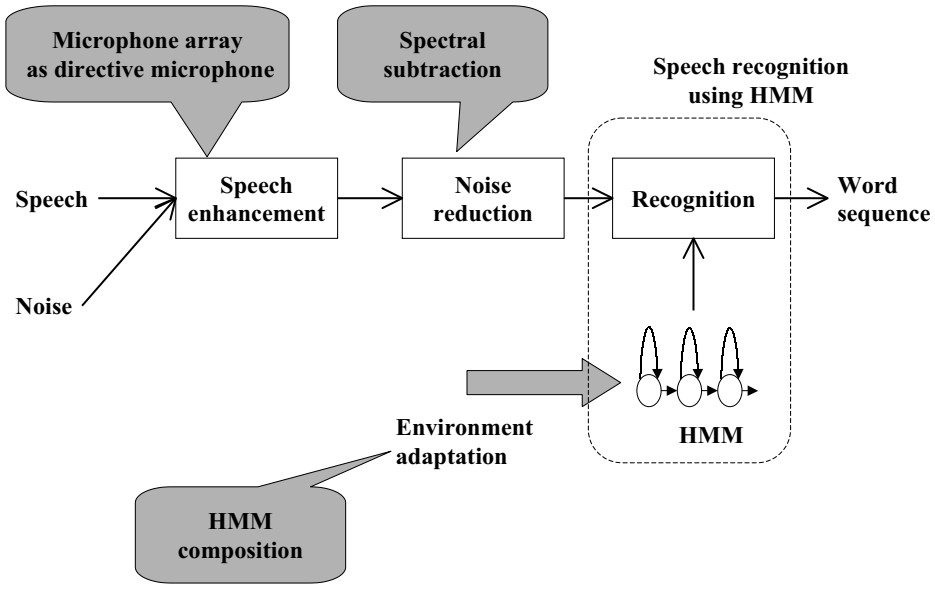

図 2.1 実環境下における音声認識

Baysian Network による確率統計的処理による手法などが開発されてきた．本節では，HMM による音声のモデル化とその学習アルゴリズム，認識アルゴリズムについて述べる．

2.2.1 HMM による音声のモデル化

HMM(隠れマルコフモデル: Hidden Markov Model) は時系列信号の確率モデルであり，複数の定常な状態の間を遷移することで非定常な時系列信号をモデル化する．この HMM を音声の特徴量時系列に適用する [57].

一般に音声認識に用いられる HMM は left-to-right モデルと呼ばれ，1 回の状態遷移に対してシンボルを 1 つずつ出力する．出力シンボルによって状態遷移先を一意に決定できないという意味において非決定性有限状態オートマトンとして

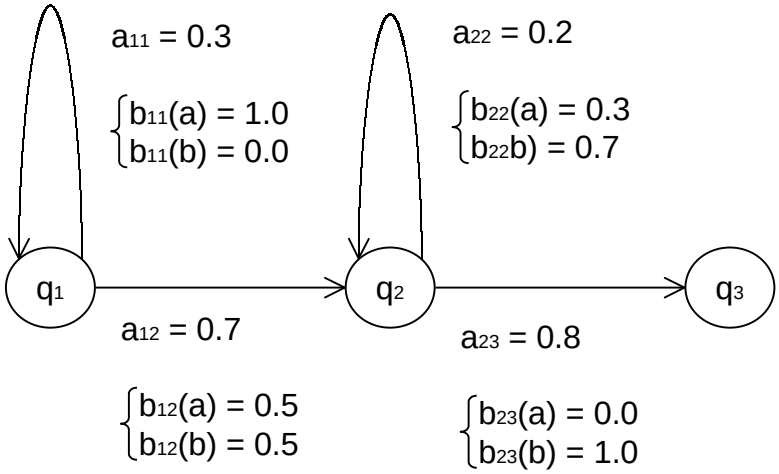

図 2.2 left-to-right 型 HMM の一例

定義される．次の遷移先やその遷移の際のシンボルの出力は遷移確率・出力確率として統計的に与えられる．一般的なマルコフモデルとの相違点は出力シンボル系列が与えられても状態遷移系列は唯一に決定できず，シンボル系列のみを観測できる点にある．この特徴から「hidden(隠れ)」マルコフモデルと呼ばれる．出力シンボルが複数存在する場合，各々のシンボルの出力確率の集合を 1 つの出力確率分布としてとらえることができる．この分布が離散分布である場合 (ベクトル量子化を行う場合) を離散出力確率分布型 HMM といい，連続分布の場合 (ベクトル量子化を行わない場合) を連続出力確率分布型 HMM という．

離散出力確率分布型 HMM の簡単な例を図 2.2 に示す．図中の a_{ij} は状態 q_i から q_j へ遷移する確率，$b_{ij}(k)$ はその遷移でシンボル k を出力する出力確率である．これらは次式の拘束条件を満たす．

$$\sum_j a_{ij} = 1, \qquad \sum_k b_{ij}(k) = 1 \tag{2.1}$$

図の例において観測シンボル系列が aab であったならば，可能な状態遷移系列は $q_1q_1q_2q_3$ と $q_1q_2q_2q_3$ の 2 つが存在する．ここで、$\boldsymbol{o} = o_1o_2 \ldots o_T$ と $\boldsymbol{x} = x_1x_2 \ldots x_T$ をそれぞれ出力と状態の系列とすると，単純マルコフ過程と非決定性の仮定から

$$P(x) = \prod_i P(x_i|\boldsymbol{x}_1^{i-1}) = \prod_i P(x_i|x_{i-1}) \tag{2.2}$$

$$P(\boldsymbol{o}|\boldsymbol{x}) = \prod_i P(o_i|\boldsymbol{x}_1^{i}) = \prod_i P(o_i|x_{i-1}, x_i) \tag{2.3}$$

により，

$$\begin{aligned} P(\boldsymbol{o}) &= \sum_{\boldsymbol{x}} P(\boldsymbol{o}|\boldsymbol{x})P(\boldsymbol{x}) \\ &= \sum_{\boldsymbol{x}} \prod_i P(x_i|x_{i-1})P(o_i|x_{i-1}, x_i) \end{aligned} \tag{2.4}$$

この例において，生起確率が 0 とならないのは前述の 2 つの状態遷移系列に限られる．これらそれぞれの生起確率は，

$$P_1(abb) = 0.3 \times 1.0 \times 0.7 \times 0.5 \times 0.8 \times 1.0 = 0.084 \tag{2.5}$$

$$P_2(abb) = 0.7 \times 0.5 \times 0.2 \times 0.7 \times 0.8 \times 1.0 = 0.0392 \tag{2.6}$$

となるから， $P(abb)$ は

$$P(abb) = P_1(abb) + P_2(abb) = 0.084 + 0.0392 = 0.1232 \tag{2.7}$$

となる．

HMM による音声生成モデルを用いて音声を認識することを考える．認識語彙 $w_1, w_2, \ldots, w_N$ に対応するモデル $\mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_N$ があり、あるシンボル系列 $\boldsymbol{o} = \{o_1, o_2, \ldots, o_t\}$ が観測されたとき，事後確率 $P(\mathrm{M}_n|\boldsymbol{o})$ が最大となる w_n を認識結果とする．この事後確率は，

$$P(\mathrm{M}_n|\boldsymbol{o}) = \frac{P(\boldsymbol{o}|\mathrm{M}_n) \cdot P(\mathrm{M}_n)}{P(\boldsymbol{o})} \tag{2.8}$$

$$P(\mathrm{M}_n|\boldsymbol{o}) \propto P(\boldsymbol{o}|\mathrm{M}_n)P(\mathrm{M}_n) \tag{2.9}$$

と書くことができる．ここで $P(\boldsymbol{o})$ は観測シンボルそのものの生起確率であるので，モデルに依存せず一定である．$P(\mathrm{M}_i)$ はモデル自身の事前確率であり，一般に言語モデルとしてモデル化される．以上により，HMM からの生起確率をモデル間比較することにより音声認識を行うことができる．

2.2.2 HMM の学習

音声認識とは各モデルから入力シンボル系列が出力される確率，すなわち各モデルに対する $p(\boldsymbol{o}|\mathrm{M}_n)$ が最大となるモデルを選択する問題である．これを効率的に計算するため，以下のような計算法を用いる．まず，前向き変数 $\alpha(i,t)$ を初期状態から始まり，$o_1, o_2, \ldots, o_t$ を生成して状態 i に達する確率と定義する．この $\alpha(i,t)$ は次式で再帰的に求められる．

$$\alpha(i,t) = \sum_j \alpha(j,t-1)a_{ji}b_{ji}(o_t)\text{ただし初期値}\alpha(i,0) = \pi_i\text{初期状態確率} \tag{2.10}$$

この結果,

$$P(\boldsymbol{o}|\mathrm{M}_n) = \sum_{q_i \in F} \alpha(i,T) \tag{2.11}$$

となる．ここで F は受理状態の集合である．このように時間方向に処理を進めていく方法を Forward アルゴリズムという．先ほどの例の場合について Forward アルゴリズムで計算した場合について図 2.3 に示す.

同様に，時間方向を逆向きに考えることで後ろ向き変数 $\beta(i,t)$ を最終状態から始まり， $o_T, \ldots, o_{t+2}, o_{t+1}$ を生成して状態 i に達する確率と定義すると

$$\beta(i,t) = \sum_j \beta(j,t+1)a_{ij}b_{ij}(o_{t+1}) \tag{2.12}$$

$$\text{ただし，初期値}\beta(i,T) = \begin{cases} 1 & q_i \in F \\ 0 & \text{otherwise} \end{cases} \tag{2.13}$$

上式を Backward アルゴリズムという．ここからは

$$P(\boldsymbol{o}|\mathrm{M}_n) = \sum_i \beta(i,0)\pi_i \tag{2.14}$$

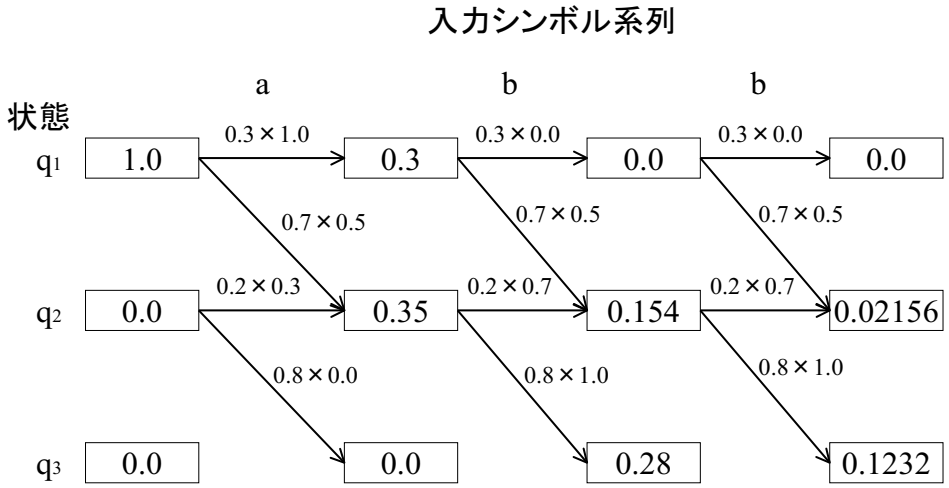

図 2.3 Forward アルゴリズム

が導出される．

この前向き確率 $\alpha(i,t)$ および後ろ向き確率 $\beta(j,t)$ を用いて HMM のパラメータ推定を行う．このアルゴリズムを Baum-Welch アルゴリズムという．

HMM のパラメータセットを $\theta = \{\pi_i, a_{ij}, b_{ij}(o)\}$ とし，この HMM がシンボル系列 $\boldsymbol{o} = \{o_1, o_2, \ldots, \o_t\}$ を出力する確率を最大化することを考える．

シンボル系列 $\boldsymbol{o}$ に対して，状態 i から状態 j への遷移が時刻 t で生じた確率を $\gamma(i,j,t)$ と表すと

$$\gamma(i,j,t) = \frac{\alpha(i,t-1)a_{ij}b_{ij}(o_t)\beta(j,t)}{\sum_{i\in F}\alpha(i,T)} \tag{2.15}$$

$$= \frac{\alpha(i,t-1)a_{ij}b_{ij}(o_t)\beta(j,t)}{\sum_{i}\alpha(i,t)\beta(i,t)} \tag{2.16}$$

となる．このとき尤度，すなわち上式を最大にすることのできる θ の推定値 $\hat{\theta}$ は

最尤推定法に基づいて次の各式で求められる．

$$\hat{\pi_i} = \frac{\sum_j \gamma(i,j,1)}{\sum_{i,j} \gamma(i,j,1)} \tag{2.17}$$

$$\begin{aligned} \hat{a_{ij}} &= \frac{\sum_t \gamma(i,j,t)}{\sum_j \sum_t \gamma(i,j,t)} \\ &= \frac{\sum_t \alpha(i,t-1) a_{ij} b_{ij}(o_t) \beta(j,t)}{\sum_t \alpha(i,t)\beta(i,t)} \end{aligned} \tag{2.18}$$

$$\begin{aligned} \hat{b_{ij}(k)} &= \frac{\sum_{t:o_t=k} \gamma(i,j,t)}{\sum_t \gamma(i,j,t)} \\ &= \frac{\sum_{t:o_t=k} \alpha(i,t-1) a_{ij} b_{ij}(o_t) \beta(j,t)}{\sum_t \alpha(i,t-1) a_{ij} b_{ij}(o_t) \beta(j,t)} \end{aligned} \tag{2.19}$$

2.2.3 Viterbi アルゴリズム

HMM による音声認識を行う場合，入力シンボル系列に対してどのような状態系列が最適かを知る必要があることがある．これは Forward アルゴリズムで求めることはできない．そこで，式 (2.4) の

$$P(\boldsymbol{o}) = \sum_{\boldsymbol{x}} \prod_i P(x_i|x_{i-1}) P(o_i|x_{i-1}, x_i)$$

の代わりに

$$P''(\boldsymbol{o}) = \max_{\boldsymbol{x}} \{ \prod_i P(x_i|x_{i-1}) P(o_i|x_{i-1}, x_i) \} \tag{2.20}$$

を用いて近似することにより，入力シンボル系列は唯一の状態遷移系列に対応付けられる．このアルゴリズムは Viterbi によって提案されたことから Viterbi アルゴリズムと呼ばれる．各時刻，状態において最大値をとるのはどの状態からの遷移であるかを保存しておくことで，状態系列を逆順で求めることができる．この処理をバックトラックという．また，こうして求められた状態系列を最適パスという．

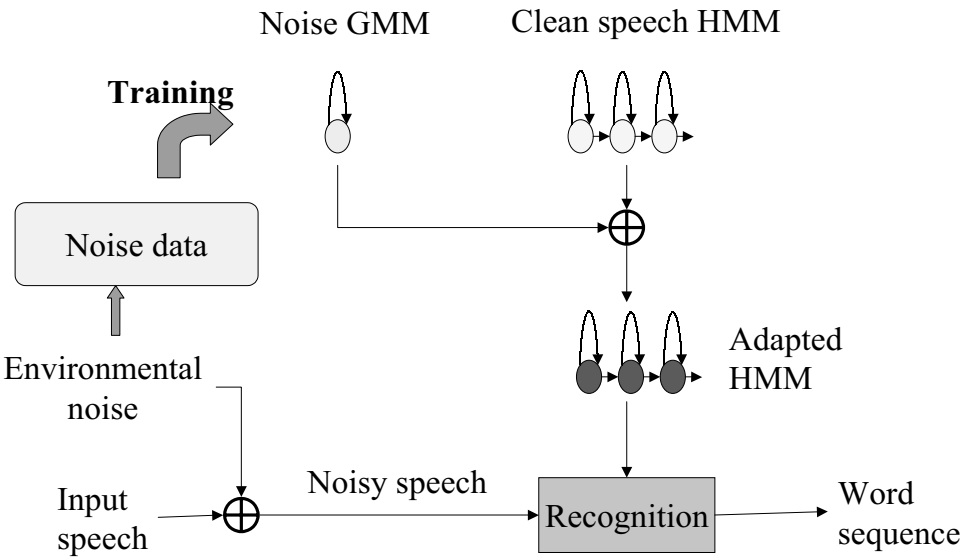

図 2.4 HMM 合成

2.3. HMM 合成法による雑音環境適応化

実環境下で音声認識システムを使用した場合，雑音が混入することで入力音声が歪むことは避けられない．入力音声の歪みは音声認識性能の低下をもたらす．そこで，入力音声の歪みに対してロバストな音響モデルの生成法が必要になる．本節では，音響モデルを雑音が混入する環境に適応化する手法の一つである HMM 合成法について解説する．

HMM 合成法 [58, 59] は，事前に雑音を含まない音声データを用いて学習を行った音素の音響モデルと，環境雑音のモデルとを合成することで，モデル化された環境雑音に適応した音響モデルを作成する方法である．HMM 合成による雑音環境下の音声認識を図 2.4 に示す．ここでは加法性の雑音のみを仮定する．観測される入力音声のパワースペクトルを Y とし，これを環境雑音のパワースペクトル

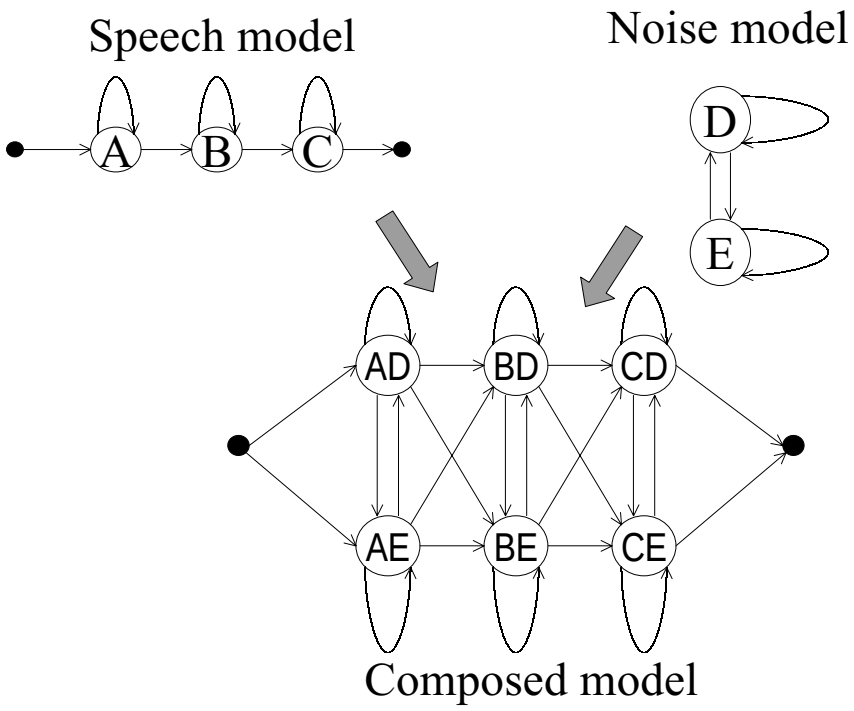

図 2.5 合成 HMM の構造

N と雑音のない音声のパワースペクトル S で表す．環境雑音の加法性は線形スペクトル領域において成立し．

$$Y_{linspc} = S_{linspc} + N_{linspc} \tag{2.21}$$

一方，音響モデルは一般的にケプストラムにより特徴抽出されているので，

$$Y_{cep} = \Gamma^{-1} \log[\exp\{\Gamma(S_{cep})\} + k_{SNR} \exp\{\Gamma(N_{cep})\}] \tag{2.22}$$

