論文 / 著書情報
Article / Book Information

論題(和文)	話者認識研究の現状と展望
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	テレビジョン学会技術報告, Vol. 20, No. 41, pp. 19-24
Citation(English)	ITEJ Technical Report, Vol. 20, No. 41, pp. 19-24
発行日 / Pub. date	1996, 7
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) 1996 Institute of Image Information and Television Engineers.

1996年7月16日(火) 発表　テレビジョン学会技術報告　ITE Technical Report Vol. 20, No. 41, PP. 19 ~24.
MIP' 96-56, NIM' 96-78(July. 1996)

話者認識研究の現状と展望

Recent speaker recognition research and future work

松井知子 † 古井貞熙 ††
Tomoko Matsui † Sadaoki Furui ††
† NTT ヒューマンインタフェース研究所 ‡ 東京工業大学
† NTT Human Interface Laboratories ‡ Tokyo Institute of Technology

あらまし 話者認識は、発声者が誰であるかを自動的に判定する技術である。本論文では、話者認識の基本的な説明をするとともに、テキスト指定型話者照合方法について解説する。テキスト指定型話者照合方法は近年注目を集めている方法で、装置を用いるたびに装置側から新しい言葉を指定することによって、テープレコーダなどによる悪用を防ぐことができる。更に近年、高品質マイク、静かな環境で録音したデータに関しては、多くの研究機関が99%以上の高い話者認識率を報告するようになったが、背景雑音、回線歪み、事前閾値などに関する課題が残されていることを紹介する。最後に、話者認識研究の展望について述べる。

和文キーワード: 話者認識、話者照合、話者識別、テキスト指定型、事前閾値、隠れマルコフモデル

Abstract Speaker recognition is the process of automatically recognizing who is speaking. This paper introduces basic speaker recognition methods and explains a text-prompted speaker verification method. The text-prompted method, in which the system prompts each user with a new key sentence every time the system is used, is widely used these days because it is unnecessary to worry about the system being fooled by recordings of key words or sentences uttered by the user. Moreover, this paper shows that recent speaker recognition techniques achieve high performance with speech data recorded using a high-quality microphone in a noise-free room, but still have some technical problems related to background noise, line distortion, and a priori threshold. It also introduces some technical issues that remain to be solved in the future.

Key words: speaker recognition, speaker verification, speaker identification, text-prompted, a priori threshold, hidden Markov model

---

† 東京都武蔵野市緑町 3-9-11
TEL: 0422-59-2387, FAX: 0422-60-7808
E-Mail: matsui@splab.hil.ntt.jp, furui@splab.hil.ntt.jp
‡ 東京都目黒区大岡山 2-12-1
TEL, FAX: 03-5734-3480

1 はじめに

一般に話者認識とは、コンピュータなどにより、発声者が誰であるかを自動的に判定することをいう。これが実現できれば、車や建物の入口の鍵、電話を用いたバンキングや買い物サービス、情報提供サービス、コンピュータのリモートアクセス、クレジットカードコールなどにおいて、音声による本人確認ができるようになる。

話者認識技術は、図1に示すように話者識別と話者照合に分類することができる[1]。話者識別では、入力された音声が登録話者中の誰であるかを判定する。話者照合では、ユーザは音声を入力するとともに、自分は誰であるかを申告する。そして、その音声が本当にその申告された話者の声であるかどうかを判定する。現在考えられている話者認識の応用例のほとんどは、話者照合に該当する。上述の車や建物の入口の鍵、バンキングサービスにおける本人確認なども、話者照合の応用例である。

また、話者認識技術は、テキスト依存型[2]、独立型[2]、指定型[3]の三つに分類することができる。テキスト依存型では、ユーザは“ひらけごま”のような決まった言葉（キーワード）を発声する。テキスト独立型では、ユーザは任意の言葉を発声してよい。しかし、これらの二つの方法では、テープレコーダなどによってユーザの声を録音し、認識装置の前で再生すれば、装置がそれを受理してしまう危険性がある。テキスト指定型では、認識のたびに、任意のテキストを装置側から指定する。この方法では、本人が指定されたテキストを正しく発声したと判定できる時のみ、それを受理する。このため、テープレコーダなどによる悪用を防ぐことができる。

本論文では、まず話者認識方法について概説した後、近年注目を集めているテキスト指定型話者照合方法について説明する。そして、話者認識研究の現状としていくつかの研究例を紹介するとともに、その展望について述べる。

