論文 / 著書情報
Article / Book Information

論題(和文)	自由テキストによる発声内容依存型話者認識法
Title(English)	
著者(和文)	古井 貞熙, 松井 知子
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会1992年秋季講演論文集, , , pp. 51-52
Citation(English)	, , , pp. 51-52
発行日 / Pub. date	1992, 10

1—7—3　自由テキストによる発声内容依存型話者認識法*

○古井貞熙　松井知子（NTTヒューマンインタフェース研究所）

1. まえがき

話者認識の方法には、あらかじめ定められたキーワードを用いる発声内容依存型と、任意の言葉を用いる発声内容独立型がある。前者は、認識系の構成が比較的簡単で、しかも、音韻や音節に依存した個人性情報を用いることができるため認識精度が高いという特徴がある。しかしこの方法には、キーワードを発声した音声をテープレコーダなどで録音し、それを認識装置の前で再生すれば、その本人に簡単になりすますことができるという欠点がある。

これに対処するため、複数の単語をキーワードとし、認識のたびに、それらの単語を新たな順序で並び変えた系列を認識装置の側から示して、それを発声させるという方法が、試みられてきた[1][2]。しかしこれでも、近年の電子化した記憶装置を用いれば、キーワードの任意の系列を再生でき、十分とは言えない。

ここでは、各話者の音韻クラスを基本的な単位として用いることによって、認識のたびに任意の新しい文章を装置側から指定し、本人がその文章を正しく発声した時だけ受理する方法を提案する。この方法により、高い性能で話者が認識できるとともに、本人の音声でも、指定とは異なる内容の録音音声を棄却することができる。次章以下で、その具体的方法を提案する。この場合の重要な課題の一つは、限られた量の学習用音声から、如何にしてその話者の音声に適合した音韻モデルを作成するかということである。

2. 基本的構成

2.1 方法1

図1は、方法1のブロック図を示したものである。この方法では、まず各話者の学習音声を用いて、テキストに独立な話者モデルを、1状態混合ガウス分布HMMで表現する。次に学習音声を再度用いて、音韻クラスモデルを作成する。音韻クラスの数と種類については、あらかじめ決めておく。学習音声のテキストは既知であるので、そのテキストを音韻クラスの接続で表現すれば、学習音声から、各音韻クラスモデルを自動的に学習することができる[3]。このとき、初期モデルとして、その話者の混合ガウス分布を用い、その各ガウス分布の重みのみを学習する。従って、各音韻クラスモデルも、1状態の混合ガウス分布HMMで表現される。

認識をする場合は、まず装置の側から、発声すべきテキスト（学習時には発声しなかった言葉でよい）を指示し、話者はこの指示に従って発声する。装置の側では、音韻クラスモデルを接続して、テキスト全体のHMMを作成する。このHMMに対する入力音声の尤度を計算し、その値に基づいて、話者および発声内容の判定を行なう。

上記では、話者の特徴を表現する方法として、混合ガウス分布を用いる場合について説明したが、ベクトル量子化の符号帳を用いる場合も、同様に構成することができる。

2.2 方法2

上で説明した方法1では、各話者の学習音声のみを用いて音韻クラスモデルを作成している。このため方法1では、学習音声の量の制限から、音韻クラスの数をあまり大きくすることができない。これに対して、不特定話者の音声認識系で作成した音韻モデルを各話者の音声に適応化する方法で、その話者の音韻モデルを作成するのが方法2である。

本方法では、まず方法1と同様に、各話者の学習音声を用いて、テキストに独立な話者モデルを、1状態混合ガウス分布HMMで表現する。次に、不特定話者の音韻モデル（混合ガウス分布を用いた連続型HMM）を、上記の話者モデル、あるいはその話者の学習音声を再度用いて、その話者に適応化する。こうして、話者に適応化した音韻モデルができれば、認識の過程は、方法1と同様に行なうことができる。方法1と同様に、話者モデルとして、混合ガウス分布を用いる代わりにベクトル量子化の符号帳を用いることもできる。

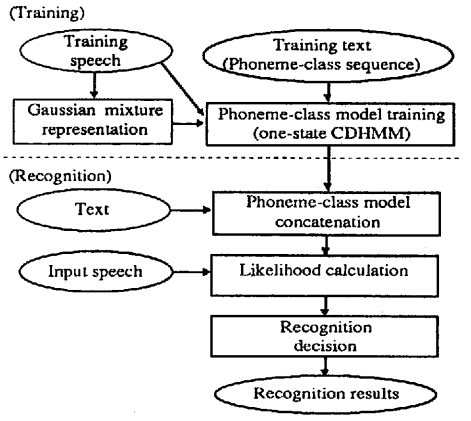

Fig. 1 - Block diagram of Method 1.