となる．Γ, Γ^{-1} はコサイン変換およびコサイン逆変換，k_{SNR} は SN 比に応じて決定する係数である．式 (2.22) を HMM に適応した場合，合成 HMM の構造は図 2.5 に示すように各 HMM の直積で表される．遷移確率は対応する遷移確率の積で求められ，出力確率分布は各状態において結合される．出力確率分布の合成手順を図 2.6 に示す．これにしたがって，以下のように適応化 HMM を得る．

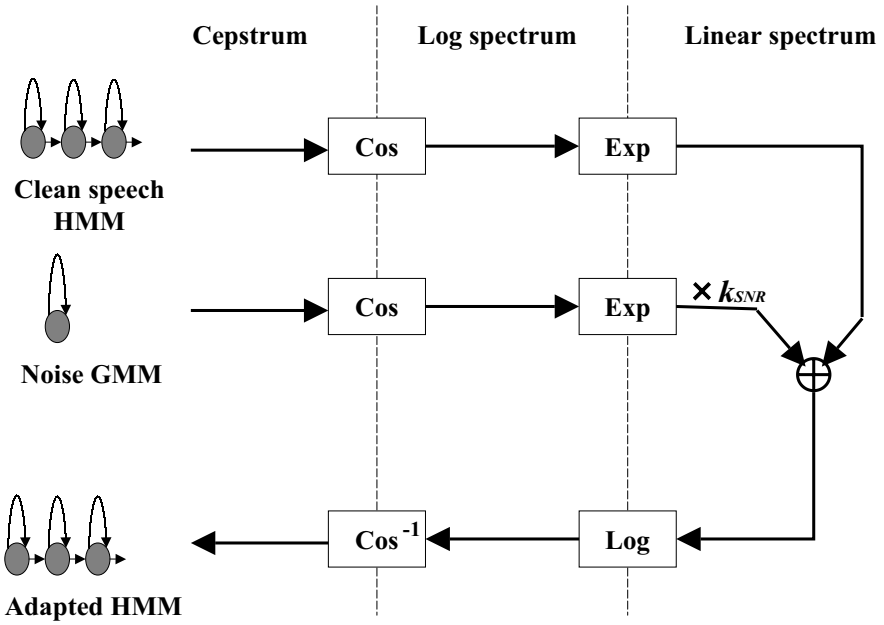

図 2.6 出力確率の合成アルゴリズム

1. ケプストラムを用いて音声と雑音のモデルのパラメータを推定する．
2. 各 HMM の平均と分散を以下の式によりコサイン変換する．μ_{cep}, Σ_{cep} はケプストラム領域における平均値と共分散行列，μ_{log}, Σ_{log} は対数パワースペクトル領域における平均値と共分散行列である．

$$\mu_{log} = \Gamma \mu_{cep} \tag{2.23}$$

$$\Sigma_{log} = \Gamma \Sigma_{cep} \Gamma^T \tag{2.24}$$

3. 指数変換を用いて線形スペクトル領域に変換する．正規分布に従う確率変数を指数変換すると対数正規分布に従う．

$$\mu_{(lin),i} = \exp\{\mu_{(log),i} + \frac{\sigma^2_{(log),ii}}{2}\} \tag{2.25}$$

$$\sigma^2_{(lin),ij} = \mu_{(log),i}\mu_{(log),j} \exp(\sigma^2_{(log),ij} - 1) \tag{2.26}$$

4. 各確率変数が独立であると仮定し，式 (2.21) に従って音声と雑音の分布を合成する．$\mu_{(lin\_S)}, \Sigma_{(lin\_S)}, \mu_{(lin\_N)}, \Sigma_{(lin\_N)}$ をそれぞれ線形スペクトル領域における音声の分布の平均および共分散行列，雑音の分布の平均および共分散行列とし，$\mu_{(lin\_Y)}, \Sigma_{(lin\_Y)}$ を合成した分布の平均および共分散行列とすれば以下の式が導かれる．

$$\mu_{(lin\_Y)} = \mu_{(lin\_S)} + k_{SNR} \cdot \mu_{(lin\_N)} \tag{2.27}$$

$$\Sigma_{(lin\_Y)} = \Sigma_{(lin\_S)} + k^2_{SNR} \cdot \Sigma_{(lin\_N)} \tag{2.28}$$

5. 合成分布の対数変換を行う．

$$\mu_{(log\_Y),i} = \log \mu_{(lin\_Y),i} - \frac{1}{2}\{\frac{(\sigma_{(lin\_Y),ij})^2}{(\mu_{(lin\_Y),i})^2} + 1\} \tag{2.29}$$

$$\sigma^2_{(log\_Y),ij)} = \log\{\frac{\sigma^2_{(lin\_Y),ij}}{\mu_{(lin\_Y),i}\mu_{(lin\_Y),j}} + 1\} \tag{2.30}$$

6. 逆フーリエ変換によりケプストラム領域に戻す．

$$\mu_{cep\_Y} = \Gamma^{-1} \mu_{(log\_Y)} \tag{2.31}$$

$$\Sigma_{cep\_Y} = \Gamma^{-1} \Sigma_{(log\_Y)} (\Gamma^{-1})^T \tag{2.32}$$

2.4. マイクロホンアレーによる指向性受音

社会システム機器に音声認識を応用することを考えた場合，ヘッドセットマイクのような話者に装着するようなマイクロホン素子を用いることができない．そこでマイクロホンを機器本体に設置せねばならない．音源から離れた機器本体に設置したマイクロホンで高い SN 比の音声収録を行うには，鋭い指向性を備えたマイクロホンを使う必要がある．鋭い指向性を単一のマイクロホンで実現したものにガンマイクと呼ばれる超指向性マイクロホンがあるが，一般に市販の超指向性マイクロホンは大型かつ高価である．さらに，指向性を形成するためにマイクロホン周囲に空間を必要とするなど，実装上の物理的制約が非常に大きい．以上の観点から市販の超指向性マイクロホンは情報端末の筐体内部に実装することは非常に困難で不向きである．

そこで，筐体に実装することが可能な指向性マイクロホンの形態としてマイクロホンアレーを用いる．マイクロホンアレーとは複数個のマイクロホンにより受音した音声に音源位置などの音響空間情報を用いて信号処理を行うことで，特定方向から到来する信号を強調・抑圧することで全体としてマイクロホンの指向性を形成するものである．アレーを構成するマイクロホン素子に安価な小型無指向性素子を利用でき，設置条件に関する制限が少ないため筐体実装に適している．マイクロホンアレーは信号処理方式により次の二つに大別される．一方は目的とする音声の方位に指向特性の山を形成して目的方向の音声を選択的に抽出する遅延和形アレーであり，もう一方は特定の雑音の方向に指向特性の谷を形成して雑音の入力を抑圧する死角型アレーである．情報端末などでの利用を考えた場合，騒音環境は不特定多数の方向から到来する雑音であると想定できる．ゆえに大部分の雑音に指向性がない．また，目的とする音声の方位，すなわちユーザの位置を 1ヶ所に特定できるというシステムの利用形態の特徴からも，ユーザの方向に対して有効な指向特性を持つ遅延和形アレーを用いるのが最適である．遅延和形アレー処理の原理図を図 2.7 に示す．複数個のマイクロホン素子の入力音声信号に対して，音源からの到達時間差に応じて時間差がなくなるように遅延を与える．これにより特定の方向から到来した音声のみ位相が一致する．この位相が一致した信号を加算することで，特定の方向からの音声を強調することができる．その

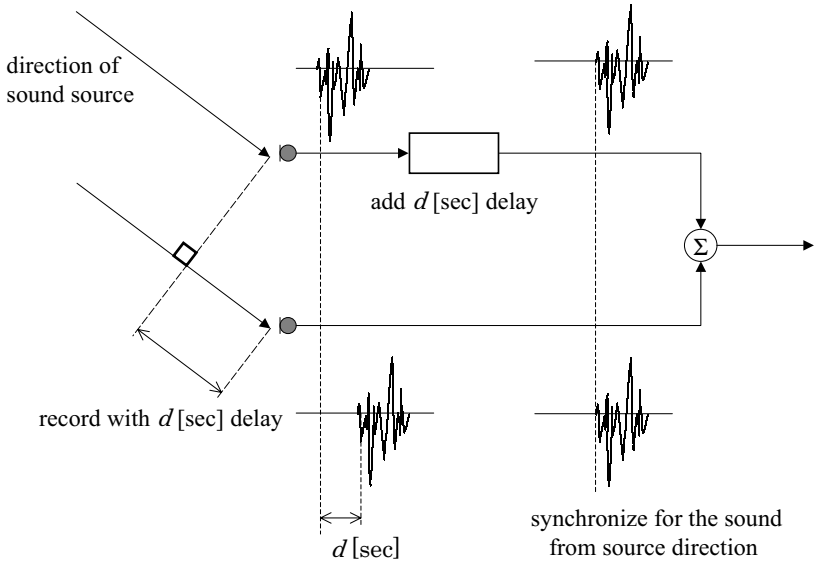

図 2.7 遅延和形アレー処理における信号の同相化

結果，特定の方向以外の音声 (雑音) は相対的に減少することになり，鋭い指向特性を形成することができる [21, 22].

遅延和形アレ―処理の原理を説明する．ここでは図 2.8 に示すような等間隔直線配列マイクロホンアレーを考える．マイクロホンの素子間隔を d ，目的音声の方向を θ_L とする．各遅延器に次式の遅延時間 D_i を受音信号に与える．ここで c は音速である．D_0 は各遅延量をディジタルフィルタで実現する際に精度低下を防ぐための固定遅延量である．

$$D_i = D_0 - (i-1)\tau_L \qquad i = 1, 2, \cdots, M \tag{2.33}$$

$$\tau_L = (d \sin \theta_L)/c \tag{2.34}$$

各マイクロホンで受音される目的音声信号 $x_i(t)$ は

$$x_i(t) = x_1(t - (i-1)\tau_L) \tag{2.35}$$

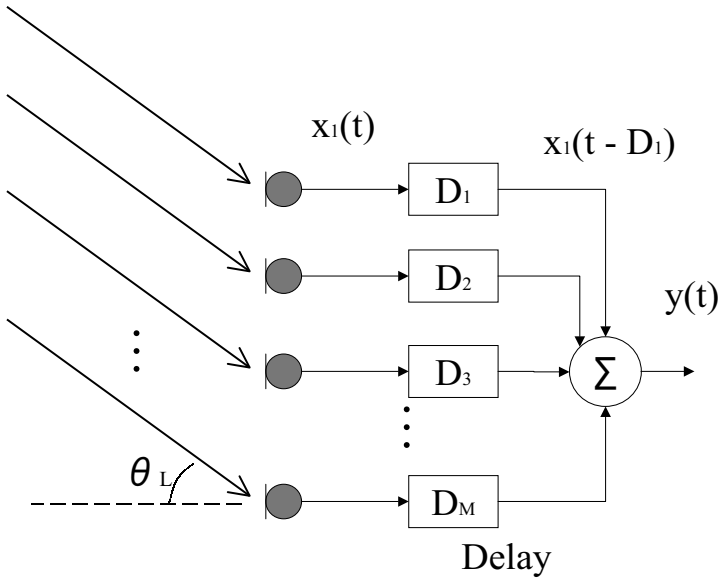

図 2.8 遅延和形アレー処理の原理図

と表される．これに式 (2.33) の遅延を与えると，

$$x_i(t - D_i) = x_1(t - (i-1)\tau_L - D_i) = x_1(t - D_0) \tag{2.36}$$

となり，マイクロホンの位置 i に依存しない信号となる．すなわち，式 (2.33) によって θ_L 方向から到来する信号の到達時間差が補正され，同相化されていることがわかる．同相化された各信号を加算することにより，目的音声信号が強調される．一方，θ_L とは異なった方向から到来する音は τ_L とは異なった時間差で各マイクロホンに到達する．この信号に対して式 (2.33) による遅延を与えたとしても信号は同相化されない．同相化されていない信号を加算しても音声信号は強調されない．以上の結果，目的音声の方向 θ_L 方向から到来する音声のみが強調される．

この方法で構成したマイクロホンアレーの特性の一般的な例として，10 個の

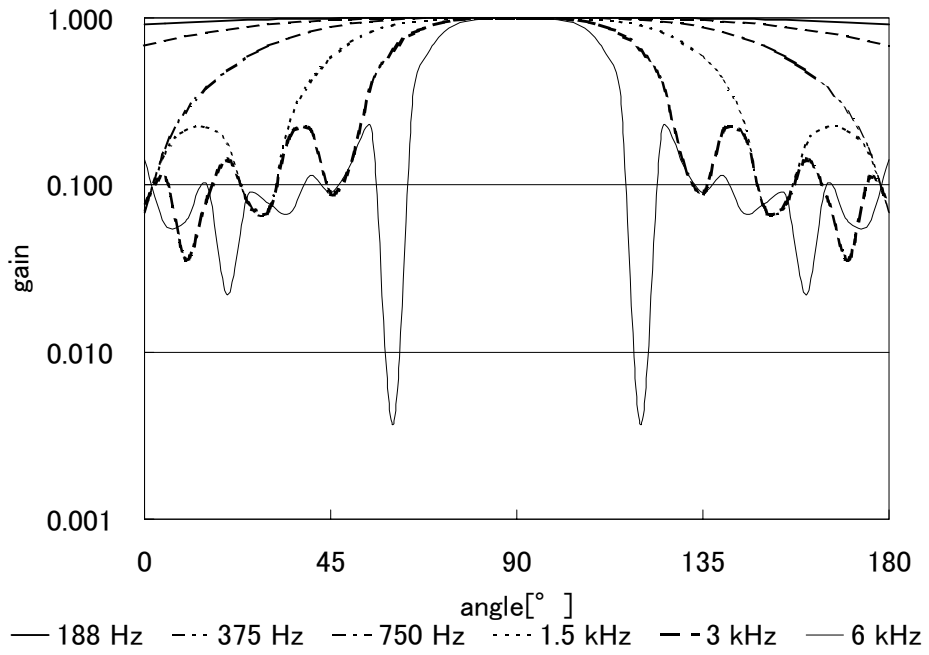

図 2.9 マイクロホンアレーの指向特性

無指向性マイクロホンを間隔 42.45 mm で直線上に配置したマイクロホンアレー (全長 382.05 mm) を用いた場合の指向特性を図 2.9 に示す．遅延和形アレー方式を用いる長所として目的方向が既知の場合，容易に方向の制御を行うことが可能である点があげられる．しかし，指向特性に周波数依存性を持ち，高い周波数ほど鋭い指向性を持つ．低周波域において鋭い指向特性を与えるにはアレーの規模が大きくなるため，情報端末機器等に用いる際に欠点となる．

2.5. スペクトル減算による雑音除去

遅延和形アレー処理により，目的音声方向に対する指向性を確保することができる．しかしながら，目的音声方向と同一方向から到来する雑音を抑制することができない．また，アレー規模に制限がある場合には低周波域において雑音抑制

の効果を期待できない。そこで，さらなる雑音抑制を行うため，スペクトル減算 (Spectral Subtraction: 以下 SS と呼ぶ) を併用する．SS は雑音が時間的に定常であるとの仮定に基づく方法で，非音声区間の信号から雑音の特徴を推定しておき，入力された音声の特徴から雑音の成分を取り除く方式である [37]．入力音声を $y(t)$ に対して，音声/非音声の判定を行う．非音声区間の線形スペクトルを平滑化したものを雑音のスペクトルとし，音声区間の線形スペクトルから雑音成分を次式によって減算して，雑音を除去する．

$$\hat{S}(\omega;t_n) = \begin{cases} Y(\omega;t_n) - \alpha\hat{N}(\omega;t_n) \\ \qquad \text{if } Y(\omega;t_n) - \alpha\hat{N}(\omega;t_n) > \beta Y(\omega;t_n) \\ \beta Y(\omega;t_n) \qquad \text{otherwise} \end{cases} \tag{2.37}$$

$$\hat{N}(\omega;t_n) = \begin{cases} \hat{N}(\omega;t_{n-1}) \\ \qquad \text{if } Y(\omega;t_n) - \alpha\hat{N}(\omega;t_n) > \beta Y(\omega;t_n) \\ \gamma\hat{N}(\omega;t_{n-1}) + (1-\gamma)Y(\omega;t_n) \\ \qquad\qquad \text{otherwise} \end{cases} \tag{2.38}$$

ここで，α は over-estimation factor，β は flooring factor，γ は smoothing factor である．

単純に雑音のスペクトルを減算しているだけであるから，0 dB に限りなく近いような劣悪な雑音環境の場合や，human-speech like noise の場合，目的音声の成分も消去することになる．このため雑音除去後の音声に歪みを生じる．また，雑音の動的な変化に対して十分に追随できないために musical noise と呼ばれる残存スペクトルが残る．

マイクロホンアレー処理と SS を併用した先行研究として [60]，[61]，[62]，[63] などがある．これらの研究においては，マイクロホンアレー処理部と SS 部が雑音成分の推定・抽出を行う部分で密接に結合しており，複雑な構成になっている．

これに対し本手法ではそれぞれ独立に動作するため簡単な構成で実現できる長所を持っており，実用向きといえる．しかしながら，複数マイクによる入力を用いた高精度な雑音成分の推定 [61] などの先行研究の長所を利用する際に問題が残る．

第3章

雑音GMMの適応化とSN比別マルチパスモデルを用いたHMM合成による雑音環境適応化

3.1. はじめに

音声認識システムを実環境下で使用することを考えたとき，入力音声が歪むことによる認識性能の低下を避けることができない．そこで入力音声の歪みに対してロバストな音響モデルの生成法が求められている．入力音声の歪みの原因は二つに分けることができる．一方はマイクロホンの特性や室内空間音響特性や伝送路の電気的音響特性などの伝送歪みであり，もう一方は周囲の環境に依存する雑音が入力音声に混入する加法性雑音である．本章では後者について取り上げる．

加法性の雑音に対する改善策として，従来より音響モデルの環境適応化が用いられてきた [47, 48, 50, 51]．しかしながら，適応データとして雑音の混入した音声が必要なためにユーザの負担が生じたり，近似を用いているために雑音環境の大きな変動に対する誤差が大きくなるといった弱点を持っている．

加法性雑音の問題点は，(1) 混入する雑音の種類が未知である問題と (2) SN 比が未知である問題の二つに分けて考えることができる．そこで，音響モデルの適応化手法として HMM 合成法に着目する．HMM 合成法は適応データに雑音データのみを用いる長所があるが，雑音のモデル化に十分な量の雑音データが必要であることと合成の際に入力音声の SN 比が既知である必要があることの二つの短

所がある．本章では上記の二つの欠点を克服した，HMM 合成法に基づいた新しい音響モデル適応化手法を提案する．(1) の問題に対してあらかじめさまざまな雑音による初期雑音モデルを必要に応じて適応化する HMM 合成法を提案し，(2) の問題に対して，SN 比ごとに音響モデルの経路を複数個用意してマルチパス化する方法を用いる．

