

博士論文

部分空間に基づく話者性の抽出と
その応用に関する研究

2002 年 1 月

龍谷大学大学院 理工学研究科

博士後期課程 電子情報学専攻

T99D003 西田 昌史

博士論文

部分空間に基づく話者性の抽出と
その応用に関する研究

2002 年 1 月

龍谷大学大学院 理工学研究科
博士後期課程 電子情報学専攻
T99D003 西田 昌史

部分空間に基づく話者性の抽出とその応用に関する研究

論文要旨

音声データには、発話内容を表す音韻性と、話者が誰であるかを表す話者性が含まれている。この音韻性と話者性を分離し、話者性だけを持つ音声データを基に話者モデルを設計することができれば、発話内容に依存しない頑健な話者認識が可能になると考えられる。このことから、本論文では、音韻性と話者性の分離に基づいた話者性の抽出に関する研究を行った。

話者認識の手法としては、従来、音声データを統計的にモデル化するVQやGMMといった方法がよく用いられてきた。しかし、これらの手法は、音韻性と話者性が分離されていないため音韻性に起因する特徴変動が生じることから、精度の高い認識が困難であると考えられる。また、多くの学習データが必要であり、モデルのパラメータ数も比較的多いといった問題がある。

そこで、本論文では、音韻性と話者性を分離する手法として、主成分分析に基づいた部分空間法に注目した。主成分分析に基づいた部分空間法は、学習パターンの分散が最大となるように、部分空間を設計する手法である。また、パラメータ数が少なく、高速に計算が可能な手法である。本論文では、学習パターンを主成分分析して得られる低次の部分空間は、その話者の音声データをよく表現した空間であり、話者固有音声空間と呼ぶ。話者の認識は、学習データの平均ベクトルと入力特徴ベクトルとの差ベクトルを求め、この差ベクトルと話者固有音声空間との投影距離を求め、この距離が最も短い話者を認識結果とする。

学習データの平均ベクトルと入力特徴ベクトルとの距離が離れていた場合、異なる話者である可能性が高い。しかし、従来の主成分分析に基づく部分空間法では、入力特徴ベクトルと話者固有音声空間との距離のみで話者を識別している。そこで、入力特徴ベクトルと平均ベクトルとの距離を考慮した音声の特徴変動に頑健な話者性の抽出手法として、話者固有音声空間における動的・静的特徴統合による話者照合法を提案した。従来の部分空間法と本手法に対して話者照合実験を行った結果、本手法は、従来の部分空間法に比べて約75%の照合精度の改善が得られ、音声の特徴変動に対して頑健な手法であることがわかった。

話者固有音声空間における動的・静的特徴統合法では、静的特徴距離と動的特徴距離を統合化することで話者性を表していた。この話者固有音声空間は、学習パターンのお話者の音声データをよく表現した空間であるが、音韻性と話者性が含まれている。発話内容による音声の特徴変動は、音韻性、すなわち音声中に含まれている音韻のばらつきにより生じると考えられる。これに対して、発話内容に依存しない話者性が存在するのではないかと考えられる。この点から、本論文では、話者内分散の大きい空間を音韻空間、話者内分散の小さい空間を話者空間として、主成分分析に基づいた部分空間分離による音韻性と話者性の分離手法を提案した。さらに、本手法に基づいて、音韻性を抑え話者性を精度よく抽出する手法として、話者空間への射影に基づく統計的モデルによる話者認識法を提案した。従来のGMMによる手法と本手法に対して話者識別実験を行った結果、本手法は4混合分布の時に、従来の64混合分布のGMMに比べると、約43%の識別精度の改善が得られた。したがって、本手法は、少量のパラメータ数で頑健な話者モデルを構成でき、高速に学習および識別を行えることがわかった。

これら部分空間により抽出された話者性の応用として、アナウンサー発話の自動抽出と音声認識を統合化したニュース記事の自動分類システムを提案し構築する。NHK5分間のニュース45日分に対して記事分類実験を行った結果、アナウンサー発話の抽出率93.2%、記事分類率70.8%が得られ、本システムが有効であることがわかった。また、特定の話者がある話題について話しているシーンの検索を目的として、話者認識と音声認識を統合した話者の人名付与システムを提案し構築する。NHK15分間のニュース4日分に対して人名付与実験を行った結果、話者交替検出率80.3%、話者同定率89.2%、人名付与率70.0%が得られ、本システムが有効であることがわかった。

A Study on Speaker Information Extraction and Its Application Based on Subspace

Abstract

Speech data includes both of phonetic information and speaker information. If these two kinds of information are separated each other and a speaker model can be constructed by only the speaker information, sentences independent robust speaker recognition will be realized. This thesis studied on speaker information extraction based on separation of the phonetic information and speaker information included in the speech data.

In speaker recognition, VQ and GMM have been conventionally used and are statistically constructed using features of speech data. The conventional methods are based on a statistical method but not based on the separation of the phonetic information and speaker information. Therefore, the conventional methods can not obtain a high recognition performance because speech feature varies depending on the phonetic information. Moreover, the conventional methods have a problem that many training data are required and the number of the model parameters is relatively large.

This thesis noticed a subspace method based on PCA to separate the phonetic information and speaker information. This method can be constructed a subspace where variance of a training data is maximized by a few model parameters and real time training. A lower order subspace obtained by PCA represents a speech data of a training speaker and this thesis calls the subspace "speaker specific eigenvoice space". Speaker recognition can be carried out by computing a distance between an input feature vector and the speaker specific eigenvoice space.

If a distance is large between a mean vector of training data and an input feature vector, the speaker will be different from a speaker of training data. However, a conventional subspace method based on PCA identifies a speaker based only on the distance between a speaker specific eigenvoice space and the input feature vector. This thesis proposes a speaker verification method by integrating dynamic features in the speaker space and static features of the distance between the mean vector and the input

feature vector in order to robustly extract the speaker information. The proposed method was shown to be effective by speaker verification experiments. As a result, Equal Error Rate was reduced by 75% by the proposed method compared with the conventional subspace method and the proposed method was shown to be robust for the speech feature variation.

The speaker specific eigenvoice space represents a speech data of a training speaker but includes both of phonetic information and speaker information. The speech feature variation is mainly caused by the variation of the phonetic information included in speech data. On the otherhand, the speaker information does not depend on sentences so that the variation is small. According to this insight, this thesis proposes a method to separate the phonetic information and speaker information by PCA, under the assumption that within-speaker variance is large in a lower order subspace, so called "phonetic space" and within-speaker variance is small in a higher order subspace, so called "speaker space". Moreover, this thesis proposes a speaker recognition method by the statistical model based on projection to the speaker space where the phonetic information is already suppressed. The proposed method was shown to be effective by speaker identification experiments. As a result, Identification Error Rate was reduced by 43% by the proposed method(4 mixtures) compared with the conventional method based on GMM(64 mixtures). Therefore, it can be said that the proposed method can construct a robust speaker model by a few model parameters and a training data and also, training and identification can be performed in a real time.

In application using speaker information extracted by the subspace, this thesis proposes an news article classification system by integrating an automatic speech recognition and automatic extraction of announcer utterance. As a result of article classification experiments by the proposed method, the extraction rate of announcer utterance was 93.2% and article classification rate was 70.8% so that the system was shown to be effective. Further, this thesis proposes a speaker name indexing system by integrating a speaker recognition and a speech recognition for simultaneous retrieval of a speaker and utterances. As a result of speaker name indexing experiments by the proposed method, detection rate of speaker changes was 80.3% and a speaker verification rate was 89.2%, a name indexing rate was 70.0% so that the system was shown to be effective.

目次

論文要旨	i
Abstract	iii
図目次	ix
表目次	xi
1 序論	1
1.1 はじめに	1
1.2 本論文の構成	5
2 話者認識	7
2.1 まえがき	7
2.2 話者認識について	7
2.2.1 話者識別と話者照合	7
2.2.2 話者認識の分類	9
2.2.3 特徴抽出	10
2.3 従来手法	13
2.3.1 VQ 歪み	13
2.3.2 HMM	15
2.3.3 GMM	19
2.3.4 判別分析	21
2.3.5 部分空間法	24
2.3.6 識別学習	30

2.4	まとめ	34
3	話者認識における距離尺度	35
3.1	まえがき	35
3.2	ベイズ識別	36
3.3	疑似ベイズ	37
3.4	マハラノビス距離	38
3.5	疑似マハラノビス距離	38
3.6	投影距離	39
3.7	改良マハラノビス距離	39
3.8	重み付きマハラノビス距離	40
3.9	話者照合実験	41
3.9.1	実験条件	41
3.9.2	実験結果と考察	42
3.10	まとめ	44
4	話者固有音声空間における動的・静的特徴統合による話者照合	45
4.1	まえがき	45
4.2	ケプストラム平均値正規化	46
4.3	話者の動的・静的特徴	46
4.4	話者の動的・静的特徴統合	48
4.5	話者固有音声空間への投影距離との話者照合実験	49
4.5.1	実験条件	49
4.5.2	実験結果	50
4.5.3	時期差の考察	52
4.5.4	次元数の考察	52
4.6	GMMとの話者照合実験	54
4.6.1	実験条件	54
4.6.2	実験結果	54
4.6.3	処理時間の考察	56
4.7	まとめ	57

5	話者空間への射影に基づく統計的モデルによる話者認識	59
5.1	まえがき	59
5.2	補空間	60
5.3	音韻空間と話者空間の分離	61
5.4	話者空間への射影に基づく統計的モデル	63
5.5	話者認識実験	65
5.5.1	実験条件	65
5.5.2	実験結果と考察	65
5.6	識別学習による話者空間の最適化	68
5.7	話者照合実験	70
5.8	まとめ	71
6	アナウンサー発話の自動抽出と大語彙連続音声認識によるニュース記事分類	73
6.1	まえがき	73
6.2	アナウンサーの発話区間の抽出	74
6.2.1	アナウンサーの話者区間抽出法	74
6.2.2	アナウンサーの話者区間抽出実験	77
6.3	大語彙連続音声認識実験	80
6.3.1	実験条件	80
6.3.2	実験結果	81
6.4	ニュース記事分類	82
6.4.1	記事分類の概要	82
6.4.2	新聞記事からのキーワード選択法	84
6.4.3	記事の分類方法	84
6.5	ニュース音声記事の分類実験	85
6.5.1	実験条件	85
6.5.2	実験結果と考察	86
6.6	まとめ	87
7	話者認識と音声認識を統合した話者の人名付与システム	89
7.1	まえがき	89
7.2	話者セグメンテーション	90

7.2.1	話者セグメンテーション法	90
7.2.2	話者セグメンテーション実験	91
7.3	大語彙連続音声認識	93
7.3.1	認識システムの構成	93
7.3.2	大語彙連続音声認識実験	93
7.3.3	大語彙連続音声認識による人名, キーフレーズ検出	94
7.4	人名スポッティング	96
7.4.1	キーワードスポッティングによる人名の検出	96
7.4.2	人名スポッティング実験	98
7.5	人名インデキシング	100
7.6	まとめ	101
8	結論	103
8.1	まとめ	103
8.2	今後の課題	107
	謝辞	109
	参考文献	111
	発表論文	121

目次

2.1	話者識別	8
2.2	話者照合	9
2.3	基本周期とスペクトル包絡の抽出手順	12
2.4	ベイキス型HMMの例	16
2.5	エルゴディックHMMの例	18
2.6	判別分析による空間変換	23
2.7	CLAFIC法	27
2.8	主成分分析法	28
2.9	投影距離	30
3.1	各識別関数の関係	36
3.2	話者照合結果(1)	42
4.1	話者の動的・静的特徴	47
4.2	話者固有音声空間への投影距離との比較結果(1)	51
4.3	話者固有音声空間への投影距離による照合結果	53
4.4	話者固有音声空間への投影距離でテストデータをCMNした場合の照合結果	54
4.5	話者の動的・静的特徴統合による照合結果	55
4.6	平均累積寄与率	56
5.1	音韻ベクトルと話者ベクトル	63
5.2	話者空間への射影	64
6.1	話者セグメンテーション法	76

6.2	最初の話者区間の継続時間分布	79
6.3	音声区間の継続時間分布	80
6.4	最初の話者区間の継続時間削減率と発話区間抽出率	81
7.1	認識システム	93
7.2	キーワードスポッティング用言語モデル	97
7.3	2つのモデルでの尤度区間の対応	98
7.4	人名スポッティング結果(1)	99

表目次

3.1 分析条件 41

3.2 話者照合結果(2) 43

4.1 分析条件 50

4.2 話者固有音声空間への投影距離との比較結果(2) 52

4.3 GMMとの比較結果 55

4.4 処理時間とテンプレートサイズ 56

5.1 分析条件 66

5.2 話者識別結果(%) 66

5.3 話者照合結果(%) 67

5.4 各空間におけるFisher比 68

5.5 分析条件 70

5.6 話者照合結果(%) 71

6.1 実験条件 78

6.2 アナウンサーの話者区間切り出し結果(%) 78

6.3 音響分析とHMM 82

6.4 評価用ニュース音声データ 83

6.5 ニュース音声認識実験の結果(%) 83

6.6 キーワード「日米」の3つの分野に対する関連度と分類寄与率の例 . . 85

6.7 記事の分野の分布 86

6.8 ニュース音声記事の分類結果(%) 87

7.1 分析条件 92

7.2 話者セグメンテーション結果 92

7.3 音響分析とHMM 94

7.4 大語彙連続音声認識結果 (%) 95

7.5 キーフレーズリスト 95

7.6 人名, キーフレーズの検出結果 96

7.7 人名スポッティング結果 (2) 99

7.8 人名スポッティング結果 (3) 100

第1章

序論

1.1 はじめに

近年，話者認識技術は，コンピュータやネットワークの発達により，インターネットや電話によるバンキング，ショッピング，情報提供等のサービスの利用における個人認証やセキュリティへの応用として，盛んに研究されている．また，ニュースあるいは対談の書き起こしや話者のインデキシング [1]–[7]，検索，話者適応 [8]，話者クラスタリング [9]，対話ロボット等，発話者の認識といった応用として研究されている．

音声データには，発話内容を表す音韻性と，話者が誰であるかを表す話者性が含まれている．この音韻性と話者性を分離し，話者性だけを持つ音声データを基に話者モデルを設計することができれば，発話内容に依存しない頑健な話者認識が可能になると考えられる．また，音韻性だけを持つ音声データを基に音声モデルを設計することができれば，話者の違いに依存しない頑健な音声認識が可能になると考えられる．このことから，本論文では，音声データに含まれている音韻性と話者性の分離に基づいた話者性の抽出に関する研究を行った．

話者認識の手法としては，従来，音声データから抽出した特徴を統計的にモデル化する VQ (Vector Quantization) や GMM (Gaussian Mixture Model) といった方法がよく用いられてきた．しかし，これらの従来法は，音韻性と話者性が含まれている音声データを直接統計的にモデル化する手法であり，音韻性と話者性が分離されていないため音韻性に起因する特徴変動が生じることから，精度の高い認識が困難であると考えられる．また，モデルのパラメータを頑健に推定するために，多くの学習データが必要であり，モデルのパラメータ数も比較的多いといった問題がある．

そこで、本論文では、音韻性と話者性を分離する手法として、主成分分析に基づいた部分空間法に注目する。主成分分析に基づいた部分空間法は、学習パターンの分散が最大となるように、部分空間を設計する手法である。また、パラメータ数が少なく、高速に計算が可能な手法である。本論文では、学習パターンを主成分分析して得られる低次の部分空間は、その話者の音声データをよく表現した空間であり、話者固有音声空間と呼ぶ。話者の認識は、学習データの平均ベクトルと入力特徴ベクトルとの差ベクトルを求め、この差ベクトルと話者固有音声空間との投影距離を求め、この距離が最も短い話者を認識結果とする。

学習データの平均ベクトルと入力特徴ベクトルとの距離が離れていた場合、異なる話者である可能性が高い。しかし、投影距離は、入力特徴ベクトルと平均ベクトルとの距離を考慮していないので、認識を誤ってしまう場合がある。そこで、話者を識別する上で、入力特徴ベクトルと平均ベクトルとの距離が短く、かつ投影距離も短くなるように識別関数を定義する必要がある。これら2つの距離の変動を抑制するために、それぞれの距離を分散で正規化する。分散で正規化した入力特徴ベクトルと平均ベクトルとの距離はマハラノビス距離であり、分散で正規化した投影距離とこのマハラノビス距離を統合する。これらを踏まえて、本論文では話者認識の距離尺度として、2つの距離に重み付けすることで、話者性を強調した重み付きマハラノビス距離法を提案する。本手法と投影距離に基づく手法に対して話者照合による比較実験を行い、話者を識別する上で、入力特徴ベクトルと平均ベクトルとの距離を考慮することが有効であることを示す。これを踏まえて、音声の特徴変動に頑健な話者性の抽出手法として、話者固有音声空間における動的・静的特徴統合による話者照合法を提案する。

従来の部分空間法では、発話内容や時期差による特徴変動に起因して、学習データの平均ベクトルと入力特徴ベクトルとの間にずれが生じ、話者の認識精度が劣化する。話者固有音声空間における動的・静的特徴統合法は、この音声の特徴変動を正規化する手法である。話者の動的特徴とは、入力音声の特徴の時間的変化量であり、動的な話者性を表す。本論文では、この動的特徴と話者固有音声空間との距離を話者の動的特徴距離と呼ぶ。話者の静的特徴とは、入力音声の平均ベクトルを表す。この静的特徴は、入力音声の静的な話者性を表しているが、時期差の影響を受けて学習データの平均ベクトルとずれが生じる。つまり、静的特徴と学習データの平均ベクトルとの差（距離）には時期差が含まれており、本論文では、これを話者の静的特徴距離と呼ぶ。そこで、静的特徴距離に重み付けすることで時期差による変動を抑制し、動的特徴距

離と統合することによって、音声の特徴変動に頑健な話者照合が可能になると考えられる。本手法と従来の部分空間法に対して話者照合による比較実験を行い、本手法は、音声の特徴変動に対して頑健な手法であることを示す。

話者固有音声空間における動的・静的特徴統合法では、静的特徴距離と動的特徴距離を統合化することで話者性を表していた。この話者固有音声空間は、学習パターンのお話者の音声データをよく表現した空間であるが、音韻性と話者性が含まれている。発話内容による音声の特徴変動は、音韻性、すなわち音声中に含まれている音韻のばらつきにより生じると考えられる。これに対して、発話内容に依存しない話者性が存在するのではないかと考えられる。この点から、話者内分散に基づいて音韻性と話者性を分離することができると考えられる。つまり、各話者の音声データの分散が大きい空間では音韻性が強く表れ、分散が小さい空間は、音韻性が抑えられ話者性が表れているとみなす。そこで、本論文では、音韻性と話者性を分離する手法として、主成分分析に基づいた部分空間分離による話者性の抽出法を提案する。本手法では、学習パターンを主成分分析して得られる分散が大きい低次の部分空間を音韻空間と呼ぶ。また、分散が小さい高次の部分空間は、音韻空間に対する直交補空間であり、話者空間と呼ぶ。さらに、本手法に基づいて、話者空間に学習データを射影し、各話者の話者空間上でGMMにより話者性を統計的にモデル化する話者認識法を提案する。したがって、本手法により音韻性が抑えられ、話者性を精度よく抽出することができると考えられる。本手法と従来のGMMによる手法、判別分析によって変換された空間におけるGMMによる手法に対して、話者識別および話者照合による比較実験を行い、本手法は、音韻性を抑えることができ、話者性を精度よく抽出できることを示す。さらに、少量のパラメータ数で頑健な話者モデルを構成でき、高速に学習および識別、照合を行えることを示す。また、本手法で構成された話者空間は、他者との相関を考慮していないので、必ずしも他者との識別に最適な空間になっているとは限らない。そこで、学習パターンに対する誤識別率を最小化する識別学習による話者空間の最適化法を提案する。話者空間の最適化は、最小分類誤り学習（Minimum Classification Error）／一般化確率的降下法（Generalized Probabilistic Descent）により行う。本手法と従来のGMMによる手法、最適化していない話者空間におけるGMMによる手法に対して、話者照合による比較実験を行い、MCE/GPDによる話者空間の最適化が有効であることを示す。

次に、これら部分空間により抽出された話者性の応用について述べる。近年、放送

のデジタル化により、ユーザは見たい番組や関心のあるシーンを検索したり、番組の内容を要約して見たいといった要求を持つようになる。これらの要求に対応するには、番組の内容に対して音声や映像といったメディアの解析により、索引情報を自動付与しておく必要がある。そこで、本論文では、ニュース映像の構造化と検索に関する研究を行った。ニュース映像の構造化に関しては、話者セグメンテーションと音声認識を統合化したニュース記事の自動分類システムを提案し構築する。話者セグメンテーションでは、人手であらかじめ話者のモデルを学習しておくことは、自動化の点から望ましくない。そこで、本論文では、特定の話者をあらかじめ学習することなく、入力音声から話者モデルを自動学習し、話者照合に基づいて自動的に話者のセグメンテーションを行う方法を提案する。また、話者空間への射影に基づく統計的モデルとして、VQを用いた話者照合法を提案する。ニュース記事の分類では、話者セグメンテーションにより抽出したアナウンサー発話に対して音声認識を行い、認識結果から χ^2 法によりキーワードを自動選択し、選択されたキーワードと記事の分類分野との関連度を計算する。この関連度を分類分野毎に計算し、関連度が最大となる分野に、ニュース音声記事を自動分類する。NHK5分間のニュース45日分に対して記事分類実験を行い、本システムが有効であることを示す。

ニュース映像の検索に関しては、「クリントン大統領が、情報スーパーハイウェイについて話しているシーンを見たい」といった、特定の話者がある話題について話しているシーンの検索を目的としている。この処理を実現するために、本論文では、話者認識と音声認識を統合した話者の人名付与システムを提案し構築する。ニュース映像では、「(人名)は、次のように述べました」といった発話の交替を示すキーフレーズが存在する。したがって、このようなキーフレーズを含む発話の前後において、この話者とは異なる話者の発話を、インタビューされている人の発声とみなして、話者の名前を付与する。つまり、アナウンサー発話の抽出で用いた話者セグメンテーション法により、話者の交替を検出し、話者の同定を行う。アナウンサー発話の抽出における話者同定では、学習データ量を考慮せずに閾値を設定していたため、学習データ量が少ない場合、学習話者の閾値が小さく設定され、本人を誤って棄却しやすくなる。また、学習データ量が多くなると、学習話者の閾値が大きくなり、他人を誤って受理しやすくなってしまったといった問題があった。そこで、本論文では、学習データ量を考慮した閾値の設定手法を提案する。話者セグメンテーションを行った後、ニュース音声に対して音声認識を行い、発話の交替を示すキーフレーズと人名を検出し、自動構

築された話者モデルに対して、話者の名前を付与する。NHK15分間のニュース4日分に対して人名付与実験を行い、本システムが有効であることを示す。

1.2 本論文の構成

本論文の構成は、以下の通りである。

第2章では、話者識別と話者照合、話者認識における特徴量について述べる。また、VQ歪み、HMM、GMMによる統計的手法、ならびに判別分析、部分空間法による空間表現手法、識別学習といった従来話者認識手法について述べる。

第3章では、話者認識における識別関数について述べる。部分空間法がベイズ識別法の近似解であることから、ベイズ識別から投影距離（部分空間法）までの導出について述べる。さらに、入力特徴ベクトルと学習データの平均ベクトルとの距離と、投影距離をそれぞれ分散で正規化して統合した重み付きマハラノビス距離法について述べる。また、本手法と従来の各識別関数（ベイズ識別、疑似ベイズ識別、マハラノビス距離、疑似マハラノビス距離、投影距離）、さらに改良マハラノビス距離に対する話者照合実験の結果を示す。

第4章では、発話内容や時期差による音声の特徴変動を正規化する手法として、話者固有音声空間における動的・静的特徴統合法について述べる。本手法における話者の静的特徴と動的特徴を定義し、話者固有音声空間との静的特徴距離、動的特徴距離を求め、それぞれの距離の統合について述べる。また、本手法と従来の部分空間法に対する話者照合実験の結果を示す。さらに、従来法であるGMMに対する話者照合実験の結果も示す。

第5章では、主成分分析により話者内分散の大きい空間を音韻空間、話者内分散の小さい空間を話者空間とみなして、音韻性と話者性を分離する手法について述べる。さらに、本手法に基づいて話者空間上で話者性を統計的に表現する、話者空間への射影に基づく統計的モデルによる話者認識法について述べる。また、本手法と観測音声空間における従来のGMMによる手法、判別分析によって変換された空間におけるGMMによる手法に対する話者識別および話者照合実験の結果を示す。さらに、他者との相関を考慮したMCE/GPDによる話者空間の最適化法について述べる。また、本手法と観測音声空間における従来のGMM、最適化していない話者空間におけるGMMに対す

る話者照合実験の結果を示す。

第6章では、部分空間により抽出された話者性の応用として、アナウンサー発話の自動抽出と音声認識を統合化したニュース記事の自動分類システムについて述べる。アナウンサー発話の抽出では、特定の話者をあらかじめ学習することなく、入力音声において話者モデルを自動学習し、学習した話者モデルとの話者照合に基づいた話者セグメンテーション法について述べる。また、話者空間への射影に基づく統計的モデルとして、VQを用いた話者照合法について述べる。ニュース記事の分類では、自動抽出したアナウンサー発話の音声認識結果から、 χ^2 法により選択されたキーワードを元にニュース記事を分類する手法について述べる。また、本手法によるアナウンサー発話の抽出、ならびにニュース音声記事の分類実験の結果を示す。

第7章では、部分空間により抽出された話者性の応用として、話者認識と音声認識を統合した話者の人名付与システムについて述べる。インタビュー区間の検出では、アナウンサー発話の抽出における話者セグメンテーション法を用いており、本手法に対して学習データ量を考慮した話者同定における閾値設定手法について述べる。発話の交替を示すキーフレーズと人名の検出では、大語彙連続音声認識とキーワードスポットティングによる手法について述べる。その後、検出したインタビュー区間に、話者の人名を付与する手法について述べる。また、本手法による話者セグメンテーション、ならびに話者の人名付与実験の結果を示す。

第2章

話者認識

2.1 まえがき

本章では、話者認識の原理と従来手法について述べる。話者認識の形態は、話者識別と話者照合に分けることができ、それぞれの処理内容ならびに話者認識における特徴抽出について述べる。また、話者認識の従来手法としては、統計的手法であるVQ歪み、HMM、GMM、空間表現手法である判別分析、部分空間法、その他の手法として識別学習に基づいた手法等について述べる。

2.2 話者認識について

2.2.1 話者識別と話者照合

音声波に含まれる個人性情報を用いて、誰の声であるかを自動的に判定することを話者認識という[10]。主な用途は、電話やインターネットによるバンキングや買い物サービス、クレジットカードコール、コンピュータのリモートアクセスなどである。車や家や特殊な場所への非登録者の侵入防止のための音声キーなども検討されている。

話者認識の形態は、話者識別と話者照合に分けることができる。話者識別とは、入力音声があらかじめ登録されている人の中で誰の声であるかを判定するものである[10]。図2.1に示すように、入力音声と登録されている話者の標準パターンとの類似度を調べて、最も類似度の高い話者を選び、その話者が発話者であると判定する。話者識別の性能は、登録話者の内で本人以外の話者が選択される誤り率で評価する。当然ながら

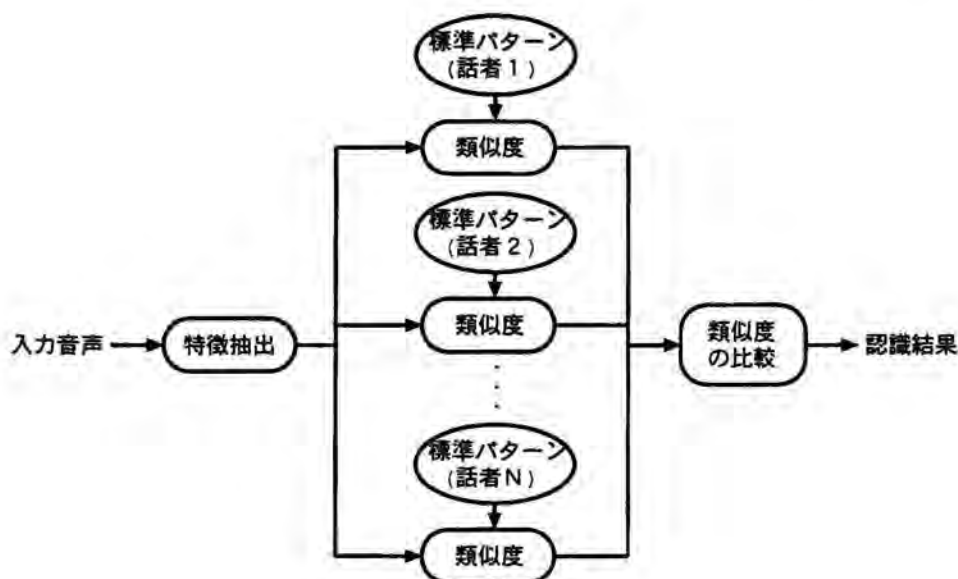


図 2.1: 話者識別

登録話者の数が多くなればそれだけ難しくなるので、話者識別の誤り率は、登録話者の数が増えるにつれて単調に増加する。

話者照合とは、入力音声と同時に自分が誰であるかのIDを入力して、その音声が本当にそのIDに対応する人の発話であるかを判定するものである[10]。図2.2に示すように、本人の標準パターンとの類似度（尤度）が、閾値よりも大きければ本人の発話であると判定し、それ以外の場合は他人の発話であると判定する。しかし、テキストや発声時期の違いによって、入力音声と本人の標準パターンとの尤度の値が大きくばらつき、閾値を安定して設定することが難しい。この問題に対処するために、尤度比尺度[11][12]や事後確率の概念に基づいた尤度の正規化[13]–[15]が行われている。

話者照合における尤度の正規化法として、コホート正規化が有効である[16]。このコホート正規化は、登録話者毎に事前に登録話者の尤度に近い話者をコホート話者として選択し、登録話者に対する尤度をコホート話者に対する尤度で正規化する方法である。しかし、一般的にコホートの話者数を一意に決定することは難しい。

尤度比尺度に基づいた尤度の正規化は、入力音声 x に対する申告話者のモデルの尤度を、申告話者を除いた登録話者のうちの最も高い尤度で割って正規化を行う。式で表すと次のようになる。

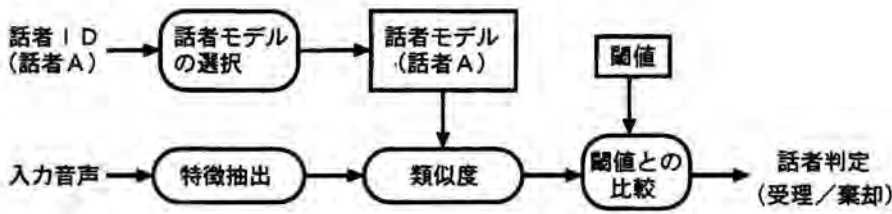


図 2.2: 話者照合

$$norm(S_c) = \frac{S_c}{\max_{i \neq c} S_i} \tag{2.1}$$

ここで、 S_c は入力音声 x に対する本人であると申告された話者のモデルの尤度、 S_i は入力音声 x に対する各登録話者のモデルの尤度、 $norm(S_c)$ は入力音声 x に対する本人であると申告された話者のモデルの正規化した尤度を示す。また、事後確率の概念に基づいた尤度の正規化は、入力音声 x に対する申告話者のモデルの尤度を、全話者のモデルの尤度の総和で割って正規化を行う。式で表すと次のようになる。

$$norm(S_c) = \frac{S_c}{\sum_i S_i} \tag{2.2}$$

話者照合の性能は、本人が棄却される誤り率と詐称者が受理される誤り率で評価される。閾値を設定する場合、閾値を大きくすれば本人棄却率が大きくなり、小さくすれば詐称者受理率が大きくなる。このように両者の誤りにはトレードオフの関係があるので、閾値は2種類の誤りの相対的な重要性（本人棄却率と詐称者受理率のどちらを重要視するか）に従って設定される。特に目的がない場合は、本人棄却率と詐称者受理率が等しくなる値が最適な閾値となる。

2.2.2 話者認識の分類

話者認識の方法はさらに、あらかじめ決まっている言葉（キーワード）を発声するテキスト（発声内容）依存型[17]と、任意の言葉を発声してよいテキスト独立型[18]–[21]に分類できる。テキスト依存型において、キーワードを話者によって違えておけば、そ

の違いも手がかりとして用いることができるので、それだけ認識性能を高めることができる。

話者認識、特に話者照合の応用の多くにおいては、キーワードをそれぞれの人について固定することが可能であるが、どのような言葉を発声しても話者を認識することができれば、応用の分野は一層広がると期待される。犯罪捜査などにおいては、必ずしも同じ言葉どうしを比較できず、テキスト独立型が要求される場合もある。

テキスト依存型、独立型に共通して、詐称者（本人以外の話者）が本人の音声を録音して装置の前で再生すれば、本人になりすまして装置を容易にだますことができるという問題がある。このためテキスト依存型において、数字や決められたキーワードの発声順序を、認識のたびに装置の側から指定するという方法が試みられている。しかしこの方法でも、最近の電子化された録音装置を用いれば、比較的容易に任意の単語系列を再生することができるので、万全ではない。このため、テキスト指定型話者認識法が提案されている[22]–[26]。

