

論文 / 著書情報
Article / Book Information

論題(和文)	音声認識における話者適応化技術
Title(English)	
著者(和文)	古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	テレビジョン学会誌, Vol. 43, No. 9, pp. 929-934
Citation(English)	, Vol. 43, No. 9, pp. 929-934
発行日 / Pub. date	1989,
権利情報 / Copyright	本著作物の著作権は映像情報メディア学会に帰属します。 Copyright (c) 1989 Institute of Image Information and Television Engineers.

音声認識における話者適応化技術

古井 貞 熙†

音声認識の高度化のための重要な課題のひとつである話者適応化技術、すなわち、認識系を話者による声質の違いに自動的に適応させる技術の研究が活発化しつつある、ここでは特にベクトル量子化法を基礎とし、教師あり／教師なし学習による符号帳のマッピングによって適応化を行う方法を中心にとりあげ、研究の動向について解説を行う。

1. まえがき——音声の個人差——

音声認識装置、すなわち、音声によって機械と対話するための音声入力装置に関しては、古くから研究が行われており、最近では、限定された用途ではあるが、実用的な装置が実現され使われるようになって来た。将来的には、自動翻訳電話や音声タイプライタの実現が期待されている。音声認識の難しさには種々の側面があるが、そのひとつの重要な要素に音声の個人差がある。同じ言葉を発声しても、話し方の違いや発

声器官の違いのため、音声の音響的・物理的性質は話者によって大きく異なる。音声の物理的性質は通常、スペクトルによって表現されるが、異なる話者の異なる音素が同じスペクトルを有するというものしばしば生ずる。

音素を特徴づける優勢なスペクトル成分は、ホルマントと呼ばれる。普通、声帯が振動して発声される有声音、特に母音には3個程度の特徴的なホルマントがあって、周波数の低い方から順に、第1、第2、第3ホルマントと呼ぶ。

図1は、第1ホルマント周波数 F_1 を横軸に、第2ホルマント周波数 F_2 を縦軸にした、いわゆる F_1 - F_2 平面上に、男性・女性それぞれ約30名が発声した単独5母音の分布を示したものである¹⁾。各母音ごとに、男性・女性それぞれの平均値と標準偏差が示してある。

発声者や年齢による変動は、平均的には、5母音の相対的位置関係を維持したまま対数周波数で平行移動、すなわち、周波数が比例的に変化する傾向があるが、一人一人について見ると現象はかなり複雑であり、音韻性情報と個人性情報の交互作用、すなわち、音素による個人性情報の含まれ方の相違があることも確認されている。

また、図1は単独に発声した母音に関する結果であるが、単語や文章を発声した音声の中の母音は、いわゆる調音結合の影響のため、さらに大きく変動する。このため、実際の連続音声の中の各母音の広がり、図1に示すよりもかなり大きく、個人差による重なりも極めて大きい。

したがって、認識対象語彙が少なく、しかもそれら

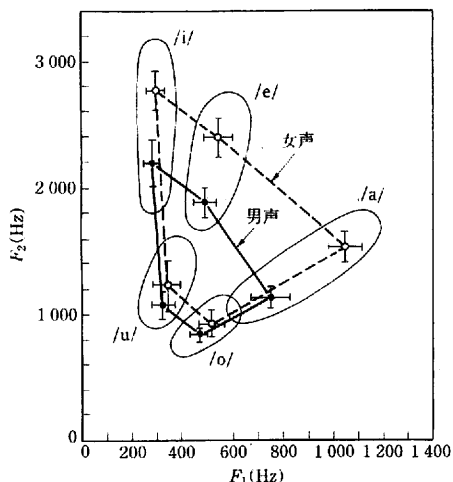


図1 男性および女性の単独5母音の第1、第2ホルマント周波数の分布

† NTT ヒューマンインタフェース研究所

"Speaker Adaptation Techniques for Speech Recognition" by Sadaaki Furui (NTT Human Interface Laboratories, Tokyo)

のスペクトルが互いに大きく異なる場合には、単に多数の話者の平均的な標準パターンを各語彙ごとに用意するだけでも十分な認識率を得ることができるが、語彙数が比較的大きい場合や、語彙数が小さくても似た語彙がある場合、あるいは連続音声のように変動が大きい場合には、高い認識率を得るために特別な工夫が必要である。

