論文 / 著書情報
Article / Book Information

論題(和文)	VoIP 携帯端末を利用した音声認証技術の検証
Title(English)	
著者(和文)	土屋 直樹, 山口 泰広, 福本 博文, 岩野 公司, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本人間工学学会 モバイル人間工学研究部会 シンポジウム「ケータイ・カーナビの利用性と人間工学」論文集, Vol. , No. , pp. 109-114
Citation(English)	, Vol. , No. , pp. 109-114
発行日 / Pub. date	2004, 3
Note	rights: ここに掲載した著作物の利用に関する注意　本著作物の著作権は一般社団法人日本人間工学会に帰属します。本著作物は著作権者である日本人間工学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」に従ってください。

VoIP 携帯端末を利用した音声認証技術の検証

○土屋 直樹*，山口 泰広*，福本 博文*，岩野 公司**，古井 貞熙**
*オムロンソフトウェア株式会社、**東京工業大学大学院　情報理工学研究科

Speech Identification in VoIP (Voice over IP) System

Naoki TSUCHIYA*, Yasuhiro YAMAGUCHI*, Hirofumi FUKUMOTO*
Koji IWANO**, Sadaoki FURUI**
*OMRON SOFTWARE Co., Ltd, **Tokyo Institute of Technology

Abstract: In this paper, we present an evaluation of speech identification in VoIP (Voice over IP) system.

Recently, VoIP system grows popular because of the inexpensive communication costs. As we can create some additional information in VoIP data, we expect to be able to provide personalized services using VoIP system. But there is no method to identify users in VoIP system. So, we need some identification method to achieve such services.

Consequently, we evaluated speech identification in VoIP system, and we made sure that the speech identification is useful as a identification method in VoIP system.

Keywords: VoIP, Speech Identification

キーワード:VoIP, 音声認証

1. はじめに

近年、IP 網を利用し、音声データを伝送する VoIP（Voice over IP）が、従来の公衆回線に変わる新しい通話方式として、注目を集めている[1]。これは、IP 網を利用することにより、通信コストが削減されることや、伝送データに付加情報を付与することが可能であるためである。

特に、伝送データに付加情報を付与する特長により、今後、VoIP システムを利用した個人向けにカスタマイズされたサービスの提供が期待されている。しかしながら、現在、VoIP システム上で利用することができる個人の識別手段は少ない。これは、たとえばクレジット決裁など、本人を確定しなければ提供できないサービスを実現するためには不可欠な要素である。そこで、今回は、音声を利用した個人の識別（音声認証）を評価した。

2. 認証システム

本節では認証システムについて述べる。図 1 に開発したシステム構成図を示す。

本システムは以下のステップにより認証を行う。

① VoIP プロトコルによりサーバと端末との通話を確立する。

② ユーザは端末に向け既定の 6 桁数字を発話する。

③ サーバは前記発話音声を検出し、サーバ上にファイル出力する。

④ サーバは前記発話音声ファイルを認証する。

⑤ サーバは認証結果を端末に伝送する。

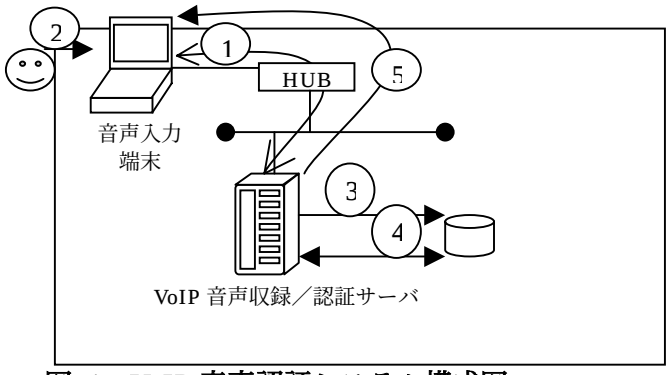

図 1　VoIP 音声認証システム構成図

上記認証処理では音声に含まれる 5 つの個人識別に有効と思われる要素（以下、認証子）、すなわち声紋（発話音声の周波数特性）、言語（発話内容）、アクセント、ピッチ（音声の高低）、発話長（ひとつの数字の発話に要する発話時間）を利用して認証を行う。音声に含まれる複数の認証子を利用するのは、複数の要素を利用することで、音声認識の精度が向上するという報告があり[2]、これらの認証子が有効であると想定するためである。以降、各々の認証子を利用した認証方式を、声紋認証、言語認証、アクセント認証、ピッチ認証、発話長認証と呼ぶ。