テキスト指定型話者認識法では、認識装置を用いるたびに、装置の側から新しいテキストを、文字あるいは合成音声でユーザに提示し、ユーザが本人の声でそのテキストを正しく発声したと判定できる時のみ、話者照合の受理判定をする。その言葉としては、初めてでも比較的発声しやすい、短い文章を用いる。決められたキーワードの発声順序を変えるのではなく、発声する言葉をその都度根本的に変えてしまう。予想のできない言葉が文字表示あるいは音声合成によって指定されるので、テーブルコードによってだますことは極めて難しくなる。

2.2.3 特徴抽出

音声の性質を調べる際、波形を直接扱わず、自己相関関数を含め周波数スペクトルに関連した特徴に変換する。音声の短時間スペクトルは、周波数とともにゆるやかに変化する成分、すなわちスペクトル包絡と、細かく周期的（有声音の場合）または非周期的（無声音の場合）に変化する成分、すなわちスペクトル微細構造の積に分解して考えることができる。スペクトル包絡は、声道の共振・反共振特性のほか、音源のスペクトル概形、口唇・鼻孔などからの放射特性などを含めたスペクトルの大局的特徴である。一方、スペクトル微細構造は、音源の周期性に関連している。このスペクトル包絡は、音韻性を表しているが話者性も表しており、話者認識の特徴量のひとつと

してよく用いられている。ここでは、スペクトル包絡の抽出法としてよく用いられているケプストラム分析について述べる。

音声波形 x_n は、音源波 g_n と声道のインパルス応答 h_n の畳み込みとして表わされる。

$$x_n = \sum_{k=-\infty}^{\infty} g_k \cdot h_{n-k} = g_n * h_n \quad (2.3)$$

ケプストラム $c(\tau)$ は、波形の短時間振幅スペクトル $|X(\lambda)|$ の対数の逆フーリエ変換として定義される [27][28]。音源と声道のインパルス応答のフーリエ変換を $G(\lambda)$ および $H(\lambda)$ で表わすと、線形分離等価回路モデルにより、次の関係が得られる。

$$X(\lambda) = G(\lambda)H(\lambda) \quad (2.4)$$

この対数をとると、

$$\log |X(\lambda)| = \log |G(\lambda)| + \log |H(\lambda)| \quad (2.5)$$

となる。このようにして、声道のスペクトル包絡と音源のスペクトル微細構造が和の形で表わされる。次にこれをフーリエ逆変換すると、

$$c(\tau) = \mathcal{F}^{-1} \log |X(\omega)| = \mathcal{F}^{-1} \log |G(\omega)| + \mathcal{F}^{-1} \log |H(\omega)| \quad (2.6)$$

となり、これがケプストラムである。離散フーリエ変換 (DFT) で求めると、

$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} \log |X(k)| e^{\frac{j2\pi kn}{N}} \quad (2.7)$$

となる。この場合、波形の標本化の場合と同様に折り返しが生ずるので、変換の基底の大きさ N を十分大きくとる必要がある。

ケプストラムという言葉は、スペクトルを逆変換するという意味から、spectrum をもじって作った造語であり、その変数は frequency をもじってケフレンシー (quefrensy) と呼ばれる。ケフレンシーの次元は、周波数領域からの逆変換であるから時間になる。

式 (2.5) の右辺第1項はスペクトルの微細構造に対応し、第2項はスペクトル包絡に対応する。第1項は高ケフレンシー部のピークとなり、第2項は0から2～4ms程度の低ケフレンシー部に集中する。したがって、高ケフレンシー部のピーク的位置によって音源の基本周期が求まり、低ケフレンシー部のみを用いてフーリエ変換すれば対数スペクトル包絡が、それを指数変換すればスペクトル包絡が求まる。この手順を図2.3

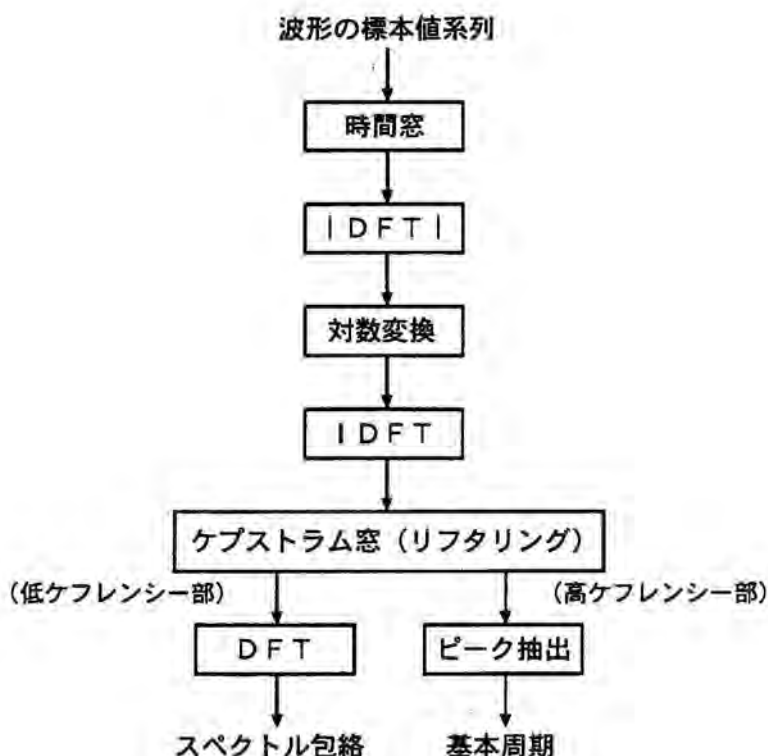


図 2.3: 基本周期とスペクトル包絡の抽出手順

に示す。求まるスペクトル包絡のなめらかさの度合いは、低ケフレンシー部のどれだけの成分を用いるかによって決まる。種々のケフレンシー成分を分離する操作は、フィルタ (filter) をもじってリフタ (lifter) と呼ばれる。

このように、ケプストラム分析は、音声波形のスペクトルから、声道スペクトルの共振によるスペクトル包絡と、音源の波形の周期性によるスペクトルの微細構造との分離を可能にするノンパラメトリックな分析手法である。

次に、話者認識における特徴抽出に関する研究について述べる。

宮島らは、特徴抽出において2次オールパスシステムにおける周波数伸縮パラメータと、話者モデルのパラメータとを最小分類誤り基準で同時最適化する手法を提案している [29]。

早川らは、線形予測残差スペクトルの調波構造に個人性情報が含まれていることを示し、調波構造を定量化するパラメータとして帯域内強度差を用いた話者認識手法を提案している [30]。また、線形予測残差を用いたその他の手法としては、高次の項を含

めた線形予測残差のケプストラムを用いた話者照合法 [31], 線形予測残差とピッチを統合した話者認識法が提案されている [32].

2.3 従来手法

2.3.1 VQ歪み

音声波形の瞬時値のデジタル化はスカラ量子化と呼ばれる。これに対して、複数の値の組（ベクトル）をまとめて一つの符号で表現する量子化方法を、ベクトル量子化 (VQ: Vector Quantization) [33] と呼ぶ。これは、波形の時系列や音声の特徴パラメータ (LPC パラメータなどによるベクトル表現) に適用され、ベクトルの低ビット表現化が可能であることから、主にデータ量の削減化法として用いられてきた。また、音声波形やパラメータの組には連続性や分布の偏りがあり、ベクトル量子化ではそれを利用するため、音声を少ない情報量で表現できる。例として、ベクトル量子化法による LPC パラメータの高能率符号化法について考えてみる。音声波形は、LPC 分析により短時間区間ごとに LPC パラメータに変換され、ベクトルの時系列 $x_1, x_2, \dots, x_N$ として表現される。各ベクトル x_t はあらかじめ求められている典型的なベクトル集合 $\{y_1, y_2, \dots, y_K\}$ (コードブック) の一つの要素であるベクトル y_k (コードベクトル) に置換される。式で表すと次のようになる。

$$\hat{k} = \arg \min_k \text{dist}(x_t, y_k) \quad (2.8)$$

もし、コードブックサイズが $K = 2^k$ であれば、各ベクトル x_t は k ビットで表現されることになる。当然、 K を大きくすれば量子化誤差（歪み）は小さくなるが、 K を固定してもコードベクトルの選び方によって量子化誤差は変化する。母集団の分布の偏りを利用して、ベクトル量子化による歪みがサンプル全体として最小になるコードブックを求めることが必要になる。

多数の訓練サンプルから、これらのサンプル集合を最もよく近似する K 個のコードベクトルを求める方法は、パターン認識の分野で研究されているクラスタリング手法にほかならない。音声認識におけるマルチテンプレート作成のためのクラスタリング技法としては、 k -means 法がよく知られている。代表的なコードブック作成アルゴリズム

ムに Linde らのアルゴリズム (LBG アルゴリズム) [33] がある。ここで、コードブックの代表ベクトル (セントロイドベクトル) x を次式で定義する。

$$\hat{x} = C(x_1, x_2, \dots, x_N) = \arg \min_x \frac{1}{N} \sum_{t=1}^N \text{dist}(x, x_t) \quad (2.9)$$

VQ 歪みによる話者認識法は、まず各登録話者が発声した学習データである任意の文章の短時間スペクトルの集合を、クラスタ化して各クラスタの重心ベクトル (セントロイド) を求め、これから符号帳 (コードブック) を構成する。話者識別では、入力音声の特徴ベクトルの集合 $\{x_t\} (1 \leq t \leq N)$ と、各登録話者 i に対して最も近いセントロイド c_k^i との距離を求める。これは、コードブックでベクトル量子化した際の量子化歪みとみなすことができる。ただし、ベクトル間の距離 $\text{dist}(x_t, c_k^i)$ としてユークリッド距離を用いている。次式により、各登録話者 i との最も近いセントロイド c_k^i との距離のうち、最小距離をもつ登録話者 s を識別結果とする。

$$s = \arg \min_i \frac{1}{N} \sum_t \text{dist}(x_t, c_k^i) \quad (2.10)$$

話者照合では、入力音声の特徴ベクトルの集合 $\{x_t\} (1 \leq t \leq N)$ と、本人 s と申告された登録話者に最も近いセントロイド c_k^s との距離を求める。この距離を次式のように、特徴ベクトルの集合 x_t と各登録話者に最も近いセントロイド c_k^i との距離の総和で正規化する。

$$\hat{\text{dist}}(x_t, c_k^s) = \frac{\text{dist}(x_t, c_k^s)}{\sum_i \text{dist}(x_t, c_k^i)} \quad (2.11)$$

この正規化した距離が閾値より小さければ本人の音声として受理し、それ以外は棄却する。

次に、VQ 歪みを用いた話者認識に関する研究について述べる。

松井らは、テキスト独立型話者認識において、学習データ量が十分でない場合には、VQ 歪みの方が連続 HMM よりも認識精度が高いことを示している [34]。

特徴量に関する研究としては、音源と声道に関する特徴量を、発話内容に依存せず、時期的な変動に対して頑健に組み合わせる方法 [35]、電話回線や異なるマイクロホンで収録した音声に対して、LPC と PLP といった特徴量を用いた話者認識 [36]、線形と非線形予測モデルからの残差信号を VQ でモデル化する方法が提案されている [37]。

モデル化に関する研究としては、ランダムに特徴ベクトルを選択してヒューリスティックに識別誤り最小化基準でコードブックを再学習する方法が提案されている[38][39]。

2.3.2 HMM

HMM (Hidden Markov Model: 隠れマルコフモデル) [10][33] は、出力シンボルによって一意に状態遷移先が決まらないという意味で、非決定性有限状態オートマトンとして定義される。また、HMMは複数の状態を持ち、それら相互の状態間の遷移確率により定義される単純マルコフモデルに、各状態からの複数シンボルの出力確率を加えたモデルである。観測できるのは出力されるシンボル系列だけであり、モデルの状態を直接観測することはできないことから隠れマルコフモデルとよばれている。

HMMは、単語や音素を確率状態遷移機械（マルコフモデル）で表現し、スペクトル時系列の統計的変動をモデルのパラメータに反映させることができる特徴があるが、逆にモデルパラメータを決定するための学習処理がやや複雑になる。

音声認識に用いられるHMMは、left-to-right型で一つの初期状態と一つの最終状態がある構造が多く、図2.4は最もよく用いられるベイキスモデルと呼ばれる型の例である。図の a_{ij} は、状態 q_i から状態 q_j への状態遷移確率を表し、状態数を S とすると $S \times S$ の行列で表現できる。各状態 q_i の初期状態を π_i で表し、最終状態の集合を F で表す。 $b_{ij}(y)$ は、状態 q_i から状態 q_j への遷移で、スペクトルパターン y が観測（出力）される観測（出力）確率を表し、 $\{b_{ij}(y)\}$ を出力確率行列と呼ぶ。出現するスペクトルパターンに関しては、連続値として表す連続分布モデル（連続HMM）と、有限個のシンボルの組み合わせで表現する離散分布モデル（離散HMM）がある。

次に、HMMにおける話者認識アルゴリズムについて述べる。 $y = \{y_1, y_2, \dots, y_T\}$ を観測（出力）系列とする。このとき、各HMMモデルによって y が生起する確率（尤度） $P(y|S)$ （ここで S はHMMによって表される話者に対応）を求め、最大確率（最大尤度）を与えるモデルを選んで、これを認識結果とする。 $q = q_{i0}, q_{i1}, \dots, q_{iT}$ を状態遷移系列（ただし $q_{iT} \in F$ ）とすれば、

$$P(y|S) = \sum_{i_0, i_1, \dots, i_T} P(y|q, S) \cdot P(q|S) \quad (2.12)$$

と書ける。一般に $P(y|S)$ の値は、各遷移での確率値を対数化することで出現確率（尤度）を加算化したビタビアルゴリズム[40]によって求められる。

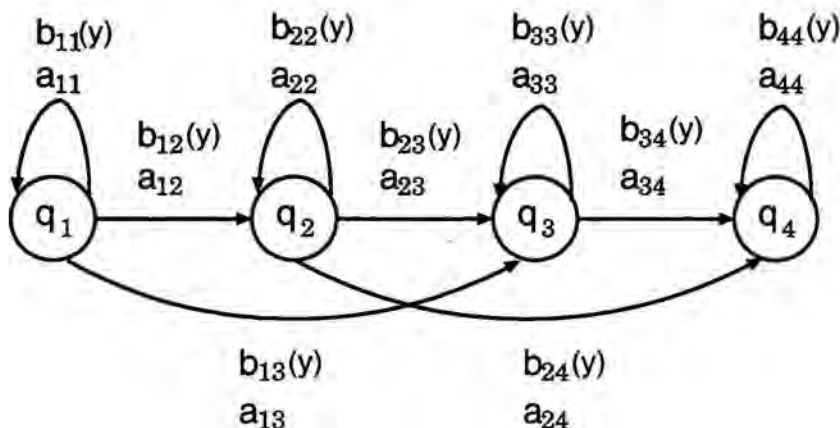


図 2.4: ベイキス型HMMの例

次に、話者認識に用いられる連続HMMにおけるパラメータ推定法について述べる。連続分布モデルで出現確率を表す方法としては、単一ガウス分布（正規分布）または混合ガウス分布が用いられるが、ここでは混合ガウス分布について述べる。学習用音声の観測（出力）系列を $y = \{y_1, y_2, \dots, y_T\}$ と表すと、 y_t は多次元のベクトルである。混合ガウス分布の出現確率は、次のように表される（ガウス分布の数を M とする）。

$$b_{ij}(y) = \sum_{m=1}^M \lambda_{ijm} b_{ijm}(y) \quad (2.13)$$

ここで、

$$\sum_{m=1}^M \lambda_{ijm} = 1 \quad (2.14)$$

$$\int b_{ijm}(y) dy = 1 \quad (2.15)$$

である。このときの初期確率 π_i と遷移確率 a_{ij} の推定式は、

$$\hat{\pi}_i = \frac{\sum_j \gamma(i, j, 1)}{\sum_i \sum_j \gamma(i, j, 1)} \quad (2.16)$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^T \alpha(i, t-1) \cdot a_{ij} \cdot b_{ij}(y_t) \cdot \beta(j, t)}{\sum_t \sum_j \alpha(i, t) \cdot \beta(j, t)} = \frac{\sum_t \gamma(i, j, t)}{\sum_t \sum_j \gamma(i, j, t)} \quad (2.17)$$

で表される。

$$\hat{\lambda}_{ijm} = \frac{\sum_{t=1}^T \gamma(i, j, t, m)}{\sum_{t=1}^T \gamma(i, j, t)} \quad (2.18)$$

$$\hat{\mu}_{ijm} = \frac{\sum_{t=1}^T \gamma(i, j, t, m) y_t}{\sum_{t=1}^T \gamma(i, j, t, m)} \quad (2.19)$$

$$\hat{\Sigma}_{ijm} = \frac{\sum_{t=1}^T \gamma(i, j, t, m) (y_t - \mu_{ijm})(y_t - \mu_{ijm})^t}{\sum_{t=1}^T \gamma(i, j, t, m)} \quad (2.20)$$

ただし、

$$\gamma(i, j, t, m) = \alpha(i, t-1) \cdot a_{ij} \cdot \lambda_{ijm} \cdot b_{ijm}(y_t) \cdot \beta(j, t) \quad (2.21)$$

で、 m 番目の分布関数の遷移 $q_i \rightarrow q_j$ の確率を表している。 $\alpha(i, t)$ は時刻 t で符号ベクトル y_t を出力して状態 q_i にある確率、 $\beta(i, t)$ は時刻 t に状態 q_i にあって、以後符号ベクトル y_{t+1}^T を出力する確率を表している。これらのパラメータは、期待値の計算と尤度最大化の計算を交互に繰り返すことで、局所的最大値に収束させる EM アルゴリズムによって推定する [40]。

HMM による話者認識は、各話者の任意の発声内容の音声を図 2.5 に示すようなエルゴディックな HMM で表現し、このモデルから入力音声のスペクトル系列が出現する尤度を計算して、話者の認識を行う。エルゴディックな HMM は、任意の状態からスタートして、任意の状態に遷移し、任意の状態を終端とすることができる。HMM の学習は、各話者毎に学習用の文章や多数の単語を発声した音声を用いて自動的に行なうこ

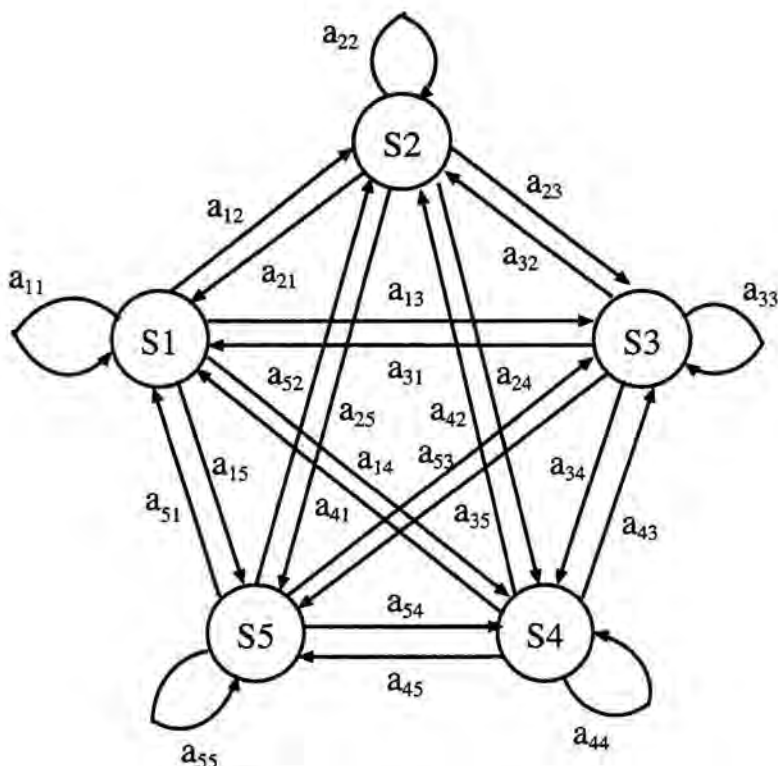


図 2.5: エルゴディック HMM の例

とができるが、図 2.5 に示す五つの状態がほぼ、鼻音、摩擦音、破裂音、および 2 種類の有声音に対応するように設定されることが確かめられている。状態間の遷移は発声内容によって変化するので、テキスト独立の場合は結果的に遷移確率は評価に含まれず、出力確率のみを用いて計算した尤度によって、話者の判定を行う場合と同じになることが多い。この場合は結局、全状態でのガウス分布の総数さえ同じなら、いくつの状態に分けても同じ結果になり、1 状態の HMM、すなわち単なる混合ガウス分布による方法でも同じ結果が得られる。

HMM による話者認識法は、まず各登録話者の学習用音声データから抽出した特徴ベクトルの時間変動を、状態間の遷移と特徴ベクトルの出力確率によって表す。各状態における出力確率を混合ガウス分布で近似して、特徴ベクトルの時系列を各話者ごとに、HMM によってモデル化する。話者識別では、入力音声の特徴ベクトルの集合 $\{x_t\} (1 \leq t \leq N)$ と、各登録話者の HMM との尤度 $P(x_t|S_i)$ を求める。次式により、各登録話者 i の HMM との尤度 $P(x_t|S_i)$ から、最大尤度をもつ登録話者 s を識別結果とする。

$$s = \arg \max_i \frac{1}{N} \sum_t P(x_t | S_i) \quad (2.22)$$

話者照合では、入力音声の特徴ベクトルの集合 $\{x_t\} (1 \leq t \leq N)$ と、本人 s と申告された登録話者の HMM との尤度 $P(x_t | S_s)$ を求める。この尤度を次式のように、各登録話者の HMM に対する特徴ベクトルの集合 $\{x_t\}$ の尤度を求め、この総和で正規化する。

$$\hat{P}(x_t | S_s) = \frac{P(x_t | S_s)}{\sum_i P(x_t | S_i)} \quad (2.23)$$

正規化した尤度が閾値より大きければ本人の音声として受理し、それ以外は棄却する。次に、HMM を用いた話者認識に関する研究について述べる。

特徴量に関する研究としては、学習と認識時における音響環境の変動に対して、スペクトル減算法を用いた頑健な話者認識法 [41][42]、認識データを線形変換することで、学習データと同じ音響環境に変換する方法 [43]、周波数領域を帯域分割し、分割された帯域毎に抽出したケプストラムを用いた話者照合法 [44] が提案されている。

モデル化に関する研究としては、HMM のパラメータの推定において EM アルゴリズム、Segmental K-Means、MAP 推定といった手法による比較 [45]、HMM の連続関数を多項式関数で表現することで初期モデルのパラメータを頑健に推定する方法 [46] が提案されている。

その他に関する研究としては、磯部らは、話者照合における尤度正規化法として、コホートモデルを HMM の音素、状態、分布が表す局所的な音響情報に基づいて選択される近傍話者の HMM から合成する方法を提案している [47]。また、コンテキスト依存型 HMM を用いた数字列指定型の話者照合 [48]、入力音声を音声認識して得られた言語情報によるテキスト指定型話者照合法 [49]、ビタビアルゴリズムにおけるビームサーチの探索空間を部分空間により狭めることで、デコーダの処理を高速化する方法が提案されている [50]。

2.3.3 GMM

近年、話者認識の手法としては、音声から抽出した特徴を統計的にモデル化する GMM (Gaussian Mixture Model) がよく用いられている。GMM は、2.3.2 節で述べた 1 状態の HMM、すなわち混合ガウス分布により音声データをモデル化する方法である。

音声データの n 次元特徴ベクトルを x_t 、話者 s のモデルを λ_s とすると、話者モデル λ_s の混合分布は、次のように表される。

$$b(x_t|\lambda_s) = \sum_{m=1}^M w_{sm} N(x_t|\mu_{sm}, \Sigma_{sm}) \quad (2.24)$$

式(2.24)に示すように、GMMは n 次元 m 混合のガウス分布 $N(x_t|\mu_{sm}, \Sigma_{sm})$ を重み w_{sm} で線形結合した確率モデルとなる。

$$N(x_t|\mu_{sm}, \Sigma_{sm}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x_t - \mu_{sm})^t \Sigma_{sm}^{-1} (x_t - \mu_{sm})\right\} \quad (2.25)$$

ここで、 μ_{sm} は話者モデル λ_s の平均ベクトル、 Σ_{sm} は話者モデル λ_s の分散共分散行列を表している。また、混合分布の重み w_{sm} は、

$$\sum_{m=1}^M w_{sm} = 1 \quad (2.26)$$

である。

GMMによる話者認識法は、まず各登録話者の学習用の音声データから抽出した特徴ベクトルの時系列から、EMアルゴリズム[40]によって尤度が最大となるように、話者モデル λ_s のパラメータを推定する。話者識別では、入力音声の特徴ベクトルの集合 $\{x_t\}(1 \leq t \leq N)$ と、各登録話者 i のGMMとの対数尤度 $\log P(x_t|\lambda_i)$ を求める。次式により、各登録話者 i のGMMとの対数尤度のうち、最大対数尤度をもつ登録話者 s を識別結果とする。

$$s = \arg \max_i \frac{1}{N} \sum_t \log P(x_t|\lambda_i) \quad (2.27)$$

話者照合では、入力音声の特徴ベクトルの集合 $\{x_t\}(1 \leq t \leq N)$ と、申告話者 s のGMMとの対数尤度 $\log P(x_t|\lambda_s)$ を求める。この対数尤度に対して次式のように、尤度比尺度に基づく正規化を行う。

$$\log \hat{P}(x_t|\lambda_s) = \frac{1}{N} \sum_t \log P(x_t|\lambda_s) - \max_{i \neq s} \frac{1}{N} \sum_t \log P(x_t|\lambda_i) \quad (2.28)$$

これは、申告話者 s を除いた全登録話者の GMM との最大尤度を引くことによって正規化を行う方法である。この正規化した対数尤度が閾値より大きければ、本人の音声であると判定する。

次に、GMM を用いた話者認識に関する研究について述べる。

Reynolds らは、GMM を用いたテキスト独立型話者識別法を提案している [51]。

特徴量に関する研究としては、学習と認識時の音響環境の変動に対して、背景話者モデルを用いた尤度正規化による頑健な話者認識法 [52]、話者モデルの分散の補正による頑健な話者認識法 [53]、話者性が多く存在する周波数帯域を RASTA により強調した話者照合法 [54] が提案されている。

モデル化に関する研究としては、音声中の動的な話者性を考慮する方法として、フレーム間の連鎖確率を導入した GMM による話者識別法 [55]、スペクトルとピッチをモデル化する方法として、多空間上の確率分布に基づく GMM による話者認識法が提案されている [56]。

その他の研究としては、GMM を用いた話者照合における尤度の正規化法の比較 [57] について検討されている。

2.3.4 判別分析

ここでは、特徴空間の線形変換法として、多クラス ($c > 2$) に対する線形判別法について述べる [58]。多クラスに対する線形判別法では、 n 次元の特徴空間は変換によって、 $k (= c - 1)$ 次元の空間へ次元が削減される。

観測音声空間で観測される話者 i の学習音声データを n 次元特徴ベクトルの集合 $\{x_t^{(i)}\}$ ($t = 1, 2, \dots, N^{(i)}$) とする。この学習データの平均ベクトルを $\mu^{(i)}$ 、分散共分散行列を $\Sigma^{(i)}$ とする。話者 i の事前確率を $P(\omega^{(i)})$ 、全話者の平均ベクトルを μ とすると、クラス内共分散行列 Σ_W およびクラス間共分散行列 Σ_B は、次のように定義される。

$$\begin{aligned} \Sigma_W &= \sum_{i=1}^c P(\omega^{(i)}) \Sigma^{(i)} \\ &= \sum_{i=1}^c \frac{P(\omega^{(i)})}{N^{(i)}} \sum_{t=1}^{N^{(i)}} (x_t^{(i)} - \mu^{(i)})(x_t^{(i)} - \mu^{(i)})^T \end{aligned} \quad (2.29)$$

$$\Sigma_B = \sum_{i=1}^c P(\omega^{(i)}) (\mu^{(i)} - \mu)(\mu^{(i)} - \mu)^T \quad (2.30)$$

ここで、各話者の事前確率 $P(\omega^{(i)})$ は、各話者の学習データ数 $N^{(i)}$ がその話者の事前確率を反映していると仮定して、 $P(\omega^{(i)}) = N^{(i)}/N$ とした。また、 $N^{(i)}$ は話者 i の学習データ数、 N は全話者の学習データの総数を表す。

線形変換した空間上でのクラス内共分散行列 $\tilde{\Sigma}_W$ およびクラス間共分散行列 $\tilde{\Sigma}_B$ は、変換行列を A とすると、

$$\tilde{\Sigma}_W = A^T \Sigma_W A \quad (2.31)$$

$$\tilde{\Sigma}_B = A^T \Sigma_B A \quad (2.32)$$

となる。線形変換後の空間上でのクラス間の分離度を表す Fisher の評価基準 $J_\Sigma(A)$ は、クラス内分散・クラス間分散比として次のように定義される。

$$J_\Sigma(A) = \frac{|\tilde{\Sigma}_B|}{|\tilde{\Sigma}_W|} = \frac{|A^T \Sigma_B A|}{|A^T \Sigma_W A|} \quad (2.33)$$

この評価基準 $J_\Sigma(A)$ を最大にする A を求める問題は、 $\tilde{\Sigma}_W = A^T \Sigma_W A = I$ という制約条件のもとで、 $|A^T \Sigma_B A|$ を最大にすることとなり、次の固有値問題

$$\Sigma_B A = \Sigma_W A \Lambda \quad (2.34)$$

に帰着する。ここで、 Λ は、固有値 $\lambda_d (d = 1, \dots, k, \dots, n)$ を対角成分にもつ対角行列である。 Σ_W が正則行列であれば、変換行列 A は、 $\Sigma_W^{-1} \Sigma_B$ を固有値分解することで得られる。したがって、固有値が大きい上位 k 個の固有ベクトルが、線形変換後の空間を張る基底となる。

次に、判別分析による空間変換に基づく話者認識法について述べる。観測音声空間で観測される話者 s の学習音声データを n 次元特徴ベクトルの集合 $\{x_t^{(s)}\}$ とする。全話者に対して線形判別分析を行い、式 (2.34) の固有値分解から得られた固有ベクトルを $a_d (d = 1, \dots, k, \dots, c-1)$ とし、この固有ベクトルを列ベクトルとする変換行列を A とすると、この a_i で張られる部分空間は、全話者間の分離度が大きい空間を構成している。この部分空間に学習データ $\{x_t^{(s)}\}$ を次式のように射影する。

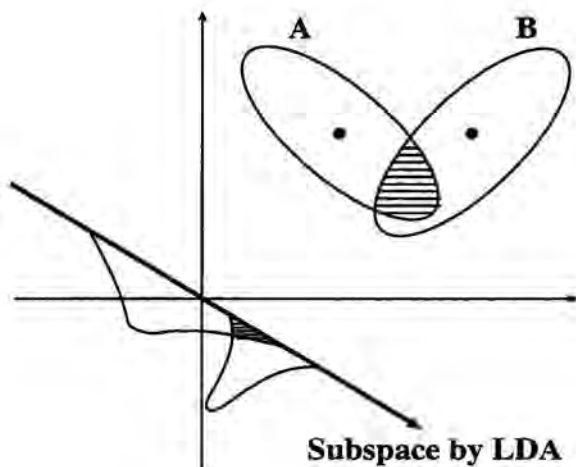


図 2.6: 判別分析による空間変換

$$\hat{x}_t^{(s)} = A^T x_t^{(s)} \quad (2.35)$$

式(2.35)により、各話者の学習データを射影し、射影した部分空間でGMMにより話者を学習する。

判別分析による空間変換の概念図を図2.6に示す。図2.6に示すように、判別分析(LDA: Linear Discriminant Analysis)により得られた部分空間へ射影することで、話者間の分離度が大きくなる。

話者識別では、入力特徴ベクトルの集合 $\{x_t\}$ を式(2.35)により部分空間に射影して $\{\hat{x}_t\}$ を求める。次に、各登録話者 i のGMMとの対数尤度 $\log P(\hat{x}_t | \lambda^{(i)})$ を求める。次式により、各登録話者 i のGMMとの対数尤度のうち、最大対数尤度をもつ登録話者 s を識別結果とする。

$$s = \arg \max_i \frac{1}{N} \sum_{t=1}^N \log P(\hat{x}_t | \lambda^{(i)}) \quad (2.36)$$

話者照合では、入力特徴ベクトルの集合 $\{x_t\}$ を式(2.35)により部分空間に射影して $\{\hat{x}_t\}$ を求める。次に、申告話者 s のGMMとの対数尤度 $\log P(\hat{x}_t | \lambda^{(s)})$ を求める。この対数尤度に対しては、式(2.37)に示す尤度比尺度に基づく正規化を行う。

$$\log \hat{P}(\hat{x}_t | \lambda^{(s)}) = \frac{1}{N} \sum_{t=1}^N \log P(\hat{x}_t | \lambda^{(s)}) - \max_{i \neq s} \frac{1}{N} \sum_{t=1}^N \log P(\hat{x}_t | \lambda^{(i)}) \quad (2.37)$$

これは、申告話者 s を除いた全登録話者の GMM との最大尤度を引くことによって正規化を行う方法である。この正規化した対数尤度が閾値より大きければ、本人の音声であると判定する。

次に、判別分析を用いた話者認識に関する研究について述べる。対数スペクトルに対して LDA により得られる話者の分離度が高い特徴量，ならびに LSP といった特徴量による話者認識の比較 [59] について検討されている。また，不特定話者モデルから適応データを用いて MLLR により特定話者モデルを学習する際に，特徴空間を LDA により線形変換する方法が提案されている [60]。

2.3.5 部分空間法

部分空間法 [61][62] は，本来線形特徴抽出とデータ圧縮の考え方を発展させたものである。ベクトルの形で与えられる実データの多くは，比較的少数の主成分方向にそのエネルギーの大半が集中しているので，各クラスに対してそれぞれ異なる線形のデータ圧縮を行うことができる。一度このようにデータを圧縮してしまえば，新しい入力信号に対する実際の識別は，クラスのこのように圧縮表現された特徴と単に比較するだけで済むことになり，複雑な識別方法は必要でなくなる。また，多次元の観測音声空間において，特徴ベクトルは低次元の部分空間に偏って分布しており，識別を行う際には，データが分布している部分空間だけに着目すればよいことになる。つまり，部分空間法は，特徴抽出と識別を分離することなく，特徴空間の線形変換そのものを利用して識別するという手法である。識別方法は，各クラス毎にそのクラスを表現する低次元の部分空間を用意し，未知パターンがどの部分空間で最もよく近似表現できるかを比較することにより行う。

次に，部分空間法の決定規則について述べる。対象とする C 個のクラスを， $\omega^{(1)}, \dots, \omega^{(C)}$ で表す。各クラスは， n 次元パターン空間内の部分空間によって表現される。これらの部分空間を $L^{(1)}, \dots, L^{(C)}$ で表し，その次元を $k^{(i)} = \dim(L^{(i)})$, $i = 1, \dots, C$ とする。各部分空間は， $\mathbf{R}^n$ 内の $k^{(i)}$ 個の正規直交基底ベクトル $\varphi_d^{(i)}$, $d = 1, \dots, k^{(i)}$ により張られる。基底は一意ではないが，射影行列

$$P^{(i)} = \sum_{d=1}^{k^{(i)}} \varphi_d^{(i)} \varphi_d^{(i)T} \in R^{n \times n} \quad (2.38)$$

は一意に決まり、部分空間 $L^{(i)}$ と等価になる。

部分空間法の決定規則は、次のように定義される。もしすべての $j \neq i$ に対して、

$$x^T P^{(i)} x > x^T P^{(j)} x \quad (2.39)$$

ならば、入力特徴ベクトル x をクラス $\omega^{(i)}$ に識別する。行列 $P^{(i)}$ の性質により、二次形式 $x^T P^{(i)} x$ は次のように表現できる。

$$x^T P^{(i)} x = x^T P^{(i)T} P^{(i)} x = \|P^{(i)} x\|^2 \quad (2.40)$$

ここで、 $P^{(i)} x$ は、入力特徴ベクトル x の部分空間 $L^{(i)}$ への正射影である。ゆえに、式 (2.39) の決定規則では、各入力特徴ベクトル x をユークリッドノルムの意味で、最大の射影成分をもつ部分空間のクラスに識別することになる。

次に、部分空間法に基づく話者性の抽出について述べる。観測音声空間で観測される話者 i の音声データを $X^{(i)}$ とする。音声データは、話者 i の音声波形を時刻 t で短時間スペクトル分析した n 次元特徴ベクトルの集合 $\{x_t^{(i)}\}$ である。音声データ $X^{(i)}$ は、特徴ベクトル $x_t^{(i)}$ ($t = 1, 2, \dots, N$) を行ベクトルとする行列である。

この音声データの行列 $X^{(i)}$ は一般に正方行列ではないので、特異値分解によって次のように因子分解される。

$$X^{(i)} = U^{(i)} \Lambda^{(i)} V^{(i)T} \quad (2.41)$$

ここで、 $U^{(i)}$ と $V^{(i)}$ は、それぞれ行列 $X^{(i)} X^{(i)T}$ と行列 $X^{(i)T} X^{(i)}$ の固有ベクトルを列ベクトルとする行列である。 $\Lambda^{(i)}$ は、 $X^{(i)}$ の特異値を対角成分に持つ対角行列である。したがって、 $X^{(i)} X^{(i)T}$ と $X^{(i)T} X^{(i)}$ は、次のような固有値分解として表現される。

$$X^{(i)} X^{(i)T} = U^{(i)} \Lambda_1^{(i)} U^{(i)T} \quad (2.42)$$

$$X^{(i)T} X^{(i)} = V^{(i)} \Lambda_2^{(i)} V^{(i)T} \quad (2.43)$$

$X^{(i)T} X^{(i)}$ は、話者 i の特徴ベクトル $x_t^{(i)}$ の相関行列であり、相関行列の固有ベクトルは特徴ベクトル集合 $\{x_t^{(i)}\}$ の正規直交基底ベクトルとなることが示されている [61]。

さらに、この正規直交基底ベクトルは、特徴ベクトルの集合 $\{x_t^{(i)}\}$ との距離の総和を最小にすることも示されている [63][64].

