論文 / 著書情報
Article / Book Information

論題(和文)	テキスト指定形話者認識システムの試作
Title(English)	
著者(和文)	松井 知子, 古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	1994年日本音響学会秋季講演論文集, , No. 1-8-14, pp. 27-28
Citation(English)	, , No. 1-8-14, pp. 27-28
発行日 / Pub. date	1994, 10

テキスト指定形話者認識システムの試作 *

⊙ 松井知子　　古井貞熙
(NTT ヒューマンインタフェース研究所)

1　まえがき

筆者らは、信頼性の高い話者認識を実現するために、テキスト指定形話者認識法を提案してきた [1]-[3]。この方法では、認識のたびに、システム側から任意のテキストを指定し、本人がそのテキストを正しく発声したと判定できる時のみ、その声を受理する。そのために、テープレコーダ等による録音音声によって、システムがだまされる危険性が少ない。

この度、テキスト指定形話者認識システムを試作し、現在、研究所内の展示ホールのドアの開閉に用いている。本論文では、システムの概要について述べるとともに、実環境でその性能を評価した結果について述べる。

2　システム概要

2.1　システム構成

本システムは、インターフェース装置、主装置の二つの部分によって構成される（図1）。インターフェース装置はパネル PC（コンテック社製）、主装置は DECpc（DEC 社製、α XP150）によって制御し、専用 DSP 等は一切使用していない。なお、マイクはエレクトレット型の単一指向性マイク（プリモ社製）を使用した。

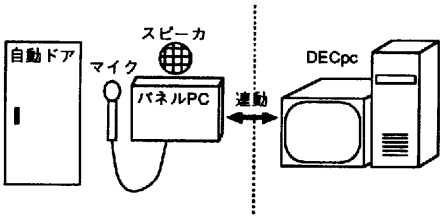

図 1: システム構成

2.2　登録時の処理

図2に登録時の処理フローを示す。登録処理には次の四つのモードがある。

- 話者登録
- 削除
- 正規化用モデル計算
- しきい値計算

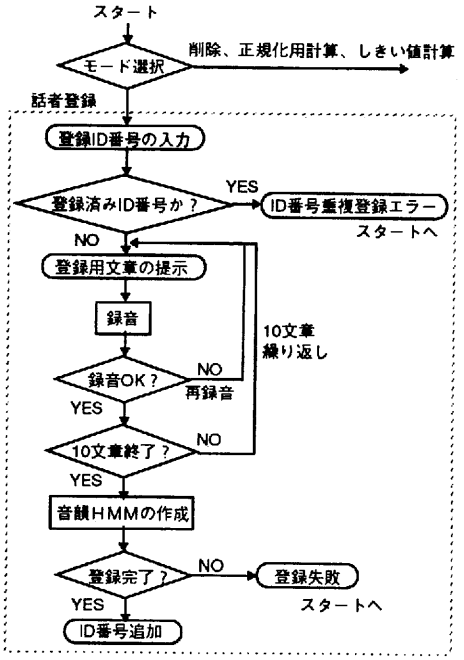

図 2: 登録時の処理

登録モードでは、まず登録する ID 番号を入力する。次にシステム側から、登録用の 10 文章（約 4 秒／文）を逐次提示して、ユーザに発声してもらい、そのユーザの音韻 HMM を作成する。削除モードでは、削除したい ID 番号を入力する。実際には、安全のために削除の意思を一度確認する。正規化用モデル計算モードでは、尤度正規化用の音韻・話者独立モデルを作成する（3 節参照）。しきい値計算モードでは、数文章をユーザに発声してもらい、その時の尤度の値からしきい値を調整する。

2.3　認識時の処理

図3に認識時の処理フローを示す。まず、ユーザは自分の ID 番号を入力する。システム側ではその ID 番号の登録を確認した後、あらかじめ定めた認識用の多数文章（約 1 秒／文）から無作為に一文を選択し、ユーザに提示するとともに、そのテキストに従って、そのユーザの音韻 HMM を連結し、そのテキストのモデルを生

*Implementation of a text-prompted speaker recognition system.
By Tomoko Matsui and Sadaoki Furui (NTT Human Interface Laboratories)