次に各認証方式の詳細を示す。

3. 認証方式

以下に、音声に含まれる認証子の認証方式を示す。認証方式は、大きく 3 つのステップにより構成する。

1. 各認証子のスコアを算出する。
2. 1 で算出したスコアから、言語認証と声紋認証の重みつき和を計算し、既定の閾値との大小関係により、第一段階の認証結果を判定する。
3. 第一段階の認証結果により、本人であると判定することが難しい場合、さらに残りの 3 つの認証子により得られるスコアの重みつき和を計算し、既定の閾値との大小関係により、第二段階の認証結果を判定する。なお、第一段階で認証を行った結果が受理された場合、第二段階の認証は行わない。

次に、上記 3 つのステップの詳細を示す。

3.1 各認証子におけるスコアの算出

以下に、各々の認証子（声紋、言語、アクセント、ピッチ、発話長）におけるスコアの計算方法を示す。

3.1.1. 声紋認証

音声の声紋情報である個人の周波数特性を利用した認証を行う。認証方法は以下の通りである。

(1) 6 桁数字が任意の組み合わせ w で発話されると想定した特定話者モデル S_w による発話 V の出現確率 $P(V|S_w)$ を算出する。

(2) 6 桁数字が任意の組み合わせ w で発話されると想定した不特定話者モデル W_w による発話 V の出現確率 $P(V|W_w)$ を算出する。

(3) $P(V|S_w)$ と $P(V|W_w)$ は、出現確率を対数値に変換した値で出力し、前者より後者を減算することで正規化を行う。この値を声紋認証スコア S_v として算出する。

$$S_v = \log P(V|S_w) - \log P(V|W_w)$$

3.1.2. 言語認証

音声の言語情報である発話内容を利用した認証である。認証方法は以下の通りである。

(1) 6 桁数字が特定の組み合わせ c で発話されると想定した不特定話者モデル W_c による発話 V の出現確率 $P(V|W_c)$ を算出する。

(2) 6 桁数字が任意の組み合わせ w で発話されると想定した不特定話者モデル W_w による発話 V の出現確率 $P(V|W_w)$ を算出する。

(3) $P(V|W_c)$ と $P(V|W_w)$ は、出現確率を対数値に変換した値で出力し、前者より後者を減算することで正規化を行う。この値を言語認証スコア S_l として算出する。

$$S_l = \log P(V|W_c) - \log P(V|W_w)$$

3.1.3. アクセント認証

音声のアクセント情報を利用した認証を行う。アクセントの情報として、ここでは、話者の抑揚の大きさ（観測されるピッチの範囲）のみに注目する。特徴量としては、有声区間のピッチの分散値を用いる。認証方法は以下の通りである。

(1) 特定の数字の組み合わせが入力されると想定し、各数字の有声音区間を切り出し、i 番目に発話された数字の有声音区間のピッチ分散 $x_{\delta i}$ を算出する。

(2) 同一発話者が過去に発声した数字 k のピッチ分散の平均値 μ_{as_k}、ピッチ分散の分散値 δ_{as_k} を算出する。

(3) 上記の μ_{as_k} と δ_{as_k} から特定話者の k に対するピッチ分散の正規分布 $f_{as}(x,k)$ を導出する。

$$f_{as}(x,k) = \frac{1}{\sqrt{2\pi}\delta_{as_k}} \exp\left\{-(x-\mu_{as_k})^2 / 2\delta_{as_k}{}^2\right\}$$