式 (2.41) の特異値分解を解釈すると、相関行列 $X^{(i)T} X^{(i)}$ の固有ベクトルは、話者 i の音声データを表現する固有の空間（部分空間）を形成している。本論文では、この部分空間を話者固有音声空間と呼ぶ。したがって、 $V^{(i)}$ は、観測音声空間の座標系で表現された話者 i の話者固有音声空間を表しており、観測音声空間から話者固有音声空間への変換行列となっている。また、 $\Lambda^{(i)}$ の中から大きな特異値のみを k 個選ぶことにより、 $V^{(i)}$ の k 本の固有ベクトルは、 k 次元の話者 i の話者固有音声空間を張る正規直交基底ベクトルとなっている。

ここで述べた部分空間法には、CLAFIC 法と主成分分析法がある。次に、これらの手法について述べる。

CLAFIC 法 (CLAss-Featuring Information Compression) は、平均二乗誤差最小基準に基づいて学習データとの距離が最小になる部分空間を求める手法である。図 2.7 に CLAFIC 法の概念図を示す。

まず、CLAFIC 法による話者認識における学習法について述べる。観測音声空間で観測される話者 i の学習音声データを n 次元特徴ベクトルの集合 $\{x_t^{(i)}\}$ ($t = 1, 2, \dots, N$) とする。この学習データから、相関行列 $R^{(i)}$ を次式で求める。

$$R^{(i)} = \frac{1}{N} \sum_{t=1}^N x_t x_t^T \quad (2.44)$$

この相関行列 $R^{(i)}$ を固有値分解すると、次式のようになる。

$$R^{(i)} = V^{(i)} \Lambda^{(i)} V^{(i)T} \quad (2.45)$$

ここで、 $\Lambda^{(i)}$ は、相関行列 $R^{(i)}$ の固有値 $\lambda_d^{(i)}$ ($d = 1, \dots, k, \dots, n$) を対角成分にもつ対角行列である。また、 $V^{(i)}$ は、相関行列 $R^{(i)}$ の固有ベクトル $\varphi_d^{(i)}$ ($d = 1, \dots, k, \dots, n$) を列ベクトルとする行列である。

固有ベクトルの集合 $\varphi_d^{(i)}$ は正規直交基底ベクトルとして、話者 i の音声データを表現する話者固有音声空間を形成している。話者 i の話者固有音声空間を構成する場合、固有値 $\lambda_d^{(i)}$ が大きい上位 k 個の固有ベクトル $\varphi_d^{(i)}$ を正規直交基底ベクトルとして、 k 次元話者固有音声空間を構成する。したがって、 n 次元観測音声空間における話者 i のパターン分布を k 次元話者固有音声空間で近似することができる。

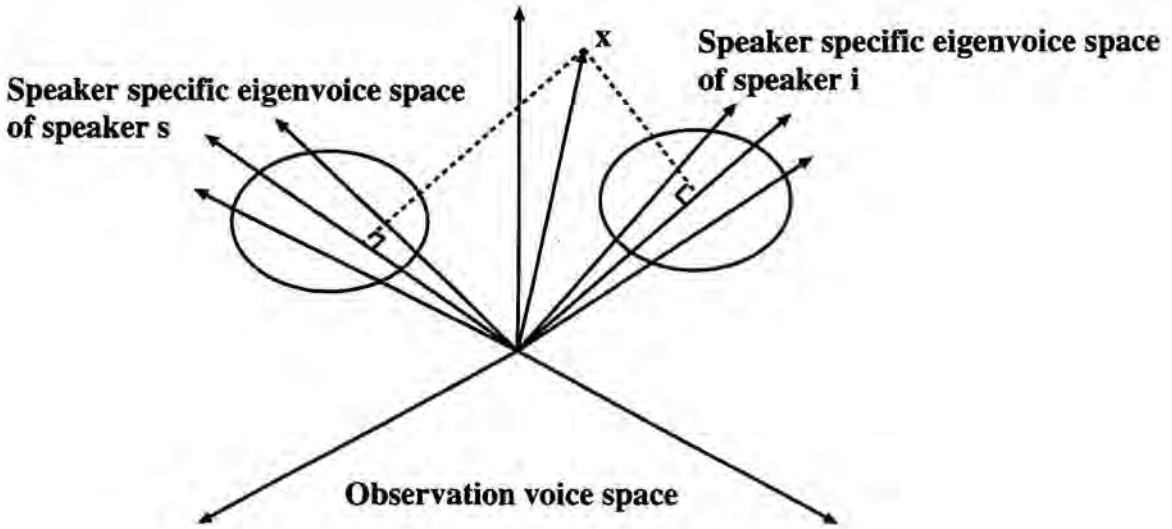


図 2.7: CLAFIC 法

次に, CLAFIC 法による話者認識について述べる. 話者識別では, 入力特徴ベクトルの集合 $\{x_t\}$ を各登録話者 i の話者固有音声空間に射影し, 次式により平均射影量を求める.

$$\begin{aligned}
 PL^{(i)} &= \frac{1}{N} \sum_{t=1}^N x_t^T P^{(i)} x_t \\
 &= \frac{1}{N} \sum_{t=1}^N \sum_{d=1}^k (x_t^T \varphi_d^{(i)})^2
 \end{aligned} \tag{2.46}$$

ここで, N は入力音声の総フレーム数である. 式 (2.47) により, 各登録話者 i の話者固有音声空間に対する射影量のうち, 最大射影量をもつ登録話者 s を識別結果とする.

$$s = \arg \max_i PL^{(i)} \tag{2.47}$$

話者照合では, 入力特徴ベクトルの集合 $\{x_t\}$ を本人 s であると申告された話者固有音声空間に射影し, 式 (2.46) により平均射影量を求める. 尤度比尺度に基づく正規化法の定式化を参照し, 求めた射影量に対して式 (2.48) に示す射影量の比に基づく正規化を行う.

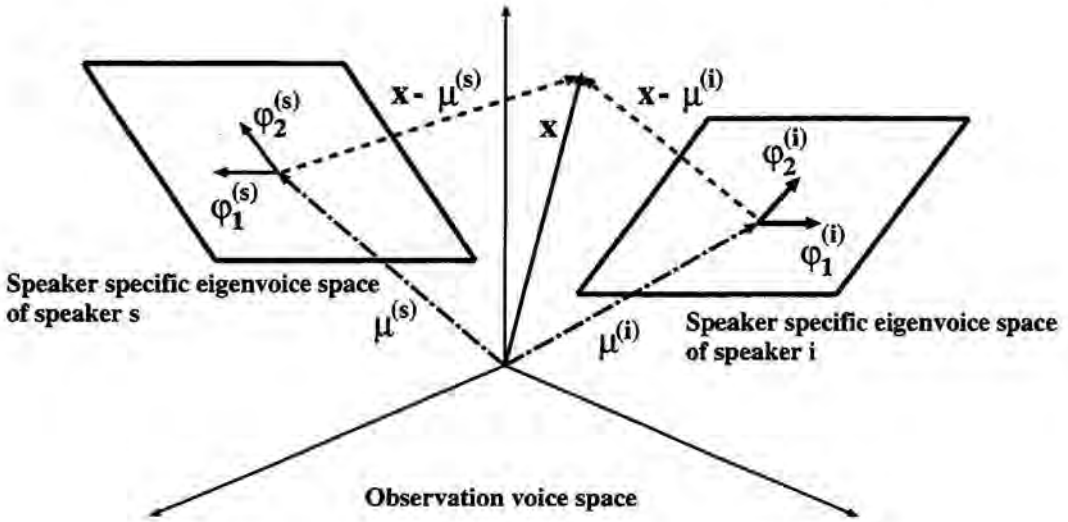


図 2.8: 主成分分析法

$$\hat{PL}^{(s)} = \frac{PL^{(s)}}{\max_{i \neq s} PL^{(i)}} \quad (2.48)$$

これは、申告話者 s を除いた全登録話者の話者固有音声空間との最大射影量で割ることによって正規化を行う方法である。正規化した射影量が閾値より大きければ、本人の音声であると判定する。

次に、主成分分析に基づく部分空間法について述べる。本論文では、この主成分分析に基づく部分空間法を主成分分析法と呼ぶ。

主成分分析 (PCA: Principal Component Analysis) は、多変量解析における次元削減のための手法である。この手法では変数の数を減らすために、分散の小さな一次結合成分を取り除き、大きな分散をもつ成分のみを扱うことができる。CLAFIC法では、平均二乗誤差最小基準に基づいて学習データとの距離が最小になる話者固有音声空間を求めているのに対して、主成分分析法は分散最大基準に基づいて学習データから話者固有音声空間を求める手法である。また、主成分分析法は、平均値移動を伴う場合へ CLAFIC 法を拡張したものである。図 2.8 に主成分分析法の概念図を示す。

まず、主成分分析法による話者認識における学習法について述べる。観測音声空間で観測される話者 i の学習音声データを n 次元特徴ベクトルの集合 $\{x_t^{(i)}\} (t = 1, 2, \dots, N)$

とする。この学習データから、平均ベクトル $\mu^{(i)}$ および分散共分散行列 $\Sigma^{(i)}$ を次式で求める。

$$\mu^{(i)} = \frac{1}{N} \sum_{t=1}^N x_t^{(i)} \quad (2.49)$$

$$\Sigma^{(i)} = \frac{1}{N} \sum_{t=1}^N (x_t^{(i)} - \mu^{(i)})(x_t^{(i)} - \mu^{(i)})^T \quad (2.50)$$

この分散共分散行列 $\Sigma^{(i)}$ を固有値分解すると、次式のようになる。

$$\Sigma^{(i)} = V^{(i)} \Lambda^{(i)} V^{(i)T} \quad (2.51)$$

ここで、 $\Lambda^{(i)}$ は、分散共分散行列 $\Sigma^{(i)}$ の固有値 $\lambda_d^{(i)} (d = 1, \dots, k, \dots, n)$ を対角成分にもつ対角行列である。また、 $V^{(i)}$ は、分散共分散行列 $\Sigma^{(i)}$ の固有ベクトル $\varphi_d^{(i)} (d = 1, \dots, k, \dots, n)$ を列ベクトルとする行列である。

図2.8に示すように、固有ベクトル $\varphi_d^{(i)}$ は正規直交基底ベクトルとして、話者 i の音声データを表現する話者固有音声空間を形成している。話者 i の話者固有音声空間を構成する場合、固有値 $\lambda_d^{(i)}$ が大きい上位 k 個の固有ベクトル $\varphi_d^{(i)}$ を正規直交基底ベクトルとして、 k 次元話者固有音声空間を構成する。したがって、 n 次元観測音声空間における話者 i のパターン分布を k 次元話者固有音声空間で近似することができる。

次に、主成分分析法による話者認識について述べる。話者識別では、入力特徴ベクトルの集合 $\{x_t\}$ を各登録話者 i の話者固有音声空間に射影し、式(2.52)により投影距離 [65] を求める。投影距離の概念図を図2.9に示す。

$$PD^{(i)} = \frac{1}{N} \sum_{t=1}^N \{ \|x_t - \mu^{(i)}\|^2 - \sum_{d=1}^k (x_t - \mu^{(i)}, \varphi_d^{(i)})^2 \} \quad (2.52)$$

ここで、 N は入力音声の総フレーム数である。式(2.53)により、各登録話者 i の話者固有音声空間に対する投影距離のうち、最小距離をもつ登録話者 s を識別結果とする。

$$s = \arg \min_i PD^{(i)} \quad (2.53)$$

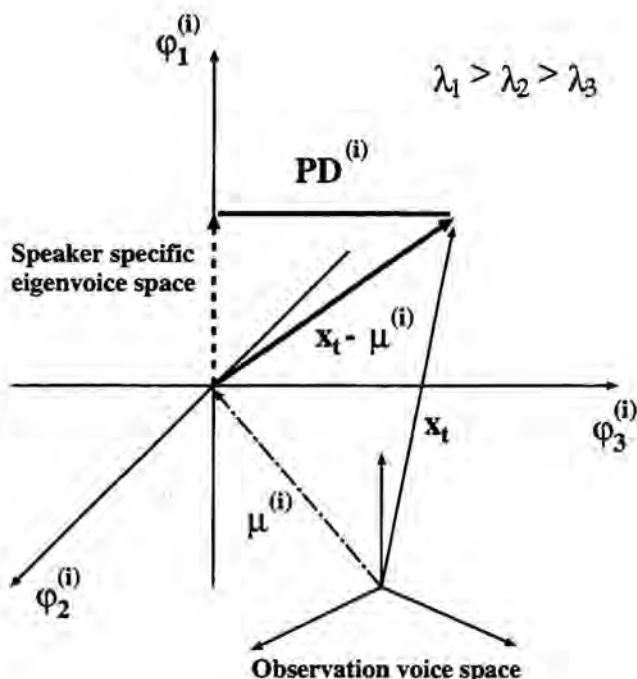


図 2.9: 投影距離

話者照合では，入力特徴ベクトルの集合 $\{x_t\}$ を本人 s であると申告された話者固有音声空間に射影し，式 (2.52) により投影距離を求める．尤度比尺度に基づく正規化法の定式化を参照し，求めた投影距離に対して式 (2.54) に示す距離の比に基づく正規化を行う．

$$PD^{(s)} = \frac{PD^{(s)}}{\min_{i \neq s} PD^{(i)}} \quad (2.54)$$

これは，申告話者 s を除いた全登録話者の話者固有音声空間との最小投影距離で割ることによって正規化を行う方法である．正規化した距離が閾値より小さければ，本人の音声であると判定する．

2.3.6 識別学習

ここでは，学習パターンに対する誤識別率を最小化する識別学習法として，最小分類誤り学習（MCE）／一般化確率的降下法（GPD）[66] について述べる．

最小分類誤り学習 (MCE) は、学習パターンに対する誤識別率最小化基準に基づいた学習法である。クラス $\omega_i (1 \leq i \leq c)$ に対する識別関数を $g_i(x; \theta)$ とする。 θ は、識別関数を規定するパラメータである。識別関数法による決定規則では、 c 次元ベクトル

$$\Psi(x; \theta) = (g_1(x; \theta), g_2(x; \theta), \dots, g_c(x; \theta)) \quad (2.55)$$

に対し、最大の要素のインデックスが x のクラスとなる。すなわち、

$$\max_i \{g_i(x; \theta)\} = g_k(x; \theta) \quad (2.56)$$

のとき、 $x \in \omega_k$ である。

$x \in \omega_i$ のパターンに対して、 θ に関する連続な誤分類尺度として

$$d_i(x) = -g_i(x; \theta) + \left\{ \frac{1}{c-1} \sum_{j \neq i} g_j(x; \theta)^\eta \right\}^{1/\eta} \quad (2.57)$$

を考える。ここで、 η は正定数で、この値が大きくなるにつれて式 (2.57) の右辺第2項は、 $g_j(x; \theta), j \neq i$ の中で最も値の大きなものが支配的となる。その極限として $\eta \rightarrow \infty$ のとき、式 (2.57) は

$$\begin{aligned} d_i(x; \theta) &= -g_i(x; \theta) + g_k(x; \theta) \\ k &= \arg \max_{j \neq i} g_j(x; \theta) \end{aligned} \quad (2.58)$$

となる。

この誤分類尺度による損失関数を次式のように定義する。

$$l_i(x; \theta) = \frac{1}{1 + \exp(-\xi d_i)} \quad (2.59)$$

ここで、 ξ は正定数である。この損失関数は、 $d_i(x)$ が小さくなるにつれて損失が0に漸近し、逆に $d_i(x)$ が大きくなるにつれて損失が1に漸近し、 $d_i(x) = 0$ 付近では分類結果の正解、不正解にかかわらず同程度の損失が付与されることになる。これにより、ク

ラス境界付近に位置し、クラスラベルがベイズ決定と異なる学習パターンに対しても適切な損失が与えられ、滑らかな識別境界が得られることになる。滑らかさの度合いは、パラメータ ξ や η に依存する。

次に、一般化確率的降下法（GPD）について述べる。一般化確率的降下法は、期待損失最小化を実現するためのパラメータ θ の推定法である。損失関数を $l_i(x; \theta)$ とすると、期待損失は次のように書ける。

$$\begin{aligned} L(\theta) &= E_{x, \omega_i} \{l_i(x; \theta)\} \\ &= \sum_{i=1}^c \int l_i(x; \theta) P(\omega_i | x) p(x) dx \end{aligned} \quad (2.60)$$

したがって、最適な θ は $\partial L / \partial \theta$ の解として得られる。しかし、実際の応用では $p(x)$ や $P(\omega_i | x)$ が未知のため、直接計算することはできない。そこで、 L の代わりに経験損失の最小化を考える。

具体的には、式(2.60)の $p(x)$ を n 個のパターン分布を表す経験分布

$$p(x) = \frac{1}{n} \sum_{p=1}^n \delta(x - x_p) \quad (2.61)$$

で近似し、さらに式(2.60)の $P(\omega_i | x)$ を、与えられたクラスラベルに基づき

$$P(\omega_i | x) = \begin{cases} 1 & \text{if } x \in \omega_i \\ 0 & \text{otherwise} \end{cases}$$

とおくと、経験損失 $L_e(\theta)$ は式(2.60)より

$$\begin{aligned} L_e(\theta) &= \frac{1}{n} \sum_{i=1}^c \sum_{p=1}^n \int l_i(x; \theta) 1(x \in \omega_i) \delta(x - x_p) dx \\ &= \frac{1}{n} \sum_{p=1}^n \sum_{i=1}^c l_i(x_p; \theta) 1(x_p \in \omega_i) \end{aligned} \quad (2.62)$$

となる。 L_e の θ に関する微分は

$$\frac{\partial L_e}{\partial \theta} = \frac{1}{n} \sum_{p=1}^n \sum_{i=1}^c \frac{\partial l_i(x_p; \theta)}{\partial \theta} 1(x_p \in \omega_i) \quad (2.63)$$

となるので, θ は最急降下法により

$$\begin{aligned}\theta(t+1) &= \theta(t) - \rho(t) \frac{\partial L_e}{\partial \theta} \\ &= \theta(t) - \rho(t) \frac{1}{n} \sum_{p=1}^n \sum_{i=1}^c \nabla l_i(x_p; \theta(t)) 1(x_p \in \omega_i)\end{aligned}\quad (2.64)$$

として逐次推定できる.

確率的降下法では, パターンが逐次的に与えられるたびにパラメータ θ を適応的に推定する方法である. つまり, 経験損失 L_e に関する期待値 $E\{L_e\}$ の減少方向に, パラメータ θ が修正される.

いま, t 回目の反復における θ の推定値を $\theta(t)$ とし, $x(t)$ が提示されたときに, $(t+1)$ 回目で $\delta\theta$ だけ修正したとする. すなわち

$$\theta(t+1) = \theta(t) + \delta\theta(t) \quad (2.65)$$

ここで, $\delta\theta(t)$ が微小とすると, $\delta\theta(t)$ に伴う L_e の変化分は

$$\begin{aligned}\delta L_e(t) &= L_e(\theta(t) + \delta\theta(t)) - L_e(\theta(t)) \\ &\approx \delta\theta(t)^t \nabla L_e(\theta(t))\end{aligned}\quad (2.66)$$

となる. 式(2.66)の期待値をとると

$$E\{\delta L_e(t)\} = E\{L_e(\theta(t))\}^t \nabla L_e(\theta(t)) \quad (2.67)$$

を得る. 確率的な降下を実現するためには $E\{\delta L_e(t)\} < 0$ であればよいので, $E\{L_e(\theta(t))\}$ が任意の正定値行列 C を用いて

$$E\{\delta\theta(t)\} = -\rho(t) C \nabla L_e(\theta(t)) \quad (2.68)$$

と書ければよい. ここで, $\rho(t)$ は正定数である.

式(2.60)から $\nabla L_e(\theta(t)) = E\{\nabla l_i(x(t); \theta(t))\}$ の関係があるので、これを式(2.68)に代入すると

$$E\{\delta\theta(t)\} = -\rho(t)C\nabla l_i(x(t); \theta(t)) \quad (2.69)$$

を得る。したがって、 $x(t) \in \omega_i$ に対しては

$$\delta\theta(t) = -\rho(t)C\nabla l_i(x(t); \theta(t)) \quad (2.70)$$

と修正する。最終的に、確率的降下法によりパラメータ θ は

$$\theta(t+1) = \theta(t) - \rho(t)C \sum_{i=1}^c \nabla l_i(x(t); \theta(t)) 1(x(t) \in \omega_i) \quad (2.71)$$

として逐次推定できる。

識別学習を用いた話者認識に関する研究としては、HMMを用いた話者照合において、照合誤り率が最小になるように、モデルのパラメータを推定する方法が提案されている[67]。また、HMMを用いた話者識別において、分類誤り率が最小になるように、モデルのパラメータを推定する方法が提案されている[68]。

2.4 まとめ

従来の話者認識手法は、一般的に音声認識と同じ特徴量ならびにモデル化手法が用いられており、音韻性と話者性が含まれている音声データを直接統計的にモデル化している。したがって、音韻性と話者性が分離されていないため音韻性に起因する特徴変動が生じることから、精度の高い認識が困難であると考えられる。この音韻性と話者性を分離し、話者性だけを持つ音声データを基に話者モデルを設計することができれば、発話内容に依存しない頑健な話者認識が可能になると考えられる。つまり、話者認識に対して特有の特徴抽出ならびにモデル化を行う必要があると考えられる。そこで、次章から音声データに含まれている音韻性と話者性の分離という観点から、話者性の抽出に関する研究を行う。

第3章

話者認識における距離尺度

3.1 まえがき

本章では，主成分分析に基づいた部分空間法における距離尺度について述べる．パターン認識における代表的な手法である部分空間法と疑似ベイズ識別法は，ガウス分布をパターン分布モデルとして導くことができる．このとき，部分空間法は，ベイズ識別法の近似解であることが明らかとなっている [69]．これは，ガウス分布から最初の識別関数であるベイズ識別が得られ，ベイズ識別から第1近似により疑似ベイズ識別，疑似ベイズ識別から第2近似により疑似マハラノビス距離，疑似マハラノビス距離から第3近似により投影距離（部分空間法）が得られるという流れになっている．そこで，本論文では，ガウス分布から得られる最初の識別関数であるベイズ識別からスタートし，投影距離までの導出について述べる．

学習データの平均ベクトルと入力特徴ベクトルとの距離が離れていた場合，異なる話者である可能性が高い．しかし，投影距離は，入力特徴ベクトルと平均ベクトルとの距離を考慮していないので，認識を誤ってしまう場合がある．そこで，話者を識別する上で，入力特徴ベクトルと平均ベクトルとの距離が短く，かつ投影距離も短くなるように識別関数を定義する必要がある．これら2つの距離の変動を抑制するために，それぞれの距離を分散で正規化する．分散で正規化した入力特徴ベクトルと平均ベクトルとの距離はマハラノビス距離であり，分散で正規化した投影距離とこのマハラノビス距離を統合する．これらを踏まえて，本研究では話者認識の距離尺度として，2つの距離に重み付けすることで統合した重み付きマハラノビス距離法を提案する．図3.1に，各識別関数の関係図を示す．本手法の有効性を示すために，従来の各識別関数（ベ

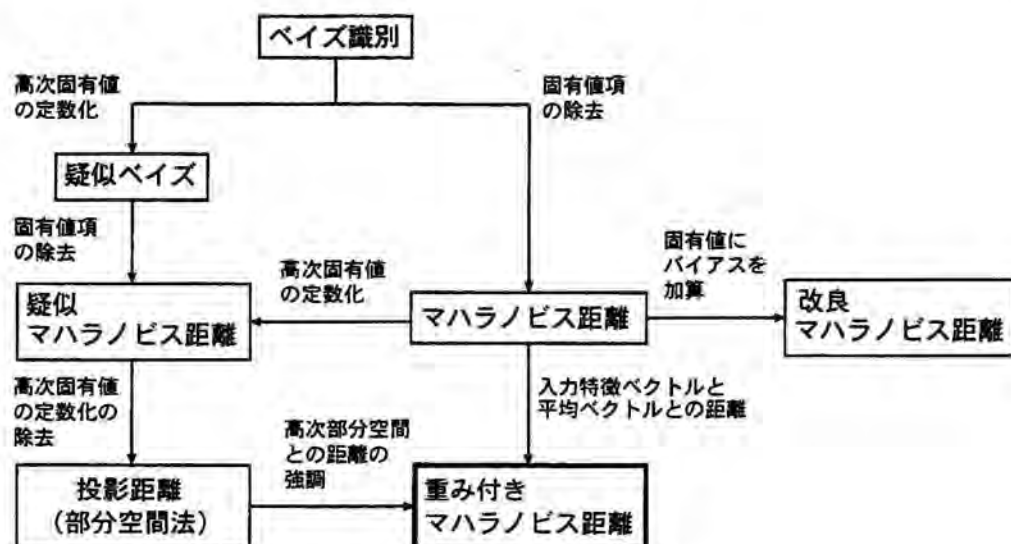


図 3.1: 各識別関数の関係

イズ識別, 疑似ベイズ識別, マハラノビス距離, 疑似マハラノビス距離, 投影距離), さらに改良マハラノビス距離 [70] に対して, 話者照合の比較実験を行った。

3.2 ベイズ識別

パターン x のクラス ω における確率密度関数 $p(x|\omega)$ が, 次式のようなガウス分布で表される場合を考える。

$$p(x|\omega) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right\} \quad (3.1)$$

このとき, n 次元特徴ベクトル x に対するクラス ω の 2 次識別関数は, 次式のように表される。ここで, 式 (3.2) はベイズ識別に相当する。

$$g(x) = (x - \mu)^T \Sigma^{-1}(x - \mu) + \log |\Sigma| - 2 \log P(\omega) \quad (3.2)$$

式 (3.2) において, 事前確率 $P(\omega)$ が全クラスで等しいとして省き, さらに分散共分散行列 Σ とその固有値 λ_d , 固有ベクトル φ_d の関係

$$\Sigma = V \Lambda V^T = \sum_{d=1}^n \lambda_d \varphi_d \varphi_d^T \quad (3.3)$$

から式(3.2)のベイズ識別は、次式のように書き直すことができる。

$$D(x) = \sum_{d=1}^n \frac{(x - \mu, \varphi_d)^2}{\lambda_d} + \sum_{d=1}^n \log \lambda_d \quad (3.4)$$

この式(3.4)の $D(x)$ が最小になるカテゴリーに入力パターン x を分類する。

3.3 疑似ベイズ

疑似ベイズ法は、ピーキング現象の原因となる分散共分散行列の推定誤差を改善するため、推定誤差が大きく表れる高次の固有値を、ある定数 δ で置き換えた手法である[71]。

ベイズ識別の式(3.4)において高次の固有値を定数 δ で置き換えると、

$$D(x) = \sum_{d=1}^k \frac{(x - \mu, \varphi_d)^2}{\lambda_d} + \sum_{d=k+1}^n \frac{(x - \mu, \varphi_d)^2}{\delta} + \sum_{d=1}^k \log \lambda_d + \sum_{d=k+1}^n \log \delta \quad (3.5)$$

となる。さらに、次式

$$\sum_{d=1}^n (x - \mu, \varphi_d)^2 = \|x - \mu\|^2 \quad (3.6)$$

を用いて式(3.5)を変形すると、

$$D(x) = \|x - \mu\|^2 - \sum_{d=1}^k \left(1 - \frac{\delta}{\lambda_d}\right) (x - \mu, \varphi_d)^2 + \delta \sum_{d=1}^k \log \lambda_d + \delta \sum_{d=k+1}^n \log \delta \quad (3.7)$$

となる。この式(3.7)の $D(x)$ が最小になるカテゴリーに入力パターン x を分類する。

3.4 マハラノビス距離

マハラノビス距離は、式(3.2)の2次識別関数の第1項

$$D(x) = (x - \mu)^T \Sigma^{-1} (x - \mu) \quad (3.8)$$

に相当する。式(3.3)の分散共分散行列とその固有値、固有ベクトルの関係から式(3.8)を変形すると、

$$D(x) = \sum_{d=1}^n \frac{(x - \mu, \varphi_d)^2}{\lambda_d} \quad (3.9)$$

となる。式(3.9)のマハラノビス距離は、式(3.4)のベイズ識別において、第2項 $\sum_{d=1}^n \log \lambda_d$ の固有値項を省いた近似式になっている。

3.5 疑似マハラノビス距離

疑似マハラノビス距離は、疑似ベイズと同様にピーキング現象の原因となる分散共分散行列の推定誤差を改善するため、推定誤差が大きく表れる高次の固有値をある定数 δ で置き換えた手法である [71]。

