NAIST-IS-DD0561207

博士論文

音響空間可視化手法を応用した効率的な音声コーパス構築フレームワーク

奈木野 豪秀

2008年9月24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。

奈木野 豪秀

審査委員：

鹿野 清宏 教授　　（主指導教員）
木戸出 正繼 教授　（副指導教員）
猿渡 洋 准教授　　（副指導教員）

音響空間可視化手法を応用した効率的な音声コーパス構築フレームワーク*

奈木野 豪秀

内容梗概

隠れマルコフモデルのような統計モデルをベースとした，統計的パターン認識では，統計モデルの技術要素の一つである，教師信号としてのデータベースの質がその後の性能を決めていると言っても過言ではない．音声認識の分野においては，周辺の環境（雑音や残響），利用者の年齢層や発話様式，発話内容等の，音声認識を利用する状況にその性能は大きく依存する（タスク依存性）．そのため，隠れマルコフモデルをベースとした，現在実用化されている音声認識アプリケーションの多くが，実際にアプリケーションを使用する実環境下で音声データを収集し，目的とするタスクに特化した音響モデル（隠れマルコフモデル）を作成することで，入力データとモデルとの整合性を保っている．しかしながら，実環境でのデータの収集は，データ収集システムの開発，運営にかかるコストや，収集後のデータ整理等のコストが膨大となり，企業での音声認識アプリケーションの開発コスト全体を圧迫している．

本論文は，以上の課題を踏まえ，低コストで効果的な音声コーパスを構築するためのフレームワークを実現することを目的としたものである．本論文では，まず，タスク間，タスク内の音響的変動や，実環境で発生する雑音と音声との違いを直感的に把握するために，音響モデルの分布を可視空間に写像することで，音響空間上でのデータ間の関連性やその拡がりを把握する手法（COSMOS 法）を

*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-DD0561207, 2008 年 9 月 24 日．

提案した．実験では，従来の可視化手法である主成分分析法や，SOM 法と比較し，明確にその優位性を示した．

次に，目的タスクの少量データを用い，既存音声コーパス群のタスクと目的タスクとの関係性を COSMOS 法により可視化することで，既存音声コーパス群の目的タスクに対する再利用性を視覚的に判定する手法を提案した．従来は，既存音声コーパス群から，最も音響的に特徴の近い音声データを選択することは出来ても，その再利用性が十分に高いかどうかを判定する基準がなく，選択後作成された音響モデルの性能を保証することが困難であったが，提案手法を用いることで，目的タスクと既存タスクとの COSMOS Map 上の分布の重なり具合から，直感的に再利用性を把握することが可能となることを実験的に示した．

次に，低コストで従来と同等以上の性能を実現する音声コーパス構築手法を提案した．提案に先立ち，高い認識性能を実現するためには，話者の多様性を確保することが重要であることから，まず，COSMOS 法により可視空間に写像された音響空間上の話者分布において，話者分布を囲うように，周辺部分に位置する話者を選択することが，多様性のある話者セットを構築することと等価であることを証明した．証明に基づき，収集対象の候補話者の少量音声データから，COSMOS 法を用いて音声認識性能向上に寄与する話者を予備選択し，選択された話者の音声データを収集することで，より低コストで効果的な音声コーパスを構築する手法を提案した．音声認識実験では，従来の無作為に話者を選択する手順と比較し，より高い性能を示すだけでなく，60%程度のコスト削減を実現する等，提案手法の有効性を示した．

キーワード

音響モデル，音声コーパス，タスク依存性，再利用性，データ収集コスト，可視化

The Framework of Building Effective Speech Corpus using Acoustic Spacce Visualizing Technique*

Goshu Nagino

Abstract

In using a statistical model such as a hidden Markov model (HMM) for pattern recognition, the quality of the training database is one of the most important issues for maximizing performance. In speech recognition, performance is still extremely sensitive to environmental conditions such as speaker characteristics and style, any background noise and the task domain. Together, these issues are called the task dependency. Practical Automatic Speech Recognition (ASR) applications using HMM as the acoustic model often collect the training speech corpus in the real environment where the application is to be used. Other times, this corpus is recorded in a recording booth and the real environment simulated by adding noise and echo. In both cases, in order to provide precise acoustic models for higher recognition performance, building a large-scale speech corpus is indispensable. However, doing so comes at great expense - required for developing the recording system, system maintenance, checking recorded speech data and labeling. These costs put enormous pressure on the development of any ASR application.

This paper describes the framework for building an effective speech corpus but with lower cost. First, in order to analyze task dependency, a method of visualizing acoustic models to grasp the acoustic space was proposed. The visualization technique named COSMOS method showed the relationship between several speech corpus and contrast

*Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0561207, September 24, 2008.

of speech and non-stational noise in the application environment. In comparing the proposed method with the Principal Component Analysis (PCA) and Self-Organizing Map (SOM) visualizing techniques, the effectiveness of the proposed COSMOS technique for analysis was shown.

After this, a technique to judge the reusability of existing speech corpora for a target task by using COSMOS method was proposed. This method showed a high correlation with the cross task recognition performance and judged the reusability of existing speech corpora correctly for the target task with a small quantity of speech data.

Lastly, a technique for building a speech corpus with lower cost by using the COSMOS method was proposed. At first, it was shown that speakers in the COSMOS map who are located in the periphery of and surround the distribution contribute most to improving speech recognition performance. In the proposed method, a small quantity of speech data is collected for each speaker; collected utterances are then used to train speaker-adapted acoustic models. Next, the models are mapped into two-dimensional space by using the COSMOS method. Speakers located in the periphery of the distribution - in the COSMOS map are selected and a full speech corpus is built using enough speech data for the selected speakers. As a result, the corpus built using this proposed method showed higher cost-effectiveness (about 60% reduction).

Keywords:

Acoustic model, Speech corpus, Task dependency, Reusability, Recording cost, Visualization

目 次

略語リスト（一般）

ASR	Automatic Speech Recognition
ATR	Advanced telecomunication research
CIAIR	Center for Integrated Acoustic Information Research
COSMOS	COmprehensive Space Map of Objective Signal
CW	Continuous Word
DTW	Dinamic Time Warping
EM	Expectation-maximization
HMM	Hidden Markov Model
IW	Isolated Word
MAPLR	Maximum A Posteriori - Linear Regression
MFCC	Mel Frequency Cepstrum Coefficient
ML	Maximum likelihood
MLLR	Maximum Likelihood Liner Regression
PCA	Principal Component Analysis
PDA	Personal Digital Assistants
SD	Speaker Dependent
SNR	Signal to Noise Ratio
SOM	Self-Organizing Map

略語リスト（音声コーパス関連）

ATR-APP	ATR-APPointment scheduling task
CIAIR-HCC	CIAIR-Human actions while Car Controlling
CSJ	Corpus of Spontaneous Japanese
JNAS	Japanese Newspaper Article Sentences
S-JNAS	Senior-JNAS

図 目 次

表 目 次

1. 緒言

1.1 はじめに

近年，インテリジェントなセンシングアプリケーションの中核となる技術として，統計的パターン認識技術が注目を浴びている．統計的パターン認識は，確率的な振る舞いを持つ任意の入力データに対しても，高精度にその入力データが属するクラスを識別する技術である [1, 2]．既に実用化されている分野として音声認識 [3, 4, 5, 6] や画像認識 [7, 8] などが挙げられるが，今後は様々なセンサーデバイスから入力されるデータのセンシングに有効な技術であると考えられている（例えば [10]）．とは言え，既に実用化されている音声認識アプリケーションでさえ，未だ多くの課題を残している．最も大きな課題は実環境における性能である．パーソナルコンピューター [11, 12] やカーナビゲーションシステム [17]，また，携帯電話 [18] や PDA[19] 等の組み込み機器 [13, 14, 15, 16] で実用化が進められている音声認識アプリケーションにおいては，「静かな環境で」，「アプリケーションが受け付けることのできる語彙の範囲内で」，「一般的な人が」，「丁寧に話せば」，ほぼ 100%に近い性能を示すことができる技術レベルにはあるが，実際にこのような恵まれた状況はありない．一般に，音声認識を利用する状況（タスク）と音響的に整合性の高い音響モデルを実装することで高い認識性能が見込むことができる．しかしながら，タスクには音声認識を利用する際の周辺の環境（雑音や残響），ドメイン（語彙），利用者の年齢層や発話様式等が含まれており，実環境ではこれらの要因が複雑に絡み合っているため，あらゆるタスクに対して高い認識性能を保証することは容易ではない．

過去の研究には，適応手法や雑音抑圧手法等（[20, 21, 22, 23, 24, 25]），それぞれの要因による影響を軽減するための手法も数多く提案され，性能は年々向上しているものの，複雑に絡み合ったこれらの要因を完全に消し去ることは非常に困難であり，未だ十分な性能とは言い難い．

隠れマルコフモデル（以下 HMM）のような統計モデルをベースとした，統計的パターン認識では，統計モデルの技術要素として，1）教師信号としてのデータベースの構築，2）データベースを確率的に表現する統計モデルの学習，3）入力

データとモデルとの照合，の 3 つが挙げられる．一般にデータベースは大規模であり，そこから学習されるモデルはデータベースの多様性を確率的に表現することが期待されるが，その性能は，結局のところ，教師信号であるデータベースの実環境に対する被覆性，具体的には，認識対象とする入力データとデータベースとの整合性にかかっており，あまりに実環境とかけ離れた特徴のデータをいくら集めても，実環境における入力データとモデルとの不整合により，高い性能を示すことは出来ない．いくら後段の学習プロセスや照合プロセス技術が優れていようとも，全て水泡に帰す．そのため，現在実用化されている音声認識アプリケーションの多くが，実際にアプリケーションを使用する実環境下で音声データを収集し，目的とするタスクに特化した音響モデル（統計モデル）を作成することで，入力データとモデルとの整合性を保っている．しかしながら，実環境でのデータの収集は，データ収集システムの開発，運営にかかるコストや，収集後のデータ整理等のコストが膨大となり，企業での音声認識アプリケーションの開発コスト全体を圧迫している（全体のコストに対する音声コーパス構築にかかるコストの割合に関しては 5 章参照）．コスト，工期が不確定であることを嫌う企業では，アプリケーションの使用状況を予め想定し，構築するデータベースにおける，収録語彙，話者数，発話数といった要件の設計を行い，設計した仕様に従い，収録室等での収録を行う（背景環境（雑音，残響等）に関しては別途、計算機上でシミュレートする）ことで，コスト，工期を管理することが多い．また，どちらの収録手順においても，データ収集システムの開発や，構築するデータベースの設計等には，音声認識の専門知識を持つ人間の経験則に頼るところが大きいのが現状であり，コスト同様，音声認識アプリケーション開発のボトルネックとなっている．

これまで，教師信号であるデータベースの構築に関する研究報告としては，主に実環境での，または実環境に近いデータベースを膨大なコストと時間をかけて収集した後，分析，モデリング，評価を通し，構築したデータベースの重要性を論じるものが多かった．実際に，構築されたデータベースからは，実環境特有の特徴（発話速度，発話の怠け等）を確認することができ，それらの知見は音声認識技術の発展に大きく貢献した（本論文ではこれらに関する詳細は省略する）[26, 27, 28, 29, 30, 31, 32, 33, 34]．一方，データベース構築の効率化，認識性能の

向上を考慮した収集プロセスに関する研究は非常に少ない．しかしながら，統計的パターン認識技術をコア技術として普及させ，実用化レベルにまで発展させるためには，データベース構築の効率化，認識性能の向上を考慮した収集プロセスの検討は不可避の課題と言える．

1.2 本論文の目的

本論文では，前節の課題を踏まえ，目的タスクにおける少量の音声データ（開発用データ，図 1.1 中の Development data）を用い，保持する既存の音声データベース（以下，音声コーパス）群に，目的タスクと合致した再利用性の高い音声データが存在するかを判断する手法 [66] を提案する．更に，新規に音声データを収集する場合におけるコストの低減を実現するために，予備収録という概念を取り入れた，より低コストでの音声データ収集手法 [62, 70] を提案する．これらの技術を導入することで図 1.1 に示すような，目的とするタスクに特化した音響モデルを低コストで作成できる枠組みを実現することができる．

まず，開発用データを用いて，既存音声コーパス群に目的タスクに合致した再利用性の高い音声データが存在するかを判断する．高い精度で再利用性の判定が可能となれば，再利用性の高い音声データが存在する場合には，目的タスクに合致した音声データを既存音声コーパスから選択することで目的タスクの音声コーパスを構築することができるため，音声データの収集コストは低く，従来よりも低いコストで目的のタスクに特化した音響モデルを作成することができる．なお，開発用データを用いて既存音声コーパスから目的タスクに合致した音声データを選択し，モデルを作成する手法（図 1.1 の Data Selection）としては，[47] が提案されている．再利用性の高い音声データが存在しないと判断された場合は，本論文で提案する，効率的な音声コーパスの構築手法を適用する．このように，図 1.1 に示す枠組みを実現することで，既存音声コーパスの目的タスクに対する再利用性の有無によらず，従来よりも低コストで目的タスクに特化した音響モデルを作成することが可能となる．

また，本論文では，上述の手法の提案に先立ち，音声コーパスまたは音響モデルが表現する多次元の音響空間の拡がりを直感的に把握する技術として，音響モ

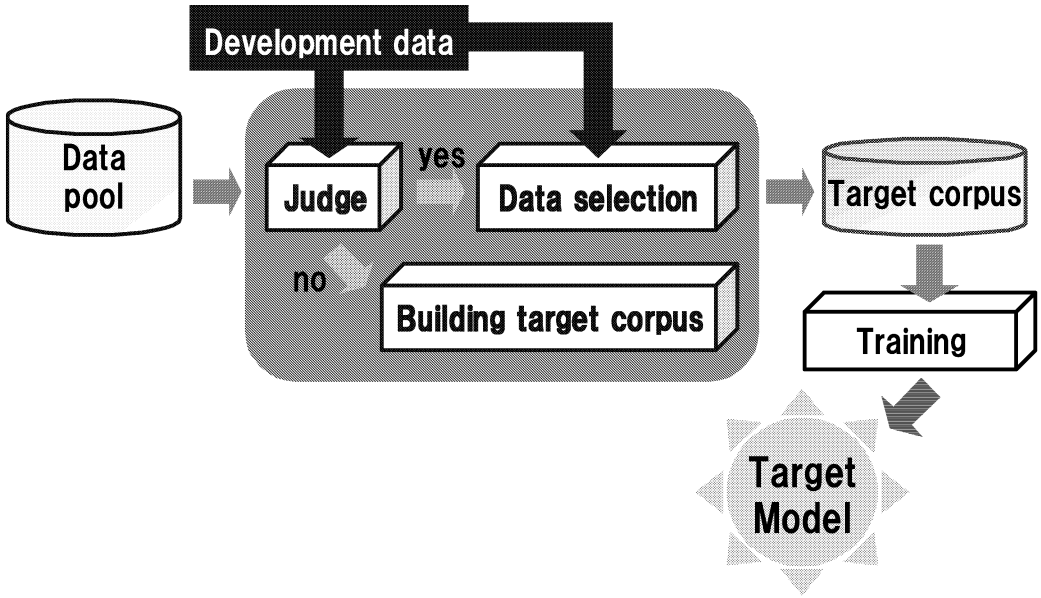


図 1.1 効率的な音声コーパス構築の流れ

デル群を可視空間に写像する手法を提案する [59, 61]．本手法により，保持する音声コーパス群（またはそこから作成される音響モデル群）の多様性やタスク間の関係性を直感的に把握することができる．可視化技術は，本来不可視である多次元データを，2 次，3 次といった可視空間に写像することで，多次元データの分布や構造を把握を可能とし，データ処理に有用な知識（アイデア）の抽出を補助する効果も有することから，更なる新規技術開発の促進や，音声認識の専門家の持つ知識，経験則のない人であっても，音声認識アプリケーションの開発に携わる機会を与える技術として期待できる．

1.3 本論文の構成

本論文の構成について説明する．本論文は本章を含めた 6 章から構成されている．まず第 2 章では，音声認識の基礎知識の導入として，音声認識の原理と構成及び，各要素技術の概要説明行う．また，本論文の重視する要素である音響モデル及び音響モデルを作成するもととなる音声コーパスについてその詳細を説明す

る．また，国内の大規模音声コーパスを用いたクロスタスクの音声認識実験を通し，タスク間の音響的不整合による生じる性能劣化を確認することで，タスク依存性の重要性の確認を行うと共に，タスク依存性の分析方法と課題について議論を行う．次に，第3章では，音声コーパスまたは音響モデルが表現する多次元の音響空間の拡がりを直感的に把握する技術として，音響モデル群を可視空間に写像する手法を提案し，その手法の詳細について説明する（図1.1のData Poolの分析）．なお，可視空間としては2次または3次が考えれるが，本論文をはじめ多くの表示媒体が2次元に従うものであることから，本論文では可視空間を2次元に限定し，議論を行うものとする．第4章では，提案する可視化手法を用いた，既存の音声コーパス群の目的タスクに対する再利用性の判定手法を提案し，その手法の詳細について説明する（図1.1のJudge）．第5章では，提案する可視化手法を用いた，新規音声コーパス構築における効率的な音声データ収集手法を提案し，その手法の詳細について説明する（図1.1のBuilding target corpus）．最後に，第6章で，本論文のまとめを行う．

2. 音声認識とタスク依存性

2.1 はじめに

本章では，音声認識の基礎知識の導入として，音声認識全般における概要を説明する [3, 4, 5, 6]．また，音声認識の実用化における課題の一つであるタスク依存性について議論を行う．まず，2.2 節で，音声認識の原理と構成を説明する．ここでは，音声認識の原理と，音声認識の構成要素である音響モデル，言語モデルの説明及び，音響モデルを用いた音声認識処理に関する説明を行う．次に，2.3 節では，本論文の議論の中心となる音響モデルの作成手順に関して具体的に説明を行う．作成手順の説明では，モデル作成時の具体的な学習アルゴリズムの紹介に加え，音響モデルを作成する材料とも言える音声コーパスの構築に関して，音声データの収集の一般的な手順に関する説明を行う．次に，2.4 節では，国内の大規模音声コーパスを用いたクロスタスクの音声認識実験を通し，タスク間の音響的不整合による生じる性能劣化を確認することで，タスク依存性の重要性の確認を行うと共に，タスク依存性の分析方法と課題について議論を行う．最後に 2.5 節で本章のまとめを行う．

2.2 音声認識の原理とシステム構成

音声認識は，入力音声の特徴ベクトルの時系列パターン $\mathbf{X}$ に対し，$\mathbf{X}$ を観測して最も尤度の高い単語列である $\mathbf{W}$ を探索する問題として考えることができる．ここで，S は単語数，T はフレーム数である．

$$\mathbf{X} = x_1, x_2, \cdots, x_T \tag{2.1}$$

$$\mathbf{W} = w_1, w_2, \cdots, w_S \tag{2.2}$$

これは，事後確率 $\mathcal{P}(\mathbf{W}|\mathbf{X})$ を最大にする単語列 $\mathbf{W}$ を探索する問題と捉えることが出来る．ここで $\mathcal{P}(\mathbf{W}|\mathbf{X})$ に対してベイズの定理を用いると次式を得ることがで

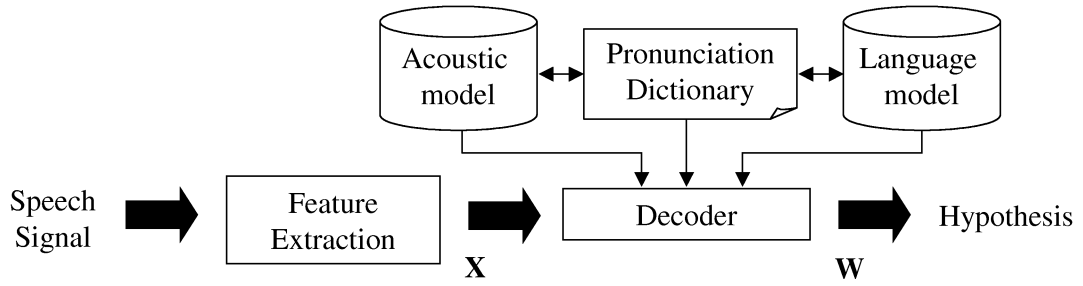


図 2.1 連続音声認識の流れ

きる．

$$\mathcal{P}(\mathbf{W}|\mathbf{X}) = \frac{\mathcal{P}(\mathbf{W}) \cdot \mathcal{P}(\mathbf{X}|\mathbf{W})}{\mathcal{P}(\mathbf{X})} \tag{2.3}$$

$\mathcal{P}(\mathbf{X})$ は，入力の特徴ベクトル系列の生起確率であり，単語列 $\mathbf{W}$ には無関係である．よって，音声認識は次式を最大にする $\mathbf{W}$ を求める問題と考えることができる．

$$\mathcal{P}(\mathbf{W}|\mathbf{X}) = \mathcal{P}(\mathbf{W}) \cdot \mathcal{P}(\mathbf{X}|\mathbf{W}) \tag{2.4}$$

式 (2.4) の $\mathcal{P}(\mathbf{W})$ は，単語列の事前確率であり，入力 $\mathbf{X}$ とは無関係な確率である．この単語列の出現確率を与えるモデルが言語モデルである．$\mathcal{P}(\mathbf{X}|\mathbf{W})$ は，単語列 $\mathbf{W}$ を発生したときに，特徴ベクトル系列 $\mathbf{X}$ が観測される確率で，この計算に用いるモデルは音響モデルと呼ばれる．

図 2.1 に音声認識の流れを図示する．音声認識は音響解析部（Feature Extraction）と照合処理部（Decoder）から構成される．音響解析部では入力された音声波形から短時間周波数分析によって特徴ベクトルが抽出される．特徴ベクトルとしては，MFCC（Mel Frequency Cepstrum Coefficient）を用いることが多く，本論文でもこの MFCC を用いる（特徴抽出に関する議論は，本論文の研究対象から外れるので，詳細な説明は [3, 4, 5, 6] に譲り，ここでは省略する.）．照合処理部は抽出された特徴ベクトルを入力として，言語モデル，音響モデル，辞書を用いた尤度計算により，入力音声に整合する最尤な単語列を探索し，テキストを出力する．

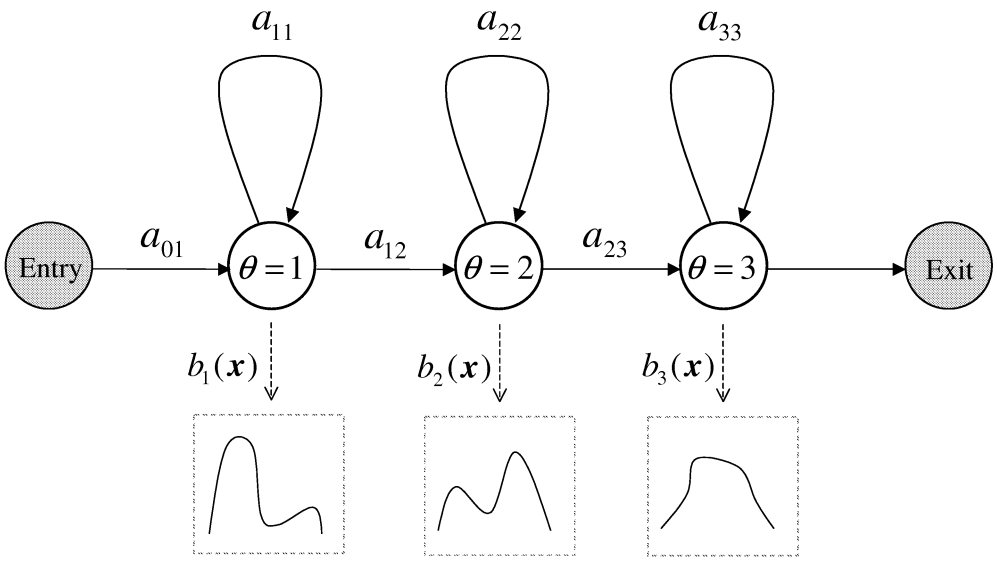


図 2.2 HMM 音響モデルの例

2.2.1 音響モデル

近年の音声認識では音響モデルに HMM（Hidden Markov Model : 隠れマルコフモデル）を適用したものが主流となっている．HMM は，観測可能な出力シンボルにより一意に状態遷移が決まらないという意味での非決定性有限状態オートマトンとして定義される．音声認識では，入力された音声（実際には特徴抽出後の特徴ベクトル系列）に対し，シンボル系列を状態遷移しながら生成する，音声の生成モデルとして扱われ，遷移する状態の集合，状態間の遷移確率，状態遷移の際のシンボル出力確率から構成される．また，HMM の形状には図 2.2 に示すような，初期状態と最終状態を持った，left-to-right 型をとることが一般的である．

図 2.2 において，状態 i から次の状態 j への遷移確率を a_{ij} と表し，自己遷移の確率と次の状態への遷移の確率の和は 1 となる．状態 i から次の状態 j への遷移の際，状態 j で特徴ベクトルを出力する出力確率密度を $b_j(\boldsymbol{x})$ と表し，出力確率密度関数としては多次元正規分布を用いることが多い．多次元正規分布は K 次の特徴ベクトル $\boldsymbol{x}$，平均ベクトル $\boldsymbol{\mu}$，と共分散行列 Σ を用いて以下で表される．な

お，′は転置を表す．

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{K/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{2.5}$$

また，出力確率の分布が一個の多次元正規分布で表すには複雑である場合には，M 個の多次元正規分布の重み付け和（重み ψ ここで $\sum_{m=1}^{M}\psi_m = 1$）で表すことが有効である．M 個の正規分布は混合正規分布と呼ばれる．ある時刻 t に観測された特徴ベクトルを生成する出力確率密度は混合正規分布を用いて次式で表される．

$$b(\boldsymbol{x}_t) = \sum_{m=1}^{M} \psi_m \mathcal{N}(\boldsymbol{x}_t,\boldsymbol{\mu}_m,\Sigma_m) \tag{2.6}$$

HMM は構成単位として，音素や単語を持つことができ，どの構成単位を用いるかは音声認識アプリケーションの仕様に依存する．例えば，0〜9 までの 10 個の数字を連続的に認識する場合や，数十単語のコマンドを認識する場合は，単語（数字）を構成単位として用いることが多い．但し，HMM の各状態は，定常性を仮定した短時間表現のモデルとなっているため，表現したい単位が時間的に長ければ，状態数は必然的に長くなる．大語彙の音声認識を目的とした場合，全ての単語について HMM を作成するのは非現実的であるため，音素を構成単位として用いることが多い（日本語の場合，音節を構成単位として用いる場合もある）．音素を構成単位とすることで，複数単語間で共通の HMM を用いることができ，各単語は音素の連結として表現されることとなる．日本語の音素は，母音，無声破裂音，有声破裂音，無声摩擦音，有声摩擦音，鼻子音，流音，半母音など 20 数種類の音素から構成されており，拗音や無声化母音，調音結合の強い連続母音等を考慮すると 50 数種類となる．これらの音素に対し HMM を定義したモデルは，monophopne HMM と呼ばれる．monophopne HMM を拡張した高精度な HMM として，音素の前後環境を考慮した biphopne HMM や triphopne HMM がある（biphopne HMM は前もしくは後ろの音素環境のみを考慮）．triphopne HMM の場合，HMM の種類が膨大な数になり，学習データが不足することや，triphopne が学習データに出現しないといった問題が生じるため，異なった HMM 間でパラメータを共有することが行なわれる．まず，同一の中心音素をもつ triphopne HMM に分類する．次に，分類した triphopne HMM の各状態をクラスタリングし，クラスタリングされ

た状態間でHMMのパラメータを共有する．このようにして構築されたモデルを状態共有型triphopne HMMと呼ぶ[37]．また，異なるモデル，状態間で混合正規分布のための正規分布を共有することで，さらに効率的なモデルを構成することが出来る．特に，全てのモデル間で同一の数百から千程度の正規分布を共有するHMMをTied Mixture型と呼び，このTied Mixture型の中で，特に中心音素が同一であるtriphopneの間だけで分布の共有を行う方法を音素内Tied Mixtureモデル（PTMモデル）[38]と呼ぶ．

2.2.2 言語モデル

現在連続音声認識において広く用いられている言語モデルとして，正規文法と統計的言語モデルであるN-gram言語モデルがある．どちらの言語モデルを用いるかは音声認識アプリケーションの仕様に依存し，一般に，孤立単語認識及び，小，中語彙（数〜数百単語）の連続音声認識には正規文法を用い，数千，数万単語以上の連続音声認識にはN-gram言語モデルを用いることが多い．

正規文法はもともと文を生成するモデルとして作られたものである．正規文法は，文法が生成可能な文に従う発声（文法が受理可能な発声）に対して高い認識精度を示すことが知られている．コマンド操作，対話システムのような，タスクが設定されている状況においては，音声認識アプリケーションの使用者が用いる語彙・言い回しを予め予測することが比較的容易であるため，言語モデルに正規文法が用いられることが多い．また，意味理解などを行う対話制御の観点からも，対話制御が求める知識情報や制約を認識処理に反映させることでき，対話制御の効率化につながる．

N-gram言語モデルは語彙サイズが数万，数十万といった，大語彙での連続音声認識に広く用いられている言語モデルで，文（単語列）を単語・品詞の複数組の生起確率で近似するモデルである．大量の学習コーパスが存在すれば，半自動的に作成することができる．$N = 3$の場合をトライグラム言語モデルと呼ぶ．特に，学習データに現れなかったN-gramに対して$(N-1)$-gram確率から推定するバックオフ平滑化手法をとるものを，バックオフトライグラム言語モデルと呼ぶ．

言語モデルに関する議論は，本論文の研究対象から外れるので，詳細な説明は

[3, 4, 5, 6] に譲り，ここでは省略する．また，本論文で行う音声認識実験では正規文法のみを使用している．

2.2.3 HMM による音声認識

モデル Λ が観測系列である T フレームの特徴ベクトル系列 $\mathbf{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ を生成する確率 $\mathcal{F}(\mathbf{X}|\Lambda)$（以下，尤度と呼ぶ.）を求めるには，まず長さ T の全ての状態系列に対して，確率の計算を行なうことが考えられる．この場合，各時刻に対し，ありうる状態が N 個あり，かつ各状態系列において $2T$ の計算が必要となるため，計算量は $O(2T \cdot N^T)$ となり現実的ではない．計算量を削減した実用的なアルゴリズムとして Forward アルゴリズムがある。

時刻 t に観測系列 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ を出力し，状態 j にいる確率を次のように定義する．

$$\alpha_t(j) = \mathcal{F}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \theta_t = j|\Lambda) \tag{2.7}$$

$\mathcal{F}(\mathbf{X}|\Lambda)$ は $\alpha_t(j)$ の漸化式を次のような手順で計算することによって求めることができる．

1. 初期化

   全ての状態 j に対して初期状態確率 π を与える．

   $$\alpha_0(j) = \pi_j \tag{2.8}$$

2. 導出過程

   各時刻 t における全ての状態 j に対して $\alpha_t(j)$ を求める．

   $$\alpha_t(j) = \sum_{j=1}^{N} \alpha_{t-1}(i) a_{ij} b_{ij}(\boldsymbol{x}_t) \tag{2.9}$$

3. 結果

   $$\mathcal{F}(\mathbf{X}|\Lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{2.10}$$

Forward アルゴリズムでは，直前のフレームにおける確率 $\alpha_{t-1}(i)$ から $\alpha_t(i)$ を求めている．$\alpha_t(i)$ を前向き確率と呼ぶ．なお，Forward アルゴリズムでの計算量は

$O(N^2T)$ となる.

Forward アルゴリズムが，前向き確率を初期状態から前向きに計算するのに対し，モデル Λ において，時刻 t に状態 j に停留し，時刻 $t+1$ から観測系列 $\boldsymbol{x}_t+1$, $\ldots$,$\boldsymbol{x}_T$ を出力する確率を後向き確率と呼び，最終状態から後向きに計算を行うこの方法を Backward アルゴリズムと呼ぶ．後向き確率 $\beta_t(i)$ を次のように表す.

$$\beta_t(i) = \mathcal{F}(\boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_T, \theta_t = i|\Lambda) \tag{2.11}$$

$\beta_t(i)$ は次の手順で求めることができる.

1. 初期化

   全ての状態 i に対して確率 1 を与える.

   $$\beta_T(i) = 1 \tag{2.12}$$

2. 導出過程

   各時刻 t における全ての状態 i に対して $\beta_t(i)$ を後向きに求める.

   $$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_{ij}(\boldsymbol{x}_t) \beta_{t+1}(j) \tag{2.13}$$

3. 結果

   $$\mathcal{F}(\mathbf{X}|\Lambda) = \sum_{i=1}^{N} \pi\beta_0(i) \tag{2.14}$$

尤度 $\mathcal{F}(\mathbf{X}|\Lambda)$ は，前述の Forward アルゴリズムにより効率的に求めることができるが，ある時刻 t における状態 θ_t の確率として，その状態にたどり着く全ての状態系列の確率の和を用いる Forward アルゴリズムに対し，最も確率の高い状態系列の確率を用いる Viterbi アルゴリズムが提案されている.

モデル Λ が観測系列である T フレームの特徴ベクトル系列 $\mathbf{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ に対する最尤な状態系列 $\Theta = \theta_1, \ldots, \theta_T$ ををを求めるために，時刻 t で状態 i に至るまでの最尤状態確率を $\varrho_t(i)$ を定義する.

$$\varrho_t(i) = \max_{\theta_1, \ldots, \theta_{t-1}} \mathcal{F}(\theta_1, \ldots, \theta_t = i, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_t|\Lambda) \tag{2.15}$$

時刻 $t+1$ における最尤状態の確率は次式となる.

$$\varrho_{t+1}(i) = \max_i[\varrho_t(i)a_{ij}]b_{ij}(\boldsymbol{x}_{t+1}) \tag{2.16}$$

時刻 t，状態 i において，生成確率を最大にする経路（最尤経路）を記憶するための遷移元の状態を記憶するバックポインタを $\varpi_t(i)$，最尤経路の生成確率を $\check{\mathcal{F}}$ 最尤経路上の最終状態を $\check{s}_T$ とすると，最尤経路およびその生成確率は以下の手順で求めることができる.

1. 初期化

$$\varrho_0(i) = \pi_i \tag{2.17}$$

$$\varpi_0(i) = 0 \tag{2.18}$$

2. 導出過程

$$\varrho_t(j) = \max_{i=1,\ldots,N}[\varrho_{t-1}(i)a_{ij}b_{ij}(\boldsymbol{x}_t)] \tag{2.19}$$

$$\varpi_t(j) = \operatorname*{argmax}_{i=1,\ldots,N}[\varrho_{t-1}(i)a_{ij}b_{ij}(\boldsymbol{x}_t)] \tag{2.20}$$

3. 結果

$$\check{\mathcal{F}} = \max_{i=1,\ldots,N}[\varrho_T(i)] \tag{2.21}$$

$$\check{s}_T = \operatorname*{argmax}_{i=1,\ldots,N}[\varrho_T(i)] \tag{2.22}$$

4. バックトラック

バックポインタから時間 t に沿って，逆順に状態系列を求める.

$$\check{s}_t = \varpi_{t+1}(\check{s}_{t+1}) \tag{2.23}$$

Viterbi アルゴリズムは Forward アルゴリズムの近似手法であり，計算量が少なく，性能も同程度であることが知られており，実際の音声認識アプリケーションで用いられることが多い．また，確率の積を直接に計算するときには，演算のアンダーフローが生じる恐れがあるため，式 (2.19) において，確率の積の代わりに対数確率（対数尤度）の和を用いることが一般的である.

