

論文 / 著書情報
Article / Book Information

論題(和文)	不特定話者音声の認識
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	電子情報通信学会誌, Vol. 73, , pp. 1286-1290
Citation(English)	, Vol. 73, , pp. 1286-1290
発行日 / Pub. date	1990, 12
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 1990 Institute of Electronics, Information and Communication Engineers.

小特集

2-3

2. 音声認識技術の動向

不特定話者音声の認識

古井貞熙

古井貞熙：正員 NTT ヒューマンインタ
フェース研究所Speaker-Independent Speech Recognition Techniques. By Sadaoki FURUI, Member (NTT Human In-
terface Laboratories, Musashino-Shi 〒180).

あぶすとらくと

音声にはさまざまな個人差があり、これに対処して不特定話者の音声を高い精度で認識できるようにすることは、音声認識の重要な課題の一つである。この具体的方法としては、話者による変動の比較的少ない音素特徴や尺度を用いる方法、話者による変動を複数のクラス（テンプレート）や確率モデルで表現する方法、新しい話者が発声した少数の単語や短文音声を用いて、複数の標準話者クラスターのいずれに近いかを選択する方法、標準話者の音声との対応関係（マッピング）を自動的に学習して、標準音声スペクトルを新しい話者に適応化する方法などが提案されている。

1. 音声に含まれる個人性情報

音声波には、声道長などの発声器官の違いや、アクセントなどの発話習慣の違いに起因する個人情報が、複雑にからみあった形で含まれている。発声者や年齢による母音のホルマント周波数の変動には、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⁽¹⁾。

音声認識において、話者が変わっても高い認識性能を得る最も確実な方法は、その話者の音声

を用いて学習をする方法（特定話者型）であるが、すべての単語をあらかじめ発声しなければならないというのでは不便であるし、電話を用いたバンキングサービス、番号案内などのサービスでは、このようなことは不可能な場合が多い。このため、不特定話者の音声を認識するための方法の研究が、これまでに多くの研究者によって行われて来た。

2. 不特定話者音声認識法

2.1 話者による変動の少ない特徴や距離尺度を用いる方法

個人差の少ない有効な特徴、すなわちどのような人が発声しても同じ音素あるいは単語である限り共通に含まれていて、しかも異なる単語を区別するのに有効な音韻性的特徴を見だし、それを手がかりに認識を行うのが、最も基本的な方法である。スペクトログラムリーディングなどに基づいた知識工学的な方法で、このような特徴を見いだす研究も行われている⁽²⁾。但し、類似した方言で通常容易に理解できる音声であっても、すべての人の音声にこのような共通の物理的特徴があるという保証は必ずしもな

用語解説

スペクトログラムリーディング スペクトルの時間変化を2次元的に表示したスペクトログラムを目で見て、種々の音韻特徴を読みとること。

削除内挿法 限られた標本の集合から母集団を推定する問題において、標本の集合より一部を削除した残りの部分から特徴量を求め、削除部分の内挿の良否によって、その特徴量の評価や重み付けをする方法。

い。

これまでに有効性の高い特徴として確認されているものに、スペクトルの動的特徴⁽³⁾がある。これは既に多くの不特定話者音声認識システムに用いられ、認識性能の向上が確認されている。相互情報量などに基づいて、不特定話者の音素を識別するのに有効な判別の特徴や距離尺度を構築し、これを用いて認識を行う方法も試みられている。

2.2 マルチテンプレート法

図1に示すように、各単語ごとに、多数話者の学習用音声のクラスタ化によって、個人差による変動をカバーするような複数 (M) のテンプレートを作成し、未知入力音声とこれらすべてとの類似度を計算する方法⁽⁴⁾をマルチテンプレート法とよぶ。認識対象語彙数が大きい場合にこの方法を用いると、装置の規模が膨大になってしまう問題があるが、これに対してはSPLIT法⁽⁵⁾が提案されている。SPLIT法では、音素標準パターンはすべての話者に共通で、単語辞書のみ各単語について複数用意する。距離計算は入力音声と音素標準パターンとの間で行