最も確実な方法は、発声者が変わるたびにすべての語彙を発声してもらって、標準パターン(認識系)をその話者に合わせることであるが、語彙数が大きくなると発声者の負担が大きくなり、非現実的となってしまう。このような背景から、音声の個人差に対処するための種々の方法が試みられて来た¹⁾。

2. 音声認識における個人差への対処法

音声の個人差への対処として最も基本的なことは、話者による変動が比較的小さく、音韻性を強調するうえで有効な物理特徴を探索して用いることであり、スペクトルの動的な特徴²⁾などが有効であることが確かめられている。しかし、この方法だけでは限界があるので、次の2種類のもののいずれか、あるいは両方との組合せが試みられている。

(1) 個人差による変動がカバーできるように、標準パターンあるいは認識モデルに広がりを持たせる。具体的には、各語彙あるいは音素について複数の標準パターン(マルチテンプレート)³⁾を用意したり、個人差による変動を確率的なモデル⁴⁾⁵⁾で表現したり、複雑な識別関数⁶⁾を作成する。

(2) 話者が変わるたびに、少数の単語あるいは短い文章を発声してもらい、その音声を用いて、その人の声に自動的に適応できる学習機能、すなわち、話者適応化機能を認識系に持たせる。

(1)の方法の有効性は種々のケースについてすでに確認されているが、まえがきで述べたように、異なる話者の異なる音素の物理的特徴に重なりがある場合には、この方法では完全に対処することができない。さらに実際の音声認識においては、個人差の他に、電話機やマイクロホンの種類、環境騒音、室内音響特性、周波数帯域制限などの影響に対しても対応できることが必要になる。このため究極的には、(2)の方法のように何らかの形で適応化機能を持たせることが望ましいと考えられる。

人間にもこのような機能、すなわち初め言葉がわかりにくくても、5音節程度の音声によって、すぐにその人の声に慣れてわかるようになる学習機能のあるこ

とが、実験的に確かめられている⁷⁾。以下では、このような機能を認識系に持たせるための種々の試みを紹介する。

3. 話者適応化法の分類

音声認識系に話者適応化機能を持たせる方法としては、これまでに次のような試みが行われている。

(1) 声帯波スペクトルと声道長の個人差を正規化する方法

異なる音素に共通に含まれる個人差の要因として、声帯波スペクトルと声道長の個人差が考えられる。これらをそれぞれ音声スペクトルの平均的概形(長時間平均スペクトルの概形)と音声スペクトルの周波数軸方向の伸縮係数で近似し、音声波から自動的に抽出して、入力音声の個人差を正規化(除去)したり、標準パターンを入力音声に適応させる方法である⁸⁾⁹⁾。

(2) 一部の音素から個人に適応した全音素のスペクトルを推定する方法

個人差によるスペクトル変動の統計的モデル化に基づいて、認識対象の一部の単語を発声した音声から抽出した音素特徴から、その人の他の音素のスペクトルを重回帰式によって推定し、その人の音声に適応した音素標準パターンを生成する方法である¹⁰⁾¹¹⁾。

(3) 音韻性情報と個人性情報の自動分離法

個人差のモデル化に基づいて、繰り返し演算により、音声スペクトルから音韻性情報と個人性情報を自動的に分離して抽出する方法である¹²⁾¹³⁾。

(4) 話者に適合する標準パターンセットを選択する方法

多数の話者を声質によってクラスタ化(分類)して、あらかじめ多数の音素標準パターンセットを用意しておき、入力音声と各セットとの類似度に基づいて、話者に最も適合したセットを選択する方法である¹⁴⁾¹⁵⁾。

(5) ベクトル量子化における符号帳のマッピングによる方法

ベクトル量子化に基づく音声認識において、あらかじめ登録されている符号帳と新しい話者の符号帳の要素間の対応関係(マッピング)を学習音声を用いて求め、話者に適応した符号帳におきかえる方法である^{16)~24)}。