* Text-dependent speaker recognition methods using free texts
by Sadaoki FURUI and Tomoko MATSUI (NTT Human Interface Laboratories)

2.3 方法3

これまでの方法1と2では、各話者の音韻クラスモデル、あるいは音韻モデルを共通に用いて、話者認識と発声内容の判定を行なっている。これに対して、話者の判定とテキストの判定を並列的に別々に行なって、両者の類似度から総合的に話者の判定を行なうのが方法3である。

この方法では、まず各話者の学習音声を用いて、テキストに独立な話者モデルを作成する。認識の過程では、まず、話者モデルに対する入力音声の尤度を計算する。同時に、入力音声のテキストに従って、不特定話者の音韻モデルを接続したモデルを作成し、それに対する入力音声の尤度を計算する。この両者の尤度を論理的、あるいは数値的に組み合わせて、最終的な話者の判定を行なう。具体的には、(a)両者の尤度の重み付き平均値を計算する、(b)後者の尤度が、あるしきい値よりも大きい場合のみ、前者の尤度に従って話者の判定を行なう、などの方法が考えられる。

3. 方法1による認識実験[4]

3.1 実験方法

次のようにして、方法1に関する認識実験を行なった。実験には、約6か月にわたる3時期（A, B, C）に、15名（男性10名、女性5名）が発声した種々の文章音声を用いた。1文章の平均的長さは4秒である。特徴パラメータとしては、16次のLPCケプストラムを用いた。まず、各話者の時期Aの音声（10文章）を用いて、発声内容に独立な話者モデルを、対角共分散、1状態64混合ガウス分布HMMによって作成した。次に、音韻クラスの数を、1、6、15、25、の4通りに変えて、各話者の音韻クラスモデルの学習を行なった。

認識には、時期BとC（学習音声との時期差は3および6か月）の種々の文章音声を用いた。テストサンプルの総数は150（5文章×15話者×2時期）である。

指定したテキストと異なる内容の音声を棄却するために、尤度の値にしきい値を設定した。このしきい値は、指定したテキスト通りに発声した音声が棄却されることがないという条件を満たす最大値に、事後的に設定した。しかし尤度の値は認識すべき音声のテキストや、声質の時期的変化によって大きく変動するため、異なる時期の音声に共通の絶対的なしきい値を決定するのは極めて難しい。ここでは、入力音声と各話者との組み合わせ毎に、テキストの制約を与えなかったときの尤度、すなわち各話者の混合ガウス分布をそのまま用いたときの尤度の値を差し引くことによって、尤度のばらつきの正規化を試みた。

3.2 実験結果

(1) 正しい文章を発声した場合の結果

話者照合の平均誤り率を表1に示す。照合の判定のしきい値は、各登録話者ごとに、2種類の誤り（詐称者受理率と本人棄却率）が等しくなる値に、事後的に設定した。この結果から、15〜25の音韻クラスに分けることによって、平均誤り率が約4／5に低下することがわかる。話者識別誤り率は、音韻クラス数を変えても変化がなく、平均誤り率は4.7%であった。

Table 1 - Verification error rates

	Number of phoneme classes		
	1	15	25
Session B	4.4%	3.4%	3.4%
Session C	4.1%	3.7%	3.7%
Average	4.3%	3.6%	3.6%

(2) 異なる文章を発声した場合の結果

HMMに対する尤度を直接用いた場合に、本人が、指定したテキストと異なる文章を発声した音声を棄却できる割合を表2に示す。音韻クラスを増やすと棄却性能が向上するが、クラス数を25にしても、尤度の正規化を行なわないと、半分程度しか棄却することができない。尤度の正規化を行なえば、誤りが正規化前の約半分に低下し、異なる発声内容の音声の約80%を棄却することができる。

Table 2 - Different sentence rejection rates

	Number of phoneme classes			
	6	15	25	25*
Session B	41.3%	63.7%	69.0%	82.0%
Session C	37.7%	54.0%	60.7%	81.3%
Session B + Session C	24.3%	39.7%	48.3%	80.7%

* With likelihood normalization

4. 結び

本報告では、認識のたびに任意の文章を指定することができる、信頼性の高い話者認識法を提案した。実験により、ここで提案した方法のうち、方法1が高い話者認識性能を有し、しかも、本人の音声でも指定したテキストとは異なる場合には、高い精度で棄却できることが確認できた。今後、方法2および3に関する評価実験の他、安定した個人性情報を豊富に持つ音韻を重視した距離尺度（類似度）等について検討して行きたい。

謝辞

熱心に討論して下さった研究室の諸氏に感謝する。

文献

[1] A.E.Rosenberg & F.K.Soong,in *Advances in Speech Signal Processing*, pp.701-738 (1992)

[2] J.M.Naik, *IEEE Comm. Magazine*, 28, pp.42-48 (1990)

[3] K.F.Lee & H.W.Hon, *Proc. IEEE ICASSP*, pp.123-126 (1988)

[4] 松井・古井、信学秋季大会 (1992)