2.3 音響モデルの作成

統計モデルであるHMM音響モデルを高精度に作成するためには，学習データである音声コーパスの構築，モデルパラメータの推定を行う必要がある．モデルパラメータの推定手法が優れていようとも，目的タスクの環境下での入力データに対し，音響的に不整合な学習データでは高精度な音響モデルを作成できない．また，目的タスクの環境下での入力データに対し，音響的に整合性のとれた学習データを大量に保有していたとしても，モデルパラメータの推定（モデル構造の決定も含む）が上手くいかなければ，同様に高精度な音響モデルを作成できない．

本節では，音響モデルの作成工程が逆にはなるが，まず，モデルパラメータの推定に関する紹介を行い，本論文の課題である音声コーパスの構築に関する議論を後に行う．

2.3.1 モデルパラメータの推定

一般に，音響モデルのパラメータ推定の基準としては，ML（Maximum Likelihood：尤度最大）基準がよく用いられる．ML基準による音響モデルのパラメータ推定は，学習データの特徴ベクトル系列に対して尤度$\mathcal{F}(\mathbf{X})$が最大となるようパラメータを推定するものである（以下最尤推定と呼ぶ）．但し，HMMのパラメータは，状態遷移系列が非観測であることから，直接，最尤推定するこができない．このため，EMアルゴリズム[9]用いた繰り返しアルゴリズムにより，パラメータΛを推定する．

観測系列であるTフレームの特徴ベクトル系列$\mathbf{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$を生成する全ての状態系列の中で，フレーム$t$で状態$i$から状態$j$に遷移する回数の期待値を求め，これを全体の確率で正規化した生起確率の期待値$\gamma_t(i, j)$を次のように定義する．

$$\gamma_t(i, j) = \frac{\alpha_{t-1}(i) a_{ij} b_{ij}(\boldsymbol{x}_t) \beta_t(j)}{\mathcal{F}(\mathbf{X}, \Lambda)} \tag{2.24}$$

EMアルゴリズムにより，HMMの各パラメータの再推定値はそれぞれの条件付きの相対確率で表される．

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T} \alpha_{t-1}(i) a_{ij} b_{ij}(\boldsymbol{x}_t) \beta_t(j)}{\sum_{t=1}^{T} \alpha_t(i) \beta_t(i)} \tag{2.25}$$

$$\hat{\boldsymbol{\mu}}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j) \boldsymbol{x}_t}{\sum_{t=1}^{T} \gamma_t(i,j)} \tag{2.26}$$

$$\hat{\Sigma}_{ij} = \frac{\sum_{t=1}^{T} \gamma_t(i,j) (\boldsymbol{x}_t - \boldsymbol{\mu}_{ij})(\boldsymbol{x}_t - \boldsymbol{\mu}_{ij})'}{\sum_{t=1}^{T} \gamma_t(i,j)} \tag{2.27}$$

複数の正規分布（M 個）とその重み ψ で表現される混合正規分布による出力確率を考えた場合の m 番目の分布の各モデルパラメータは次のようになる．

$$\gamma_t(i,j,m) = \frac{\alpha_{t-1}(i) a_{ij} \cdot \psi_{ijm} b_{ijm}(\boldsymbol{x}_t) \beta_t(j)}{\mathcal{F}(\mathbf{X}, \Lambda)} \tag{2.28}$$

$$\hat{\psi}_{ijm} = \frac{\sum_{t=1}^{T} \gamma_t(i,j,m)}{\sum_{t=1}^{T} \gamma_t(i,j)} \tag{2.29}$$

$$\hat{\boldsymbol{\mu}}_{ijm} = \frac{\sum_{t=1}^{T} \gamma_t(i,j,m) \boldsymbol{x}_t}{\sum_{t=1}^{T} \gamma_t(i,j,m)} \tag{2.30}$$

$$\hat{\Sigma}_{ijm} = \frac{\sum_{t=1}^{T} \gamma_t(i,j,m) (\boldsymbol{x}_t - \boldsymbol{\mu}_{ijm})(\boldsymbol{x}_t - \boldsymbol{\mu}_{ijm})'}{\sum_{t=1}^{T} \gamma_t(i,j,m)} \tag{2.31}$$

パラメータの更新により，常にが次式が成り立つことが証明されており，更新を繰り返すことで，学習データに対しより頑健なモデルを構築することができる．

$$\mathcal{F}(\mathbf{X}, \Lambda) \leq \mathcal{F}(\mathbf{X}, \hat{\Lambda}) \tag{2.32}$$

なお，実際の学習時は学習回数を予め設定しておくか，閾値を設定することで，更新を終了している．

ML 基準によるモデルパラメータ推定は，推定に十分な量の学習データが存在することを前提としており，学習データが不十分のまま学習を進めると，モデルパラメータの推定精度が低下するばかりでなく，偶発的に発生した特異なデータの影響を受け，本来推定すべきモデルとはかけ離れたモデルが作成される恐れがある．また，学習データの音響的な挙動が最も適切に表現されるようにモデルパラメータは推定されるが，パターン認識の基本原理である，他のクラス（HMM）

との識別性に関しては，一切考慮されていない．これらの課題から，近年ではML基準を改良したモデルパラメータ推定手法 [39, 40] や，基準の全く異なる，ベイズ的基準 [41, 42] や最小識別誤り基準 [43, 44, 45] によるモデルパラメータ推定手法も提案されている．

2.3.2 音声コーパスの構築

音響モデルの作成には元となる学習用の音声コーパスが必要となる．一般に音声コーパスは大規模であり，そこから学習される音響モデルは音声コーパスの多様性を確率的に表現することが期待されるが，音響モデルの性能は，教師信号である音声コーパスの出来不出来，つまりは，目的とするタスクにおける入力音声データと音声コーパスとの整合性にかかっている．学習データ量が十分に存在し，優れた学習プロセスを適用したとしても，あまりにかけ離れた特徴の音声データからは，入力音声データに対して整合性の高いモデルを作成することは出来ず，高い性能を期待することが出来ない．アプリケーションを使用する状況下での音声データと整合性のとれた学習用音声コーパスを構築することが重要である．

収録は一般に，実際にアプリケーションを使用しながらアプリケーション利用者の発話を収録する手順（収録手順Ⅰ）と，発話内容を収録する側が予め決定しておき，募集した発話者による発話リストの読み上げ音声を防音室等で収録する手順（収録手順Ⅱ）の二つに大別される．前者は，収録環境が実際に音声認識アプリケーションを使用する状況に近いため，その環境で収録された音声コーパスを用いて学習された音響モデルは音響的にも整合性が高く，より高い性能を期待することができる．これに対し，後者の手順では，収録した音声に対し，実際に音声認識アプリケーションを使用する状況下における，雑音環境や残響環境といった背景環境をシミュレートすることで，音響的な整合性を高めている．音響的な整合性は前者の手順に劣るものの，音声データの収集期間，工数の見積もりが明確であるという利点がある．前者及び後者の利点を考慮した収録手順として，募集した発話者に収録室もしくはそれに準ずる静かな環境で，実際に音声認識アプリケーションを使用させたり，それを模擬するような発話を促すことで，背景環境以外の音響的な特徴に関して，整合性の高い音声データの収録を行う場合もあ

表 2.1 音響解析条件

	Setting 1	Setting 2
Sampling frequency	11.025 kHz	16 kHz
Feature parameter	10 MFCC +10 ΔMFCC +ΔLogPower	12 MFCC +12 ΔMFCC +ΔLogPower
Pre-mphasis	$1 - 0.97z^{-1}$	
Frame length	20 ms	25 ms
Frame shift	10 ms	
Window type	Hamming	

る（収録手順Ⅲ）．但し，いずれの手順においても，音声コーパスの構築には膨大な時間とコストがかかり，音声認識アプリケーション開発におけるボトルネックとなっている．

現存する国内の大規模音声コーパスは全て，大別すると，これらの手順に従い収録されている．但し，コスト，工期が不確定であることを嫌う企業では，収録手順Ⅱもしくは Ⅲにより音声データを収集することが多い．本論文で取り扱う音声コーパスの内，次節及び 4 章，5 章で用いる音声コーパスの概要及び収録手順，音響解析条件を付録 A に示す．なお，本論文で用いる音響解析条件を表 2.1 に示す．

2.4 タスク依存性

緒言で述べたように，タスクには音声認識を利用する際の周辺の環境（雑音や残響），ドメイン（語彙），利用者の年齢層や発話様式等が含まれており，実環境ではこれらの要因が複雑に絡み合っているため，あらゆるタスクに対して高い認識性能を保証することは容易ではない．タスクが異なることで，これらの音響的要因が変動し，音響的な不整合が生じてしまう．本章では，タスクに依存した音響的変動，性能の変動をタスク依存性と呼ぶ．近年，複数のタスクを用いたクロスタスクの音声認識実験，タスク間の移植性（Portability）の調査，物理量分析

による性能劣化要因（タスクの難易度を表す要因）の分析等，タスク依存性に関する報告が増えており [33, 34, 36, 46, 47]，タスク依存性に関する関心の高さ，重要性が伺える．

本節では，複数のタスクの音声コーパスを用い，クロスタスクの音声認識実験を行うことで，実際にタスク依存性の確認を行う．

2.4.1 実験用音声コーパス概要

本章ではタスク依存性の分析に，国内でよく知られた 5 つの大規模日本語音声コーパスを用いる．用いる音声コーパスは，ATR-APP[27]，JNAS[28]，S-JNAS[29]，CIAIR-HCC[30, 31]，CSJ[32, 33, 34] であり，それぞれ，研究開発を目的とした音声コーパスや，実際のアプリケーションを想定した音声コーパス等，その目的は多岐にわたる（詳細は付録 A 参照）．各音声コーパスは更に複数のタスクに分類されている．本論文では各音声コーパスから計 12 個のタスクを抽出し，実験に用いる．ATR-APP からは，バランス文読み上げタスク（ATR-APP_balance）の 1 タスクのみを用いる．JNAS からは，バランス文読み上げタスク（JNAS_balance），新聞記事読み上げタスク（JNAS_news）の 2 タスクを用いる．S-JNAS からは，バランス文読み上げタスク（S-JNAS_balance），新聞記事読み上げタスク（S-JNAS_news），情報検索文読み上げタスク（S-JNAS_infoseek）の 3 タスクを用いる．CIAIR-HCC からは，バランス文読み上げタスク（CIAIR-drive_balance），人対 ASR の対話タスク（CIAIR-drive_dialogA），人対 WOZ の対話タスク（CIAIR-drive_dialogW），人対人の対話タスク（CIAIR-drive_dialogH）の 4 タスクを用いる．CSJ からは，学会講演タスク（CSJ_academic），模擬講演タスク（CSJ_speech）の 2 タスクを用いる．

後述の音声認識実験のために，各タスクの話者を学習話者と評価話者に分割する．本論文では学習及び評価ともに男性話者のみを用いるものとする．評価話者は，タスク毎に男性 15 名とし，この 15 名は無作為に選択される．実験に用いる音声コーパスには，一人の話者が複数のタスクを発声している場合があるため，あるタスクで評価セットとして選択された話者は他のタスクの学習データからは除外する．各タスクの延べ話者数とデータサイズを表 2.2 に示す．なお，表 2.2 に

表 2.2 各タスクのデータサイズ（単位は時間 [h]）

Task	(ID)	Number of speakers train/total	Data size train/total
ATR-APP_balance	(AAb)	1364/1379 males	45.7/46.2
JNAS_balance	(Jb)	124/151 males	6.9/8.5
JNAS_news	(Jn)	124/151 males	18.1/22.4
S-JNAS_balance	(SJb)	126/151 males	18.7/22.1
S-JNAS_news	(SJn)	161/202 males	31.0/38.6
S-JNAS_infoseek	(SJi)	34/51 males	2.5/3.6
CIAIR-drive_balance	(Cdb)	261/314 males	5.1/6.1
CIAIR-drive_dialogA	(CdA)	245/297 males	1.9/2.3
CIAIR-drive_dialogW	(CdW)	247/298 males	3.2/3.8
CIAIR-drive_dialogH	(CdH)	258/310 males	4.8/5.7
CSJ_academic	(Ca)	785/804 males	70.8/72.9
CSJ_speech	(Cs)	774/805 males	57.6/62.7

示されているデータサイズは Viterbi アライメントによる発話区間検出結果から得られた発話区間の長さをもとに算出されたものである．また，CIAIR の全タスクに関しては有効発話のタグがつけられた発声のみ使用し，CSJ の両タスクに関しては「笑い」や「咳」といった特別なタグが付けられた発声は除外している．

2.4.2 クロスタスクの音声認識実験

タスク依存性には，そのタスクの難しさ及び，他のタスクとの関係性の二つの観点がある．タスクの難しさ（以下，タスク難易度）は，あるタスクの評価データに対し，タスクが一致している音響モデルを用いた場合（Closed 評価）での性能であり，性能が低い程タスク難易度は高いと定義される．他のタスクとの関係性は，ある目的タスクの評価データに対する他のタスクの音響モデルを用いた場合での性能であり，目的タスクの Closed 評価における性能と比較し，性能が低い

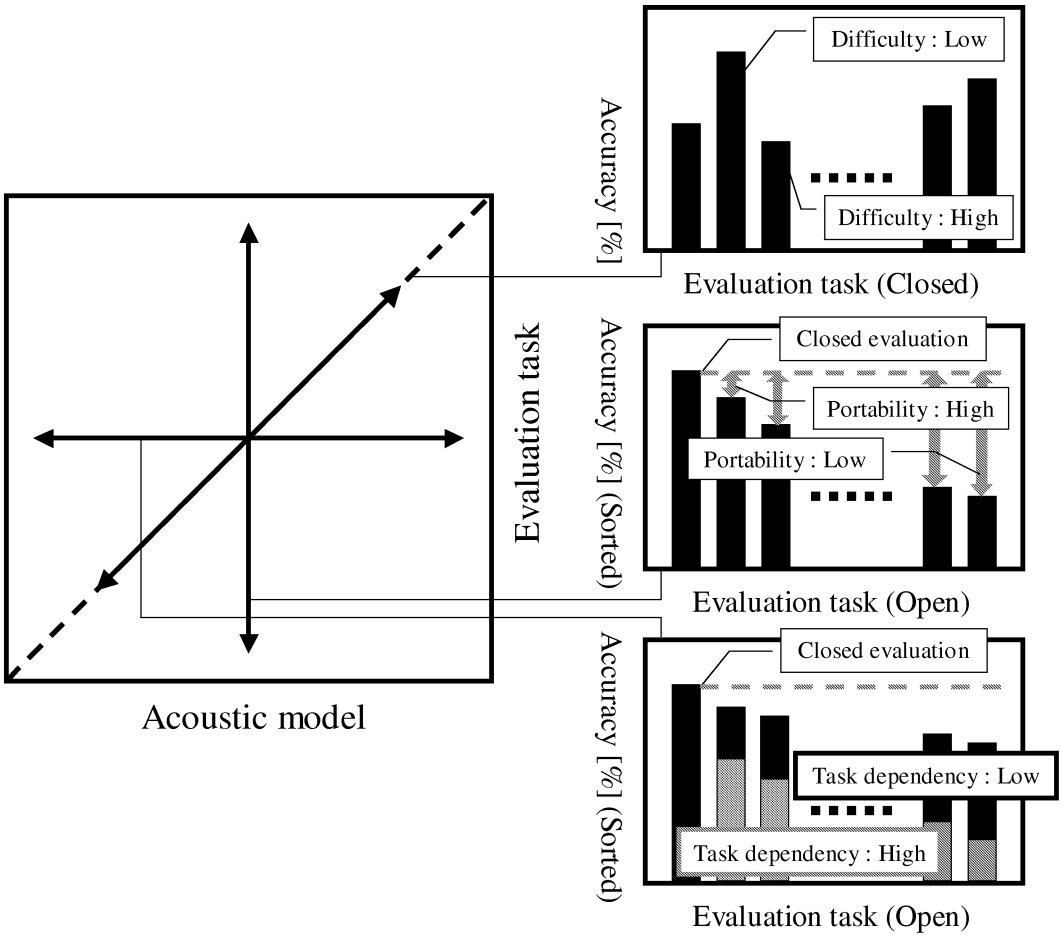


図 2.3　クロスタスクによる音声認識実験結果から読み取るタスク依存性

程，その目的タスクへの移植性（Portability）が低いと定義される．また，目的タスクの Closed 評価における性能と比較し，他のどのタスクの音響モデルを用いても十分な性能を示すことができない場合，その目的タスクは他のタスクと音響的に大きく異なり，独立性が高いことから，タスク依存性が強いと定義される．図 2.3 に，クロスタスクの音声認識実験結果から読み取ることの出来る，タスク難易度，他のタスクとの関連性（他のタスクに対する移植性），タスク依存性の例を示す．本節では，前節で紹介した 12 個のタスクの音声データを用い，クロスタスクの音声認識実験を行い，タスク難易度や他のタスクとの関連性，タスク依

存性の強さに関する分析を行う．

各タスクの学習データを用いて学習された 12 個のタスク依存音響モデルに対して，各タスクの評価セットを用い，クロスタスクの音素認識実験を行う．音素認識実験では，日本語の音素配列構造の制約を用いた音素タイプライタを用いるものとする．また，音響モデルのモデル構造がタスク依存性に与える影響を調べるために，本論文では音響モデルに monophone HMM 及び triphone HMM（状態共有型）の両方を用いて評価を行う．音素は 43 種類とする．全ての音素において状態数を 3 とする．monophone HMM では各状態における混合分布数を 8 とし，状態共有型 triphone HMM は総状態数を 2000，各状態における混合分布数は 8 とする．また，各正規分布の共分散行列は対角共分散行列とする．

monophone HMM 及び triphone HMM によるクロスタスクの音素認識実験結果を表 2.3 に示す．また，表 2.3 の結果を等高線として表した結果を図 2.4 に示す．図 2.4 において，横軸はタスク依存音響モデルの種別，縦軸は評価タスクの種別を表す．また，図中のタスク名は省略名（表 2.2 中の”ID”を参照）である．表 2.3 及び図 2.4 の認識実験結果より，モデル構造の違いや，学習データ量によらず，全てのタスクでタスクが一致している場合の性能が最も高いことがわかる．タスク依存性の存在とその重要性を確認することができる．また，monophone HMM と比較し，triphone HMM は，よりタスク依存性が強くなる傾向にあることがわかる．triphone HMM が前後の音素環境を考慮したモデル構造として定義されているため，音素環境の偏りの影響が顕著に現れることが，タスク依存性が強くなる要因の一つとして考えられる．また，モデルパラメータが増え，音響モデルがより精密化されることで，話者性や発話様式等の音素環境以外の音響的特徴もより精密に表現されることもタスク依存性が強くなる要因の一つと考えられる．反対に，タスク依存性が強くなることで，各タスクの他のタスクに対する移植性は低下する傾向にあることがわかる．

各タスクにおけるタスク依存性の議論に関しては，音素バランス文読み上げタスクや新聞記事読み上げタスクが，比較的タスク難易度が低く，他のタスクに対する移植性が高い結果となっていることや，対話タスクや自然発話タスクが，比較的難易度が高く，他のタスクに対する移植性が低い結果となっていることなど，

表 2.3 クロスタスクによる音素認識実験結果（上：monophone，下：triphone）*

Acoustic model	Evaluation task											
	AAb	Jb	Jn	SJb	SJn	SJi	Cdb	CdA	CdW	CdH	Ca	Cs
AAb	67.26	60.3	60.27	46.21	46.59	48.36	57.63	53.38	47.25	46.58	52.36	47.59
Jb	62.59	65.89	64.88	50.71	52.5	50.99	62.24	53.96	47.36	48.01	50.43	48
Jn	61.21	64.11	65.44	50.01	53	51.1	60.51	54.03	48	48.48	50.54	48.28
SJb	58.74	59.91	61.42	57.8	60.06	59.27	58.22	52.9	45.75	45.96	48.53	48.01
SJn	57.99	59.03	61.75	56.66	60.57	58.35	56.65	52.13	46.04	46.26	48.2	48.16
SJi	57.58	59.69	60.32	56.1	58.07	67.39	59.04	56.16	48.73	48.6	47.8	48.76
Cdb	55.36	61.4	58.54	42.47	43.94	48.04	69.66	57.71	49.42	48.48	48.88	44.96
CdA	46.02	47.09	46.22	27.18	28.73	31.93	55.02	65.64	53.1	51.52	44.28	35.75
CdW	43.08	44.83	44.91	26.34	27.45	32.65	52.05	58.36	52.65	51.49	43.62	34.9
CdH	44.49	45.92	45.87	27.15	28.51	32.65	52.93	57.5	52.16	51.84	44.96	36.48
Ca	52.92	49.83	51.13	33.54	35.15	38.97	50.95	50.45	46.01	44.94	55.47	44.83
Cs	55.18	53.44	54.8	41.07	43.4	45.93	49.68	50.88	44.45	45.86	51.01	50.83

Acoustic model	Evaluation task											
	AAb	Jb	Jn	SJb	SJn	SJi	Cdb	CdA	CdW	CdH	Ca	Cs
AAb	78.05	68.23	63.88	43.17	39.5	46.26	59.96	47.95	46.3	45.91	57.61	48.2
Jb	69.83	73.86	68.96	50.14	49.4	52.3	66.4	51.22	49.58	49.78	57.16	50.59
Jn	65.88	70.24	69.95	45.44	46.81	49.95	62.61	50.52	48.16	48.24	56.34	48.82
SJb	65.98	69.21	67.76	62.35	61.25	64.13	62.7	50.15	48.3	47.73	54.48	51.72
SJn	65.05	68.18	68.48	61.25	63.34	63.6	62.19	51.66	47.61	47.61	55.12	52.73
SJi	46.63	48.85	51.07	44.6	45.99	73.47	43.36	46.07	41.8	40.89	41.35	43.75
Cdb	61.48	66.03	59.95	42.43	40.99	52.81	75.15	57.83	52.07	52.49	53.93	48.45
CdA	30.67	29.61	29.37	10.19	10.98	16.01	37.98	64.29	53.12	50.53	33.29	22.81
CdW	34.2	32.76	33.61	13.1	14.55	23.28	40.99	58.61	58.65	55.25	39.49	28.69
CdH	36.32	36.68	37.8	14.46	16.29	25.56	45.06	57.75	57.51	57.93	42.84	32.5
Ca	61.45	58.4	54.93	31.03	31.26	38.3	52.71	49.26	45.81	47.17	63.85	48.47
Cs	63.45	62.51	60.1	39.96	40.59	47.26	54.14	49.84	47.35	49.28	59.9	56.2

様々な傾向を見て取ることができる．

このように，タスクの違いによる性能変動は小さくなく，音響モデルや音声コーパスの再利用などを考えた場合に，非常に大きな問題となることが伺える．音響的な不整合による性能劣化を事前に防ぐためには，タスク依存性の分析を行い，予め目的とするタスクの特性や既存タスクとの関連性を知っておくことが重要と言える．

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

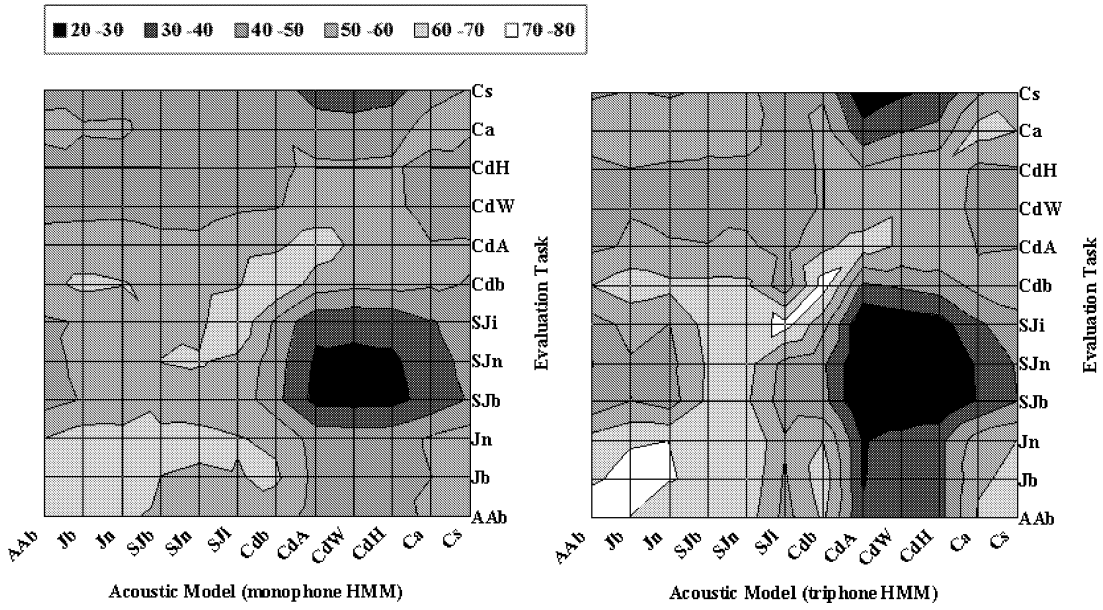


図 2.4 クロスタスクによる音素認識実験結果（等高線）

2.4.3 タスク依存性の分析と課題

音声認識アプリケーションを使用する状況（タスク）に依存し，音響的特徴は異なるため，異なるタスク間での音声コーパス，音響モデルを安易に流用すると，音響的な不整合により，大幅な性能劣化を起こす場合がある（2.4.2 節のクロスタスクの音素認識実験で確認）．タスクが異なることによる，音響的な変動を，各音響的変動要因を表す物理量で表現することが可能であり，それぞれの物理量は，大小はあるが，認識性能と相関を持つことが分かっている（文献 [27, 29, 31, 32, 33, 34, 36] 及び付録 B 参照）．将来的には，保持する音声コーパスの物理量をデータベース化することで，音声コーパス間，またはそこから作成される音響モデル間の整合・不整合といった相関関係を把握することができると期待される．しかしながら，実環境においては，様々な音響的変動要因が複雑に絡み合っており，現在定義できる物理量のみで表現することが可能とは言い難い．また，目的タスクにおける音声認識性能には，タスク内での音響的変動（話者，発話様式等）や，実環境で発生する雑音（誤動作を引き起こす非定常雑音）も考

慮されるべきである．

音声認識における入力音声データと音響モデルとの照合処理は，音響分析により得られる特徴ベクトルの集合が張る音響空間内で行われる処理であり，あらゆる音響的変動要因はこの音響空間内に集約される．つまりは，音声データの特徴ベクトルから，データ間の違いを表現することができれば，直接的に，タスク間及びタスク内に限らず，また，音声及び非音声に限らず，データ間の関連性の把握が可能となる．音声データから抽出される特徴ベクトルのような多次元のデータのデータ集合分布や構造を直感的に把握する技術として，可視化技術が知られている．可視化技術は，本来不可視である多次元データを，2 次または 3 次といった可視空間に写像することで，多次元データの分布や構造を把握を可能とし，データ処理に有用な知識（アイデア）の抽出を補助する効果も有することから，主にデータマイニングの分野で幅広く適用され，その有効性が示されている．音声コーパスの可視化が可能となれば，直感的にタスク間の関連性，ひいてはタスク依存性を把握することができると期待できる．しかしながら，従来の可視化手法は全て，多次元ベクトルを低次元可視空間に写像する手法であり，音声コーパスのような膨大な量のデータ（数万〜）の可視化は，計算量の観点からも現実的でない．実際に，これらの可視化手法により可視化されるデータの多くは数十〜数千までの規模のものが多い．

本論文では，上記の課題を踏まえ，タスク依存性の分析に可視化手法を用いることを前提に，従来の可視化手法の問題点を解決する新たな可視化手法を，次章で提案する．

なお，付録 B に，2.4.2 節でのクロスタスクの音声認識実験で用いた音声コーパスに対し，物理量分析を行った結果を参考情報として示す．国内にはこれまで様々な目的で構築された大規模音声コーパスが複数存在するが，音声コーパス間のタスク依存性に関する報告は少ない．関連する報告として，国内の既存音声コーパスを用いた性能劣化要因の分析は行われているが [27, 29, 31, 33, 34]，複数の音声コーパスの横断的な分析は成されていない．今後の日本語における音声認識技術の向上に，タスク依存性の問題は重要な課題の一つであり，日本語の複数の大規

模音声コーパスを用いてのタスク依存性の分析は必須と考えられる．また，国内の既存の大規模音声コーパスの有効利用という観点から，それらを用いたタスク依存性の分析は資料的価値があるものと考えられる．

2.5 本章のまとめ

本章ではまず，音声認識の基礎知識の導入として，音声認識全般における概要として，音声認識の原理と構成及び，構成要素である音響モデル，言語モデル，HMM による音声認識に関する説明を行った．次に，本論文の議論の中心となる音響モデルに関して，モデル学習の具体的な学習アルゴリズム及び，音響モデルを作成する材料である音声コーパスの構築における音声データの収集の一般的な手順に関する説明を行い，高い認識性能を実現するためには，目的とするタスクにおける音声データと音響的に整合性の高い音響モデルの作成，つまりは，音声コーパスの構築が重要であることを述べた．次に，実際に，国内に現存する大規模日本語音声コーパスを用い，クロスタスクの音声認識実験を行った．クロスタスクの音声認識実験では，モデル構造の違いや，学習データ量によらず，全てのタスクでタスクが一致している場合の性能が最も高く，他のタスクに対しては性能が劣化し，その度合いはタスクにより異なる等，タスク依存性とその重要性を確認した．また，monophone HMM と比較し，triphone HMM はよりタスク依存性が強くなる傾向にあることを確認した．最後に，タスク依存性の分析における，従来の物理量解析の問題点及び新しい分析手法としての可視化手法の可能性とその課題について議論を行った．次章では，大規模な音声コーパスの可視化を可能とする新しい可視化手法を提案し，その有効性について議論を行う．

3. 音響空間の可視化手法：COSMOS 法

3.1 はじめに

可視化は多次元のデータの分布や構造を把握し，有用な知識を抽出するための重要な技術の一つである．主にデータマイニングの分野で幅広く適用され，その有効性が示されている．多次元データを低次元空間に写像することで，情報は欠落する恐れはあるが，多次元データの全体像を捉えることができれば知識抽出は容易となる．多次元データの可視化手法は大別して，対象となるデータの属性情報（クラスラベル）を用いた教師有の手法とクラスラベルを用いない教師無の手法に分けることができる．教師有の代表的な手法としては，判別分析法 [48]，Aladjem 法 [49]，ニューラルネットワークによる手法 [50]，グラフを利用した手法 [51] 等がある．また，教師無の代表的な手法としては主成分分析法 [52]，多次元尺度構成法 [53]，SOM（Self-Organizing Map）法 [54, 55]，射影追跡法 [56]，数量化IV類による手法 [57] 等がある．

これらの手法は全て，多次元ベクトルを低次元可視空間に写像する手法である．しかしながら，2.4.3 節で述べたとおり，音声コーパスのような膨大な量のデータ（数万〜）の可視化は，計算量の観点からも現実的でない．そこで，本論文では，音響モデル群を可視空間に写像し，音響モデル群が表現する音響空間の拡がりを把握する可視化手法を提案する．音響モデルは音声データから抽出された特徴ベクトル集合を統計的に表現するものであり，音声コーパスの近似表現と捉えることができる．音響モデル群を可視空間に写像することで，大規模な多次元情報の可視化が実現し，保持する音声コーパス群の多様性やタスク間の関係性を直感的に把握できることが期待される．過去の報告では，主成分分析法を用いた音響モデルの可視化手法が提案されている [52]．この手法では，音響モデルの平均ベクトルの連結ベクトルの第 1 主成分と第 2 主成分を利用して，音響モデルを可視空間である 2 次元平面上に写像している．しかしながら報告では，第 2 主成分までの累積寄与率は 9 %程度であり，主成分分析の累積寄与率の目安とされている 80 %に比べて著しく小さく，得られた 2 次元平面上の散布図は，元の多次元正規分布が表現する情報（音響）空間を忠実に表現しているとは言い難い．音響モデル

の多次元正規分布により表現される集合を，できるだけ情報の欠落なく，可視空間である 2 次元平面上へ写像する手法が求められる．

次節以降では，まず 2 節で，従来の可視化手法の技術的な紹介として，主成分分析法，多次元尺度構成法，SOM 法の説明を行う．3 節では，本論文で提案する可視化手法である COSMOS（COmprehensive Space Map of Objective Signal）法の説明を行う．4 節では，従来の可視化手法と提案する可視化手法との比較を行い，提案手法の有効性を示す．5 節では，提案手法による可視化の例として，様々な音声コーパスを用い，年代，性別，信号雑音比依存の音響モデル群や，音声群と非音声（非定常性雑音）群といった，様々な特徴を有する音響モデル群の 2 次元空間への写像結果を紹介する．最後に，6 節で本章のまとめを行う．

3.2 従来の可視化手法

本節では，様々な分野で用いられている，従来の可視化手法の紹介を行う．本論文では，音声コーパスのようなクラスラベルのないデータベースの可視化を目的としていることから，教師無の可視化手法のみ紹介を行う．以降，従来の代表的な教師無の可視化手法である，主成分分析法，多次元尺度構成法，SOM 法の紹介を行う．

3.2.1 主成分分析法

主成分分析は，K 次元の T 個の入力データ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$ が与えられたときの標本平均を $\bar{\boldsymbol{x}}$，ベクトルのノルムを $\|\cdot\|$ で表すとして，単位ベクトル $\|\boldsymbol{\nu}\| = 1$ の中で次式を最小にする $\boldsymbol{\nu}$ を求める形で定式化される．ここで $'$ は転置を表す．

$$\sum_{t=1}^{T}\left[\|\boldsymbol{x}_t - \bar{\boldsymbol{x}}\|^2 - \{\boldsymbol{\nu}'(\boldsymbol{x}_t - \bar{\boldsymbol{x}})\}^2\right] \tag{3.1}$$

平均を差し引く操作は以下の考察では本質的でないので，以降は標本平均は 0，

あるいは 0 となるように全体を平行移動するとして次式を考える.

$$\sum_{t=1}^{T}\left[\|\boldsymbol{x}_t\|^2 - \{\boldsymbol{\nu}'\boldsymbol{x}_t\}^2\right] \tag{3.2}$$

上式の最小化は，入力データに含まれる $\boldsymbol{\nu}$ 方向の成分だけを用いて，出来るだけ元の入力データの分散をうまく表現するように方向 $\boldsymbol{\nu}$ を選ぶということを意味している.

更に，第 1 項はベクトルのノルムの 2 乗和であり，物理的には K 次元信号 $\boldsymbol{x}_t$ のパワーを表しているとみなすことができ，これに対して $\boldsymbol{\nu}'\boldsymbol{x}_t$ は $\boldsymbol{x}_t$ を $\boldsymbol{\nu}$ 方向に射影したもののパワーを表していることから，式 (3.2) は次式で書き換えることが出来る.

$$\sum_{t=1}^{T}\|\boldsymbol{x}_t - (\boldsymbol{\nu}'\boldsymbol{x}_t)\boldsymbol{\nu}\|^2 \tag{3.3}$$

結果，第一主成分を求める問題は入力の多次元データ $\boldsymbol{x}_t$ を，1 つの定ベクトル $\boldsymbol{\nu}$ でうまく表現する場合に，最小 2 乗誤差の意味で最も良く近似する $\boldsymbol{\nu}$ を求める問題と解釈することができる.

次に，$\boldsymbol{\nu}'\boldsymbol{x}_t$ を $\xi(t)$ とし，次式の最小化を考える.

$$\sum_{t=1}^{T}\|\boldsymbol{x}_t - \xi(t)\boldsymbol{\nu}\|^2 = \sum_{t=1}^{T}\left\{\|\boldsymbol{x}_t\|^2 - \|\xi(t)\boldsymbol{\nu}\|^2\right\} \tag{3.4}$$

これは次式の最大化と等価である.

$$\sum_{t=1}^{T}\|\xi(t)\boldsymbol{\nu}\|^2 \tag{3.5}$$

$\xi(t)$ を代入し，次式を得ることができる.

$$\sum_{t=1}^{T}\|(\boldsymbol{\nu}'\boldsymbol{x}_t)\boldsymbol{\nu}\|^2 = \sum_{t=1}^{T}\|\boldsymbol{\nu}(\boldsymbol{x}_t'\boldsymbol{\nu})\|^2 = \boldsymbol{\nu}'\left(\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)\boldsymbol{\nu} \tag{3.6}$$

ここで，標本分散行列が次式で表現された場合，$\boldsymbol{\nu}'V\boldsymbol{\nu}$を最大にする大きさ1のベクトルを求める問題，すなわち最大固有値を求める問題に帰着される．

$$V = \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{x}(t)\boldsymbol{\nu}(t)' \tag{3.7}$$

分散行列を直交行列Uによって対角化し，次式を得る．

$$U'VU = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix} \tag{3.8}$$

但し，$\lambda_1 \geq \cdots \geq \lambda_d$である．この時，$U$の$\boldsymbol{i}$番目の列ベクトルは第$\boldsymbol{i}$番目の固有値$\lambda_i$に対応する固有ベクトルとなっている．大きさ1のベクトルを$\boldsymbol{o} = (o_1, \ldots, o_d)'$として，この対角行列を両側から挟みこむことで次式が得られる．

$$\boldsymbol{o}'U'VU\boldsymbol{o} = \boldsymbol{o}' \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix} \boldsymbol{o} = \lambda_1 o_1^2 + \cdots + \lambda_d o_d^2 \tag{3.9}$$

この式を最大にする大きさ1のベクトルは明らかに$\boldsymbol{e}_1 = (1, 0, \ldots, 0)'$（または$\boldsymbol{e}_1 = (-1, 0, \ldots, 0)'$）である．従って，

$$\nu = U\boldsymbol{e}_1 \tag{3.10}$$

とすればVを両側から挟んで最大にするベクトルが求められる．つまり，Vの最大固有値の固有ベクトルになっているUの第1ベクトルを選んでやればよいことが分かる．

以上の議論は第2主成分以降にも拡張され，次式の近似が得られる．

$$\boldsymbol{x}(t) \sim \xi_1(t)\nu_1 + \xi_2(t)\nu_2 + \cdots \tag{3.11}$$

すなわち，主成分が求められた元の観測値から主成分方向の成分を取り除いた残差（次式）に対して，同じように主成分分析を行う．

$$\boldsymbol{x}(t) - \xi_1(t)\nu_1 \tag{3.12}$$

この操作を 3 番目以降の固有ベクトルにも順次繰り返すことで次式の分解を得ることができる．

$$\boldsymbol{x}(t) = \xi_1(t)\nu_1 + \xi_2(t)\nu_2 + \cdots + \xi_d(t)\nu_d \tag{3.13}$$

なお，$\xi_i(t)$ の分散が大きいものから第 1 主成分，第 2 主成分，…，第 d 主成分となっていることに注意する．

3.2.2 多次元尺度構成法

多次元尺度構成法は対象間の計量的関係を空間的に表現する方法の一つであり，元空間の位置関係（距離関係）をできるだけ保持するように低次の空間に再配置する手法である．多次元尺度構成法には様々なものがあるが，代表的な考え方を以下に示す．今，T 個の多次元データ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \boldsymbol{x}_j, \ldots, \boldsymbol{x}_T$ が存在し，対象 i，j 間の元空間における距離 $\mathcal{D}(i, j)$ が定義されているものとする．空間上に配置される i 及び j の座標ベクトルを $\boldsymbol{z}(i)$，$\boldsymbol{z}(j)$ とし，配置後の距離 $\hat{\mathcal{D}}(i, j)$ を次式で定義する．

$$\hat{\mathcal{D}}(i, j) = \|\boldsymbol{z}(i) - \boldsymbol{z}(j)\| \tag{3.14}$$

この時，$T(T-1)/2$ 個の像点間の距離 $\hat{\mathcal{D}}(i, j)$ をできるだけ元の空間での距離 $\mathcal{D}(i, j)$ に近づける $\boldsymbol{z}(1), \ldots, \boldsymbol{z}(h), \ldots, \boldsymbol{z}(T)$ の配置を求めたい．全ての i，j に対して $\mathcal{D}(i, j) = \hat{\mathcal{D}}(i, j)$ を満たす配置を見つけることは通常不可能であるため，ある配置が他の配置より優れていると判定する基準が必要となる．以下に示す誤差 2 乗和関数は，評価関数（判定基準）としてよく用いられており，損失関数と呼ばれる．

$$\mathcal{E}_\tau^{I} = \sum_{i=1}^{T-1}\sum_{j=i+1}^{T}\frac{\{\mathcal{D}(i,j)-\hat{\mathcal{D}}_\tau(i,j)\}^2}{\mathcal{D}(i,j)^2} \tag{3.15}$$

$$\mathcal{E}_\tau^{II} = \sum_{i=1}^{T-1}\sum_{j=i+1}^{T}\Big[\frac{\mathcal{D}(i,j)-\hat{\mathcal{D}}_\tau(i,j)}{\mathcal{D}(i,j)}\Big]^2 \tag{3.16}$$

$$\mathcal{E}_\tau^{III} = \frac{1}{\sum_{i=1}^{T-1}\sum_{j=i+1}^{T}\mathcal{D}(i,j)}\sum_{i=1}^{T-1}\sum_{j=i+1}^{T}\Big[\frac{\{\mathcal{D}(i,j)-\hat{\mathcal{D}}_\tau(i,j)\}^2}{\mathcal{D}(i,j)}\Big] \tag{3.17}$$