3.2. 音響モデルの雑音環境適応化

雑音環境下における音声認識で，最も簡単かつ理想的な音響モデルの構築法は入力音声と同じ雑音環境下での学習データを用いて音響モデルを構築する方法である．この方法では混入する雑音が既知であるという制約が必要である．以下，入力音声と同一の雑音環境下の学習データで作成した音響モデルを特定環境 (Environment dependent) 音響モデルと呼ぶ．本節では，混入する雑音環境が未知である条件下において特定環境音響モデルを用いた場合の問題点を検証するため，音響モデルの学習データと評価データの雑音環境が一致している場合と一致していない場合の音声認識性能の比較を行う．

特定環境音響モデルの性能評価のため，旅行対話タスク [64] の音声データベースによる評価実験を行う．評価データ数は 42 話者，計 551 発声である．旅行対話データベースの学習セットに電子協騒音データベース [65] の展示会場雑音を SN 比が 15 dB となるよう重畳し，展示会場音響モデルを作成する．ベースライン音響モデル (Baseline HMM)，展示会場音響モデル (Exhibition hall HMM) いずれも総状態数 1400，各状態 5 混合分布，性別依存，話者非依存状態共有 HMM を用いる．評価データは展示会場雑音を重畳したもの (雑音環境一致) と自動車雑音を重畳したもの (雑音環境不一致) を作成し，比較する．結果を図 3.1 に示す．雑音環境が一致しない場合の認識性能はベースラインと同等であることがわかる．

同様に，AURORA2[14, 15, 16] データベースによる認識実験を行う．AURORA2 データベースは ETSI STQ-AURORA DSR Working Group が作成した，雑音環境下における音声認識システム評価用データベースである．詳細を表 3.1 に示す．

特定環境音響モデル (Environment dependent HMM) の学習には，学習セット

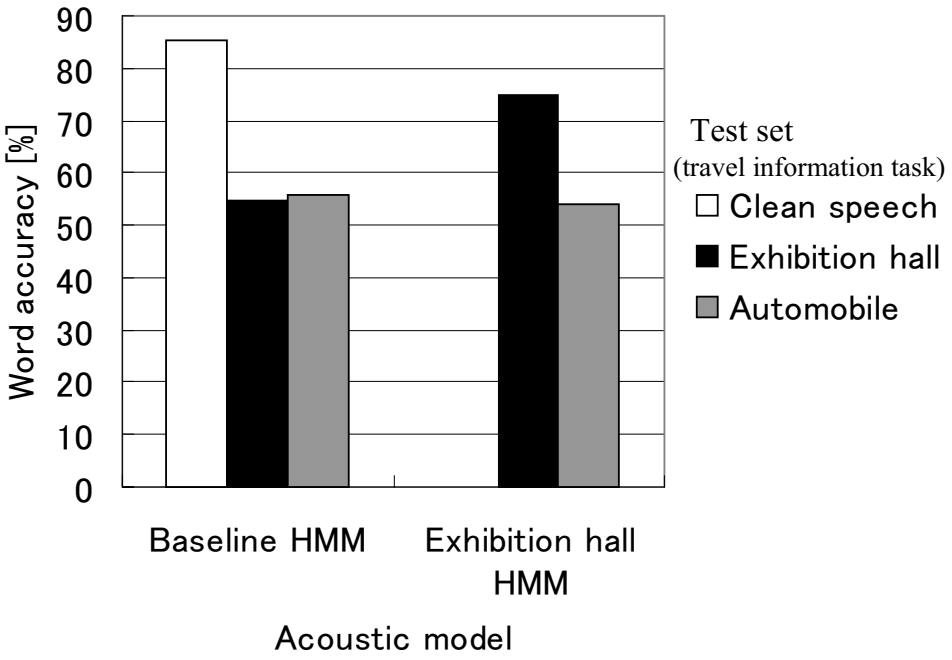

図 3.1 ベースライン，特定環境音響モデルの旅行対話タスクによる性能評価

表 3.1 AURORA2 データベース

タスク: TI-digit (連続数字認識)
サンプリング周波数: 8kHz
16bit PCM / モノラル
training set
雑音: subway, babble, car noise, exhibition hall
SN 比: 5dB, 10dB, 15dB, 20dB, clean
成人男性 55 話者，成人女性 55 話者
全発話数: 8840
test set A
雑音: subway, babble, car noise, exhibition hall
SN 比: -5dB, 0dB, 5dB, 10dB, 15dB, 20dB, clean
成人男性 52 話者，成人女性 52 話者
全発話数: 28028
test set B
雑音: restaurant, street, airport, train station
SN 比: -5dB, 0dB, 5dB, 10dB, 15dB, 20dB, clean
成人男性 52 話者，成人女性 52 話者
全発話数: 28028

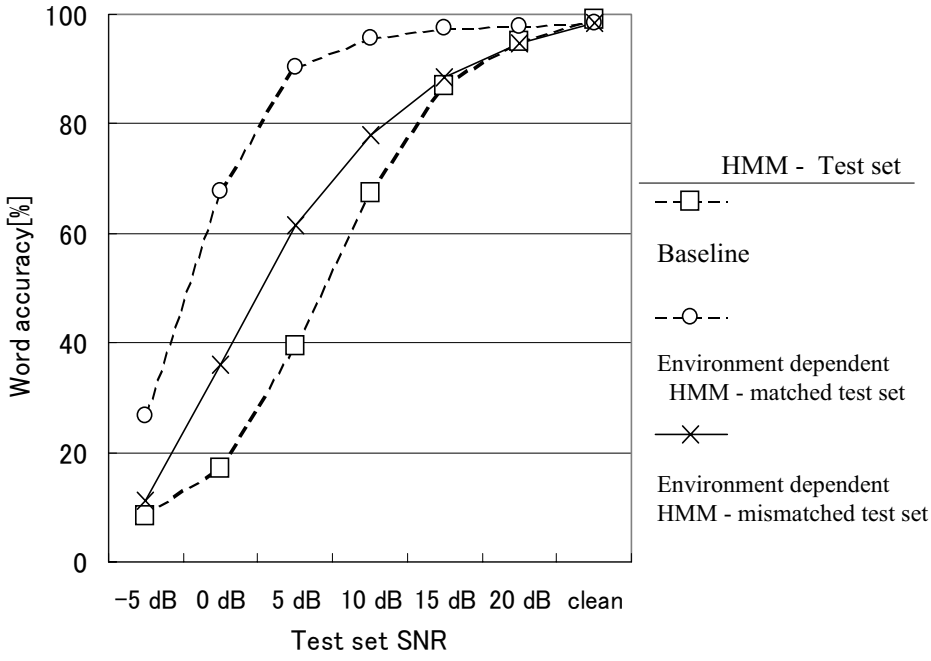

図 3.2 ベースライン，特定環境音響モデルによる AURORA2 タスクによる性能評価

のうち 1 種類の雑音の混入したサブセットを用い，評価には A セットのうち対応した雑音が混入した音声データを用いる．これらの平均を特定環境音響モデルの性能とする．ベースライン音響モデル，特定環境音響モデルいずれも，各数字ごとに 16 状態，各状態 3 混合の HMM を用いる．音響モデル学習データと評価データの雑音環境が異なっている場合の評価として，上記特定環境音響モデルと A セットのうち対応しない雑音の混入した音声データを用いて評価する．これらの場合の平均の認識率を図 3.2 に示す．学習データと評価データの雑音環境が一致していない場合，入力音声の SN 比低下に伴って大幅に認識性能が低下している。

次に，学習データとしてさまざまな雑音を含んだ音声データを用いて音響モデルの学習を行った場合 (マルチコンディションモデル) について検討する．マルチコンディションモデルの学習データには AURORA2 の学習セット全てを用いる．

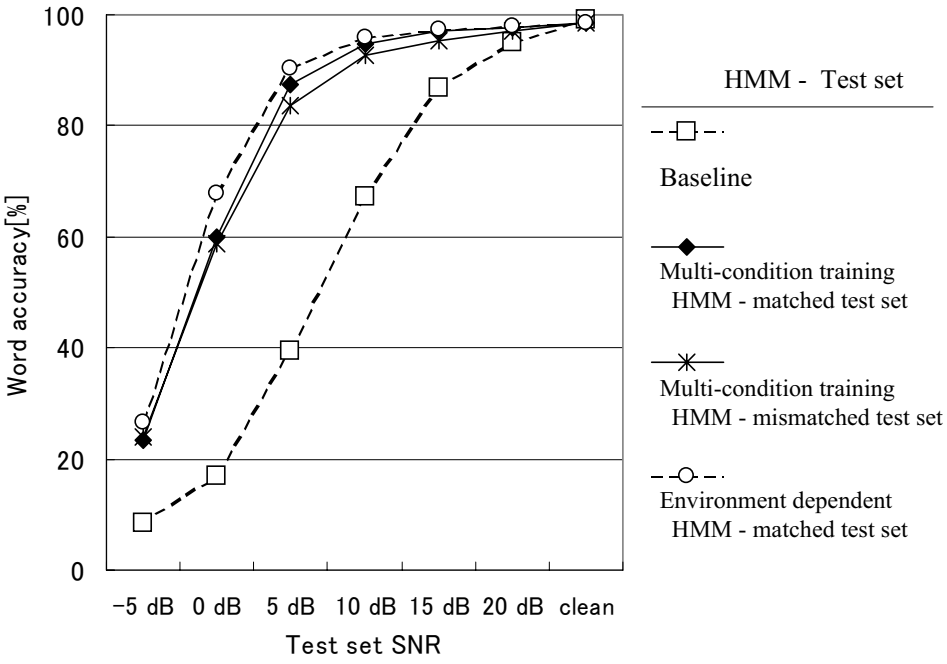

図 3.3 マルチコンディションモデルによる認識結果

ベースライン音響モデルと同じく，各数字ごとに16状態，各状態3混合のHMMを用いる．評価実験として，学習セットと同じ雑音環境であるAセット(雑音環境既知)と学習セットに含まれないBセット(雑音環境未知)を用いて比較する．結果を図3.3に示す．マルチコンディションモデルを用いた場合には，雑音環境が未知であるBセットの場合，認識性能はAセットに比べて劣る．雑音環境が既知であるAセットの場合は，特定環境音響モデルを用いた場合に比べて劣化の幅が小さい．実環境での音声認識を考えた場合，現場の雑音環境は，時々刻々変化し，新しい雑音の出現なども起こりえる．このような場合，マルチコンディションモデルの性質を備えたまま，対象の雑音の性質に合わせて逐次適応する手法が有望となる．

表 3.2 旅行対話タスクにおける実験条件

音声データ	:	サンプリング周波数 16 kHz
		16 bit PCM monaural
特徴ベクトル	:	12Cep + 12 Δ Cep + Δ power
		(power は合成時のみ使用)
		フレーム長 20 ミリ秒 ハミング窓
		フレームシフト 10 ミリ秒
		CMN なし
Clean speech HMM	:	性別依存/ 話者非依存 トライフォン
		1 音素あたり 4〜5 状態
		全 1400 状態の HMnet
		ATR 音声データベース 計 3619 話者を用いて学習
Noise GMM	:	1 状態 {1,2,4,8} 混合
		{100, 10, 5, 3, 1} 秒 の電子協展示会場雑音 (ブース内) で学習
評価データ/タスク	:	ATR 旅行対話評価セット 42 対話
		連続単語音声認識
		電子協展示会場雑音 (ブース内) を SN 比=15 dB になるよう重畳

3.3. HMM 合成法における実騒音データ量と認識性能

2.3 に示す通り，従来法の HMM 合成は使用環境の実騒音データを用いて環境音モデルの学習を行っており，性能は実騒音データの量と性質に依存する．本節では予備実験として実騒音データ量と認識性能の関係について調べる．

実験条件を表 3.2 に，実験結果を図 3.4 に示す．表中，Cep はケプストラム係数を表す．実験結果の図より，学習データ量が少ない場合，十分にある場合に比べて音声認識性能が劣化し，その傾向は混合数が大きいほど顕著に表れることがわかる．また，データ量に関しては，この評価環境を GMM で表すためには 10 秒の実雑音データが必要であるといえる．しかしながら，実際の使用環境として変動する雑音環境を考えた場合，雑音環境に変化がみられるたびに 10 秒 の実雑音データを取得することは難しく，適応化に必要な実雑音データ量を削減する必要

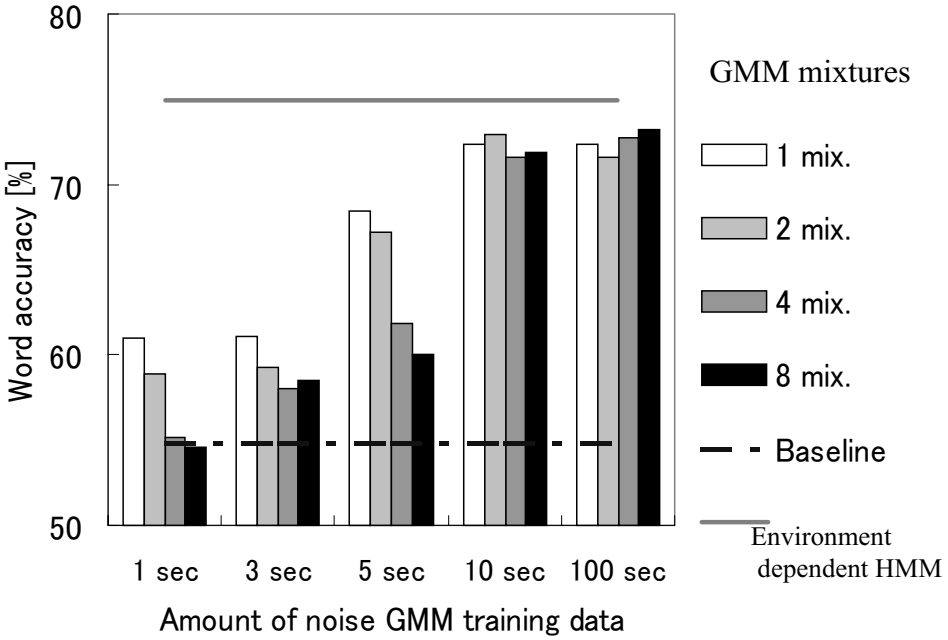

図 3.4 雑音 GMM 学習データ量と認識性能

がある.

3.4. 雑音モデル適応化を用いたHMM合成

本節では，少量の雑音データで環境適応化を行う方法として，雑音 DB とモデル適応化を用いる HMM 合成法を提案する．提案法の概略を図 3.5 に示す．従来の HMM 合成においては，環境雑音のモデル化に十分な量の雑音データが必要である．ここでは環境雑音を集めて確率モデルにモデル化することが重要である．そこで，提案法では雑音環境のモデル化を行う際に，あらかじめ用意した雑音 DB の先験知識を利用することで，少量の雑音データで環境適応化が可能になる．

まず，あらかじめ多様な雑音を含む DB を用いて初期雑音 GMM の学習を行う．また，のちの計算簡単化のため，初期雑音 GMM とクリーン音声 HMM を HMM

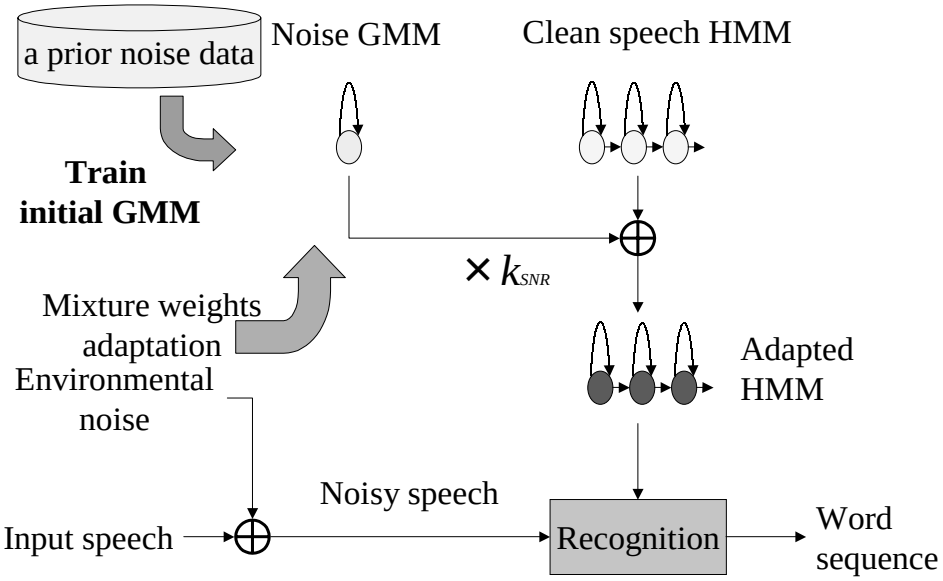

図 3.5 雑音 GMM 重み適応化を用いた HMM 合成

合成した初期合成 HMM も準備しておく．環境適応化の際には，少量の実雑音データを取得して初期雑音 GMM に混合重み適応化を施し，適応化雑音 GMM を得る．適応化には MAP 推定 [47] を用いる．事前分布 $g(\theta)$ は Dirichlet 分布に従う．これより，混合重みの推定式は以下のようになる．

$$\tilde{w} = \frac{(\nu_k - 1) + \sum_{t=1}^{T} c_{kt}}{\sum_{k=1}^{K}(\nu_k - 1) + \sum_{k=1}^{K}\sum_{t=1}^{T} c_{kt}}$$

$$c_{kt} = \frac{w_k N_k(x_t)}{\sum_{l=1}^{K} w_l N_l(x_t)}$$

適応化を GMM の混合重み係数に限定しているので，適応化を行った上で HMM 合成した適応化 HMM と，初期合成 HMM の間で，各確率分布の平均や分散が変化することはなく，環境適応化により変化するのは重み係数のみである．したがって，GMM 適応化で得た重み係数を合成後のモデルに直接反映することで適

応化 HMM を得ることができ，計算量を大きく削減できる．この関係を図 3.6 に示す．図中，パワースペクトル領域でモデル化されたクリーン音声 HMM と環境音 GMM を合成する場合を考える．説明のため，クリーン音声 HMM の第 1 状態の出力確率分布が 2 混合の混合正規分布

$$w_{S11}P_{S11} + w_{S12}P_{S12}$$

で与えられ，環境音 GMM も同様に

$$w_{N1}P_{N1} + w_{N2}P_{N2}$$

で与えられるとする．w はそれぞれ混合重み係数である．$P_{S11}, P_{S12}, P_{N1}, P_{N2}$ はそれぞれ正規分布であり，平均 μ と分散 Σ で表される．

$$P_{S11} = N(\mu_{S11}, \Sigma_{S11})$$

$$P_{S12} = N(\mu_{S12}, \Sigma_{S12})$$

$$P_{N1} = N(\mu_{N1}, \Sigma_{N1})$$

$$P_{N2} = N(\mu_{N2}, \Sigma_{N2})$$

このとき合成後の HMM の第 1 状態の出力確率分布は次式で表される．

$$\begin{aligned} & w_{S11}w_{N1}(P_{S11} \oplus k_{SNR}P_{N1}) \\ + \quad & w_{S12}w_{N1}(P_{S12} \oplus k_{SNR}P_{N1}) \\ + \quad & w_{S11}w_{N2}(P_{S11} \oplus k_{SNR}P_{N2}) \\ + \quad & w_{S12}w_{N2}(P_{S12} \oplus k_{SNR}P_{N2}) \end{aligned} \tag{3.1}$$