(6) 符号帳のマッピングを確率モデルに取り込む方法

あらかじめ登録されている符号帳と新しい話者の符号帳の要素間のマッピングを、確率的モデルとしてHMM (Hidden Markov Model) に組み込む方法である²⁵⁾²⁶⁾。

一般に話者適応化の方法は、学習音声の発声内容が

既知であることが前提になっている“教師ありの方法”と、任意の音声を用いて適応化することができる“教師なしの方法”に分類できる。上述の方法のうち、(2)(6)は教師あり、(1)(3)(4)は教師なし、(5)は教師あり／なしの両方が試みられている。ただし、まえがきでも述べたように、個人性情報の含まれ方は音素によって必ずしも共通していないため、教師なしの方法でも、適応化性能を上げるためには、学習の過程で音声認識を行って陰にあるいは陽に音素を決定し、その情報を用いる場合が多い。完全に発声内容に依存しない個人性情報を抽出して用いる適応化方法が確立できれば、その用途は大きく広がると思われる。人間の話者適応機能は受聴される発声内容には依存しないことが確かめられている⁷⁾。

以下では、上述の中から最近の主な方法について解説する。

4. 話者に適合する標準パターンセットを選択する方法

まえがきで述べたように、 F_1 - F_2 平面における5母音の相対的な位置関係は、話者が変わっても比較的保たれる性質がある。そこで、5母音を1つのまとまった構造を持ったものと考えて1つの組(セット)として扱い、話者による変動を多数のセットの集合で表現する試みが行われている¹⁴⁾¹⁵⁾。学習音声中の母音区間を最もよく近似できる母音パターンセットを自動的に選択することによって、教師なし学習による母音の話者適応化を行うことができる。具体的には、すべての母音区間について、各セットの中の最も近い母音標準パターンで置き換えたときの歪みを計算し、全体として最も歪みの小さい(最適な)セットを選択する。各セットを、ベクトル量子化の符号帳と考え、最も量子化歪みの小さい符号帳を選択する手続きとみなすこともできる。

100都市名单語を用いた認識実験の結果、各話者について単独5母音を学習に用いた場合の認識率が88.4%であったのに対し、この方法によって5単語を用いて学習を行ったところ93.3%に向上することが確認された¹⁴⁾。母音の認識結果の確実性が高い音声区間のスペクトルを用いて、母音の標準パターンを、その話者に合うように修正する試みも行われている。

5. ベクトル量子化における符号帳のマッピングによる方法—1(教師あり学習による方法)

5.1 概要

ベクトル量子化は、本来音声や画像信号の情報圧縮

のために考え出された手法であるが、これを音声認識に適用することによって、認識性能を低下させることなく、大幅に処理量や記憶容量が削減できる(SPLIT法)²⁷⁾²⁸⁾。この方法による単語音声認識の場合、各単語標準パターンは、符号帳の要素を示す記号の列(単語辞書)で表現されるので、未知入力音声に対しては、10~20 ms程度の細かい時間間隔(フレーム)ごとに、対象語彙数に関係なく符号帳の各要素との類似度(あるいは距離)を計算すればよく、この結果(距離行列)と単語辞書を用いた動的計画法(DP)によって、各語彙との類似度を算出することができる。

このベクトル量子化を用いた音声認識の枠組みにおいて、話者適応化の問題を、話者間での有限個のスペクトルの写像問題としてとらえ、符号帳を新しい話者(未知話者)に適応化する方法が考え出された^{16)~20)}。単語辞書は話者による発声の仕方の違いを反映しているともいえるが、ここでは単語辞書はすべての話者に共通とし、符号帳のみを適応化している。この方法のブロック図を図2に示す¹⁸⁾。

この方法では、あらかじめ標準話者の音声をベクトル量子化し、符号帳を作成するとともに、各単語をその符号帳要素の系列で表現しておく。次に未知話者の学習音声を同様にベクトル量子化して、符号帳を作成するとともに、学習音声を符号帳要素の系列で表現する。未知話者と標準話者の同じ単語どうしについて、時間軸の対応関係を求める。