評価関数が決まれば，最適な配置は，評価関数が最小となる配置として，最急降下法等の最適化手法により求められる．ここで，τ は更新回数を示している．これらの評価関数は，点間の距離のみを含んでいるため，配置の剛体運動に対して不変である．更に，全て正規化されているために，標本点の膨張に対しても最小値は不変である．$\mathcal{E}_I$ は大きな誤差を強調し，$\mathcal{E}_{II}$ は大きな誤差率を強調する．$\mathcal{E}_{III}$ は $\mathcal{E}_I$ と $\mathcal{E}_{II}$ のバランスをとり，大きな誤差と誤差率の損を強調するものである．なお，各評価関数の勾配は以下ように容易に求めることが出来る．

$$\nabla z(h)_\tau^{\mathcal{E}_I} = \frac{-2}{\sum_{i=1}^{T-1}\sum_{j=i+1}^{T}\mathcal{D}(i,j)^2}\sum_{j=1,j\neq h}^{T}\Big\{\acute{\mathcal{D}}(i,j)\cdot\frac{z_\tau(h)-z_\tau(j)}{\hat{\mathcal{D}}_\tau(h,j)}\Big\} \tag{3.18}$$

$$\nabla z(h)_\tau^{\mathcal{E}_{II}} = -2\sum_{j=1,j\neq h}^{T}\frac{\acute{\mathcal{D}}(i,j)}{\mathcal{D}(h,j)^2}\cdot\frac{z_\tau(h)-z_\tau(j)}{\hat{\mathcal{D}}_\tau(h,j)} \tag{3.19}$$

$$\nabla z(h)_\tau^{\mathcal{E}_{III}} = \frac{-2}{\sum_{i=1}^{T-1}\sum_{j=i+1}^{T}\mathcal{D}(i,j)}\sum_{j=1,j\neq h}^{T}\frac{\acute{\mathcal{D}}(i,j)}{\mathcal{D}(h,j)}\cdot\frac{z_\tau(h)-z_\tau(j)}{\hat{\mathcal{D}}_\tau(h,j)} \tag{3.20}$$

$$\acute{\mathcal{D}}(i,j) = \mathcal{D}(i,j)-\hat{\mathcal{D}}_\tau(i,j) \tag{3.21}$$

初期配置は任意に設定が可能であり．一般に，乱数や主成分分析法の結果を用いることが多い．

3.2.3 SOM 法

SOM（Self-Organizing Map：自己組織化）法は人間の脳の仕組みを模倣した情報処理機構として知られるニューラルネットワークの一種であり，中間層の無い

2 階層型の教師無し競合学習モデルである．競合近傍学習と呼ばれる学習アルゴリズムによって，似た特徴を持つデータは近くに，そうでないものは離れた位置に配置されるようなマップを形成する．今，次元数が K である多次元ベクトル $\boldsymbol{x}$ が入力層に存在するならば，競合層の個々のノード i は同様に K 次元で表現されるベクトル $\boldsymbol{u}(i)$ を持つ．このベクトルを参照ベクトルと呼ぶ．但し，参照ベクトルの $\boldsymbol{u}(i)$ の k 次元目の要素 $u(i,k)$ はノード i と入力データ $\boldsymbol{x}$ の k 次元目の要素 $x(k)$ の間の重みであり，自己組織化過程で少しずつ修正される．具体的には，入力ベクトル $\boldsymbol{x}$ が与えられたとき，まず，その入力を全てのノードの参照ベクトルと比較し，ユークリッド距離の最も小さいノードを活性化する．その活性化されたノードを勝者ノードと呼ぶ．すなわち，勝者ノード ζ は以下のように選ばれる．

$$\zeta = \underset{i}{\operatorname{argmin}}\{\|\boldsymbol{x} - \boldsymbol{u}(i)\|\} \tag{3.22}$$

そして，自己組織化が大局的に行われるように，勝者ノードの近傍ノードも活性化させ，スムージングを行う．これは，活性化された全てのノードに対し，それらの参照ベクトルを入力ベクトルに近づくように修正を行うことを意味している．

$$\boldsymbol{u}_{\tau+1}(i) = \boldsymbol{u}_{\tau}(i) + \mathcal{I}_{\tau}(\zeta, i)[\boldsymbol{x}_{\tau} - \boldsymbol{u}_{\tau}(i)] \tag{3.23}$$

ここで，τ は学習回数で，$\mathcal{I}_{\tau}(\zeta, i)$ は近傍関数である．近傍関数には様々なものがあるが，本論文では次式を用いる．

$$\mathcal{I}_{\tau}(\zeta, i) = \epsilon_{\tau} \exp(-\frac{\|\boldsymbol{v}_{\zeta} - \boldsymbol{v}_i\|^2}{2\delta_{\tau}^2}) \tag{3.24}$$

$\boldsymbol{v}_{\zeta}$ 及び $\boldsymbol{v}_i$ はそれぞれ，勝者ノード ζ と傍ノード i の位置ベクトルである．従って，$\|\boldsymbol{v}_{\zeta} - \boldsymbol{v}_i\|$ は，近傍ノード i が勝者ノード ζ から離れていくにつれ，$\mathcal{I}(\zeta, i)$ が小さくなり，$\boldsymbol{u}_{\tau}(i)$ の修正量が小さくなることを意味する．また，ϵ_{τ} は学習率で，δ_{τ} は近傍の大きさ（半径）である．これらは時間とともに単調に減少していく関数である．本論文では次式を用いる．

$$\epsilon_{\tau} = \epsilon_0 \frac{H - \tau}{H} \tag{3.25}$$

$$\delta_\tau = 1 + (\delta_\tau - 1)\frac{H - \tau}{H} \tag{3.26}$$

但し，ϵ_0 は初期値で，H は総学習回数である．自己組織化が終わった後のマッピング処理は，単に入力ベクトル $\boldsymbol{x}$ に対する勝者ノードを選び出すことである．従って，これまで述べた自己組織化は，入力ベクトルにマッチするノードの整列過程とみなすことができる．すなわち，マッチするノードを入力に近づける過程である．

なお，本論文では，SOM 法による可視化に，SOM_PAK[55] を用いている．

3.3 COSMOS 法の概要

本節では，教師無の手法の一つである多次元尺度構成法を応用した音響モデルの可視化手法として，COSMOS（COmprehensive Space Map of Objective Signal）法を提案する [59, 60, 61, 63]．多次元尺度構成法は対象間の計量的関係を空間的に表現する手法である．多次元尺度構成法には様々なものがあるが，COSMOS 手法では，古典的多次元尺度構成法の一つである Sammon 法 [58] に統計モデル間距離を適用することで，音響モデルの可視化を可能としている．以下に Sammon 法の概要及び統計モデル間距離について述べる．

3.3.1 Sammon 法

Sammon 法は，多次元空間上における 2 つのベクトル i と j の相互距離 $\mathcal{D}(i, j)$ と，非線形写像により低次元空間上に配置された写像位置座標の相互ユークリッド距離 $\hat{\mathcal{D}}_l(i, j)$ との誤差を全ベクトル間で求め，その総和が最小となるように最急降下法によって 2 次元または 3 次元の可視空間上の写像位置の座標を最適化する手法である．最小化すべき損失関数 $\mathcal{E}_\tau$ は次式で与えられる．

$$\mathcal{E}_\tau = \frac{1}{\iota}\sum_{i=1}^{T-1}\sum_{j=i+1}^{T}\left[\frac{\{\mathcal{D}(i,j)-\hat{\mathcal{D}}_\tau(i,j)\}^2}{\mathcal{D}(i,j)}\right] \tag{3.27}$$

$$\iota = \sum_{i=1}^{T-1}\sum_{j=i+1}^{T}\mathcal{D}(i,j) \tag{3.28}$$

ここで T は総ベクトル数，τ は最急降下法の更新回数である．$\hat{\mathcal{D}}_\tau(i,j)$ はユークリッド距離であり次式で与えられる．

$$\hat{\mathcal{D}}_\tau(i,j) = \sqrt{\sum_{k=1}^{K}(z_\tau(i,k)-z_\tau(j,k))^2} \tag{3.29}$$

ここで K は写像される低次元空間の次元数であり，k は所定の次元である．2 次元空間上に写像する場合は $K=2$ となる．また，$\hat{\mathcal{D}}_0(i,j)$ は低次元空間上の写像位置の初期座標から上式に従い算出されるが，初期座標としては一般的に，乱数や主成分分析法の結果が用いられる．

最急降下法により $\tau+1$ 回更新後の第 k 次元の h 番目のベクトルの座標値は次式で算出される．

$$z_{\tau+1}(h,k) = z_\tau(h,k)-\epsilon\cdot\Delta z_\tau(h,k) \tag{3.30}$$

$$\Delta z_\tau(h,k) = \frac{\dfrac{\partial\mathcal{E}_\tau}{\partial z_\tau(h,k)}}{\left|\dfrac{\partial^2\mathcal{E}_\tau}{\partial z_\tau(h,k)^2}\right|} \tag{3.31}$$

$$\frac{\partial\mathcal{E}_\tau}{\partial z_\tau(h,k)} = \frac{-2}{\iota}\sum_{j=1,j\neq h}^{T}\frac{\acute{\mathcal{D}}(i,j)}{\mathcal{D}(h,j)}\cdot\frac{z_\tau(h,k)-z_\tau(j,k)}{\hat{\mathcal{D}}_\tau(h,j)} \tag{3.32}$$

$$\frac{\partial^2\mathcal{E}_\tau}{\partial z_\tau(h,k)^2} = \frac{-2}{\iota}\sum_{j=1,j\neq h}^{T}\frac{1}{\mathcal{D}(h,k)\cdot\hat{\mathcal{D}}_\tau(h,j)}\cdot \left[\{\acute{\mathcal{D}}(i,j)\}-\frac{\{z_\tau(h,k)-z_\tau(j,k)\}^2}{\hat{\mathcal{D}}_\tau(h,j)}\cdot\left\{1+\frac{\acute{\mathcal{D}}(i,j)}{\mathcal{D}(h,j)}\right\}\right] \tag{3.33}$$

$$\acute{\mathcal{D}}(i,j) = \mathcal{D}(i,j)-\hat{\mathcal{D}}_\tau(i,j) \tag{3.34}$$

COSMOS 法は $\mathcal{D}(i,j)$ に統計モデル間の距離を導入することで音響モデル群の可視化を実現している．

3.3.2 統計モデル間距離

本論文では，統計モデル（以下，音響モデル）に HMM を適用する．音響モデルは複数の音響単位（本論文では音素または単語）の HMM からなる HMM セットとして定義される．音響モデル i と音響モデル j の距離 $\mathcal{D}(i,j)$ は次式で与えられる．

$$\mathcal{D}(i,j) = \sum_{r=0}^{R-1} \mathcal{H}(i,j,s_r) \cdot \mathcal{W}(r) \tag{3.35}$$

$$\mathcal{W}(r) = \frac{\omega(i,r) + \omega(j,r)}{2} \tag{3.36}$$

ここで R は音響単位の総数，$\omega(i,r)$ は音響モデル i の r 番目の音響単位 s_r の重みであり，総和が 1 となるよう設定される．注目すべき音響単位にはより大きな重みが設定されることになる．本論文では学習データに現れる各音素の出現頻度を，総和が 1 となるよう正規化したものを用いる．$\mathcal{H}(i,j,s_r)$ は音響モデル i 及び音響モデル j に含まれる r 番目の音響単位 s_r の HMM 間距離（以下，音響単位間距離と定義する）である．$\mathcal{H}(i,j,s_r)$ は次式で与えられる．

$$\mathcal{H}(i,j,s_r) = PD(s_r^i, s_r^j) \tag{3.37}$$

音響単位間距離を算出する際，各音響単位の状態数が同数であり，状態間のアライメントは一対一に対応がとれていると仮定すると，音響単位 υ と ϕ の音素間距離 $PD(\upsilon,\phi)$ は次式で定義される．

$$PD(\upsilon,\phi) = \frac{1}{N} \sum_{n=1}^{N} ND(\upsilon,\phi,n_\upsilon,n_\phi) \tag{3.38}$$

ここで N は総状態数，n_a，n_b はそれぞれ，音響単位 υ と ϕ の状態番号であり，状態間のアライメントが一対一に対応が取れている場合は $n_a = n_b$ となる．各音

響単位の状態数が異なる場合等，状態間のアライメントが一対一に対応がとれていない場合は，DTW 法 [5] により次式で示されるように状態間の対応をとる．

$$PD(\upsilon,\phi) = \frac{1}{L}\sum_{l=1}^{L}\rho_l \cdot ND(\upsilon,\phi,\mathcal{Y}(\upsilon,l),\mathcal{Y}(\phi,l)) \tag{3.39}$$

ここで，L は DTW 法により求められた状態系列の状態系列長であり，l は系列番号である．$\mathcal{Y}(\upsilon,l)$ は，系列番号 l における音響単位 υ の状態番号 n_υ を出力する関数である．ρ_l は系列番号 l における重み係数である（DTW 法では一般に傾斜制限が設けられる）．

状態間の距離 $ND(\upsilon,\phi,n_\upsilon,n_\phi)$ は次式で定義される．なお，状態には混合正規分布が付与されているものとする．

$$ND(\upsilon,\phi,n_\upsilon,n_\phi) = \min_{1\le p,q\le M} BD(g^{\upsilon}_{n,p}(x), g^{\phi}_{n,q}(x)) \tag{3.40}$$

ここで，$g^{\upsilon}_{n,p}(x)$ は音響単位 υ の状態 n の p 番目のガウス分布である．M はガウス分布の総数であり，ここでは対応する状態間の分布数は同数としている．式 (3.40) からもわかるように，各状態の $M\times M$ 通りの分布間距離から，最小値を状態間距離としている．また，$BD(g(x),h(x))$ は分布間距離の一つとしてよく知られている Bhattacharyya 距離 [74] を用いる．Bhattacharyya 距離 BD は次式で定義される．

$$BD(g,h) = \frac{1}{8}(\boldsymbol{\mu}_g-\boldsymbol{\mu}_h)\left\{\frac{\boldsymbol{\Sigma}_g-\boldsymbol{\Sigma}_h}{2}\right\}^{-1}(\boldsymbol{\mu}_g-\boldsymbol{\mu}_h)' + \frac{1}{2}\log\left(\frac{|(\boldsymbol{\Sigma}_g+\boldsymbol{\Sigma}_h)/2|}{|\boldsymbol{\Sigma}_g|^{\frac{1}{2}}|\boldsymbol{\Sigma}_h|^{\frac{1}{2}}}\right) \tag{3.41}$$

ここで，$\boldsymbol{\mu}$ は平均ベクトル，$\boldsymbol{\Sigma}$ は共分散行列である．分布間距離として，Kullback-Leibler divergence[75] などの公知の距離尺度を用いることが可能であり，Kullback-Leibler divergence に関しては，Bhattacharyya 距離と比較し同程度の有効性が確認されているが [65]，本論文では Bayes 誤り率と関連深い Bhattacharyya 距離を用いるものとする．

分布パラメータである平均及び分散を考慮した分布間距離に基づいた統計モデル間の距離を導入することで，音響モデル間の距離関係を精度良く表現すること

が可能となり，低次元空間への写像（可視化）の精度が向上することが期待できる．また，一般的に可視化の対象となる入力データの系列長は一定であることが前提となっているが，COSMOS 法（Sammon 法も同様）では距離を写像における尺度としているため，式 (3.39) のような，系列長の異なるデータ間の距離を算出することのできる定義を導入することで，容易に系列長の異なるデータ群の写像が可能となるという利点も持つ．

3.4 従来の可視化手法との比較

本節では，提案手法である COSMOS 法と，COSMOS 法と同様に教師無しの手法である，主成分分析による手法，SOM 法との比較を行う．前述の通り，これまでに音響モデルの可視化手法として主成分分析による手法が提案されており，音響モデルの平均ベクトルの連結ベクトルから，第 1 主成分と第 2 主成分を抽出し，音響モデルを可視空間である 2 次元平面上に写像している．SOM 法はデータマイニングの分野で幅広く用いられている手法であり，主成分分析による手法と同様に，音響モデルの平均ベクトルの連結ベクトルを，2 次元平面上に写像することが可能である．また，音響モデル間距離に分散値を考慮することの有意性を調査するために，音響モデルの平均ベクトルの連結ベクトルに Sammon 法を適用した場合との比較を行う．

3.4.1 実験データ概要

ATR で収録された 5240 単語 [26] を 175 単語からなる 30 の単語リストに分割し，成人の女性性話者 126 名が 3 又は 4 つののリストをそれぞれ別の発話様式により発声した音声データを実験データとして用いる（のべ話者数は 457 名）．なお，発話リストの分割は，分割後のリスト間で，出現する音素の頻度に大きな違いが出ないよう，50 音順に並び替えられた 5240 単語を上から順番に割り振るように行った．各発話様式により発声した音声データを収録する際に話者に提示した指示内容の概要を表 3.1 に示す．音響解析条件は表 2.1 の設定 1 に従うものとする．

表 3.1 各発話様式の概要

Speaking style	Instruction
Normal	Read utterance list at normal speed of conversation.
Fast	Read utterance list at faster than normal speed of speech.
High	Read utterance list in a high key (high-pitched voice).
Whisper	Read utterance list at a level not to be overheard by near-by persons.
Loud	Read utterance list at a level to be heard by persons at some distance.
Lomberd	Read utterance list among an ambient car noise.
Syllable enhanced	Read utterance list by enhancing the Japanese syllables.

3.4.2 評価尺度

一般的に，可視化手法は，得られた可視化結果の可視空間上（2 次元平面上）でのデータの配置や分離度の視覚的な主観評価により，その優劣を比較されることが多い．本論文で用いる実験データにおいては，発話様式以外の条件（性別，年代，背景雑音，残響）が同一であることから，発話様式の違いが，音響的特徴の変動に最も大きく影響を及ぼす要因であると考えられる．そこで，本論文では，発話様式の分離度を評価尺度とし，提案手法の評価及び，従来手法との比較を行う．分離度は，主観評価だけでなく，客観評価として，以下の手順により評価される．

1. 写像後の 2 次元平面上の座標値に対し，原点（0,0）が中心となるよう平行移動処理を行う．

2. 1. で得られた座標値に対し，中心からの距離の平均が 1 となるよう正規化処理を行う．

3. 2. で得られた座標値に対し，発話様式毎に 2 次元正規分布を求める．

4. 3. で得られた各発話様式の分布間距離を式 (3.41) に従い算出する．

3.4.3 節では，上記の手順に従い，各可視化手法の比較を行う．

3.4.3 写像実験

前述の実験データを用い，特定話者音響モデルを学習する．音響モデルには biphone HMM を用いる．monophone の音素数は 29 とし，biphone は日本語の接続規則に従い定義される．各 HMM の状態数は 3 とし，各状態の分布数は 1 とする．なお，各正規分布の共分散行列は対角共分散行列とする．

主成分分析法，SOM 法，Sammon 法，COSMOS 法を用いて，全 457 名の特定話者音響モデルを 2 次元平面上に写像した結果を図 3.1 及び図 3.3 に示す．2 次元平面上の各点が各特定話者音響モデルを表している．図 3.1 における記号は各話者の発話様式を表しており，発話様式の違いによる音響空間上での配置を確認することができる．また，図 3.3 における記号は各話者の認識性能を表しており，認識性能の違いによる音響空間上での配置を確認することができる．なお，認識性能は全 457 話者の音声データから不特定話者音響モデルを学習し，各特定話者音響モデルの学習用 175 発声を評価した際の性能である（Closed 評価）．

図 3.1 では，全ての手法において発話様式毎の偏りが確認できる．なお，図 3.1 において，Normal と High，Loud と Lombard はそれぞれ，手法によらず明確な分離が見られなかったため，同一の記号を用いている．各手法において，発話速度の軸（Fast↔Syllable enhanced）及び発話音量の軸（Whisper↔Loud，Lombard）が形成されていることが確認できる．従来の手法はこれらの 2 つの軸がほぼ同じ軸上に存在しているのに対し，COSOMS 法ではこれらの軸がほぼ 90 度で交差しており（発話速度の軸（左下：Fast↔ 右上：Syllable enhanced），発話音量の軸（左上：Whisper↔ 右下：Loud，Lombard）），Fast 及び Syllable enhanced 発声話者の分離度という観点では，COSMOS 法の優位性を明確に確認できる．次に，発

話様式間の分布の分離度を視覚的な主観評価だけでなく，客観的な評価を行うため，3.4.2 節で紹介した手順により各発話様式間の分布間距離を求める．表 3.2 に，各可視化手法における発話様式間の分離度（分布間距離）を示す．また，手法毎に，分離度の平均を算出した結果（平均分離度）を図 3.2 に示す（図中では主成分分析法を PCA と表記）．表 3.2 及び図 3.2 より，COSMOS 法が，主成分分析法，SOM 法，Sammon 法に対し，有意な差を示しており，その有効性が確認できる．

COSMOS 法以外の手法では，発話様式以外の要因が軸を形成している可能性もあるが，発話様式が重要な要因であることは明確である．そこで，主成分分析法において，第 3，7，13 主成分まで考慮した場合の各発話様式の分離度（平均）を調査する．第 3 主成分以降の主成分を考慮することで，発話速度の軸と発話音量の軸が分離され，各発話様式の分離度が向上することが期待できる．結果を表 3.3 に示す．また，参考として，累積寄与率を合わせて表 3.3 に示す．表 3.3 より，分離度は向上するものの，2 次元における COSMOS 法が示した値と比較し小さいことが分かる．第 13 主成分以降を考慮することで大幅に分離度が向上することは考え難いことからも，COSMOS 法の優位性を示す結果と言える．

それぞれの手法は 3.2 及び 3.3 節で紹介したように，同じ可視化手法でありながら異なるアプローチをとっており，当然得られる結果は異なる．線形写像である主成分分析法は，データの分布が線形部分空間に乗っているような比較的単純な場合には分布形状をほぼそのまま表現できる長所がある反面，分布形状が複雑で非線形である場合にはどうしても構造を表現できないといった欠点がある．逆に，非線形写像である SOM 法や Sammon 法及び COSMOS 法は，複雑な分布形状を表現できる可能性があるものの，過度な変形により局所解に陥り，本来と異なる構造を表現してしまう恐れがある．音響モデルのモデルパラメータの平均ベクトルのみを用いた場合の比較となる，主成分分析法，SOM 法及び Sammon 法の比較では，図 3.2 からは，やや主成分分析法が発話様式間分離度という観点で上回っている．表 3.2 の結果を詳細に分析すると，発話様式間分離度の大きい Whisper-Loud 間で SOM 法及び Sammon 法に対し明らかな有意差を確認することができる．SOM 法及び Sammon 法共に，元空間上において近距離にあるデータ間の修正量（配置精度）を優遇するため，遠距離にあるデータの位置関係に誤

差が生じやすいという特徴がある．なお，SOM 法においては式 (3.24)，Sammon 法においては式 (3.27) がその特徴を表している．これに対し，主成分分析法では Whisper 及び Loud の元空間におけるデータの分布形状を表現する軸を精度良く形成することができていると考えられる．この違いが，Whisper-Loud 間の発話様式間分離度における主成分分析法と，SOM 法及び Sammon 法との差異の要因と考えられる．但し，全発話様式間の平均の分離度における，COSMOS 法の値との開きと比べ，これらの 3 つの手法の差は大きいとは言えず，写像手法自体に明らかな有意差はないと考えられる．結果，COSMOS 法の優位性は，同じ写像手法を用いている Sammon 法との違いに集約することができ，音響モデルのモデルパラメータの分散ベクトルを考慮していることが COSMOS 法の優位性の主要因であると言える．実際に分散値は音声認識性能に大きく影響を与えるモデルパラメータであることからもその重要性が伺える．

次に図 3.3 では，Sammon 法及び COSMOS 法において，低性能話者が分布の周辺付近に位置することが確認できるが，主成分分析法及び SOM 法においては低性能話者の位置に特徴的な傾向を確認することができない．また，COSMOS 法で得られる分布上では，Sammon 法と比較し，低性能話者がより分布の周辺に位置していることが分かる．低性能話者の分析は，認識性能向上のための重要な観点の一つであり，COSMOS 法が低性能話者の音響的な特徴の違いを捉えることができることからも，COSMOS 法の分析手法としての有効性が確認できる．

以上より，音響モデル集合を 2 次元平面上に可視化する手法として，COSMOS 法の優位性が確認することができた．また，本実験では特定話者音響モデルを可視化の対象としているが，これは，不特定話者音響モデルが張る話者空間の拡がりを特定話者音響モデルの集合分布により近似的に表現することを目的としている．更に，音響モデルは学習元である音声コーパスに統計処理を施し得られることから，特定話者音響モデルの集合分布は音声コーパスが張る音響空間全体の近似表現とも捉えることもできる．

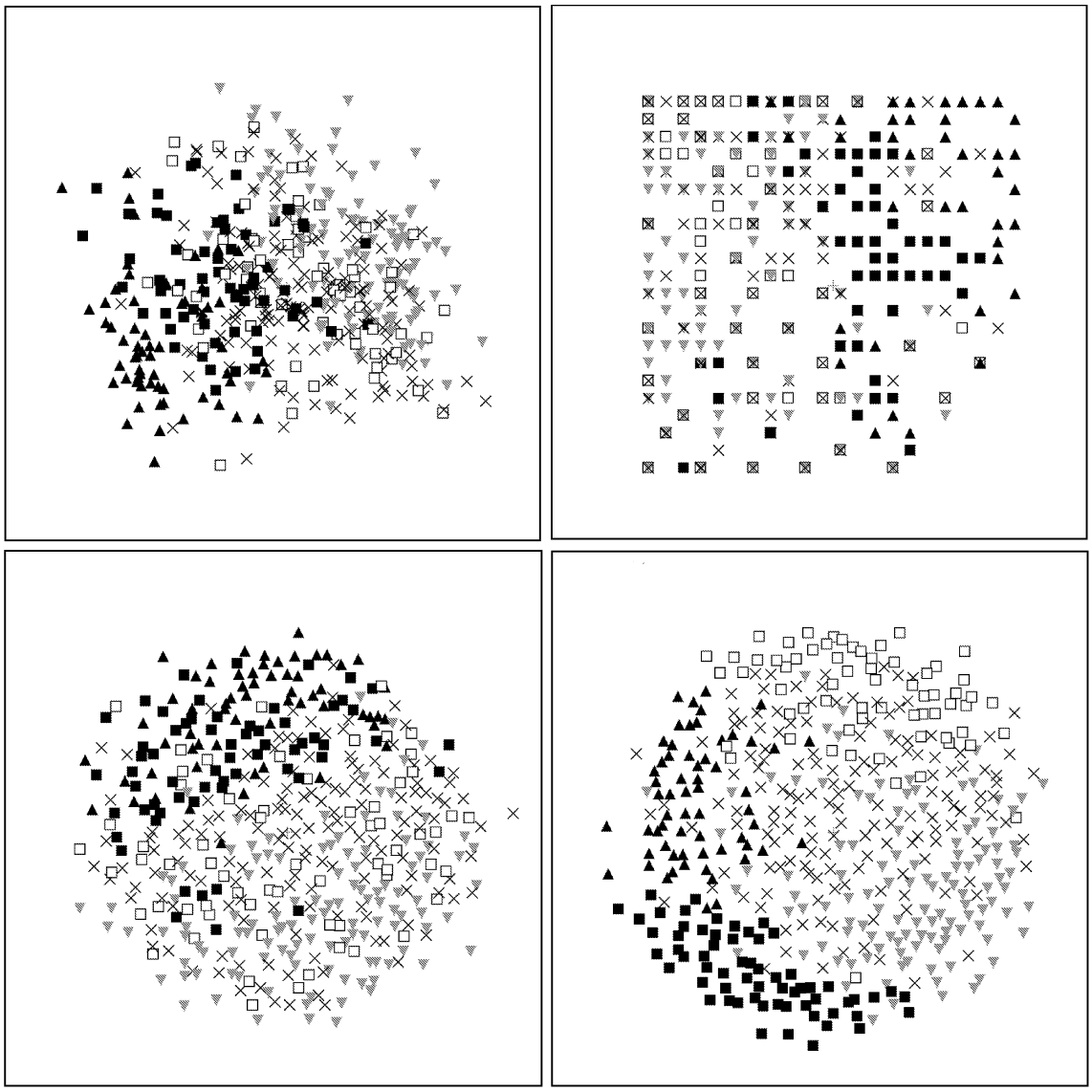


図 3.1　各可視化手法の比較：発話様式 （左上：主成分分析法，右上：SOM 法，左下：Sammon 法，右下：COSMOS 法）*

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．

表 3.2 各可視化手法の比較：発話様式間分離度（上から，COSMOS 法，主成分分析法，Sammon 法，SOM 法）

	Fast	High	Lombard	Loud	Syllable	Normal	Whisper
Fast	0.00	2.54	6.77	11.22	38.4	3.42	12.78
High	2.54	0.00	1.10	2.50	14.96	0.49	9.19
Lombard	6.77	1.10	0.00	0.29	6.86	1.19	13.47
Loud	11.22	2.50	0.29	0.00	9.26	3.03	22.08
Syllable	38.43	14.96	6.86	9.26	0.00	9.24	15.25
Normal	3.42	0.49	1.19	3.03	9.24	0.00	5.59
Whisper	12.78	9.19	13.47	22.08	15.25	5.59	0.00

	Fast	High	Lombard	Loud	Syllable	Normal	Whisper
Fast	0.00	0.28	0.55	1.40	1.54	0.09	6.86
High	0.28	0.00	0.78	0.86	3.24	0.69	9.75
Lombard	0.55	0.78	0.00	0.49	3.88	1.09	13.46
Loud	1.40	0.86	0.49	0.00	7.81	2.70	21.96
Syllable	1.54	3.24	3.88	7.81	0.00	1.05	2.60
Normal	0.09	0.69	1.09	2.70	1.05	0.00	6.19
Whisper	6.86	9.75	13.46	21.96	2.60	6.19	0.00

	Fast	High	Lombard	Loud	Syllable	Normal	Whisper
Fast	0.00	0.24	0.63	0.87	2.11	0.08	5.57
High	0.24	0.00	0.74	0.51	3.20	0.50	6.41
Lombard	0.63	0.74	0.00	0.24	5.79	1.28	12.75
Loud	0.87	0.51	0.24	0.00	6.70	1.65	12.81
Syllable	2.11	3.20	5.79	6.70	0.00	1.45	0.97
Normal	0.08	0.50	1.28	1.65	1.45	0.00	4.37
Whisper	5.57	6.41	12.75	12.81	0.97	4.37	0.00

	Fast	High	Lombard	Loud	Syllable	Normal	Whisper
Fast	0.00	1.46	0.50	0.57	1.66	0.51	5.77
High	1.46	0.00	1.15	1.25	4.87	0.94	9.81
Lombard	0.50	1.15	0.00	0.01	4.24	1.28	10.08
Loud	0.57	1.25	0.01	0.00	5.51	1.57	13.01
Syllable	1.66	4.87	4.24	5.51	0.00	1.16	1.72
Normal	0.51	0.94	1.28	1.57	1.16	0.00	4.17
Whisper	5.77	9.81	10.08	13.01	1.72	4.17	0.00

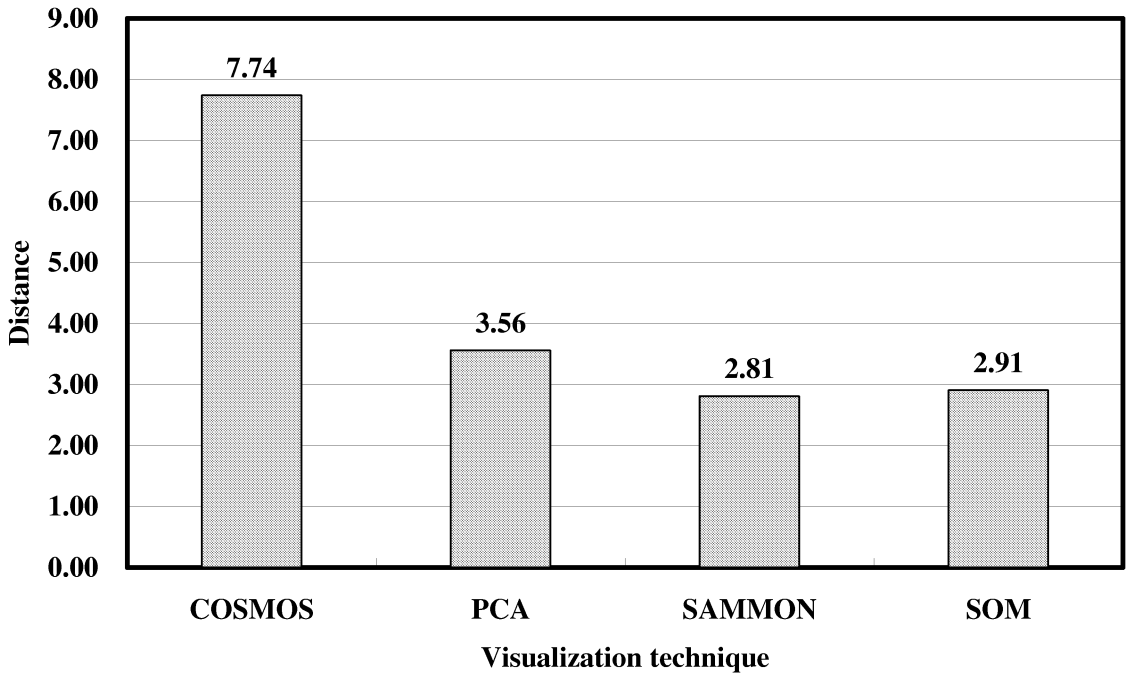


図 3.2　各可視化手法の比較：発話様式の分離度（平均）

表 3.3　主成分分析法における各主成分での発話様式間分離度（平均）及び累積寄与率

	2nd	3rd	7th	13th
Distance	3.56	4.18	4.47	5.21
Cumulative Proportion [%]	14.36	14.86	15.03	15.17

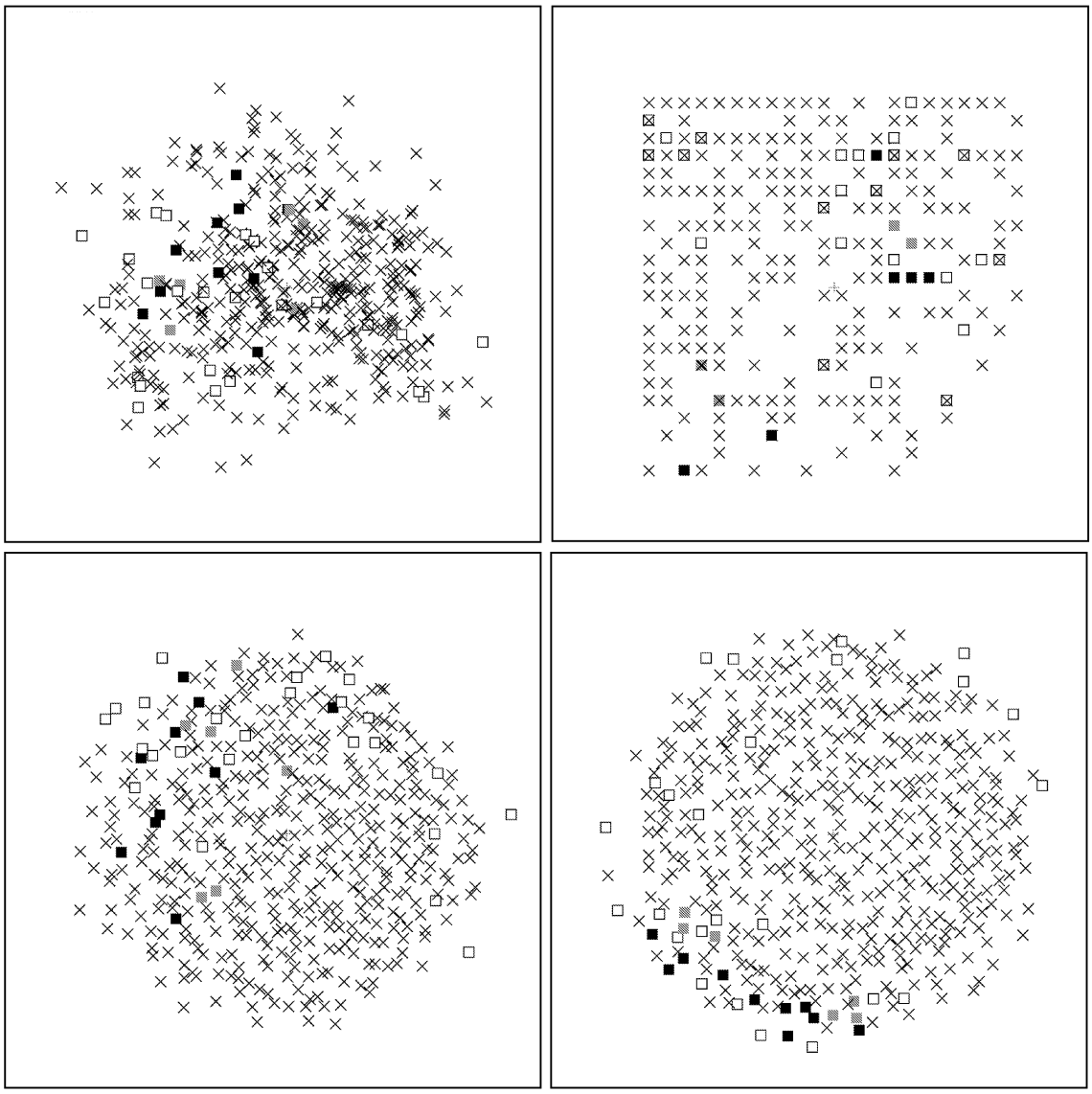


図 3.3　各可視化手法の比較：認識性能 （左上：主成分分析法，右上：SOM 法，左下：Sammon 法，右下：COSMOS 法）*

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．

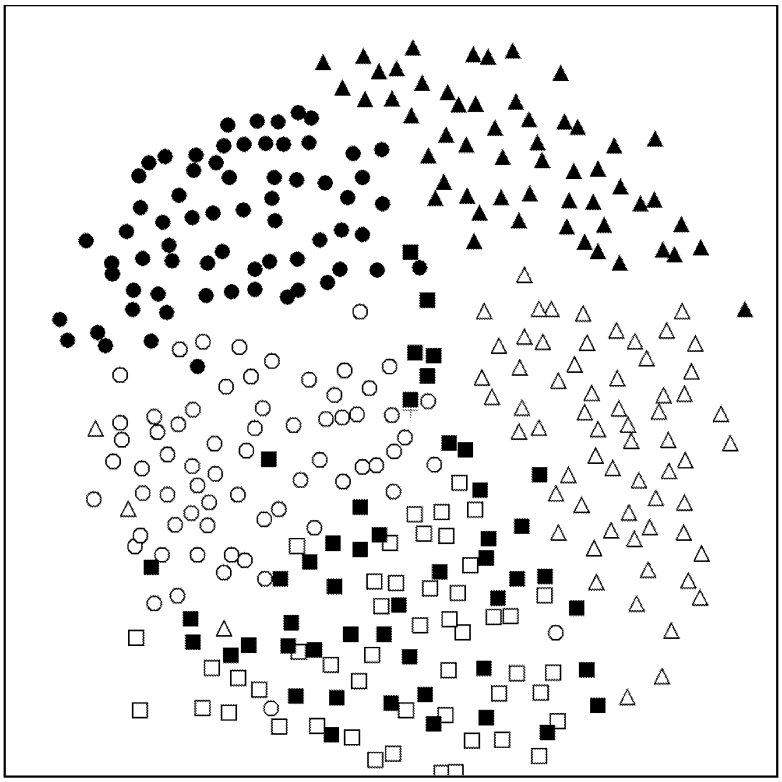
3.5 様々な音響モデル群の可視化

本節では，COSMOS 法を用い，様々な特徴を有する音響モデル群の可視化を行い，その結果を紹介する．以降，COSMOS 法により得られる，音響モデル群が張る音響空間の，2 次元平面上への写像（配置）結果を COSMOS Map と呼ぶ．まず，年代，性別，語彙，信号雑音比といったタスク依存性に関する分析を行う．次に，音声と非音声の音響空間上での違いに関して分析を行う．本節で分析に用いる音声コーパス，非音声コーパスは，付録 A に明示的に紹介されているものを除き，収録は収録条件 B（2.3.2 節参照）により行われており，音響解析条件は表 2.1 の設定 1 を用いるものとする．また，音響モデルには monophone HMM を用い，音素数は 28（pause を除く），各音素の状態数は 3，各状態の分布数は 1 とし，各正規分布の共分散行列は対角共分散行列とする．