ここで k_{SNR} は SN 比に応じた係数である．正規分布の和は，平均・分散それぞれの和として表される．たとえば，

$$\begin{aligned} & P_{S11} \oplus k_{SNR}P_{N1} \\ & = N(\mu_{S11}, \Sigma_{S11}) + k_{SNR}N(\mu_{N1}, \Sigma_{N1}) \\ & \quad = N(\mu_{S11} + k_{SNR}\mu_{N1}, \Sigma_{S11} + k_{SNR}\Sigma_{N1}) \end{aligned} \tag{3.2}$$

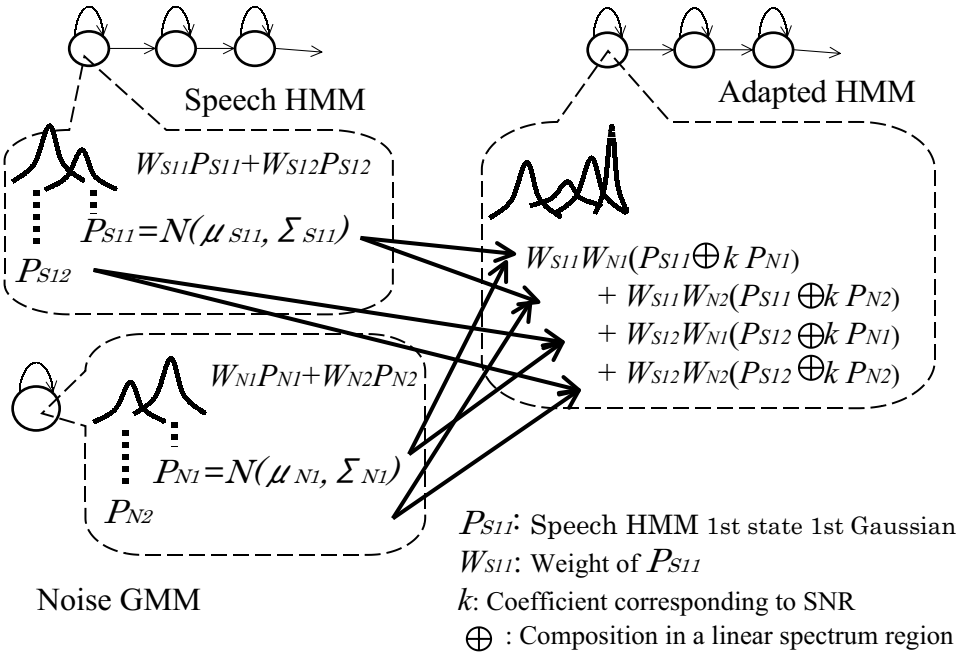

図 3.6 提案法における混合重み適応化の計算

雑音 GMM の混合重み適応化を行った場合，w_{N1}, w_{N2} のみが適応化されるので，合成後のモデルにおいても式 (3.1) の各項の確率分布の平均や分散は変化しない．したがって，適応化された GMM の重みを反映することで適応化 HMM を得ることができ，適応化の都度，HMM 合成を行う必要はない．これに対し従来法では，各項の確率分布が環境音モデルを学習するたびに変わるため，その都度式 (3.2) にしたがった計算が必要になる．

3.5. SN 比別マルチパスモデル

HMM 合成法には式 (2.22) に示す通り SN 比が既知であるという制約が存在する．SN 比の推定については，南らが，適応データによる SN 比の推定を用いた HMM 合成法を提案しているが，SN 比の推定に繰り返し計算が必要であるため，

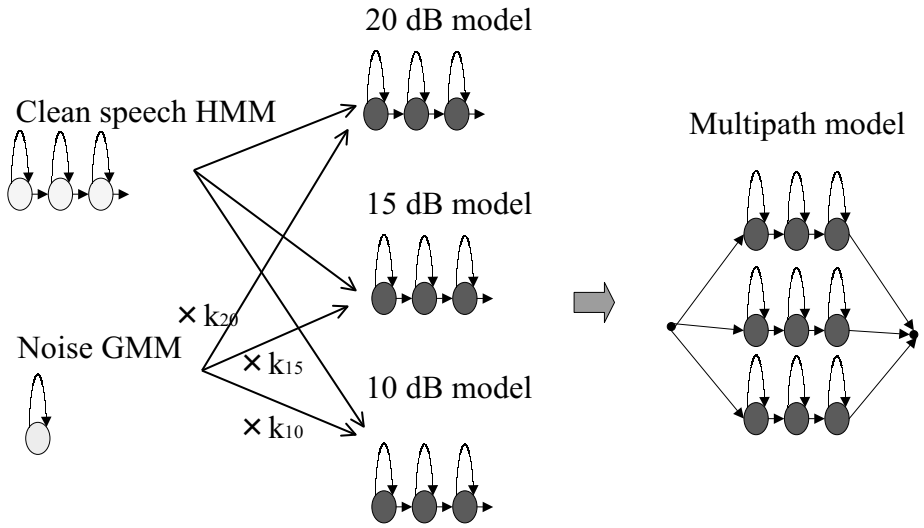

図 3.7 SN比別マルチパスモデル

本論文で取り扱っているSN比が時々刻々変化する場合にはSN比が変動するたびにSN比の再推定を行うことは難しい [67]．本論文では，入力音声のSN比がわからない問題点の解決のため，複数のSN比に対応した適応化HMMを1つのモデル中に並列に構築する手法を用いる．この手法は自由発話音声の認識における音響特徴の変動や多様化に対する方法として用いられている [68, 69]．本手法の概略図を図3.7に示す．雑音モデルを合成する際に，入力音声として予測される範囲内のいくつかのSN比に対応した複数の合成HMMを得る．図3.7の例では、まずSN比を 10 dB, 15 dB, 20 dBとし，(2.22) 式に基づいて入力音声のモデルおよび環境雑音モデルのケプストラム係数の0次項が想定SN比となるよう k_{10}, k_{15}, k_{20} を決定する [59]．次に各SN比ごとにモデルの合成を行う．図の例では，10 dB model, 15 dB model 20 dB model の3つのモデルをHMM合成により作成し，並列に並べて1つのSN比別マルチパスモデルとして構成する．認識

の際，入力音声の SN 比はわからないが，デコーディング時にもっとも尤度の高いパスが選択される．

3.6. 評価実験

3.6.1 旅行対話タスクによる評価実験

雑音モデルの適応化を用いた HMM 合成法と SN 比別マルチパスモデルの性能評価のため，雑音を重畳した音声データを用いた認識実験を行う．評価データは，旅行対話に関するタスク 42 対話 [64] を用い，これに電子協騒音 DB[65] の雑音を各 SN 比に合わせて重畳する．音響モデルとして，

- ベースライン HMM (総状態数 1400，各状態 5 混合)
- 特定環境音響モデル: Environment dependent HMM (総状態数 1400，各状態 5 混合)
- 10 秒の実雑音データで学習した雑音モデルを SN 比 = 15 dB として HMM 合成した音響モデル: 10 sec GMM train-composition 15 dB (総状態数 1400，各状態 40 混合)
- 雑音モデルを 1 秒の実雑音データで重み適応化し SN 比 = 10 dB として HMM 合成した音響モデル: 1 sec GMM adapt-composition 10 dB (総状態数 1400，各状態 40 混合)
- 雑音モデルを 1 秒の実雑音データで重み適応化し SN 比 = 15 dB として HMM 合成した音響モデル: 1 sec GMM adapt-composition 15 dB (総状態数 1400，各状態 40 混合)
- 雑音モデルを 1 秒の実雑音データで重み適応化し SN 比 = 20 dB として HMM 合成した音響モデル: 1 sec GMM adapt-composition 20 dB (総状態数 1400，各状態 40 混合)

- 雑音モデルを 1 秒の実雑音データで重み適応化し SN 比 = 10 dB, 15 dB, 20 dB として HMM 合成し, マルチパス化した音響モデル: multipath (総状態数 4200，各状態 40 混合)

- 雑音モデルに重み適応化を行わずに (初期雑音モデルのまま) SNR = 15 dB として HMM 合成した音響モデル (総状態数 1400，各状態 40 混合)

を用いる．雑音モデルは 1 状態 8 混合の GMM を用い，初期雑音 GMM の作成には電子協騒音 DB より 25 種 計 250 秒を用いた．この 25 種には評価用データに重畳した雑音は含まれない．雑音モデルの混合数を決めるため，雑音モデルに重み適応化を行わずに (初期雑音モデルのまま) SNR = 15 dB として HMM 合成した音響モデルを用いて予備実験を行う．上記騒音 DB を用いて 1 混合，2 混合，4 混合，8 混合の雑音モデルを作成し HMM 合成した音響モデルによる評価実験を行った。図 3.8 に実験結果を示す．混合数を大きくした方が認識精度がよく，混合数を 4 から 8 にした場合の性能向上幅が小さくなっていることから初期雑音モデルの混合数を 8 として以後の評価実験を行う．

評価データには展示会場雑音および自動車雑音を，それぞれ SN 比 = 0 dB, 5 dB, 10 dB, 15 dB, 20 dB となるよう重畳する．音響モデルと評価データの雑音環境が一致している場合と一致していない場合の両方について評価実験を行う．

まず，雑音 GMM 重み適応化を用いた HMM 合成法の効果を検証するため，SN 比を 15 dB 固定とした評価実験を行う．結果を図 3.9 に示す．雑音環境が一致している場合においては，雑音 GMM 重み適応化を用いる場合の方が雑音 GMM の学習を行う場合に比べて 2%程度認識精度が劣化するが，適応データ量は 10 分の 1 しか必要としない．雑音環境が一致していない場合には，特定環境音響モデルや従来法の HMM 合成では性能が低下しているのに対し，提案法では初期 GMM 作成に多様な雑音環境の情報を与えているので性能低下が抑えられている．

次に，SN 比別マルチパスモデルの効果を検証する．音響モデルとして，1 秒の実データで重み適応化した雑音 GMM を SN 比 = 10 dB, 15 dB, 20 dB となるよう HMM 合成した音響モデル，およびこの 3 つのモデルをマルチパス化した音響モデルを用いる．結果を図 3.10 に示す．各 SN 比に対応したモデルは，それぞれの SN 比において良い認識性能を示し，それ以外の SN 比において性能が低下

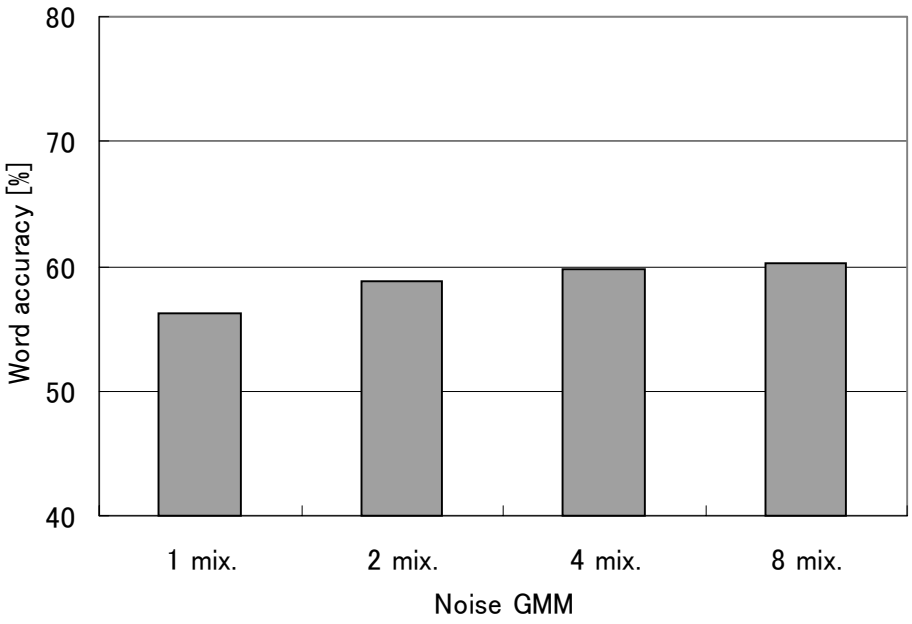

図 3.8 初期雑音 GMM の混合数

しているのに対し，マルチパス化した音響モデルは 10〜20 dB の間において各モデルの認識性能の頂点を結ぶような安定した認識性能を示す.

雑音 GMM の重み適応化を用いた HMM 合成と SN 比別マルチパスモデルを併用した場合における未知雑音に対するロバスト性の評価実験を行う．結果を図 3.11 に示す．雑音環境一致モデルおよび従来の HMM 合成法によるマルチパスモデルでは異なった雑音環境において大幅に性能が低下しているのに対し，雑音 GMM の重み適応化を用いる方法では性能低下の幅が小さい.

3.6.2 AURORA2 タスクによる評価実験

さらに多様な雑音環境下での性能評価とマルチコンディションモデルとの比較を行うため，AURORA2 DB による認識実験を行う．音響モデルとして，

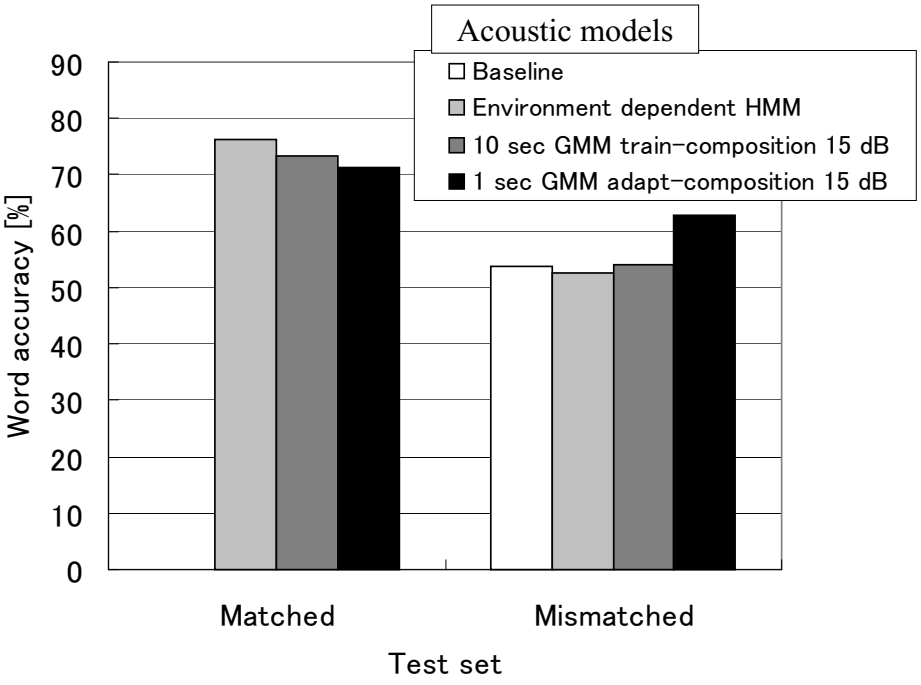

図 3.9 SN比既知 (15 dB) の場合における評価実験

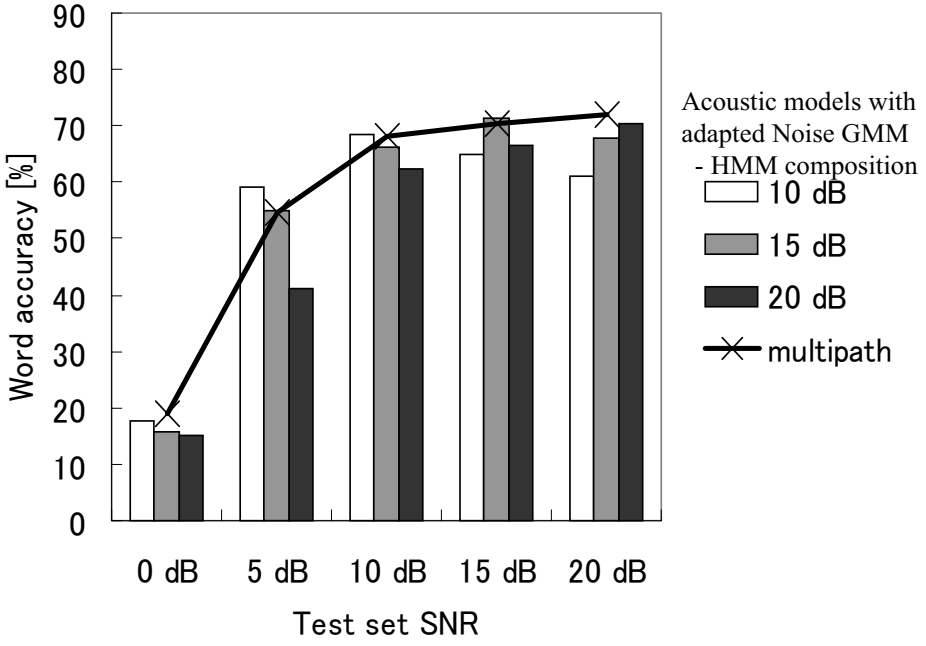

図 3.10 SN 比未知の場合における評価実験

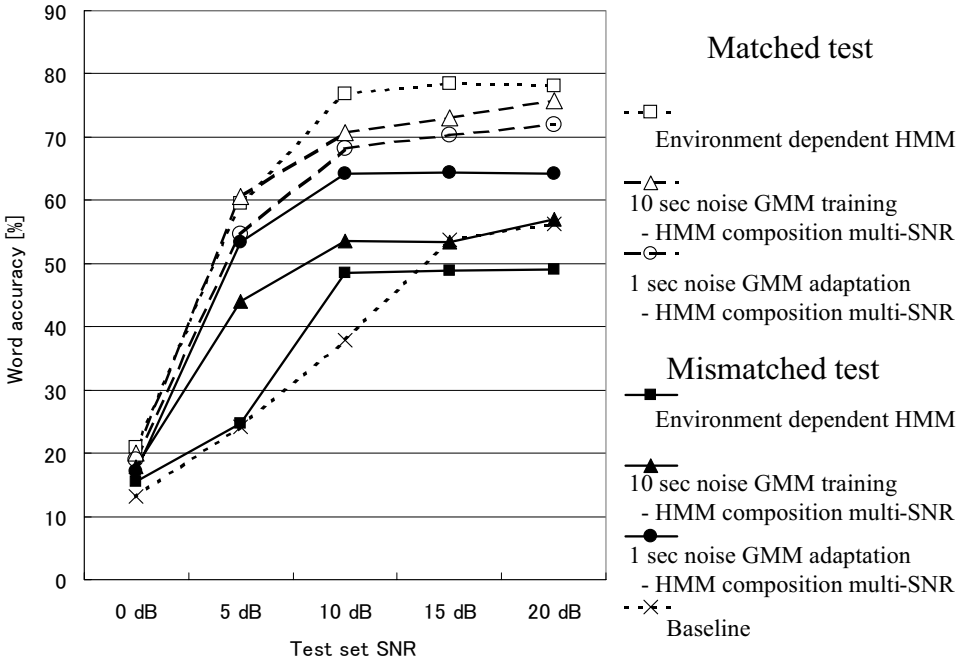

図 3.11 雑音の種類とSN比の両方が未知の場合における評価実験

- クリーンHMM(ベースライン) (各数字16状態，3混合)
- マルチコンディションモデル (各数字16状態，3混合)
- 10秒の実雑音データで学習した雑音モデルをSN比 = 15 dBとしてHMM合成した音響モデル (各数字16状態，24混合)
- 10秒の実雑音データで学習した雑音モデルをSN比 = 5 dB, 10 dB, 15 dB, 20 dB, ∞ dBとしてHMM合成し，マルチパス化した音響モデル (各数字80状態，24混合)
- 雑音モデルを1秒の実雑音データで重み適応化しSN比 = 15 dBとして