われるので、認識のための計算量と標準パターンのための記憶容量が少なくすむ特徴がある。この方法は、NTTのバンキング用ANSER音声認識応答システムで用いられている。

ニューラルネットの技術を用いて、ネットワークに種々のパターンを自動的に覚えこませ、同様の機能を実現する研究も行われている⁽⁶⁾。

2.3 HMMによる方法

HMM (隠れマルコフモデル) (本小特集「統計的手法による認識」参照)の学習に多数話者の音声を用いることにより、個人差による音素やその遷移の変動を、確率モデルとして柔軟にかつ能率よく表現できる。HMMによる単語認識では、各単語がその音素数程度の少ない状態数で表現されるので、上記のマルチテンプレート法よりも、認識のための処理量が少なくすむ特徴がある。HMMを用いた不特定話者連続音声認識で最も成功した例は、カーネギーメロン大学のSPHINXシステム⁽⁷⁾であり、特定話者なみの認識率が報告されている。

3. 話者適応・正規化

3.1 話者適応・正規化法の分類

上で述べたように、同じ単語や音素を発声した音声でも、すべての話者に共通の物理的特徴がある保証はなく、話者が変わると異なる音素が同じスペクトルを持つということもあるので、完全な不特定話者の認識には限界がある。このため最近、話者適応化 (正規化) 技術、すなわち発声者が変わるたびに少数の単語あるいは短い文章の音声を用いて、自動的に学習を行う方法が盛んに研究されている。

話者適応化の方法は、あらかじめ学習音声として用いる言葉を決めておく教師ありの方法と、任意の音声 (単語、短文章など) を用いて学習を行う教師なしの方法に分類できる。後者の場合は、未知入力音声をそのまま学習に用いることもできるので、ユーザにとっては、あたかも不特定話者の音声認識システムであるかのように動作させることができる。個人性情報の

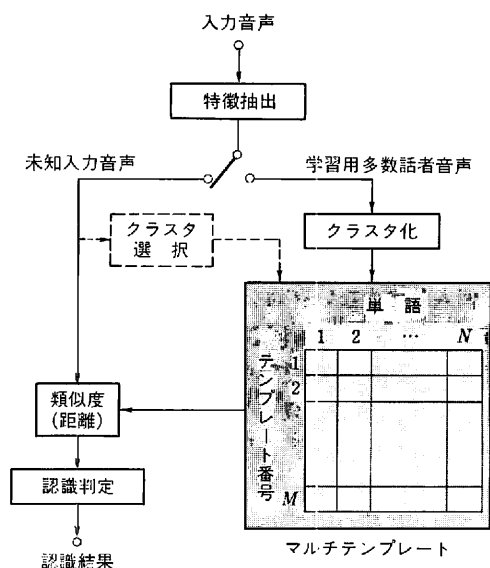


図1 マルチテンプレート法による単語音声認識のブロック図
図 学習音声のクラスタ化により、各単語ごとに複数のテンプレート (クラスタ) を作成して用いる。破線部は、後述のクラスタ選択法の場合に付加される部分を示す。

含まれる方は、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。

3.2 話者クラスタ選択法

2.2で述べたマルチテンプレート法の一種の発展として、図1に破線の部分を追加した方法が研究されている。似た話者を同じクラスタに分類しておき、入力音声に対して最も近いクラスタを選択して、そのクラスタの符号帳やモデルを用いる方法⁽⁸⁾である。話者選択機能を持ったニューラルネットを構成する試み⁽⁵⁾も行われている。

3.3 特徴量内挿法

多数話者から求めた不特定話者用のパラメータは、推定の信頼性は高いが特定の話者に必ずしも適合せず、その話者について学習したパラメータは、その話者への適合性は高いが推定の信頼性が低いので、両者を適切に内挿するのが一般的に望ましい。この内挿の度合を自動的に決定する具体的方法として、削除内挿法⁽¹⁰⁾が提案され、有効性が確認されている。