3.5.1 年代及び性別依存音響モデル群の可視化

話者の年代及び性別の違いによる音響的特徴の違いを COSMOS 法により分析する．実験に用いる音声コーパスとして，子供，成人，高齢者が音声コマンド（孤立単語）を発声した音声コーパスを用いる．但し，年代により語彙セットは異なる．話者数は，子供が女性 50 名及び男性 50 名，成人が女性 67 名及び男性 67 名，高齢者が女性 66 名及び男性 55 名である．なお，子供は 12 歳以下，高齢者は 60 歳以上とする．1 話者あたりの発話数は，子供が 206 発話，成人が 472 発話，高齢者が 180 発話である．話者毎に特定話者音響モデルを学習し，全 354 話者を COSMOS 法により 2 次元空間上に写像した結果を図 3.4 に示す．

図 3.4 より，成人及び高齢者の年代において女性と男性がそれぞれ明瞭に分布が分離していることから，女性と男性では，音響的特徴が大きく異なることがわかる．これは，一般に不特定話者音響モデルを女性用と男性用に分けて作成する方法の妥当性を裏付けている．子供の年代では一般に声変わりが始まっていないため，女性と男性で音響的特徴に大きな違いがないことがわかる．また，各年代も性別と同様にそれぞれ明瞭に分布が分離していることがわかる．

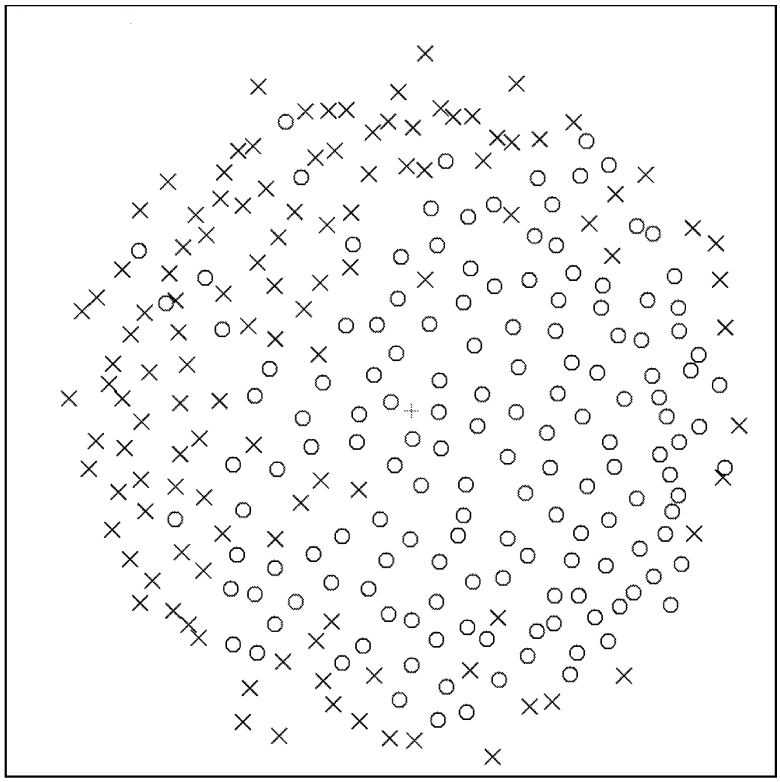


○: Adult/Female, □: Elder /Female, △: Kids/Female,
●: Adult/Male, ■: Elder /Male, ▲: Kids/Male

図 3.4 年代及び性別依存 COSMOS Map*

但し，各年代の分布が分離している要因として，年代の違いによる音響的特徴の他に，語彙セットが異なることも要因として考えられる．そこで，成人による日本語大規模音声コーパス JNAS[28] 及び高齢者の日本語大規模音声コーパス S-JNAS[29] を用い，同様の実験を行う（子供，成人，高齢者で同一の語彙を発声している音声コーパスは存在しないため，本論文では 3 つの年代の比較は行わない）．JNAS，S-JNAS は共にバランス文読み上げタスク及び新聞記事読み上げタスクの発話を収録している．本実験では，JNAS の男性 120 名及び S-JNAS の男性

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．



×：JNAS, ○：S-JNAS

図 3.5 JNAS/S-JNAS を用いた COSMOS Map*

150 名の新聞記事の読み上げ（JNAS，S-JNAS 共に 100 文）タスクにおける音声データを用いる．話者毎に特定話者音響モデルを学習し，全 270 話者を COSMOS 法により 2 次元空間上に写像した結果を図 3.5 に示す.

図 3.5 より，成人と高齢者の分布がそれぞれ明瞭に分離していることから，男性話者において，成人と高齢者では，音響的特徴が大きく異なることがわかる．これは，不特定話者音響モデルを成人用と高齢者用に分けて作成することが望ましいという過去の報告 [29] と一致している（但し，収録環境が異なることから，収録環境の違いによる分離である可能性も否定できない）．

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．

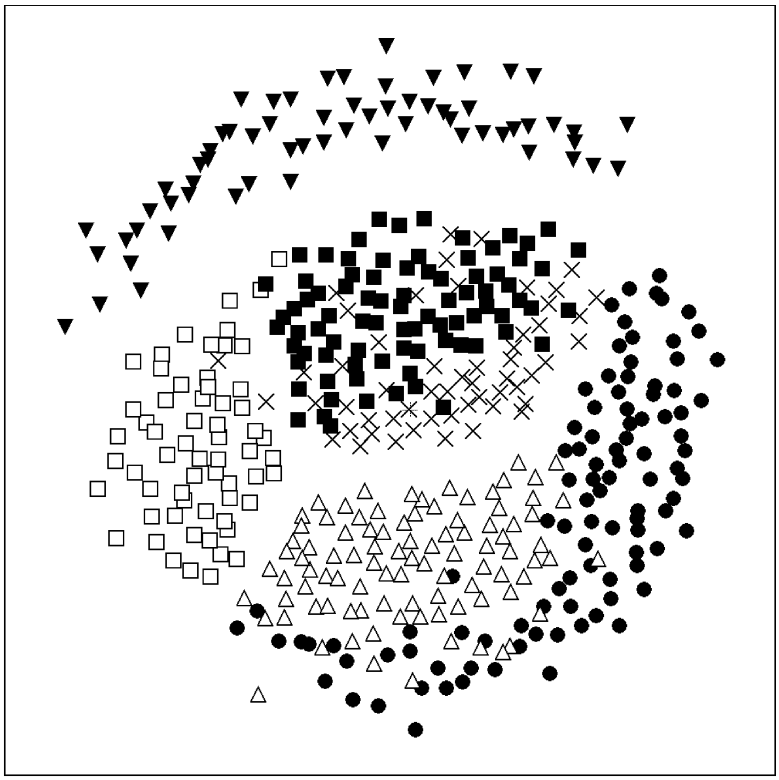
3.5.2 語彙依存音響モデル群の可視化

語彙セットの違いによる音響的特徴の違いをCOSMOS法により分析する．実験に用いる音声コーパスとして，カーナビゲーション用コマンド（COM），都市名（CITY），人名（NAME），外来語（FWORD），4桁数字（DIGIT），仮名（KANA）の6つの単語音声コーパスを用いる．それぞれ，車内での音声認識の利用を想定した孤立単語発声のタスクとなっている．各音声コーパスの話者数はCOMが女性67名，CITYが女性76名，NAMEが女性85名，FWORDが女性84名，DIGITが女性72名，KANAが女性75名である．1話者あたりの発話数はCOMが472発話，CITYが433発話，NAMEが256発話，FWORDが351発話，DIGITが140発話，KANAが220発話である．話者毎に特定話者音響モデルを学習し，全459話者をCOSMOS法により2次元空間上に写像した結果を図3.6に示す.

図3.6より各音声コーパスの分布が，DIGIT，KANA，COMとFWORD，NAMEとCITY，4つの分布に明瞭に分離していることがわかる．一般に不特定話者音響モデルを仮名発声用，数字発声用，単語発声用とに分けて作成する方法の妥当性を裏付けていることのみならず，同じ単語発声でも語彙セットが異なれば不特定話者音響モデルも分けて作成することの有効性を示唆しているものと考えられる．過去の報告では，語彙セットに依存し，不特定話者音響モデルを作成することの有効性が報告されている[64].

3.5.3 信号雑音比依存音響モデル群の可視化

信号雑音比の違いによる音響的特徴の違いをCOSMOS法により分析する．実験に用いる音声コーパスとして，成人が音声コマンド（孤立単語）を発声した音声コーパスを用いる．話者数は，女性48名及び男性45名である．1話者あたりの発話数は，128発話である．この音声に展示会雑音を信号雑音比（SNR:Signal-to-Noise Ratio)を変化させて重畳した．話者/SNR毎に特定話者音響モデルを学習し，のべ465話者をCOSMOS法により2次元空間上に写像した結果を図3.7に示す.

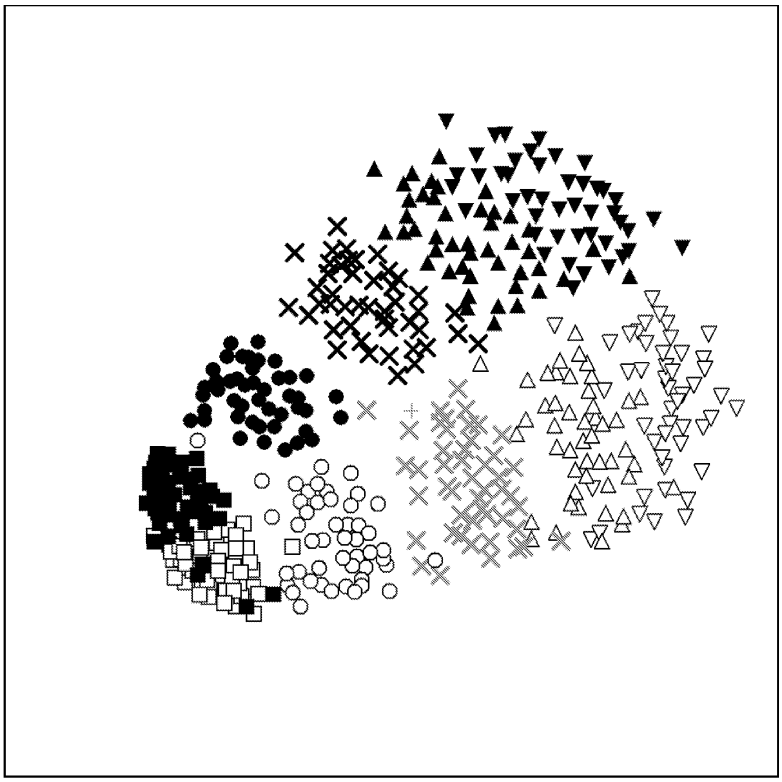


×：COM, □：DIGIT, ■：FWORD,
△：NAME, ●：CITY, ▼：KANA

図 3.6 語彙依存 COSMOS Map*

図 3.7 より，SNR が下がるにつれて，分布が縮小し，女性の分布と男性の分布が接近していることから，SNR が下がるにつれて，話者の音響的な拡がりは縮小傾向を示し，男女間の音響的な特徴の差異も小さくなることがわかる．一般に不特定話者音響モデルを SNR 別に分けて作成する方法の妥当性を裏付けている [80, 81]．また，SNR が低い条件では，性別非依存の不特定話者音響モデルでも性能上問題がない可能性を示している．

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．



□ : –10dB/Female, ○ : 0dB/Female, × : 10dB/Female,
△ : 20dB/Female, ▽ : 30dB/Female,
■ : –10dB/Male, ● : 0dB/Male, × : 10dB/Male,
▲ : 20dB/Male, ▼ : 30dB/Male

図 3.7 信号雑音比依存 COSMOS Map*

3.5.4 音声と非音声音響モデル群の可視化

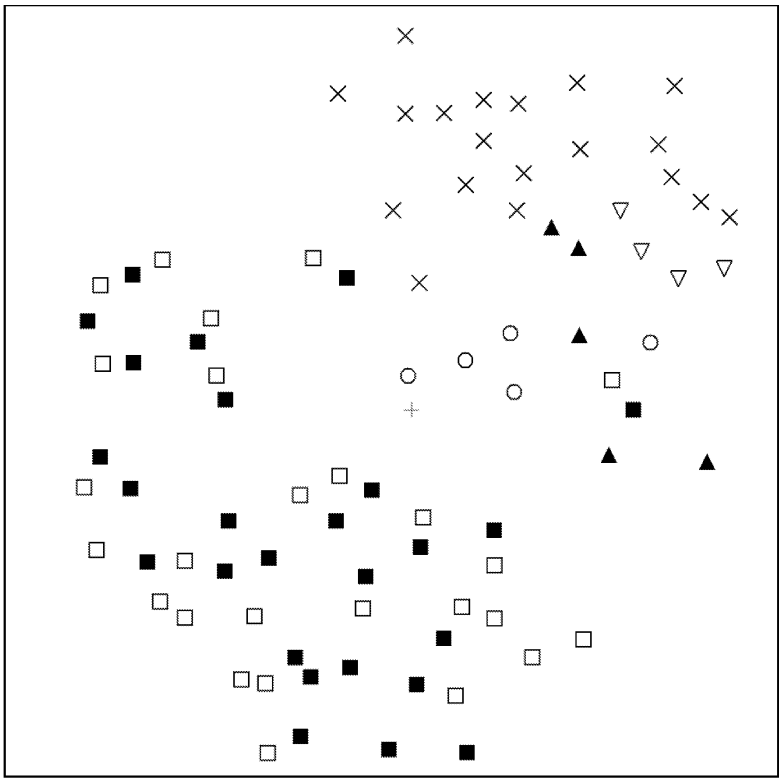
COSMOS 法は，人間の音声信号に止まらず，屋内で発生する雑音信号や，自動車，鉄道の音響信号などの非音声の可視化にも有効であることが知られている [71, 82]．音声と非音声（雑音等）の音響的特徴の違いを COSMOS 法により分析

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．

する．本実験では，非音声の音源として突発性雑音等の非定常な雑音を用いる．コップ，湯呑み等の食器類を置く，もしくは重ねる音（Hit），カーテン，窓，ドア，引き戸等の開閉音（Open/Close），階段や廊下を歩行時に発生する音（Footsteps），本やペン等が堅い床やカーペットに落ちる音（Drop Strongly/Weakly）等，計 33 種類の雑音源を用いる．各雑音は 10 回収録されており，5 回分を学習データとして用い，音響モデルを作成する．各雑音の時間長は，短いもので 100msec 以下，長いもので 1 秒程度であるため，長さに応じて状態数を 3〜12 とした．各状態の分布数は 1 とする．また，各正規分布の共分散行列は対角共分散行列とする．

音声用のモデルとして，成人男女各 67 名が音声コマンドを発声した（1 話者あたり 472 発話）音声コーパスから学習話者男女各 52 名を用いて作成した性別依存不特定話者音響モデルを用いる．音声には雑音を収録した環境での背景雑音を SNR25dB で重畳している（SNR は雑音収録時の入力ゲインから算出）．全雑音モデルと男女の不特定話者音響モデルの全音素モデルを合わせた計 89 個の音モデルを COSMOS 法により 2 次元空間上に写像した結果を図 3.8 に示す．

図 3.8 より，音声モデルの分布と雑音モデルの分布が明瞭に分離していることから，音声と本実験で用いたような非定常雑音では，音響的特徴が大きく異なることがわかる．なお，音声モデルの一部が非音声モデル群の分布上に配置されているが，これらの音声モデルは撥音及び促音である．これらの結果は，音響モデルを用いた音声と突発性雑音の識別が比較的容易であることを示唆しているものと考えられる．実際に，評価用の音声データ（男女 12 名によるコマンド発声）及び雑音データ（学習に用いていない 5 回分の収録音）を用い，音声と非音声の識別実験を行ったところ，互いに 100%の識別率を示した．但し，実環境には，音声との識別が困難な非定常雑音が存在すると考えられ，その場合には，音声モデル及び雑音モデルの分布は重なり合うことが予想される．



□: Speech/Female, ■: Speech/Male, ▲: Footsteps,
▽: Drop Sounds (Weakly), ○: Open/Close Sounds,
×: Drop and Hit Sounds (Strongly)

図 3.8 音声及び非音声 COSMOS Map*

3.6 まとめ

本章では，タスク間，タスク内の音響的変動や，実環境で発生する雑音と音声との違いを直感的に把握するために，音響モデルの分布を可視空間に写像することで，音響空間上でのデータ間の関連性やその拡がりを把握する手法（COSMOS 法）を提案した．また，主成分分析法や SOM 法等の従来法との比較を行い，統

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．

計的音響モデルが張る音響空間の分析手法としての提案手法の有効性を示した．次に，提案手法を用い，年代，性別，語彙，信号雑音比依存の音響モデル群や音声群と非音声（非定常雑音）群の音響モデルを 2 次元空間上に写像し，その結果を紹介した．実験では，年代，性別，語彙，信号雑音比の違いや，音声及び非音声の違いにより，音響空間上での配置や形状が異なることを確認した．

本手法を用いることで，大規模な音声コーパスの可視化が可能となるため，保持する全ての音声コーパス群のタスク間，タスク内の音響的変動や，音声と非音声（雑音）の音響空間上での関連性を直感的に把握することができる．データ間の関連性を把握することができれば，タスクが異なる音声コーパス間の移植性や，タスク内における性能劣化要因，音声と非音声の識別可能性等の予測が可能となると期待できる．特に，国内に現存する既存の音声コーパス群の将来的な活用 [86] における，音声コーパス間の関連性の分析手法として期待が大きく，本論文に関わる報告 [67, 68, 69, 72] だけでなく，他の研究機関でも研究開発が進められている [83, 84, 85]．また，可視化により直感的にデータの分布や構造を把握することができ，新たなアイデアの創出を補助する効果も期待できる．他のセンシング分野においてもその有効性は実証されており [10, 73, 82]，今後も，他分野への適用範囲が更に拡がることが期待される．

今後は，写像結果の精度を向上させるために，混合正規分布間距離や HMM 間距離に関する最近の報告 [76, 77, 78, 79] を取り入れ，写像手法の高精度化の検討を行う予定である．また，本手法のツール化，フリーツールとしての配布の要求もあることから，汎用性，ユーザビリティを考慮した，ツール開発を進め（現在開発されている分析ツールはあくまで研究用途），フリーツールとして配布することも検討したいと考えている．

4. COSMOS法を用いた既存音声コーパスの再利用性の判定

4.1 はじめに

低コストで高性能な音響モデルを作成するためには，既存の音声コーパスの有効な再利用が望まれる．しかしながら，2.4節で述べたように，音響的特徴が大きく異なる音声コーパスを学習に用いた場合，タスク依存性の問題から，性能が劣化する恐れがある．目的とするタスクに対する既存音声コーパスの再利用性を予め判定することができれば，性能劣化のリスクを回避し，同時に効果の高い（性能が高い）音響モデルを低コストで導入することが可能となる．適用効果の高い既存音声コーパスの選択手法として，開発用データ（目的タスクの少量音声データ）を用いて，認識性能や尤度から判断する手法が考えられ，これらの尺度を用いることで，特徴の似た再利用性（性能）の高い音声コーパスを選択することが可能である[47]．しかしながら，これらの尺度により選択された音声コーパスはあくまで，保持する音声コーパス群の中で最適な音声コーパスであり，選択された音声コーパスの再利用性を十分に保証するものではない．目的タスクの音声データが十分に存在すれば，そのデータを用いて学習した音響モデルの性能から，再利用性を判断することができるが，それでは意味がない．以上の議論から，目的タスクに対し高い性能を示す音響モデルを作成するためには，音声コーパスを選択する手法だけでなく，再利用性を予め判定する手法が求められる．再利用性を判断できてようやく，音声コーパスの構築にコストをかけるべきか否かの投資判断を下すことができることからも，音声認識アプリケーションの開発において，既存音声コーパスの再利用性の判定手法の重要性が伺える．

また，企業において，その投資判断を行う人物が必ずしも，音声認識技術に関して深い知識があるとは限らないため，投資判断を行う人物への説明材料として，誰でも直感的に再利用性を把握することができればなお良い．

本章では，3章で紹介したCOSMOS法を用いた視覚的再利用性判定手法を提案し[64, 66]，その有効性について論じる．まず，2節で提案手法に関する説明を行う．提案手法は，COSMOS法を用いて既存音声コーパス群を2次元平面上に可

視化し，各タスクの分布の重なり具合から再利用性を議論するものである．3 節では，2.4 節で用いた国内の既存音声コーパス群を用い，提案手法の有効性の検証を行う．また，検証実験では，開発用データ（目的タスクの少量音声データ）の規模による提案手法の頑健性に関しても調査を行う．開発用データの規模はその収集に音声データ収集と同様のコストがかかるため，人数，発話量ともにできるだけ少人数，少量であることが望ましい．最後に 4 節で本章のまとめを行う．

4.2 視覚的再利用性判定手法

提案する再利用判定手法のブロック図を図 4.1 に示す．ブロック C で，目的タスクの音声データとして，任意に選ばれた N 名の話者から T 秒の音声データを収集する．この際，実際の音声コーパス構築と比較し，話者の数，発話数は十分に少ないことを想定すると，収集コストは十分に少ないと言える（収集コストの詳細に関しては次章で述べる）．次に，ブロック D で，収集された音声データを用いて特定話者音響モデルを作成する．この際，特定話者音響モデルを作成するにあたり，全てのモデルパラメータを推定するために十分な発話量が得られていない可能性もあるため，少量の音声データでも比較的頑健にモデルパラメータを推定することのできる話者適応技術を用いることが望ましい（例えば MLLR 法 [20] 等）．次にブロック A で，既存音声コーパス群（M 個）から，タスク別に N 名の話者を無作為に選択し，更にそれぞれの話者から T 秒の音声データを無作為に抽出する．次に，ブロック B で抽出された音声データを用いて特定話者音響モデルを作成する．既存音声コーパスを用いての特定話者音響モデルの作成は，N 名以上，T 秒以上の音声データを用いることが可能であるが，目的タスクと条件が異なることによる不整合を排除するため，条件を統一している．特に，話者数に関しては，ブロック E で用いられる可視化手法（COSMOS 法）の評価関数が，写像対象である特定話者音響モデル群全体の写像誤差を最小とすることを尺度としていることから，話者数の多いタスクが全体の写像誤差の最小化に寄与してしまう恐れがあるため，各タスクの話者数を同一としている．次に，ブロック E でこれらの特定話者音響モデルに対して COSMOS 法を適用し，既存タスク群と目的タスクの COSMOS Map を作成する．この際，写像される特定話者音響モデルの

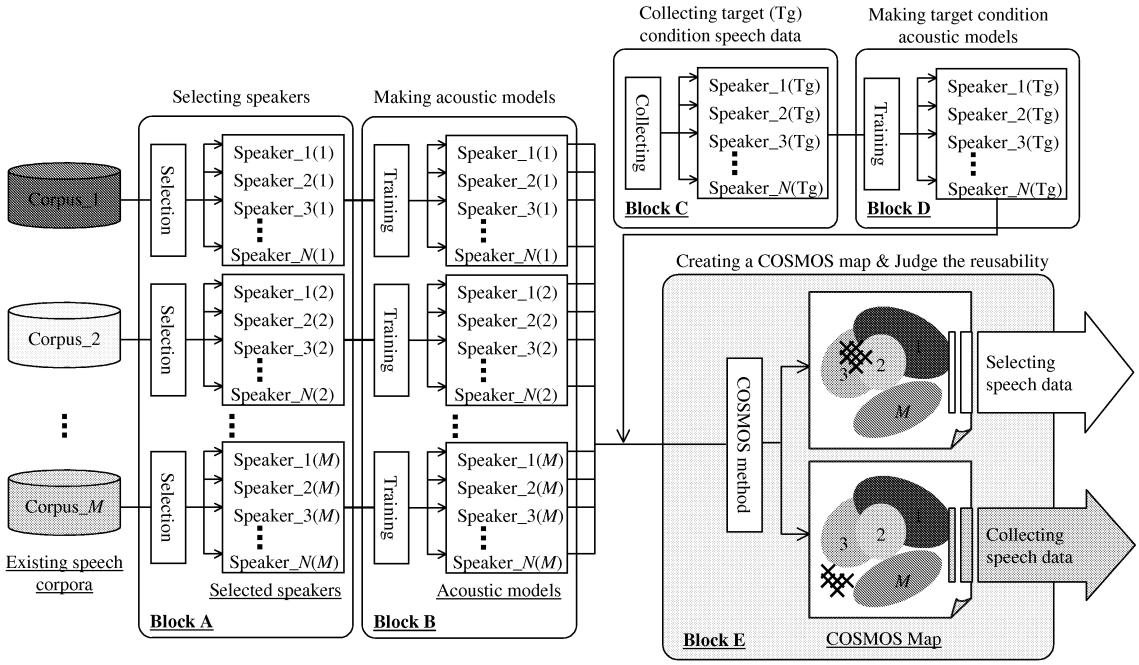


図 4.1 視覚的再利用性判定手法のブロック図

総数は $(M+1) \times N$ となる．最後に，目的タスクと重なり合う既存タスクの有無により，既存音声コーパス群の目的タスクに対する再利用性を判定する．判定では，重なり合う既存タスクがなければ，保持する既存音声コーパス群に再利用性の高い音声コーパスは存在しないと判断される．逆に，重なり合う既存タスクがあれば，保持する既存音声コーパス群に再利用性の高い音声コーパスが存在すると判断される．

4.3 検証実験

本節では，実験用音声コーパスとして，2.4節で紹介した全12タスクからなる国内の大規模音声コーパス群を用いる．本実験では，目的タスクとしてJNAS_news（ディクテーションタスク），CIAIR-drive_dialogH（車内での自然対話タスク），CIAIR-drive_balance（車内での一般的なタスク）の3つのタスクを用いた3通りの実験を行う．JNAS_newsを目的タスクとする場合は，CIAIR-drive_dialogH,

CIAIR-drive_balance を含む残りの 11 タスクを既存タスクとし，実験を行う．

まず，2.4 節でのクロスタスクの音素認識実験結果から本実験で定めた目的タスクにおけるクロスタスクの認識実験結果を抜粋し，表 4.1 に示す．表中の括弧及び太字で表記されている数値は，タスクが一致している場合の性能を示している．実際は，目的タスクのタスク依存音響モデルは大量の学習データなしには作成出来ないため，この性能は未知である．また，表 4.1 を図示した結果を図 4.2 から図 4.4 に示す．なお，表 4.1 中の太字は図中の灰色の結果と一致している．

再利用性という観点では，JNAS_news には JNAS_balance が，CIAIR-drive_dialogH には CIAIR-drive_dialogW が，CIAIR-drive_balance には JNAS_balance が高い再利用性を示している．実際に，JNAS_news に対する JNAS_balance の性能，CIAIR-drive_dialogH に対する CIAIR-drive_dialogW の性能は，タスクが一致している場合での性能に近い性能を示している．しかしながら，CIAIR-drive_balance に対する JNAS_balance の性能は，タスクが一致している場合での性能に比較し，著しく低く，再利用性が高いとは言えない．CIAIR-drive_balance で高い性能を示すためには，CIAIR-drive_balance 用に，新たに音声コーパスを構築する必要があることが分かる．既に述べたように，タスクが一致している場合の性能（目標性能）を予め知ることは出来ない．つまり，クロスタスクの音声認識実験では，既存のタスクにおける，目的タスクに最も近いタスクを選択することは出来ても，選択されたタスクの目的タスクに対する再利用性を保証することが出来ないということが分かる．

次に，各目的タスクに対し，提案手法を適用する．本実験においては目的タスク数は 1 であり，既存タスク数 M は 11 となる．まず，ブロック C で，$N = 10, 30, 50$ 名の話者から 1 人あたり $T = 30$ 秒の発話を目的タスク環境下で収録する．次に，ブロック D で，これらの収録発話を用い，話者毎に特定話者音響モデルを作成する．この際，各特定話者音響モデルには monophopne HMM を用い，音素数は 43，各 HMM の状態数は 3，状態毎の分布数は 1 としている．また，各正規分布の共分散行列は対角共分散行列とする．次に，ブロック A で，既存の音声コーパス群の各タスクから，$N = 10, 30, 50$ 名の話者を無作為に選択し，話者毎に $T = 30$ 秒

表 4.1 目的タスクに対する既存のタスク依存音響モデルの性能（上 : monophone, 下 : triphone）*

Acoustic model	Target Task		
	Target1(Jn)	Target2(CdH)	Target3(Cdb)
AAb	60.27	46.58	57.63
Jb	64.88	48.01	62.24
Jn	**(65.44)**	48.48	60.51
SJb	61.42	45.96	58.22
SJn	61.75	46.26	56.65
SJi	60.32	48.6	59.04
Cdb	58.54	48.48	**(69.66)**
CdA	46.22	51.52	55.02
CdW	44.91	51.49	52.05
CdH	45.87	**(51.84)**	52.93
Ca	51.13	44.94	50.95
Cs	54.8	45.86	49.68

Acoustic model	Target Task		
	Target1(Jn)	Target2(CdH)	Target3(Cdb)
AAb	63.88	45.91	59.96
Jb	68.96	49.78	66.40
Jn	**(69.95)**	48.24	62.61
SJb	67.76	47.73	62.70
SJn	68.48	47.61	62.19
SJi	51.07	40.89	43.36
Cdb	59.95	52.49	**(75.15)**
CdA	29.37	50.53	37.98
CdW	33.61	55.25	40.99
CdH	37.80	**(57.93)**	45.06
Ca	54.93	47.17	52.71
Cs	60.10	49.28	54.14

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

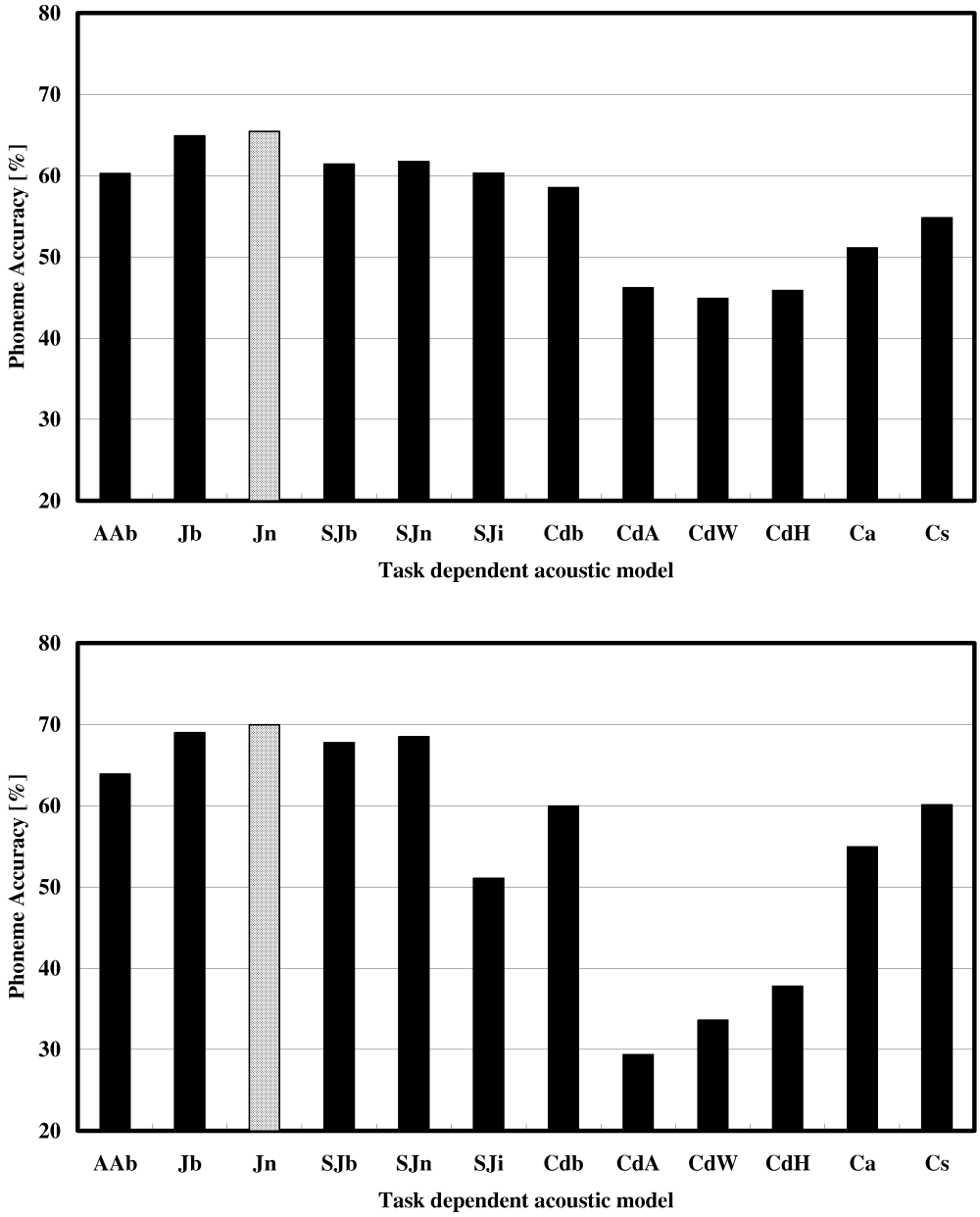


図 4.2 目的タスク JNAS_news に対する既存のタスク依存音響モデルの性能（上：monophone，下：triphone）*

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

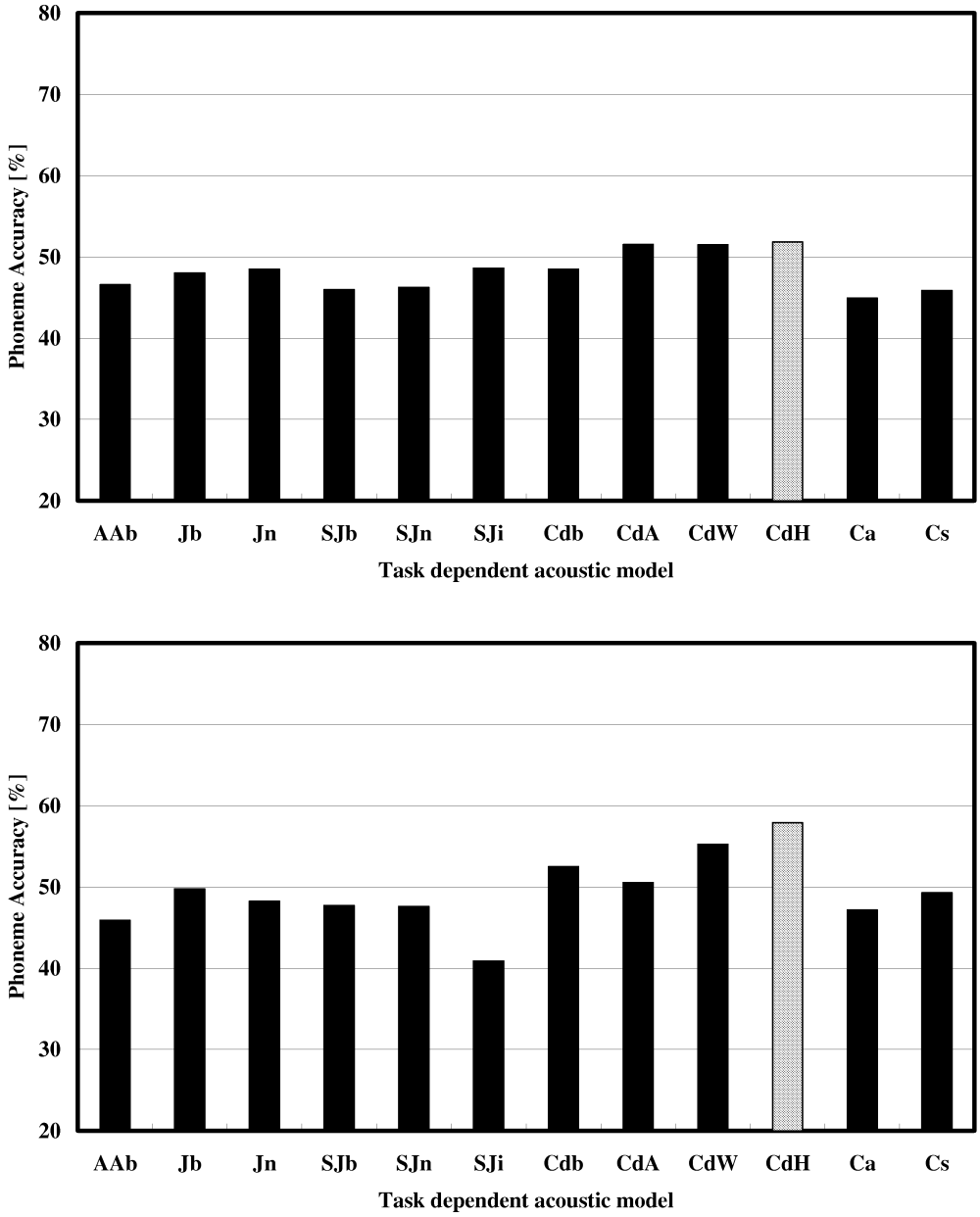


図 4.3 目的タスク CIAIR-drive_dialogH に対する既存のタスク依存音響モデルの性能（上：monophone，下：triphone）*

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

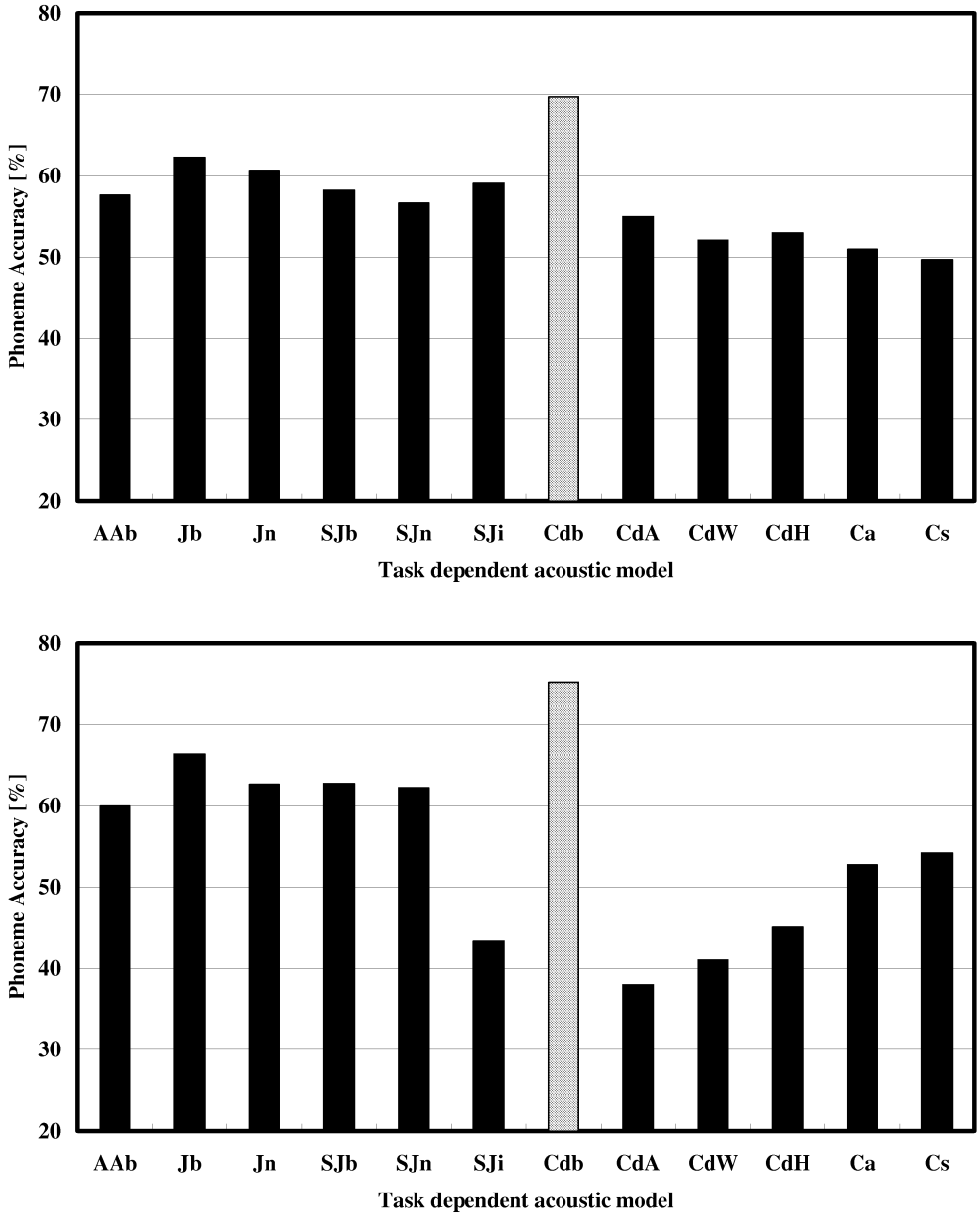


図 4.4 目的タスク CIAIR-drive_balance に対する既存のタスク依存音響モデルの性能（上：monophone，下：triphone）*

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

の発話を無作為に抽出する．次に，ブロック B で特定話者音響モデルを作成する．次に，ブロック E で，これらの特定話者音響モデルのモデル間距離を式 (3.35) に従い計算する．この時，特定話者音響モデルの総数は，$N = 10$ の場合は $(11+1)\times 10 = 120$，$N = 10$ の場合は $(11 + 1) \times 30 = 360$，$N = 10$ の場合は $(11 + 1) \times 50 = 600$ となる．最後に，COSMOS 法を適用することで，COSMOS Map を得ることが出来る．提案手法により得られる COSMOS Map の結果を図 4.5 から図 4.7 に示す．このとき，本実験では，目的タスクと既存タスクの話者数，発話数が同数，同量であるため，得られる COSMOS Map は目的タスクによらず同じものとなる．なお，各図の上図は，目的タスクのみの形状を定義しているものである．