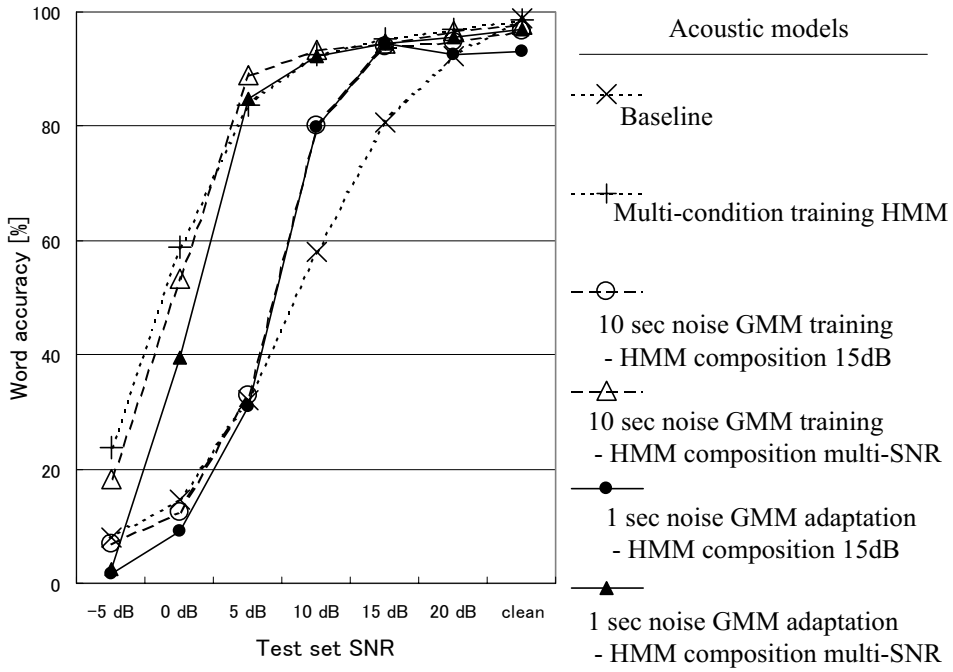

図 3.12 AURORA2 タスクによる提案手法の評価実験

HMM 合成した音響モデル (各数字 16 状態，24 混合)

・雑音モデルを 1 秒の実雑音データで重み適応化し SN 比 = 5 dB, 10 dB, 15 dB, 20 dB, ∞ dB として HMM 合成し，マルチパス化した音響モデル (各数字 80 状態，24 混合)

を用いる．雑音モデルは 1 状態 8 混合の GMM を用い，初期雑音 GMM の作成には電子協騒音 DB より 25 種 計 250 秒を用いた．この 25 種には評価用データに重畳した雑音は含まれない．認識タスクは TI digit 連続数字認識タスクである．評価には B セットを用いる．これはマルチコンディションモデルの学習データに対してオープンな条件である．実験結果を図 3.12 に示す．雑音 GMM 重み適応

化とマルチパスモデルを用いることで，マルチコンディションモデルと同等の認識性能を得ることができた．ベースラインモデルに対して，SN 比 = 15 dB 固定の場合 14%の性能向上，マルチパスモデルにすることで SN 比 = 5 dB において 53%の性能向上である．このときに用いた適応データ量は 1 秒であり，雑音 GMM 重み適応化を用いない場合に比べて 10 分の 1 に削減が可能である．提案法を用いることで状態数，混合数ともに増加し，認識時における計算量が増加するが，適応データ量の削減，ならびに SN 比が変動するタスクにおける認識率改善が実験を通して明らかになった．

一方，雑音 GMM の重み適応化に従来の HMM 合成を用いた場合と同量の適応データを与えた場合について調べる．以下の 3 つの音響モデルについて比較する．

- 10 秒の実雑音データで学習した雑音モデルを SN 比 = 5 dB, 10 dB, 15 dB, 20 dB, ∞ dB として HMM 合成し，マルチパス化した音響モデル (各数字 80 状態，24 混合)
- 雑音モデルを 1 秒の実雑音データで重み適応化し SN 比 = 5 dB, 10 dB, 15 dB, 20 dB, ∞ dB として HMM 合成し，マルチパス化した音響モデル (各数字 80 状態，24 混合)
- 雑音モデルを 10 秒の実雑音データで重み適応化し SN 比 = 5 dB, 10 dB, 15 dB, 20 dB, ∞ dB として HMM 合成し，マルチパス化した音響モデル (各数字 80 状態，24 混合)

実験結果を図 3.13 に示す．雑音 GMM の重み適応化に用いるデータ量を増加させた場合にも，大幅な性能の向上は認められなかった．10 秒のデータを用いた場合，雑音 GMM の重み適応化に用いた場合よりも雑音 GMM をはじめから学習した場合の方がわずかながら性能が勝っている現象が見られた．このことから，雑音環境が比較的定常であり，十分な雑音データ量を取得することが可能な場合には両者を組み合わせて使う方が有効と考えられる．

AURORA2 タスクを用いたスペクトル減算の性能評価を行った先行研究 [66] と本章の提案手法の比較を図 3.14 に示す．提案法は 1 秒の適応データを用いて雑音 GMM 適応化を行い SN 比別マルチパスモデルとしたものである．スペクトル減算

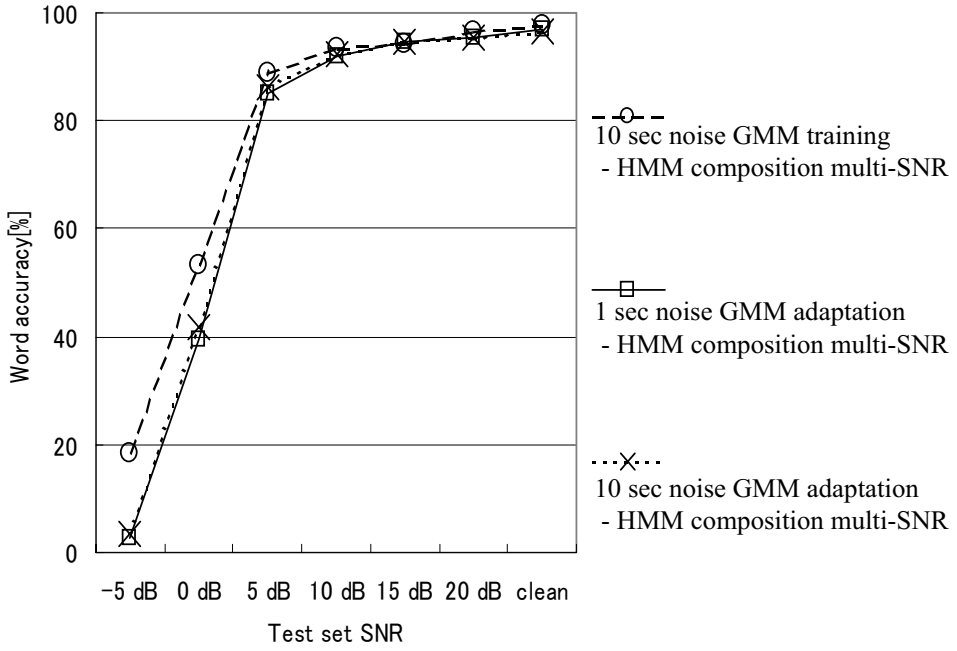

図 3.13 提案法による適応データ量と認識性能の比較

は入力音声の冒頭 1 秒を用いて雑音スペクトルの推定を行い，式 (2.37) および式 (2.38) により雑音除去を行い，クリーン HMM で認識したものである．式 (2.37) および式 (2.38) において $\alpha = 2.0, \beta = 0.1, \gamma = 0.95$ とした．結果より，特に低 SN 比の領域において本手法の効果が大きいことがわかる．本論文の 4 章では雑音除去アルゴリズムとしてスペクトル減算を用いている．これは実際の情報提供端末に実装する上で，計算コストの制約を考慮したためである．将来，計算コストの問題を解決し，本章の提案法を導入することでさらに性能向上が見込まれる．

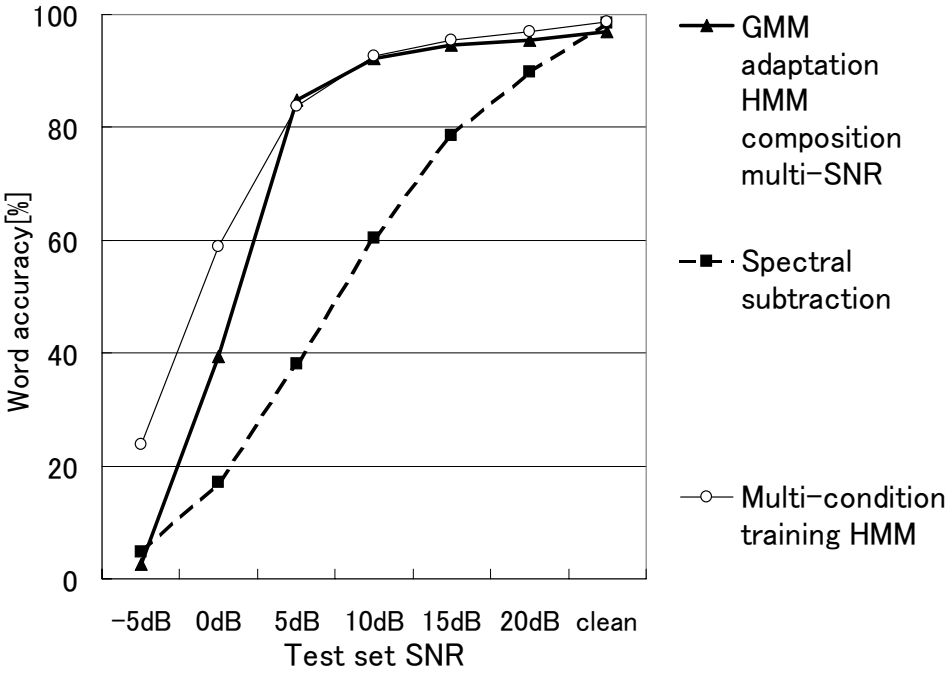

図 3.14 提案法とスペクトル減算の認識性能の比較

3.7. まとめ

本章では多様な雑音が混入する音声認識における問題点として，雑音の種類が未知である問題とSN比が未知である問題を取り上げた．その解決策として，雑音GMMの重み適応化によるHMM合成法とSN比別マルチパスモデルを用いる方法について検討した．性能評価は旅行対話タスクの音声データと，AURORA2データを用いた認識実験を行った．実験の結果，AURORA2タスク，SN比 = 5 dBに対して1秒の実雑音データを適応データとして用いることで53%の認識性能向上を得た．これは雑音GMM重み適応化を行わない場合において10秒の実雑音データを使った場合に相当し，適応データ量で10分の1への削減となる．提案法を

用いることで状態数，混合数ともに増加するが，適応データ量の削減，ならびにSN 比が変動するタスクにおける認識率改善が評価実験を通して明らかになった．

しかしながら，雑音データを十分に取得できる環境においては，雑音 GMM 重み適応化を用いない従来手法の方が優れた性能を示すことも確認された．このことから，雑音 GMM の重み適応化を行う提案法と，はじめから雑音 GMM の学習を行う従来法とを雑音環境の定常性に基づいて選択的に用いる機構について検討を行う必要がある．

第4章

据え置き型情報端末向き雑音処理を用いた音声入力インタフェース

4.1. はじめに

本章では，情報提供端末やコンビニエンスストアに置かれるようなオンラインショッピング端末，鉄道の券売機などの社会システム機器に音声認識を組み込むことを想定した音声入力インタフェースについて検討を行う．まず情報提供端末機の大きさなどの物理的制約と設置環境の音響特性を考慮した入力インタフェースの設計について述べ，次に実騒音環境において収録した評価用データによってその音声入力性能と音声認識性能を評価した．

先行研究においては，各手法の性能評価に主としてシミュレーション実験によって有効性が示されてきた．性能評価のための騒音・音声データベースやシミュレーションのためのツール群が整備され，実環境における音声認識技術の研究に広く用いられている [16, 36]．シミュレーションの方法としては再現する環境により，次の三つの方法が用いられる．第一に加法性雑音を擬似的に作り出す方法として，クリーン (無騒音) な環境で録音したテストデータに騒音データを計算機上で重畳してテストデータを作成する方法．第二に実環境の空間音響特性による音声の歪みを再現する方法として，空間音響特性のインパルス応答を畳み込んでテストデータを作成する方法．第三に擬似的な実騒音環境を作り出す方法として，実験室内に置いた複数個のスピーカから騒音を再生し，そこで被験者に発声させることでテストデータを作成する方法である．しかし，いずれの方法によっても実際

の騒音環境を再現することは一般に困難であるため，シミュレーション実験で示された有効性が実環境においても有効に機能することを確認する必要がある．

4.2. ハンズフリー音声認識を備えた情報提供端末機の試作

ハンズフリー音声入力インタフェースを備えた情報提供端末機の試作を行った．基本となる端末機はオムロン社製の多機能端末機 Cyber Gate [70] を用いる．Cyber Gate は web ベースのアーキテクチャを採用しており，多様な範囲のシステム機器として利用することができる．

試作機の外観を図 4.1 に示す．ハンズフリー音声入力用のマイクロホンアレーはタッチパネルを備えた表示画面の上下に，水平方向かつ話者に対して直角に設置される．端末機の全幅約 45 cm に収まるよう，素子間隔 42.45 mm で 8 つのマイクロホン素子を直線上に配置したアレーを 2 セット用いる．設置一の床面からの高さは，表示画面の下部に取りつけたものが 990 mm，上部に取りつけたものが 1420 mm である．話者は表示画面に対して中央に正対するので話者から上下の各アレーは等距離になるよう設置される．各 8 素子直線マイクロホンアレーには話者方向を正面固定として，アナログ回路による遅延を与えることにより水平面方向の遅延を与え指向性を形成する．2 つの直線アレーの出力に対し，ディジタル遅延によって垂直方向の指向性が与えられる．以上により，本システムのハンズフリー音声入力インタフェースが構成される．ハンズフリー入力と従来の受話器による入力の比較評価を行うために受話器型のマイクロホンも取りつけられている．本試作機は，従来からある情報提供端末機に音声入力インタフェースを追加する形で実現された．試作機のシステム構成を図 4.2 に示す．コンテンツは CD(コンパクトディスク) や書籍のオンラインショッピング，チケット予約，旅行案内など多岐にわたって用意されている．ユーザは音声とタッチパネルによるマルチモーダル入力により，システムと対話を行い，種々のコンテンツを呼び出すことでタスクが達成される．

この情報提供端末でサービスが行われるコンテンツの一例を図 4.3 に示す．コ

図 4.1 音声認識情報提供端末機の外観

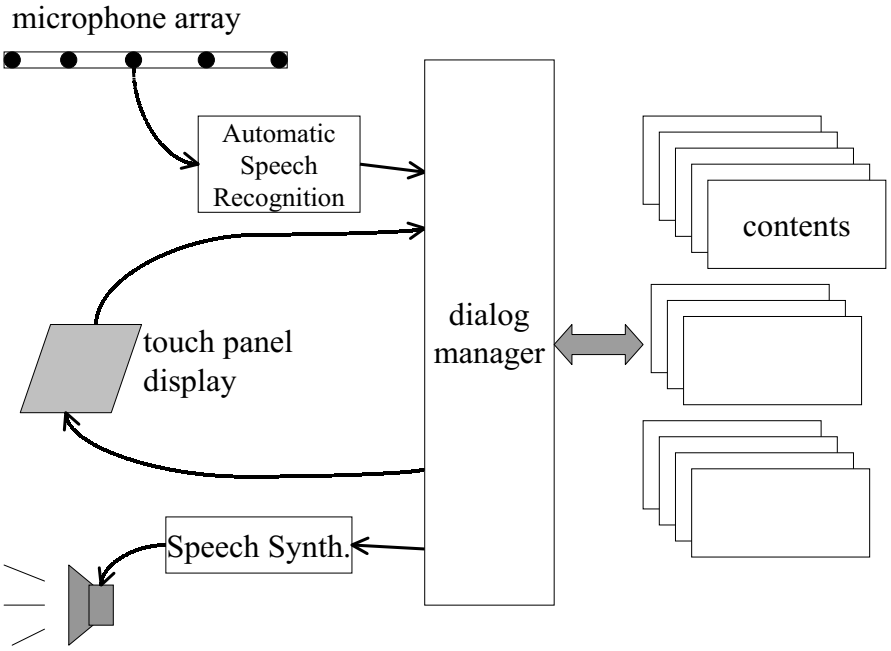

図 4.2 音声認識情報提供端末機のシステム構成

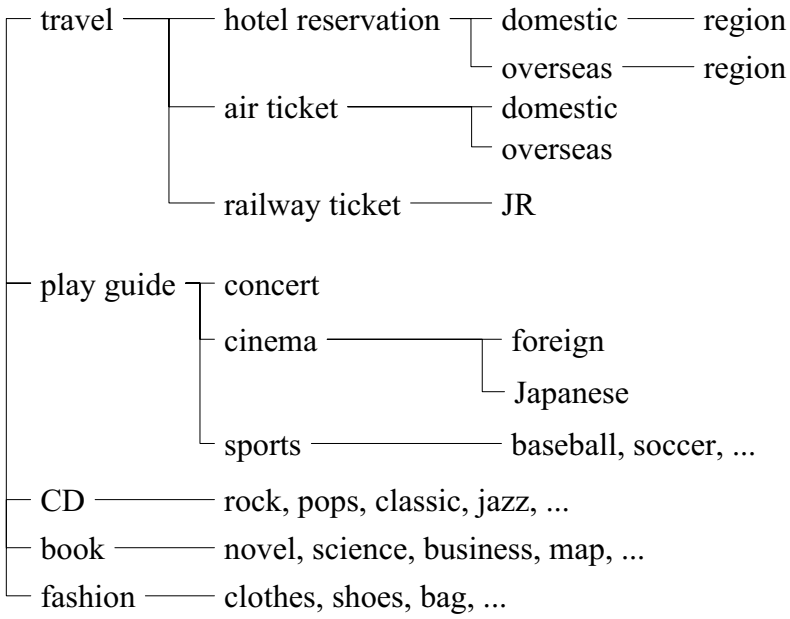

図 4.3 情報提供端末のコンテンツ構成例

ンテンツは情報を提供する側の意図で分類されてツリー状に構築されている．従来のタッチパネルによる方法でタスクを遂行する場合，音楽のジャンル (ポップス・演歌) や男女別，五十音順，などの階層的に分類されたツリー構造を逐一たどって目的のコンテンツを探す必要がある．この場合コンテンツが多種多量になればなるほどに階層構造が多重化するため，ユーザの操作量が増加する欠点があった．これに対し，音声認識を用いた方法ではツリーのリーフに相当する目的のコンテンツを直接選択することが可能で，これに簡単な確認の対話を行うことでタスクを達成することが可能である．