3.4 符号帳適応・正規化法

種々のスペクトルをベクトル量子化の符号帳によって表現し、標準の符号帳またはある特定話者の符号帳と、新しい未知話者の符号帳の要素との対応関係（マッピング）を学習音声を用いて求めて、一方を他方に合せる適応化方法が研究されている。各単語の単語辞書あるいはモデルに関しては、すべての話者に共通の1種類のモデルを用いる場合と、話者による発声の仕方の違いを代表する複数のモデル（マルチテンプレート）を用意する場合とがある。

符号帳のマッピング規則を学習する方法として、教師ありの方法と教師なしの方法が提案されている。教師ありの方法⁽¹¹⁾のブロック図を図2に示す。この方法では、未知話者と標準話者の同じ単語の時間軸の対応関係を調べ、これから両話者の符号帳要素の対応関係を求める。全学習単語に関して、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

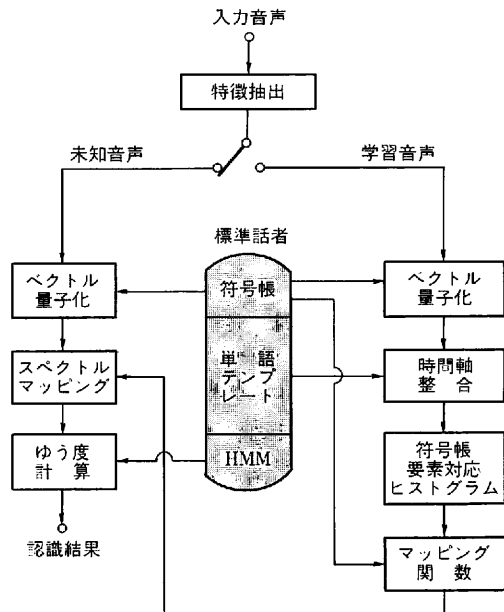


図2 教師あり学習による符号帳の話者適応化を含む単語音声認識のブロック図 未知話者と標準話者の同じ単語の時間軸整合に基づいて、符号帳のマッピング関数を求める。

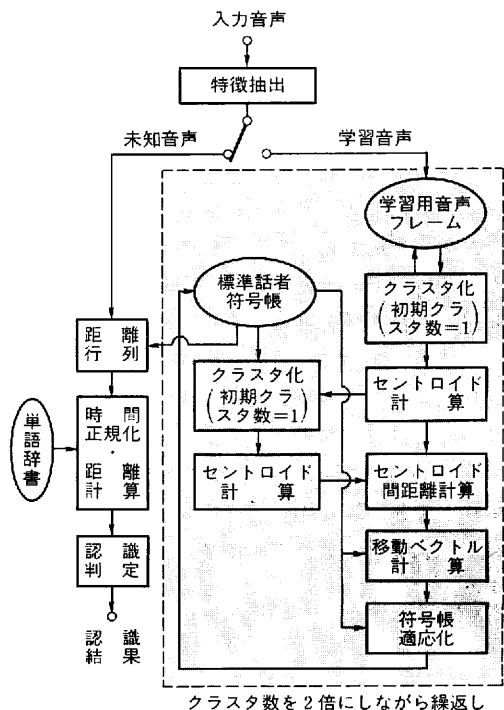


図3 教師なし学習による符号帳の話者適応化を含む単語音声認識のブロック図 未知話者のスペクトルと標準話者の符号帳の階層的クラスタ化に基づいて符号帳の適応化を行う。

よる重みからなるマッピング関数(行列)を決定する。認識すべき未知入力音声に対しては、ベクトル量子化したのち、各フレームごとにマッピングを行って標準話者の音声に近付ける(適応化する)。その後標準話者の各単語との類似度を調べて、何の単語が発声されたかを決定する。