図 4.5 から図 4.7 より，JNAS_news の分布と JNAS_balance の分布の大部分が重なり合っており，同様に，CIAIR-drive_dialogH の分布と CIAIR-drive_dialogW の分布も大部分が重なり合っていることが分かる．また，CIAIR-drive_balance の分布に対し，近くに配置されているタスク（JNAS_balance）はあるが，明確に重複するタスクが存在しないことが分かる．この結果は先ほどのクロスタスクの認識実験結果から得られた知見に一致していると言える．

ここで，提案手法の有効性を客観的に検証するために，提案手法で得られた COSMOS Map 上の目的タスク及び既存タスクの分布の位置関係と，目的タスクの評価データに対する既存タスクのタスク依存音響モデルの認識性能との相関を調査する．位置関係には，目的タスク及び既存タスクの分布の分布間距離を用いる．具体的には，各タスクにおけるタスク依存特定話者音響モデル群の写像後の座標値を用い，タスク間で座標間の距離を総当りで計算し，その平均をタスク間距離としている．この場合，重複が大きいタスク間の分布間距離は小さな値となり，反対に重複が小さいタスク間の分布間距離は大きな値となる．なお，3.4.2 節で述べたような，座標値からタスク毎に 2 次元正規分布を求め，Bhattacharrya 距離（式 (3.41)）のような分布間距離を用いることも当然可能である．但し，正規分布による距離では，写像誤差による悪影響を軽減することを目的としており，本実験のように N の数が小さい場合には，一般に写像誤差が小さいため，逆に精度が劣化する恐れがあることから，直接座標値を用いている．COSMOS 法は，距離の遠い要素間と比較し，距離の近い要素間の写像を重要視しているため，写像

結果に誤差が生じることから，遠近含めた関係で高い相関を示すことは難しいと考えられるが，COSMOS 法の適用範囲の検証も含め，調査を行う．話者数別の分布間距離を表 4.2 から表 4.4 に示す．

COSMOS Map 上の目的タスク及び既存タスクの分布の分布間距離と，目的タスクの評価データに対する既存タスクのタスク依存音響モデルの認識性能との相関を図 4.8 に示す．図 4.8 より，monophone では，どの目的タスクにおいても比較的高い相関を示している（平均で約 0.7）．なお，図 4.8 では，相関係数の大きさを比較することを目的としているため，相関係数は正の値を示しているが，実際は，負の値となっている（分布間距離が小さい程，高い性能を示すため）．triphone では，JNAS_news 及び CIAIR-drive_dialogH が目的タスクである場合の相関は 0.7 以上と高いが，CIAIR-drive_balance タスクが目的タスクである場合の相関が 0.5 前後と，低い．これは，COSMOS Map 上で近傍に位置する CIAIR の対話タスクの triphone HMM によるタスク依存音響モデルが，CIAIR-drive_balance タスクの評価データに対し，十分な性能を示すことができていないことが大きな要因である．2.4 節でも述べたように，triphone 化することで音素環境の偏りの影響が顕著に受け，タスク依存性が強く現れている．写像に monophohe HMM を用いているため，この影響を追従できていないことによる不整合と考えられる．

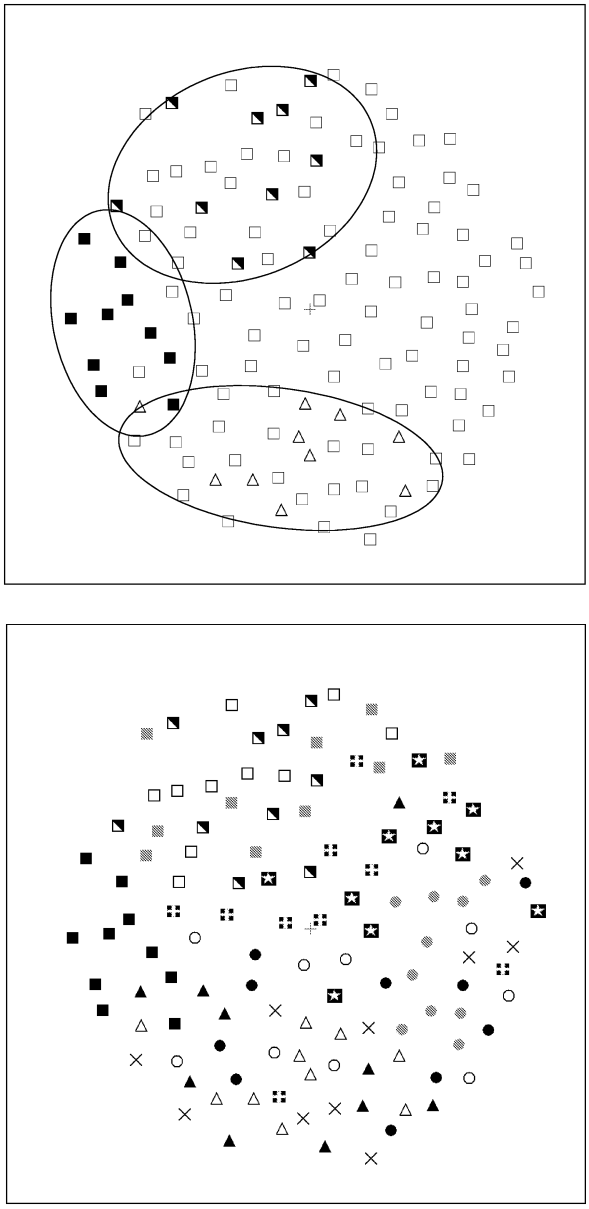
また，図 4.8 からは，話者数の違いによる相関係数の変動を確認することも出来る．話者数による大きな変動がないことから，提案手法の頑健性を確認することができる．話者数はコストにも大きな影響を与えるため，少ない話者数でも，高精度に再利用性を判定することができることは，コストの観点からも有効性の高い技術として期待できる．

4.4 まとめ

本章では，開発用データを用い，既存音声コーパス群のタスクと目的タスクの関係性を COSMOS 法に用いて可視化することで，既存音声コーパス群の目的タスクに対する再利用性を視覚的に判定する手法を提案した．従来は，既存音声コーパス群から，最も音響的に特徴の近い音声データを選択することは出来ても，その再利用性が十分に高いかどうかを判定する基準がなく，選択後作成された音響

モデルの性能を保証することが困難であったが，提案手法を用いることで，目的タスクと既存タスクとの COSMOS Map 上の分布の重なり具合から直感的に再利用性を把握することが可能となることを実験的に示した．また，提案手法の有効性を客観的に評価するために，クロスタスクの音声認識性能と，COSMOS Map 上の目的タスクと既存タスクの分布間の位置関係との相関性の調査を行い，monophone HMM におけるクロスタスクの認識性能と高い相関があることを示し，写像時のモデル構造と同一となる条件においては，提案手法の信頼性を確認することができた．今後は，タスク依存性が強く現れる triphone HMM におけるクロスタスクの認識性能との相関性を高めるための検討が必要と考えられるため，写像時のモデル構造に triphone HMM を用いることを検討する予定である．また，目的タスクの音響空間上での配置を把握するために必要な開発用データに関しては，収集する話者数による大きな変動がないことから，提案手法の頑健性を確認することができた．開発用データの収集にコストを割くことは一般的に敬遠されており，話者数はコストにも大きな影響を与えるため，少ない話者数でも，高精度に再利用性を判定することができることは，コストの観点からも有効性の高い技術として期待できる．

本章では，視覚的に目的タスクに対する既存音声コーパス群の再利用性を判定する手法を提案したが，本提案手法により得られる COSMOS Map は実験結果からも分かるようにクロスタスクの音声認識性能と比較的高い相関を示しており，triphone HMM への対応により，その精度を更に向上することが出来れば，性能予測技術への応用が期待できる．性能予測技術は，音声認識アプリケーションの性能保証や，仕様を決定する際に重要となる指標となるため，近年研究が活発に行われており [87, 88, 89, 90]，音声認識技術の今後の重要な技術分野となると考えられている．性能予測技術への応用を考えた場合，付録 B で紹介している物理量分析により得られる複数の物理量と合わせることで，更なる精度向上が期待できるため，今後，再利用性の判定技術の実用化とあわせ検討を行う予定である．



AAb:×, Jb:▲, Jn:△, SJb:●, SJn:○, SJi:◍,
Cdb:■, CdA:□, CdW:▩, CdH:◹, Ca:★, Cs:⁘

図 4.5 再利用性判定用タスク COSMOS Map（各タスク 10 名）*

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．

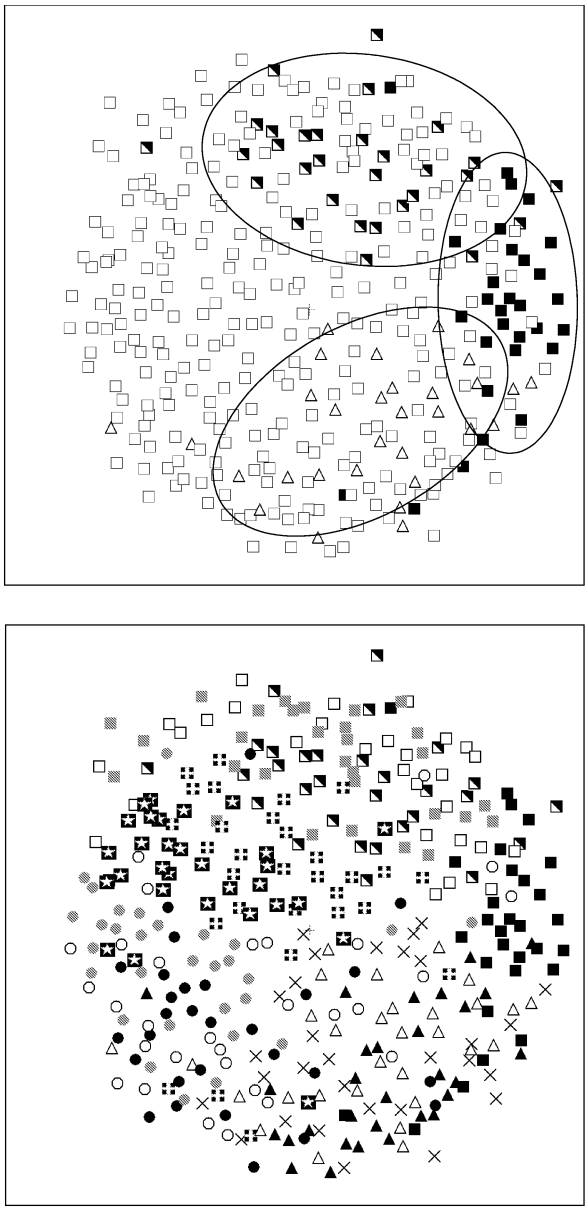
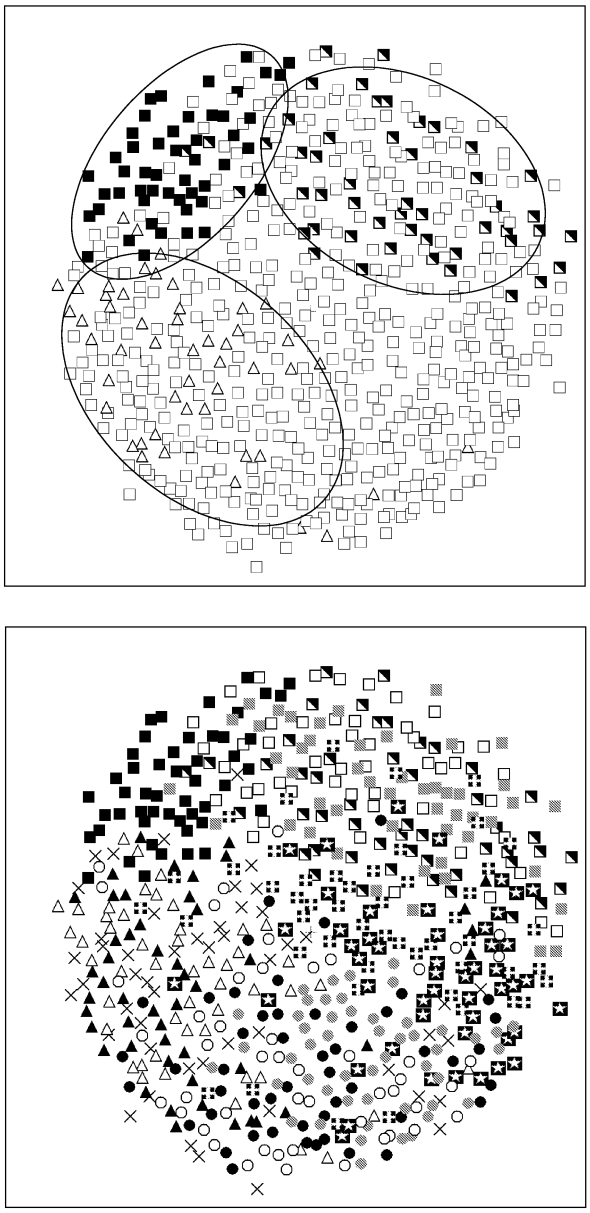


図 4.6 再利用性判定用タスク COSMOS Map（各タスク 30 名）*

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．



AAb:×, Jb:▲, Jn:△, SJb:●, SJn:○, SJi:◍,
Cdb:■, CdA:□, CdW:▩, CdH:◹, Ca:★, Cs:✚

図 4.7 再利用性判定用タスク COSMOS Map（各タスク 50 名）*

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．

表 4.2 再利用性判定用タスク COSMOS Map における目的タスク JNAS_news と既存タスクとの分布間距離（上：話者 10 名，中：話者 30 名，下：話者 50 名）*

Existing	Target Task		
task	Target1(Jn)	Target2(CdH)	Target3(Cdb)
AAb	0.1256	0.2100	0.1859
Jb	0.1188	0.1972	0.1493
Jn	(0.0000)	0.1926	0.1658
SJb	0.1218	0.1940	0.1826
SJn	0.1236	0.1765	0.1712
SJi	0.1332	0.1808	0.2208
Cdb	0.1658	0.1488	(0.0000)
CdA	0.2018	0.0783	0.1475
CdW	0.1963	0.0884	0.1624
CdH	0.1926	(0.0000)	0.1488
Ca	0.1440	0.1269	0.1609
Cs	0.1556	0.1384	0.2069

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

表 4.3　再利用性判定用タスク COSMOS Map における目的タスク CIAIR-drive_dialogH と既存タスクとの分布間距離（上 : 話者 10 名，中 : 話者 30 名，下 : 話者 50 名）*

Existing	Target Task		
task	Target1(Jn)	Target2(CdH)	Target3(Cdb)
AAb	0.1109	0.1892	0.1551
Jb	0.1036	0.1963	0.1495
Jn	(0.0000)	0.1912	0.1477
SJb	0.1301	0.1919	0.2089
SJn	0.1312	0.1920	0.2024
SJi	0.1432	0.1758	0.2196
Cdb	0.1477	0.1531	(0.0000)
CdA	0.1942	0.0952	0.1317
CdW	0.1947	0.0906	0.1491
CdH	0.1912	(0.0000)	0.1531
Ca	0.1552	0.1187	0.1700
Cs	0.1705	0.1342	0.2113

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

表 4.4　再利用性判定用タスク COSMOS Map における目的タスク CIAIR-drive_balance と既存タスクとの分布間距離（上：話者 10 名，中：話者 30 名，下：話者 50 名）*

Existing task	Target Task		
	Target1(Jn)	Target2(CdH)	Target3(Cdb)
AAb	0.0901	0.1771	0.1373
Jb	0.0870	0.1935	0.1388
Jn	(0.0000)	0.1778	0.1368
SJb	0.1272	0.1884	0.1862
SJn	0.1286	0.1760	0.1757
SJi	0.1475	0.1722	0.1930
Cdb	0.1368	0.1466	(0.0000)
CdA	0.1915	0.0964	0.1578
CdW	0.1944	0.0845	0.1648
CdH	0.1778	(0.0000)	0.1466
Ca	0.1429	0.1090	0.1566
Cs	0.1646	0.1305	0.1863

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

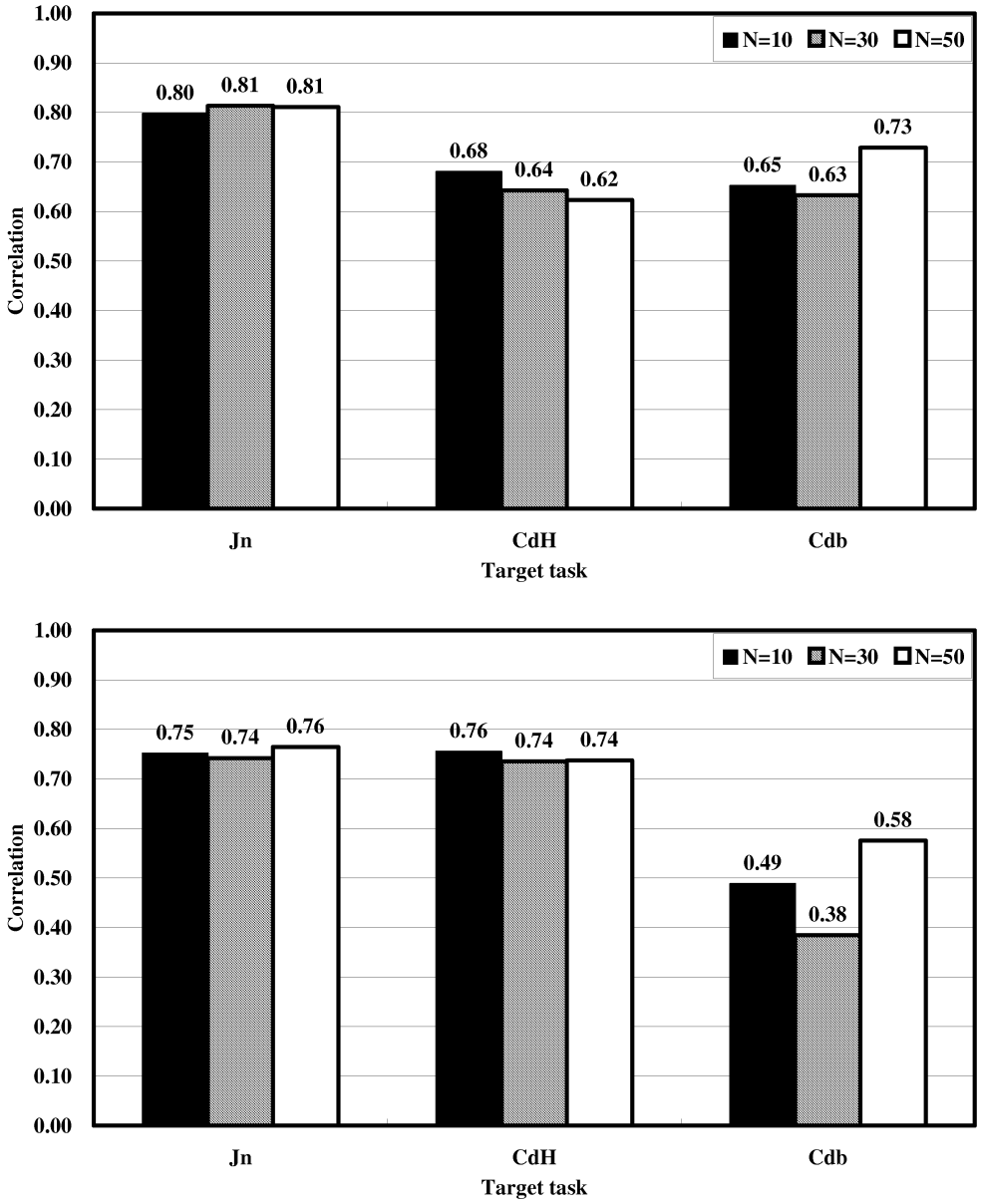


図 4.8 分布間距離とクロスタスクの音素認識実験結果との相関係数（上：mono-phone，下：triphone）*

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

5. COSMOS法を用いた効率的な音声コーパス構築手法

5.1 はじめに

保持する音声コーパス群に，目的のタスクに対して，再利用性の高い音声コーパスが存在しない場合，新規に音声コーパスを構築する必要がある．音声認識性能に大きな影響を与える音響モデルは，その作成に膨大な量の音声データを必要とするため，音声コーパスの構築には音声データの収録及びその後の編集作業（切り出し，ラベリング等）を含め，膨大な時間とコストを必要とする．収録にかかるコストはその規模や収録方法，内容に依存し，例えば，収録話者に負担が大きい環境下では更にコストが大きくなることがある．報告によれば，音声データの収録及び，その後の編集作業にかかるコストは，対話システムの開発コストの約50％を占めるとされている [91]．また，音声認識アプリケーション自体に変更はなく，目的タスクに特化した音響モデルの開発のみが必要な場合は，音声コーパス構築と音響モデル開発のコストのみがかかる．この場合，音響モデル開発コストをおおよそ100万円程度とし（200万円/1人月の開発者が半月従事することを想定），100名規模の収録を行う場合，音声コーパス構築にかかるコストは全体のコストの8割以上となる（音声コーパス構築のコストの詳細は以下を参照）．

収録は一般に，実際にシステムを運用しながらシステム利用者の発話を収録する手順と，発話内容を収録する側が予め決定しておき，発話者は与えられた発話リストを発話するという手順の二つに大別される．前者は，目的とするタスクの音声データに，より音響的に整合性の高い音声コーパスを構築することが出来るが，構築にかかるコストが大きい．これに対し，後者の手順は，音響的な整合性は前者の手順に劣るものの，音声データの収集期間、工数の見積もりが明確であると言う利点がある（詳細は2.3.2節参照）．実際に，企業での音声データ収集では，大学等の研究機関と比較し，予算と期間に対する制限が強いことから，後者の手順を踏むことが多い．本章ではこれらの背景に基づき，後者の手順に関してコスト削減を観点に議論を行う．なお，過去の報告では，音声コーパス構築にかかるコストを削減するために，全ての音声データを学習に用いず，一部の音声データのみを選択的に学習に用いる手法が提案されているが [92, 93, 94]，これ

らの技術は収録後の編集作業（主にラベリング）のコストを下げるためのものであり，後者の手順で必要とされる，収録自体にかかるコストを削減するものではない．

音声コーパスの構築は，まず予算に応じて収録話者数及び収録内容を決定した後に，収録話者を確保，選定し，収録のスケジュールを決定する．収録の際には，収録者の指示により，収録話者は予め決められた内容の発話を発声する．この時，各話者に対し，ある一定の拘束時間の中で収録が進められることになる．収録をスムーズに進め，拘束時間内に予定する量の音声データを収録するために，収録介助者（収録機材の設定や収録話者への指示，発話内容の確認等を行う者）を用意することが多い．また，収録室を所有していない場合，別途収録室の賃料がかかる等，収録話者に支払う謝礼以外にもコストが発生することとなる．例として，話者 120 名，1 人あたり 1 時間（200 単語発声程度）とした場合の音声データ収集にかかるコストの内訳を表 5.1 に示す．この例では，1 人あたり 1 時間（200 発声程度），1 日 6 人，計 20 日間の収録，収録補助者 1 人，発話者 1 人に対し謝礼を 12000 円（内 2000 円は交通費），収録室の賃料を 1 日あたり 100000 円，収録補助者 1 名の 1 日あたりの費用を 30000 円，その他諸経費 200000 円としている．結果，総コストは 120 人× 12000 円＋ 20 日× 100000 円＋ 20 日× 2 人× 30000 円＋ 200000 円＝ 4840000 円となり，1 人あたりのコストは 4840000 円／ 120 人 ≒ 40000 円となる．例からも分かるとおり，音声コーパス構築にかかるコストは収録話者数の増加に伴い，表 5.1 に示すそれぞれの項目の対してコストが増加するため，話者数に比例して収集コストも嵩むこととなる．しかしながら，不特定の話者が使用することを目的とした音声認識アプリケーションを想定した場合，できるだけ多くの話者を収録することが望ましく，コスト削減の実現が難しい．結果的に，音声コーパス構築にかかるコストは膨大となり，音声認識アプリケーションの開発にかかるコストを圧迫することとなる．本章では，これらの問題をふまえ，低コストで従来と同等以上の性能を実現する音声コーパス構築手法として，新規に音声データを収集する際に，収集対象の候補話者の少量音声データから COSMOS 法を用いて音声認識性能向上に寄与する話者を選択し，選択された話者の音声データを収集する技術を紹介する [70]．まず，次節で，COSMOS 法を

表 5.1 収録コスト内訳

Cost item	Cost (yen)	Amounts
Reward for speaker	12000	120 speakers
Rental for recording booth	100000	20 days
Reward for recording staff	30000	2 staffs & 20 days
Overhead expenses	200000	-

用い，音声認識性能向上に寄与する音響空間の分析を行う．3 節で，提案手法の具体的な説明を行い，4 節で，認識実験による提案法の有効性の検証を行う．最後に 5 節で本章のまとめを行う．

5.2 話者空間の分析

音声データの収集において，話者の選択は，無作為に行なわれることが多く，特徴の似た話者を多数収集している可能性がある．ある音響空間において，統計的に十分な量の音声データが既に収集されている場合，更に同じ空間の話者の音声データを収集しても音声認識性能の向上に対する寄与度は小さく，無駄にコストをかけることになる．音声データの収集前に音声認識性能の向上に寄与する話者かどうかの判定を行うことができれば，低コストで音声コーパスを構築することができる．また，音声認識性能の向上に寄与する話者を選択することで構築された音声コーパスは，無作為に話者を選択することで構築された音声コーパスと比較し，音声認識性能が向上することが期待される．本節では COSMOS 法を用い，特定話者音響モデル群を 2 次元平面上に写像し，特定話者音響モデル群が張る，話者空間の分析を行うことで，音声認識性能向上に寄与する話者空間について議論する．

5.2.1 実験用音声コーパス概要

本章での実験に用いる音声コーパスの概要について述べる．本実験では2つの日本語音声コーパスを実験に用いる．一つは孤立単語音声コーパス（IW-Corpus）であり，もう一方は連続音声コーパス（CW-Corpus）である（詳細は付録A参照）．IW-Corpusは，音声コーパスATR5240の全5240単語を175単語からなる複数の単語リストに分割したものを561名の男性話者が発話した音声コーパスである．また，この561名の話者は533名の学習用話者と，28名の評価用話者に分類される．次に，CW-Corpusは，音声コーパスATR503の全503の音素バランス文を20文からなる複数のリストに分割したものを1379名の男性話者が発話した音声コーパスである．また，この1379名の話者は1179名の学習用話者と，200名の評価話者に分類される．

それぞれ評価話者は無作為に抽出される．また，実験結果の信頼性を上げるために，評価話者の抽出は3回行うものとする．この時，学習話者は評価話者の変更に伴い，変更されるものとする．IW-Corpusの評価では評価話者が発声する175単語の孤立単語認識を行い，CW-Corpusの評価では音素認識実験を行う．音素認識実験では，日本語の音素配列構造の制約を用いた音素タイプライタを用いるものとする．IW-Corpusは，組み込み機器上で動作する音声認識アプリケーションに導入する音響モデルの学習用音声コーパスとして想定しており，音響モデルのサイズも限定されることから，音響解析条件には表2.1の条件1を用い，音響モデルの構成も一般に用いられるtriphone HMMではなくbiphone HMMを用いる．これに対し，CW-Corpusは，PC上で動作する音声認識アプリケーションに導入する音響モデルの学習用音声コーパスとして想定しており，音響モデルのサイズに強い制限がかかることがないことから，音響解析条件には表2.1の条件2を用い，音響モデルの構成も一般に用いられるtriphone HMMを用いる．各音声コーパスにおける音響モデルの構成を表5.2に示す．

表 5.2 HMM の構造

	IW-Corpus	CW-Corpus
Acoustic unit	biphone	triphone
Number of mixtures	8	32
Total state number	1263	3000

5.2.2 可視化実験

各音声コーパスにおいて，全話者（学習話者及び評価話者）を用い，特定話者音響モデルを作成する．この時，音響モデルには monophopne HMM を用い，各 HMM の状態数は 3，各状態の分布数は 1 とする．また，各正規分布の共分散行列は対角共分散行列とする．これらの特定話者音響モデルを，COSMOS 法により可視空間である 2 次元平面上に写像し，得られた配置図（以下，COSMOS Map）を図 5.1 に示す．図 5.1 より，各音声コーパスの特定話者音響モデルの分布を確認することが出来る．また，各音声コーパスにおいて，全話者を用いて不特定話者音響モデルを作成し，Closed 評価を行った結果，認識性能の低い話者（IW-Corpus では，単語正解精度 80%以下の話者，CW-Corpus では，音素正解精度 60%以下の話者）を，“ × ”で示し，それ以外の話者を“ ○ ”で示している．図 5.1 より，認識性能の低い話者が分布の中心から離れた周辺部分に位置していることが分かる．また，3 章での分析（3.4.3 節図 3.1 及び図 3.3）では，特殊な発話様式の話者が分布の周辺に位置しているという結果が得られている．以上より，特殊な発話様式の話者や，認識性能が低い話者が，COSMOS Map 上の話者分布において，分布の中心付近よりも周辺付近に位置する傾向にあると考えられる．

COSMOS 法により得られた 2 次元平面図上では，特徴の似た音響モデル同士は近傍に配置されていることが期待される．しかしながら，COSMOS 法をはじめとする全ての可視化手法は，例外なく写像誤差の問題を有している．図 5.2 は IW-Corpus の COSMOS Map であり，写像時に生じる誤差の例が示されている．図 5.2 では，COSMOS Map 上の話者分布で中心付近に位置する 1 名の話者及び周辺

部分に位置する 1 名の話者に対して，元の多次元空間上で近傍に存在する話者 30 名をそれぞれ線分で指している．図 5.2 より，周辺部分の話者が指す，多次元空間上での近傍 30 名の話者には，必ずしも COSMOS Map 上の話者分布での近傍話者とはなっておらず，中心付近に位置する話者も指していることが分かる．反対に，中心付近の話者が指す，近傍 30 名の話者は，おおよそ COSMOS Map 上の話者分布での近傍話者を指しており，周辺部分の話者を指すことはない．以上より，COSMOS Map が写像誤差を含み，その誤差は周辺部分程大きいことが分かるが，周辺部分の話者の多次元空間上での近傍話者が COSMOS Map 上の話者分布で中心部分にも存在していることから，COSMOS Map 上の話者分布で周辺部分に位置する話者は，周辺部分に位置する話者の特徴だけでなく，中心付近に位置する話者の特徴も含んでいると考えられる．反対に，COSMOS Map 上の話者分布で中心付近に位置する話者は，近傍の，中心付近に位置する話者の特徴のみを含んでいると考えられる．音響モデルの観点では，多くの話者に対し，高い性能を示すために，学習に用いる音声コーパスに含まれる話者の多様性はできるだけ大きいことが望ましい．よって，COSMOS Map 上の話者分布で周辺に位置する話者を選択し，選択された話者の音声データを収集することが，効率の高い音声コーパス構築手法と考えられる．

以上の議論から，COSMOS Map 上の話者分布の周辺部分が，音声認識性能の向上に寄与する話者空間であり，COSMOS Map 上の話者分布の周辺部分に位置する話者の音声データを収集することで，より効率的に学習用音声コーパスを構築することができると推察される．

5.2.3 音声認識実験

本節では前節での仮説を証明するために，IW-Corpus を用いた音声認識実験を行う．実験では比較のために，COSMOS Map 上の話者分布で周辺部分に位置する話者だけでなく，中心付近に位置する話者，無作為に抽出された話者の評価も