マハラノビス距離の式(3.9)において高次の固有値を定数 δ で置き換えると、

$$D(x) = \sum_{d=1}^k \frac{(x - \mu, \varphi_d)^2}{\lambda_d} + \sum_{d=k+1}^n \frac{(x - \mu, \varphi_d)^2}{\delta} \quad (3.10)$$

となる。さらに、式(3.6)を用いて式(3.10)を変形すると、

$$D(x) = \|x - \mu\|^2 - \sum_{d=1}^k \left(1 - \frac{\delta}{\lambda_d}\right) (x - \mu, \varphi_d)^2 \quad (3.11)$$

となる。式(3.11)の疑似マハラノビス距離は、式(3.7)の疑似ベイズにおいて、第3項および第4項の固有値項を省いた近似式になっている。

3.6 投影距離

2.3.5節の図2.9は、3次元の観測音声空間における投影距離の概念図を示している。図2.9の φ_d は固有ベクトルを表しており、各固有ベクトルの固有値を λ_d とすると、

$$\lambda_1 > \lambda_2 > \lambda_3 \quad (3.12)$$

のような関係が成り立っている。図2.9から投影距離は、入力特徴ベクトル x と平均ベクトル μ の差ベクトル $x - \mu$ と、 $x - \mu$ を固有ベクトル φ_1, φ_2 で構成される話者固有音声空間へ射影した射影ベクトルとの距離として、式(3.13)のように定義できる。

$$PD(x) = \|x - \mu\|^2 - \sum_{d=1}^k (x - \mu, \varphi_d)^2 \quad (3.13)$$

式(3.13)の投影距離は、式(3.11)の疑似マハラノビス距離において、固有値 $\lambda_d \gg \delta$ であると仮定して高次の固有値を省いた形になる。さらに、式(3.6)を用いて式(3.13)を変形すると、

$$PD(x) = \sum_{d=k+1}^n (x - \mu, \varphi_d)^2 \quad (3.14)$$

となる。これは、話者固有音声空間の直交補空間への射影量を表している。

3.7 改良マハラノビス距離

式(3.9)で表されるマハラノビス距離は、分散（固有値）が0に近い軸が数多く存在し、これらの軸が冗長である。また、分散（固有値）の小さい軸が距離に過大な影響を与えてしまい、認識率の低下をもたらすといった問題がある。つまり、距離を計算する場合、固有値が分母となっているため、小さい固有値の方がより大きい影響をもつことになり、特徴量の変動特性をそれほど反映しない軸が距離を左右してしまう危険がある。これを防ぐために、改良マハラノビス距離は、式(3.9)の固有値にバイアスを加える方法で次式のようなになる[70]。

$$D(x) = \sum_{d=1}^n \frac{(x - \mu, \varphi_d)^2}{\lambda_d + b} \quad (3.15)$$

本研究での改良マハラノビス距離は、話者固有音声空間の直交補空間における距離として式(3.16)のように定義した。

$$D(x) = \sum_{d=k+1}^n \frac{(x - \mu, \varphi_d)^2}{\lambda_d + b} \quad (3.16)$$

3.8 重み付きマハラノビス距離

学習データの平均ベクトルと入力特徴ベクトルとの距離が離れていた場合、異なる話者である可能性が高い。しかし、投影距離は、入力特徴ベクトルと平均ベクトルとの距離を考慮していないので、認識を誤ってしまう場合がある。そこで、話者を識別する上で、入力特徴ベクトルと平均ベクトルとの距離が短く、かつ投影距離も短くなるように識別関数を定義する必要がある。入力特徴ベクトルと平均ベクトルとの距離の変動を抑制するために、距離を分散で正規化する。この正規化された距離はマハラノビス距離であり、分散で正規化した投影距離とこのマハラノビス距離を統合する。

これら2つの距離に重み付けすることで統合した重み付きマハラノビス距離法を提案する。これは、まず差ベクトル $x - \mu$ のノルムと線形結合したものとして、話者固有音声空間の直交補空間とのマハラノビス距離を重み α で式(3.17)のように定義する。

$$D(x) = \|x - \mu\|^2 + \alpha \sum_{d=k+1}^n \frac{(x - \mu, \varphi_d)^2}{\lambda_d} \quad (3.17)$$

式(3.17)の第1項をマハラノビス距離に換算すると、

$$D(x) = \sum_{d=1}^k \frac{(x - \mu, \varphi_d)^2}{\lambda_d} + (1 + \alpha) \sum_{d=k+1}^n \frac{(x - \mu, \varphi_d)^2}{\lambda_d} \quad (3.18)$$

となる。これは、マハラノビス距離の計算において、話者固有音声空間の直交補空間との距離を強調した距離尺度になっている。

表 3.1: 分析条件

サンプリング周波数	12kHz
高域強調	なし
フレーム長	20ms
フレーム周期	5ms
窓タイプ	ハミング窓
特徴パラメータ	LPC ケプストラム (16 次)

3.9 話者照合実験

従来の各識別関数（ベイズ識別，疑似ベイズ，マハラノビス距離，疑似マハラノビス距離，投影距離，改良マハラノビス距離）と，提案手法である重み付きマハラノビス距離による話者照合実験を行った。

3.9.1 実験条件

使用したデータは，男性 10 名・女性 5 名が約 10 ヶ月に渡る 3 時期（時期 1，時期 2，時期 3）に発声した文章データである．分析条件を表 3.1 に示す．

話者 15 名（男性 10 名・女性 5 名）を登録話者として，各登録話者に対する詐称者はその他の登録話者 14 名とした．学習には時期 1 に発声した 10 文章を用い，評価には時期 2 および時期 3 に発声した各 15 文章を 1 文章ずつ用いた．1 文章は，平均約 4 秒である．

本研究では，式 (3.19) により各識別関数における距離の正規化を行った．式 (3.19) は，本人 s であると申告された話者モデルとの距離を $dist^{(s)}$ とし，全登録話者のモデルとの距離総和で正規化を行う．この正規化した距離が閾値 θ より小さければ，本人の音声であると判定する．

$$dist^{(s)} = \frac{dist^{(s)}}{\sum_i dist^{(i)}}$$

(3.19)

本研究では，閾値 θ は，話者毎に本人棄却率と詐称者受理率の 2 つの誤り率が等しくなるように，事後的に設定している．

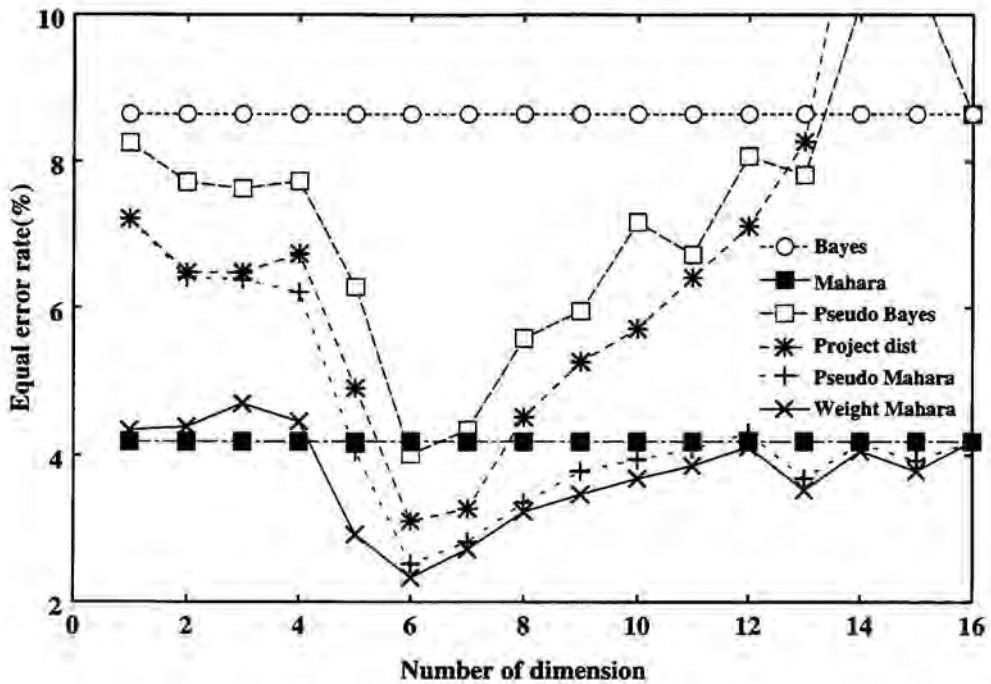


図 3.2: 話者照合結果 (1)

3.9.2 実験結果と考察

話者照合実験を行った結果を図3.2および表3.2に示す。図3.2は、従来の各識別関数（ベイズ識別、疑似ベイズ、マハラノビス距離、疑似マハラノビス距離、投影距離）と、提案手法である重み付きマハラノビス距離による話者照合結果である。縦軸は等誤り率、横軸は固有ベクトルの数（打ち切り次元数）を表している。図3.2中の *Bayes* はベイズ識別、*Pseudo Bayes* は疑似ベイズ、*Mahara* はマハラノビス距離、*Pseudo Mahara* は疑似マハラノビス距離、*Project dist* は投影距離、*Weight Mahara* は重み付きマハラノビス距離を示している。

表3.2は、従来の各識別関数（ベイズ識別、疑似ベイズ、マハラノビス距離、疑似マハラノビス距離、投影距離、改良マハラノビス距離）と、提案手法である重み付きマハラノビス距離において、各固有ベクトル数のうち最も照合精度がよかった場合の等誤り率を示している。

また、予備実験により、疑似ベイズの高次固有値 $\delta = 2$ 、疑似マハラノビス距離の高次固有値 $\delta = 2$ 、改良マハラノビス距離のバイアス $b = 1$ 、重み付きマハラノビス距離の重み $\alpha = 5$ に設定した。

表 3.2: 話者照合結果 (2)

	等誤り率 (%)
ベイズ識別	8.64
マハラノビス距離	4.18
疑似ベイズ ($\delta=2$)	4.02
投影距離	3.1
改良マハラノビス距離 ($b=1$)	2.61
疑似マハラノビス距離 ($\delta=2$)	2.52
重み付きマハラノビス距離 ($\alpha=5$)	2.33

図3.2および表3.2から照合精度は、重み付きマハラノビス距離、疑似マハラノビス距離、改良マハラノビス距離、投影距離、疑似ベイズ、マハラノビス距離、ベイズ識別という順番で精度がよいという結果になり、提案手法である重み付きマハラノビス距離が有効であることがわかった。また、疑似ベイズ、投影距離、疑似マハラノビス距離、重み付きマハラノビス距離において、固有ベクトルの数（打ち切り次元数）が6次元の時に最も照合精度がよい結果が得られた。このことから、固有ベクトルの数が6次元の時に、話者の音声データを最もよく表す話者固有音声空間が構成できるといえる。

表3.2における各識別関数の照合結果について考察する。ベイズ識別は、8.64%と誤り率が高い結果となった。ベイズ識別は、固有値を加えることで距離が変動し、閾値が不安定になってしまうため、照合精度が低くなったと考えられる。

マハラノビス距離は4.18%、疑似ベイズは4.02%に対して、投影距離は3.1%、改良マハラノビス距離は2.61%、疑似マハラノビス距離は2.52%という結果が得られた。このことから、小さい固有値に対応する固有ベクトルで構成される話者固有音声空間が、話者の識別に大きく寄与していることが確認できる。疑似ベイズでは、ベイズ識別と同様に固有値を加えることで距離が変動し、閾値が不安定になってしまう。それに対して、疑似マハラノビス距離は固有値項を省き、推定誤差が大きく影響する高次の固有値を定数で置き換えることが、有効に働いていると考えられる。

3.10 まとめ

本章では、主成分分析に基づいた部分空間法における距離尺度として、ガウス分布から得られる最初の識別関数であるベイズ識別からスタートして、投影距離までの導出について述べた。投影距離は、入力特徴ベクトルと平均ベクトルとの距離を考慮していないので、認識を誤ってしまう場合がある。そこで、話者を識別する上で、入力特徴ベクトルと平均ベクトルとの距離が短く、かつ投影距離も短くなるような識別関数として、2つの距離に重み付けすることで統合した重み付きマハラノビス距離法を提案した。

本手法の有効性を示すために、従来の各識別関数（ベイズ識別、疑似ベイズ、マハラノビス距離、疑似マハラノビス距離、投影距離、改良マハラノビス距離）と、提案手法である重み付きマハラノビス距離による話者照合の比較実験を行った。その結果、投影距離に基づく手法では等誤り率3.1%であるのに対して、本手法では等誤り率2.3%が得られ、約25%の照合精度の改善が得られた。したがって、話者を識別する上で、入力特徴ベクトルと平均ベクトルとの距離を考慮する必要があることがわかった。また、話者固有音声空間の直交補空間との距離を強調することで照合精度が向上したことから、話者固有音声空間の直交補空間は、話者性をよく表していることがわかった。

第4章

話者固有音声空間における動的・静的特徴統合による話者照合

4.1 まえがき

2.3.5節で述べた部分空間法では，発話内容や時期差による特徴変動に起因して，平均ベクトルと入力特徴ベクトルとの間にずれが生じ，話者の認識精度が劣化する．これらの特徴変動を正規化する方法として，特徴量を正規化する方法と，尤度を正規化する方法が提案されている．前者の方法としては，ケプストラム平均値正規化（Cepstrum Mean Normalization）法[72]が有効であることが確認されている．後者の方法としては，尤度比[11][12]あるいは事後確率の概念に基づいた正規化法[13]–[15]が有効であることが確認されている．特徴量の正規化は，発話内容や発声時期の違いによる物理特徴の変動，つまり特徴量の話者内変動を正規化する方法である．また，尤度の正規化は，申告話者のモデルに対する尤度を背景話者のモデルとの尤度で正規化することにより，尤度のばらつきを抑える方法である．したがって，尤度の正規化と特徴量の正規化を組み合わせることにより，話者内の変動を抑制することができ，話者の判定のための閾値をより安定して設定することができると考えられる．また，発声した少量の更新用データを用いて，発声変動に対して頑健になるように，話者のモデルを更新する方法，およびその話者のモデルの頑健性を考慮しながら，話者を判定する閾値を適切に設定する方法が提案されている[73]．

本研究では，前述したように，話者の判定のための閾値をより安定して設定するために，尤度の正規化と特徴量の正規化を組み合わせ，話者照合を行う．また，3章で

述べたように、話者を識別する上で、入力特徴ベクトルと学習データの平均ベクトルとの距離を考慮する必要がある。これらを踏まえて、本章では、音声の特徴変動に対して、より頑健な話者性の抽出手法として、話者固有音声空間における動的・静的特徴統合による話者照合法を提案する。

本手法の有効性を示すために、2.3.5節で述べた部分空間法による話者照合法との比較実験を行った。さらに、2.3.3節で述べたGMMを用いた話者照合法との比較実験も行った。

4.2 ケプストラム平均値正規化

話者認識において、電話回線等の回線特性や伝送歪み、音声入力系の周波数特性の差異、音声の特徴変動に対して、ケプストラム平均値正規化 (Cepstrum Mean Normalization) の有効性が確認されている [72][74][75]。CMNは、次式のように表される。

$$\hat{x}_t = x_t - \frac{1}{N} \sum_{t=1}^N x_t \quad (4.1)$$

ここで、 N は入力音声の総フレーム数であり、 x_t は入力音声の時刻 t における特徴ベクトルである。

CMNは、入力特徴ベクトル x_t の各次数毎に、短文章の音声区間全体での平均ベクトルを計算し、この平均ベクトルを各時点の入力特徴ベクトルから差し引くことによって逆フィルタ処理を行っている。ケプストラム上での減算は、スペクトル上での除算に対応するので、この処理により、伝送歪み、スペクトル包絡の話者内（長期間）変動を正規化することができる。しかし、この平均ベクトルは伝送歪みを表しているが、入力音声の長時間平均であり静的な話者性も表している。したがって、CMNを行うことで、静的な話者性が失われるという問題がある。そこで、本研究では、静的な話者を補う手法として、話者固有音声空間における動的・静的特徴統合法を提案する。

4.3 話者の動的・静的特徴

まず、話者固有音声空間における動的・静的特徴について述べる。

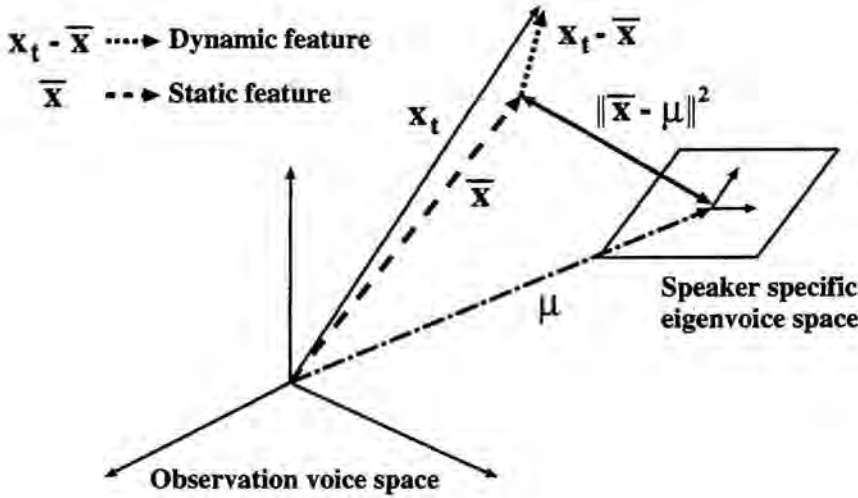


図 4.1: 話者の動的・静的特徴

CMNは、入力音声区間全体での平均ベクトルを計算し、この平均ベクトルを各時点の特徴ベクトルから差し引いている。この差ベクトルは、入力音声の特徴の時間的変化量であるため動的な話者性を表しており、ここでは話者の動的特徴と定義する。本研究では、この動的特徴と話者固有音声空間との距離を話者の動的特徴距離と呼ぶ。また、入力音声区間全体で計算した平均ベクトルを、ここでは話者の静的特徴と定義する。この静的特徴は、入力音声の静的な話者性を表しているが、時期差の影響を受けて学習データの平均ベクトルとずれが生じる。つまり、静的特徴と学習データの平均ベクトルとの差（距離）には時期差が含まれており、本研究では、これを話者の静的特徴距離と呼ぶ。話者の動的・静的特徴の概念図を図4.1に示す。

図4.1中の x_t は入力特徴ベクトル、 $\bar{x}$ は入力音声区間全体での平均ベクトル、 μ は話者の学習データの平均ベクトルを表す。また、入力特徴ベクトル x_t と入力音声区間全体での平均ベクトル $\bar{x}$ との差ベクトル $x_t - \bar{x}$ は話者の動的特徴、入力音声区間全体での平均ベクトル $\bar{x}$ は話者の静的特徴を表す。この静的特徴と学習データの平均ベクトルとの距離 $\|\bar{x} - \mu\|^2$ は、話者の静的特徴距離を表す。

話者の動的特徴距離 $Dist_{dynamic}$ を式(4.2)に、また、話者の静的特徴距離 $Dist_{static}$ を式(4.3)に示す。

$$Dist_{dynamic} = \frac{1}{N} \sum_{t=1}^N \{ \|x_t - \bar{x}\|^2 - \sum_{d=1}^k (x_t - \bar{x}, \varphi_d)^2 \} \quad (4.2)$$

$$Dist_{static} = \|\bar{x} - \mu\|^2 \quad (4.3)$$

ここで、 N は入力音声の総フレーム数である。式(4.2)は、時期的変動を取り除いた話者固有の動的特徴に関する距離を表しており、式(4.3)は、話者の静的特徴と各話者の学習データの平均ベクトルとの距離を表している。

したがって、話者の動的特徴と静的特徴を統合することによって、動的かつ静的な話者性を表現することができ、音声の特徴変動に頑健な話者照合が可能になると考えられる。

4.4 話者の動的・静的特徴統合

次に、話者固有音声空間における動的・静的特徴統合について述べる。本研究では、距離の変動を抑制し、話者判定のための閾値を安定して設定するため、話者の動的特徴距離、および静的特徴距離をそれぞれ対数化する。さらに、静的特徴距離に重み付けすることで時期差による変動を抑制し、動的特徴距離と統合する。

対数関数は、引数の値が大きくなるほど緩やかな勾配曲線を描くので、対数化すると値の変動を抑制することができる。本研究の実験データでは、話者の動的特徴距離は平均約26.5、分散約20.6であり、話者の静的特徴距離は、平均約11.1、分散約37.8である。したがって、それぞれの距離を対数化することで、距離の変動（分散）を抑制し、より安定した閾値設定が可能になると考えられる。ここで、これらの距離は、話者固有音声空間の次元数が5次元のときの学習データから求めた値であり、データや話者固有音声空間の次元数により一般的に異なった値となる。

話者の動的特徴距離、および静的特徴距離をそれぞれ対数化して重み付けした線形和による統合距離 $Dist_{integrated}$ を式(4.4)に示す。

$$Dist_{integrated} = \log Dist_{dynamic} + \alpha \log Dist_{static} \quad (4.4)$$

ここで、話者の静的特徴対数距離 $\log Dist_{static}$ の分散 σ^2_{static} は、話者の動的特徴対数距離 $\log Dist_{dynamic}$ の分散 $\sigma^2_{dynamic}$ に比べて大きいので、時期差を含んでいる話者の静的特徴距離により、話者の動的・静的特徴の統合距離 $Dist_{integrated}$ が変動してしまう。そこで、動的特徴対数距離の分散 $\sigma^2_{dynamic}$ と、静的特徴対数距離の分散 σ^2_{static} の比を求め、次式のように重み α を決定する。

$$\alpha = \frac{\sigma_{dynamic}^2}{\sigma_{static}^2} \quad (4.5)$$

したがって、式(4.4)において重み α は、動的な話者性と静的な話者性を統合化し、さらに静的特徴距離に含まれている時期差を抑制していることになる。

話者照合における距離の正規化法としては、ここでは距離の比に基づく正規化を行う。話者の動的特徴距離、および静的特徴距離の統合距離における正規化法を式(4.6)に示す。

$$\hat{Dist}_{integrated}^{(s)} = Dist_{integrated}^{(s)} / \min_{i \neq s} Dist_{integrated}^{(i)} \quad (4.6)$$

これは、申告話者 s の話者固有音声空間との統合距離 $Dist_{integrated}^{(s)}$ を、申告話者を除いた全登録話者の話者固有音声空間との最小距離で割ることによって、正規化を行う方法である。

4.5 話者固有音声空間への投影距離との話者照合実験

4.5.1 実験条件

使用したデータは、男性10名・女性5名が約10ヵ月に渡る3時期（時期1、時期2、時期3）に発声した文章データである。時期2は時期1から7ヶ月後、時期3は時期1から10ヶ月後に発声した文章データである。分析条件を表4.1に示す。

話者15名（男性10名・女性5名）を登録話者として、各登録話者に対する詐称者は、その他の登録話者14名とした。学習には時期1に発声した10文章を用い、評価には時期2および時期3に発声した各15文章を1文章ずつ用いた。実験に用いたテスト文章は、登録話者1名に対して、15名 $\times$ 2時期 $\times$ 15文章 = 450文章であり、テストの総数は、450文章 $\times$ 15名 = 6750である。1文章は、平均約4秒である。

閾値は、話者毎に本人棄却率と詐称者受理率の2つの誤り率が等しくなるように、事後的に設定した。

話者の動・静的特徴統合の式(4.4)の重み α は、学習データから式(4.5)を用いて、 $\alpha = 0.06$ に設定した。

表 4.1: 分析条件

サンプリング周波数	12kHz
高域強調	$1 - 0.97z^{-1}$
フレーム長	20ms
フレーム周期	5ms
窓タイプ	ハミング窓
特徴パラメータ	LPC ケプストラム (16 次)

4.5.2 実験結果

提案手法である話者固有音声空間における動的・静的特徴統合の有効性を示すために、2.3.5節で述べた話者固有音声空間への投影距離による話者照合との比較実験を行った。ここで、主成分分析では式(2.50)より、学習データの平均ベクトルを学習データの各フレームの特徴ベクトルから差し引いているので、学習データに対してCMNを行っていることになる。投影距離では式(2.52)より、学習データの平均ベクトルをテストデータの各フレームの特徴ベクトルから差し引いているが、これはテストデータをCMNしたことにならない。したがって、話者固有音声空間への投影距離においては、テストデータをCMNしない場合と、CMNした場合に分けて実験を行っている。テストデータをCMNした場合は、話者の動的特徴のみを用いた場合に対応する。また、話者の動的・静的特徴統合は、その導出過程から、学習データとテストデータをCMNしていることになる。

各手法において、時期2および時期3のテストデータを照合した結果を図4.2および表4.2に示す。図4.2は、話者固有音声空間の次元数を1次元から16次元まで変化させたときの各次元における等誤り率を示す。図中のPCAは話者固有音声空間への投影距離、PCA+CMNは話者固有音声空間への投影距離においてテストデータをCMNした場合（動的特徴のみの場合）、Dynamic and staticは話者固有音声空間における動的・静的特徴統合の結果を示す。

図4.2の結果から、話者固有音声空間においてCMNによる音声の特徴変動の正規化に対する有効性が確認できた。さらに、話者固有音声空間における動的・静的特徴統合が有効であることがわかった。また、話者固有音声空間の次元数を変化させて実験した

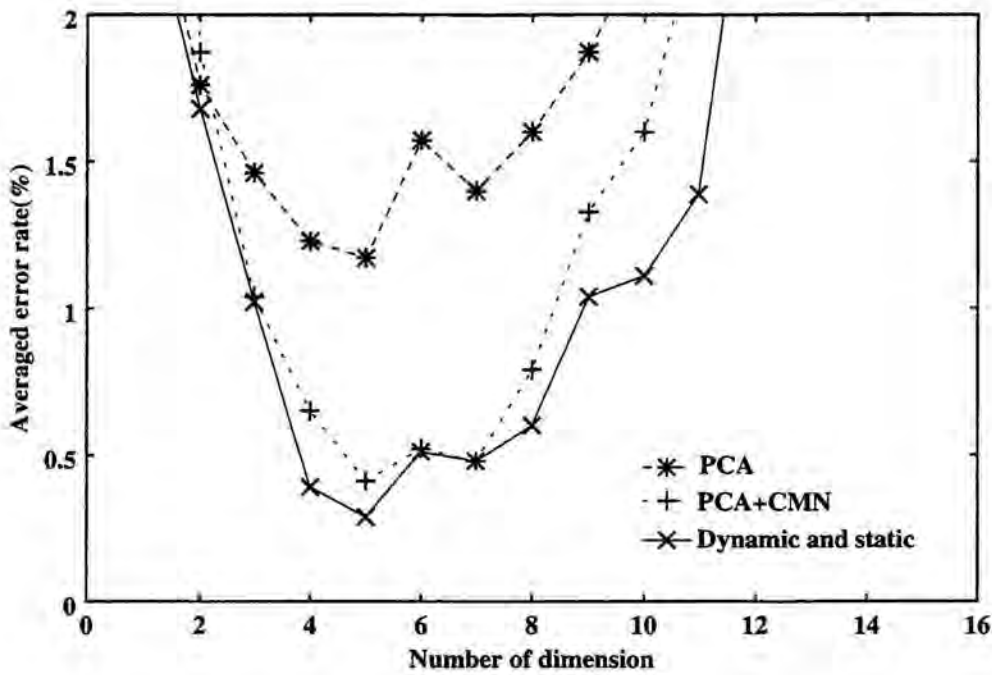


図 4.2: 話者固有音声空間への投影距離との比較結果 (1)

結果、話者固有音声空間の次元数は5次元が最適であり、この場合に話者の音声データを最もよく表す話者固有音声空間が構成できるといえる。それ以上に次元数を上げてしまうと、各話者の話者固有音声空間との重なりが大きくなり、照合精度が低下する。

表 4.2は、話者固有音声空間への投影距離および提案手法において、最も照合精度がよかった場合の等誤り率と、そのときの話者固有音声空間の次元数を示す。表 4.2から、話者固有音声空間への投影距離の場合で等誤り率 1.17% に対して、テストデータに対して CMN を行った場合（動的特徴のみの場合）で等誤り率 0.41% が得られ、CMN を行うことによって約 65% の照合精度の改善が得られた。さらに、話者固有音声空間への投影距離の場合で等誤り率 1.17% に対して、話者固有音声空間における動・静的特徴統合により等誤り率 0.29% が得られ、約 75% の照合精度の改善が得られた。また、話者の動・静的特徴統合における重みは 0.06 と小さいが、重みが 0、つまり動的特徴のみの場合では等誤り率 0.41% となり、重みが 0.06 でも充分意味があり、話者の動・静的特徴を統合することにより、約 30% の照合精度の改善が得られた。

表 4.2の結果から、提案手法である話者固有音声空間における動・静的特徴統合法は、音声の特徴変動に対して頑健な手法であることがわかった。

表 4.2: 話者固有音声空間への投影距離との比較結果 (2)

	等誤り率 (%)
PCA (5次元)	1.17
PCA (5次元) + CMN	0.41
動的・静的特徴統合した PCA (5次元, $\alpha=0.06$)	0.29

4.5.3 時期差の考察

次に、各手法に対する時期差の影響について調べた。各手法において、時期2、時期3それぞれのテストデータを照合した結果を図4.3、4.4、4.5に示す。図4.3は話者固有音声空間への投影距離、図4.4は話者固有音声空間への投影距離でテストデータをCMNした場合（動的特徴のみの場合）、図4.5は話者の動的・静的特徴統合における照合結果を示す。図4.3、4.4、4.5は、各時期のテストデータに対して、話者固有音声空間の次元数を1次元から16次元まで変化させたときの各次元における等誤り率を示す。

図4.3の結果から、話者固有音声空間への投影距離においては、各次元数において時期差による影響が大きく、全体的に誤り率が高い。これに対して、図4.4の結果から、話者固有音声空間への投影距離においてテストデータをCMNすることにより、各次元数における時期差の影響を抑えることができている。また、図4.5の結果から、話者の動的・静的特徴統合により、4~6次元において、時期差の影響を抑え、誤り率をより改善することができている。この話者固有音声空間の次元数は、全話者に対して共通に設定している。図4.5の結果から、5次元のときに誤り率が最小値を示しているが、各話者の各次元数における誤り率を見ると、最も時期差の影響が抑えられ、誤り率が最小値を示す次元数は5次元であるとは限らず、話者によって最適な次元数は異なっている。

4.5.4 次元数の考察

次に、話者固有音声空間の次元数について述べる。話者固有音声空間の次元数は、図4.2の結果からもわかるように、照合精度に影響する。最適な話者固有音声空間の次元数 k を決定する一つの方法として、累積寄与率（Cumulative proportion）を用いる方

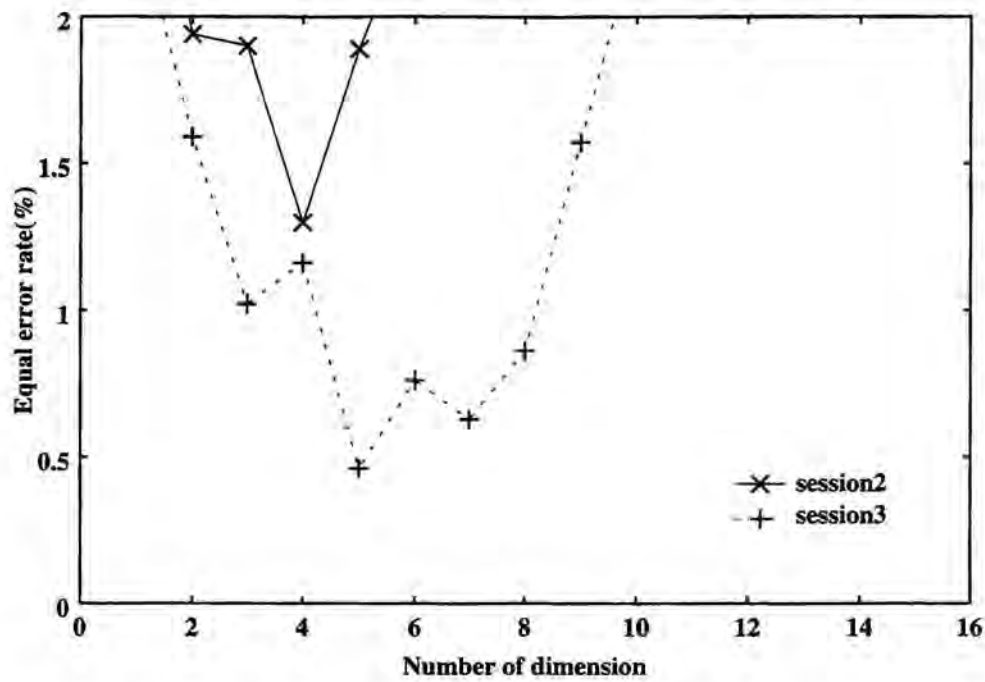


図 4.3: 話者固有音声空間への投影距離による照合結果

法がある。累積寄与率 cp_k は、固有値 $\lambda_d (\lambda_1 \geq \dots \geq \lambda_k \geq \dots \geq \lambda_n)$ を用いて

$$cp_k = \frac{\sum_{d=1}^k \lambda_d}{\sum_{d=1}^n \lambda_d} \tag{4.7}$$

と定義される。この累積寄与率が一定値以上となる次元数で打ち切ること、話者固有音声空間の次元数を決定する。式 (4.7) により、全登録話者の平均累積寄与率を求めた結果を図 4.6 に示す。