$$for \quad k = 0,1,\ldots,9.$$

(4) 多数の発話者が過去に発声した数字 k のピッチ分散のピッチ分散の平均値 μ_{aw_k}、ピッチ分散の分散値 δ_{aw_k} を算出する。

(5) 上記の μ_{aw_k} と δ_{aw_k} から不特定話者の k に対するピッチ分散の正規分布 $f_{aw}(x,k)$ を導出する。

$$f_{aw}(x,k)=\frac{1}{\sqrt{2\pi}\delta_{aw_k}}\exp\left\{-(x-\mu_{aw_k})^2/2\delta_{aw_k}{}^2\right\}$$
$$for \quad k=0,1,...,9.$$

(6) $x_{\delta i}$ と $f_{as}(x,k)$ をもとに $x_{\delta i}$ の出現確率 $f_{as}(x_{\delta_i},k_i)$ を求め、対数値に変換し、i 番目の数字 k_i の特定話者ピッチ分散尤度 S_{as_i} とする。

$$S_{as_i} = \log \quad f_{as}(x_{\delta_i}, k_i)$$

(7) i 番目に発話された数字のピッチ分散 $x_{\delta i}$ と不特定話者ピッチの正規分布 $f_{aw}(x,k)$ をもとに $x_{\delta i}$ の出現確率 $f_{aw}(x_{\delta_i},k_i)$ を求め、対数値に変換し、i 番目の数字 k_i の不特定話者ピッチ分散尤度 S_{aw_i} とする。

$$S_{aw_i} = \log \quad f_{aw}(x_{\delta_i}, k_i)$$

(8) 各数字の特定話者ピッチ分散尤度 S_{as_i} から不特定話者ピッチ分散尤度 S_{aw_i} を減算し、正規化を行い、全ての数字の総和を取ることで、アクセント認証スコア S_a とする。

$$S_a = \sum_{i=1}^{6}\{S_{as_i} - S_{aw_i}\}$$

3.1.4. ピッチ認証

音声のピッチ情報を利用した認証を行う。認証方法は以下の通りである。

(1) 特定の数字の組み合わせが入力されると想定し、各数字の有声音区間を切り出し、i 番目に発話された数字の有声音区間のピッチ平均 x_{μ_i} を算出する。

(2) 同一発話者が過去に発声した数字 k のピッチの平均値 μ_{ps_k}、ピッチの分散値 δ_{ps_k} を算出する。

(3) 上記の μ_{ps_k} と δ_{ps_k} から特定話者の数字 k に対する特定話者ピッチの正規分布 $f_{ps}(x,k)$ を導出する。

$$f_{ps}(x,k)=\frac{1}{\sqrt{2\pi}\delta_{ps_k}}\exp\left\{-(x-\mu_{ps_k})^2/2\delta_{ps_k}{}^2\right\}$$
$$for \quad k=0,1,...,9.$$

(4) 多数の発話者が過去に発声した数字 k のピッチの平均値 μ_{pw_k}、ピッチの分散値 δ_{pw_k} を算出する。

(5) 上記の μ_{pw_k} と δ_{pw_k} から特定話者の数字 k に対する特定話者ピッチ平均の正規分布 $f_{pw}(x,k)$ を導出する。

$$f_{pw}(x,k)=\frac{1}{\sqrt{2\pi}\delta_{pw_k}}\exp\left\{-(x-\mu_{pw_k})^2/2\delta_{pw_k}{}^2\right\}$$
$$for \quad k=0,1,...,9.$$

(6) i 番目に発話された数字の $x_{\mu i}$ と $f_{ps}(x,k)$ をもとに $x_{\mu i}$ の出現確率 $f_{as}(x_{\mu_i},k_i)$ を求め、対数値に変換し、i 番目の数字 k_i の特定話者ピッチ分散尤度 S_{ps_i} とする。

$$S_{ps_i} = \log \quad f_{ps}(x_{\mu_i}, k_i)$$

(7) i 番目に発話された数字の $x_{\mu i}$ と $f_{pw}(x,k)$ をもとに $x_{\mu i}$ の出現確率 $f_{pw}(x_{\mu_i},k_i)$ を求め、対数値に変換し、i 番目の数字 k_i の不特定話者ピッチ平均尤度 S_{pw_i} とする。