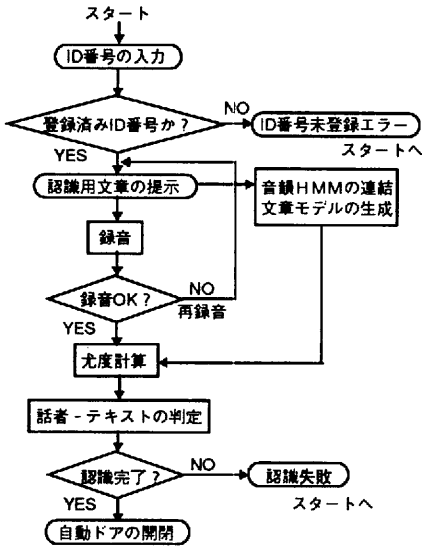

図 3: 認識時の処理

成する。次いで、ユーザの入力音声をそのテキストのモデルに与えた時の尤度を計算し、話者およびテキストの判定を行う。その文がそのユーザによって正しく発声されたと判定できる場合のみ、システムに接続されている自動ドアが開く。

3 要素技術 — 尤度正規化 —

テキストに応じて連結した HMM の尤度の値は、テキストや発声時期の違いによって大きく変動する。このことは、話者・テキスト照合において、全ての音声に共通なしきい値を安定して設定することを難しくする。

ここでは、尤度の値のばらつきを次式で表される事後確率に基づく尤度正規化法 [2][3] によって正規化する。

$$p(s_c,t_c|x) = \frac{p(x|s_c,t_c)\times p(s_c,t_c)}{\sum_i\sum_j\{p(x|s_i,t_j)\times p(s_i,t_j)\}} \approx \frac{p(x|s_c,t_c)}{\sum_i\sum_j p(x|s_i,t_j)}$$

x は入力音声を、s_i は話者の識別子、特に s_c は申告された話者を示す。t_j はテキストの識別子、特に t_c はシステム側から指定されたテキストを示す。$p(s_i,t_j)$ は、話者 s_i、テキスト t_j の同時確率であり、全話者、全テキストの組み合わせについて共通の値であると仮定し、キャンセルする。$p(x|s_c,t_c)$ は、申告話者 s_c の音韻 HMM をシステム側から指定されたテキスト t_c に従って連結した HMM に対する入力音声 x の尤度によって表される。$\sum_i\sum_j p(x|s_i,t_j)$ を求めるために必要とされる計算量は全登録話者数に比例するので、単一の音韻・話者独立モデルを作成し、その一つのモデルに対する尤度で上記の累積尤度を置き換える。この置き換えにより、正規化に必要な計算量は従来の累積尤度を計算する場合に比べ、全登録話者数分の一となる。

表 1: 話者・テキスト照合平均誤り率 (%).

話者	男1	男1	女2	女2	平均
誤り率	0.0	5.7	0.0	7.1	3.2

4 実験

システムの評価は男 5 名、女 5 名で男女別に行った。そのうち男 2 名、女 2 名を登録話者として用い、各登録話者に対する詐称者は、その他の男 3 名／女 3 名とその話者以外の登録話者の計 4 名とした。特徴パラメータとして、標本化周波数 12kHz、フレーム長 32ms、フレーム周期 16ms、LPC 分析次数 16 でケプストラムを抽出した。登録には、ある 1 時期に発声した 10 文章（1 文章長は平均約 4 秒）を用い、認識では、それとは異なる 20 文章（1 文章長は平均約 1 秒）から 1 文章づつ、10 数回無作為に選択して用いた。なお、登録作業はテストの 2 〜 3 カ月前に行った。音韻数は 41 である。音韻 HMM として、3 状態、256 混合分布の半連続型 HMM を用いた。HMM パラメータは Baum-Welch アルゴリズムを用いて推定した。しきい値は、本実験では各話者ごとに、2 種類の誤り（詐称者受理率、本人棄却率）が等しくなる値に事後的に設定した。なお、音声区間の平均 SNR は 32.8dB であった。

5 結果

表 1 に話者・テキスト照合誤り率を示す。この平均誤り率が、筆者らが今まで行ってきた実験結果（平均誤り率 1 %以下）[1]-[3] と比べ、大きい原因としては、テスト用の文章長が短いこと（従来：約 4 秒／文 → 今回：約 1 秒／文）などが考えられる。

6 まとめ

今回、テキスト指定形話者認識システムを試作し、実環境での評価を行った。話者 4 名で、登録から数カ月後に評価した結果、96.8 %の話者・テキスト照合率が得られた。

今後はしきい値を事前に設定する方法について検討していきたい。

参考文献

[1] 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 (1993)*

[2] 松井知子、古井貞熙、“テキスト指定形話者認識のための事後確率に基づく尤度正規化法、” 日本音響学会秋季研究発表会、*1-7-20*、山梨 *(1993)*

[3] 松井知子、古井貞熙、“音韻・話者独立モデルによる話者照合尤度の正規化、” 信学技報、*SP94-22 (1994)*