本研究のデータおよび実験の結果では、提案手法における話者固有音声空間の次元数は 5 次元が最適であり、図 4.6 から累積寄与率が 65% 程度で次元数を打ち切ることにより、最適な話者固有音声空間の次元数を決定できることがわかった。しかし、一般的に、部分空間を用いたパターン認識においては、データにより部分空間との距離が大きく変動するため、最適な部分空間の次元数を事前に決定することは困難である。したがって、実験により部分空間の次元数を決定しているのが現状である。最適な部分空間の次元数を事前に決定する方法については、今後の課題として残されている。

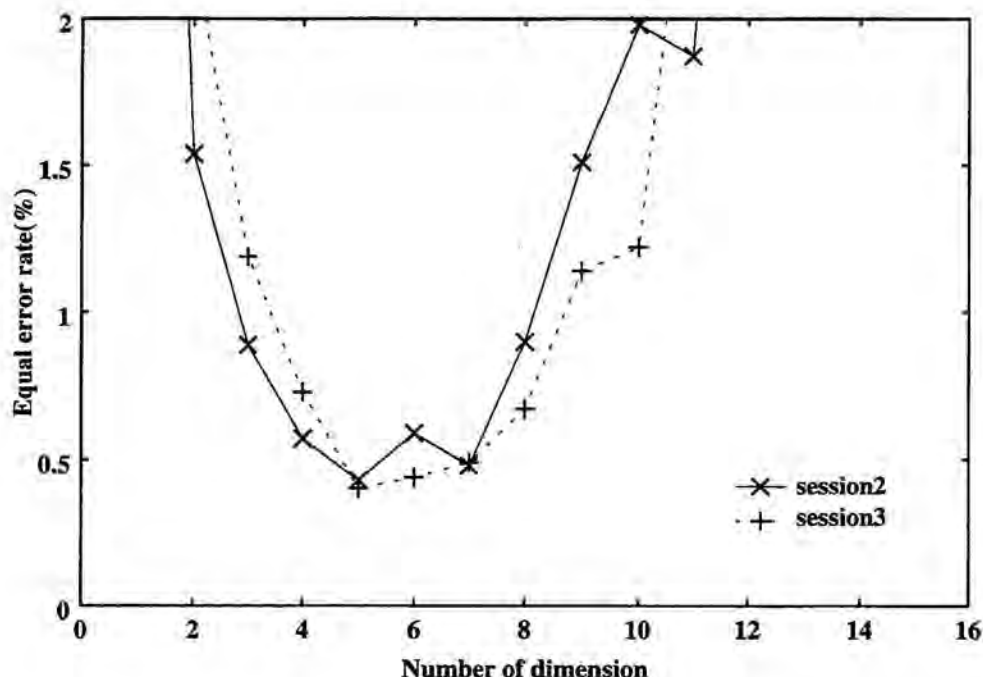


図 4.4: 話者固有音声空間への投影距離でテストデータを CMN した場合の照合結果

4.6 GMM との話者照合実験

4.6.1 実験条件

実験に使用したデータ、実験データの分析条件や閾値の設定については、4.5.1節で示した実験条件と同じである。GMM の学習および照合には、HTK (Hidden Markov Model Toolkit V2.1) [76] を用いた。また、GMM は 64 混合分布、対角共分散行列で実験を行った。実験には、Sun Ultra30 (CPU:248MHz, Memory:128MB) を使用した。

4.6.2 実験結果

提案手法である話者固有音声空間における動的・静的特徴統合の有効性を示すために、2.3.3節で述べた GMM による話者照合法との比較実験を行った。具体的には、GMM において、学習データとテストデータを CMN しない場合と、CMN した場合の 2 つの方法で実験を行った。実験結果を表 4.3 に示す。ここで、GMM における尤度の正規化には、式 (2.37) に示す尤度比尺度に基づいた正規化法を用いた。

表 4.3 は、GMM (64 混合分布) における等誤り率、および GMM において、学習データ

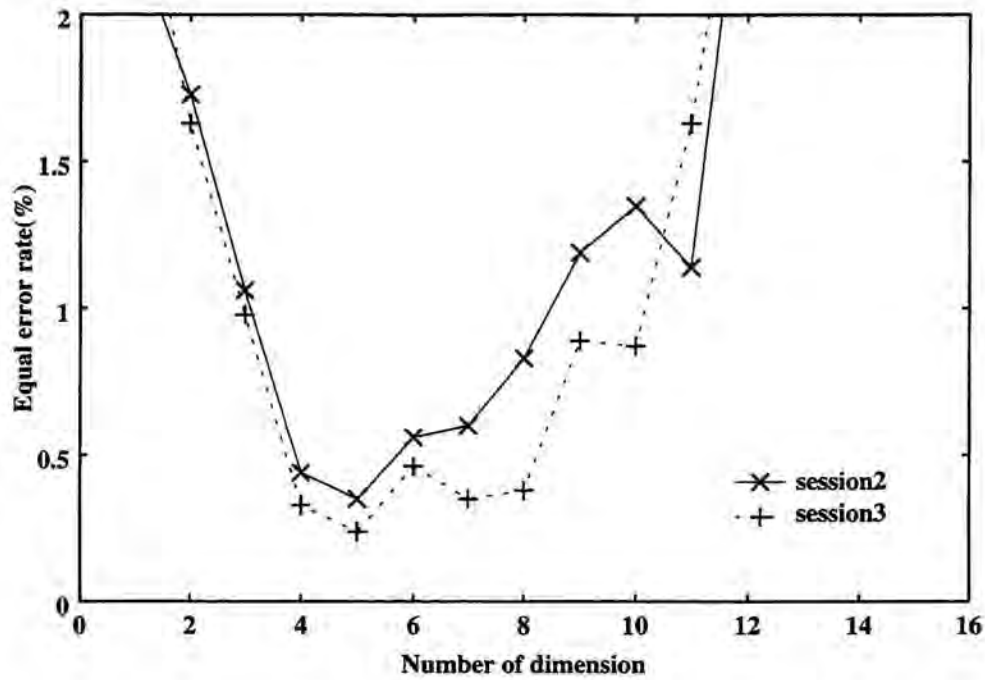


図 4.5: 話者の動的・静的特徴統合による照合結果

とテストデータを CMN した場合の等誤り率を示す。表 4.3 から、GMM において、CMN をしない場合で等誤り率 0.68% に対して、CMN を行った場合で等誤り率 0.34% が得られ、CMN を行うことによって約 50% の照合精度の改善が得られた。さらに、GMM において CMN を行った場合で等誤り率 0.34% に対して、提案手法により等誤り率 0.29% が得られている。

表 4.3: GMM との比較結果

	等誤り率 (%)
GMM (64md)	0.68
GMM (64md) + CMN	0.34
動的・静的特徴統合した PCA (5次元, $\alpha=0.06$)	0.29

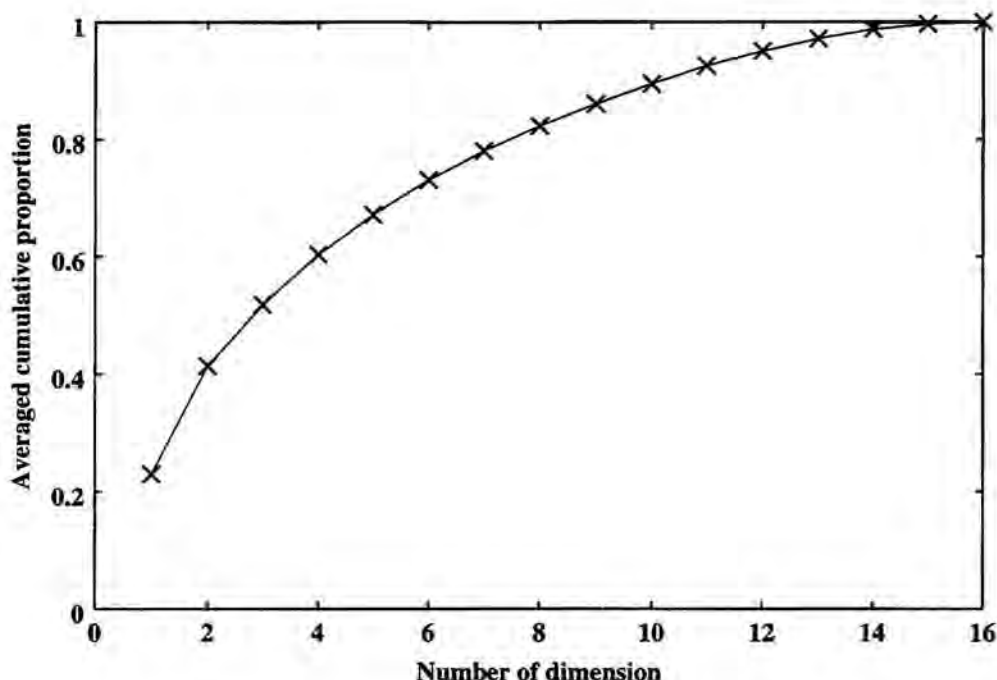


図 4.6: 平均累積寄与率

4.6.3 処理時間の考察

表4.4に、GMM（64混合分布）を基準として、話者固有音声空間（5次元）における動的・静的特徴統合に対する学習および照合処理時間、テンプレートサイズを示す。表中の学習時間は、登録話者1名10文章（1文章平均4秒）で学習するのに要する処理時間、照合時間は、1文章を照合するのに要する処理時間（登録話者1名のモデルに対する距離または尤度の計算時間であり、距離または尤度の正規化の処理時間は含まれていない）、テンプレートサイズは、登録話者1名のモデルのサイズを示す。

表 4.4: 処理時間とテンプレートサイズ

	学習時間	照合時間	テンプレート サイズ
GMM（64md）	1.0	1.0	1.0
動的・静的特徴統合した PCA（5次元）	0.0056	0.25	0.03

表4.4から、話者固有音声空間における動的・静的特徴統合法はGMMに比べて、学習処理時間が0.0056、照合処理時間が0.25、テンプレートサイズが0.03であることが分かり、より計算量が少なく、メモリ容量を削減することができる。また、話者固有音声空間における動的・静的特徴統合法では、学習処理時間が約1.5秒（10文章に対して）、照合処理時間が約0.1秒（1文章約4秒に対して）、テンプレートサイズが約1.0KB（登録話者1名に対して）であり、リアルタイムな話者照合が可能である。

以上の結果から、提案手法である話者固有音声空間における動的・静的特徴統合法は、話者照合において有効であることがわかった。

4.7 まとめ

本章では、音声の特徴変動に対して、より頑健な話者性の抽出手法として、話者固有音声空間における動的・静的特徴統合による話者照合法を提案した。

本手法の有効性を示すために、従来の話者固有音声空間への投影距離による話者照合法との比較実験を行った。その結果、従来の話者固有音声空間への投影距離では等誤り率1.17%であったが、本手法により等誤り率0.29%が得られ、約75%の照合精度の改善が得られた。さらに、従来法であるGMMを用いた話者照合法との比較実験も行った。その結果、GMMにおいてCMNを行った場合で等誤り率0.34%に対して、本手法により等誤り率0.29%が得られた。本手法はGMMに比べて、より計算量が少なく、メモリ容量を削減することができ、リアルタイムな話者照合が可能である。

以上の結果から、提案手法である話者固有音声空間における動的・静的特徴統合法は、音声の特徴変動に対して頑健な手法であり、話者照合において有効であることがわかった。また、本手法は、CMNを考慮した尺度となっている。したがって、電話回線等の回線特性、伝送歪みに頑健な話者照合が可能になると考えられる。

第5章

話者空間への射影に基づく統計的モデルによる話者認識

5.1 まえがき

4章で述べたように、話者固有音声空間における動的・静的特徴統合法では、静的特徴距離と動的特徴距離を統合化することで話者性を表していた。この話者固有音声空間は、学習パターンのお話者の音声データをよく表現した空間であるが、音韻性と話者性が含まれている。また、従来のGMMによる話者認識は、統計的手法であり、音韻性と話者性が分離されていないため、音韻性による特徴変動を含んでおり、認識精度がよくないと考えられる。また、モデルのパラメータを頑健に推定するために、多くの学習データが必要であり、モデルのパラメータ数も比較的多い。

発話内容による音声の特徴変動は、音韻性、すなわち音声中に含まれている音韻のばらつきにより生じると考えられる。これに対して、発話内容に依存しない話者性が存在するのではないかと考えられる。この点から、音韻性と話者性を分離する手法として、主成分分析に基づいた部分空間分離による話者性の抽出法を提案する。本手法では、学習パターンを主成分分析して得られる分散が大きい低次の部分空間を音韻空間と呼ぶ。また、分散が小さい高次の部分空間は、音韻空間に対する直交補空間であり、話者空間と呼ぶ。さらに、本手法に基づいて、話者空間に学習データを射影し、各話者の話者空間上でGMMにより話者性を統計的にモデル化する話者認識法を提案する。したがって、本手法により音韻性が抑えられ、話者性を精度よく抽出できると考えられる。

本手法の有効性を示すために、観測音声空間においてGMMにより話者識別および話者照合を行う方法、また、判別分析によって変換された空間においてGMMにより話者識別および話者照合を行う方法との比較実験を行った。

また、本手法で構成された話者空間は話者性を表す空間であるが、他者との相関を考慮していない。そこで、他者との違いをより強調した空間を構成するために、学習パターンに対する誤識別率を最小化する識別学習による話者空間の最適化法を提案する。話者空間の最適化は、2.3.6節で述べた最小分類誤り学習（Minimum Classification Error）／一般化確率的降下法（Generalized Probabilistic Descent）[66]により行った。

話者空間の最適化の有効性を示すために、観測音声空間における従来のGMM、最適化していない話者空間への射影に基づくGMMとの話者照合による比較実験を行った。

5.2 補空間

ここでは、まず一般的な補空間の概念について述べる。ベクトル集合 $A = (a_1, a_2, \dots, a_p)$ によって張られる部分空間を V_A 、ベクトル集合 $B = (b_1, b_2, \dots, b_q)$ によって張られる部分空間を V_B とすると、部分空間 V_A と V_B の共通部分が零ベクトルのみからなるとき、その部分空間は独立であるという。このとき、 $V_{A \cup B} = V_A \oplus V_B$ と表され、和空間 $V_{A \cup B}$ は、 V_A と V_B の直和に分解される。 n 次元全空間 E^n が二つの空間 V と W の直和、すなわち

$$E^n = V \oplus W \quad (5.1)$$

と表されるとき、 W は V の補空間といい、 $W = V^C$ と表す。

さらに、 W に属する任意のベクトルが V に属する任意のベクトルと直交するとき、 W は、 V の直交補空間といい、 $V^\perp$ と表す。このとき、全空間 E^n は V と $V^\perp$ の直和となり、次のように表される。

$$E^n = V \oplus V^\perp \quad (5.2)$$

また、直交補空間 $V^\perp$ は、次のように定義される。

$$V^\perp = \{v^\perp | (v^\perp, v) = 0, \forall v \in V\} \quad (5.3)$$

したがって、ベクトル $x \in E^n$ は、 V に属するベクトル v と $V^\perp$ に属するベクトル $v^\perp$ によって、 $x = v + v^\perp$ に一意的に分解される。全空間 E^n に含まれる p 個の 1 次独立なベクトル a_i を列ベクトルにもつ行列を $A = (a_1, a_2, \dots, a_p)$ とすると、部分空間 V_A への射影行列 P は、 $P = A(A^t A)^{-1} A^t$ と表される。この射影行列 P を用いて、ベクトル v と $v^\perp$ は、それぞれ $v = Px$ と $v^\perp = (I - P)x$ で求めることができる。

次に、投影距離と直交補空間への射影量との関係について述べる。2.3.5 節の図 2.9 は、3 次元の観測音声空間における投影距離の概念図を示している。図 2.9 で話者 i の話者固有音声空間は、正規直交基底ベクトル $\varphi_1^{(i)}$ で張られる空間として表されている。このとき、正規直交基底ベクトル $\varphi_2^{(i)}$ と $\varphi_3^{(i)}$ で張られる空間は、話者 i の話者固有音声空間 $(\varphi_1^{(i)})$ に対する直交補空間である。話者 i の学習データの平均ベクトル $\mu^{(i)}$ と入力特徴ベクトル x_t との差ベクトル $x_t - \mu^{(i)}$ を、2 つのベクトルに分解することを考える。いま、 $x_t - \mu^{(i)}$ を話者固有音声空間へ射影した射影ベクトルを $\hat{x}_t$ 、直交補空間へ射影した射影ベクトルを $\hat{x}_t^\perp$ とすると、差ベクトル $x_t - \mu^{(i)}$ は、射影ベクトル $\hat{x}_t$ と $\hat{x}_t^\perp$ によって、 $x_t - \mu^{(i)} = \hat{x}_t + \hat{x}_t^\perp$ として一意的に分解される。

直交空間において

$$\sum_{d=1}^n (x_t - \mu^{(i)}, \varphi_d^{(i)})^2 = \|x_t - \mu^{(i)}\|^2 \quad (5.4)$$

が成立することから、式 (2.52) に示す投影距離は、次式のように変形することができる。

$$PD^{(i)} = \frac{1}{N} \sum_{t=1}^N \left\{ \sum_{d=k+1}^n (x_t - \mu^{(i)}, \varphi_d^{(i)})^2 \right\} \quad (5.5)$$

これは、直交補空間への射影量を表しており、投影距離は、話者 i の話者固有音声空間に対する直交補空間への射影量と一致している。

5.3 音韻空間と話者空間の分離

発話内容による音声の特徴変動は、音韻性、すなわち音声中に含まれている音韻のばらつきにより生じると考えられる。これに対して、発話内容に依存しない話者性が

存在するのではないかと考えられる。この点から、話者内分散に基づいて音韻性と話者性を分離することができると考えられる。つまり、各話者の音声データの分散が大きい空間では音韻性が強く表れ、分散が小さい空間では、音韻性が抑えられ話者性が表れていると考える。そこで、本研究では、音韻性と話者性を分離する手法として、主成分分析に基づいた部分空間分離による話者性の抽出法を提案する。本手法では、各話者の音声データに対して主成分分析を行い、主成分である低次の軸によって構成された部分空間は音韻性が強く、高次の軸によって構成された補空間は、音韻性が抑えられ話者性を表していると考えられる。この点から、低次の軸で構成された分散の大きい部分空間を音韻空間、高次の軸で構成された分散の小さい部分空間を話者空間と呼ぶ。

次に、本手法による音韻空間と話者空間の構成法について述べる。観測音声空間で観測される話者 i の学習音声データを n 次元特徴ベクトルの集合 $\{x_t^{(i)}\}$ ($t = 1, 2, \dots, N$) とする。この学習データから、平均ベクトル $\mu^{(i)}$ および分散共分散行列 $\Sigma^{(i)}$ をそれぞれ式(2.49)、式(2.50)により求める。この分散共分散行列 $\Sigma^{(i)}$ を式(2.51)により固有値分解する。

固有値分解で得られた固有値 $\lambda_d^{(i)}$ は、対応する正規直交基底ベクトル $\varphi_d^{(i)}$ 上のデータの分散を表しており、本研究では、固有値が大きい上位 k 個の正規直交基底ベクトルにより張られた空間が音韻空間となる。また、残りの固有値が小さい $n-k$ 個の正規直交基底ベクトルにより張られた空間が話者空間となる。したがって、入力音声音を音韻空間と話者空間へそれぞれ射影することで、音韻性と話者性に分離することができる。

3次元の観測音声空間における音韻空間と話者空間への射影の概念図を図5.1に示す。図5.1で正規直交基底ベクトル $\varphi_1^{(i)}$ で張られる空間が音韻空間、正規直交基底ベクトル $\varphi_2^{(i)}$ と $\varphi_3^{(i)}$ で張られる空間が話者空間を表している。また、 $\hat{x}_{phoneme}^{(i)}$ は音韻空間への射影ベクトルである音韻ベクトル、 $\hat{x}_{speaker}^{(i)}$ は話者空間への射影ベクトルである話者ベクトルを表している。

入力特徴ベクトル x_t は、音韻空間と話者空間へ射影することで式(5.6)と式(5.7)により、音韻ベクトルと話者ベクトルに分解することができる。

$$\hat{x}_{phoneme}^{(i)} = \sum_{d=1}^k (x_t - \mu^{(i)}, \varphi_d^{(i)}) \varphi_d^{(i)} \quad (5.6)$$

$$\hat{x}_{speaker}^{(i)} = \sum_{d=k+1}^n (x_t - \mu^{(i)}, \varphi_d^{(i)}) \varphi_d^{(i)} \quad (5.7)$$

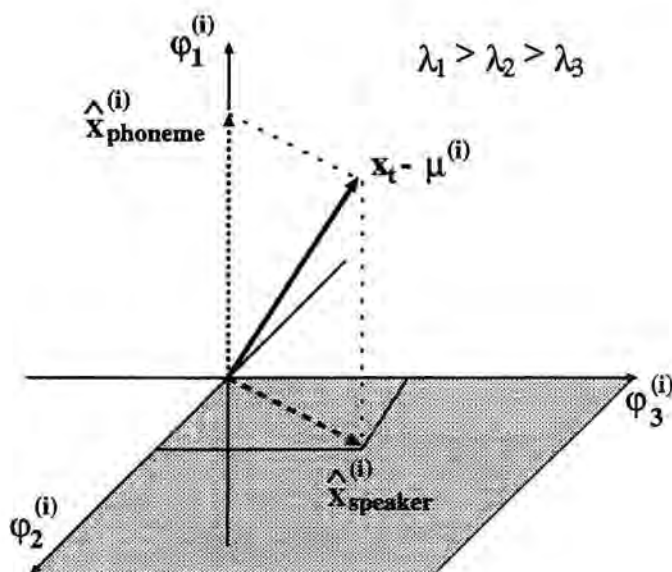


図 5.1: 音韻ベクトルと話者ベクトル

5.4 話者空間への射影に基づく統計的モデル

話者空間は、学習データを主成分分析して得られた分散が大きく音韻性が強い低次の軸を除いて、分散が小さく話者性を表す高次の軸で構成しているため、学習データを話者空間に射影すると、元の観測音声空間に比べて話者内分散が小さくなる。したがって、話者間分散と話者内分散の比で表す話者間の分離度は、話者空間へ射影することで大きくなり、より話者性を表現することができる。話者空間への射影の概念図を図 5.2 に示す。図中の P_A , P_B はそれぞれ話者 A, 話者 B の話者空間を表している。図 5.2 に示すように、話者空間は分散の小さい軸で構成された空間であるため、各話者 A, B の学習データを各話者空間に射影すると、観測音声空間に比べて話者間分散は固定のまま、話者内分散が小さくなる。

そこで、本研究では、以下の方法で話者識別および話者照合を行った。

1. 各話者の学習データを主成分分析することにより、高次の軸を選択し、各話者の話者空間を構成する。
2. 学習データを各話者の話者空間へ射影する。
3. 話者空間への射影ベクトルを逆射影し、観測音声空間で GMM により話者モデル

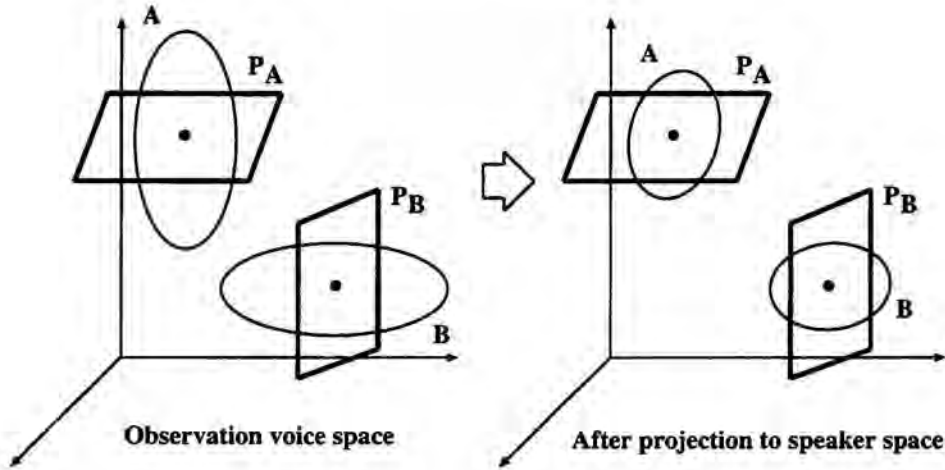


図 5.2: 話者空間への射影

を構成する。

本研究では、提案手法である話者空間への射影に基づく統計的モデルとして、GMMを用いた話者認識法をSSGMM(Speaker Space GMM)と呼ぶ。

話者認識で一般的に用いられている特徴量として、MFCC(Mel-Frequency Cepstral Coefficient)がある。MFCCは、メルフィルタバンク(MF)に対してDCT(Discrete Cosine Transform)を行うことで得られる。しかし、DCTは、データの分布を考慮せずに空間変換を行っており、特徴次元において相関が残る。そこで、本研究では、メルフィルタバンク(MF)に対してDCTの代わりに主成分分析を話者毎に行う。こうすることにより、特徴次元を無相関化し、分散共分散行列を対角化することができるので、対角成分を用いたGMMによる話者認識の精度を向上させることができる。また、信号処理においてDCTの代わりに主成分分析を行うことで、各話者の話者空間を構成する。したがって、この主成分分析は、分散共分散行列の対角化と話者空間の構成といった二つの役割を担っている。そして、本手法で構成された話者空間に学習データを射影し、話者空間でGMMにより話者モデルを学習する。

話者識別では、入力特徴ベクトルの集合 $\{x_t\}$ を式(5.7)により各登録話者 i の話者空間に射影し、射影ベクトル $\hat{x}_t$ とGMMとの対数尤度 $\log P(\hat{x}_t|\lambda^{(i)})$ を求める。式(2.36)により、各登録話者 i のGMMとの対数尤度のうち、最大対数尤度をもつ登録話者 s を識別結果とする。

話者照合では，入力特徴ベクトルの集合 $\{x_t\}$ を式 (5.7) により申告話者 s の話者空間に射影し，射影ベクトル $\hat{x}_t$ と GMM との対数尤度 $\log P(\hat{x}_t|\lambda^{(s)})$ を求める．この対数尤度に対して，式 (2.37) に示す尤度比尺度に基づく正規化を行い，この正規化した対数尤度が閾値より大きければ，本人の音声であると判定する．

5.5 話者認識実験

5.5.1 実験条件

使用したデータは，男性 20 名・女性 10 名が約 10 ヶ月に渡る 2 時期（時期 1，時期 2）に発声した文章データである．1 文章は，平均 4 秒である．分析条件を表 5.1 に示す．本研究では，以下に示す 3 つの方法で話者認識実験を行った．

1. 21 次元 MFCC の観測音声空間における GMM（従来法）
2. 21 次元 MFCC の観測音声空間から判別分析で得られた部分空間における GMM（従来法）
3. 24 次元メルフィルタバンクの観測音声空間から得られた話者空間における GMM（提案手法）

話者識別実験では，登録話者は男性 20 名・女性 10 名で，学習には時期 1 に発声した 5 文章を用い，評価には，時期 1 から約 10 ヶ月後の時期 2 に発声した各 15 文章を 1 文章ずつ用いた．話者照合実験では，登録話者は男性 10 名・女性 5 名で，詐称者は登録話者と異なる男性 10 名・女性 5 名である．学習には時期 1 に発声した 5 文章を用い，評価には，時期 2 に発声した各 15 文章を 1 文章ずつ用いた．

5.5.2 実験結果と考察

話者識別実験の結果を表 5.2，話者照合実験の結果を表 5.3 に示す．表 5.2 の GMM は，21 次元 MFCC に対して GMM で学習した場合の識別誤り率を示している．LDA は，21 次元 MFCC に対して判別分析による空間変換を行い，その後 GMM を学習した場合で，最も誤り率が小さかった 21 次元のときの結果を示している．SSGMM は，24 次メルフィルタバンクに対して提案手法である話者空間において GMM を学習した SSGMM

表 5.1: 分析条件

サンプリング周波数	16kHz
高域強調	$1 - 0.97z^{-1}$
フレーム長	20ms
フレーム周期	5ms
窓タイプ	ハミング窓
特徴パラメータ	24次メルフィルタバンク 21次ケプストラム

表 5.2: 話者識別結果 (%)

mixture number	GMM	LDA (21dim)	SSGMM (4-24dim)
2	18.44	16.67	6.44
4	12.00	11.33	5.11
8	9.56	8.22	6.44
64	8.89	6.00	7.78

の場合で、最も誤り率が小さかった4-24次元のときの結果を示している。4-24の意味は、主成分分析した結果、低次の1-3次元で音韻空間を作成し、4-24次元で話者空間を作成したことを示している。

表5.2の結果から、混合分布数が4のときに、提案手法である話者空間におけるGMMでは識別誤り率5.11%が得られ、観測音声空間で作成されたGMMの識別誤り率12.00%に比べると、57%の識別精度の改善が得られた。また、判別分析におけるGMMの識別誤り率11.33%に比べると、55%の識別精度の改善が得られた。さらに、提案手法は4混合分布のときに、64混合分布の観測空間で作成されたGMMの識別誤り率8.89%に比べると、43%の識別精度の改善が得られた。また、64混合分布の判別分析におけるGMMの識別誤り率6.00%に比べると、15%の識別精度の改善が得られた。

次に、話者照合実験の結果について述べる。表5.3のGMMは、21次元MFCCに対してGMMを学習した場合の等誤り率を示している。LDAは、21次元MFCCに対し

表 5.3: 話者照合結果 (%)

mixture number	GMM	LDA (14dim)	SSGMM (4-24dim)
2	3.96	4.52	2.44
4	2.98	3.91	1.79
8	2.28	3.07	1.88
64	2.00	2.68	1.70

て判別分析による空間変換を行い、その後GMMを学習した場合で、最も誤り率が小さかった14次元のときの結果を示している。SSGMMは、24次メルフィルタバンクに対して提案手法である話者空間においてGMMを学習した場合で、最も誤り率が小さかった4-24次元のときの結果を示している。

表 5.3の結果から、提案手法である話者空間における GMMにおいては、混合分布数が4のときに等誤り率1.79%が得られ、観測音声空間で作成された GMMに比べると、40%の照合精度の改善が得られた。判別分析においては、観測音声空間で作成された GMMに比べると、話者照合実験では照合精度の改善が得られなかった。また、提案手法は、4混合分布のときに、64混合分布の観測音声空間で作成された GMMに比べると、11%の照合精度の改善が得られた。

表 5.2および表 5.3の結果から、混合数増加による性能の変化に関して、従来法 (GMM, LDA) と提案手法 (SSGMM) では異なる傾向が見られる。まず、従来法では、音韻性の影響によりデータ分布の分散が大きく、他者との分布の重なりが多いため、混合分布数を増やすことで、他者との境界付近に識別に有効な分布が増えて、認識精度が向上していると考えられる。それに対して、提案手法 (SSGMM) では、主成分分析で得られた分散が大きく音韻性が強い低次の軸を除くことにより、分散が小さく話者性を表す高次の軸で話者空間を構成している。したがって、話者空間に学習データを射影すると、データ分布の分散が小さくなり、図 5.2のように学習データが局所的にまとまった状態で存在しているため、少ない混合分布数で話者性を十分表現することができる。また、ある一定以上に混合数を増やすと、各分布に含まれるデータ数が少なくなり、分散等のパラメータの推定精度が低下するため、認識精度が劣化していると考えられる。

表 5.4: 各空間における Fisher 比

	GMM	LDA (21dim)	SSGMM (4-24dim)
F-ratio	0.16	0.15	0.7

次に，提案手法における話者空間の性質を考察するために，従来法と提案手法における学習データの分布について調べた．学習データの分布は，各空間に学習データを射影し，話者内分散と話者間分散の比である Fisher 比で評価を行った．その結果を表 5.4 に示す．表 5.4 の *GMM* は観測音声空間（21 次元）での F 比，*LDA* は判別分析で得られた部分空間（21 次元）での F 比，*SSGMM* は提案手法における話者空間（4-24 次元）での F 比を表している．

表 5.4 の結果から，提案手法における話者空間が最も Fisher 比の大きい空間を構成していることがわかった．したがって，提案手法における話者空間は，話者内分散が小さくなり，他者との分布の重なりが削減されて，より話者間の分離度が高い空間を構成することができていると考えられる．

提案手法である話者空間への射影により，高い識別および照合精度が得られたことから，話者内分散の大きい空間を音韻空間，話者内分散の小さい空間を話者空間とみなすことで，話者空間は，音韻性を抑えた話者性を表す空間であるということがわかった．また，提案手法は，少量の学習データにより，少量のパラメータ数で頑健な話者モデルを記述でき，高速に学習，識別および照合を行えることがわかった．

5.6 識別学習による話者空間の最適化

次に，MCE/GPD による話者空間の最適化について述べる．MCE/GPD による話者空間の最適化は，話者 s の話者空間への射影行列 $P^{(s)}$ を最小誤り基準で最適化する．すなわち， k 次元話者空間への射影量を識別関数として，最小誤り基準で $P^{(s)}$ を最適化する．識別関数を

$$g_s(x; \theta) = \|P^{(s)}(x - \mu_s)\|^2 \quad (5.8)$$

として, $x \in \omega_s$ についての誤分類尺度を

$$\begin{aligned} d_s(x; \theta) &= g_s(x; \theta) - g_j(x; \theta) \\ j &= \arg \min_{i \neq s} g_i(x; \theta) \end{aligned} \quad (5.9)$$

と定義する. 誤分類尺度による損失関数をシグモイド関数を用いて, 次式のように定義する.

$$l_s(x; \theta) = \frac{1}{1 + \exp(-\xi d_s(x; \theta))} \quad (5.10)$$

ここで, ξ は, シグモイド関数の傾きを表す正定数である.