2 話者認識方法

話者認識システムでは一般的に、まず各話者ごとに、予め決められた複数の単語や文章を発声してもらい、その音声データを用いて各話者のモデルを生成（学習）する。話者識別では、ユーザが認識用に発声した音声と、全ての登録話者のモデルとの類似度を調べ、そのユーザは一番大きい値の登録話者であるとする。話者照合では、ユーザが発声した音声と、申告された話者のモデルとの類似度を調べ、その値を一定の閾値と比較する。そして、その値が閾値よりも大きければ本人の声として受理し、小さければ他人（詐称者）の声として棄却する。なお、ここ数年、話者のモデルとして、隠れマルコフモデル（hidden Markov model; HMM）[4]がよく利用されている。HMMは統計的な手法の一つで、各話者ごとに、その学習データを用いて生成されたHMMは、発声変動や時間的な伸縮を統計的な形でモデル内部に反映できるという特徴を持つ。

また、話者識別における誤りは、入力音声の話者として、登録話者中の本人以外の話者が選ばれる誤りのみであるが、話者照合では、本人を他人と判定する誤り（本人棄却率）と他人を本人と判定する誤り（詐称者受理率）の二種類の誤りを考える必要がある。図2は、本人棄却率と詐称者受理率の関係を表したものである。例えば閾値を大きくすると、詐称者受理率は低くなるが、本人棄却率は高くなる。

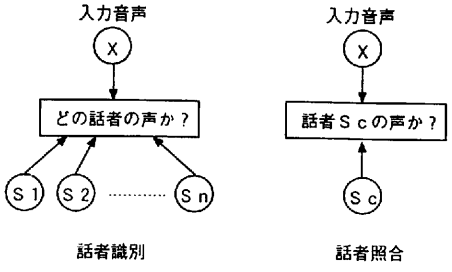

図 1: 話者識別と話者照合

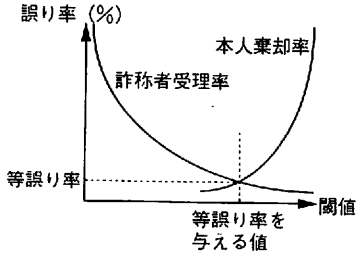

図 2: 本人棄却率と詐称者受理率の関係

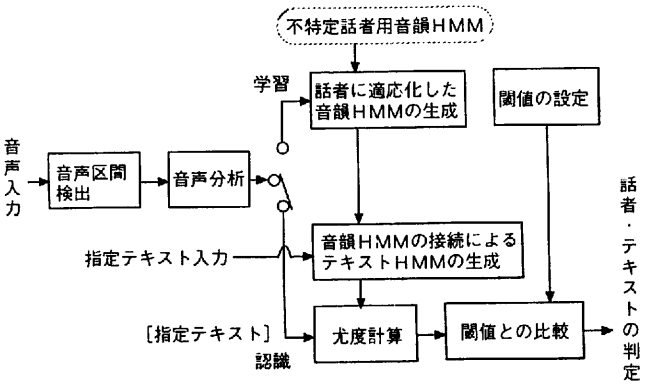

図 3: テキスト指定型話者照合システムの構成

話者照合システムの目的によっては、本人棄却率の方が詐称者受理率よりも重要であったり、またその逆であったりするが、予めその目的が明確でない場合には、本人棄却率と詐称者受理率が等しくなるように閾値を設定するのが一般的である。（このことはベイズ識別の意味で最適な判別を行うことに相当する。）

閾値の設定方法に関しては、これまでいくつかの研究例 [14][15] はあるが、閾値を事前に安定して設定することは難しく、現在のところ、決定的な方法はまだない。そのために話者照合実験においては、認識データから計算した本人棄却率と詐称者受理率が等しくなる値（等誤り率を与える値）に閾値が仮に設定できたものとして、その等誤り率で評価する場合が多い。（なお、この閾値は“事後的に設定した閾値”としばしば言われる。）その場合、話者照合率は 100 %から等誤り率を差し引いた値となる。しかしながら、実際には閾値は事前に設定しておく必要がある。