$$S_{pw_i} = \log \quad f_{pw}(x_{\mu_i}, k_i)$$

(8) 各数字の特定話者ピッチ分散尤度 S_{ps_i} から不特定話者ピッチ分散尤度 S_{pw_i} を減算し、正規化を行い、全ての数字の総和を取ることで、アクセント認証スコア S_p とする。

$$S_p = \sum_{i=1}^{6}\{S_{ps_i} - S_{pw_i}\}$$

3.1.5. 発話長認証

音声の発話長を利用した認証を行う。認証方法は以下の通りである。

(1) 特定の数字の組み合わせが入力されると想定し、各数字の有声音区間を切り出し、各数字の有声音区間の発話長 x_{s_i} を算出する。

(2) 同一発話者が過去に発声した数字 k の発話長の平均値 μ_{ss_k}、発話長の分散値 δ_{ss_k} を算出する。

(3) 上記の μ_{ss_k} と δ_{ss_k} から数字 k について、特定話者の特定話者発話長の正規分布 $f_{ss}(x,k)$ を導出する。

$$f_{ss}(x,k)=\frac{1}{\sqrt{2\pi}\delta_{ss_k}}\exp\left\{-(x-\mu_{ss_k})^2/2\delta_{ss_k}{}^2\right\}$$
$$for \quad k=0,1,...,9.$$

(4) 多数の同一発話者が過去に発声した数字 k の発話長の平均値 μ_{sw_k}、発話長の分散値 δ_{sw_k} を算出する。

(5) 上記の μ_{sw_k} と δ_{sw_k} から数字 k について、不特定話者発話長の正規分布 $f_{sw}(x,k)$ を算出する。

$$f_{sw}(x,k)=\frac{1}{\sqrt{2\pi}\delta_{sw_k}}\exp\left\{-(x-\mu_{sw_k})^2/2\delta_{sw_k}{}^2\right\}$$
$$for \quad k=0,1,...,9.$$

(6) i 番目に発話された数字 k_i の発話長 x_{s_i} と特定話者発話長の正規分布 $f_{ss}(x,k)$ をもとに x_{s_i} の出現確率 $f_{ss}(x_{s_i},k_i)$ を求め、対数値に変換し、i 番目の数字 k_i の特定話者発話長平均尤度 S_{ss_i} とする。

$$S_{ss_i} = \log \quad f_{ss}(x_{s_i}, k_i)$$

(7) i 番目に発話された数字 k_i の発話長 x_{s_i} と不特定話者発話長の正規分布 $f_{sw}(x,k)$ をもとに x_{s_i} の出現確率 $f_{sw}(x_{s_i},k_i)$ を求め、対数値に変換し、i 番目の数字 k_i の不特定話者発話長平均尤度 S_{sw_i} とする。

$$S_{sw_i} = \log \quad f_{sw}(x_{s_i}, k_i)$$

(8) 各数字の特定話者発話長分散尤度から不特定話者発話長分散尤度を減算し、正規化を行い、総和をとり、発話長認証スコア S_s とする。

$$S_s = \sum_{i=1}^{6}\{S_{ss_i} - S_{sw_i}\}$$

3.2 言語認証および声紋認証を利用した認証

本節では、前節における、言語認証および声紋認証を利用し、各々のスコアを評価する方式を示す。

評価方法は、以下の手順に従う。

① 第一段階における総合的なスコア S_1 を言語認証スコア S_l および声紋認証スコア S_v の重みつき和により計算する。

$$S_1 = \sum_m \omega_m S_m \qquad for \quad m = l, v$$
$$where \quad \sum_m \omega_m = 1 \qquad for \quad m = l, v.$$