一般に音声の時間軸は非線形に伸縮するが、その最適な対応関係は、DPを用いることによって能率良く求めることができる。この時間軸の対応関係から、両話者の符号帳要素の対応関係を求める。各符号帳要素が種々の単語音声の中に出現するので、全学習単語に関する標準話者と未知話者の時間軸の対応関係から、符号帳要素(ベクトル)の対応関係のヒストグラムを求め、その頻度による重みからなるマッピング行列を決定する。

認識すべき未知入力音声に対しては、まずその話者の符号帳を用いてベクトル量子化し、上で得られた結果を用いて、各フレーム毎にマッピングを行って標準話者の音声に近づける(適応化する)。その後、標準話者の各単語との類似度を調べることによって、何の単語が発声されたかを決定する。

5.2 実験結果

216単語の認識において、100単語を学習に用い、他の100単語を評価用の未知単語とした実験の結果、標準話者と未知話者が同性の場合は単語認識率が93.5%から95.9%に、異性の場合は67.0%から92.3%

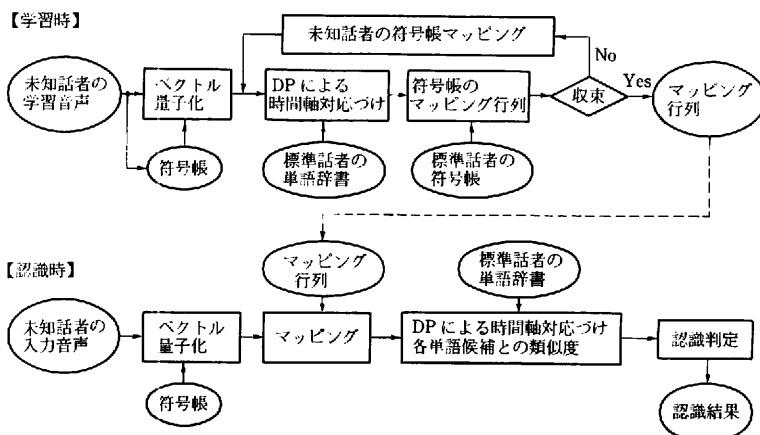


図 2 符号帳のマッピングによる教師あり話者適応化法を含む単語音声認識系のブロック図

に向上した¹⁸⁾。符号帳の大きさを比較的小さく保って精度を向上させる目的で、ファジーベクトル量子化法を取入れた試みも行われ、学習単語数を 1/4 程度に減少できることが確認されている¹⁹⁾。

6. ベクトル量子化による符号帳のマッピングによる方法-2 (教師なし学習による方法)

6.1 概要

上で述べた方法が、発声内容が既知の学習音声を用いる教師あり学習であるのに対して、学習音声の発声内容を特定せずに、任意の単語あるいは短い文章音声を用いて新しい話者に適応化する方法も研究されている^{21)~24)}。発声内容を限定しないので、未知入力音声をそのまま学習に用いることもできる。この方法は、セグメント符号化に用いる符号帳を発声者に適応化するために考案された方法²⁵⁾を基礎にしている。この話者適応化法を含む SPLIT 法に基づく単語音声認識系全体のブロック図を図 3 に示す²²⁾。

上でも述べたように、符号帳のみの適応化では解決できない個人差として時間構造の個人差があるため、これに対処する方法として、各単語毎に複数の符号系列 (マルチテンプレート) を蓄え、細かい時間軸変動には DP による時間伸縮マッチングによって対処している。

この適応化法のポイントは、学習音声スペクトルと符号帳を少数のクラスタから多数のクラスタへ、階層的にクラスタ化し、対応する各クラスタ重心間のずれにもとづいて、スペクトルの大局的な話者特性から局所的な話者特性へと徐々に適応化を進めていくところにある。この階層的処理により、学習音声と符号帳の音韻性を自動的に対応付けながら、話者適応を進める