話者空間への射影行列 $P^{(i)}$ は, 次式によって損失を最小化するように, 繰り返し更新する.

$$P_{t+1}^{(i)} = P_t^{(i)} - \epsilon \nabla_{P^{(i)}} l_s(x; \theta) \quad (5.11)$$

ここで, ϵ は, 更新幅を制御する正定数である. この場合, $\nabla_{P^{(i)}} l_s(x; \theta)$ は,

$$\nabla_{P^{(i)}} l_s(x; \theta) = \begin{cases} 2l_s(1 - l_s)P^{(s)}(x - \mu_s)(x - \mu_s)^T & \text{for } i = s \\ -2l_s(1 - l_s)P^{(j)}(x - \mu_j)(x - \mu_j)^T & \text{for } i = j \end{cases}$$

となる.

本手法による話者照合は, 入力特徴ベクトルの集合 $\{x_t\}$ を式 (5.7) により最適化した話者空間に射影して $\{\hat{x}_t\}$ を求める. 次に, 申告話者 s の GMM との対数尤度 $\log P(\hat{x}_t | \lambda^{(s)})$ を求める. この対数尤度に対しては, 式 (5.12) に示す尤度比尺度に基づく正規化を行う.

$$\log \hat{P}(\hat{x}_t | \lambda^{(s)}) = \frac{1}{N} \sum_{t=1}^N \log P(\hat{x}_t | \lambda^{(s)}) - \max_{i \neq s} \frac{1}{N} \sum_{t=1}^N \log P(\hat{x}_t | \lambda_s^{(i)}) \quad (5.12)$$

登録話者 s の話者空間を最適化した際には, その話者にとって最適化された他の登録話者の話者空間も得られる. 式 (5.12) に示すように, その他登録話者の話者空間でモデル化した GMM (GMM のパラメータ $\lambda_s^{(i)}$) に対する尤度を正規化する際に用いている.

表 5.5: 分析条件

サンプリング周波数	12kHz
高域強調	$1 - 0.97z^{-1}$
フレーム長	20ms
フレーム周期	5ms
窓タイプ	ハミング窓
特徴パラメータ	24次メルフィルタバンク 18次ケプストラム

5.7 話者照合実験

提案手法の有効性を示すために、従来法である観測音声空間における GMM、最適化していない話者空間における GMM、提案手法である MCE/GPD により最適化された話者空間における GMM による話者照合実験を行った。

使用したデータは、男性 20 名・女性 10 名が約 10 ヶ月に渡る 3 時期（時期 1、時期 2、時期 3）に発声した文章データである。分析条件を表 5.5 に示す。

登録話者は男性 10 名・女性 5 名で、詐称者は登録話者と異なる男性 10 名・女性 5 名である。学習には、時期 1 で発声した 10 文章を用い、登録話者毎に時期 2 で発声した 5 文章を用いて、MCE/GPD による話者空間の最適化を行った。評価には、時期 1 から約 10 ヶ月後の時期 3 に発声した各 15 文章を 1 文章ずつ用いた。

閾値は、話者毎に本人棄却率と詐称者受理率の 2 つの誤り率が等しくなるように、事後的に設定した。

話者照合結果を表 5.6 に示す。表 5.6 の GMM は従来法である観測音声空間における GMM、SSGMM は最適化していない話者空間における GMM、SSGMM+MCE/GPD は提案手法である MCE/GPD により最適化された話者空間における GMM の結果を等誤り率で示している。

表 5.6 から、MCE/GPD により最適化された話者空間における GMM は、4 混合分布のときに等誤り率 0.62% が得られ、観測音声空間で作成された GMM に比べると、63% の照合精度の改善が得られた。さらに、最適化していない話者空間における GMM に比べると、41% の照合精度の改善が得られた。また、観測音声空間で作成された 64 混合

表 5.6: 話者照合結果 (%)

mixture number	GMM (18dim)	SSGMM (4-18dim)	SSGMM +MCE/GPD
2	2.96	1.10	0.65
4	1.7	1.05	0.62
8	1.45	1.08	0.81
16	1.32	0.98	0.74

分布の GMM での等誤り率は 0.87% であった。したがって、MCE/GPD による話者空間の最適化が有効であることがわかった。

5.8 まとめ

本章では、話者内分散の大きい空間を音韻空間、話者内分散の小さい空間を話者空間とみなして、主成分分析により音韻性と話者性を線形分離し、音韻性を抑えた話者空間上で GMM により話者モデルを学習する話者識別および話者照合法を提案した。

本手法の有効性を示すために、観測音声空間において GMM により話者識別および話者照合を行う方法、また、判別分析によって変換された空間において GMM により話者識別および話者照合を行う方法との比較実験を行った。その結果、話者識別において、本手法は 4 混合分布のときに識別誤り率 5.11% が得られ、64 混合分布の観測音声空間で作成された GMM の識別誤り率 8.89% に比べると、43% の識別精度の改善が得られた。また、話者照合において、本手法は 4 混合分布のときに等誤り率 1.79% が得られ、64 混合分布の観測音声空間で作成された GMM に比べると、11% の照合精度の改善が得られた。

したがって、話者内分散の大きい空間を音韻空間、話者内分散の小さい空間を話者空間とみなすことで、話者空間は、音韻性を抑えた話者性を表す空間であることがわかった。また、本手法は、少量の学習データにより、少量のパラメータ数で頑健な話者モデルを記述でき、高速に学習、識別および照合を行えることがわかった。

さらに、本手法で構成された話者空間は話者性を表す空間であるが、他者との相関

を考慮していない。そこで、他者との違いをより強調した空間を構成するために、最小分類誤り学習（MCE）／一般化確率的降下法（GPD）に基づく識別学習による話者空間の最適化法を提案した。

話者空間の最適化の有効性を示すために、観測音声空間で作成されたGMM、最適化していない話者空間におけるGMMとの話者照合による比較実験を行った。その結果、MCE/GPDにより最適化された話者空間におけるGMMは、4混合分布のときに等誤り率0.62%が得られ、観測音声空間で作成されたGMMに比べると、63%の照合精度の改善が得られた。さらに、最適化していない話者空間におけるGMMに比べると、41%の照合精度の改善が得られた。したがって、MCE/GPDによる話者空間の最適化が有効であることがわかった。

第6章

アナウンサー発話の自動抽出と大語彙連続音声認識によるニュース記事分類

6.1 まえがき

近年、放送の多チャンネル化により、多くのニュース番組が放映されるようになった。これを受けて視聴者には、知りたいニュースだけを見たいという要求が生じている。この要求に対応するには、ニュース記事を分類してデータベースを構築しておく必要がある。このとき、人手でニュース記事を分類することは不可能であり、機械によるニュース記事の自動分類が望まれる。この点から、ニュース音声に対する話題分類、話題同定の研究が行われている [77]–[80]。

本研究では、ニュース音声に対して大語彙連続音声認識を行い、キーワードを抽出することにより、自動的に記事分類を行うことを目的としている。ニュース音声は、レポーターあるいはインタビュアーの発話部分に比較的雑音が重畳している場合が多いので、レポーターあるいはインタビュアーの発話を正しく音声認識することは困難である。したがって、比較的雑音が少ないアナウンサーの発話区間のみを音声認識することにより、記事の分類に必要なキーワードを精度良く抽出できると考えられる。また、アナウンサーの発話のみをディクテーションすることは、処理の短縮にもつながる。この点から、本研究ではアナウンサーの発話区間のみを自動的に抽出して音声認識し、ニュース記事を自動分類する方法を提案する。

本研究では、ニュース音声からアナウンサーの発話区間を自動的に抽出する方法として、セキュリティ応用を目的に研究されている話者照合の技術を用いている。現在、話者照合は、技術的に高い照合精度が得られており、セキュリティ応用上、時期差

に対するロバストネスが研究課題となっている [73][81]。したがって、この話者照合技術は、時期差を問題としない限り、音声メディア処理として応用可能な技術である。例えば、座談会などで特定の人の発言を拾い出して聞くといった目的にも利用可能である [82]。この点に関しては、討論会において話者毎に発話区間を切り出すインデキシングの研究が報告されている [83]。この方法は、あらかじめ討論会の参加者について話者モデルを学習しておき、話者モデルを用いた話者認識により、インデキシングを行う方法である。しかし、話者モデルをあらかじめ学習しておくことは、自動化の点から望ましくない。特に、ニュース番組では、アナウンサーが日によって替ることもあり、またあらかじめアナウンサーの話者モデルを学習しておく、時期差に対処する必要がある。そこで、本研究では、特定の話をあらかじめ学習することなく、入力音声の中から各話者の話者モデルを自動学習し、話者照合に基づいて自動的に話者のセグメンテーションを行う方法を提案する。また、話者空間への射影に基づく統計的モデルとして、VQを用いた話者照合法を提案する。提案手法の有効性を示すために、NHK5分間のニュース45日分に対して、アナウンサーの発話区間の切り出し実験を行った。

ニュース記事を分類するには、記事中のキーワードを抽出し、キーワードと分類分野との関係を用いるのが一般的である。ここで問題になるのは、ニュース記事の分類性能が、ニュース音声記事からキーワードを抽出する精度だけではなく、キーワードと分類分野との関係をどのように設計するかに依存している点である。これまで、人手で与えたキーワードを基にニュース記事を分類する研究が行われてきた [84]。これに対して本研究では、 χ^2 法を用いてキーワードの自動選択を行い、選択されたキーワードと分類分野との関連度を計算した。記事の分類は、単語 bigram と不特定話者 HMM により、ニュース音声に対して音声認識することから始まる。音声認識結果から得られたキーワードをもとに、分類分野毎にその関連度を積算する。最後に、積算された関連度が最大となる分野にニュース音声記事を分類する。

6.2 アナウンサーの発話区間の抽出

6.2.1 アナウンサーの話者区間抽出法

ここでは、話者空間への射影に基づく VQ を用いた話者照合法ならびに、本手法による話者セグメンテーションについて述べる。話者セグメンテーション法を図 6.1 に示

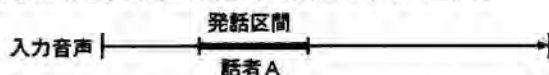
し、処理内容を以下に述べる。ここで、話者を判定する閾値は、話者モデルごとに設定している。

本研究では、異なる話者が発話したそれぞれの区間のことを話者区間と呼ぶ。同一話者がポーズをおいてしゃべっても、同一話者の区間とみなす必要がある。

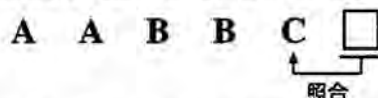
1. 入力音声の1秒毎に、1秒間の平均パワーを求め、閾値処理により音声か無音かを判定する。本研究では、閾値の設定方法は、評価データとは異なるNHK5分間のニュース5日分を用いて、無音区間と音声区間の平均パワーを求め、音声区間の抽出率が100%になるように、閾値を設定した。これにより抽出された無音から無音までの区間を発話区間と呼び、この発話区間に対して、以下の処理を行う。
2. 最初に抽出した発話区間で話者モデルを学習するとともに、話者照合のための閾値を式(6.1)により設定する。この発話区間が最初の話者区間となる。
3. 次の発話区間に対して、直前の発話区間の話者（以後、直前の話者と呼ぶ）と照合する。
 - (a) 直前の話者と判断されれば、直前までの話者区間と今抽出した発話区間を合わせて、その話者の新しい話者区間とする。この話者区間に対して話者モデルの再学習、閾値の再設定を行う。
 - (b) 直前の話者でないと判断されれば、直前の話者以外と照合し、閾値を下回った話者の中で最も距離が小さい話者に判定する。この場合、(a)と同様に、話者モデルの再学習、閾値の再設定を行う。一方、直前の話者以外と照合した結果、閾値を下回る話者がいなければ、新しい話者と判断する。今抽出した発話区間をその話者の話者区間とし、話者モデルの学習、閾値の設定を行う。

話者セグメンテーションを行う際、各話者の発話区間の継続時間は短い場合が多く、継続時間は発話区間によって大幅に変動する。発話区間の継続時間が短い場合、その話者の分散の推定精度は低くなる。また、学習データが多くなると、その話者の分散が大きくなり、話者照合を行う際に学習データの多い話者に受理されやすくなってしまふといった問題がある。つまり、話者セグメンテーションを行う際は、話者の分散を考慮しない方が話者の照合精度が高くなると考えられる。したがって、本研究では、話

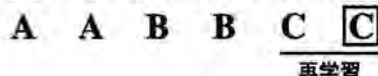
- (1) 入力音声から発話区間を抽出
 (2) 最初に抽出した発話区間で話者モデルを学習



- (3) 以後同様に発話区間を抽出し、直前の話者と照合



- (a) 受理された場合、話者モデルを再学習



- (b) 棄却された場合、それ以外の話者と照合



- 受理された場合、最も近い話者のモデルを再学習



- すべて棄却された場合、新たな話者モデルを学習

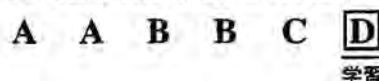


図 6.1: 話者セグメンテーション法

者空間における統計的モデルとして、VQを用いた話者照合に基づく話者セグメンテーション法を提案する。

本手法による話者照合では、話者モデルの学習において、まず話者空間を作成する。学習データを話者空間に射影し、コードブックを作成する。話者照合では、入力音声の特徴ベクトルを話者空間に射影し、最も近いセントロイドとの距離を求める。ここで、入力特徴ベクトルとセントロイドとの距離に対して、式(2.11)に示す正規化は行っていない。なぜならば、本研究では、あらかじめ話者の数や話者モデルが既知ではなく、処理の進行につれて話者モデルの数が増えるためである。さらに、直前の話者のモデルと常に照合を行うため、時期差による特徴変動の影響を考慮する必要がないからである。また、話者モデルの学習において、学習データ量が30秒を越えた場合、再学習は行わない。これは、話者の発話時間に偏りができ、学習データ量が多い特定の

話者モデルに照合されやすくなることを避けるためである。

同一話者か新しい話者かを判定する話者照合の閾値 θ は、学習に用いた音声区間中の特徴ベクトル集合と、学習時に作成された話者空間におけるセントロイドとの平均距離 μ ，ならびに標準偏差 σ を求めて、次式のように設定している。

$$\theta = \mu + w\sigma \quad (6.1)$$

この閾値は、NHK5分間のニュース5日分を設定データに用いて、重み w の値を変えて実験を行い、切り出し精度が最も高かったときの重み $w = 0.5$ として決定している。

このように最初に抽出した発話区間で話者のモデルを学習し、以後抽出した発話区間に対して、直前の話者モデルとの照合を繰り返すことで、事前に話者モデルを作成しておくことなく、話者交替の検出と話者同定を行うことが可能になる。

本研究では、ニュース音声の中のアナウンサーの話者区間を自動的に抽出することを目的としている。したがって、ニュース一日分毎に図6.1に示した方法により、話者区間を切り出すが、どの話者区間がアナウンサーの話者区間であるかを判定する基準を持っていない。そこで、話者区間を話者毎に抽出した後、最も長時間発話している話者をアナウンサーと判定している。

6.2.2 アナウンサーの話者区間抽出実験

NHKの5分間のニュース45日分を用いて、アナウンサーの話者区間を切り出す実験を行った。音声データの作成方法は、NHK5分間のニュースを8mmテープに録画し、8mmテープからSGI社のOctaneにより、チャンネルモノ、サンプリング周波数44.1kHz、量子化ビット数16ビットで音声を録音した後、サンプリング周波数12kHzにダウンサンプリングを行った。

音声区間検出のための閾値および話者空間の次元、コードブック数は、式(6.1)で求められる話者の閾値の設定データと同じで、評価データとは異なるNHK5分間のニュース5日分を用いて設定した。その結果、最も切り出し率が高かった時の値として、話者空間の次元を3-16次元、コードブック数を4に設定した。また、実験に用いた音声データの平均SNRは、20.5dBであった。実験条件を表6.1に示す。

実験の評価は、アナウンサーの話者区間の切り出し率と適合率で評価した。これらは次式で定義される。また、切り出された発話区間において、アナウンサーの発話とその他の発話が混ざった区間は、アナウンサーの発話として評価している。

表 6.1: 実験条件

データ	NHK5 分間のニュース 45 日分
サンプリング周波数	12kHz
フレーム長	20ms
フレーム周期	5ms
窓タイプ	ハミング窓
特徴パラメータ	LPC ケプストラム (16 次)
話者空間の次元	3-16
コードブック数	4
閾値 θ	$\theta = \mu + 0.5\sigma$

表 6.2: アナウンサーの話者区間切り出し結果 (%)

切り出し率	93.2
適合率	93.5

$$\text{切り出し率} = \frac{\left\{ \begin{array}{l} \text{アナウンサーとして正しく同定} \\ \text{した発話区間数} \end{array} \right\}}{\left\{ \begin{array}{l} \text{全ニュース中のアナウンサーの} \\ \text{発話区間数} \end{array} \right\}}$$

(6.2)

$$\text{適合率} = \frac{\left\{ \begin{array}{l} \text{アナウンサーとして正しく同定} \\ \text{した発話区間数} \end{array} \right\}}{\left\{ \begin{array}{l} \text{アナウンサーとして切り出された} \\ \text{発話区間数} \end{array} \right\}}$$

(6.3)

実験に用いた NHK5 分間のニュース 45 日分すべてに、レポータあるいはインタビュアーが含まれている。また、5 分間のニュース 45 日分に含まれる総記事数は、136 記事であり、その中でレポータあるいはインタビュアーを含む記事は 48 記事である。ニュース音声中的アナウンサーの数は 1 名であった。

アナウンサーの話者区間切り出し実験を行った結果を表 6.2 に示す。アナウンサーの話者区間の切り出し率、適合率はそれぞれ 93.2%、93.5% と高い値を得ることができた。

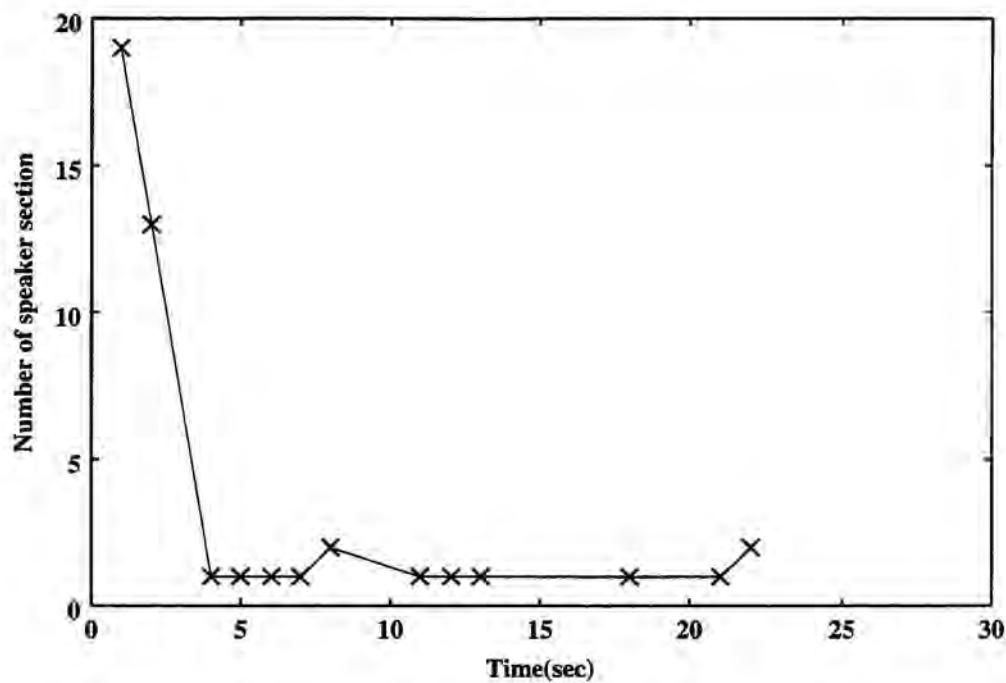


図 6.2: 最初の話者区間の継続時間分布

これは、アナウンサーの話者モデルを1日のニュース毎に自動作成していることから、時期差の問題を回避できたためと考えられる。本手法は、最初の話者区間の継続時間がアナウンサーの発話区間の抽出精度に大きく依存する。したがって、最初の話者区間（45 日分）の継続時間が問題になるので、この分布を図 6.2 に示す。図 6.2 より全体の 75% は、最初の話者区間の継続時間が 5 秒以下であり、かなり短い区間が多く含まれていることがわかる。また、抽出された音声区間の継続時間の分布を図 6.3 に示した。音声区間の平均継続時間は、11 秒であった。

さらに、最初の話者区間の継続時間に対する本手法の性能の依存性を調べるために、最初の話者区間の継続時間を擬似的に 80, 60, 40, 20% にそれぞれ削減した場合のアナウンサーの発話区間抽出率を図 6.4 に示す。削減していない場合は抽出率が 93.2% であるのに対して、最初の話者区間の継続時間を 80% に削減すると抽出率が 85.9% となった。

切り出し誤りが生じる原因としては次のような場合があった。各話者の話者空間の学習に用いる音声データが短かすぎると、一般に閾値が低く設定されるため、以後その話者の音声は棄却されやすくなる。また、話者空間の作成後、話者照合のための発話区間が短い場合にも、同一話者の発話を棄却したり、異なる話者の発話を同一話者

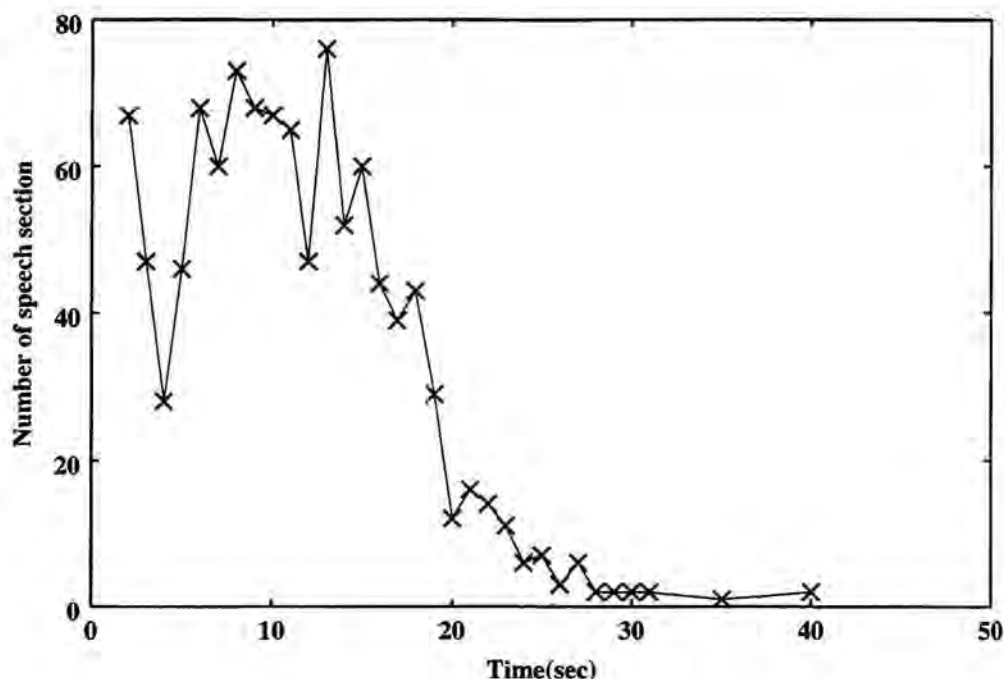


図 6.3: 音声区間の継続時間分布

として受理しやすくなる傾向がある。さらに、発話区間に雑音が重畳すると、話者照合を誤る場合があった。

6.3 大語彙連続音声認識実験

6.3.1 実験条件

用いた言語モデルは、語彙サイズ 20,000 の back-off bigram で、毎日新聞 CD-ROM 版の 45ヶ月分 (91 年 1 月～94 年 9 月) の記事から学習した。Back-off smoothing には wittenbell の推定を用い、bigram に対する cut-off は 1 とした。

音響モデルは、男性不特定話者 HMM で、単語間の音素文脈依存も考慮した cross-word triphone モデルである。学習には、日本音響学会新聞記事読み上げ音声コーパス (JNAS) のうち、男性話者 137 人分の 21,782 発話を用いた。音響特徴量には 39 次元の MFCC 特徴パラメータ (12 次元のメルケプストラム係数とパワー、およびそれぞれの Δ 、 $\Delta\Delta$ 係数) を用いた。実験条件を表 6.3 に示す。

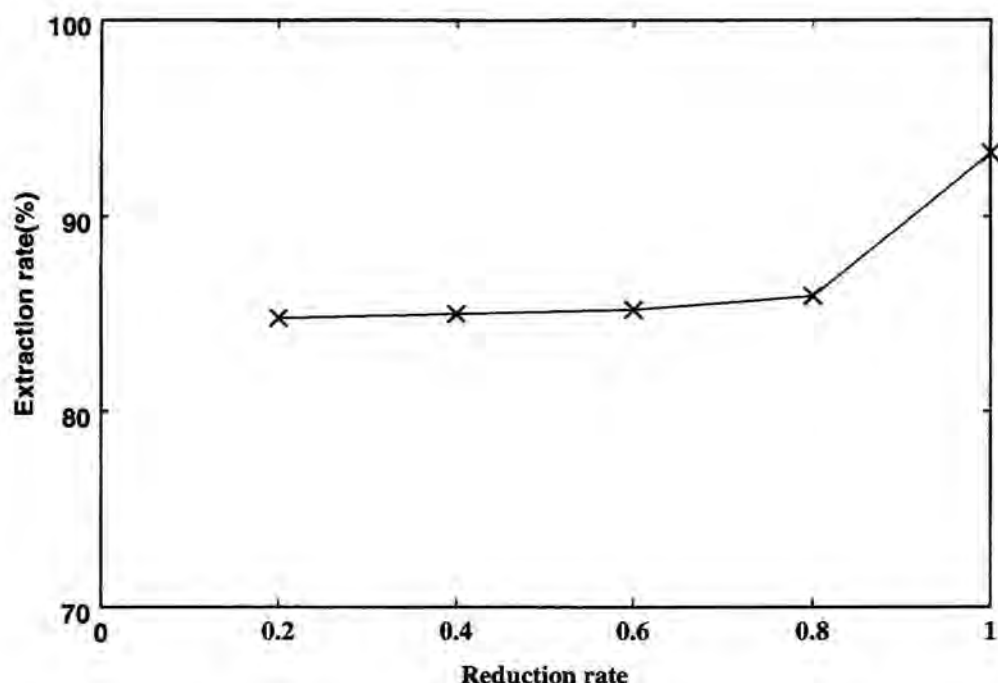


図 6.4: 最初の話者区間の継続時間削減率と発話区間抽出率

評価用音声データは、NHK5分間のニュース45日分からレポートあるいはインタビュアーを含む48記事を用いた。この48記事に対して評価したアナウンサーの話者区間切り出し率は97.4%、適合率は82.9%であった。こうして切り出されたアナウンサーの発話に対して大語彙連続音声認識を行う。

6.3.2 実験結果

大語彙連続音声認識の評価用データは、話者区間抽出実験により切り出されたアナウンサーの話者区間(Anchor)、アナウンサー以外のインタビュアーやレポート等の話者区間(Other)、AnchorとOther区間を両方ともに含んだ全区間(All)の3セットである。この評価用データ3セットのテストセットパープレキシティを表6.4に示す。

表6.4より、アナウンサーの話者区間におけるテストセットパープレキシティは127.4であり、アナウンサー以外の話者区間におけるテストセットパープレキシティ234.5に比べてかなり低いことがわかる。これは、アナウンサーの発話がニュース原稿を正確に読み上げているのに対して、レポートやインタビュアーの発話は、原稿の正確な読み上げとは異なっていることを示している。

表 6.3: 音響分析とHMM

音 響 分 析	サンプリング周波数	12kHz
	特徴パラメータ	MFCC (39次元)
	フレーム長	20ms
	フレーム周期	5ms
	窓タイプ	ハミング窓
H M M	状態数	5状態3ループ
	タイプ	Triphone HMM
	混合数	8
	学習方法	連結学習
	発話数	21,782文
	発話者数	男性137人

前節で述べた音響モデルと言語モデルを用いて、ニュース音声に対して beam-search を用いたビタピデコーディングを行った。デコーダーには HTK(HMM Toolkit)[76] を用いた。実験結果を表 6.5 に示す。表 6.5 は、アナウンサーの話者区間(Anchor)、アナウンサー以外のインタビュアーやレポーター等の話者区間(Other)、Anchor と Other 区間を両方ともに含んだ全区間(All) に対する音声認識結果を示しており、単語誤り率、単語正解率、単語正解精度で評価を行った。

資料として用いた音声データは平均 SNR20.5dB で比較的 background 雑音が多く、全体的に認識精度は低い値となった。特に Other セットのインタビュアー等の区間では、平均 SNR11.2dB で特に background 雑音が多く、Anchor 区間の平均 SNR29.8dB に対して平均 SNR が約 63% 低かった。また、テストセットパープレキシティも高い値を示し、アナウンサーに対する音声認識結果に比べて単語正解精度はかなり低い結果になった。

6.4 ニュース記事分類

6.4.1 記事分類の概要

毎日新聞記事の形態素解析結果である RWC テキストデータベース中の 93 年度の新聞記事 10,000 件を、92 年度の朝日新聞記事データベース分類表索引 [85] を用いて分類

表 6.4: 評価用ニュース音声データ

	発話数	20K 未知語率	perp
Anchor	250 発話	10.8%	127.4
Other	113 発話	24.3%	234.5
All	363 発話	17.6%	178.6

perp: test-set perplexity

表 6.5: ニュース音声認識実験の結果 (%)

	単語誤り率	単語正解率	単語正解精度
Anchor	33.6	75.1	66.4
Other	77.2	39.4	22.8
All	45.6	65.8	54.4

した。この分類表索引には、約1万2千のキーワードが載っており、キーワード毎に対応する分類分野が付与されている。この分類表索引には、大分類として総類，政治，経済，労働，文化，科学，社会，事件，スポーツ，国際の10分野が設定されている。ニュース記事分類の実験では、この10分野に対して記事の分類を行った。

10,000 記事の正解分野の決定方法は次の通りである。10,000 記事に含まれるキーワードのうち、朝日新聞記事データベース分類表索引に含まれるキーワードを抽出する。このキーワードから、文献 [86] に記述されている方法でキーワードの各分野に対する寄与率を求める。次に、キーワードの音響尤度を 1 として分類確率を計算し、分類を行っている。この分類結果を用いて、キーワード選択、キーワードと分類分野の関連度計算を行った。10,000 記事の正解分野を手で行うことも可能であるが、その場合においても、10,000 件の正解分野が少し異なるだけで、キーワード選択と関連度の計算法はそのまま用いることができる。

分類した 10,000 記事を学習データとして、この中から名詞だけを抜き出し、この名詞からキーワードを選択するとともに、キーワードと分野間の関連度を求めた。次に、選択したキーワードと関連度に基づいて、ニュース音声記事の大語彙連続音声認識結果に対して分類実験を行った。

6.4.2 新聞記事からのキーワード選択法

新聞記事中の単語 w_i と、その記事が属する分野 t_j との間で関連度 γ_{ij} を求め、閾値処理を行って、関連度 γ_{ij} の大きいものをキーワードとして抽出する [87]。この関連度 γ_{ij} として、本研究では χ_{ij}^2 値を用いている。

χ^2 検定で用いられる χ^2 値は、分野における単語の偏りを示す指標として利用できる。 χ^2 値の計算方法は、まず、各単語 w_i について分野 t_j における頻度 x_{ij} を求める。次に、単語 w_i の出現確率は全分野を通じて等しいという仮説を設定し、この仮説に基づいて単語 w_i の分野 t_j における予測頻度 m_{ij} を式 (6.5) により計算する。この x_{ij} と m_{ij} を基に、式 (6.4) に従って、 χ_{ij}^2 値を求める。もし、この χ_{ij}^2 値が十分大きな値になれば特定の分野に偏って表われる単語ということになり、分野の識別に有効な単語と見なすことができる。別の見方をすれば、分野 t_j との共起が高い単語 w_i を、キーワードとして選んでいると解釈できる。

$$\chi_{ij}^2 = \frac{(x_{ij} - m_{ij})^2}{m_{ij}} \quad (6.4)$$

$$m_{ij} = \frac{\sum_{j=1}^n x_{ij}}{\sum_{i=1}^m \sum_{j=1}^n x_{ij}} \times \sum_{i=1}^m x_{ij} \quad (6.5)$$

ここで、 m は異なり単語数、 n は分野数、 x_{ij} は単語 w_i の分野 t_j における頻度、 m_{ij} は単語 w_i の分野 t_j における予測頻度を表している。