② 上記重みつき演算により得られるスコアと既定の閾値 θ_{1h}、θ_{1l} との大小関係により、認証結果 $result_1$ を判定する。ここで、認証結果は、3つ存在する。すなわち、受理（Accept：本人であると断定する。）、判定困難（Gray：本人であるか他人であるかの判定が難しい。）、拒否（Reject：本人ではないと断定する。）である。

$$result_1 = \begin{cases} accept & where \quad S_1 \geq \theta_{1h} \\ gray & where \ \theta_{1h} > S_1 \geq \theta_{1l} \\ reject & where \quad S_1 < \theta_{1l} \end{cases}$$

3.3 アクセント認証、ピッチ認証、発話長認証を利用した認証

本節では、3.2 において、判定が困難と判断した場合に、アクセント認証、ピッチ認証、発話長認証を用いて本人の判定を行う手順を示す。評価方法は、以下の手順に従う。

① 第二段階における総合的なスコア S_2 を S_a、S_p、S_s の重みつき和により計算する。

$$S_2 = \sum_m \omega_m S_m \qquad for \quad m = a, p, s$$
$$where \quad \sum_m \omega_m = 1 \quad for \quad m = a, p, s.$$

② 重みつき演算により得られるスコア S_2 と既定の閾値 θ_2 との大小関係により、認証結果 $result_2$ を判定する。ここで、認証結果は、2 つ存在する。すなわち、受理（Accept）、拒否（Reject）である。

$$result_2 = \begin{cases} accept \ where \ S_2 \geq \theta_2 \\ reject \ where \ S_2 < \theta_2 \end{cases}$$

以上 3.1 から 3.3 の処理により、$result_1$ または、$result_2$ が受理であった場合、認証結果を受理とし、$result_2$ が拒否であった場合、認証結果を拒否とする。

4. 評価条件

今回は、前節の認証方式を用いて、**表 1** の評価環境下で検証を行った。ここで、公衆回線を利用する認証では、入力端末に携帯電話を利用している。また、VoIP における入力端末には、PC を利用している。

表 1 評価環境

評価環境	室内	
伝送	VoIP	公衆回線
評価人数	5 名	65 名
学習用数字発話	50 発話	
評価用数字発話	65 発話	
収録周波数	8kHz	
収録ビット数	8bit	

また、われわれの方式では、音声に含まれるこれらの認証子を用いて総合的に認証する方式を採用しているが、本方式では、各々の認証子の総合的な認証性能への寄与が明確にできないため、次節では、各認証子を個別に評価している。なお、声紋認証、言語認証には音声認識エンジン Julius/Julian 3.3p4[3][4]を利用している。

5 評価結果

本節では、前節の条件化で、各々の認証子を評価した結果を示す。

図 2 から図 6 に、各認証方式における閾値を変動させた場合の他人受理率 FAR(False Acceptance Rate：他人を本人と誤って受理する確率)と本人拒否率 FRR(False Rejection Rate：本人を誤って拒否する確率)の推移を示す。ここで、左上部より右下部へ推移する曲線が FAR を表し、他者が FRR を表す。また、実線が VoIP 環境下での評価結果、破線が公衆回線での評価結果を示す。