以下、テキスト指定型話者照合方法について説明する。

2.1 テキスト指定型話者照合方法

図3に基本的なテキスト指定型話者照合システムの構成を示す [3]。テキスト指定型では、本人が正しく指定されたテキストを発声したか否かを判定するので、本人がそのテキストを発声した場合の音声のモデルを生成する必要がある。システムの側から任意のテキストが指定できるためには、各話者ごとにすべての基本的な音韻（/a/ や /i/ などの言葉の短い単位）のモデルを学習しておき、それを接続して、テキスト通り発声した音声のモデルを生成すればよい。このため、まず学習では、各話者ごとに 10 文章程度の発声をしてもらう。各発声について音声区間を検出し、音声分析して、特徴ベクトル（ケプストラム、Δケプストラムなど）の時系列を得る。そして、その時系列を各音韻ごとに例えば left-to-right 型の HMM によってモデル化する。left-to-right 型の HMM は、（各状態に対応した）音声の音響的事象が時間につれて逆戻りしないモデルなので、その時系列固有の時間的な構造がそのまま反映される（図4）。

その際、各音韻モデルに如何に正確に音韻性・話者性を持たせるかが重要な課題である。10 文章程度の学習データだけでは、各音韻モデル（HMM）のパラメータを推定するには十分ではない。しかし、話者認識では各話者に多量の発声を要求するような方法は現実的ではないので、これ以上文章の数を増やすのは適当ではない。そこで、予め収録しておいた多数の一般話者が発声したデータ（話者 60 名、計

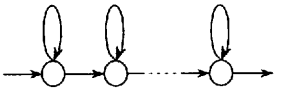

図 4: Left-to-right HMM

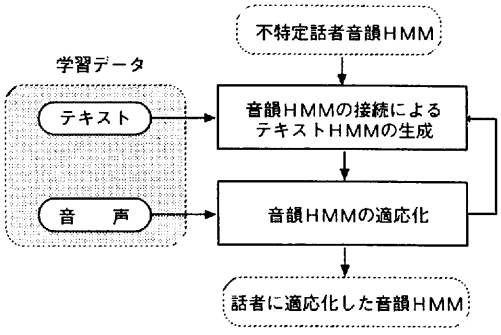

図 5: 話者に適応化した音韻モデルの生成法

9000 文章）から、一般話者（不特定話者）用音韻モデルを生成する。この不特定話者用音韻モデルは、各話者共通の音韻に関する情報がたくさん含まれたモデルであると考えられる。そのモデルを初期モデルとし、各話者の学習データを用いて、その初期モデルをその話者に適応化する。こうして、その話者の音韻モデルを生成する。

図5にその処理の流れを示す。学習データのテキストは既知なので、そのテキストに従って不特定話者音韻モデルを接続して、そのテキストのモデル（テキストモデル）を生成する。それに音声を与えて、Baum-Welch アルゴリズム [4] により、テキストモデル中のパラメータを推定する。その後、そのテキストモデルを各音韻モデルに分ける。以上を全学習データについて何度か繰り返し、各音韻モデルのパラメータを話者に適応化する。

認識では、システム側から指定したテキストに従って、本人の音韻モデルを連結し、そのテキストのモデルを作る。そして、入力音声から得られた特徴ベクトルの時系列が、そのテキストモデルから出現する尤度を計算し、その値を閾値と比較して、話者の判定を行なう。

筆者らの実験 [3] によれば、表1に示す実験条件において、事後的に設定した閾値を用いた場合、99 %の話者照合率が得られる。

表 1: 実験条件

時期差	10 カ月にわたる 5 時期
話者数	登録話者 15 名（男 10、女 5）、 詐称者は登録話者のうち本人を除く 14 名
音声帯域	0-6 kHz
特徴量	16 次 LPC ケプストラム
マイク	同一コンデンサーマイクロホン
学習	10 文章／話者
認識	1 文章（平均 4 秒）、5 回テスト／話者

3 話者認識研究の現状

現在のところ、高品質マイク、静かな環境で録音したデータに関しては、ケプストラムを特徴量として用いて、本人棄却率と詐称者受理率が等しくなる値に事後的に設定した閾値によって判定した場合、99 %以上の話者照合率が得られている。しかしながら、背景雑音がある場合、実際の電話回線を通した場合、事前に設定した閾値を用いた場合などには話者認識率は著しく低下する。それらの問題に関しては、近年盛んに研究されている。