6.4.3 記事の分類方法

1つの記事が与えられると、記事中のキーワード w_i を全て抽出する。このキーワード w_i と分野 t_j との関連度 γ_{ij} を基に、キーワード w_i が分野 t_j の分類に寄与する割合（分類寄与率） C_{ij} を式 (6.6) で計算する。例えば、キーワード「日米」が分野「政治」、「経済」、「社会」に対して表 6.6 のような関連度を持っている場合、分類寄与率 C_{ij} は表 6.6 の右欄のように表わされる。

$$C_{ij} = \frac{\gamma_{ij}}{\sum_j \gamma_{ij}} \quad (6.6)$$

表 6.6: キーワード「日米」の3つの分野に対する関連度と分類寄与率の例

分野	関連度	分類寄与率
政治	30	0.6
経済	20	0.4
社会	0	0.0

この方法は、キーワード w_i の分野 t_j に対する関連度を、キーワード間で比較可能にする効果がある。

最後に、1つの記事中に含まれているキーワードの出現回数を N_i とすると、分類寄与率 C_{ij} を次式のように総和して、記事 x と分野 t_j との類似度を求める。

$$S(x, t_j) = \sum_i N_i \cdot C_{ij} \tag{6.7}$$

この類似度の大きな分野 t_j に記事を分類する。

6.5 ニュース音声記事の分類実験

6.5.1 実験条件

自動抽出されたアナウンサーの話者区間に対して大語彙連続音声認識を行い、その結果を基に記事分類を行った。また、比較として、インタビュアー、レポーター等も含んだ場合についても記事分類を行った。記事分類に用いた48記事における分野の分布を表6.7に示す。記事分類に用いた48記事には、10分野のうち科学と文化を除く8つの分野が含まれていた。実際のニュースにおいては、政治、経済といった分野が多く存在していることから、本研究で記事分類に用いた48記事の分野の分布は、実際のニュースの分野の分布を反映していると考えられる。

キーワードと分類分野との関連度の学習には、94年の毎日新聞記事データを用いている。記事分類を行う際には、新聞記事データ中の単語に対して分野数だけ χ^2 値を求め、分野毎にある一定の閾値以上の χ^2 値をもつ単語をキーワードとして選択した。

表 6.7: 記事の分野の分布

分野	頻度	分野	頻度
政治	19	総類	3
経済	9	社会	2
国際	6	スポーツ	2
事件	6	労働	1

6.5.2 実験結果と考察

ニュース音声記事の分類結果を表 6.8 に示す。表には、 χ^2 値が 80 以上の値を持つキーワード 1376 単語を用いた記事分類率を示している。表中の分類率とは、記事の総数に対して正解分野に分類された記事数の割合である。正解分野は、NHK5 分間のニュース 45 日分から選んだ 48 記事に対して人手で付与した。

表 6.8 より、アナウンサーの話者区間のみで記事分類を行なった結果、単語正解精度 66.4% のとき記事分類率 70.8% を得た。比較として行った All セットに対する記事分類率は 66.7% であり、アナウンサーの話者区間のみで記事分類した方が、約 4% 程高い結果を得た。しかし、この差は 1 記事分の差であり、有意な差であるとは言えず、2 つのデータセットに対する記事分類率はほぼ同じであると考えた方がよい。これには次のような理由がある。(1) 評価データである 48 記事に対するアナウンサーの話者区間切り出し率が 97.4%、適合率 82.9% という結果から、レポートあるいはインタビュアーの区間を誤ってアナウンサーの区間と判定してしまった場合が多く、Anchor に対する記事分類率が低下した。(2) アナウンサーと比べてレポートあるいはインタビュアーの区間が少ないため、All に対する記事分類率が低下しない。(3) レポートあるいはインタビュアーの区間に対するディクテーションの精度は低く、記事分類に貢献するキーワードが抽出できていない。

以上のことから、Anchor に対する記事分類率を向上させるためには、話者区間の切り出し精度を更に向上させる必要がある。また、キーワードの選択や関連度の計算には、新聞記事 10,000 件に対して正解分野を機械的に付与したのに対し、ニュース音声記事 48 件に対しては正解分野を手で与えている。この新聞記事 10,000 件から抽出されたキーワードは、機械的に付与した正解分野に依存している。一方、ニュース音声記事 48 件の分類結果の評価では、人手で付与した正解分野と比較している。このように、

表 6.8: ニュース音声記事の分類結果 (%)

	単語正解率	単語正解精度	記事分類率
Anchor	75.1	66.4	70.8
All	65.8	54.4	66.7

キーワードの選択と選択されたキーワードを用いた記事分類では、新聞記事とニュース記事、機械的に付与した正解分野と人手で付与した正解分野といった違いがある。したがって、新聞記事から選択したキーワードは、必ずしもニュース音声記事の分類にベストなキーワードではない。このため分類精度には上限ができ、2つのデータセットに対する分類精度に差が出なかったものと考えられる。

6.6 まとめ

本研究では、特定の話者をあらかじめ学習することなく、入力音声から各話者の話者モデルを自動学習し、話者照合に基づいて自動的に話者のセグメンテーションを行う方法を提案した。また、話者空間への射影に基づく統計的モデルとして、VQを用いた話者照合法を提案した。NHK5分間のニュース45日分に対して、アナウンサーの話者区間抽出実験を行った結果、アナウンサーの話者区間切り出し率93.2%、適合率93.5%が得られ、本手法が有効であることがわかった。

また、自動抽出されたアナウンサーの話者区間に対して、大語彙連続音声認識を行い記事分類を行った。記事分類を行う際、 χ^2 法によりキーワードを自動選択し、キーワードと分類分野との関連度を求め、この関連度により記事の分類を行った。その結果、インタビュアー等を含めた場合に比べ、分類率を落とすことなく、処理時間を短縮できる可能性があることがわかった。実験で用いたニュースデータには、アナウンサーの話者区間にも背景雑音が重なっているものもあった。そのようなデータについては、比較的明瞭に、また文法的にも正しいアナウンサーの話者区間であっても、低い認識結果となった。今後、そのような雑音下におけるロバストな音声認識が課題として残されている。

第7章

話者認識と音声認識を統合した話者の人名付与システム

7.1 まえがき

放送がデジタル化されると、放映された番組をハードディスクに録画しておくことができるホームサーバーが、各家庭に設置されるようになる。そうすると、ユーザは、ホームサーバーから見たい番組や、その中の関心のある場面を検索したり、番組の内容を要約して見たいといった要求を持つようになる。この要求に対応するには、番組の内容に対して、音声や映像といったメディアの解析を通して、索引情報を自動付与しておく必要がある。

これまでニュース番組を対象に、音声や映像メディアを解析して、ニュース番組の構造化ならびに検索に関する研究が行われてきた[88]-[90]。ニュース番組の構造化に関しては、アナウンサー発話の自動抽出と大語彙連続音声認識を統合化したニュース記事の自動分類手法を提案した[88]。これを踏まえて、本研究では、「クリントン大統領が、情報スーパーハイウェイについて話しているシーンを見たい」といった、特定の話者がある話題について話しているシーンの検索を目的としている。この処理を実現するためには、話者の交替を検出して、各話者の名前を付与し、大語彙連続音声認識により重要語を抽出しておく必要がある。

ニュース音声や対話における話者セグメンテーションおよびクラスタリングの研究がさかんに行われている[91]-[94]。これまで、ニュース音声の中のアナウンサー発話の自動抽出[82]、座談会やドラマに対する話者セグメンテーション[82][95]に関する研究を

行ってきた。これらの研究では、特定の話者をあらかじめ学習することなく、入力音声中から話者モデルを自動学習し、部分空間法による話者照合 [96] に基づいて自動的に話者のセグメンテーションを行っている。これに対して本研究では、6.2.1節で述べた話者セグメンテーション法をベースとして、話者の学習データ量に応じて話者判定の閾値を自動設定する方法を提案する。

人名の付与に関しては、ニュース映像中の顔の名前付けを自動的に行う Name-It という手法が提案されている [97]。Name-It は、ニュース映像中から顔を検出するとともに、クロズドキャプションならびにテロップから名前を抽出し、両者を対応付ける手法である。しかし、ニュースでは顔が写っているからといって、その話者がしゃべっているとは限らない。そこで、本研究では、大語彙連続音声認識とワードスポッティングによりニュース音声中から人名を抽出し、対応する話者の名前を付与する方法を提案する。これは、ニュース音声に対して大語彙連続音声認識を行い、人名および話者の交替を促すフレーズを抽出する方法である。本研究では、話者の交替を促すフレーズをキーフレーズと呼ぶ。この際、大語彙連続音声認識では、辞書にない未知語を認識することができないため、あらかじめ用意された人名辞典により、人名スポッティングを新たに行う [98]。

したがって、本研究では、話者セグメンテーションにより自動構築された話者モデルに対して、大語彙連続音声認識とワードスポッティングにより抽出した話者の名前を付与する。この人名インデキシングと、大語彙連続音声認識により得られた重要語を基に、話者と発話内容を同時検索することが可能になる。本手法により、98年8月のNHK15分間のニュースに対して、話者セグメンテーション、大語彙連続音声認識、人名スポッティングを行い、人名インデキシング実験を行った。

7.2 話者セグメンテーション

7.2.1 話者セグメンテーション法

本研究では、6.2.1節で述べた話者セグメンテーション法を用いた。本研究で用いたニュース音声は、6.2.2節で用いたデータに比べて、無音区間が短く話者交替が頻繁に起こっているため、1秒毎に平均パワーを求めると発話区間が抽出できない場合があった。したがって、本研究では、0.5秒毎に平均パワーを求め、閾値処理により音声か無

音かを判定して、発話区間の抽出を行った。

同一話者か新しい話者かを判定する話者照合の閾値 θ は、式(6.1)における重み w を式(7.1)のように設定している。

$$w = \frac{1}{\alpha \log T} \quad (7.1)$$

式(7.1)中の T は、話者の学習データ量(秒)を表している。6.2.1節で述べた手法では、学習データ量を考慮せずに閾値を設定していた。したがって、学習データ量が少ない場合、学習話者のモデルが話者性を十分表現できていないため、閾値が小さくなり本人を誤って棄却しやすくなってしまう。また、学習データ量が多くなってくると、学習話者のモデルが話者性を十分表現することができ、閾値が大きくなり他人を誤って受理しやすくなってしまう。それに対して、本手法は、閾値 θ が重み w によって、学習データ量が少ない場合に大きくなり、本人を受理しやすくなる。また、学習データ量が多い場合に小さくなり、他人を棄却しやすくなる。このように、話者の学習データ量に応じた照合が可能になる。

また、式(7.1)中の係数 α は、NHK15分間のニュース2日分を設定データに用いて、値を変えて実験を行い、話者交替検出の精度が最も高かったときの係数 $\alpha = 0.6$ として決定している。

7.2.2 話者セグメンテーション実験

98年8月のNHK15分間のニュース4日分を用いて、話者セグメンテーション実験を行った。以後、このニュース音声に対して、大語彙連続音声認識、ワードスポッティング、人名インデキシングの実験を行った。音声の分析条件を表7.1に示す。

本実験では、音声区間検出のための閾値および話者空間の次元、コードブック数は、式(7.1)の係数 α を決定するのに用いた設定データと同じで、評価データとは異なるNHK15分間のニュース2日分を用いて設定した。その結果、最も話者交替検出率が高かった時の値として、話者空間の次元を6-18次元、コードブック数を4に設定した。

実験の評価は、話者交替検出率・適合率、話者同定率で評価した。これらは、次式で定義される。

表 7.1: 分析条件

サンプリング周波数	12kHz
高域強調	$1 - 0.97z^{-1}$
フレーム長	20ms
フレーム周期	5ms
窓タイプ	ハミング窓
特徴パラメータ	24次メルフィルタバンク 18次ケプストラム

表 7.2: 話者セグメンテーション結果

	話者境界数	%
話者交替検出率	57 / 71	80.3
話者交替適合率	57 / 71	80.3
	発話区間数	%
話者同定率	440 / 493	89.2

$$\text{話者交替検出率} = \frac{\text{正しく検出した話者間の境界数}}{\text{全ニュース中の話者間の境界数}}$$

(7.2)

$$\text{話者交替適合率} = \frac{\text{正しく検出した話者間の境界数}}{\text{検出した話者間の境界数}}$$

(7.3)

$$\text{話者同定率} = \frac{\text{正しく話者を同定した発話区間数}}{\text{全ニュース中の発話区間数}}$$

(7.4)

話者セグメンテーション実験を行った結果を表 7.2 に示す. 表 7.2 の結果から, 話者交替検出率 80.3%, 適合率 80.3%, 話者同定率 89.2% が得られ, 本手法による有効性が確認できた. 本研究では, 話者交替の検出において, 一発話分でも話者の境界がずれて検出された場合, 検出誤りと判断している. 話者の照合においては, 新たな話者の学習データが少なすぎると, 以後棄却されてしまったり, 話者が交替しているにも関わらず, 発話区間が短いために受理されてしまうといった誤りがあった.

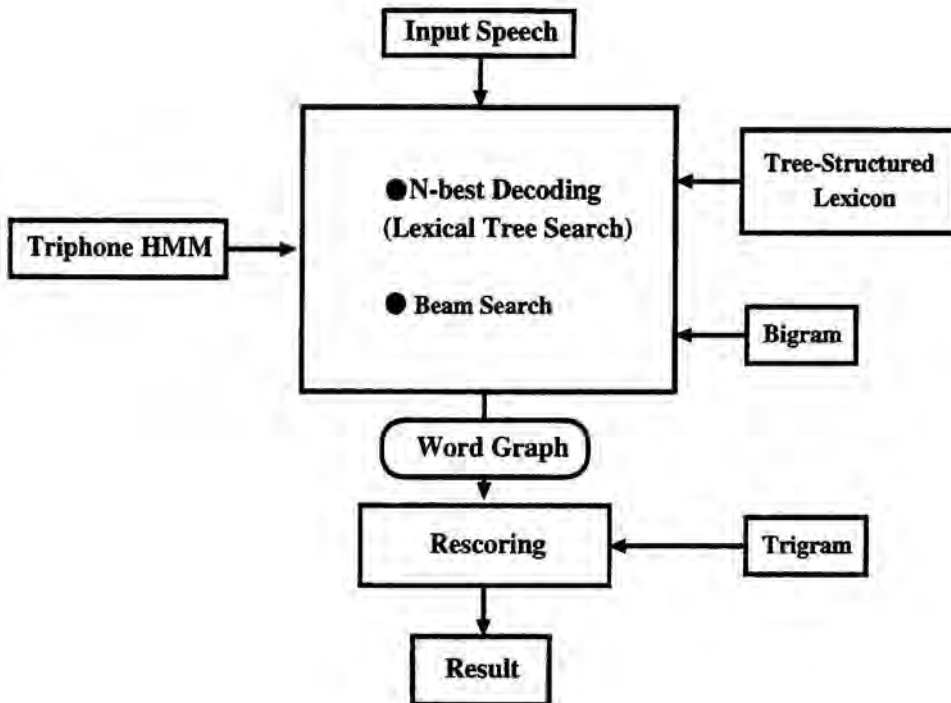


図 7.1: 認識システム

7.3 大語彙連続音声認識

7.3.1 認識システムの構成

大語彙連続音声認識システムとしては、ワードグラフを中間結果とする2-pass 構成のシステムを用いている(図7.1)[99]。まず、1st-passで単語 bigram を用いた lexical tree searchを行ない、認識結果をもとにワードグラフを作成する。このとき、最もスコアの高い単語にのみ back-off 接続を行う、最尤単語 back-off 接続を用いることにより、認識精度を落すことなく処理時間を大幅に削減している[100][101]。2nd-passでは、ワードグラフに登録された1st-passの音響尤度と trigram を用いてリスコアリングを行なう。

7.3.2 大語彙連続音声認識実験

音響モデルには、前後の音素環境を考慮した triphone HMMを用いた。音響モデルの学習には、まず ATR 連続音声データベース a~j セットの6名分のデータの視察ラベ

表 7.3: 音響分析とHMM

音 響 分 析	サンプリング周波数	16kHz
	特徴パラメータ	MFCC (39 次元)
	フレーム長	20ms
	フレーム周期	10ms
	窓タイプ	ハミング窓
H M M	状態数	5 状態 3 ループ
	タイプ	triphone
	混合数	12
	学習方法	連結学習

ルを用いて初期モデルを作成し，次に日本音響学会新聞記事読み上げコーパス (JNAS) のうち，男性話者137名分の21782発話を用いて連結学習を行った．音響分析条件とHMMのトポロジーを表7.3に示す．

言語モデルには，IPAモデル98年度版のうち，語彙サイズ20K，cut-offはbigram，trigramそれぞれに対して4-4のモデルを用いている．言語モデルの学習データは，毎日新聞記事75ヶ月分である[102]．評価用データは，98年8月のNHK15分間のニュース4日分(7.2.2節と同様)である．

大語彙連続音声認識の実験結果を表7.4に示す．ここで，"Anchor"はアナウンサー発話のみを認識した結果，"Other"はアナウンサー以外のインタビュアーやレポート発話を認識した結果を示している．全体的に認識率は低くなっているが，まずAnchor部分に関しては，言語モデルの学習データの時期的な差異などが原因と考えられる．また，アナウンサー以外の部分(Other)においては，インタビュアーなどの完全な自由発話や外国語発話なども多く含まれており，そのような発話はほとんど認識不能であったため，認識率が低下している．

7.3.3 大語彙連続音声認識による人名，キーフレーズ検出

本研究における話者交替検出，人名インデキシングシステムにおいては，音声認識により得られた人名や話者交替を促すフレーズ（キーフレーズ）が重要となる．そこで，本研究での大語彙連続音声認識システムにおいて，人名，キーフレーズがどの程

表 7.4: 大語彙連続音声認識結果 (%)

	単語正解率	単語正解精度
Anchor	80.1	78.4
Other	38.7	35.4

表 7.5: キーフレーズリスト

次のように述べています
次のように述べました
次のように発言しています
次のように発言しました
お聞きください
考えを示しました
非難しました
語りました
反論しました
考えを明らかにしました
述べました
見通しを示しました

度検出されたかを調べた。なお、人名やキーフレーズは、アナウンサーだけでなく、レポータも発話しているので、アナウンサーのみではなくレポータの発話も含めた音声認識結果から、人名とキーフレーズの検出を行った。

検出対象の人名としては、比較的著名な 50 人分の名前リスト (例えば、クリントン大統領, 野中官房長官...) をあらかじめ用意した。実際に用いたニュースデータ中で出現した人名は、50 名のうち 23 名であった。また、23 名分の人名が出現した回数は、全部で 82 回であった。

キーフレーズとしては、予備調査として評価データ以外のニュース中において、実際に話者交替時に検出されたもの 12 種類 (表 7.5) を用意した。実際の評価データ中で出現したキーフレーズの種類数は、12 種類 (あらかじめ用意したもの全て) であった。また、12 種類のキーフレーズが出現した回数は、全部で 24 回であった。

表 7.6: 人名, キーフレーズの検出結果

	検出率
人名	41.5%(34/82)
キーフレーズ	79.2%(19/24)

人名, キーフレーズの検出結果を表7.6に示す. 実験結果より, キーフレーズは比較的高い検出率を得たが, 人名の検出率は非常に低い結果となった. この原因としては, 言語モデルの時期差の関係が考えられる. すなわち, 言語モデルの学習において, 人名は言語モデル学習データの時期に強く依存するのに対し, キーフレーズはどの年代のニュースにも独立に出現する. また, 人名, キーフレーズとも, レポーターなどの背景雑音が比較的多い区間に関しては, 誤認識が多かった.

7.4 人名スポッティング

7.4.1 キーワードスポッティングによる人名の検出

ここでは, 人名検出の改善法について述べる. 先に述べたように, N -gram 言語モデルを用いた大語彙連続音声認識では, 人名においては言語モデルの学習データの時期差に強く依存するといった問題(未知語の問題)がある. そこで, 本研究においては, あらかじめ用意した人名リストを用いて人名スポッティング(キーワードスポッティング)を行い, それと7.3.3節の大語彙連続音声認識結果を併用することで人名検出の改善を行う.

人名スポッティングは, サブワードデコーダーの併用によるリジェクションを用いることで行う[103][104]. 具体的には, 図7.2に示す(a)(b)2種類の言語モデルを用いる. (a)はキーワード(ここでは人名に相当)とフィラーモデルにより構成されており, キーワードモデルと呼ぶ. フィラーモデルは, キーワード以外の区間を近似するもので, ここでは任意の音素を用いている. (b)は音素モデルの出現のみを許したもので, ここではバックグラウンドモデルと呼ぶ. (b)を用いたデコーディングは, サブワードデコーダーに相当する.

入力音声に対して, 2種類の言語モデルを利用してそれぞれビタビアルゴリズムに基

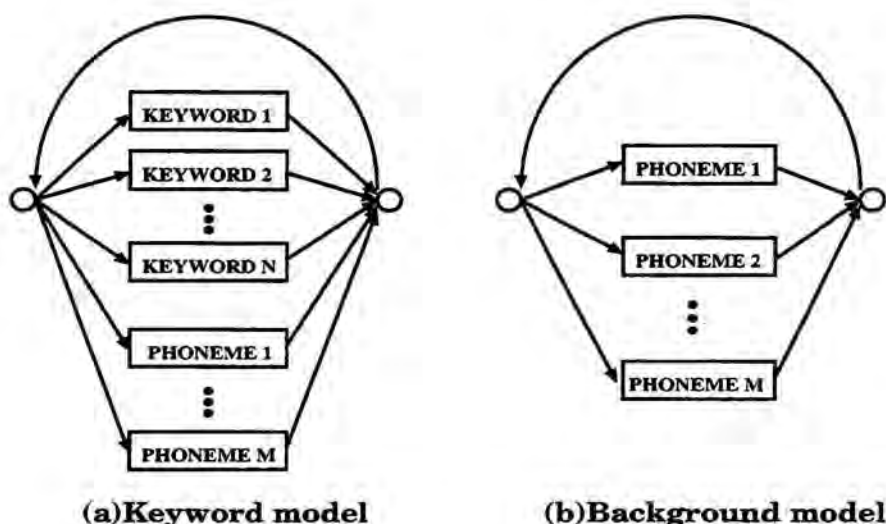


図 7.2: キーワード スポッティング 用 言語モデル

づき、次式のように尤度を算出する。

$$S_{KW}(W) = \frac{\text{単語 } W \text{ が認識されたときの対数尤度}}{\text{フレーム数}} \quad (7.5)$$

$$S_{BA}(W) = \frac{\text{単語 } W \text{ に対応する音素列の対数尤度}}{\text{フレーム数}} \quad (7.6)$$

ここで $S_{KW}(W)$ は、キーワードモデル中の単語 W が認識されたときの対数尤度を、そのときのフレーム数（図 7.3 で $f_{e-KW} - f_{s-KW}$ ）で割ったものであり、 $S_{BA}(W)$ はバックグラウンドモデルでの音素列の対数尤度を、単語 W に対応するフレーム数（図 7.3 で $f_{e-BA} - f_{s-BA}$ ）で割ったものである。図 7.3 に単語 W が認識されたときの各尤度とフレーム数の対応を示す。

湧き出し単語のリジェクションは、式 (7.7) により求めた対数尤度差 $S(W)$ が、ある一定の閾値以下の値のものに対して行った。本研究では、キーワード スポッティングにより検出された人名の湧き出しが多いと、誤って人名を付与してしまう可能性がある。そこで、実験的にリジェクションの閾値を変化させて、湧き出し誤り率が少ない場合の値を閾値として設定した。

$$\text{対数尤度差 } S(W) = S_{KW}(W) - S_{BA}(W) \quad (7.7)$$

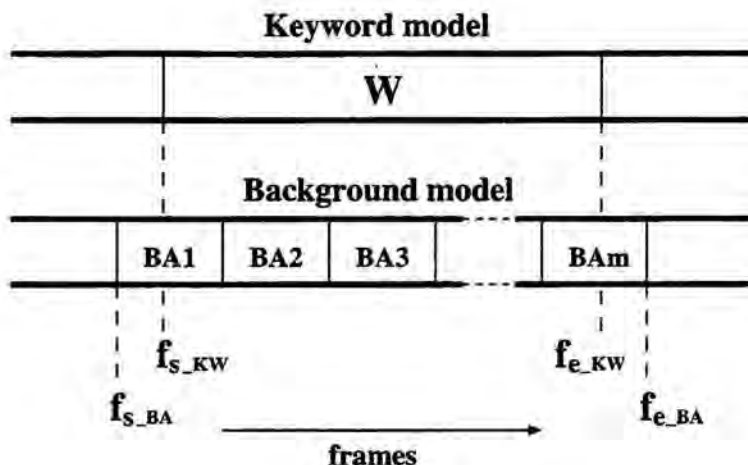


図 7.3: 2つのモデルでの尤度区間の対応

7.4.2 人名スポッティング実験

音響モデルには、音素環境独立な HMM(monophone HMM)を用いた。音素数は、無音を含む 41 種類である。学習データは、日本音響学会新聞記事読み上げコーパス (JNAS) のうちの 21782 発話 (男性話者 137 名分) である。音響分析条件等は、7.3.2 節で述べた音響モデルと同様である。

評価用データは、98 年 8 月の NHK15 分間のニュース 4 日分 (7.2.2 節と同様) で、継続時間長は 1.05 時間である。スポッティング対象単語 (人名) の出現回数は 82 回、語彙サイズは 50 である。

人名スポッティングの結果を図 7.4、表 7.7 に示す。ここで、図 7.4 中の A/W は検出率、 $FA/KW/H$ は単位時間当たりの湧き出し誤り率を表しており、湧き出し単語のリジェクションの閾値を変化させた時の、検出率と湧き出し誤り率を示している。ここでも、大語彙連続音声認識実験と同様で、レポーターやインタビュアーなどの背景雑音が比較的多い発話に関しては、精度が低く (湧き出し誤りが多く) なっていた。

後述する人名インデキシングにおいては、これまでに述べた話者交替やキーフレーズ、そしてここで述べた人名スポッティング結果を併用するため、人名スポッティング時に起こった湧き出し誤り単語を、ある程度抑制することが可能である。しかしながら、湧き出し誤り単語が大量に起こった場合、そのような方法でも人名インデキシングに悪影響を及ぼすと考えられる。したがって、人名スポッティングにおける湧き

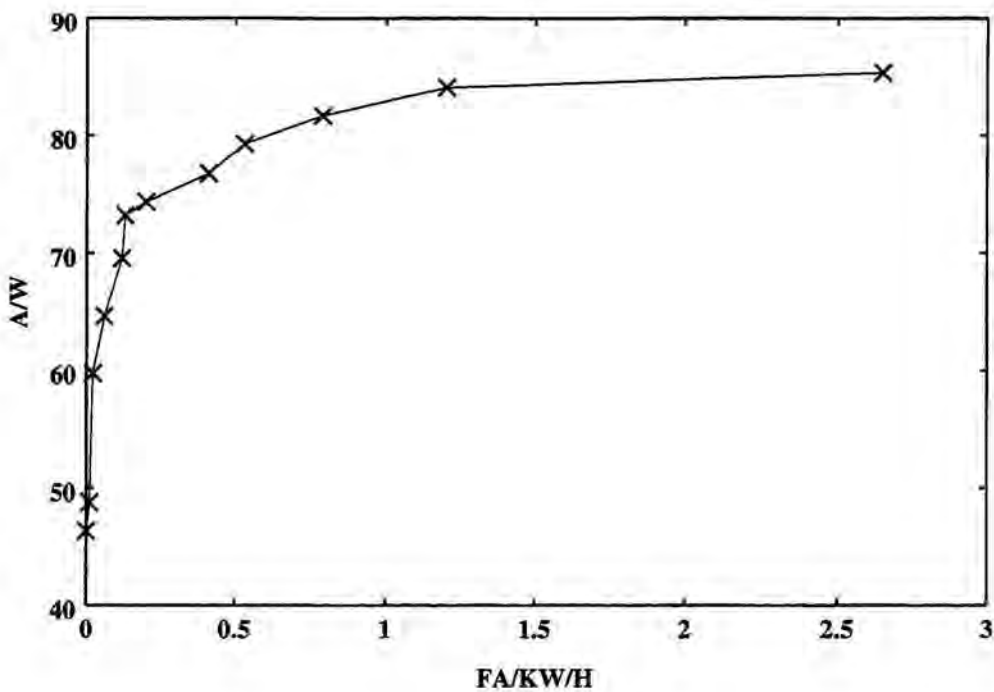


図 7.4: 人名スポッティング結果 (1)

出し誤り単語は、ある程度少ない方が望ましい。この点から、検出率 79.3%，湧き出し誤り率 0.5% のときの結果を人名スポッティング結果として、以降の実験で用いることにする。

大語彙連続音声認識結果と人名スポッティング結果との併用による人名検出実験を行った。すなわち、大語彙連続音声認識においては、個々の人名が誤って湧き出すことはほとんどないと考えられることから、大語彙連続音声認識結果と人名スポッティング結果から、どちらか一方に出現していたら、その人名の検出を行うようにする。表 7.8 に、大語彙連続音声認識結果のみ (a)，人名スポッティングのみ (b)，そして両方の

表 7.7: 人名スポッティング結果 (2)

検出率 (%)	湧き出し 誤り率 (%)
85.4	2.7
79.3	0.5
46.3	0.0

表 7.8: 人名スポッティング結果 (3)

手法	検出率 (%)
(a)	41.5
(b)	79.3
(c)	81.7

併用 (c) による人名検出結果をそれぞれ示す。実験結果より、両者の併用により人名検出率が 81.7% に向上していることがわかる。実際、評価データの大語彙連続音声認識結果中には、人名の湧き出し誤りは全くなかったことから、両者の併用による人名検出は、以降の人名インデキシング実験に非常に有効であることがいえる。

7.5 人名インデキシング

7.4 節で、大語彙連続音声認識とワードスポッティングの併用により抽出された人名と話者の交替を促すフレーズ（キーフレーズ）から、話者セグメンテーションによって得られた話者ラベルに、対応する話者の名前を付与する。

抽出された人名ならびにキーフレーズを発話している話者は、その発話の前後で発話しているインタビューされる人に関して述べており、アナウンサーあるいはレポーターであるとみなすことができる。従って、抽出された人名とキーフレーズを含む発話の前後で、その話者と異なる話者ラベルの発話をインタビューされる人とみなして、話者の名前を付与する。

この方法により、人名のインデキシング実験を行った。実験データは、98 年 8 月の NHK15 分間のニュース 4 日分 (7.2.2 節と同様) である。本研究で用いたニュース音声には、インタビューされる人の区間が 18 個あり、そのうちキーフレーズが前後に存在していたインタビューされる人の区間は、10 個であった。なお、キーフレーズがなかったインタビューされる人の区間としては、国会の答弁で複数の話者が次々に質問を行っている場面があった。したがって、この 10 個のインタビューされる人の区間に対して、大語彙連続音声認識とワードスポッティングの併用により抽出された人名とキーフレーズによる人名インデキシングを行った結果、7 個のインタビューされる人の区間に対して、話者の名前を付与することができた。残りの区間は、人名あるいはキーフレーズ

のどちらかが抽出できなかったために、話者の名前を付与することができなかった。

7.6 まとめ

本研究では、「クリントン大統領が、情報スーパーハイウェイについて話しているシーンを見たい」といった、特定の話者がある話題について話しているシーンの検索を目的として、話者の交替を検出し、各話者の名前を付与する人名インデキシングについて検討を行った。話者交替の検出においては、6.2.1節で述べた手法により話者セグメンテーションを行い、話者の学習データ量に応じた話者判定のための閾値設定法を提案した。また、大語彙連続音声認識により話者の交替を促すフレーズ（キーフレーズ）を抽出し、大語彙連続音声認識とワードスポッティングの併用により人名を抽出することで、話者セグメンテーションによって自動構築された話者モデルに、対応する話者の名前を付与する方法を提案した。NHK15分間のニュース4日分に対して人名付与実験を行った結果、話者交替検出率80.3%、話者同定率89.2%、人名付与率70.0%が得られ、本手法が有効であることがわかった。

今後の課題としては、2話者の発話が重畳した区間に対する話者セグメンテーションが残されている。また、人名の付与に関して、人名と話者モデルとの共起の利用、音声のみではなくテロップの人名を検出することで、音声と画像を統合化した人名付与が残されている。さらに、話者と発話内容の同時検索システムの構築が残されている。

第8章

結論

8.1 まとめ

音声データには，発話内容を表す音韻性と，話者が誰であるかを表す話者性が含まれている．この音韻性と話者性を分離し，話者性だけを持つ音声データを基に話者モデルを設計することができれば，発話内容に依存しない頑健な話者認識が可能になると考えられる．このことから，本論文では，音声データに含まれている音韻性と話者性の分離に基づいた話者性の抽出に関する研究を行った．

発話内容による音声の特徴変動は，音韻性，すなわち音声中に含まれている音韻のばらつきにより生じると考えられる．これに対して，発話内容に依存しない話者性が存在するのではないかと考えられる．この点から，話者内分散に基づいて音韻性と話者性を分離することができると考えられる．つまり，各話者の音声データの分散が大きい空間では音韻性が強く表れ，分散が小さい空間は，音韻性が抑えられ話者性が表れているとみなす．そこで，本論文では，音韻性と話者性を分離する手法として，主成分分析に基づいた部分空間分離による話者性の抽出法を提案した．