教師なしの方法^{(12),(13)}のブロック図を、図3に示す。この方法では、スペクトル集合の階層的クラスタ化によってマッピング関数を求める。学習音声スペクトルと符号帳を、階層的にクラスタ化し、対応する各クラスタ重心(セントロイド)間のずれに基づいて、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。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、話者適応を進めることができる。符号帳要素を繰り返し適応化する際、その周辺のクラスタの重心のずれを考慮して、クラスタ間のずれの連続性が保持されるように移動ベクトルの平滑化を行う。

3.5 ニューラルネットによる方法

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、上で述べた教師あり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ニューラルネットを構成する試み⁽¹⁴⁾も行われている。標準話者と未知話者の音声(特徴ベクトル時系列)の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。

4. む す び

話者適応・正規化法は、少数であったにせよ学習音声を必要とするという意味では、完全な不特定話者認識系に比べると使い勝手は劣るが、教師なしであればユーザに意識されずに学

習ができるので、ユーザにとっての煩わしさを避けることができる。実際の音声認識においては、個人差のほか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も適応できることが必要であるが、個人差への適応化アルゴリズムは、これらの変動要因にも適用できる場合が多いと考えられる。今後は、完全な不特定話者認識系の性能向上と共に、それに組み合わせる話者適応化法の高度化を進めて行くことが必要である。

文 献

- (1) 古井貞照: "ディジタル音声処理", 東海大学出版会(1985-09)。
- (2) N. Carbonell, et al.: "APHODEX, Design and implementation of an acoustic-phonetic decoding expert system", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Tokyo, 23.3 (1986)。
- (3) S. Furui: "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans. Acoust., Speech, Signal Process., ASSP-34, 1, pp.52-59 (1986)。
- (4) L.R. Rabiner, et al.: "Speaker-independent recognition of isolated words using clustering techniques", IEEE Trans. Acoust., Speech, Signal Process., ASSP-27, 4, pp.336-349 (1979)。
- (5) 菅村 昇, 古井貞照: "擬音韻標準パターンによる大語い単語音声認識", 信学論 (D), J 65-D, 8, pp.1041-1048 (1982-08)。
- (6) H. Sakoe, et al.: "Speaker-independent word recognition using dynamic programming neural networks", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Glasgow, SI.8 (1989)。
- (7) K.-F. Lee: "Automatic speech recognition: The development of the SPHINX system", Kluwer Academic Publishers (1989)。
- (8) 杉山雅英, 鹿野清宏: "母音標準パターンの教師なし学習法", 音響学会音声研資, S83-48 (1983-12)。
- (9) J.B. Hampshire II and A. H. Waibel: "The Meta-Pi network: Connectionist rapid adaptation for high-performance multi-speaker phoneme recognition", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Albuquerque, S.3.9 (1990)。
- (10) F. Jelinek and R. L. Mercer: "Interpolated estimation of Markov source parameters from sparse data", in Pattern Recognition in Practice, ed. E.S. Gelsema and L.N. Kanal, pp.381-397, North-Holland Publishing Co., Amsterdam (1980)。
- (11) K. Shikano, et al.: "Speaker adaptation through vector quantization", Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Tokyo, 49.5 (1986)。
- (12) 白木善尚, 譽田雅彰: "セグメント符号化における話者適応化", 信学技報, SP 87-67 (1987-10)。
- (13) 古井貞照: "マルチテンプレートと教師なし話者適

応化による音声認識”, 信学論 (A), J72-A, 10, pp.1476-1483 (1989-10).

- (14) 磯 健一, 麻生川 稔, 吉田和永, 渡辺隆夫: “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 (1989-03).



古井 貞熙 (正員)

昭 43 東大・工・計数卒。昭 45 同大学院修士課程了。電電公社電気通信研究所入社。以来、音声認識、話者認識等の研究に従事。現在、NTT 田 研究所音声情報研究部長。工博。米沢賞。論文賞。著述賞。日本音響学会論文賞。IEEE 論文賞。科技庁長官賞受賞。著書「デジタル音声処理」など。