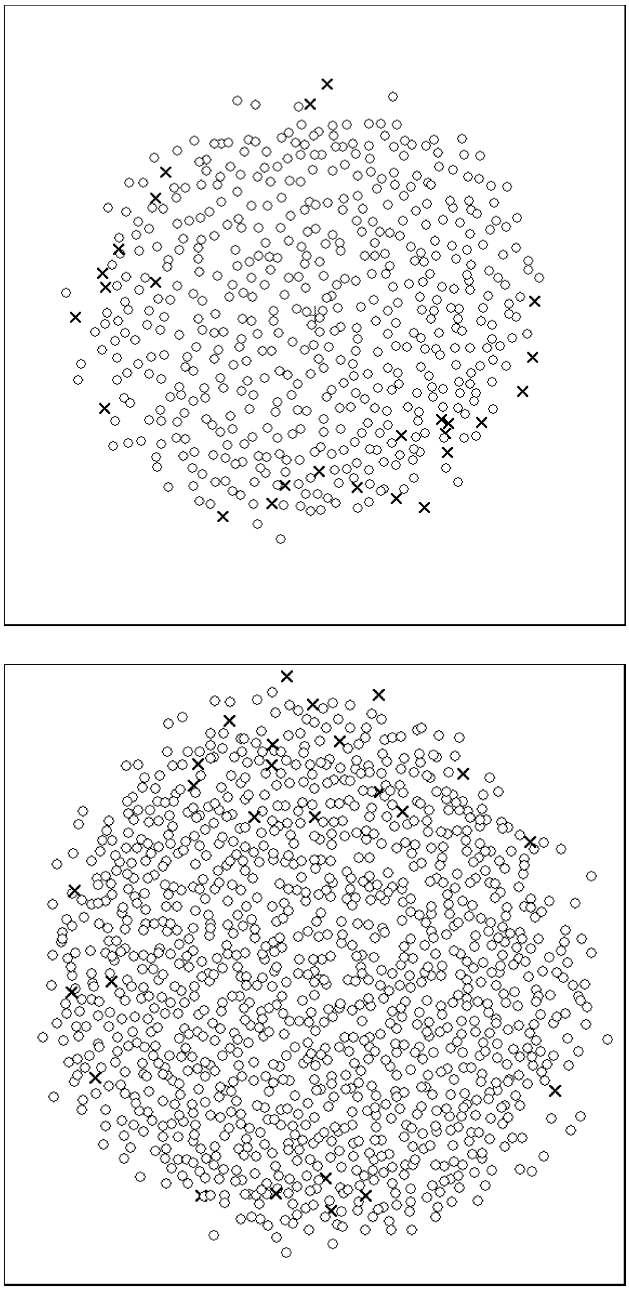


図 5.1 COSMOS Map (上：IW-Corpus, 下：CW-Corpus)*

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．

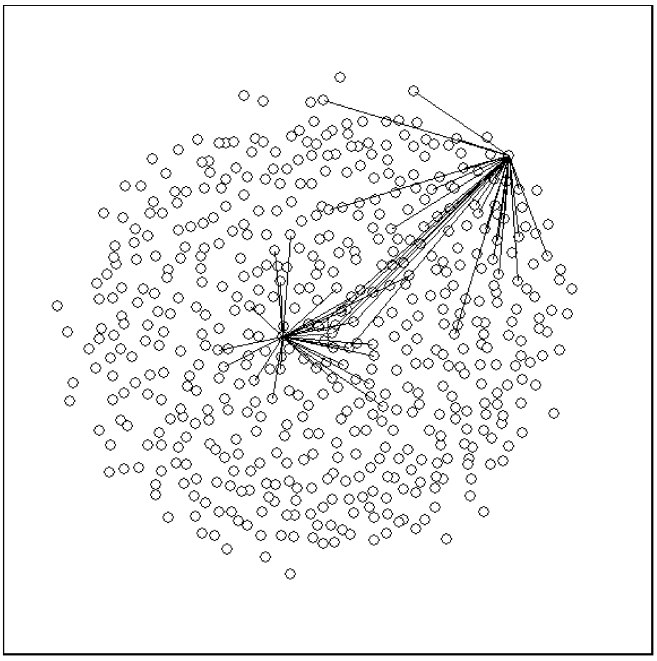


図 5.2 COSMOS Map 上の写像誤差*

行うものとする．まず，全学習話者 533 名から，COSMOS Map 上の話者分布で周辺部分に位置する話者，中心付近に位置する話者，それぞれ，分布の縁および分布の中心から N 名を選択する．選択された周辺部分の話者セットを Periphery，中心付近の話者セットを Center と呼ぶ．また，同数の話者を無作為に選択した話者セットを Random と呼ぶ．本実験では $N = 100, 150, 200, 250, 300$ とする．次に，それぞれの手法で選択された話者の音声データを用いて，不特定話者音響モデルを学習する．音響解析条件，音響モデルの構成は 5.2.1 節で述べた条件に従うものとする．また，評価は 5.2.1 節で述べたとおり，発話内容と一致する 175 単語の孤立単語認識とする．選択された話者数と認識性能の関係を図 5.3 及び表 5.3 に示す．表 5.3 は，3 通りの評価話者セットでの認識実験結果であり，図 5.3 の結果は表 5.3 の結果の平均を表したものである．図 5.3 及び表 5.3 の結果から，Center の話者セットで学習された音響モデルの性能は，話者数が同じであるにもかかわ

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．

らず，Periphery 及び Random の話者セットで学習された音響モデルよりも著しく性能が低いことが分かる．反対に，Periphery の話者セットで学習された音響モデルは，Center 及び Random の話者セットで学習された音響モデルよりも高い性能を示していることが分かる．また，図 5.3 の結果から，200 名の話者からなる Periphery の話者セットで学習された音響モデルが全話者で学習された音響モデルの性能を上回る性能を示していることが分かる．

以上の結果から，前節の仮説である，COSMOS Map 上の話者分布の周辺部分が，音声認識性能の向上に寄与する話者空間であり，効率的に音声コーパスを構築する手法として，COSMOS Map 上の話者分布の周辺部分に位置する話者の音声データを収集することの妥当性が証明されたと言える．また，高い認識性能を実現するためには，話者の多様性を確保することが重要であり，COSMOS Map 上の話者分布において，話者分布を覆うように，周辺部分に位置する話者を選択することが，多様性のある話者セットを構築することと等価であることが証明されたと言える．

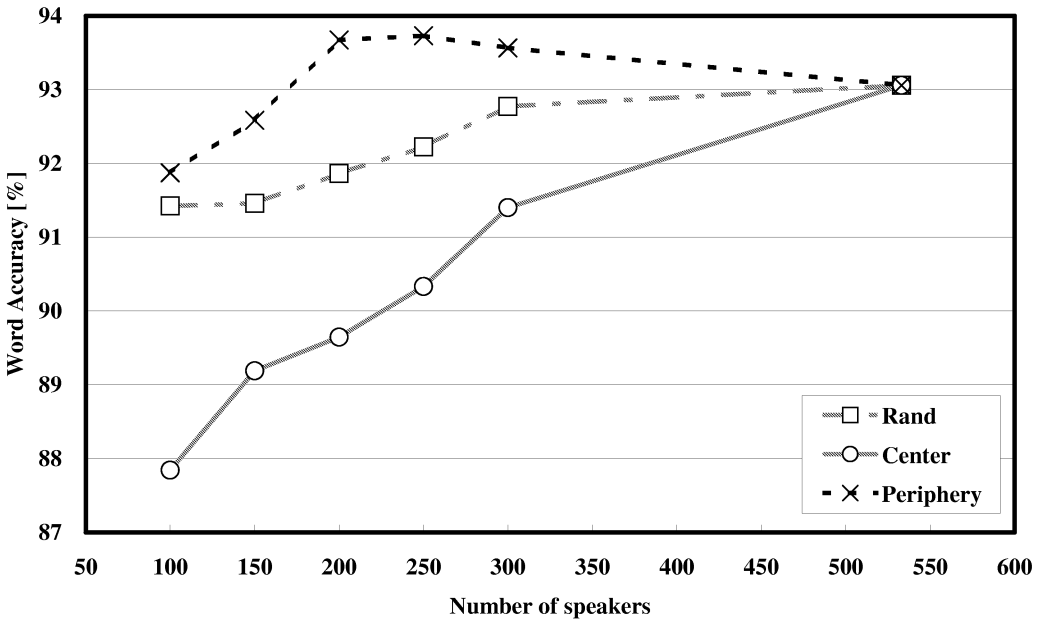


図 5.3 音声コーパスサイズの増加における性能の変動（IW-Corpus）

表 5.3 評価セット毎の音声コーパスサイズの増加における性能の変動（IW-Corpus）

Evaluation speaker set 1

Training	Number of training speakers					
speaker-set	100	150	200	250	300	533
Random	93.28	93.55	93.99	94.18	94.64	94.85
Center	90.84	92.08	92.45	92.95	93.59	94.85
Periphery	93.50	94.13	95.05	95.22	95.14	94.85

Evaluation speaker set 2

Training	Number of training speakers					
speaker-set	100	150	200	250	300	533
Random	89.43	89.27	89.94	90.11	91.29	91.62
Center	85.06	86.29	87.05	87.96	89.54	91.62
Periphery	89.17	90.34	92.52	92.19	91.86	91.62

Evaluation speaker set 3

Training	Number of training speakers					
speaker-set	100	150	200	250	300	533
Random	91.56	91.56	91.65	92.39	92.39	92.71
Center	87.64	89.20	89.44	90.09	91.08	92.71
Periphery	92.95	93.28	93.45	93.78	93.70	92.71

5.3 効率的な音声コーパス構築手法

前節の結果に基づき，COSMOS 法を用いて効率的に音声コーパスを構築する方法を提案する．本節ではまず，提案手法の概要を示し，以降は提案手法の有効性を証明する実験を行う．実験に用いる音声コーパスは前述の IW-Corpus 及び

CW-Corpus である．

5.3.1 提案手法概要

提案手法は，音声コーパス構築のための収録対象となる候補話者の少量の音声データから，COSMOS 法を用いて音声認識性能向上に寄与する話者かどうかを判定し，寄与の大きいことが期待できる話者を選択することで，低コストで音声コーパスを構築する技術である．図 5.4 に提案手法のブロック図を示す．まず，ブロック A で，音声コーパス構築のための収録対象となる候補話者から少量の音声データを収集する．発話内容（語彙）は，目的とするタスクに依存する．ブロック E での音声データ収集と異なり，収録室ではなく，インターネットや電話を介しての収集を想定しているため，ブロック E での音声データ収集と比較し，収集にかかるコストは，十分に小さいことを前提としている．また，話者の多様性を確保するために，ブロック A ではできるだけ多くの候補話者の音声データを収集することが望ましい．なお，予備収録という概念は，1.2 節で紹介した開発用データ（図 1.1 中の Development data）と似ており，実際に，1 話者あたりの発話量は少量であることが望ましいという点では同様であるが，話者の多様性を確保するという点で，話者数の捉え方が大きく異なることがわかる（開発用データはできるだけ少人数が望ましいとされている）．次にブロック B で，ブロック A で収集した候補話者の音声データをもとに，候補話者毎に MLLR 法 [20] を用いて話者適応音響モデルを作成する．話者適応音響モデルは COSMOS Map の生成に必要となる特定話者音響モデルの近似モデルとして作成されるものである．MLLR 法を用いることで，音声データに出現しない音素の適応が可能となるため，少量の音声データでも，高い精度で特定話者音響モデルを近似するモデルが作成できると期待される．次にブロック C で，全ての話者適応音響モデルを COSMOS 法により 2 次元平面上に写像し，COSMOS Map を生成する．COSMOS Map 生成の際に，式 (3.36) における重み $\omega(i,r)$ 及び $\omega(j,r)$ は，目的とするタスクの語彙から求めるものとする．次にブロック D で，COSMOS Map 上の話者分布の周辺部分の話者 N 名を選択し，最後にブロック E で，選択された話者に対して再度音声データの収集を行うことで，目的とするタスクの音声コーパスを得ることができる．

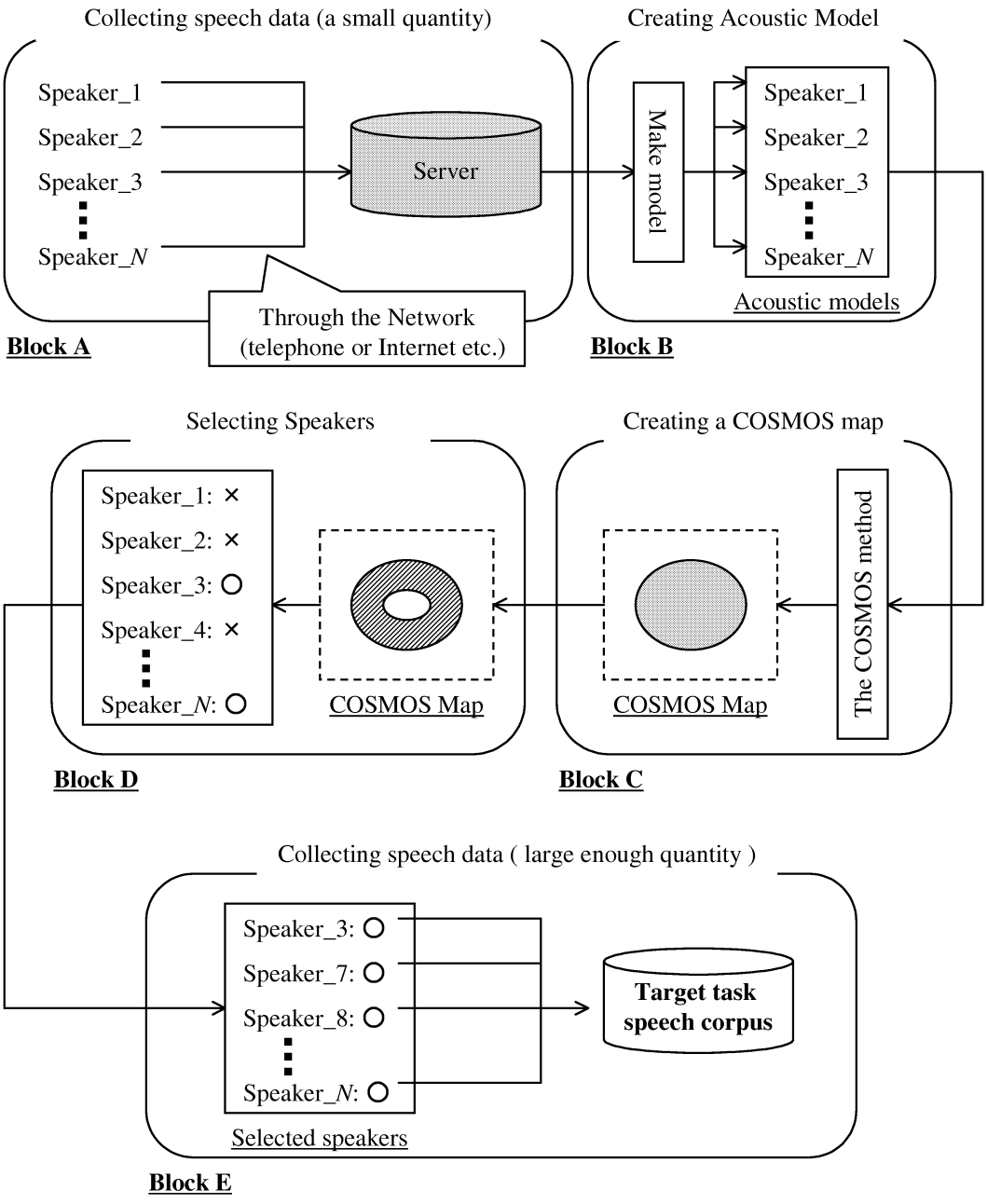


図 5.4 音声コーパス構築のブロック図

5.3.2 実験概要

IW-Corpusを用いた実験では，収録候補話者533名から10発話（単語）を収集し（ブロックA），収集した音声データを元に話者適応音響モデルを作成する（ブロックB）．次に，COSMOS法を用いて，話者適応モデルを2次元平面に写像することで，COSMOS Mapを得る（ブロックC）．COSMOS Map上の話者分布で周辺部分に位置する話者N名を選択し（ブロックD），選択された話者から更に165発話を収集する（ブロックE）．

CW-Corpusを用いた実験では，収録候補話者1179名から5発話（文）を収集し（ブロックA），収集した音声データを元に話者適応音響モデルを作成する（ブロックB）．次に，COSMOS法を用いて，話者適応モデルを2次元平面に写像することで，COSMOS Mapを得る（ブロックC）．COSMOS Map上の話者分布で周辺部分に位置する話者N名を選択し（ブロックD），選択された話者から更に20発話を収集する（ブロックE）．

最後に，提案手法の有効性を確認するために，IW-Corpus，CW-Corpusそれぞれにおいて，提案手法により選択された話者セットの音声データを元に学習した不特定話者音響モデルの認識性能と，従来の無作為に選択された話者セットの音声データを元に学習した不特定話者音響モデルの認識性能とを比較する（5.3.4節）．

5.3.3 可視化実験

少量の音声データからMLLR法を用いて作成した話者適応音響モデルの，特定話者音響モデルの近似モデルとしての精度を検証するために，話者適応音響モデル及び特定話者音響モデルから作成されたCOSOMS Mapの比較を行う．以下，特定話者音響モデルによるCOSMOS MapをSD-model COSMOS，話者適応音響モデルによるCOSMOS MapをAdapted-model COSMOSと呼ぶ．IW-CorpusとCW-Corpusそれぞれの SD-model COSMOS及びAdapted-model COSMOSを図5.5に示す．図中では，SD-model COSMOS上の話者分布の周辺部分に位置する話者を“■”で示し，同一の話者を，Adapted-model COSMOS上でも同様に“■”で示す．なお，“■”で示される話者数は，IW-Corpusでは100名，CW-Corpusで

は200名としている．SD-model COSMOS及びAdapted-model COSMOSを比較すると，両者の間に差異はあるが，SD-model COSMOS上で話者分布の周辺部分に位置する話者の多くが，Adapted-model COSMOS上においてもその周辺部分に位置していることがわかる．このことから高々10単語，5文程度の音声データで適応化された話者適応音響モデルでも，その話者が位置する音響空間の同定が可能であると考えられる．また，少なくとも音声認識性能向上に寄与する話者であるかを判断するには十分な表現力を持つモデルであると期待できる．

5.3.4 音声認識実験

提案手法で選択された話者からなる話者セット（Proposed）を用いて学習された不特定話者音響モデルの性能評価を行う．また，無作為に選択された話者からなる話者セット（Random）を用いて学習された不特定話者音響モデルとの比較も行う（IW-Corpusにおいては5.2.3節のRandomと同様の性能）．音響解析条件，音響モデルの構成は5.2.1節で述べた条件に従うものとする．また，評価は5.2.1節で述べたとおり，IW-Corpusでは，発話内容と一致する175単語の孤立単語認識を行い，CW-Corpusでは，音素認識実験を行う．選択された話者数と認識性能の関係を図5.6及び表5.4及び表5.5に示す．表5.4及び表5.5は，3通りの評価話者セットでの認識実験結果であり，図5.6の結果は表5.4及び表5.5の結果の平均を表したものである．図5.6及び表5.4，5.5の結果から，IW-Corpus及びCW-Corpusのどちらの音声コーパスを用いた実験においても，Proposedの話者セットを用いて学習された音響モデルの性能は，話者数によらずRandomの話者セットを用いて学習された音響モデルを上回る性能を示していることが分かる．これらの結果は，5.2.3節での検証実験と同様に，COSMOS Map上の話者分布の周辺部分が，音声認識性能の向上に寄与する話者空間であることを証明していると言える．また，高々10単語，5文程度の音声データで適応化された話者適応音響モデルでも，音声認識性能向上に寄与する話者であるかを判断するには十分な表現力を持つことを示した．更に，提案手法が孤立単語認識や連続音声認識といったタスクに依存せず，有効性のある手法であることが証明された．

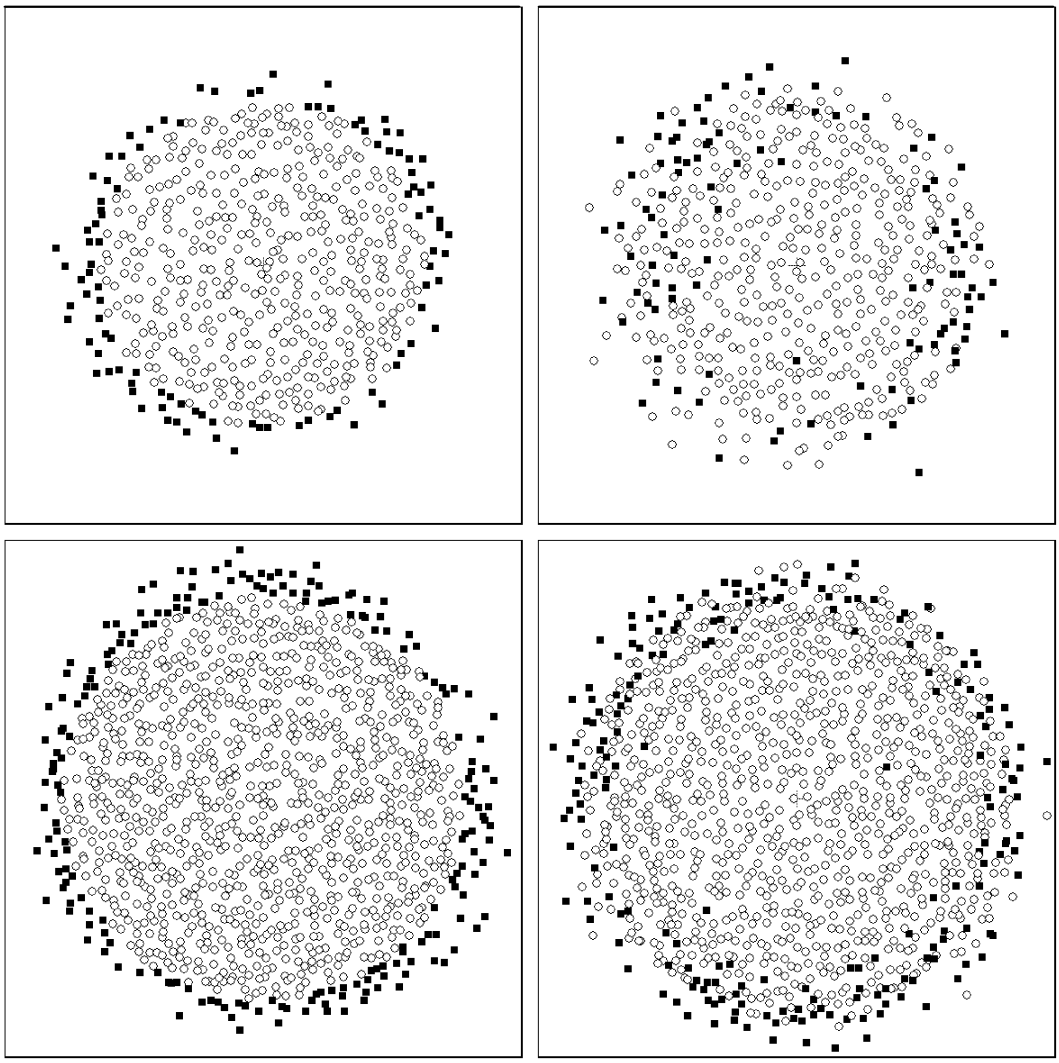


図 5.5 SD-model COSMOS と Adapted-model COSMOS の比較（上：IW-Corpus，下：CW-Corpus，左：SD-model COSMOS，右：Adapt-model COSMOS）*

コスト削減の観点においては，IW-Corpus を用いた実験では，200 名の話者セットで学習された音響モデルの性能が，全話者 533 名の話者セットで学習された音響モデルの性能を上回っており，62%以上のコスト削減を達成している．また，

*相対関係のみが重要であるため，軸は明記されていない．なお，原点を中心とし，原点からの距離の平均が 1 となるように正規化されている．

CW-Corpus を用いた実験では，500 名の話者セットで学習された音響モデルの性能が，全話者 1179 名の話者セットで学習された音響モデルの性能を上回っており，57%以上のコスト削減を達成している．以上の結果より，提案手法が従来と比べ，低コストで効果的な音声コーパスを構築する手法として有効性の高い手法であることを示したと言える．

IW-Corpus 及び CW-Corpus のどちらの結果においても，全話者セットで学習された音響モデルの性能を，提案手法で選択された，より少ない話者セットで学習された音響モデル上回っている．これは，話者分布の周辺のデータを集めることで，相対的に，モデルパラメータであるガウス分布の裾野が持ち上がり，分布境界の精度が向上しているためと考えられる．反対に，話者分布の中心のデータが増えると，相対的にモデルパラメータであるガウス分布の中心が持ち上がり，裾野が下がるため，分布境界の精度が劣化すると考えられる．

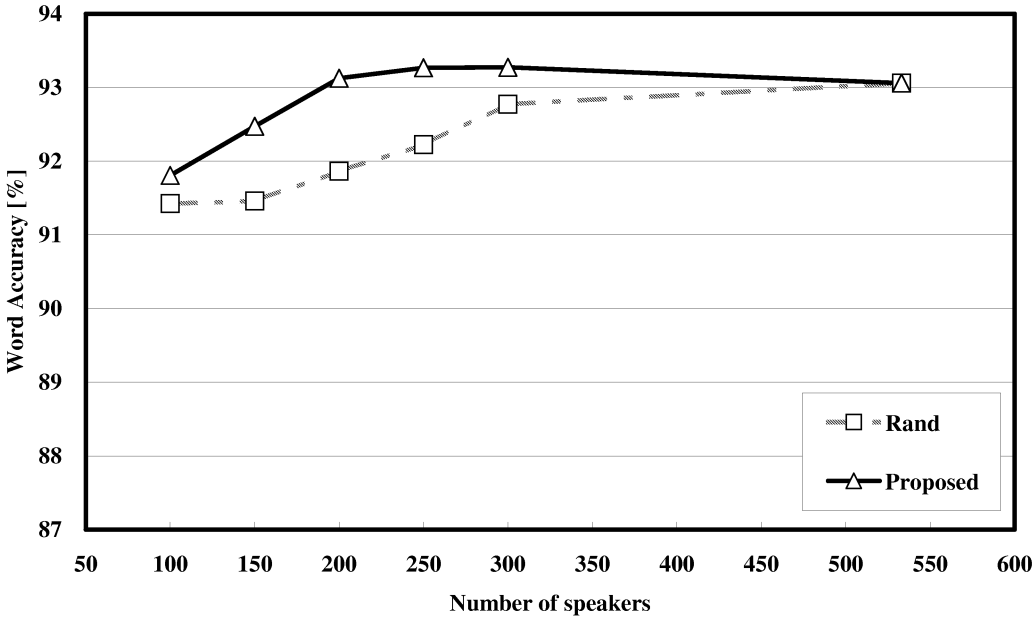
5.4 まとめ

本章では，低コストで従来と同等以上の性能を実現する音声コーパス構築手法を提案した．まず，COSMOS 法を用い，音声認識性能向上に寄与する音響空間の分析を行い，COSMOS Map 上の話者分布の周辺部分が，音声認識性能の向上に寄与する話者空間であることを確認した．また，高い認識性能を実現するためには，話者の多様性を確保することが重要であり，COSMOS Map 上の話者分布において，話者分布を覆うように，周辺部分に位置する話者を選択することが，多様性のある話者セットを構築することと等価であることを実験的に示した．

次に，提案手法として，収集対象の候補話者の少量音声データから，COSMOS 法を用いて音声認識性能向上に寄与する話者を予備選択し，選択された話者の音声データを収集することで，より低コストで効果的な音声コーパスを構築する手法を紹介し，音声認識実験において，提案手法の有効性を示した．

本章では話者選択のための少量音声データの収集方法について論じていない．コストの低い音声データの収集方法としては，電話やインターネットを介しての収集が考えられる．今後は本手法の実用化を目指し，話者選択用の音声データの収集に電話やインターネットを介した場合について検証を行う予定である．また，

予備選択における収録の際の，話者の負担を軽減するためには，発話数を減らすことが考えられる．本章で用いたMLLR法と比較し，より少ない発話数で高い適応効果を示す話者適応技術に，MAPLR法[95]やEigen-MLLR法[96]等が挙げられるが，話者の負担軽減は本手法の実用化するための重要な課題の一つであることから，これらの適応技術の有効性に関しても，検証を行う予定である．また，認識性能の観点では，予備選択において何割程度の話者を選択するかが重要となる．本章での実験結果からは，予備選択により集められた全話者数に対して約5割程度の話者数でその性能がピークに達しているが，選択する話者数を決定するための基準に関しては検証する余地があると考えられことから，今後の取り組みとして検証を行う予定である．



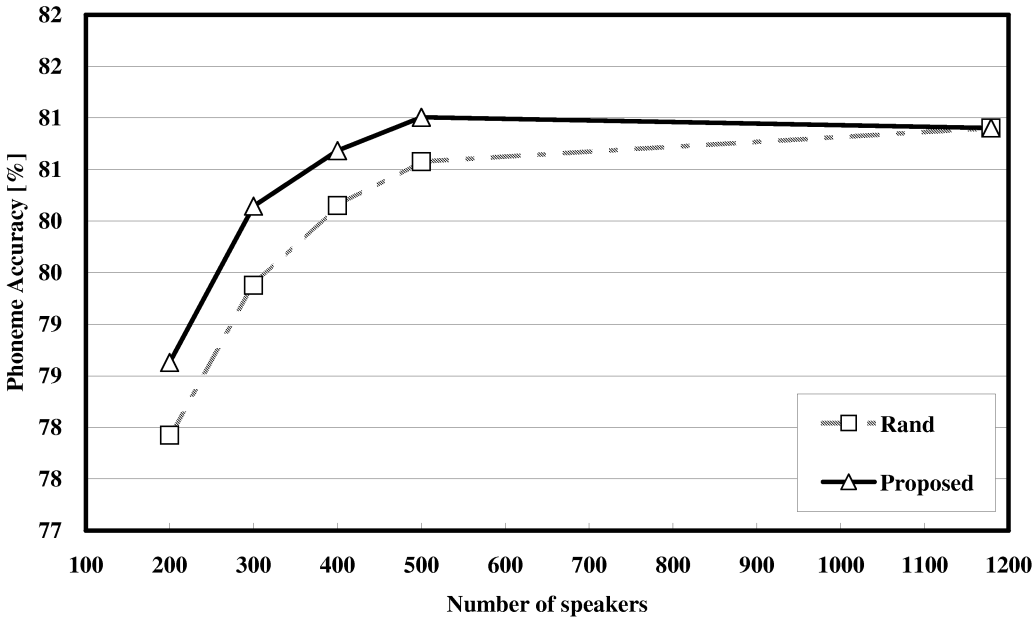


図 5.6 音声コーパスサイズの増加における性能の変動（上 : IW-Corpus, 下 : CW-Corpus）

表 5.4 評価セット毎の音声コーパスサイズの増加における性能の変動（IW-Corpus）

Evaluation speaker set 1

Training	Number of training speakers					
speaker-set	100	150	200	250	300	533
Random	93.28	93.55	93.99	94.18	94.64	94.85
Proposed	93.67	94.33	94.89	94.97	95.00	94.85

Evaluation speaker set 2

Training	Number of training speakers					
speaker-set	100	150	200	250	300	533
Random	89.43	89.27	89.94	90.11	91.29	91.62
Proposed	89.37	90.62	91.77	91.86	92.20	91.62

Evaluation speaker set 3

Training	Number of training speakers					
speaker-set	100	150	200	250	300	533
Random	91.56	91.56	91.65	92.39	92.39	92.71
Proposed	92.38	92.46	92.71	92.96	92.63	92.71

表 5.5　評価セット毎の音声コーパスサイズの増加における性能の変動（CW-Corpus）

Evaluation speaker set 1

Training	Number of training speakers				
speaker-set	200	300	400	500	1179
Random	77.83	79.24	80.00	80.40	80.70
Proposed	78.44	79.94	80.44	80.82	80.70

Evaluation speaker set 2

Training	Number of training speakers				
speaker-set	200	300	400	500	1179
Random	77.71	79.16	79.90	80.35	80.72
Proposed	78.47	79.98	80.54	80.88	80.72

Evaluation speaker set 3

Training	Number of training speakers				
speaker-set	200	300	400	500	1179
Random	78.23	79.73	80.54	80.97	81.29
Proposed	78.97	80.51	81.07	81.32	81.29

6. 結言

6.1 本論文のまとめ

本論文では，統計的パターン認識技術の根幹である統計モデルの学習データベースに着目し，タスク依存性という音声認識の大きな課題を背景に，音声コーパス構築の効率化，認識性能の向上を考慮した収集プロセスの検討を行った．

まず，2 章では，音声認識技術の概要と音声認識の課題の一つであるタスク依存性に関して説明を行った．タスク依存性の議論では，実際に国内の大規模音声コーパスを用いてクロスタスクの音声認識実験を行い，モデル構造の違いや，学習データ量によらず，全てのタスクでタスクが一致している場合の性能が最も高く，他のタスクに対しては性能が劣化し，その度合いはタスクにより異なる等，タスク依存性とその重要性を確認した．

次に，3 章では，タスク間，タスク内の音響的変動や，実環境で発生する雑音と音声との違いを直感的に把握するために，音響モデルの分布を可視空間に写像することで，音響空間上でのデータ間の関連性やその拡がりを把握する手法（COSMOS 法）を提案し，従来の可視化手法である主成分分析法や，SOM 法との比較実験では，明確な優位性を示した．この手法を用いることで，大規模な音声コーパスの可視化が可能となるため，保持する音声コーパス群のタスク間の関連性，タスク内の音響的変動を直感的に把握することができると期待できる．

次に，4 章では，目的タスクの少量音声データを用い，既存音声コーパス群のタスクと目的タスクの関係性を COSMOS 法により可視化することで，既存音声コーパス群の目的タスクに対する再利用性を視覚的に判定する手法を提案した．従来は，既存音声コーパス群から，最も音響的に特徴の近い音声データを選択することは出来ても，その再利用性が十分に高いかどうかを判定する基準がなく，選択後作成された音響モデルの性能を保証することが困難であったが，提案手法を用いることで，目的タスクと既存タスクとの COSMOS Map 上の分布の重なり具合から，直感的に再利用性を把握することが可能となることを実験的に示した．また，提案手法の有効性を客観的に評価するために，クロスタスクの音声認識性能と，COSMOS Map 上の目的タスクと既存タスクの分布間の位置関係との相関

性の調査を行い，monophone HMM におけるクロスタスクの認識性能と高い相関があることを示し，写像時のモデル構造と同一となる条件においては，提案手法の信頼性を確認することができた．また，目的タスクの音響空間上での配置を把握するために必要な少量音声データに関しては，収集する話者数による大きな変動がないことを確認し，提案手法の頑健性を示した．話者数はコストにも大きな影響を与えるため，少ない話者数でも，高精度に再利用性を判定することができることは，コストの観点からも有効性の高い技術として期待できる．

次に，5 章では，低コストで従来と同等以上の性能を実現する音声コーパス構築手法を提案した．まず，COSMOS 法を用い，音声認識性能向上に寄与する音響空間の分析を行い，COSMOS Map 上の話者分布の周辺部分が，音声認識性能の向上に寄与する話者空間であることを確認した．また，高い認識性能を実現するためには，話者の多様性を確保することが重要であり，COSMOS Map 上の話者分布において，話者分布を覆うように，周辺部分に位置する話者を選択することが，多様性のある話者セットを構築することと等価であることを実験的に示した．提案手法として，収集対象の候補話者の少量音声データから，COSMOS 法を用いて音声認識性能向上に寄与する話者を予備選択し，選択された話者の音声データを収集することで，より低コストで効果的な音声コーパスを構築する手法を紹介し，音声認識実験では，従来法と比較し，高い性能を示すだけでなく，60%程度のコスト削減を実現する等，提案手法の有効性を示した．

6.2 今後の研究課題

今後の課題としては，まず，再利用性の判定手法に関しては，タスク依存性が強く現れる triphone HMM におけるクロスタスクの認識性能との相関性を高めるための検討として，写像時のモデル構造に triphone HMM を用いることを検討する予定である．

新規の音声コーパス構築手法に関しては，話者選択のための少量音声データの収集方法について論じていない．現在，筆者が所属する旭化成株式会社新事業本部音声ソリューションビジネス推進部では，実際に本技術を用いた音声コーパス構築を行い，その効果を確認している．しかしながら，予備収録を収録室で行っ

ているため，コストが十分に小さいとは言えない．予備収録のコストを低減する音声データの収集方法としては，電話やインターネットを介しての収集が考えられる．今後は本手法の更なる利便性の向上を目指し，話者選択用の音声データの収集に電話やインターネットを介した場合について検証を行う予定である．また，予備選択における収録の際の話者の負担を軽減するために，より少ない発話数でも高い適応効果を示すMAPLR法やEigen-MLLR法等の適用に関しても検討を行う予定である．また，予備選択における，選択する話者数を決定するための基準に関しても検証を行う予定である．

再利用性の判定手法及び新規の音声コーパス構築手法はそれぞれ，3次元以上の空間での処理が可能である．一般に写像空間の次元が上がるほど写像誤差は軽減するため，配置精度が向上することが期待されることから，3次元以上の空間での処理による，それぞれの手法の精度の検証を行う予定である．反対に，直感的な把握が可能となること以外の可視化による利点を明確にすることで，提案手法の優位性をより明らかにするための検証を行う予定である．本論文では触れていないが，可視化により，Outlier（外れ値）の検出が容易となることが利点の一つとして期待できる．極端に特徴の異なるデータや，不備のあるデータは音響モデルの性能の劣化要因となる恐れがある．これらのデータは可視空間上でOutlierとして明確に配置されると考えられ，検出及び除去が容易となることが期待できることから，実環境の音声データ群を用いての検証実験を行う予定である．

また，COSMOS Map上の配置がクロスタスクの認識性能と高い相関があることを利用して，目的タスクの少量データを用いた認識性能予測技術の検討を行う予定である．そのためには，物理量の適用や，COSMOS法による可視空間上への配置精度の向上が必須であると考えられるため，これらの要素技術の改善を検討する予定である．

6.3 あとがき

最後に，本論文で提案する技術により，目的タスクに特化した音響モデル作成に必要な音声コーパス構築のコストを軽減することで，音声認識アプリケーション開発のコスト削減，工期短縮を助け，新規の音声認識アプリケーション開発へ

の参入を促進し，最終的に，音声認識業界全体の活性化につながること期待する．

謝辞

本論文は，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後期課程在学中及び旭化成株式会社新事業本部情報技術研究所及び音声ソリューションビジネス推進部での筆者の研究成果をまとめたものです．本研究を遂行するにあたり，学内，学外，社内，社外問わず数多くの方々に援助を頂きました．ここに感謝を意を表したいと思います．

はじめに，本研究を進め，論文としてまとめるにあたり，奈良先端科学技術大学院大学　情報科学研究科　音情報処理学講座　鹿野清宏教授には，多くの有益な御助言を頂きました．鹿野先生には，筆者が博士前期課程及び博士後期課程において，主指導教官として御指導頂きました．また，社会人学生として，本学で学位を得る機会を与えていただき，入学後も社会人として研究活動を進める上で，温かいお言葉，ご配慮をくださいました．本研究での研究活動を成し遂げることができたのは，鹿野先生の研究に対する深い御理解と的確かつ暖かい御助言に励まされたおかげです．心より感謝の意を表します．そして，副指導教官として本論文の執筆にあたり，有益なご助言と熱心な御指導を賜わりました，同研究科　知能情報処理学講座　木戸出正繼教授，同研究科　音情報処理学講座　猿渡洋准教授に厚く御礼申し上げます．木戸出先生，猿渡先生には，筆者が博士前期課程及び博士後期課程において，副主指導教官として御指導頂き，厳しくも温かく，そして的確な御助言を賜りましたことを深く感謝致します．

音情報処理学講座では，川波弘道助教授をはじめ，多くの諸先輩方，スタッフの方々，学生の方々に色々な面で助けて頂きました．ここに深く感謝致します．戸田智基助教授には筆者が博士前期課程及び博士後期課程において，良き先輩，また目指すべき研究者として，多くのことを教えて頂きましたこと，そして，多くの御助言を賜りました．本研究においても，熱心な議論をさせて頂き，有益な御助言を頂きましたことを深く感謝致します．宮部滋樹博士，Tobias Cicarek 博士（現ヤフー株式会社），高谷智哉博士（現トヨタ自動車株式会社）には，本研究に関して熱心な議論をさせて頂き，有益な御助言を頂きましたことを深く感謝致します．馬場朗氏（現松下電工株式会社），中山彰博士（現日本電信電話株式会社）には，筆者と同じく社会人学生として，貴重な時間を共有させて頂きまし

た．また，本研究に関して熱心な議論をさせて頂き，有益な御助言を頂きましたことを深く感謝致します．

また，本研究を進める上で，行いました学会活動の中で，多くの方々と，本研究に関して有益な議論をさせて頂き，多くのご助言，厳しい御指摘，温かいお言葉を賜りました．ここに，深く感謝致します．

また，名古屋工業大学大学院　工学研究科　李晃伸准教授，同志社大学大学院工学研究科　片桐滋教授，日本電信電話株式会社　南泰浩博士，中村篤博士，Erick Mcdermott 博士，Harman/Becker 社　Daniel Willet 博士には，筆者が博士前期課程在学中，音声認識に関して熱心な議論をさせて頂き，多くのご助言を賜りました．ここに深く感謝致します．同じく，筆者が博士前期課程在学中，多くの時間を共有した，同期の方々にも，深く感謝致します．そして，筆者が音声認識に携わるきっかけを与えてくださり，研究者としてのあるべき姿を示してくださいました，ATR 株式会社　音声言語コミュニケーション研究所　中村哲所長に深く感謝致します．

筆者は旭化成株式会社に在籍しながら，奈良先端科学技術大学院大学　情報科学研究科　博士後期課程への入学の機会を頂き，本研究をまとめることができました．業務の忙しい時期にもかかわらず入学を薦めてくださり，筆者の入学をご快諾してくださいました旭化成株式会社　新事業本部　音声ソリューションビジネス推進部　庄境誠部長（情報技術研究所長兼務）に，心からの感謝の意を表します．庄境部長には，筆者が旭化成株式会社に入社する機会を与えていただき，以来 6 年間，様々な業務を通し，研究者としてのあり方，社会人としての心構え，その他多くのことに関し，多くのご助言，ご指導を頂きました．また，本研究における筆者の最大の理解者であり，支援者でいてくださいました．本来の業務が多忙となり，長期間本研究を進めることができなっかた時期には，心温かいご配慮のお言葉を賜りましたことを忘れることはできません．重ねて深く感謝致します．

旭化成株式会社　新事業本部　情報技術研究所及び音声ソリューションビジネス推進部の諸氏には本研究を進めるにあたり，多くのご助言，ご支援を頂きました．業務が多忙な中，本研究を継続し，本論文としてまとめることができましたのは，諸氏のご協力と，ご支援があったからこそであり，全ての方々に深く感謝

致します．全ての方のお名前をここに挙げられないことをご容赦ください．本研究に関し，特に，熱心な議論をさせて頂き，また日頃の業務において，ご助言，ご協力頂きました，情報技術研究所　野口祥宏博士，石原憲氏，水嶋康和氏，宇田川健氏，嶋田敬士氏，音声ソリューションビジネス推進部　古越道昭氏，尾和邦彦氏，谷智洋氏，松本勧氏，舛田剛志氏，瀬古康之氏，眞杉裕美子氏，山田真士氏，加藤智之氏，小笹詩織氏，片桐章宏氏，田中寿子氏に深く感謝致します．また，本研究の特許化を進めるにあたり，ご尽力頂きました，情報技術研究所　小林士朗博士，河崎貴昭氏に深く感謝致します．そして，本研究における様々な実験にご協力頂き，日頃の業務でも筆者を支えてくださいました原囿友輔氏に，心より感謝の意を表します．

最後に，社会人学生として二足の草鞋を履いた3年間，家庭での務めを十分に果たせない私を許し，常に私を支えてくれた最愛の妻と娘，そして，私の家族を支えてくれた私と妻の両親に心から感謝致します．

参考文献

[1] C. M. Bishop, Pattern Recognition And Machine Learning, *Springer-Verlag*, 2006.