背景雑音がある場合について

Rose らは、話者 16 名の会話データを用いた、HMM によるテキスト独立型話者識別実験において、静かな環境では識別率は 96 %であるが、信号対雑音比（signal-to-noise ratio; SNR）10 dB の白色雑音下では 14 %に低下すると報告した [6]。この報告によると、該当する SNR で音声モデルと雑音モデルを合成することにより、耐雑音性に優れたモデルを生成する HMM 合成法 [7][8] を適用すれば、識別率は 69 %まで回復する。この報告にある方法では学習時に、各話者のモデルとして、実際に認識を行う環境とは別の静かな環境で録音した学習データを用いて混合ガウス分布モデル（1 状態の HMM）を生成する。

また、認識では、認識を行う環境で録音した雑音データから混合ガウス分布モデル（1 状態の HMM）を生成する。HMM 合成法により、話者モデルと雑音モデルの両分布を、入力音声の SNR に応じた重みで重畳し、雑音重畳話者モデルを生成する。そして入力音声を、この雑音重畳話者モデルに与えた時の尤度によって話者の判定を行う。

HMM 合成法では、SNR が既知である必要があるが、雑音が非定常な場合には、SNR を正確に測定することは困難である。そこで筆者らは、SNR が未知の場合にも有効な方法について検討した [9]。この方法では、学習時に複数の SNR で話者モデルと雑音モデルを合成したモデルを用意し、認識では入力

音声に対するそれら複数の合成モデルの尤度を計算し、その最大値を各話者の尤度とする。登録話者10名、詐称者10名の音声データを用いて、この方法の評価を行った結果、SNRが未知の場合でも、十分高い話者認識性能が得られることが確認された。

実際の電話回線を通した場合について

実際の電話回線を利用する場合、使用する回線はハンドセット（マイク）を含めて毎回必ずしも同じではない。回線が異なると、その回線の特性によって声は歪むが、その歪み方もまた異なる。Rosenbergらは、話者40名が発声した連続数字データを用いた、HMMによるテキスト指定型話者照合実験において、学習と認識とで同じ回線を通した場合には照合率は99％であるが、異なる回線を通した場合には92％に低下すると報告した[10]。この報告の中では、ケプストラム正規化技術を用いれば、異なる回線を通した場合でも95％まで回復することが示されている。

ケプストラム正規化技術は、特徴量としてケプストラムの時系列を用いた場合、回線歪みに相当するケプストラムのバイアス成分を推定し、その成分を観測ベクトル時系列から差し引くことによって、回線歪みによる影響を少なくするものである。ケプストラムのバイアス成分を推定する方法としては、各発声ごとに、その全区間にわたるケプストラムの平均を計算し、それをバイアス成分とする方法[11][14]、各発声について、ある短い区間ごとにケプストラムの平均を計算し、その短い区間ごとにバイアス成分を求める方法[12]、各発声を相当する音韻HMMに与えた時の尤度が最大となるようにバイアス成分を推定する方法[13]の三つが考えられている。

事前に設定した閾値を用いた場合

筆者らは、話者20名が発声した文章データを用いた、HMMによるテキスト指定型話者照合実験において、事後的に設定した閾値を用いた場合には照合率は99％であるが、学習データから計算した等誤り率を与える値に事前に設定した場合には66％になることを報告した[15]。この報告では、以下の事前に閾値を設定する方法を提案している。一般的に学習データが少ない場合は、認識データの本人棄却率を精度よく推定することは難しく、学習データから計算した等誤り率を与える値に事前に閾値を設定した場合には、認識データの本人棄却率は高くなる。この問題に対処するために[15]の方法では、学習データから計算した等誤り率を基準とし、そのシステムの性能から実験的に推定される高めの詐称者受理率（すなわち低めの本人棄却率）を与える値に事前に閾値を設定する。この方法に従って閾値を事前に設定した場合には、照合率は96％まで回復する。

以上述べたように、種々の条件下で話者認識率は著しく低下する。現在のところ、背景雑音に関してはHMM合成法が、マイクの違いや回線歪みに関してはケプストラム正規化技術が有効であると考えられ、これらの技術を適用することによって誤り率が減少することは確かめられている。話者照合の閾値を事前に設定する方法に関しては、[14][15]で検討されているが、それらの方法は実験的に調整しなければならないパラメータを含んでおり、それらのパラメータの調整が話者によっては難しいという問題がある。

4 話者認識研究の展望

今後は、周囲雑音や回線歪みのある実環境下でも、高い認識性能を得るための技術が一層重要視されると思われる。特に話者照合の閾値の設定方法に関しては、今後も引続き検討していく必要がある。また、話者認識方法を評価する場合に、閾値を事前に安定して設定できるかどうかも重要なことなので、今後は事前に設定した閾値を用いて評価を行うべきである。