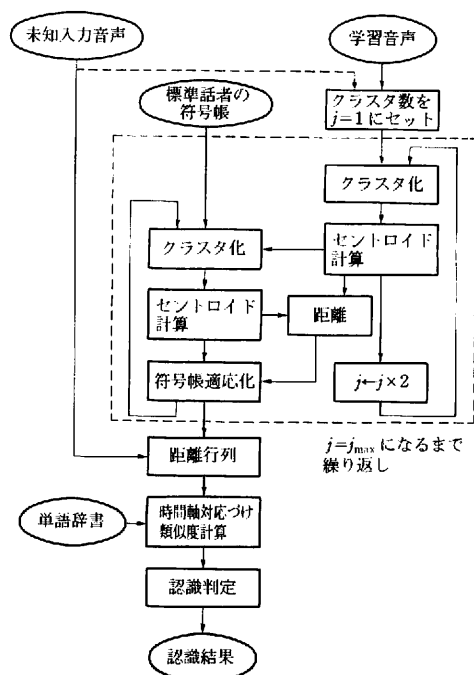


図 3 符号帳の教師なし話者適応化を含む単語音声認識系のブロック図

ことができる。

6.2 適応化の手順

具体的な適応化の手順は次の通りである²²⁾²⁹⁾。

- (1) クラスタ数の初期値を $j=1$ とし、クラスタ数の最大値を j_{max} とする。
- (2) 学習音声全フレームの重心 (セントロイド) と全符号帳要素の重心のずれ (距離ベクトル) を求める。
- (3) 全符号帳要素をこのずれに相当するだけ移動

させて重心を一致させる。

- (4) クラスタ数(j)を2倍に増やし、学習音声を用いてLBG法³⁰⁾によりクラスタ化する。
- (5) 各クラスタの重心を求め、これらの重心を用いて符号帳をクラスタ化する。
- (6) 学習音声と符号帳の対応するクラスタの重心間のずれを求める。
- (7) これらのずれをもとに、隣接クラスタ間の連続性を保ちながら符号帳要素を移動する。(具体的方法は6.3節で述べる)

- (8) $j=j_{\max}$ なら終了。 $j < j_{\max}$ なら(4)へ戻る。

このようにして(4)～(8)の処理を繰り返し、符号帳と学習音声の階層的マッピングを進める。最終状態の符号帳を用いて、SPLIT法による入力単語音声の認識を行う。

6.3 クラスタ間の連続性を保持した標準パターン移動

上でも述べたように、符号帳要素を話者に適応させて移動する際、その要素が属すクラスタの周辺の他のクラスタの重心のずれを考慮しながら、クラスタ間のずれの連続性が保持されるように移動する²²⁾²⁹⁾。

ある符号帳クラスタの重心を u_m 、対応する学習音声クラスタの重心を v_m 、このときのクラスタの総数を j とする。 u_m と v_m のずれベクトルを p_m であらわす。

$$p_m = v_m - u_m \quad (m=1, \dots, j) \quad (1)$$

各符号帳要素 c_i について、すべての符号帳クラスタの重心 u_m との距離を求め、その逆数を重み $\{w_{im}\}$ としたすべてのずれのベクトル $\{p_m\}$ の和を次のようにして求める。

$$\Delta_i = \frac{\sum_m w_{im} p_m}{\sum_m w_{im}} \quad (2)$$

$$w_{im} = 1 / \|c_i - u_m\| \quad (3)$$

この Δ_i を用いて、 c_i を、

$$c_i' = c_i + \Delta_i \quad (4)$$

へ移動する。

図4に、4クラスタまでの階層的適応化処理の様子を示す。

6.4 実験結果

日本語100都市名を認識対象とし、特徴パラメータとして、ケプストラムのみを用いる場合(Cep)と、これにケプストラムおよび対数パワーの動的特徴²⁾を組合せる場合(Cep & Δ Cep)2種類の実験を行った。

表1は認識誤り率を、適応化なし(不特定話者)、10単語音声を用いて適応化、話者毎に全100単語を用い

標準話者の符号帳

学習音声

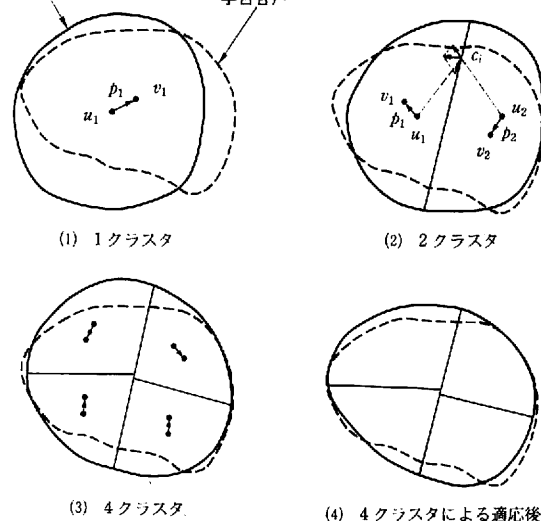


図4 隣接クラスタ間の連続性を考慮した、符号帳の階層的適応化処理

表1 単語音声認識実験の結果(平均誤認識率)