[2] 石井健一郎， 前田英作， 上田修功， 村瀬洋， わかりやすいパターン認識，オーム社， 1998.

[3] 中川聖一， 確率モデルによる音声認識， コロナ社， 1998.

[4] 古井貞熙， ディジタル音声処理， 東海大学出版会， 1985.

[5] 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.

[6] 鹿野清宏， 伊藤克旦， 河原達也， 武田一哉， 山本幹雄，音声認識システム， オーム社， 2001.

[7] 田中和之， 確率モデルによる画像処理技術入門　森北出版， 2006.

[8] R. C. Gonzalez, R. E. Woods, Digital Image Processing (Second edition), *Prentice Hall*, 2002.

[9] A. P. Dempster, N. M. Laird and D. B. Rubin, ”Maximum Likelihood from Incomplete Data via the EM Algorithm,” Journal of the Royal Statistical Society. Series B (Methodological), vol.39, no.1, pp. 1-38, 1977.

[10] 宇田川健， 山崎裕二， 庄境誠， “赤外線センサに HMM を適用した物体の認識手法 ”， 情報処理学会研究報告, MBL2005-28, pp. 179-186, 2005.

[11] http://japan.nuance.com/viavoice/

[12] http://japan.nuance.com/naturallyspeaking/

[13] http://japan.nuance.com/vocon/

[14] http://www.nec.co.jp/middle/VisualVoice/

[15] http://www.asahi-kasei.co.jp/vorero/jp/

[16] 庄境誠, “組み込み向け音声認識ミドルウェアＶＯＲＥＲＯの開発”, 日本音響学会講演論文集, 1-8-13, pp.31-32, Mar, 2004.

[17] 赤堀一郎, “カーナビ音声認識の商品開発”, 情報処理学会研究報告, SLP-103, pp. 31-32, 2005.

[18] 河井恒, “音声認識を利用した携帯電話サービスの開発”, 情報処理学会研究報告, SLP-103, pp. 35-36, 2005.

[19] 磯健一, 磯谷亮輔, 石川晋也, 江森正, 三木清一, 花沢健, 渡辺隆夫, “携帯端末向け大語彙連続音声認識システム”, 電子情報通信学会論文誌, vol.J87-D-II, no.2, pp. 487-494, 2004.

[20] C. J. Leggretter 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.

[21] T. Anastasakos, J. Mcdonough, R. Schwartz, and J.Makhoul, ”A compact model for speaker adaptive training,” *International Conference on Spoken Language Processing (ICSLP)*, pp.1137-1140, 1996.

[22] M. J. F. Gales and S. Young, ”An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*92, pp.233-236, 1992.

[23] Boll, S. F., “Suppression of Acoustic Noise in Speech using Spectral Subtraction,” *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 27, pp.113-120, 1979.

[24] M. Shozakai, S. Nakamura and K. Shikano, ”A speech enhancement approach E-CMN/CSS for speech recognition in car environments,” *IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp.450-457, 1997.

[25] Y. Ephraim and D. Malah, ”Speech enhancement using a minimum mean square error short-time spectral amplitude estimator,” *lEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 32, pp.1109-1121, 1984.

[26] 武田一哉， 匂坂芳典， 片桐滋， 阿部 匡伸， 桑原尚夫， ”研究用日本語音声データベース利用解説書”, *ATR Technical Report*, TR-I-0028, 1988.

[27] 奥田浩三， 松井知子， 内藤正樹， 匂坂芳典， 中村　哲， ”大規模日本語音声データベースの構築と評価，” 日本音響学会論文誌， vol.58， no.9， pp.569-578， 2002.

[28] http://www.mibel.cs.tsukuba.ac.jp/jnas/

[29] 馬場　朗， 芳澤伸一， 山田実一， 李　晃伸， 鹿野清宏， ”高齢者向け音響モデルによる大語彙連続音声認識，” 電子情報通信学会論文誌， vol.J85-D-II, no.3， pp.390-397， 2002.

[30] 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， vol.J84-D-II， no.6， pp.909-917， 2001.

[31] K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, S. Matsubara and F. Itakura, ”Construction and Evaluation of a Large In-Car Speech Corpus,” *IEICE Transactions on Information and Systems*, vol.E88-D, no.3, pp. 553-561, 2005.

[32] 篠崎隆宏， 古井貞熙， ”日本語話し言葉コーパスを用いた講演音声認識，” 情報処理学会論文誌， vol.43， no.7， pp.2098-2107， 2002.

[33] S. Furui, ”Why is recognition of spontaneous speech so hard?,” *International Conference on Text, Speech and Dialogue (TSD)*, pp.9-22, 2005.

[34] 南條浩輝， 河原達也， ”講演音声認識のための教師なし言語モデル適応と発話速度に適応したデコーディング，” 電子情報通信学会論文誌， vol.J87-D-II, no.8, pp.1581-1592, 2004.

[35] S. Yoo, J. R. Boston, J. D. Durrant, K. Kovacyk, S. Karn, S. Shaiman, A. El-Jaroudi and L. Ching-Chung, "Relative Energy And Intelligibility Of Transient Speech Information," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp.69-72, 2005.

[36] 山本一公，中川聖一，"発話スタイルによる話速・音韻間距離・ゆう度の違いと音声認識性能の関係，" 電子情報通信学会論文誌，vol.J83-D-II，no.11，pp.2438-2447，2000.

[37] J. J. Odel, "The Use of Context in Large Vocabulary Speech Recognition", D. Phil dissertation, Cambridge University, 1995.

[38] 李晃伸, 河原達也, 武田一哉, 鹿野清宏, "Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識", 電子情報通信学会論文誌, vol. J83-DII, no. 12, pp. 2517-2525, 2000.

[39] N. Ueda, "EM algorithm with split and merge operations for mixture models (invited)," *IEICE Transactions on Information and Systems*, vol. E83-D, no. 12, pp. 2047-2055, 2000.

[40] N. Ueda and R. Nakano, "Deterministic annealing EM algorithm," *Neural Networks*, vol.11, no. 2, pp.271-282, 1998.

[41] S. Waterhouse, D. MacKay, and T. Robinson, "Bayesian methods for mixtures of experts," *Advances in Neural Information Processing Systems (NIPS 7)*, MIT Press, 1995.

[42] S. Watanabe,Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian Estimation and Clustering for Speech Recognition," *IEEE transaction on Speech and Audio Processing*, vol. 12, pp.365-381, 2004.

[43] S. Katagiri, B-H. Juang, and C-H. Lee, "Discriminative Learning for Minimum Error Classification", *IEEE Transactions on Signal Processing*, vol.40, no.12, pp. 3043-3053, 1992.

[44] E.. McDermott, "Discriminative Training for Speech Recognition", D. Phil dissertation, Waseda University, 1997.

[45] D. Povey and P. C. Woodland, " Minimum phone error and I-smoothing for improved discriminative training, " *IEEE International Conference on Acoustics, Speechand Signal Processing (ICASSP)*, pp. 105-108, 2002.

[46] F. Lefeve, J.L. Gauvain and L. Lamel, "Genericity and portability for task-independent speech recognition," *Computer speech and language*, vol.19, pp.345-363, 2005.

[47] T. Cincarek, T. Toda, H. Saruwatari and K. Shikano, "Utterance-based Selective Training for the Automatic Creation of Task-Dependent Acoustic Models," *IEICE Transactions on Information and Systems*, vol.E89-D, no.3, pp.962-969, 2006.

[48] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol.7, no.2, pp.179-188, 1936.

[49] M. Aladjem, "Multiclass discriminant mappings," *Signal Processing*.,vol.35, pp.1-18, 1994.

[50] J. Mao and A.K. Jam, "Artificial neural networks for feature extraction and multivariate data projection," *IEEE Transactions on Neural Networks*, vol.6, no.2, pp.296-317, 1995.

[51] Y. Mori, M. Kudo, J. Toyama and M. Shimbo, "Comparison of low-dimensional mapping techniques based on discriminatory information," *International ICSC Symposium on Advances in Intelligent Data Analysis (AIDA)*, pp.1724-166, 2001.

[52] A. Nagorski, L. Boves and H. Steeneken, "Optimal selection of speech data for automatic speech recognition system," *International Conference on Spoken Language Processing (ICSLP)*, pp.2473-2476, 2002.

[53] A.K. Jain, R.P.W. Duin and J. Mao, ”Statistical pattern recognition: a review,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.22, no.1, pp.4-37, 2000.

[54] T. Kohonen, ”Self-Organizing Maps,” *Springer Series in Information Sciences*, vol.30, 1995.

[55] T. Kohonen, J. Hynninen, J. Kangas and J. Laaksonen, SOM_PAK: The Self-Organizing Map Program Package, http://www.cis.hut.fi/research/som_lvq_pak.shtml, 1996.

[56] J. H. Freidman and J. W. Tukey, ”A projection pursuit algorithm for exploratory data analysis,” *IEEE Transactions on Computers*, vol.C-23, no.9, pp.881-889, 1974.

[57] 田中豊, 脇本和昌, 多変量統計解析法, 現代数学社, 1983.

[58] J.W. Sammon, ”A nonlinear mapping for data structure analysis,” *IEEE Transactions on Computers*, vol.C-18, no.5, pp.401-409, 1969.

[59] 庄境誠, 奈木野豪秀, ”データ処理装置及びデータ処理装置制御プログラム,” 特許番号:1-669-979(DE,FR,GB), 国際公開番号:WO2005/034086, 国際出願番号:PCT/JP2004/010390.

[60] M. Shozakai and G. Nagino, “Analysis of speaking styles by two-dimensional visualization of aggregate of acoustic models,” *International Conference on Spoken Language Processing (ICSLP)*, pp.717-720, 2004.

[61] 庄境誠, 奈木野豪秀, “多次元尺度構成法による音響空間の 2 次元可視化,” 情報処理学会研究報告, SLP2004-74, pp. 129-136, 2004.

[62] G. Nagino and M. Shozakai, “Building an effective corpus by using acoustic space visualization (COSMOS) method,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp.449-452, 2005.

[63] M. Shozakai and G. Nagino, "Acoustic space analysis method utilizing statistical multidimensional scaling technique," *International Workshop on Nonlinear Signal and Image Processing (NSIP)*, 2005.

[64] G. Nagino and M. Shozakai, "Analyzing Reusability of Speech Corpus based on Statistical Multidimensional Scaling Method," *International Conference on Spoken Language Processing (ICSLP)*, pp.161-164, 2006.

[65] G. Nagino and M. Shozakai, "Distance Measure between Gaussian Distributions for Discriminating Speaking Styles," *International Conference on Spoken Language Processing (ICSLP)*, vol., pp.657-660, 2006.

[66] G. Nagino and M. Shozakai, Kiyohiro Shikano, "How to Judge Reusability of Existing Speech Corpora for Target Task by Utilizing Statistical Multidimensional Scaling," *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1302-1305, 2007.

[67] 奈木野豪秀, 庄境誠,"COSMOS 法を用いた既存音声コーパスの相関分析," 日本音響学会講演論文集, 1-P-15, pp.181-182, Sep, 2005.

[68] 庄境誠, 奈木野豪秀, 鹿野清宏, "多数日本語音声コーパスからの日本語音響空間地図の作成," 日本音響学会講演論文集, 1-11-12, pp.45-46, Mar, 2006.

[69] 奈木野豪秀, 鹿野清宏, 庄境誠,"国内既存音声コーパスの音響空間配置図の作成," 日本音響学会講演論文集, 2-P-15, pp.123-124, Sep, 2006.

[70] G. Nagino, M. Shozakai, T.Toda, H.Saruwatari and K.Shikano, "Building An Effective Speech Corpus by Utilizing Statistical Multidimensional Scaling Method," *IEICE Transactions on Information and Systems*, vol. E91-D, no. 3, pp. 607-614, March 2008.

[71] 庄境誠, "実環境で雑音と音声を聞き分ける," 電子情報通信学会誌, vol. 87, no. 3, pp. 168-174, 2004.

[72] 庄境誠, ”複数音声コーパスの俯瞰的分析,” 電子情報通信学会技術研究報告, SP2005-112, no.12, pp.43-48, 2005.

[73] T. Haderlein, D. Zorn, S. Steidl, E. N¨ oth, M. Shozakai, and M. Schuster, ”Visualization of Voice Disorders Using the Sammon Transform,” *International Conference on Text, Speech and Dialogue (TSD)*, pp. 589-596, 2006.

[74] K. Fukunaga, ”Introduction to statistical pattern recognition (Second edition),” *Academic Press*, 1990.

[75] H. Akaike, ”A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, vol.19, no.6, pp.716-723, 1974.

[76] J. Goldberger and H. Aronowitz, ”A Distance measure Between GMMs Based on the Unscented Transform and its Application to Speaker Recognition,” *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1985-1988, 2005.

[77] P. Olsen and J. Hershey, “Bhattacharyya error and divergence using variational importance sampling,” *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 46-49, 2007.

[78] J. Chen, P. Olsen and J. Hershey, “Word Confusability - Measuring Hidden Markov Model Similarity,” *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 2089-2092, 2007.

[79] B. Mohanty, J. Hershey, P. Olsen, S. Kozat and V. Goel, ”Optimizing Speech Recognition Grammars Using a Measure of Similarity between Hidden Markov Models,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp.4953-4956, 2008.

[80] 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.

[81] S. Matsuda, T. Jitsuhiro, K. Markov and S. Nakamura, "ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles," *IEICE Transaction on Information and Systems*, vol.E89-D, no.3, pp.989-997, 2006.

[82] 柴崎一郎, 庄境誠, "鉄道分野における音声・音響センサの応用," 電気学会論文誌（E）, vol.127, no.11, pp.493-498, 2007.

[83] 山川仁子, 松井知子, 板橋秀一, "非計量多次元尺度構成法を用いた複数音声コーパスの可視化," 日本音響学会講演論文集, 1-P-20, pp. 447-448, 2007.

[84] 山川仁子, 松井知子, 板橋秀一, "多次元尺度構成法を用いた複数音声コーパス可視化法の検証," 日本音響学会講演論文集, 3-Q-10, pp. 395-396, 2008.

[85] K. Yamakawa, T. Matsui and S. Itahashi, "Visualization of various speech corpora by multidimensional scaling," *International Conference on Speech Databases and Assessment (COCOSDA)*, no.45, 2007.

[86] 桑原尚夫, 板橋秀一, 山本幹雄, 中村哲, 竹澤寿幸, 武田一哉："国内における音声データベースの現状－開発, 管理及び音声研究への利用－," 日本音響学会誌, 59, 2, pp. 99-103, 2003.

[87] T. Yamada, M. Kumakura and N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2006-2013, 2006.

[88] 橋本倫和, 山田武志, 北脇信彦, "雑音下音声認識の性能推定のためのひずみ尺度の検討," 情報処理学会研究報告, SLP-69-4, pp. 19-24, Dec. 2007.

[89] 加藤智之, 岡本淳, 庄境 誠, "自動車運転行動中発話の日本語音声コーパスの物理量と認識性能の相関分析" 日本音響学会講演論文集, 3-Q-25, pp. 267-268, Sep, 2007.

[90] 小林哲則、中野鐵兵、庄境誠、加藤智之、岡本淳、石川泰、佐藤幹、岩沢透、杉山昭彦、高野陽介、藤田善弘, ”パネルディスカッション : 音声認識実用化に向けて,” 情報処理学会研究報告, SLP-068, pp. 31-42, 2007.

[91] Y. Gao, L. Gu and H.-K.Jeff Kuo, ”Portability challenges in developing interactive dialogue systems,” Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1017.1020, 2005.

[92] T.M. Kamm and G.G.L.Meyer, ”Robustness aspects of active learning for acoustic modeling,” Proc. *International Conference on Spoken Language Processing (ICSLP)*, pp. 1095-1098, 2004.

[93] Tomoyuki Kato, Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano, “Transcription Cost Reduction for Constructing Acoustic Models Using Acoustic Likelihood Selection Criteria,” *International conference on Language Resources and Evaluation (LREC)*, pp. 789–792, 2006

[94] D. Hakkani-Tur, G. Riccardi, and A. Gorin. Active learning for automatic speech recognition. Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3904-3907, 2002.

[95] C. Chesta, O. Siohan, and C.H. Lee, ”Maximum a posteriori linear regression for hidden Markov model adaptation,” *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp.211-214, 1999.

[96] K. Chen, W. lian, H. Wang, and L. Lee, ”Fast speaker adaptation using Eigenspace-based maximum likelihood linear regression,” Proc. *International Conference on Spoken Language Processing (ICSLP)*, pp.742-745, 2000.

付録

A. 音声コーパス概要

- ATR-APP：ATR-APP 多数話者音声コーパス [27]
  <収録手順 ： 手順Ⅱ及びⅢ，音響解析条件 ： 設定 2>
  全国規模の不特定話者を対象とする音声認識技術の性能向上を目的とした音声コーパスであり，全国各地（8 地方，47 都市）の 10 代〜60 代の男性 1381 名，女性 2390 名による模擬会話の自然音声データ及び ATR 音素バランス文の読み上げ音声データを収録している．地域性を網羅した数少ない大規模音声コーパスの一つである．

- JNAS：JNAS 読み上げ音声コーパス [28]
  <収録手順 ： 手順Ⅱ，音響解析条件 ： 設定 2>
  大語彙連続音声認識研究を目的とした読み上げ音声コーパスであり，成人男女各 153 名による ATR 音素バランス文，新聞記事の読み上げ音声データを収録している．収録は 39 の機関により行われており，ローパス・フィルタや AD 変換の特性は一致していない等，収録条件にはばらつきが生じている．国内の音声認識の研究において最も利用されている大規模音声コーパスの一つである．

- S-JNAS：S-JNAS 高齢者音声コーパス [29]
  <収録手順 ： 手順Ⅱ及びⅢ，音響解析条件 ： 設定 2>
  高齢者による音声認識を目的とした高齢者による読み上げ音声コーパスであり，音響モデル学習用話者として男性 151 名，女性 150 名による ATR 音素バランス文，新聞記事の読み上げ音声データ，及び，評価用話者として男性 51 名，女性 50 名による新聞記事，情報検索文の読み上げ音声データを収録している．高齢者の音声コーパスとしては最大規模の音声コーパスである．

- CIAIR-HCC：CIAIR 車内音声コーパス [30, 31]
  <収録手順 ： 手順Ⅰ，音響解析条件 ： 設定 2>

走行車内での音声認識を目的とした実車内音声コーパスであり，男女計800名に上る被験者に対して，さまざまな運転状況下での音声を収録している．走行車内という実環境下における音声コーパスとしては最大規模の音声コーパスである．本論文では，接話マイクで収録された運転中運転席での音声データに対してのみ分析を行うものとする．運転中運転席の音声データとしては，ATR音素バランス文の読み上げと，人，WOZ（Wizard Of Oz）システム，ASR（音声認識）システムとの対話がある．対話は施設等の検索タスクである．

- CSJ：CSJ話し言葉音声コーパス [32, 33, 34]
  <収録手順 ： 手順Ⅰ及びⅢ，音響解析条件 ： 設定2>
  より自然な発声（話し言葉）の研究を目的とした音声コーパスであり，男女計1417名による，学会講演，模擬講演といった自発的に発声する話し言葉での音声を収録している．話し言葉の音声コーパスとしては最大規模の音声コーパスである．

- IW-Corpus：Isolated Word（孤立単語発声）音声コーパス
  <収録手順 ： 手順Ⅱ，音響解析条件 ： 設定1>
  IW-Corpusは，組み込み機器上で動作する音声認識アプリケーションに導入する音響モデルの学習用音声コーパスとして，旭化成株式会社[15]で構築された音声コーパスであり，ATR5240[26]の全5240単語を175単語からなる30の単語リストに分割したものを561名の男性話者が読み上げ発話により発声した音声が収録されている．

- CW-Corpus：Continues Word（孤立単語発声）音声コーパス
  <収録手順 ： 手順Ⅱ，音響解析条件 ： 設定2>
  ATR-APP（ATR-APP多数話者音声コーパス）を使用．

B. 物理量によるタスク依存性の分析

本付録では，各タスクの物理量の分析を行い，2.4.2 節で紹介したクロスタスクの音声認識実験の実験結果に示された，タスク依存性との関連について議論を行う．1 節で各タスクの物理量の分析を行い，その結果を示すと共に結果に関して考察を行う．2 節では物理量とクロスタスクの音声認識実験結果との相関について分析を行い，その結果を示すと共に結果に関して考察を行う．最後に 3 節で，まとめを行う．

B.1 物理量の分析

各タスクの音響的な差異を比較するために，SNNR（Signal with Noise to Noise Ratio[dB]）[35]，発話速度，音素間距離，音素環境の偏りに関して分析を行う．SNNR は Viterbi アライメントによる発話区間検出結果をもとに，発話区間における音声信号と雑音信号の平均パワーと非発話区間の雑音信号の平均パワーとの比により算出する．単位は [dB] である．

発話速度は SNNR と同様に，Viterbi アライメントによる発話区間検出結果をもとに，発話中の総モーラ数を発話区間長で除することで算出する．単位は [mora/sec] である．

次に音素間距離 [36] について説明する．音素間距離（PD : Phoneme Distance）は，各タスクの音声コーパスから作成した monophone HMM によるタスク依存音響モデルを用いて算出される．本論文では平均音素間距離（APD : Average of PD）及び音素間の識別誤りに着目した最小音素間距離（MPD : Minimum of PD）に関して分析を行う．音素 ϕ の平均音素間距離 APD_ϕ 及び最小音素間距離 MPD_ϕ は，式 (3.38)（または式 (3.39)）における音素間距離 PD により次式で定義される．

$$APD_\phi = \frac{1}{R-1} \sum_{r=1, s_r \neq \phi}^{R} PD(\phi, s_r) \tag{B.1}$$

$$MPD_\phi = \min_{1 \leq r \leq R} PD(\phi, s_r) \tag{B.2}$$

音素環境の偏りは音素の出現頻度のエントロピー SE により表現するものとし，

次式で求められる．以下では，音素として biphone を用いる．単位は [bit] である．

$$SE = -\sum_{r=1}^{R} \mathcal{P}(s_r) \log_2 \mathcal{P}(s_r) \tag{B.3}$$

ここで，R は定義されている音素の総数，s_r は r 番目の音素，$\mathcal{P}(s_r)$ は $\sum_{r=1}^{R} \mathcal{P}(s_r) = 1$ を満たす音素 s_r の出現確率である．

発話単位に算出された SNNR，発話速度のタスク平均及び各タスクのエントロピーを表 B.1 に示す．同様に，タスク依存音響モデル毎に算出された平均音素間距離及び最小音素間距離の母音平均及び子音平均を表 B.1 に示す．また，物理量毎の各タスクの違いを図 B.1 から図 B.5 に示す．

表 B.1 及び図 B.1 から図 B.5 の結果はそれぞれ，常識として知られている事実，既に知られている事実，新しく発見した事実（経験上予測は可能であるが，事実として報告されていないものを含む）を含んでいる．それぞれを以下に挙げる．

- 常識として知られている事実

  1)　収録室で収録されたタスクの SNNR は，背景雑音が存在する実環境で収録されたタスクの SNNR と比較し高い．

  2)　エントロピーが大きくなるよう設計されたバランス文読み上げタスクや，新聞記事読み上げタスク，CSJ タスク等の発話内容が多岐にわたるようなタスクのエントロピーは大きく，言い回しが限定される検索タスク（S-JNAS_infoseek や CIAIR の対話タスク）のエントロピーは小さい．

- 既に知られている事実

  1)　読み上げタスクと比較し，自然発話タスクである CIAIR の各対話タスク，CSJ の各タスクの発話速度は早く，そのばらつきも大きい [36, 34]．

  2)　自然発話音声が読み上げ音声と比較し，発話の怠けにより平均音素間距離が小さくなることは過去の知見と一致するものである [36, 33]．

- 新しく発見した事実及びその原因に関する考察

  1)　高齢者タスクは，発話速度が遅く，SNNR のばらつきが小さい．また，音素間距離が小さい．発話内容が同じ JNAS_balance や JNAS_news と比較し

ても，その傾向が顕著に現れていることから，高齢者特有の特徴と考えられる．加齢につれて，発話はゆっくりとなる傾向にあるが，スペクトルが崩れる傾向にあることが示唆される．

2)　運転中の対話タスクは他のタスクと比較し，SNNR のばらつきが大きい．同じ自然発話タスクである CSJ の各タスク，同じ検索タスクの S-JNAS_infoseek，同じ運転行動中である CIAIR-drive_balance 等，各要因が同じであるタスクと比較しても大きいことからも，自然発話，発話内容，運転行動のどれか一つを選び，SNNR のばらつきが大きいことの原因とすることはできない．発話速度のばらつきに関しても，これらのタスクと比較して顕著に大きく，自然発話タスクであることのみが原因とは言い難い．「運転行動中における検索を目的とした対話（自然発話）」という，複数の要因が複雑に絡み合った特殊なタスクであることが原因であり，絡み合った各要因を分離することは難しいと考えられる．

3)　CIAIR-drive_balance は CIAIR の各対話タスクと比較し，SNNR は高く，そのばらつきも小さい．また，発話速度は遅く，そのばらつきも小さい．運転中にテキストを読み上げることは安全面上問題であることから，被験者は一文を複数の文節に区切った音声を聞き，鸚鵡返し発声を行っている．そのため，教師信号であるガイダンス音声に発声が誘導されてしまうことが原因と考えられる．

4)　運転中の ASR との対話タスクである CIAIR-drive_dialogA は他のタスクと比較して，著しくエントロピーが小さい．他の運転中での対話タスクと比較しても小さいことから，ASR との対話であることが原因として考えられる．実際に収録された音声データには，単語発声化が頻発するなど，語彙や言い回しが特に少なくなっている．但し，ASR の基本性能により，発話者の語彙数や言い回しは変化するため，基本性能も原因の一つとして考慮に入れる必要があると考えられる．

これらの事実からもわかる通り，各タスク間において全ての物理量の傾向が完全に一致することはなく，たとえ同一コーパス内でもタスクの設定により，一部の物理量が大きく異なっている．複数の要因が複雑に絡み合っている場合に，各

表 B.1 各タスクの物理量*

Task	SNNR [dB] Avg/Std	Speed [mora/sec] Avg/Std	Entropy [bit]	Average of PD		Minimum of PD	
				Vowel	Consonant	Vowel	Consonant
AAb	38.75/6.6	7.46/0.93	7.8	1.82	2.52	1.08	0.68
Jb	33.16/7.04	7.08/0.91	7.76	1.86	2.63	1.22	0.78
Jn	32.75/7.02	7.06/0.91	7.61	1.75	2.57	1.07	0.70
SJb	33.78/5.61	5.82/0.92	7.78	1.3	1.91	0.78	0.46
SJn	32.87/5.5	5.84/0.92	7.67	1.3	1.99	0.75	0.48
SJi	33.14/5.39	6.54/1.13	7.28	1.33	2.18	0.74	0.68
Cdb	30.13/6.39	6.26/1.07	7.8	2.05	2.64	1.17	0.77
CdA	27.53/9.61	8.38/2.46	6.99	0.85	2.07	0.57	0.72
CdW	25.11/9.55	8.69/2.41	7.32	0.98	2.32	0.67	0.73
CdH	24.08/9.02	8.66/2.46	7.35	1.01	2.24	0.66	0.77
Ca	24.75/6.18	8.16/1.72	7.69	0.88	1.93	0.73	0.60
Cs	28.94/6.57	7.48/1.71	7.62	0.9	1.93	0.63	0.57

物理量の差異の原因を特定することは難しいが，各物理量はそれぞれ，音声認識において最も重要な情報源であるスペクトル形状に影響を与える物理量であることから，これらの物理量の差異が，クロスタスクの認識実験におけるタスク依存性の要因として，高い相関関係にあると期待できる．次節では，2.4.2 節及び本節で得られた，クロスタスクの音声認識実験結果と物理量分析結果から，各物理量と認識性能（に表れるタスク依存性）との相関に関して調査を行う．

B.2 音声認識実験と物理量との相関分析

本節ではまず，前述のクロスタスクの音声認識実験結果及び物理量の分析結果から，各物理量のクロスタスクの音声認識性能に対する相関係数を求める．

相関係数は，2 つの変数間の関係性を表す係数であり，-1 から $+1$ までの値をとる．変数 $\boldsymbol{x}$ と $\boldsymbol{y}$ の対からなる標本 $\{x_i, y_i\}(i = 1, 2, \ldots, I)$ が与えられた場合，相関係数 κ は次式で定義される．

$$\kappa = \frac{\sum_{i=1}^{I}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{I}(x_i - \bar{x})^2 \sum_{i=1}^{I}(y_i - \bar{y})^2} \tag{B.4}$$