第2章では，VQ歪み，HMM，GMMによる統計的手法，ならびに判別分析，部分空間法による空間表現手法，識別学習といった従来の話者認識手法について述べた．

第3章では，主成分分析に基づいた部分空間法をベースとして，話者認識における距離尺度について検討した．学習データの平均ベクトルと入力特徴ベクトルとの距離が離れていた場合，異なる話者である可能性が高い．しかし，投影距離は，入力特徴ベクトルと平均ベクトルとの距離を考慮していないので，認識を誤ってしまう場合がある．そこで，話者を識別する上で，入力特徴ベクトルと平均ベクトルとの距離が短

く、かつ投影距離も短くなるように識別関数を定義する必要がある。これら2つの距離の変動を抑制するために、それぞれの距離を分散で正規化した。分散で正規化した入力特徴ベクトルと平均ベクトルとの距離はマハラノビス距離であり、分散で正規化した投影距離とこのマハラノビス距離に重み付けして統合した重み付きマハラノビス距離法を提案した。投影距離に基づく手法と本手法に対して話者照合による比較実験を行った結果、投影距離に基づく手法では等誤り率3.1%であるのに対して、本手法では等誤り率2.3%が得られ、約25%の照合精度の改善が得られた。したがって、話者を識別する上で、入力特徴ベクトルと平均ベクトルとの距離を考慮する必要があることがわかった。また、話者固有音声空間の直交補空間との距離を強調することで照合精度が向上したことから、話者固有音声空間の直交補空間は、話者性をよく表していることがわかった。

第4章では、入力特徴ベクトルと平均ベクトルとの距離を考慮した音声の特徴変動に頑健な話者性の抽出手法として、話者固有音声空間における動的・静的特徴統合法について述べた。従来の部分空間法では、発話内容や時期差による特徴変動に起因して、学習データの平均ベクトルと入力特徴ベクトルとの間にずれが生じ、話者の認識精度が劣化する。そこで、本手法では、入力音声の特徴の時間的変化量を動的特徴とし、話者固有音声空間と動的特徴の距離を話者の動的特徴距離と定義した。また、入力音声の平均ベクトルを静的特徴とし、この静的特徴は、時期差の影響を受けて学習データの平均ベクトルとずれが生じる。この静的特徴と学習データの平均ベクトルとの差（距離）には時期差が含まれており、これを話者の静的特徴距離と定義した。そこで、静的特徴距離に重み付けし動的特徴距離と統合することで、時期差による変動を抑制した。従来の部分空間法と本手法に対して話者照合による比較実験を行った結果、従来の部分空間法では等誤り率1.17%であったが、本手法により等誤り率0.29%が得られ、約75%の照合精度の改善が得られた。したがって、本手法は、音声の特徴変動に対して頑健な手法であることがわかった。

第5章では、主成分分析により話者内分散の大きい空間を音韻空間、話者内分散の小さい空間を話者空間とみなして、音韻性と話者性を分離する手法について述べた。さらに、本手法に基づいて話者空間上で話者性を統計的に表現する、話者空間への射影に基づく統計的モデルによる話者認識法について述べた。従来のGMMによる手法と本手法に対して話者識別および話者照合による比較実験を行った。その結果、話者識別では、本手法は4混合分布の時に識別誤り率5.11%が得られ、従来の64混合分布の

GMMにおける識別誤り率8.89%に比べると、約43%の識別精度の改善が得られた。また、話者照合では、本手法は4混合分布の時に等誤り率1.79%が得られ、従来の64混合分布のGMMにおける等誤り率2.0%に比べると、約11%の照合精度の改善が得られた。したがって、本手法により音韻性を抑えることができ、話者性を精度よく抽出することがわかった。また、少量のパラメータ数で頑健な話者モデルを構成でき、高速に学習、識別および照合を行えることがわかった。さらに、本手法で構成された話者空間は話者性を表す空間であるが、他者との相関を考慮していない。そこで、他者との違いをより強調した空間を構成するために、最小分類誤り学習（MCE）／一般化確率的降下法（GPD）に基づく識別学習による話者空間の最適化法について述べた。本手法の有効性を示すために話者照合実験を行った結果、本手法は4混合分布の時に等誤り率0.62%が得られ、最適化していない話者空間における手法に比べると、41%の照合精度の改善が得られた。したがって、MCE/GPDによる話者空間の最適化が有効であることがわかった。

第6章では、部分空間により抽出された話者性の応用として、アナウンサー発話の自動抽出と音声認識を統合化したニュース記事の自動分類システムについて述べた。アナウンサー発話の抽出では、特定の話者をあらかじめ学習することなく、入力音声において話者モデルを自動学習し、学習した話者モデルとの話者照合に基づいたオンライン話者セグメンテーション法について述べた。また、話者空間への射影に基づく統計的モデルとして、VQを用いた話者照合法について述べた。ニュース記事の分類では、自動抽出したアナウンサー発話の音声認識結果から、 χ^2 法により選択されたキーワードを元にニュース記事を分類する手法について述べた。NHK5分間のニュース45日分に対して記事分類実験を行った結果、アナウンサー発話の抽出率93.2%、記事分類率70.8%が得られ、本システムが有効であることがわかった。

第7章では、部分空間により抽出された話者性の応用として、話者認識と音声認識を統合した話者の人名付与システムについて述べた。インタビュー区間の検出では、アナウンサー発話の抽出における話者セグメンテーション法を用いており、本手法に対して学習データ量を考慮した話者同定における閾値設定手法について述べた。発話の交替を示すキーフレーズと人名の検出では、大語彙連続音声認識とキーワードスポッティングによる手法について述べた。その後、検出したインタビュー区間に、話者の人名を付与する手法について述べた。NHK15分間のニュース4日分に対して人名付与実験を行った結果、話者交替検出率80.3%、話者同定率89.2%、人名付与率70.0%が得

られ、本システムが有効であることがわかった。

8.2 今後の課題

従来の音声認識と話者認識では、同じ特徴量とモデル化手法が主に用いられており、音韻性と話者性が混在したまま処理されているため、音声認識と話者認識精度に限界があると考えられる。したがって、音声認識と話者認識精度をより向上させるためには、音声認識と話者認識のそれぞれに対して特有の特徴抽出ならびにモデル化を行う必要があると考えられる。今後の課題としては、さらに音声データに含まれる音韻性と話者性の分離という観点から、音声認識と話者認識のそれぞれに対して、特有の特徴抽出ならびにモデル化に関する研究が残されている。つまり、音韻性だけを持つ音声データを基に音声モデルを設計することにより、話者の違いに依存しない頑健な音声認識が実現できると考えられる。

また、話者空間への射影に基づく統計的モデルによる話者認識において、最適な話者空間の次元数を事前に設定する方法が課題として残されている。話者の人名付与システムにおいては、2話者の発話が重畳した区間に対する話者セグメンテーションが課題として残されている。また、人名の付与に関して、人名と話者モデルとの共起の利用、音声のみではなくテロップの人名を検出することで、音声と画像を統合化した人名付与が残されている。さらに、話者と発話内容の同時検索システムの構築が残されている。

謝辞

本研究を行うに当たって，常日頃より御指導を頂き，この研究を行う機会を与えて下さった有木康雄教授に深く感謝致します．さらに，本研究に関して，御意見及び御指導を下された熊野雅仁先生，及び有木研究室の皆様に改めて，心から感謝致します．

また，本研究で用いた話者認識用データベースを提供して頂いたATRの松井知子氏，ならびにNTTサイバースペース研究所の関係諸氏に深く感謝致します．

参考文献

- [1] L.Wilcox, F.Chen, D.Kimber and V.Balasubramanian, "Segmentation of Speech Using Speaker Identification", Proc.ICASSP, pp.161-164, 1994.
- [2] D.Roy and C.Malamud, "Speaker Identification Based Test to Audio Alignment for An Audio Retrieval System", Proc.ICASSP, Vol.2, pp.1099-1103, 1997.
- [3] D.Roy, "Speaker Indexing Using Neural Network Clustering of Vowel Spectra", International Journal of Speech Technology, Vol.1, No.2, pp.143-149, 1997.
- [4] A.E.Rosenberg, I.Magrin-Chagnolleau, S.Parthasarathy and Q.Huang, "Speaker Detection in Broadcast Speech Database", Proc.ICSLP, pp.1339-1342, 1998.
- [5] P.Delacourt and C.Wellekens, "Audio Data Indexing: Use of Second-Order Statistics for Speaker-Based Segmentation", Proc.ICMCS, pp.959-963, 1999.
- [6] S.E.Johnson, "Who Spoke When? - Automatic Segmentation and Clustering for Determining Speaker Turns", Proc.EUROSPEECH, Vol.5, pp.2211-2214, 1999.
- [7] M.Viswanathan, F.Maali, Homayoon S.M.Beigi and A.Tritschler, "Multimedia Information Access Using Multiple Speaker Classifiers", Proc.RIAO, pp.96-105, 2000.
- [8] 張 志鵬, 古井 貞熙, "話者交代検出を含むオンライン話者適応の検討", 音講論集, Vol.1, pp.45-46, 1999.
- [9] 森 一将, 山本 一公, 中川 聖一, "発話間のVQ歪みを用いたオンライン話者交替識別と話者クラスタリング", 信学技報, SP2000-18, pp.17-24, 2000.
- [10] 古井 貞熙, "音声情報処理", 森北出版 1998.

- [11] A.Higgins, L.Bahler and J.Porter, "Speaker Verification Using Randomized Phrase Prompting", Digital Signal Processing, Vol.1, pp.89-106, 1991.
- [12] 松井 知子, "HMMによる話者認識", 信学技報, SP95-111, pp.17-24, Jan. 1996.
- [13] 松井 知子, 古井 貞熙, "音韻・話者独立モデルによる話者照合の正規化", 信学技報, SP94-22, pp.61-66, June 1994.
- [14] T.Matsui and S.Furui, "Concatenated Phoneme Models for Text-Variable Speaker Recognition", Proc.ICASSP, Vol.2, pp.391-394, 1993.
- [15] T.Matsui and S.Furui, "Speaker Adaptation of Tied-Mixture-Based Phoneme Models for Text- Prompted Speaker Recognition", Proc.ICASSP, Vol.1, pp.125-128, 1994.
- [16] A.E.Rosenberg, J.Delong, C-H.Lee, B-H.Juang and F.K.Soong, "The Use of Cohort Normalized Scores for Speaker Verification", Proc.ICSLP, Vol.1, pp.599-602, 1992.
- [17] 古井 貞熙, 松井 知子, "自由テキストによる発声内容依存型話者認識法", 音講論集, 1-7-3, pp.51-52, Oct. 1992.
- [18] マルコフ コンスタンティン, 中川 聖一, "フレームレベルで尤度処理を用いるテキスト独立話者識別システム", 信学技報, SP96-17, pp.37-44, June 1996.
- [19] 早川 昭二, 武田 一哉, 板倉 文忠, "音声の高域を用いた発声内容独立型話者認識", 信学技報, SP95-68, pp.17-24, Oct. 1995.
- [20] 松村 理恵子, 岡本 敏雄, "ピッチを利用したテキスト非依存形の話者識別", 信学技報, SP95-133, pp.33-40, Feb. 1996.
- [21] 長内 隆, 木戸 博, 鎌田 敏明, 谷本 益巳, "判別関数を用いた不特定話者母音認識", 信学技報, SP96-65, pp.31-36, Nov. 1996.
- [22] 松井 知子, 古井 貞熙, "連結音韻モデルによる話者認識", 信学技報, SP92-128, pp.41-47, Jan. 1993.

- [23] 松井 知子, 古井 貞熙, “テキスト指定形話者認識のための話者適応化法”, 音講論集, 3-7-5, pp.97-98, Mar. 1994.
- [24] 松井 知子, 古井 貞熙, “テキスト指定型話者認識”, 信学論, Vol.J79-D-II, No.5, pp.647-656, May 1996.
- [25] 内部 利明, 黒岩 眞吾, 樋口 宜男, “数字を用いた話者照合方式の検討”, 信学技報, SP98-68, pp.1-8, Oct. 1998.
- [26] 内部 利明, 黒岩 眞吾, 樋口 宜男, “話者モデル学習時の尤度上昇幅を用いた話者照合閾値設定法”, 信学技報, SP99-63, pp.9-14, Aug. 1999.
- [27] 鹿野 清宏, 中村 哲, 伊勢 史朗, “音声・音情報のデジタル信号処理”, 昭晃堂 1997.
- [28] 古井 貞熙, “音響・音声工学”, 近代科学社 1992.
- [29] 宮島 千代美, 渡辺 秀行, 徳田 恵一, 北村 正, 片桐 滋, “識別的特徴抽出に基づく話者認識－メルケプストラム特徴の最適化－”, 信学技報, SP99-62, pp.1-8, 1999.
- [30] 早川 昭二, 武田 一哉, 板倉 文忠, “線形予測残差スペクトルの調波構造に含まれる個人性情報を用いた話者認識”, 信学論, Vol.J80-A, No.9, pp.1360-1367, Sept. 1997.
- [31] P.Thevenaz and H.Hugli, “Usefulness of the LPC-residue in text-independent speaker verification”, Speech Communication, Vol.17, pp.145-157, Aug. 1995.
- [32] K.P.Markov and S.Nakagawa, “Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition”, J.Acoust.Soc.Jpn.(E)20, pp.281-291, 1999.
- [33] 中川 聖一, “確率モデルによる音声認識”, 電子情報通信学会編 1988.
- [34] 松井 知子, 古井 貞熙, “VQひずみ、離散/連続HMMによるテキスト独立型話者認識法の比較検討”, 信学論, Vol.J77-A, No.4, pp.601-606, Apr. 1994.
- [35] 松井 知子, 古井 貞熙, “音源・声道特徴を用いたテキスト独立形話者認識”, 信学論, Vol.J75-A, No.4, pp703-709, Apr. 1992.

- [36] S.V.Vuuren, "Comparison of Text-independent Speaker Recognition Methods on Telephone Speech with Acoustic Mismatch", Proc.ICSLP, Vol.3, pp.1788-1791, 1996.
- [37] M.Faundez-Zanuy and D.Rodriguez-Porcheron, "Speaker Recognition Using Residual Signal of Linear and Nonlinear Prediction Models", Proc.ICSLP, Vol.2, pp.121-125, 1998.
- [38] J.He, L.Liu and G.Palm, "A New Codebook Training Algorithm for VQ-Based Speaker Recognition", Proc.ICASSP, Vol.2, pp.1091-1094, 1997.
- [39] Q.Jin and A.Waibel, "A Naive De-lambing Method for Speaker Identification", Proc.ICSLP, Vol.2, pp.466-469, 2000.
- [40] 北 研二, 中村 哲, 永田 昌明, "音声言語処理", 森北出版 1996.
- [41] J.Ortega-Garcia and J.Gonzalez-Rodriguez, "Providing Signal and Multi-Channel Acoustical Robustness to Speaker Identification Systems", Proc.ICASSP, pp.1107-1110, 1997.
- [42] D.Hardt and K.Fellbaum, "Spectral Subtraction and RASTA-Filtering in Text-Dependent HMM-Based Speaker Verification", Proc.ICASSP, pp.867-870, 1997.
- [43] Q.Li, S.Parthasarathy and A.E.Rosenberg, "A Fast Algorithm for Stochastic Matching with Application to Robust Speaker Verification", Proc.ICASSP, pp.1543-1546, 1997.
- [44] P.Sivakumaran and A.M.Ariyaeinia, "The Use of Sub-Band Cepstrum in Speaker Verification", Proc.ICASSP, Vol.2, pp.1073-1076, 2000.
- [45] M.J.Carey, E.S.Parris, S.J.Bennett and H.Lloyd-Thomas, "A Comparison of Model Estimation Techniques for Speaker Verification", Proc.ICASSP, pp.1083-1086, 1997.
- [46] S.Euler, R.Langlitz and J.Zinke, "Comparison of Whole Word and Subword Modeling Techniques for Speaker Verification with Limited Training Data", Proc.ICASSP, pp.1079-1082, 1997.

- [47] 磯部 俊洋, 高橋 淳一, “話者照合における HMM の局所的音響情報に基づく尤度正規化”, 信学技報, SP98-105, pp.69-74, Dec. 1998.
- [48] J.D.Veth, G.Gallopin and H.Bourland, “Limited Parameter Hidden Markov Models for Connected Digit Speaker Verification over Telephone Channels”, Proc.ICASSP, Vol.2, pp.247-250, 1993.
- [49] Q.Li and B-H.Juang, “Speaker Verification Using Verbal Information Verification for Automatic Enrollment, ” Proc.ICASSP, Vol.1, pp.133-136, 1998.
- [50] Q.Li, “Search-Space Reduction for Fast, Optimal HMM Decoding in Speaker Verification”, Proc.ICASSP, Vol.2, pp.1189-1192, 2000.
- [51] D.A.Reynolds and R.C.Rose, “Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, IEEE Trans.SAP, Vol.3, No.1, pp.72-83, 1995.
- [52] L.P.Heck, and M.Weintraub, “Handset-Dependent Background Models for Robust Text-Independent Speaker Recognition”, Proc.ICASSP, pp.1071-1074, 1997.
- [53] F.Beaufays and M.Weintraub, “Model Transformation for Robust Speaker Recognition from Telephone Data”, Proc.ICASSP, pp.1063-1066, 1997.
- [54] S.V.Vuuren and H.Hermansky, “On the Importance of Components of the Modulation Spectrum for Speaker Verification”, Proc.ICSLP, Vol.7, pp.3205-3208, 1998.
- [55] W.H.Tsai, C.Che and W.W.Chang, “Text-Independent Speaker Identification Using Gaussian Mixture Bigram Models”, Proc.ICSLP, Vol.2, pp.314-317, 2000.
- [56] 宮島 千代美, 徳田 恵一, 北村 正, “多空間確率分布 GMM に基づく話者識別モデルにおけるストリーム重みの推定”, 音講論集, 1-3-3, pp.5-6, Mar. 2001.
- [57] D.Tran and M.Wagner, “A Proposed Likelihood Transformation for Speaker Verification”, Proc.ICASSP, Vol.2, pp.1069-1072, 2000.

- [58] E.Battle, C.Nadeu and Jose.A.R.Fonollosa, "Feature Decorrelation Methods in Speech Recognition, A Comparative Study", Proc.ICSLP, Vol.3, pp.951-954, 1998.
- [59] I.Magrin-Chagnolleau, G.Gravier, M.Seck, O.Boeffard, R.Blounet and F.Bimbot, "A Further Investigation on Speech Features for Speaker Characterization", Proc.ICSLP, Vol.3, pp.1029-1032, 2000.
- [60] C.Tadj, P.Dumouchel, M.Mihoubi and P.Ouellet, "Environment Adaptation and Long Term Parameters in Speaker Identification", Proc.EUROSPEECH, Vol.2, pp.1015-1018, 1999.
- [61] エルツキ・オヤ著, 小川 英光, 佐藤 誠訳, "パターン認識と部分空間法", 産業図書 1986.
- [62] 石井 健一郎, 上田 修功, 前田 英作, 村瀬 洋, "わかりやすい パターン認識", オーム社 1998.
- [63] Y.Ariki and K.Doi, "Speaker Recognition based on Subspace Method", Proc.ICSLP, Vol.4, pp.1859-1862, 1994.
- [64] Y.Ariki, S.Tagashira and M.Nishijima, "Speaker Recognition and Speaker Normalization by Projection to Speaker Subspace", Proc.ICASSP, Vol.1, pp.319-322, 1996.
- [65] 池田 正幸, 田中 英彦, 元岡 達, "手書き文字認識における投影距離法", 情報学論, Vol.24, No.1, pp.106-112, 1983.
- [66] B-H.Juang and S.Katagiri, "Discriminative Learning for Minimum Error Classification", IEEE Trans. Signal Processing, Vol.40, No.12, pp.3043-3054, Dec. 1992.
- [67] A.E.Rosenberg, O.Siohan and S.Parthasarathy, "Speaker Verification Using Minimum Verification Error Training", Proc.ICASSP, pp.105-108, 1998.
- [68] O.Siohan, A.E.Rosenberg and S.Parthasarathy, "Speaker Identification Using Minimum Classification Error Training", Proc.ICASSP, pp.109-112, 1998.

- [69] 黒沢 由明, “球面ガウス分布から導出される部分空間法”, 信学論, Vol.J81-DII, No.6, pp.1205-1212, June 1998.
- [70] 加藤寧, 安倍正人, 根元義章, “改良型マハラノビス距離を用いた高精度な手書き文字認識”, 信学論, Vol.J79-DII, No.1, pp.45-52, Jan. 1996.
- [71] 木村文隆, 高階健治, 鶴岡信治, 三宅康二, “2次識別関数のピーキング現象とその防止に関する考察”, 信学論, Vol.J69-D, No.9, pp.1328-1334, Sept. 1986.
- [72] S.Furui, “Cepstral Analysis Technique for Automatic Speaker Verification”, IEEE Trans.ASSP, Vol.29, No.2, pp.254-272, 1981.
- [73] 松井 知子, 西谷 隆, 古井 貞熙, “話者照合におけるモデルとしきい値の更新法”, 信学論, Vol.J81-D-II, No.2, pp.268-276, 1998.
- [74] A.E.Rosenberg, C-H.Lee and F.K.Soong, “Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification”, Proc.ICSLP, Vol.4, pp.1835-1838, 1994.
- [75] J.Gonzalez-Rodriguez and J.Ortega-Garcia, “Robust Speaker Recognition Through Acoustic Array Processing and Spectral Normalization”, Proc.ICASSP, pp.1103-1106, 1997.
- [76] S.J.Young, et al, “HTK: Hidden Markov Model Toolkit V2.1”, Entropic Research Laboratory Inc., Dec. 1997.
- [77] 横井 謙太郎, 河原 達也, 堂下 修司, “キーワードスポッティングに基づくニュース音声の話題同定”, 情処研報, 95-SLP-6-3, pp.15-20, 1995.
- [78] 大附 克年, 松岡 達雄, 松永 昭一, 古井 貞熙, “ニュース音声を対象とした大語彙連続音声認識と話題抽出”, 信学技報, SP97-27, pp.67-74, 1997.
- [79] K.Ohtsuki, T.Matsuoka, S.Matsunaga and S.Furui, “Topic Extraction Multiple Topic-Words in Broadcast-News Speech”, Proc.ICASSP, pp.329-332, 1998.
- [80] 恒川 俊克, 山下 洋一, 溝口 理一郎, “キーワードスポッティングに基づくニュース音声の話題分類”, 情処研報, 98-SLP-20-11, pp.61-68, 1998.

- [81] 松井 知子, 相川 清明, “時期差による発声変動を考慮した話者モデルの生成法”, 音講論集, 1-1-23, pp.45-46, Sept. 1997.
- [82] 西田 昌史, 有木 康雄, “自動学習による話者セグメンテーション”, 信学技報, SP97-57, pp.1-6, 1997.
- [83] 三村 正人, 河原 達也, 堂下 修司, “パネル討論音声の話者と話題に関する自動インデキシング”, 情処研報, 96-SLP-11-3, pp.13-18, 1996.
- [84] 緒方 淳, 森 晴, 有木 康雄, “単語 bigram を用いた日本語ニュースディクテーションによる記事分類”, 音講論集, 2-Q-16, pp.151-152, Mar. 1998.
- [85] 朝日新聞社ニューメディア本部, “朝日新聞記事データベース分類表索引”, 1992.
- [86] 櫻井 光康, 有木 康雄, “キーワードスポッティングによるニュース音声の分類と索引付け”, 信学技報, SP96-66, pp.37-44, 1996.
- [87] 鷹尾 誠一, 緒方 淳, 有木 康雄, “ニュース音声の記事分類におけるキーワード選択法の比較”, 情処研報, 98-SLP-22-15, pp.75-82, 1998.
- [88] 西田 昌史, 緒方 淳, 有木 康雄, “アナウンサー発話の自動抽出と大語彙連続音声認識によるニュース記事分類”, 情処学論, Vol.40, No.4, pp.1482-1490, 1999.
- [89] 鷹尾 誠一, 緒方 淳, 有木 康雄, “ニュース音声記事データベースにおける観点の自動抽出と構造化”, 信学技報, DE2000-12, pp.89-96, 2000.
- [90] 鷹尾 誠一, 船本 純一, 有木 康雄, 緒方 淳, “ニュース音声データベースに対するクロスメディア検索”, MIRU2000, Vol.1, pp.457-462, 2000.
- [91] 村井 則之, 小林 哲則, “MLLRによる話者適応と統計的発話交代モデルを用いた複数話者対話音声の認識”, 信学技報, SP2000-14, pp.31-38, 2000.
- [92] K.Sonmez, L.Heck and M.Weintraub, “Speaker Tracking and Detection with Multiple Speakers”, Proc.EUROSPEECH, Vol.5, pp.2219-2222, 1999.
- [93] P.Delacourt, D.Kryze and C.J.Wellekens, “Detection of Speaker Changes in an Audio Document”, Proc.EUROSPEECH, Vol.3, pp.1195-1198, 1999.

- [94] A.Tritschler and R.Gopinath, "Improved Speaker Segmentation and Segments Clustering using the Bayesian Information Criterion", Proc.EUROSPEECH, Vol.2, pp.679-682, 1999.
- [95] M.Nishida and Y.Ariki, "Speaker Indexing for News Articles, Debates and Drama in Broadcasted TV Programs", Proc.ICMCS, Vol.2, pp.466-471, 1999.
- [96] 西田 昌史, 有木 康雄, "話者固有空間における動的・静的特徴統合による話者照合", 信学論, Vol.J83-DII, No.12, pp.2536-2544, 2000.
- [97] 佐藤 真一, 中村 裕一, 金出 武雄, "Name-It: 動画像処理と自然言語処理の統合による映像内容アクセス手法", 第3回知能情報メディアシンポジウム, pp.187-194, 1997.
- [98] 西田 昌史, 緒方 淳, 有木 康雄, "話者名の自動索引付けと話者・発話内容の同時検索", 第15回人工知能学会全国大会, 1A4-04, 2001.
- [99] S.Ortmanns, H.Ney and X.Aubert, "A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition", Computer Speech and Language, Vol.11, No.1, pp.43-72, 1997.
- [100] J.Ogata and Y.Ariki, "An Efficient Lexical Tree Search for Large Vocabulary Continuous Speech Recognition", Proc.ICSLP, Vol.2, pp.967-970, 2000.
- [101] 緒方 淳, 有木 康雄, "back-off接続を考慮した大語彙連続音声認識の高速化", 音講論集, 2-8-8, pp.43-44, Mar. 2000.
- [102] 李 晃伸, 河原 達也, "大語彙連続音声認識エンジン JuliusにおけるA*探索法の改善", 情処研報, 99-SLP-27-5, pp.33-39, 1999.
- [103] K.M.Knill and S.J.Young, "Speaker Dependent Keyword Spotting for Accessing Stored Speech", CUED/F-INFENG/TR 193, 1994.
- [104] 緒方 淳, 有木 康雄, "ニュース記事分類におけるディクテーションとワードスポッティングの比較", 信学技報, SP98-32, pp.67-72, 1998.

発表論文

学術論文

- [1] 西田 昌史, 緒方 淳, 有木 康雄, “アナウンサー発話の自動抽出とディクテーションによるニュース記事分類”, 情報処理学会論文誌, Vol.40, No.4, pp.1482-1490, Apr. 1999.
- [2] 西田 昌史, 有木 康雄, “話者固有空間における動的・静的特徴統合による話者照合”, 電子情報通信学会論文誌, Vol.J83-D-II, No.12, pp.2536-2544, Dec. 2000.
- [3] 西田 昌史, 有木 康雄, “音韻性を抑えた話者空間への射影による話者認識”, 電子情報通信学会論文誌, Vol.J85-D-II, No.4, pp.-, Apr. 2002.

国際会議

- [1] Y.Sugiyama, N.Ishikawa, M.Nishida and Y.Ariki, “Indexing and Retrieval of Human Individuals on Video Data Using Face and Speaker Recognition”, Proc.IWAIT, pp.122-127, Jan. 1998.
- [2] Y.Ariki, J.Ogata and M.Nishida, “News Dictation and Article Classification Using Automatically Extracted Announcer Utterance”, Proc.AMCP, pp.78-89, Nov. 1998.
- [3] M.Nishida and Y.Ariki, “Speaker Indexing for Broadcasted TV Programs”, Proc.IWHIT, pp.1-6, Nov. 1998.

- [4] M.Nishida and Y.Ariki, "Real Time Speaker Indexing based on Subspace Method - Application to TV News Articles and Debate -", Proc.ICSLP, Vol.4, pp.1347-1350, Dec. 1998.
- [5] M.Nishida and Y.Ariki, "Speaker Indexing for News Articles, Debates and Drama in Broadcasted TV Programs", Proc.ICMCS, Vol.2, pp.466-471, June 1999.
- [6] M.Nishida and Y.Ariki, "Speaker Verification by Integrating Dynamic and Static Features Using Subspace Method", Proc.ICSLP, Vol.3, pp.1013-1016, Oct. 2000.
- [7] M.Nishida and Y.Ariki, "Speaker Recognition by Separating Phonetic Space and Speaker Space", Proc.EUROSPEECH, Vol.2, pp.1381-1384, Sept. 2001.

研究会

- [1] 西田 昌史, 有木 康雄, “自動学習による話者セグメンテーション”, 電子情報通信学会技術研究報告, SP97-57, pp.1-6, Nov. 1997.
- [2] 西田 昌史, 有木 康雄, “話者照合に基づく映像中の話者の自動索引付け”, 第3回知能情報メディアシンポジウム, 3-2, pp.55-60, Dec. 1997.
- [3] 緒方 淳, 西田 昌史, 有木 康雄, “自動抽出されたアナウンサー発話に対するニュースディクテーションと記事分類”, 情報処理学会研究報告, SLP-21-5, pp.55-60, May 1998.
- [4] 杉山 善明, 西田 昌史, 有木 康雄, “VideoAssembler: 動画像処理と音声処理の統合によるニュース記事内容の分割”, 第4回知能情報メディアシンポジウム, S-2-1, pp.23-30, Dec. 1998.
- [5] 西田 昌史, 有木 康雄, “ガウス分布から導出される部分空間法による話者照合法の検討”, 電子情報通信学会技術研究報告, SP99-19, pp.9-16, June 1999.
- [6] 西田 昌史, 有木 康雄, “部分空間法を用いた動的・静的特徴統合による話者照合”, 電子情報通信学会技術研究報告, SP99-137, pp.1-6, Jan. 2000.
- [7] 西田 昌史, 有木 康雄, “話者補空間射影による話者認識”, 電子情報通信学会技術研究報告, SP2000-12, pp.17-22, June 2000.
- [8] 西田 昌史, 緒方 淳, 有木 康雄, “話者と発話内容の同時検索に関する検討”, 人工知能学会情報統合研究会, cii14-12, Nov. 2000.
- [9] 西田 昌史, 緒方 淳, “話者の名前付けと発話内容の検索”, 第9回情報処理学会関西支部サイバーメディア研究会, Dec. 2000.
- [10] 西田 昌史, 有木 康雄, “音韻空間への射影に基づく話者正規化による音素認識”, 電子情報通信学会技術研究報告, SP2000-90, pp.89-94, Dec. 2000.
- [11] 井上 徹, 西田 昌史, 藤本 雅清, 有木 康雄, “部分空間と混合分布モデルを用いた声質変換”, 電子情報通信学会技術研究報告, SP2001-9, pp.1-6, May 2001.

全国大会

- [1] 西田 昌史, 有木 康雄, “ニュース音声における話者区間の切り出し”, 第41回システム制御情報学会研究発表講演会, pp.413-414, May 1997.
- [2] 西田 昌史, 有木 康雄, “座談会における話者区間の切り出し”, 日本音響学会平成9年度秋季研究発表会, 2-Q-12, pp.139-140, Sept. 1997.
- [3] 西田 昌史, 有木 康雄, “ニュース音声中のアナウンサー発話区間の自動切り出し”, 情報処理学会全国大会, 3J-4, pp.2-31-2-32, Sept. 1997.
- [4] 西田 昌史, 有木 康雄, “部分空間法による話者照合”, 日本音響学会平成10年度秋季研究発表会, 3-1-15, pp.111-112, Sept. 1998.
- [5] 西田 昌史, 有木 康雄, “ドラマにおける話者インデキシングの検討”, 日本音響学会平成11年度春季研究発表会, 3-Q-26, pp.169-170, Mar. 1999.
- [6] 西田 昌史, 有木 康雄, “部分空間法を用いた話者照合における動的・静的特徴統合-部分空間次元数の事前設定の検討-”, 日本音響学会2000年春季研究発表会, 3-9-2, pp.97-98, Mar. 2000.
- [7] 西田 昌史, 有木 康雄, “話者補空間における話者認識”, 日本音響学会2000年秋季研究発表会, 1-5-3, pp.5-6, Sept. 2000.
- [8] 山下 隆之, 西田 昌史, 藤本 雅清, 有木 康雄, “実環境下での話者認識におけるPMC法の効果について”, 日本音響学会2000年秋季研究発表会, 2-Q-15, pp.119-120, Sept. 2000.
- [9] 西田 昌史, 有木 康雄, “音韻空間への射影に基づく話者正規化”, 日本音響学会2001年春季研究発表会, 3-3-16, pp.125-126, Mar. 2001.
- [10] 西田 昌史, 緒方 淳, 有木 康雄, “話者名の自動索引付けと話者・発話内容の同時検索”, 第15回人工知能学会全国大会, 1A4-04, May 2001.
- [11] 西田 昌史, 有木 康雄, “部分空間分離に基づく音声認識法の検討”, 日本音響学会2001年秋季研究発表会, 1-Q-16, pp.171-172, Oct. 2001.