4.3. 評価実験

4.3.1 音声入力インタフェースの構成

今回開発した音声入力インタフェースの構成図を図 4.4 に示す．前半が遅延和形アレー処理部，後半が SS 処理部である．実験ではマルチチャンネル同期録音を行った音声データをディジタルデータとして保持し，遅延和形アレー処理，SS 処理，認識実験をそれぞれ別個に行った．以下，各部について詳細に述べる．

マイクロホンアレー　情報提供端末をはじめとする社会システム機器に設置することを考えると，マイクロホンアレー全体の大きさやマイクロホン素子の配置に制限を受ける．筐体の幅が 50 cm 程度であることとユーザ (話者) 正面中央に表示画面が置かれることを考慮し，図 4.5 のように画面の四辺を囲むように小型無指向性マイクロホンを等間隔で配置する．各素子の間隔は 42.45 mm である．マイクロホン素子には，ホシデン KUC1323 を用いた．マイクロホンとユーザ (話者) の位置関係を図 4.6 に示す．アレー面は水平面に対して 60 °の傾きをもって設置され，想定話者位置 (アレー面から 500 mm) の方向は 73 °の仰角方向になる．

遅延和形アレー処理部　遅延和形アレー処理の遅延量を算出する方式は，各マイクロホン素子に到達する音波の伝達を球面波として取り扱うか，平面波に近似して取り扱うかに分類される．情報提供端末の利用形態を考えると，ユーザは機器

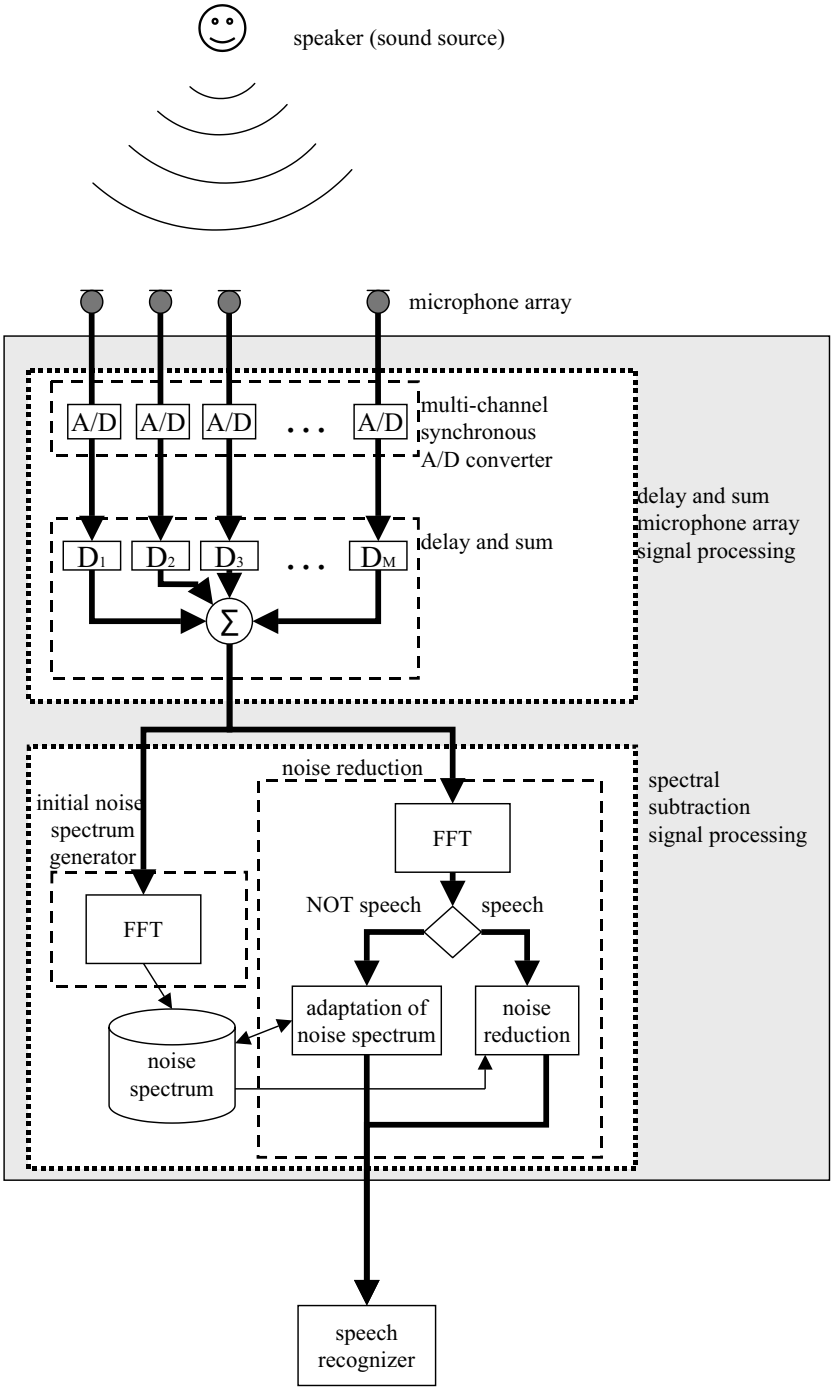

図 4.4 音声入力部の構成図

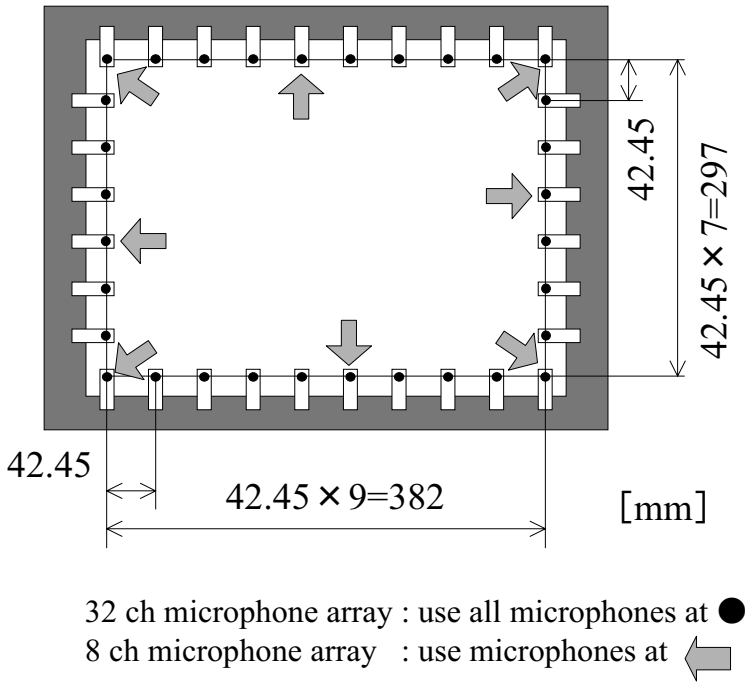

図 4.5 32 ch または 8 ch のマイクロホンアレーの配置図

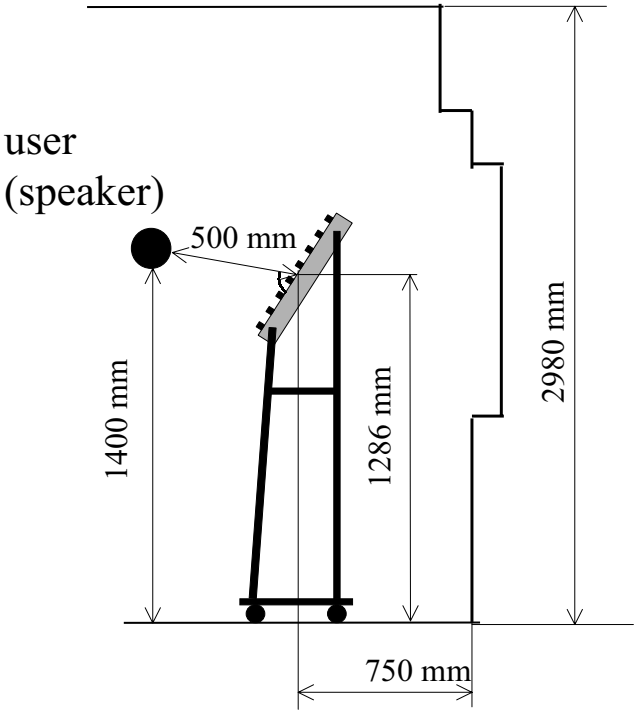

図 4.6 マイクロホンアレーと話者の位置関係

表 4.1 音源位置高さの差に関する評価実験結果

elevation angle[°]	SNR[dB]
70	6.09
71	6.11
72	6.13
73	6.15
74	6.14
75	6.17

本体に接近して操作を行う．したがって，アレーの大きさに比べてユーザの口 (音源) とアレー面との距離が小さい．この場合には音波を平面波に近似した場合の誤差が無視できない量になるため，球面波として厳密に取り扱う方式を採用する．基準点をアレー中央 (長方形の対角線の交点) とし，想定話者 (目的音源) 位置と基準点との距離を基準距離とする．図 4.5，図 4.6 より想定話者位置と各アレーマイクロホン素子との距離を求め，この距離と基準距離との距離差を音速によって時間差に変換し，時間差を吸収する遅延時間を各マイクロホンに対して設定する．遅延は FIR フィルタによって各マイクロホンの入力に与える．

本実験においては，想定話者位置を 1 点に固定して遅延和形アレー処理を行っている．この影響を予備実験として調べた．予備実験は，評価実験を行った環境と同じ駅騒音環境において成人男性 1 名 (音源位置の高さ 1400 mm) が発話した場合の SN 比を測定したものである．この場合，話者方向はアレー面に対して仰角 73 °方向であるが，遅延和アレー処理を行う際に想定話者方向を 70 °〜75 °にそれぞれ向けた場合の SN 比を示している．想定話者方向 70 °は音源位置高さ 1300 mm に，方向 75 °は 1500 mm に相当する．結果を表 4.1 に示す．これより，話者の身長差は± 10 cm 程度までにおいて影響がないことがわかる．

32 ch すべてを用いた構成のほかに，四隅と各辺中央の素子を用いた 8 ch の構成の 2 種類のマイクロホンアレーについて評価を行う．

表 4.2 音声認識器の仕様

標本化周波数	11.025 kHz
分析フレーム長	20 msec
分析フレームシフト	10 msec
特徴ベクトル	13MFCC，ΔMFCC，ΔΔMFCC， power， Δpower， ΔΔpower CMN あり
音響モデル	speaker independent context dependent 3 状態 16 混合 tied-mixture HMM 計 1085 状態
学習データ	ATR 連続音声データベース 16 話者 毎日新聞連続音声データベース 300 話者
認識辞書	ATR 音素バランス 216 単語

SS 処理部　スペクトル特徴量として FFT 短時間スペクトルを用いる．フレーム長 512 サンプル，フレームシフト 256 サンプルとした．雑音スペクトルの初期値として各発声の冒頭 5 フレーム分のスペクトルの平均値をあてる．以降，式 (2.37), (2.38) にしたがって雑音成分の除去と雑音スペクトルの更新を逐次行う．予備実験により，各パラメータに $\alpha = 2.0$ ， $\beta = 0.15$ ，$\gamma = 0.98$ を与える．

音声認識部　音声認識部は HMM を用いた不特定話者孤立単語認識器を用いる．この認識器は Linux-OS 上で動作する．表 4.2 に認識器の仕様を示す．収録音声の標本化周波数と認識器の受理する標本化周波数が異なるため，認識の前処理としてダウンサンプリングを行っている．

SS の雑音除去性能について AURORA2 タスク [14, 15, 16] で比較評価した先行研究として [66] がある．本研究の 3 章において音響モデルの適応化の評価を行っ

表 4.3 実環境音声データの収録条件

標本化周波数	24 kHz
A/D 変換	16 ビット PCM
マイクロホン	
接話マイク	Sennheizer HMD410-6
アレー	ホシデン KUC1323
	32 ch or 8 ch
タスク	ATR 音韻バランス 216 単語
話者	成人男性 2 名，成人女性 2 名

ているが，本章では音声入力フロントエンドとして，音声データの前処理による対雑音性能の評価に主眼を置く．したがって，音声認識部で環境適応化を行っていない．

4.3.2 実騒音環境評価音声データの収集

実環境での評価データは地下鉄京都駅改札口付近で収録した．収録地点の騒音レベルは約 70 dBA であった．比較対象として十分に広い防音室で同一の収録機材を用いて音声の収録を行った．一般に騒音環境下における音声認識性能の評価には，シミュレーション実験が多用される．ここでいうシミュレーション実験とは，計算機上で雑音を重畳した評価用音声データを用いた評価のみではなく，実験室内に騒音を再生して想定環境を模擬的に再現して評価を行う場合も含む．しかし，本論文で想定している情報端末などの置かれる環境はシミュレーションの困難な環境である．あらゆる方向から到来する雑音や，反射・残響が存在することがその理由である．それゆえ，実際に使用する環境で収録した音声を用いた評価実験を行うことの意義は大きい．

音声収録の条件を表 4.3 に示す．音声収録の状況を図 4.7 に示す．同一音声を用いた性能比較を行うため，マイクロホンアレーによる収録と同時にヘッドホンに装着された接話マイクを用いた収録も行った．評価用音声データは ATR 音素バ

図 4.7 地下鉄京都駅自動券売機付近における音声収録 (1999 年 1 月)

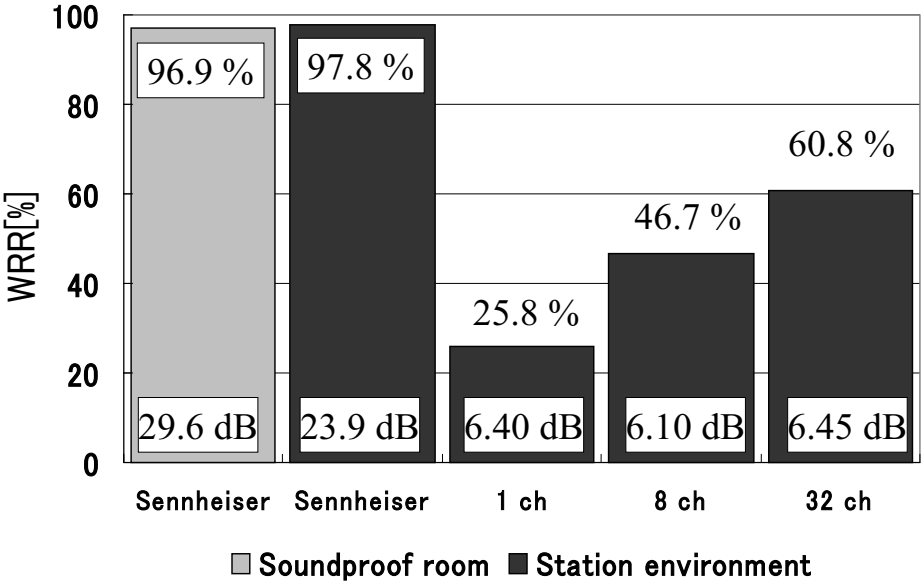

図 4.8 遅延和アレーによる雑音除去・音声認識実験結果

ランス単語セットの 216 単語の孤立単語発声を成人男女各 2 名の話者が発声したものである．

4.3.3 実験結果

実験は遅延和形アレー処理単独，SS 処理単独，および両者を併用した場合について行った．遅延和形アレー処理は 8 ch の場合と 32 ch の場合について評価する．比較対象として，接話マイク (Sennheiser と表記) とアレーを構成するマイクを単独で用いた場合 (1 ch) を用いる。1 ch の評価には最も話者に近いアレー下辺中央の 2 つのマイクをそれぞれ単独で用いた場合の評価結果の平均値を 1ch の評価とする。評価は SN 比と孤立単語認識率で比較する．SN 比は音声区間の平均パワーと音声区間以外の平均パワーの比を [dB] で表したものである．

まず，遅延和アレ―処理だけを用いた場合の雑音除去について，結果を図 4.8

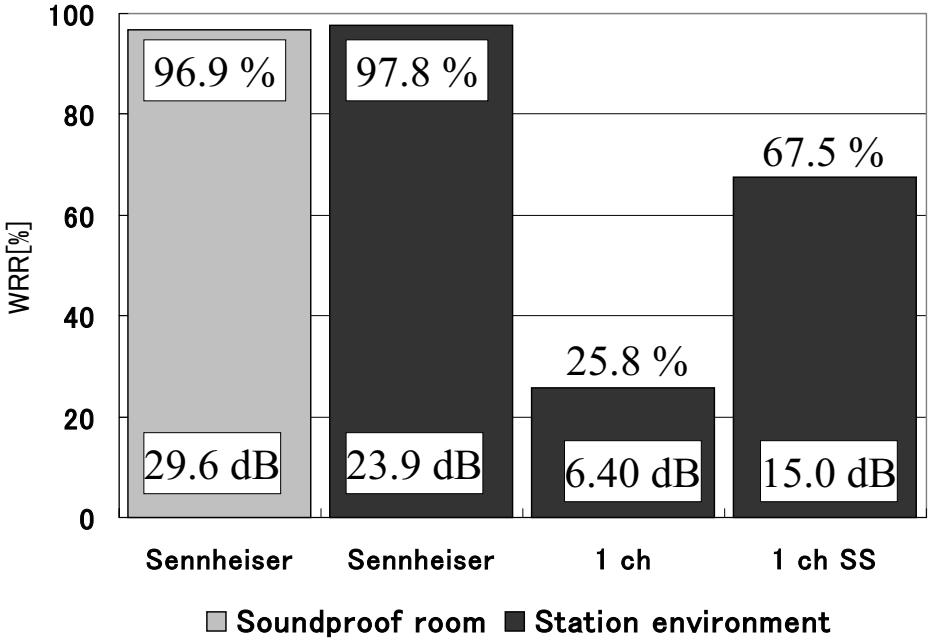

図 4.9 SS による雑音除去・音声認識実験結果

に示す．SN 比で比較した場合，ほとんど改善が見られない．8 ch の場合，むしろ SN 比は悪化しているように見える．これは 1 ch の評価に，最も話者に近いマイク素子による結果を用いているためで，8 ch・32 ch のアレーで用いている素子は 1 ch の評価で用いている素子よりも SN 比，音声認識性能ともに低いものも含んでいるためである．認識精度で比較した場合，マイクアレ―の素子数の増加とともに認識率が向上していることがわかる．しかし，遅延和アレ―処理だけでは十分な認識性能は得られていない．

次に，SS だけを用いた場合について，結果を図 4.9 に示す．SS を用いた場合，SN 比で比較すると大幅に改善している．しかし，処理前の SN 比が悪いために処理後の音声が大きく歪んでしまい，SN 比から期待されるほどの音声認識性能の向上は見られない．