一方、現在のところ、特徴量としてケプストラム、話者のモデルとしてHMMが標準的に用いられている。音声に含まれる話者性情報には、静的な特徴（声の定常的な特性、声の音色など）と動的な特徴（声の時間的な変化特性、話し方など）があるが、ケプストラムは前者を表す特徴量である。今後は、動的な特徴を有効に表す特徴量についても検討を進める必要がある。同時に、それをうまくモデル化する方法についても検討する必要がある。また、声が時間の経過につれて変わるという自然の摂理は、話者認識システムにとって重大な問題となるため、声の時期的な変動のしくみの解明、および話者のモデルの定期的な更新法[15][14][16]について引続き検討していく必要がある。更に、音声に含まれる

話者性情報と音韻性情報との関係は、今だに完全には明らかにされていない。話者性情報を音韻性情報といかに分離して抽出するかが今後の課題である。以上のように次の技術レベルに到達するために、特徴量、話者のモデル化、声の時期的な変動、話者性情報と音韻性情報との関係などに関する基礎的な研究が必要である。

5 まとめ

本論文では、話者認識法についてテキスト指定型話者照合方法を例にとって解説し、話者認識研究の現状、展望について述べた。現在の話者認識技術を用いれば、高品質マイク、静かな環境で録音したデータに関しては、事後的に設定した閾値を用いた場合に99％以上の高い話者認識率が得られる。今後は実環境下で事前に設定した閾値を用いた場合でも、高い認識性能を得るための技術に関する検討、および更なる技術進歩を目指して、特徴量、話者のモデル化、声の時間的な変動、話者性情報と音韻性情報との関係等に関する基礎的な研究を引続き行っていく必要があると思われる。

参考文献

[1] 古井貞熙、ディジタル音声処理、東海大学出版会、東京、1985.

[2] 古井貞熙、音響・音声工学、近代科学社、東京、1992.

[3] 松井知子、古井貞熙、テキスト指定型話者認識、電子情報通信学会論文誌8、J79-D-II、5、pp. 647-656、1996.

[4] 中川聖一、確率モデルによる音声認識、電子情報通信学会、1988.

[5] J.J. Webb and E.L. Rissanen, *Speaker identification experiments using HMMs*, Proc. ICASSP, pp. II-387-390, 1993.

[6] R.C. Rose, E.M. Hofstetter and D.A. Reynolds, *Integrated models of signal and background with application to speaker identification in noise*, IEEE Transactions on SA 2, pp. 245-257, 1994.

[7] F. Martin, K. Shikano, Y. Minami, and Y. Okabe, *Recognition of noisy speech by using the composition of hidden Markov models*, 電子情報通信学会技報, SP92-96, 1992.

[8] M. J. F. Gales and S. J. Young, *HMM recognition in noise using parallel model combination*, Proceedings of Eurospeech-93, Berlin, pp. II-837-840, 1993.

[9] T. Matsui, K. Kanno and S. Furui, *Speaker recognition using HMM composition in noisy environments*, Computer Speech and Language, 1996, to appear.

[10] A.E. Rosenberg, C.H. Lee and F.K. Soong, *Cepstral channel normalization techniques for HMM-based speaker verification*, Proc. ICSLP, pp. IV-1835-1838, 1994.

[11] B. S. Atal, *Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification*, J. Acoust. Soc. Am., Vol. 55, pp.1304-1312, 1974.

[12] J. de Veth, G. Gallopyn, H. Bourlard, *Limited parameter hidden Markov models for connected digit speaker verification over telephone channels*, Proc. ICASSP, pp. II-247-250, 1993.

[13] A. Sankar and C. -H. Lee, *Stochastic matching for robust speech recognition*, IEEE Signal Processing Letters, Vol. 1, pp.124-125, 1994.

[14] S. Furui, *Cepstral analysis technique for automatic speaker verification*, IEEE Trans. Acoust., Speech, Signal Processing, ASSP-29,2, pp.254-272, 1981.

[15] 松井知子、西谷隆、古井貞熙、話者認識におけるモデルとしきい値の更新法の検討、日本音響学会春季研究発表会、東京、1-5-6、1996.

[16] A. Setlur and T. Jacobs, *Results of a speaker verification service trial using HMM models*, Proc. Eurospeech, pp. 1-53-56, 1995.