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

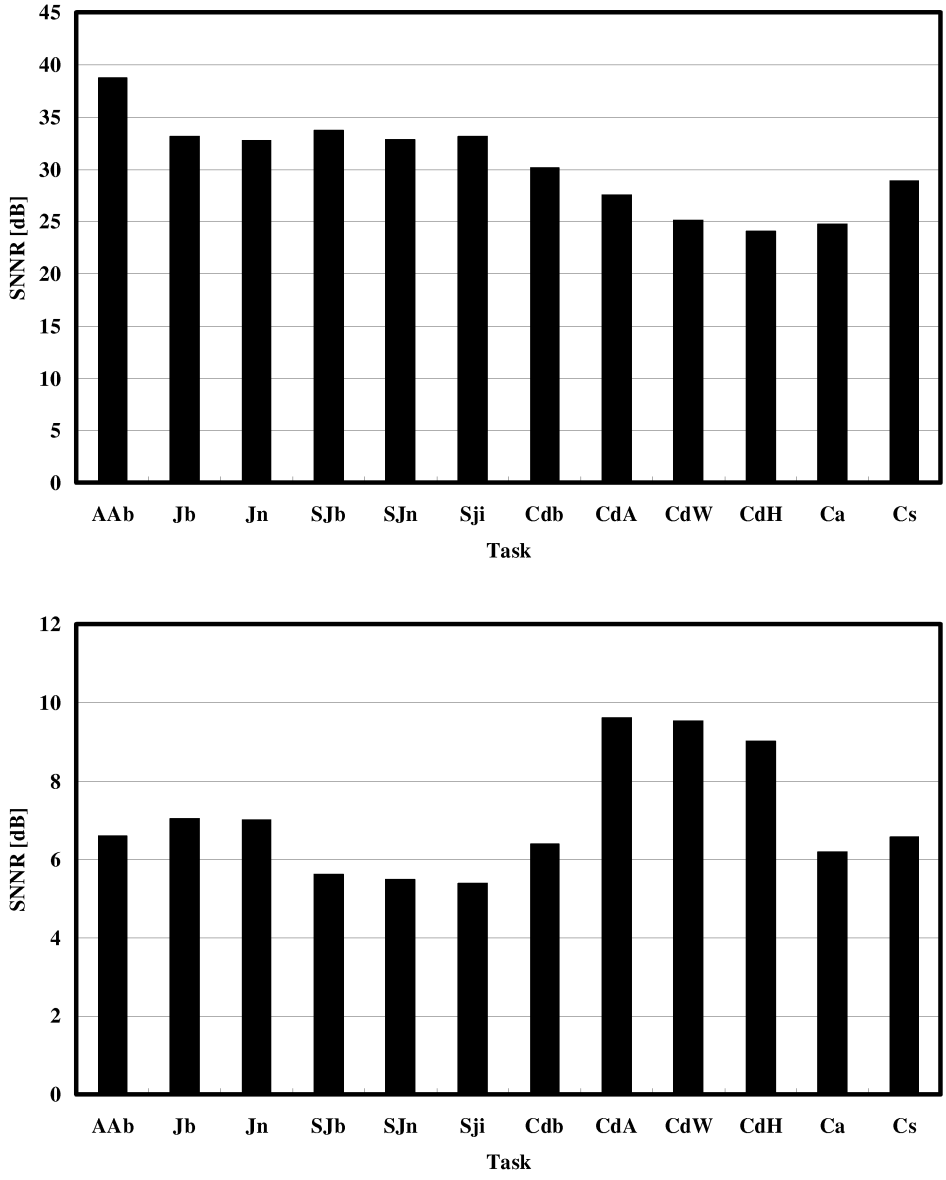


図 B.1 各タスクの SNNR（上：平均，下：標準偏差）*

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

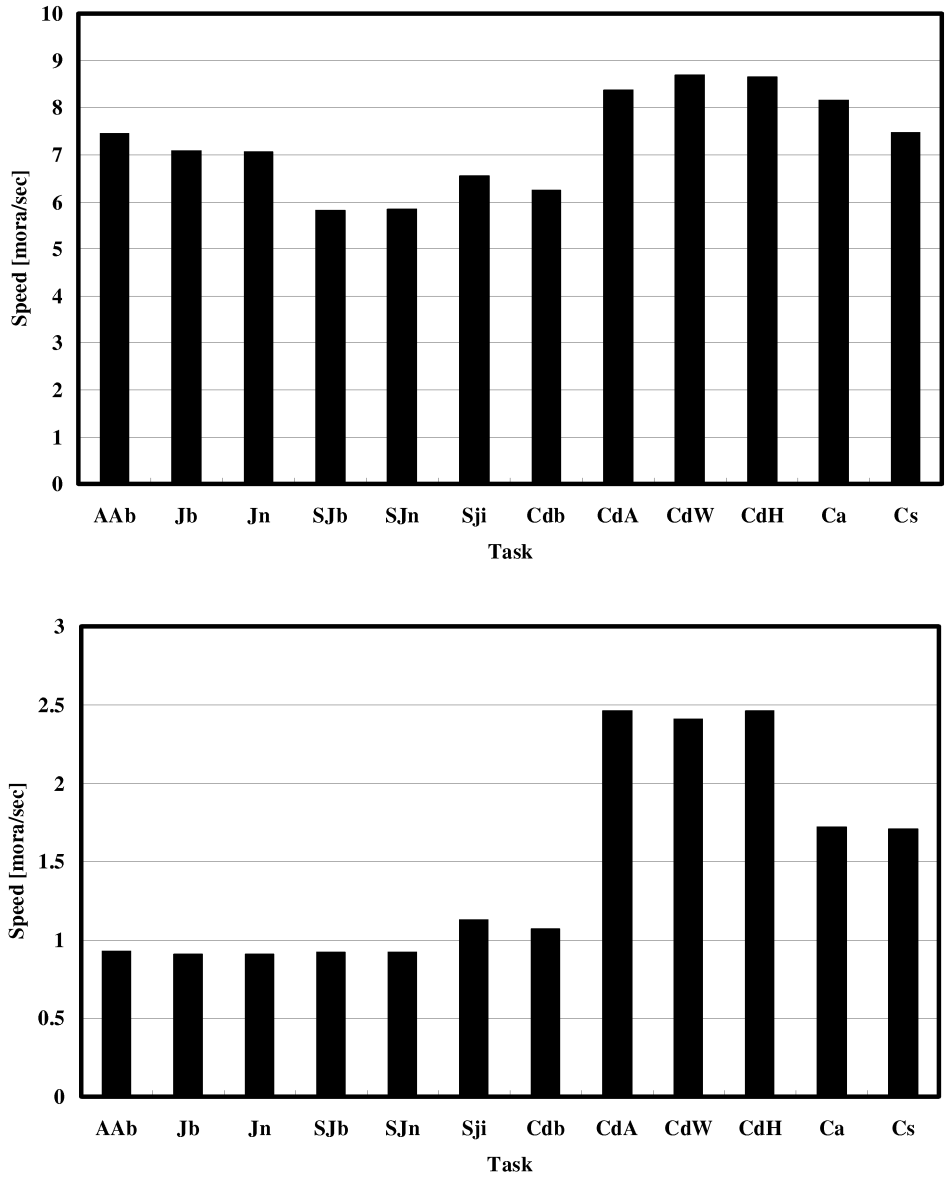


図 B.2 各タスクの発話速度（上：平均，下：標準偏差）*

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

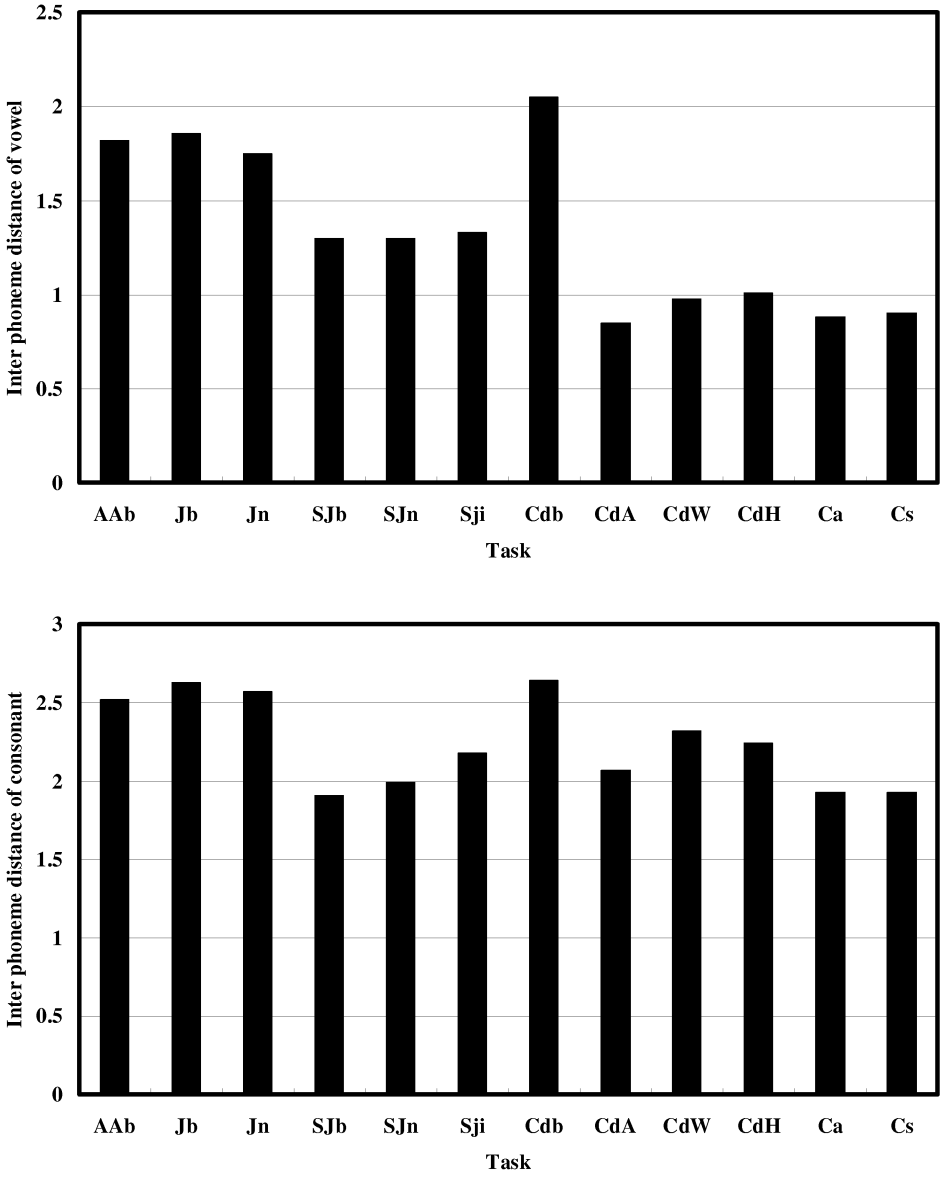


図 B.3 各タスクの平均音素間距離*

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

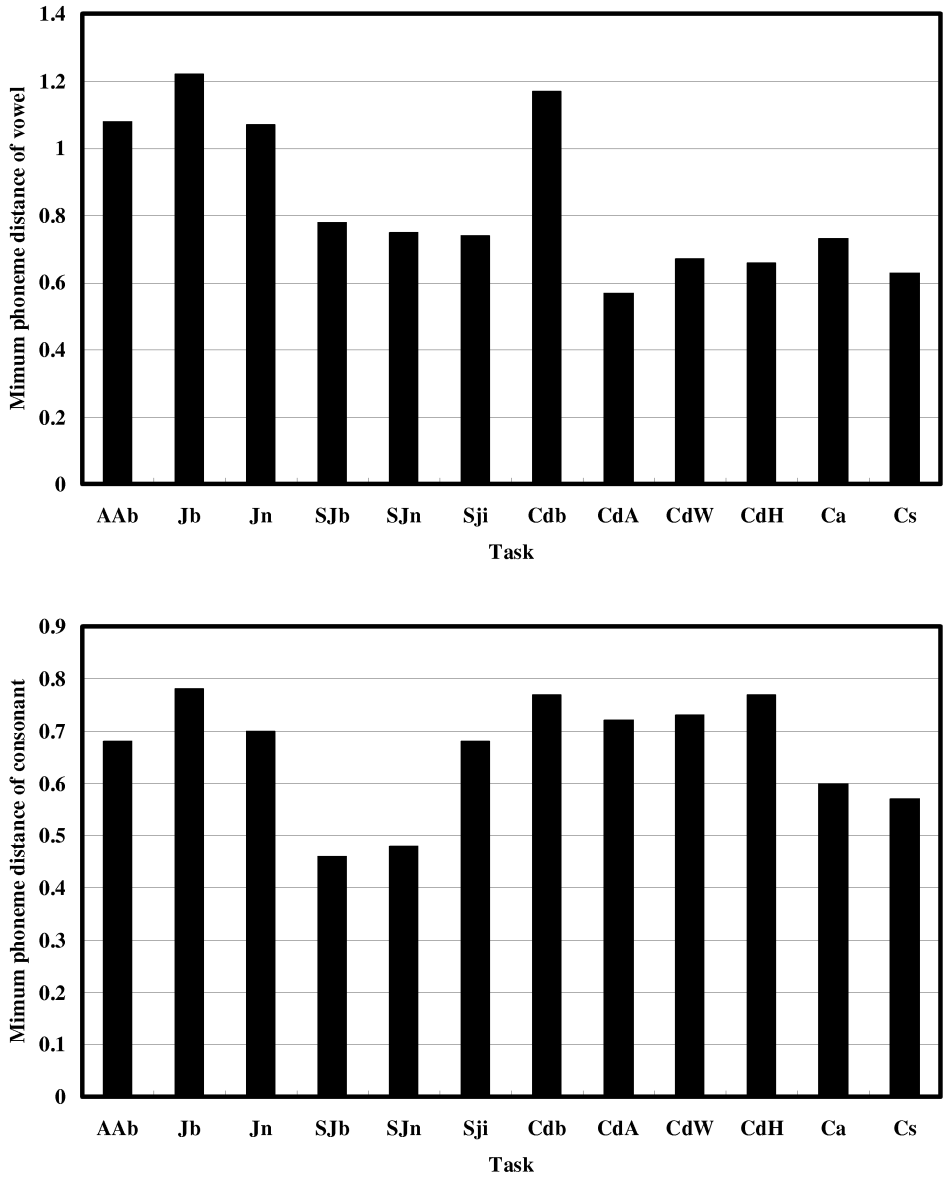


図 B.4 各タスクの最小音素間距離*

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

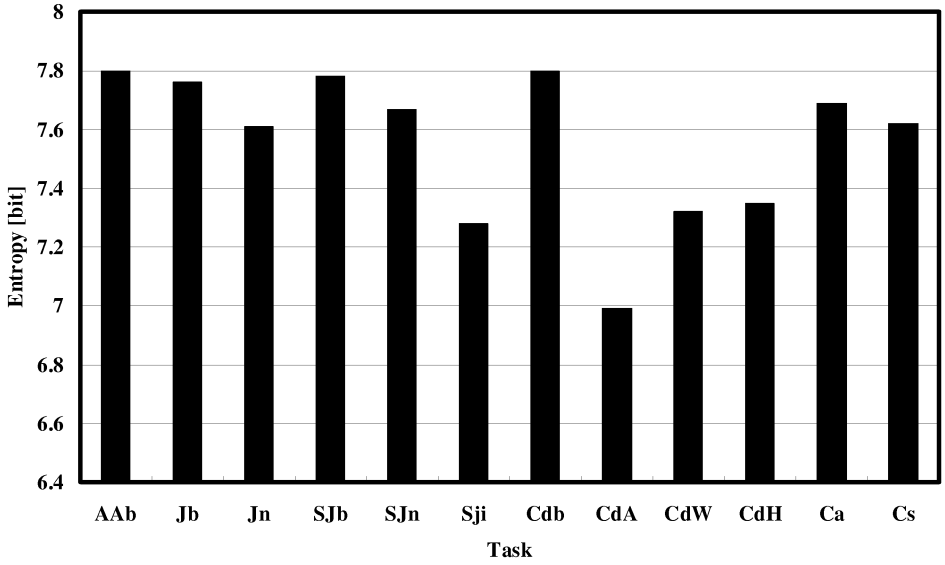


図 B.5 各タスクの語彙の偏り*

相関係数の絶対値が 1 に近ければ 2 変数間の相関は強くなり，0 となる場合を無相関と呼ぶ．本節では，この相関係数を用い，音声認識性能と各物理量との相関を調査することで，各物理量がタスク依存性に与える影響について分析を行う．まず，クロスタスクの音声認識実験結果から，タスクが一致している場合の認識性能（Closed 評価）と各物理量との相関係数を求めることで，タスク難易度と相関の高い物理量を調査する．次に，クロスタスクの音声認識実験結果から，他のタスクに対する認識性能と各物理量との相関を調査することで，タスク間の関連性と相関の高い物理量を調査する．

B.2.1 タスク難易度と物理量との相関分析

2.4.2 節のクロスタスクの音声認識実験結果における，タスクが一致している場合の認識性能（Closed 評価）と，B.1 節の各タスクの各物理量の結果から，式（B.4）に従い得られる相関係数を表 B.2 及び図 B.6 に示す．表 B.2 及び図 B.6 よ

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

表 B.2 各物理量とタスク難易度との相関係数

Model type	SNNR [dB] Avg/Std	Speed [mora/sec] Avg/Std	Entropy [bit]	Average of PD		Minimum of PD	
				Vowel	Consonant	Vowel	Consonant
monophopne	0.64/-0.27	-0.45/-0.57	0.11	0.75	0.59	0.66	0.32
triphopne	0.69/-0.38	-0.39/-0.65	0.34	0.84	0.70	0.80	0.36

り，モデル構造によらず，SNNR の平均，発話速度の標準偏差，母音及び子音の平均音素間距離，母音の最小音素間距離が比較的高い相関係数（0.6 以上）を示しており，タスク難易度と関連性の高い物理量であることが分かる．この結果より，SNNR が高い，発話速度の標準偏差が小さい，音素間距離の大きいといった特徴の持つタスクはタスク難易度が低く，これらの特徴と反対の特徴を持つタスクはタスク難易度が高くなることが分かる．また，過去の報告でも，この結果と同様に，これらの物理量がタスク難易度に影響を与えることを示す実験結果が報告されている [36, 34, 33]．但し，高齢者発話タスクである S-JNAS の各タスクはそれぞれ，音素間距離が小さいにもかかわらず，タスク難易度は低く，また，タスク難易度の高い，CIAIR の対話タスクや CSJ のタスクは，SNNR が低いだけでなく，発話速度の標準偏差が大きく，音素間距離も小さい．これらの結果から，実際は一つの物理量だけでタスク難易度を表現できるというわけでなく，複数の物理量の影響を受け，タスク難易度が表現されていることが分かる．

B.2.2 タスク間の関連性と物理量との相関分析

タスク間の関連性を示すクロスタスクの音声認識性能と，各物理量との相関を調査するために，ある一つのタスクを目的タスクとした場合の，目的タスクの評価データに対する他のタスクのタスク依存音響モデルの音声認識性能と，目的タスクの物理量に対する他のタスクの物理量との差分値（絶対値）から，相関係数を求める．

各タスクを目的タスクとした場合の，各物理量と認識性能との相関係数を表 B.3 に示す．また，各物理量における，全目的タスクの相関係数の平均値，最大値及び最小値を図 B.7 に示す．基本的に，差分値が小さい程，特徴が似ているため，性能も高くなる傾向にあることから，相関係数は負の値をとることが多いが，相

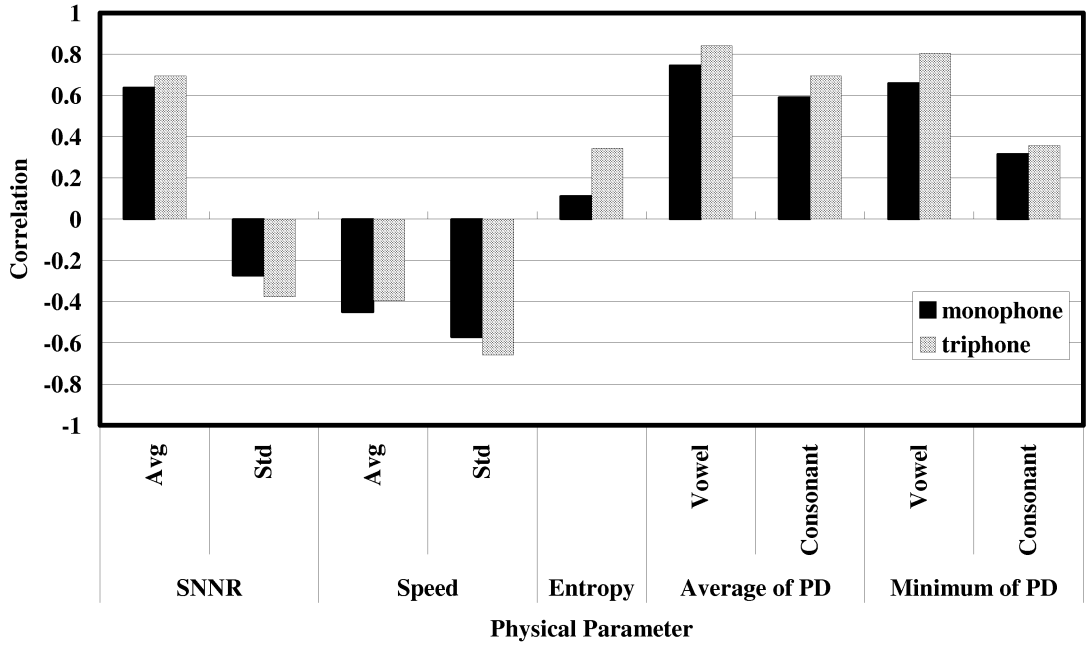


図 B.6 各物理量とタスク難易度との相関係数

関係数の値が小さい場合には例外的に正の値をとる場合もある．なお，図 B.7 は，図の見易さを考慮し，正負を逆転させた値をもとに図示している．表 B.3 及び図 B.7 より，SNNR の標準偏差及び発話速度の標準偏差が，モデル構造の違いによらず，タスク間の音声認識性能に高い相関を示していることが分かる．

特に，SNNR の標準偏差は，発話速度の標準偏差と比較し，相関係数のばらつきが小さい．これは，SNNR の標準偏差の，タスク間の音声認識性能に対する相関が，目的タスクによらず高いことを示しており，タスク依存性を要因付ける様々な物理量において，SNNR の標準偏差が重要な物理量であることが分かる．しかしながら，各タスクを目的タスクとした場合の，最も相関が高い物理量（表 B.3 中の太字）は，タスクにより異なることが分かる．これは，目的タスクに応じて，重要な物理量が異なることを示すと共に，冒頭で述べた通り，タスクには様々な要因が複雑に絡み合っているため，クロスタスクの音声認識性能に表れるタスク依存性を一つの物理量で語ることが困難であることを意味している．

タスク難易度と異なり，タスクが一致している場合だけでなく，より複雑なタスク間の関連性との相関であるため，重要とされる物理量の傾向，項目は異なるが，タスク難易度及びタスク間の関連性のどちらも，一つの物理量だけでなく，複数の物理量により表現されるということが分かる．

なお，各物理量を説明変数とし，目的変数を認識性能とした場合の，認識性能の予測式を重回帰分析により求めた．その結果を表 B.4 及び表 B.5 に示す．表 B.4 は，重相関係数及び，重回帰分析により得られる予測式における，平均予測誤差を示している．また表 B.5 は，予測式における各説明変数（物理量）の係数及び切片を示している．表 B.4 より，目的タスクによらず，高い相関及び小さい予測誤差を示しており，性能予測式の精度が高いことを示している．また表 B.5 より，目的タスクにより係数の傾向が異なることが分かる．これは，単回帰分析の結果と同様，目的タスクにより重要な物理量が異なることを示していると考えられる．認識率の推定は本論文の主旨ではないため，ここで深くは議論しないが，将来的に，複数の物理量を用いて性能を予測する技術の実現が期待できる．

B.3 まとめ

本付録では，国内に現存する大規模日本語音声コーパスから，JNAS 読み上げ音声コーパス，S-JNAS 高齢者音声コーパス，APP-BLA 多数話者音声コーパス，CIAIR 車内音声コーパス，CSJ 話し言葉音声コーパスを用い，各音声コーパスの各タスク，計 12 タスクに関してそのタスク依存性の分析として物理量の分析を行った．物理量分析では，各音声コーパス間において全ての物理量の傾向が完全に一致することはなく，たとえ同一コーパス内でもタスクの設定により，一部の物理量が大きく異なることを確認した．クロスタスクの音声認識実験と物理量分析結果との相関分析では，タスク難易度及びタスク間の関連性に影響を与える物理量を調査し，タスク依存性を分析する上で各物理量の重要性を確認するとともに，タスク依存性が，一つの物理量だけでなく，複数の物理量により表現されるということを確認し，タスク依存性を議論する上で，物理量分析の重要性を示した．

今後の課題としては，タスク依存性を表現する物理量を増やし，より高い精度でタスク依存性を表現することである．例えば，本付録での実験では，雑音環境

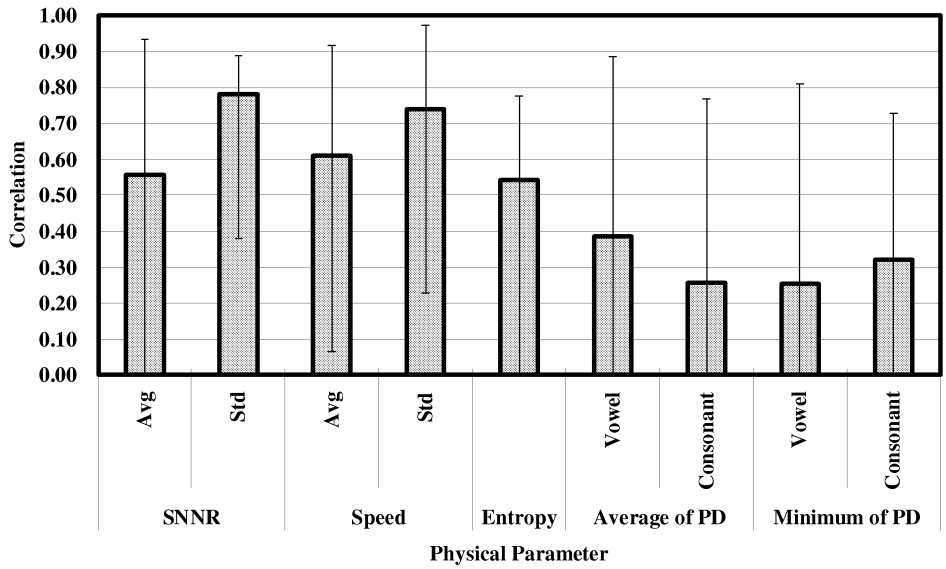
を示す物理量としてSNNRのみを用いたが，背景雑音の特性は，タスクにより異なることから，その雑音特性の違いを表現する物理量を分析する必要がある．また，ATR-APP_balanceは地域性を含んでおり，話者数も非常に多く，タスク依存性に影響を与えていることが推察されるが，その特徴を物理量で表現することは難しく，新たに物理量を定義する必要性が伺える．分析する対象のタスクの数，種類が増えるにつれ，考慮すべき物理量（雑音特性や残響特性等）が増えることとなり，分析の複雑さは増すが，タスク依存性をより高精度に表現し，理解を深めるためには，関連するであろう多数の物理量での分析をすすめ，データベース化を行うことが今後の重要な課題と考えられる．

表 B.3 各目的タスクにおけるクロスタスクの認識性能と各物理量との相関係数（上：monophone，下：triphone）*

Target task	SNNR [dB] Avg/Std	Speed [mora/sec] Avg/Std	Entropy [bit]	Average of PD		Minimum of PD	
				Vowel	Consonant	Vowel	Consonant
AAb	-0.91/-0.83	-0.43/**-0.93**	-0.71	-0.81	-0.29	-0.77	0.02
Jb	-0.85/-0.75	-0.77/**-0.96**	-0.65	-0.88	-0.47	-0.80	0.10
Jn	-0.86/-0.74	-0.76/**-0.97**	-0.55	-0.85	-0.31	-0.73	0.22
SJb	-0.93/-0.84	-0.91/**-0.92**	-0.56	-0.40	0.05	-0.03	-0.49
SJn	-0.90/-0.84	**-0.92**/-0.91	-0.52	-0.42	0.06	0.04	-0.50
SJi	-0.87/-0.87	**-0.90**/-0.88	0.21	-0.55	0.14	-0.02	0.23
Cdb	-0.41/-0.38	-0.73/-0.65	-0.38	**-0.89**	-0.69	-0.81	-0.31
CdA	-0.44/-0.73	-0.37/-0.57	**-0.78**	-0.15	-0.16	-0.22	-0.61
CdW	-0.47/**-0.84**	-0.59/-0.68	-0.55	-0.20	-0.77	-0.15	-0.73
CdH	-0.45/**-0.83**	-0.55/-0.66	-0.58	-0.21	-0.67	-0.15	-0.71
Ca	0.32/**-0.86**	-0.07/-0.51	-0.76	0.27	0.01	0.22	-0.61
Cs	0.08/**-0.89**	-0.33/-0.23	-0.70	0.44	0.01	0.38	-0.47
Average	-0.56/**-0.78**	-0.61/-0.74	-0.54	-0.39	-0.26	-0.25	-0.32

Target task	SNNR [dB] Avg/Std	Speed [mora/sec] Avg/Std	Entropy [bit]	Average of PD		Minimum of PD	
				Vowel	Consonant	Vowel	Consonant
AAb	-0.72/**-0.92**	-0.39/-0.86	-0.91	-0.67	-0.14	-0.72	0.24
Jb	-0.64/-0.86	-0.66/**-0.91**	**-0.91**	-0.70	-0.24	-0.70	0.36
Jn	-0.68/-0.83	-0.66/**-0.93**	-0.78	-0.71	-0.15	-0.68	0.43
SJb	-0.85/-0.89	-0.92/**-0.94**	-0.75	-0.30	0.01	0.00	-0.58
SJn	-0.83/-0.89	**-0.92/-0.92**	-0.70	-0.35	0.03	0.04	-0.59
SJi	-0.82/**-0.92**	-0.90/**-0.92**	0.27	-0.49	0.22	-0.01	0.30
Cdb	-0.31/-0.77	-0.73/-0.78	**-0.89**	-0.79	-0.43	-0.81	0.16
CdA	-0.65/**-0.82**	-0.42/-0.70	-0.57	-0.26	-0.13	-0.27	-0.39
CdW	-0.60/**-0.85**	-0.57/-0.69	-0.36	-0.26	-0.52	-0.15	-0.40
CdH	-0.65/**-0.81**	-0.53/-0.64	-0.30	-0.22	-0.24	-0.12	-0.44
Ca	0.32/-0.86	0.16/-0.40	**-0.95**	0.34	0.04	0.26	-0.38
Cs	0.07/**-0.92**	-0.23/-0.30	-0.86	0.42	0.00	0.38	-0.45
Average	-0.53/**-0.86**	-0.57/-0.75	-0.64	-0.33	-0.13	-0.23	-0.14

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．



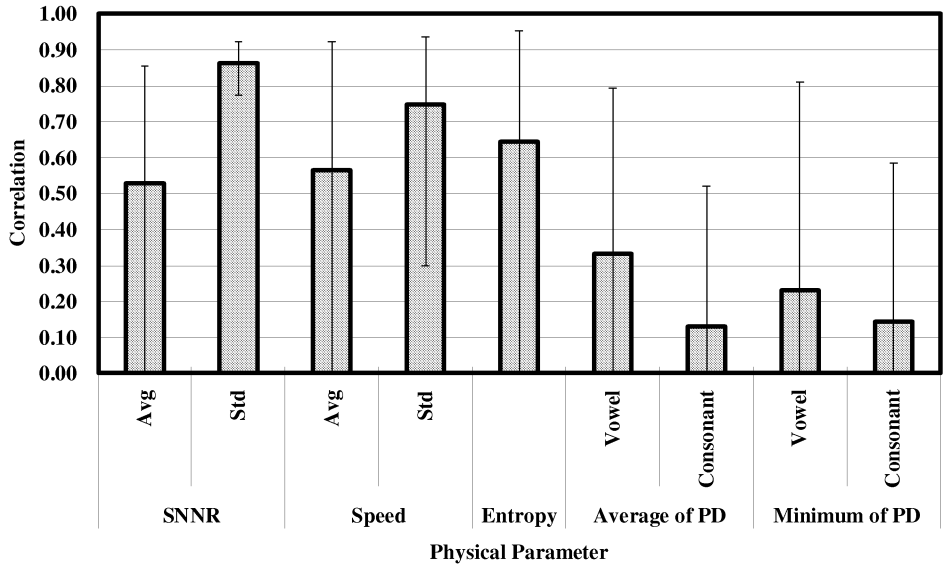


図 B.7 各目的タスクにおけるクロスタスクの認識性能と各物理量との相関係数のタスク平均（上：monophone，下：triphone）*

表 B.4 目的タスク別の重相関係数及び予測誤差（上 : monophone, 下 : triphone）

Acoustic model	Evaluation task											
	AAb	Jb	Jn	SJb	SJn	SJi	Cdb	CdA	CdW	CdH	Ca	Cs
Correlation coefficient	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.98	1.00	0.96
Prediction error	1.05	0.51	0.31	0.78	0.71	1.38	0.86	2.14	1.48	1.05	0.55	3.68

Acoustic model	Evaluation task											
	AAb	Jb	Jn	SJb	SJn	SJi	Cdb	CdA	CdW	CdH	Ca	Cs
Correlation coefficient	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.99
Prediction error	3.05	1.73	1.33	0.92	1.11	1.83	2.38	0.93	0.79	2.92	0.54	3.44

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

表 B.5 重回帰分析により算出された目的タスク別性能予測式の各変数の係数（上：monophone，下：triphone）*

Target task	Y-Intercept	SNNR [dB] Avg/Std	Speed [mora/sec] Avg/Std	Entropy [bit]	Average of PD		Minimum of PD	
					Vowel	Consonant	Vowel	Consonant
AAb	55.20	-1.76/1.11	-1.80/-0.80	-2.32	-6.19	4.85	0.43	-1.88
Jb	55.95	-2.54/-1.63	0.33/-2.22	0.44	-2.52	1.79	-1.01	-0.98
Jn	56.30	-1.79/-1.86	0.27/-2.27	-0.21	-4.72	1.80	1.58	-0.52
SJb	42.94	-4.06/-5.86	1.19/-0.77	0.61	-6.18	2.39	4.54	-2.06
SJn	44.83	-4.11/-6.27	1.59/-0.09	0.73	-7.26	2.24	5.94	-2.61
SJi	47.14	0.20/-1.08	-5.78/-2.20	0.22	-5.58	-0.27	1.45	-0.23
Cdb	57.05	1.93/2.43	-13.37/8.49	0.71	7.88	3.54	-14.60	-2.23
CdA	55.26	-1.31/-4.79	2.82/-0.19	2.75	3.14	-8.95	3.20	-4.31
CdW	48.41	-0.87/-1.75	2.31/0.06	1.81	-0.36	-1.77	-1.40	-2.09
CdH	48.17	0.30/-1.07	1.90/-2.20	-0.32	-0.62	-1.05	1.75	-0.42
Ca	48.84	-0.67/-4.74	-2.05/2.09	-0.31	-1.33	-0.15	0.76	-0.58
Cs	44.71	-0.54/-0.88	-2.70/0.52	-1.86	9.07	-5.79	-3.84	0.59

Target task	Y-Intercept	SNNR [dB] Avg/Std	Speed [mora/sec] Avg/Std	Entropy [bit]	Average of PD		Minimum of PD	
					Vowel	Consonant	Vowel	Consonant
AAb	56.58	-2.30/3.05	-6.02/-5.41	-7.63	3.59	3.04	-6.01	2.15
Jb	57.05	-2.81/-7.54	1.71/1.03	-3.18	-3.28	2.64	-4.17	4.13
Jn	55.49	-0.51/-1.52	-3.36/-1.85	-2.54	-2.50	2.60	-4.17	4.89
SJb	38.18	-3.64/-5.53	-1.59/0.51	-4.68	-7.18	0.82	7.34	-3.82
SJn	38.41	-2.54/-4.26	-3.17/-0.19	-4.68	-8.02	-1.38	8.62	-1.83
SJi	46.08	2.61/-5.64	-7.96/-4.01	-3.59	-5.69	1.08	2.99	2.73
Cdb	55.27	-4.27/0.20	4.35/2.51	-6.33	-14.47	7.42	-1.10	0.56
CdA	52.93	-3.40/-9.93	4.81/0.15	9.57	3.60	-14.32	4.14	-6.67
CdW	49.69	-1.88/-3.88	0.21/-3.79	-1.04	5.32	-2.29	1.15	5.75
CdH	49.40	-3.09/-2.08	1.73/-8.46	-3.33	4.03	-2.47	9.65	8.79
Ca	51.28	-0.27/-3.43	-2.05/1.28	-6.84	-1.16	-2.32	3.11	-0.55
Cs	44.41	0.12/3.35	-4.25/-3.37	-7.65	18.70	-13.81	-4.31	3.94

*略称に関しては表 2.2 参照．タスクの概要に関しては 2.4.1 節参照．

研究業績

学術論文

1. **Goshu Nagino**, Makoto Shozakai, Tomoki Toda, Hiroshi Saruwatari and Kiyohiro Shikano, "Building An Effective Speech Corpus by Utilizing Statistical Multidimensional Scaling Method," *IEICE Transactions on Information and Systems*, vol. E91-D, no. 3, pp. 607-614, March 2008.

2. **奈木野豪秀**, 庄境誠, 鹿野清宏, "多次元尺度法を用いた統計的音響モデルの可視化手法", 電子情報通信学会論文誌, vol. J91-D, no.11, November 2008.

国際会議

1. **Goshu Nagino** and Makoto Shozakai, "Design of ready-made acoustic model library by two-dimensional visualization of acoustic space," *International Conference on Spoken Language Processing (ICSLP)*, pp.2965-2968, 2004.

2. **Goshu Nagino** and Makoto Shozakai, "Building an effective corpus by using acoustic space visualization (COSMOS) method," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp.449-452, 2005.

3. **Goshu Nagino** and Makoto Shozakai, "Analyzing Reusability of Speech Corpus based on Statistical Multidimensional Scaling Method," *International Conference on Spoken Language Processing (ICSLP)*, pp.161-164, 2006.

4. **Goshu Nagino** and Makoto Shozakai, "Distance Measure between Gaussian Distributions for Discriminating Speaking Styles," *International Conference on Spoken Language Processing (ICSLP)*, vol., pp.657-660, 2006.

5. **Goshu Nagino**, Makoto Shozakai, Kiyohiro Shikano, "How to Judge Reusability of Existing Speech Corpora for Target Task by Utilizing Statistical Multidi-

mensional Scaling," *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1302-1305, Aug. 2007.

6. Makoto Shozakai and **Goshu Nagino**, "Analysis of speaking styles by two-dimensional visualization of aggregate of acoustic models," *International Conference on Spoken Language Processing (ICSLP)*, pp.717-720, 2004.

7. Makoto Shozakai and **Goshu Nagino**, "Acoustic space analysis method utilizing statistical multidimensional scaling technique," *International Workshop on Nonlinear Signal and Image Processing (NSIP)*, 2005.

8. Makoto Shozakai and **Goshu Nagino**, "Improving Robustness of Speech Recognition Performance to Aggregate of Noises by Two-Dimensional Visualization," *European Conference on Speech Communication and Technology (EUROSPEECH)*, pp.921-924, 2005

研究会

1. **奈木野豪秀**, Daniel Willett, 南泰浩, Erick McDermott, 中村篤, 宮崎昇, 鹿野清宏, "実対話音声を用いた有限状態トランスデューサ型認識デコーダの評価", 電子情報通信学会技術研究報告, SP2001-125, pp. 33-40, 2002.

2. **奈木野豪秀**, 庄境誠, "2 次元可視化手法に基づいた音響空間分割による音響モデルライブラリの開発," 電子情報通信学会技術研究報告, SP2004-41, pp.7-12, 2004.

3. Daniel Willett, Yasuhiro Minami, **Goshu Nagino**, Erick. McDermott, and Shigeru. Katagiri, "A Time-Synchronous Real-time Approach for Integrated Recognition and Segmentation of Unsegmented speech data", *Spontaneous Speech Science and Technology Workshop*, pp. 137-142, 2001.

4. 庄境誠, **奈木野豪秀**, "多次元尺度構成法による音響空間の 2 次元可視化," 情報処理学会研究報告, SLP2004-74, pp. 129-136, 2004. （平成 18 年度山下記

念研究賞）

大会発表

1. **奈木野豪秀**, 庄境誠,”COSMOS 法を用いた音響モデルライブラリの構築,” 日本音響学会講演論文集, 2-1-20, pp.75-76, Sep, 2004.

2. **奈木野豪秀**, 谷智洋, 庄境誠,”COSMOS 法を用いた効率的な音声コーパスの構築,” 日本音響学会講演論文集, 2-Q-21, pp.139-140, Mar, 2005.

3. **奈木野豪秀**, 谷智洋, 庄境誠,”多数音声コーパスより作成された音響空間地図の作成と音響モデルライブラリの構築,” 日本音響学会講演論文集, 2-Q-22, pp.141-142, Mar, 2005.

4. **奈木野豪秀**, 庄境誠,”COSMOS 法を用いた既存音声コーパスの相関分析,” 日本音響学会講演論文集, 1-P-15, pp.181-182, Sep, 2005.（第 15 回ポスター賞）

5. **奈木野豪秀**, 庄境誠,”複数音声コーパスのタスク依存性・再利用性の分析,” 日本音響学会講演論文集, 1-P-16, pp.176-177, Mar, 2006.

6. **奈木野豪秀**, 鹿野清宏, 庄境誠,”国内既存音声コーパスの音響空間配置図の作成,” 日本音響学会講演論文集, 2-P-15, pp.123-124, Sep, 2006.

7. **奈木野豪秀**, 原囿友輔, 庄境誠,”統計的 MDS 法を利用した性能分布近似技術に関する検討,” 日本音響学会講演論文集, 2-P-16, pp.125-126, Sep, 2006.

8. **奈木野豪秀**, 鹿野清宏, 庄境誠,”統計的 MDS 法を用いた既存音声コーパスの再利用性判定手法,” 日本音響学会講演論文集, 1-P-24, pp.171-172, Mar, 2007.

9. 庄境誠, **奈木野豪秀**,”2 次元視覚化手法を利用した音響モデルの高精度化,” 日本音響学会講演論文集, 3-Q-21, pp.185-186, Mar, 2004.

10. 庄境誠, **奈木野豪秀**,”雑音コーパスの二次元可視化による音声認識性能の耐雑音性の改善,” 日本音響学会講演論文集, 2-7-6, pp.69-70, Sep, 2005.

11. 庄境誠, **奈木野豪秀**, 鹿野清宏, ”多数日本語音声コーパスからの日本語音響空間地図の作成,” 日本音響学会講演論文集, 1-11-12, pp.45-46, Mar, 2006.

12. 庄境誠, **奈木野豪秀**,”音声コーパスの可視化手法の比較検討” 日本音響学会講演論文集, 1-P-15, pp.173-175, Mar, 2006.

受賞

1. 日本音響学会 第 15 回ポスター賞 「1-P-15 COSMOS 法を用いた既存音声コーパスの相関分析」 2005.

2. 科学・技術賞 「超大規模多次元情報群の空間可視化及び解析技術（COSMOS 法）の研究」（社内表彰） 2008.

特許

1. 庄境誠, **奈木野豪秀**,”データ処理装置及びデータ処理装置制御プログラム,” 特許番号:1-669-979(DE,FR,GB), 国際公開番号:WO2005/034086, 国際出願番号:PCT/JP2004/010390.

2. **奈木野豪秀**, 庄境誠,”標準パターン作成方法、作成装置及び作成プログラム,” 特開 2004-334024

3. 庄境誠, **奈木野豪秀**,”音声認識装置、音声認識方法、及び、プログラム,” 特開 2006-91864

4. 庄境誠, **奈木野豪秀**,”パターンモデル生成装置、パターンモデル評価装置およびパターン認識装置,” 特開 2007-65491