提案法である，遅延和アレーと SS を併用した場合の結果を図 4.10 に示す．以

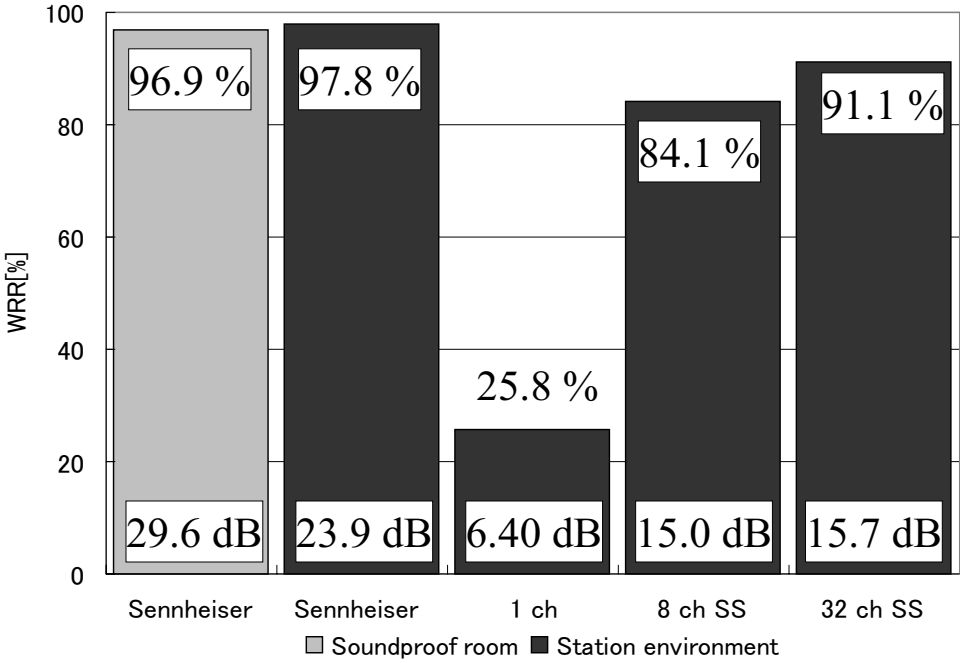

図 4.10 遅延和アレーと SS の併用による雑音除去・音声認識実験結果

上の実験結果を表 4.4，表 4.5 にまとめる．32 ch の遅延和形アレーと SS を併用することで 216 単語の不特定話者孤立単語認識において 91.1%の認識性能を達成できた．この認識性能は防音室における認識性能には至らないものの改善の度合いが大きい．SN 比に着目すると，SS を用いることにより 8 dB 以上の改善が見られ，32 ch アレー SS ありの場合において 15.73 dB になる．このことから SN 比を評価尺度とした場合には SS による改善の寄与が大きいことがわかる．これに対し，音声認識性能に着目すると，SS を用いない場合においても遅延和アレー処理を行うことにより，1 ch の場合に 25.8%であった認識率が 8 ch で 46.7%，32 ch で 60.8%に向上しており，SN 比の改善がわずかであっても遅延和アレー処理の効果が明らかに確認できる．

表 4.4 雑音除去実験結果

環境	マイク		SNR[dB]
防音室	Sennheizer		29.56
	1 ch		20.14
駅騒音	Sennheiser		23.92
	1 ch	SS なし	6.40
	1 ch	SS あり	14.97
	8 ch アレー	SS なし	6.10
	8 ch アレー	SS あり	14.99
	32 ch アレー	SS なし	6.45
	32 ch アレー	SS あり	15.73

表 4.5 孤立単語音声認識実験結果

環境	マイク		WRR[%]
防音室	Sennheizer		96.9
	1 ch		96.6
駅騒音	Sennheiser		97.8
	1 ch	SS なし	25.8
	1 ch	SS あり	67.5
	8 ch アレー	SS なし	46.7
	8 ch アレー	SS あり	84.1
	32 ch アレー	SS なし	60.8
	32 ch アレー	SS あり	91.1

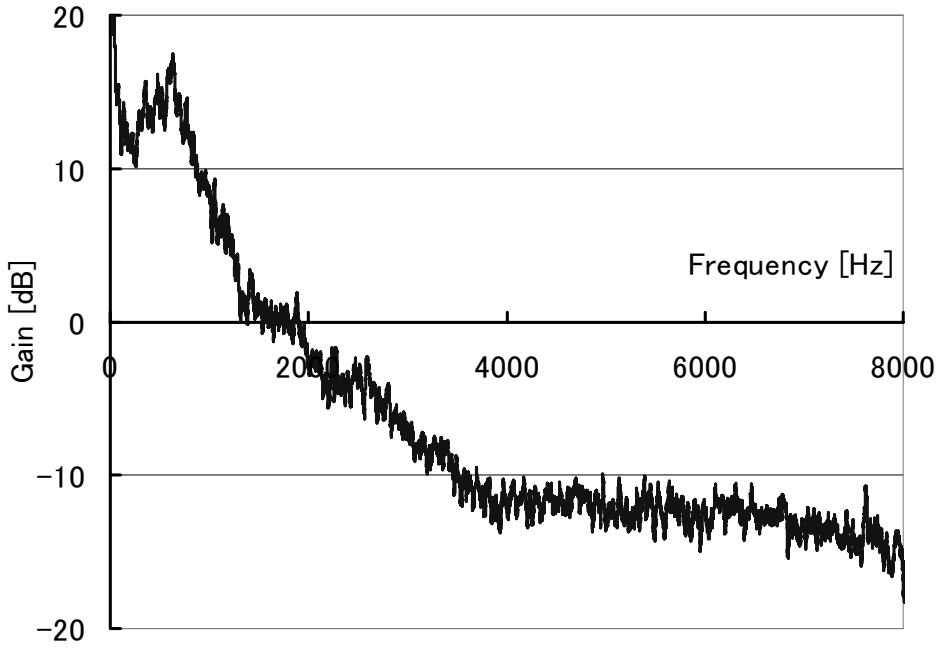

図 4.11 駅騒音の周波数特性

4.4. 考察

遅延和アレ―処理と SS を併用する効果について考察する.

駅騒音の周波数特性を図 4.11 に示す. この図は地下鉄京都駅における音声収録の際に同一機材を用いて収録した約 25 秒の騒音データを FFT 周波数特性分析した結果である. これによれば, 駅構内の騒音は低周波域に大きなパワーをもつ, 偏りのある雑音であることがわかる. 一方, 遅延和アレー処理を用いる場合には, システム機器に搭載可能な大きさの範囲内で設計を行うと, アレーそのものの大きさに制限が加わる. この結果, マイクロホン素子間の間隔を十分にとれないことから雑音除去の効果が十分に得られる周波数帯域が高周波域中心になってしまう.

成人男性話者 1 名の発話に対し, アレーを構成するマイクロホン 1 ch のみを

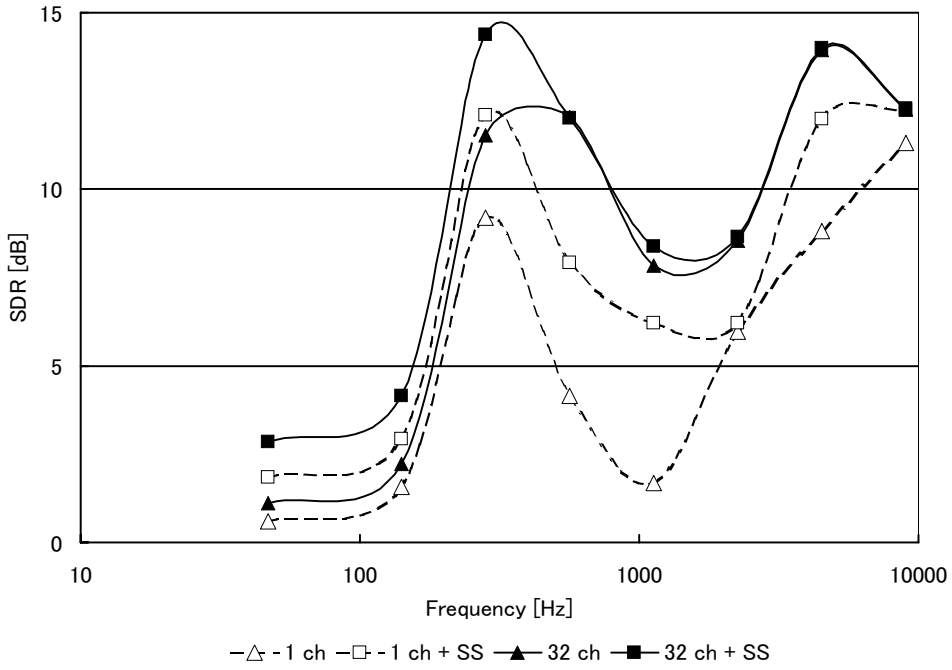

図 4.12 周波数帯域別にみた SDR

用いた場合 (1 ch)，1 ch で収録した音声に SS を施した場合 (1 ch+SS)，32 ch アレーを用いた場合 (32 ch)，および 32 ch アレーと SS を併用した場合の 4 種類の収録音声の SDR (Signal to Deviation Ratio) を周波数帯域ごとに求める．

$$SDR = 10\log_{10}\frac{\Sigma(s(n))^2}{\Sigma(s(n)-\beta x(n+\tau))^2} \qquad [\mathrm{dB}] \tag{4.1}$$

ここで，$s(n)$ は基準信号，$x(n)$ は評価信号を示す．基準信号として接話マイクを用いて同時収録した音声を用いる．β は $s(n)$ と $x(n)$ のパワー差を補正する係数であり，τ は基準信号と評価信号の時間差を補正する． 時間差 τ は接話マイクとアレーを構成するマイクの距離差とアレー処理，SS 処理により生じる遅延時間から算出する．図 4.12 に結果を示す．

マイクロホンアレーを用いることにより 500 Hz 以上の帯域において雑音が抑圧されており，低域においては抑圧の効果が小さい．全帯域における SN 比を評

価基準とした場合，駅騒音においては低域のパワーが大きいために雑音除去の効果が現れていないが，認識性能による評価では中高域の雑音が抑圧されているため，認識性能が向上している．これに対し，SS を単独で用いた場合においては低い周波数帯域において混入している雑音が大きいために音声に与える歪みが大きく，SN 比の改善量に比べて SDR は改善されていないことがわかる．この原因により，全域の SN 比において 15 dB 程度まで雑音が抑圧されているにもかかわらず認識性能の向上は 60%にとどまる．マイクロホンアレーと SS を併用することにより，システム機器に搭載する制約の範囲内において音声波形に対する歪みを低減させることと全周波数帯域における雑音除去効果の両立が可能になった．

以上より，システム機器に搭載可能な大きさの範囲内における設計では，中高域をマイクロホンアレー処理による雑音除去，低域を SS による方法の 2 つを併用する方法が有効であることがわかる．

4.5. まとめ

本章では，実際の情報提供端末に実装可能なハンズフリー音声入力インタフェースの設計と実装を行った．実環境下で収録した音声データによる評価実験の結果，32 ch の遅延和アレーと SS を併用することで 216 単語の孤立単語認識タスクにおいて 91.1% の認識性能を達成できた．遅延和アレ―処理だけ，あるいは SS だけを用いたのでは不十分な認識性能しか得られず，両者を併用する効果が大きいことが確認できた．

実環境下で収録した評価実験で 90% を超える認識性能を示せたことは，実用化するうえで非常に意義が大きい．すでに実用化されている電話音声自動応答などのサービスにおいて，90%の認識性能がひとつの目安になっているからである．一方，本章では音響モデル適応化を併用する方法について検討を行っていない．また，評価タスクとして音素バランス単語タスクを用いている．これらについて，今後評価を行う必要が残されている．

第5章

結論

5.1. まとめ

本研究では，音声認識システムの実用化における次の三点の課題について検討した．三点とは，音響モデルを雑音環境適応化する際に必要な適応データ量の削減，変動する雑音環境に対してロバストな音響モデルの構築，実際の音声対話アプリケーションにおいて利用可能な音声入力インタフェースの構築である．

第一の課題および第二の課題に対し雑音 GMM 適応化と SN 比別マルチパスモデルを用いた HMM 合成法を提案し，その認識性能を評価実験により検証した．性能評価には旅行対話タスクの音声データと AURORA2 データを用いた評価実験を行った．実験の結果，AURORA2 タスク，SN 比=5 dB の条件において，適応データとして 1 秒の実雑音データを用いた場合に 53%の認識性能向上を得た．雑音 GMM の重み適応化を行わない従来法で同等の性能向上を得るには，適応データとして 10 秒の実雑音データを必要とする．必要な適応データの量で 10 分の 1 への削減を実現した．しかしながら，実雑音データを十分に取得できる場合には，従来法の方が優れた性能を示すことも確認された．

第三の課題に対し，情報提供端末向けにハンズフリー音声入力インタフェースを開発した．音声対話が有効なアプリケーションとして券売機や情報提供端末などの社会システム機器における音声対話を取り上げ，遅延和形マイクロホンアレー処理とスペクトル減算による雑音除去を組み合わせた音声入力インタフェースを試作した．次いで，構成要素である遅延和型マイクロホンアレー処理とスペクトル減算による雑音除去について，認識性能の向上を評価実験により検証した．評

価実験には 216 単語の孤立単語認識タスクの音声データを実環境下で収録して評価データとして用いた．実際の雑音環境はシミュレーションでは再現困難な要因が多く存在する．音声対話システムの実用化を考えたとき，実環境下における音声データを用いた評価実験を行うことは非常に意義が大きい．評価実験により 32 ch の遅延和アレーとスペクトル減算を組み合わせることによって 216 単語の不特定話者孤立単語認識において 91.1%の認識率を達成することができた．70 dBA という劣悪な騒音環境下であるが，実用化の目安となる 90%を越える認識性能が得られたことは大きな成果である．一方，スペクトル減算あるいは遅延和形アレーを単独で用いたのでは十分な認識性能を得ることが難しいことも確認された．すなわち，低域における雑音除去性能を遅延和アレー処理単独で実現しようとした場合アレー長が大きくなる欠点があり，実装上の大きさの制限内でマイクロホンアレーの設計を行った場合，低域において所要の雑音除去性能を確保できない．これに対して，SS を単独で用いた場合は，駅騒音などの劣悪な環境下では音声に歪みを生じることから認識性能の低下の原因になる．したがって，情報端末などのシステム機器の音声入力インタフェースとしては互いの雑音除去性能を相補するよう両者を併用する本手法が有効であるといえる．これにより，情報端末等が実際に利用される環境において，十分な認識性能が得られることを確認した．

5.2. 今後の課題

音声インタフェースの実用化へ向けた課題として，(1) 多様な機器に利用することを目指した音声インタフェースの構築，(2) さまざまな状況に応じた雑音対策による性能向上，(3) 雑音環境下で音声インタフェースを快適に利用するための技術確立が挙げられる．以下，各々について詳細に述べる．

第一の課題は，音声インタフェースの小型化・低コスト化が最大の課題となる．近年の携帯電話を用いた種々のサービスの発展に象徴されるように，情報提供や検索，予約，購入といった従来は社会システム機器によって提供されてきたサービスが携帯端末で利用できるように変化しつつある．また，高速ネットワークの充実により，前記のサービスを家庭にいながらにして享受できる環境が整いつつ

ある．これらの流れは前記のサービスを受けるユーザ層や利用シーンを拡大することにつながり，音声認識を用いたマン-マシンインタフェースの必要性はさらに高くなる．このことから，一般家庭に導入が可能な音声入力インタフェースの形態として，たとえば家電や家電のリモコンに搭載しうる大きさや性能，あるいは携帯情報機器に搭載しうる大きさや性能が必要になる．

第二の課題について，音響モデル適応化において本研究で今後の課題とした点が二点ある．状況に応じて適応化手法を使い分ける機構の開発と，空間音響特性および回線音響特性の補償である．前者は，3章において明らかになった通り，実雑音データを十分取得できる環境やタスクにおいては従来法である雑音 GMM の重み適応化を行わない HMM 合成法の方が優れた性能を示している．また，実雑音下におけるユーザの発声を事前に取得できるタスクにおいては従来から用いられている MLLR などの音響モデル適応化手法を用いることも考えられる．このことから，雑音環境の定常性やタスクの種類，対話の進行状況に応じて最適な音響モデル適応化手法を選択する機構について検討を行う必要がある．後者については，本研究では混入する雑音として加法性の雑音についてのみ取り扱っている．しかしながら，実環境には空間音響特性による歪みや音声を伝送する回線歪みが存在する．第一の課題において述べた状況を考えると，家庭内では一般に公共の場に比べて狭く閉じられた空間であるといえる．このことは壁や天井，床での反射による歪みを無視できないことを意味する．また，携帯端末については，音声認識をサーバ側で行うと想定した場合，通信回線の歪みの影響を避けられない．これらのことから，空間音響特性や回線音響特性の歪みを補償することは音声対話システムの利用範囲の拡大に必須の課題であるといえる．

第三の課題について，本研究における音声入力インタフェースの評価実験は音声認識性能の評価に限定しており，インタフェース全体の評価のためにはタスク達成率やユーザの満足度を尺度とした評価が必要になる．音声対話として考えた場合，音声を用いた出力インタフェースの評価・検討も必要となる．また，雑音環境下で人対人の対話を考えた場合雑音のない環境における対話に比べて，発話が短く簡潔になるために冗長な情報が少なくなる，一方，必ず確認対話を行う，といった現象が起きることが経験から予想される．人対システムの対話において

も同様の現象が考えられることから，雑音環境下での音声対話システムにおけるユーザのふるまいを雑音のない環境下におけるふるまいと比較分析することで，雑音環境下における最適な音声対話のモデル化や最適な対話シナリオの作成指針が導き出せるであろう．音声を用いた出力インタフェースについて考えると，雑音環境下における情報伝達の正確さ，雑音環境下での聞き取りやすさ，などの基準による合成音声の評価は必要となるであろう．また，出力として画面を併用する際の音声出力と画面出力の同期制御についても検討する必要がある．

以上の課題について今後取り組むことで，音声対話システムや音声対話を利用したサービスの応用範囲拡大に貢献していきたい．

謝 辞

本研究を進めるにあたり，主指導教官である情報科学研究科 鹿野清宏教授には，あらゆる面で御指導をいただきました．ここに心から感謝の意を表します．また，本論文について多くの有益なご助言をいただきました情報科学研究科木戸出正繼教授，猿渡洋助教授に厚く御礼申し上げます．

株式会社国際電気通信基礎技術研究所 (ATR) 音声言語コミュニケーション研究所第一研究室 中村哲室長には，本学在任中より，研究に関するご助言をはじめ，多岐にわたりご指導いただきました．ここに深く感謝いたします．

ATR 音声言語コミュニケーション研究所 山本誠一所長，オムロン株式会社技術本部コントロール研究所荒尾真樹所長，IT 研究所舘林浩前所長，コントロール研究所知識制御グループ石田勉グループ長には本研究を行う機会を与えていただきました．ATR 音声言語コミュニケーション研究所の研究員諸氏，オムロン株式会社技術本部コントロール研究所の研究員諸氏，ならびに本学情報科学研究科音情報処理研究室の教員諸氏，学生諸氏には日々，議論いただきました．

本研究の第 3 章は，通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである．本研究の第 4 章では，実環境下におけるデータ収集において京都市交通局地下鉄京都駅，オムロン (株) 技術本部コントロール研究所森弘之係長，本学情報科学研究科 (当時) 西浦敬信博士 (現・和歌山大学助手)，三木一浩氏，岡田有加氏にご協力いただきました．