特徴 パラメータ	条件	適応なし	適応あり (10単語学習)	全単語学習
Cep		4.90 %	2.88 %	2.15 %
Cep & Δ Cep		1.95	1.55	0.70

て学習(特定話者)の3種類の条件の場合について示したものである²²⁾。(Cep)を用いる場合、適応化前に4.90%であった誤り率が、適応化により2.88%まで低下しており、この適応化法が比較的少数の音声サンプルによる学習でも高い有効性を持つことがわかる。

男性、女性、子供を含む広い範囲の話者変動に対処するための種々の検討も行われている²³⁾²⁴⁾。この適応化方法は原理的に、SPLIT法による単語音声認識だけでなく、種々の音声認識系に広く適用することが可能であり、雑音や伝送系によって変形したスペクトルへの適応などにも用いることが可能と考えられる。

7. ニューラルネットによる実現の試み

標準話者と未知話者の、発声内容が既知の音声の時間軸の対応付けを基本に用いる点は5章の方法と同じであるが、音声をベクトル量子化せずに、特徴ベクトル空間をそのまま扱って、一方から他方の話者の空間への非線形連続写像を実現する階層型ニューラルネットを構成する試みが行われている³¹⁾。標準話者と未知

話者の音声(特徴ベクトル時系列)のDPによる時間軸の対応付けの結果として、両者の話者の特徴ベクトル対を多数求め、これらをニューラルネットの入力層と出力層に与えて、バックプロパゲーション学習によってネットワークを構成する。

HMMを用いた音声認識系によって評価を行った結果、学習単語と未知単語が同じ内容(語彙)の場合に、250単語の認識率が適応化によって64.8%から95.2%に向上し、異なる語彙である場合に100単語の認識率が82.0%から99.0%に向上することが確認された。ただし今のところ学習に必要とする単語の数が非常に多いのが問題である。

8. む す び

音声認識の研究において、最近ホットな話題のひとつとなっている話者適応化技術の動向について紹介した。人間の視覚や聴覚は極めて高度な環境への適応能力あるいは順応性を有しており、この能力が視聴覚による外界からの情報の受容や、人間と人間のコミュニケーションにおいて重要な役割を果たしている。いわゆるカクテルパーティ効果なども、人間の音声知覚における適応能力の一種であると考えられることができる。一方、我々がこれまでに実現してきた種々のパターン認識アルゴリズムや装置は、このような能力に極めて乏しいものであったといえる。今後これらの装置をいっそう高度化し、人間の能力に近づけるためには、このような観点からの研究に力を入れることが重要になるであろう。

(1989年4月13日受付)