博士課程進学に際し，豊橋技術科学大学中川聖一教授，静岡大学甲斐充彦助教授，信州大学山本一公助手，和歌山大学西浦敬信助手，静岡大学立蔵洋介助手，オムロン (株) 技術本部センシング研究所諏訪正樹博士には率直で有益なアドバイスをいただきました．ここに改めて御礼申し上げます．

参考文献

[1] 立石一真, 第 1 回国際未来学会, (1970)

[2] 大須賀節雄, ヒューマンインタフェース, オーム社, (1992)

[3] J. Raskin, The humane interface, Pearson Education Japan, (2001)

[4] 中川聖一, “音声認識研究の動向,” 電子情報通信学会論文誌 (D-II), vol.J83-D-II, no.2, pp.433–457, (2000)

[5] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, “The HTK BOOK,” (1995)

[6] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山 茂樹, 伊藤克亘, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. “日本語ディクテーション基本ソフトウェア (99 年度版),” 日本音響学会誌, vol.57, no.3, pp.210–214, (2001)

[7] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, “JNAS: Japanese speech corpus for large vocablary continuous speech recognition research,” The Journal of the Acoustical Society of Japan (E), vol.20, no.3, pp.199–206, (1999)

[8] 古井貞熙, 前川喜久雄, 井佐原均, “科学技術振興調整費開放的融合研究制度：大規模コーパスに基づく『話し言葉工学』の構築,” 日本音響学会誌, 56-11, pp.752–755, 日本音響学会, (2000)

[9] 南條浩輝, 加藤一臣, 李晃伸, 河原達也, “大規模な日本語話し言葉データベースを用いた講演音声認識,” 電子情報通信学会論文誌 (D-II), Vol. J86-D-II, No.4, pp.450-459, (2003)

[10] 武田一哉, 板倉文忠, “文部省 COE プログラム統合音響情報研究拠点 (CIAIR) –音声・音響情報処理の多角的研究–,” 日本音響学会誌, vol.56, no.11, pp.748–751, (2000)

[11] 河口信夫, 牛窪誠一, 松原茂樹, 岩博之, 梶田将司, 武田一哉, 板倉文忠, “走行車室内音声対話収録システムの開発,” 電子情報通信学会論文誌 (D-II), Vol.J84-D-II, No.6, pp.909–917, (2001)

[12] http://elazar.itd.nrl.navy.mil/spine/

[13] K. Markov, T. Matsui, R. Gruhn, J. Zhang, and S. Nakamura, “Noise and Channel Distortion Robust ASR System for DARPA SPINE2 Task,” IEICE Transactions on Information and Systems Vol.E86-D, No.3, pp.497-504, (2003)

[14] http://eurospeech2001.org/ese/NoiseRobust/

[15] ETSI standerd document, “Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm,” ETSI ES 201 108 v1.1.2, (2000)

[16] H.G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluation of speech recognition system under noisy conditions,” ISCA ITRW ASR2000 “Automatic Speech Recognition: Challenges for the Next Millennium,” (2000)

[17] IBM ViaVoice, http://www-6.ibm.com/jp/voiceland/

[18] http://www.scansoft.com/naturallyspeaking/,
http://www.scansoft.co.jp/naturallyspeaking/

[19] オムロン (株) 事業開発本部 CMA プロジェクト,
http://www.omron.co.jp/cma/

[20] ユニバーサルデザインコンソーシアム, “ユニバーサルデザインの理念,”
http://www.universal-design.co.jp/what_ud/what_ud.html

[21] 大賀寿郎, 山崎芳男, 金田豊, 音響システムとディジタル処理, pp.173–218, 電子情報通信学会, (1995)

[22] 金田豊, “騒音環境下音声認識のためのマイクロホンアレー技術,” 日本音響学会誌, vol.53, no.11, pp.872–876, (1997)

[23] J.L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, “Computer-steered microphone arrays for sound transduction in large rooms,” Journal of the Acoustical Society of America, 78(5), pp.1508–1518, (1985年11月)

[24] M. Omologo and P. Svaizer, “Talker localization and speech enhancement in a noisy environment using microphone array based acquisition system,” Proc. of European Conference on Speech Communication and Technology, (1993)

[25] G.W. Elko, “Microphone array system for hands-free telecommunication,” Speech Communication, 20, pp.229–240, (1996)

[26] J.L. Flanagan, A.C. Surendran, and E.E. Jan, “Spatially selective sound capture for speech and audio processing,” Speech Communication, 13, pp.207–222, (1993)

[27] 西浦敬信, 中村哲, 鹿野清宏, “反射音を利用したマルチビームフォーミングによる音声認識,” 電子情報通信学会誌 (D-II), Vol. J83-D-II, No. 11, pp. 2198–2205, (2000)

[28] L.J. Griffith and C.W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” IEEE Trans. Antennas Propag., vol.AP-30, no.1, pp.27–34, (1982)

[29] 金田豊, “適応形雑音抑圧マイクロホンアレー (AMNOR) の指向特性,” 日本音響学会誌, vol.44, no.1, pp.23–30, (1988)

[30] M. Mizumachi, M. Akagi, and S. Nakamura, “Design of robust subtractive beamformer for noisy speech recognition,” Proc. of ICSLP2000, vol. 4, pp.57–60, (2000)

[31] S.M. Griebel and M.S. Brandstein, “Microphone array dereverberration using coarse channel modeling,” Proc. of ICASSP2001

[32] M.L. Seltzer, B. Raj, and R.M. Stern, “Speech recognizer-based microphone array processing for robust hands-free speech recognition,” Proc. ICASSP2002, Vol.1 pp.897-900, (2002)

[33] Y. Okada, T. Nishiura, S. Nakamura, T. Yamada, and K. Shikano, “A design of adaptive beamformer based on average speech spectrum for noisy speech recognition,” Acoustical Science and Technology, Acoustical letter, 23, 6, pp.323–327, (2002)

[34] 鹿野清宏, 中村哲, 伊勢史郎, 音声・音情報のディジタル信号処理, 昭晃堂, (1997)

[35] 金学胤, 浅野太, 鈴木陽一, 曽根敏男, “短時間振幅スペクトル推定を用いた2チャネル音声強調法における振幅スペクトル推定について,” 日本音響学会秋季研究発表会講演論文集, vol. I, pp.499–500, (1999)

[36] S. Nakamura, K. Hiyane, F. Asano, Y. Kaneda, T. Yamada, T. Nishiura, T. Kobayashi, S. Ise, and H. Saruwatari, “Design and collection of acoustic sound data for hands-free speech recognition and sound scene understanding,” Proc. of ICME2002(International Conference on Multimedia and Expo), vol.2, pp.161–164, (2002)

[37] S.F. Boll, “Supression of acoustic noise in speech using spectral subtraction,” IEEE Trans. on Acoustics, Speech and Signal Processing, vol.ASSP-27, no.2, pp.113–120, (1979)

[38] S. Furui, “Cepstral analysis technique for automatic speaker verification,” IEEE Trans. on Acoustics, Speech and Signal Processing, vol.ASSP6-29, no.2, pp.254–272, (1981)

[39] B.S. Atal, “Effectiveness of linear prediction characteristics of speech wave for automatic speaker identification and verification,” Journal of Acoustic Society of America, vol.55, pp.1204–1312, (1974)

[40] 庄境誠, 中村哲, 鹿野清宏, “音声強調手法E-CMN/CSSの自動車環境内での音声認識における評価,” 電子情報通信学会論文誌 (D-II), vol.J81-D-II, no.1, pp.1–9, (1998)

[41] 北岡教英, 赤堀一郎, 中川聖一, “スペクトルサブトラクションと時間方向スムージングを用いた雑音環境下音声認識,” 電子情報通信学会論文誌 (D-II), vol.J83-D-II, no.2, pp.500–508, (2000)

[42] J. Chen, K. Palliwal, T. Matsui, K. Yao, K. Markov, and S. Nakamura, “Sub-band based additive noise removal for robust speech recognition,” 電子情報通信学会技術研究報告, vol. 100, No. 522, pp.31–36, (2000)

[43] 藤本雅清, 有木康雄, “カルマンフィルタに基づく音声信号推定法を用いた雑音環境下での音声認識,” 電子情報通信学会論文誌 (D-II), vol.J85-D-II, no.1, pp.1–11, (2002)

[44] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, “Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study,” Proc. of Eurospeech 1999, pp.2407–2410, (1999)

[45] B. Raj, M. Seltzer, and R. Stern, “Reconstruction of damaged spectrographic features for robust speech recognition,” Proc. of ICSLP 2000, pp.357–360, (2000)

[46] A. Hagen, A. Morris, and H. Bourlard, “From multiband full combination to multi-stream full combination processing in robust ASR,” Proc. of ISCA ITRW ASR 2000, pp.175–180, (2000)

[47] J.L. Gauvian and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” IEEE Trans. on Speech and Audio Processing, vol.2, no.2, pp.291–298, (1994)

[48] C.J. Leggetter and P.C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, vol.9, pp.171–185, (1995)

[49] O. Siohan, C. Chesta, and C.-H. Lee, “Joint maximum a posteriori estimation of transformation and hidem Markov model parameters,” Proc. of ICASSP 2000, vol. II, pp. 965–968, (2000)

[50] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, “Jacobian approach to fast acoustic model adaptation,” Proc. of ICASSP1997, vol.II, pp.835–838, (1997)

[51] 松浪加奈子, 芳澤伸一, 馬場朗, 李晃伸, 猿渡洋, 鹿野清宏, “十分統計量を用いた教師なし話者適応および環境適応,” 情報処理学会論文誌, vol.43, no.7, pp.2038–2045, (2002)

[52] V.N. Gupta, M. Lenning, and P. Mermelstein, “Integration of acoustic information in a large vocablary word recognizer,” Proc. of ICASSP1987, vol.II, pp.697–700, (1987)

[53] 相川清明, 河原英紀, 東倉洋一, “順向マスキングの時間周波数特性を模擬した動的ケプストラムを用いた音韻認識,” 電子情報通信学会論文誌 (D-II), vol.J76-A, no.11, pp.1514–1521, (1993)

[54] 中川聖一, 平田好充, 小野義之, “固定長セグメントの統計量を用いた HMM による音声認識,” 電子情報通信学会論文誌 (D-II), vol.J75-D-II, no.5, pp843-851, (1992)

[55] 中川聖一, 山本一公, “セグメント統計量を用いた隠れマルコフモデルによる音声認識,” 電子情報通信学会論文誌 (D-II), vol.J79-D-II, no.12, pp.2032–2038, (1996)

[56] H. Hermansky and N. Morgan, “RASTA processing of speech,” IEEE Trans. on Speech and Audio Processing, vol.2, pp.578–589, (1994)

[57] 中川聖一, 確率モデルによる音声認識, コロナ社, (1988)

[58] M.J.F. Gales and S.J. Young, “HMM recognition in noise using parallel model combination,” Proc. of EUROSPEECH, pp.837–840, (1983)

[59] F. Martin, K. Shikano, and Y. Minami, “Recognition of noisy speech

by composition of hidden Markov models," Proc. of EUROSPEECH, pp.1031–1034, (1993)

[60] 中臺芳夫, 管村昇, 中津良平, "2 入力による雑音除去手法を用いた自動車内の音声認識," 電子情報通信学会技術研究報告, SP89-81, (1981)

[61] J. Meyer and K.U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," Proc of ICASSP1997, pp.1167–1170, (1997)

[62] M. Mizumachi and M. Akagi, "Noise reduction by paired-microphone using spectral subtraction," Proc of ICASSP1998, pp.1001–1004, (1998)

[63] H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Speech enhancement using nonlinear microphone array based on adaptive complementary beamforming," IEICE Trans. Fundamentals, vol. E83-A, No. 5, pp.866–876, (2000)

[64] T. Takezawa, T. Morimoto, and Y. Sagisaka, "Speech and language database for speech translation research in ATR," Proc. of EALREW, pp.148–155, (1998)

[65] 電子協騒音データベース,
http://it.jeita.or.jp/jhistory/committee/humanmed/speech/noisedbj.html

[66] M. Lieb and A. Fischer, "Experiments with the Philips continuous ASR system on the AURORA noisy digits database," Proc. of EUROSPEECH, (2001)

[67] Y. Minami and S. Furui, "A maximum likelihood procedure for universal adaptation method based on HMM composition," Proc. of ICASSP1995, pp.129–132, (1995)

[68] K. Okuda, T. Matsui, and S. Nakamura, "Towards the creation of acoustic models for stressed Japanese speech," Proc of EUROSPEECH, pp.1653–1656, (2001)

[69] T. Takiguchi, S. Nakamura, and K. Shikano, “HMM-separation-based speech recognition for a distant moving speaker,” IEEE Trans. on Speech and Audio Processing, vol.9, no.2, pp.127–140, (2001)

[70] オムロン (株) ソーシアルシステムズビジネスカンパニー, http://www.society.omron.co.jp/cyber/index.html

研究業績

学術論文

1. 中川聖一, 伊田政樹, “連続音声認識タスクの複雑さを表す新しい尺度’,” 電子情報通信学会論文誌, Vol. J81-D-II No.7, pp. 1491–1500, 1998 年 7 月
2. 伊田政樹, 森弘之, 中村哲, 鹿野清宏, “据置き型情報提供端末向き雑音処理を用いた音声入力インタフェース,” 電子情報通信学会論文誌, Vol. J84-D-II No. 6, pp. 868–876, 2001 年 6 月
3. 伊田政樹, 中村哲, “雑音 GMM の適応化と SN 比別マルチパスモデルを用いた HMM 合成による高速な雑音環境適応化,” 電子情報通信学会論文誌, Vol. J86-D-II No. 2, pp. 195–203, 2003 年 2 月

国際会議

1. Seiichi NAKAGAWA and Masaki IDA, “COMPARISON BETWEEN BEAM SEARCH AND A* SEARCH METHODS FOR ISOLATED / CONTINUOUS SPEECH RECOGNITION,” Proceedings of ASR and ASJ Third Joint Meeting, pp. 999–1004, 1996 年 12 月
2. Seiichi NAKAGAWA, Atsuhiko KAI, Toshihiko ITOH, and Masaki IDA, “An Isolated / Continuous Speech Recognition System on a Personal Computer,” Proc. 1997-China-Japan Symposium on Advanced Information Technology, pp. 72–79, 1997 年 4 月
3. Atsuhiko KAI, Masaki IDA, and Seiichi NAKAGAWA, “An acoustic look-ahead method for efficient frame-synchronous search in a large vocabulary speech recognition system,” Proceedings of ICSP1997, pp. 513–518, 1997 年 8 月
4. Masaki IDA and Ryuji YAMASAKI, “An Evaluation of Keyword Spotting Performance Utilizing False Alarm Rejection Based on Prosodic Information,” Proceedings of ICSLP1998, 1998 年 12 月
5. Masaki IDA and Satoshi NAKAMURA, “HMM Composition-based Rapid Model Adaptation Using a Priori Noise GMM Adaptation Evaluation on AURORA2 Corpus,” Proceedings of ICSLP2002, pp. 437–440, 2002 年 9 月

研究会・大会発表

1. 伊田政樹, 中川聖一, “孤立単語音声認識における全探索法・ビームサーチ法・A*探索法の比較,” 日本音響学会 1996 年春季研究発表会講演論文集, 2-5-10, pp. 77–78, 1996 年 3 月
2. 伊田政樹, 中川聖一, “音声認識におけるビームサーチ法と A* 探索法の比較,” 電子情報通信学会 音声研究会 研究技術報告, SP96-12, pp. 1–8, 1996 年 6 月
3. 伊田政樹, 中川聖一, “パープレキシティと音声認識率の関係,” 日本音響学会 1996 年秋季研究発表会講演論文集, 1-3-22, pp. 43–44, 1996 年 9 月
4. 中川 聖一, 伊田 政樹, “タスクの複雑さを表す新しい尺度― SMR-Perplexity ―,” 電子情報通信学会 音声研究会 技術研究報告, SP96-101, pp. 45–52, 1997 年 1 月
5. 甲斐充彦, 伊田政樹, 中川聖一, “大語彙連続音声認識のための音響的先読みによる高速化,” 日本音響学会 1997 年秋季研究発表会 講演論文集, 3-1-3, pp. 91–92, 1997 年 9 月
6. 伊田政樹, 森弘之, 中村哲, 鹿野清宏, “実騒音環境におけるハンズフリー単語音声認識,” 電子情報通信学会 音声研究会 研究技術報告, SP99-70, pp. 57–62, 1999 年 8 月
7. 森弘之, 伊田政樹, 中村哲, 鹿野清宏, “実騒音環境におけるマイクロホンアレーを用いた単語音声認識,” 日本音響学会 1999 年秋季研究発表会講演論文集, 1-1-21, pp. 41–42, 1999 年 9 月
8. 伊田政樹, 松井知子, 中村哲, “HMM 合成による環境音重畳音声の認識,” 日本音響学会 2000 年秋季研究発表会講演論文集, 2-5-9, pp. 67–68, 2000 年 9 月
9. 伊田政樹, 中村哲, “HMM 合成を用いた雑音環境下音声認識における環境音 GMM の適応化,” 情報処理学会 音声言語情報処理研究会 情報処理学会研究報告, 2001-SLP-37-12, pp. 67-72, 2001 年 7 月
10. 伊田政樹, 中村哲, “雑音 DB とモデル適応化を用いた HMM 合成法における雑音変動耐性の評価,” 日本音響学会 2001 年秋季研究発表会講演論文集, 1-1-7, pp. 33-34, 2001 年 10 月
11. 伊田政樹, 中村哲, “雑音 DB を用いたモデル適応化 HMM の SN 比別マルチパスモデルによる雑音下音声認識,” 電子情報通信学会 音声研究会 研究技術報告, SP2001-92, pp. 51–56, 2001 年 12 月

技術報告

1. 伊田政樹, 森弘之, “ハンズフリー音声認識技術の開発〜人にやさしい音声入力インタフェースの実現〜,” OMRON Technics, Vol. 40 No. 1, pp. 27–30, 2000 年 3 月
2. 西川憲一郎，伊田政樹，“展示会場の実データを用いた音声認識性能評価実験,” ATR テクニカルレポート, TR-S-0006, 2000 年 9 月

3. 梅田将満，伊田政樹，“マルチストリーム特徴量による雑音にロバストな音声認識,” ATR テクニカルレポート, TR-SLT-0004, 2002 年 2 月

特許

1. オムロン株式会社, 伊田政樹, “電子機器制御装置および電子機器,” 特願 2000-46856, 2000 年 2 月 24 日
2. オムロン株式会社, 伊田政樹, “音声認識装置および認識対象検出方法,” 特願 2000-75046, 2000 年 3 月 17 日
3. 株式会社エイ・ティ・アール音声言語通信研究所, 伊田政樹, 松井知子, 中村哲, “音響モデル生成装置及び音声認識装置,” 特願 2000-283516, 2000 年 9 月 19 日
4. 株式会社国際電気通信基礎技術研究所, 伊田政樹, 中村哲, “音響モデル生成装置及び音声認識装置,” 特願 2001-378546, 2001 年 12 月 12 日