【参 考 文 献】

- 1) 古井：ディジタル音声処理，東海大学出版会(1985)
- 2) S. Furui: "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Trans., ASSP-34, 1, pp. 52-59 (Feb., 1986)
- 3) L. R. Rabiner, S. E. Levinson, A. E. Rosenberg and J. G. Wilpon: "Speaker-Independent Recognition of Isolated Words Using Clustering Techniques", IEEE Trans., ASSP-27, 4, pp. 336-349 (Aug., 1979)
- 4) L. R. Rabiner, S. E. Levinson and M. M. Sondhi: "On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition", Bell Syst. Tech. J., 62, 4, pp. 1075-1105 (Apr., 1983)
- 5) K. F. Lee and H. W. Hon: "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM", Proc. ICASSP 88, New York, S3.7 (1988)
- 6) 千葉，亘理，渡辺："音声認識システム"，大型プロジェクト・パターン情報処理システム研究開発成果発表会論文集，pp. 157-165 (1980)
- 7) 加藤，古井："話者の個人性に対する聴覚の適応特性"，音響学会聴覚研資，H 85-5 (Jan., 1985)
- 8) 古井："音声認識における個人差の学習法について"，音響学会音声研資，S75-25 (Nov., 1975)
- 9) H. Matsumoto and H. Wakita: "Vowel Normalization by Frequency Warped Spectral Matching", Speech Communication, 5, 2, pp. 239-251 (June, 1986)
- 10) 好田，斎藤："部分的な学習サンプルによる単語音声の認識"，信学会パターン研資，PRL72-87 (Dec., 1972)
- 11) S. Furui: "A Training Procedure for Isolated Word Recognition Systems", IEEE Trans., ASSP-28, 2, pp. 129-136 (Apr., 1980)
- 12) 古井："音声波に含まれる声質情報と韻質情報の分離"，音響学会音声研資，S74-17 (June, 1974)
- 13) 中川，白方，三樹，坂井："個人差の教師なし学習機能を持つ60都市名の実時間認識"，音響学会音声研資，S77-11 (June, 1977)
- 14) 杉山，好田："母音標準バタンの教師なし学習法"，音響学会音声研資，S83-48 (Dec., 1983)
- 15) 溝口，木下，角所："セット化音韻テンプレートに基づく不特定話者単語音声認識システム"，信学論，J67-A, 6, pp. 572-579 (June, 1984)
- 16) 中島，高橋："大語彙音声認識における話者適応化"，秋季音学講論，1-1-6 (1983)
- 17) K. Shikano, K-F. Li and R. Reddy: "Speaker Adaptation through Vector Quantization", Proc. ICASSP 86, Tokyo, 49.5 (1986)
- 18) 中村，鹿野："セパレートベクトル量子化を用いたスペクトログラムの正規化"，音響誌，44, 8, pp. 595-602 (Aug., 1988)
- 19) 中村，鹿野："ファジィベクトル量子化を用いたスペクトログラムの正規化"，音響誌，45, 2, pp. 107-114 (Feb., 1989)
- 20) 新美，小林："ベクトル量子化のコードブックの話者適応化"，秋季音学講論，2-5-13 (1987)
- 21) 山下，松本："単語音声認識におけるベクトル量子化誤差を利用した話者適応"，音響学会音声研資，SP87-118 (Jan., 1988)
- 22) 古井："スペクトル空間の階層的クラスタ化による話者適応化"，音響学会音声研資，SP88-21 (June, 1988)
- 23) 古井："マルチテンプレートと話者適応化による音声認識"，春季音学講論，2-6-10 (1989)
- 24) S. Furui: "Unsupervised Speaker Adaptation Method Based on Hierarchical Spectral Clustering", Proc. ICASSP 89, Glasgow, S6.9 (1989)
- 25) R. Schwartz, Y-L. Chow and F. Kubala: "Rapid Speaker Adaptation Using a Probabilistic Spectral Mapping", Proc. ICASSP 87, Dallas, 15.3 (1987)
- 26) 西村，菅原，年岡，大河内："HMMにおける話者適応化"，音学講論，2-5-17 (1987)
- 27) 菅村，古井："擬音韻標準ボタンによる大語彙単語音声認識"，信学論，J65-D, 8, pp. 1041-1048 (Aug., 1982)
- 28) 鹿野："入力音声のベクトル量子化による単語音声認識"，音響学会音声研資，SP82-60 (Dec., 1982)
- 29) 白木，菅田："セグメント符号化における話者適応化"，音響学会音声研資，SP87-67 (Oct., 1987)
- 30) Y. Linde, A. Buzo and R. M. Gray: "An Algorithm for Vector Quantization Design", IEEE Trans., COM-28, 1, pp. 84-95 (Jan., 1980)
- 31) 磯，麻生川，吉田，渡辺："ニューラルネットワークによる話者適応"，春季音学講論，1-6-16 (1989)



ふるい さいおき
古井 貞昭 昭和45年、東京大学工学部計数工学科大学院修士課程修了。同年、NTT電気通信研究所に入社。以後、同研究所において音声認識、話者認識、音声知覚などの基礎研究に従事。現在、NTTヒューマンインタフェース研究所音声情報研究部長、工学博士。