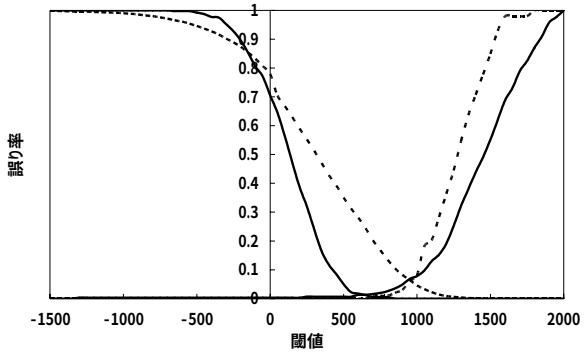

図 2 声紋認証評価結果の比較

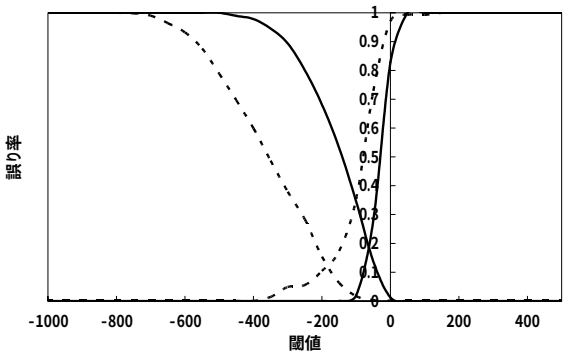

図 3 言語認証評価結果の比較

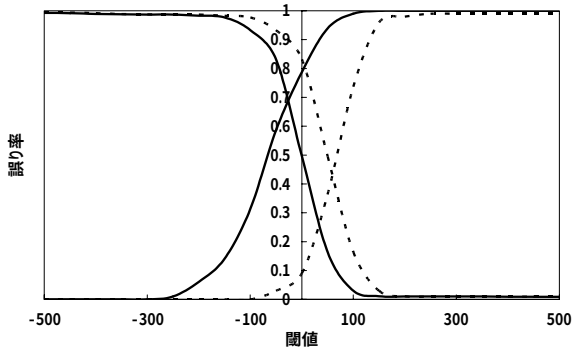

図 4 アクセント認証評価結果の比較

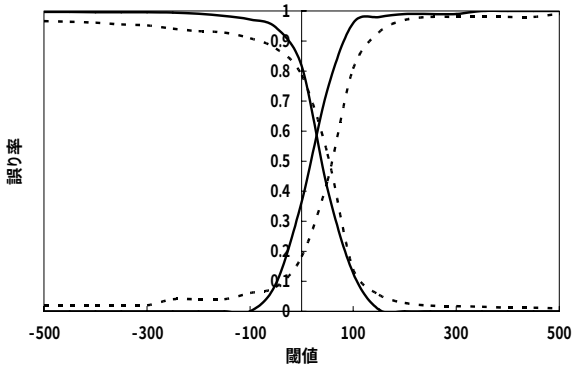

図 5 ピッチ認証評価結果の比較

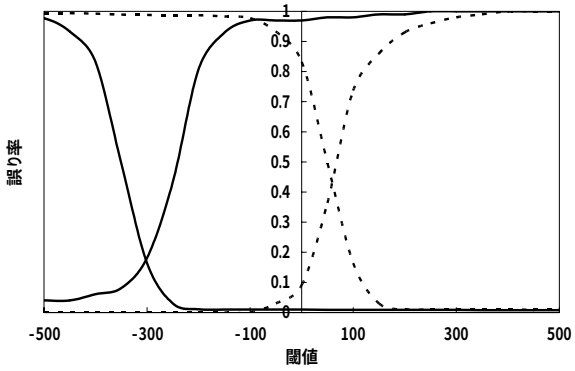

図 6 発話長認証評価結果の比較

次に、2 つの環境（VoIP、公衆回線）における提案方式第一段階における等誤り率（FAR=FRR となる閾値における認証精度）の比較を示す。

表 2 VoIP と公衆回線環境下での等誤り率比較

	等誤り率(%)		
	声紋	言語	声紋+言語
VoIP	1.01	18.7	0.19
公衆回線	6.47	10.8	0.70

6 考察と課題

一般的に、認証システムを評価する場合、FAR、FRR 曲線が重ならないことが理想とされており、これらの重なり度合いによりシステムの精度を判定する。

前節の結果を評価すると、図 2 の声紋認証において、公衆回線環境化と比べ、VoIP 環境下は、良好な結果が確認できる。これは、現在の公衆回線の音声符号化よりも、VoIP の音声符号化が声紋の特性を忠実に再現することによるものと思われる。しかし、図 3 の言語認証では、性能が低下することがわかる。これは、入力端末のマイク特性により、雑音が含まれることによるものと思われる。

また、図 3 から図 6 の認証子については、FAR, FRR の重なりが大きいため、単体での高い認証性能を実現することは難しいと思われる。

しかし、今回提案した手法、すなわち、認証子を二段階に分離し、認証する方式、を利用することで、前述の第一段階で本人を拒否する現象が発生した場合に、第二段階の認証を本人の救済を目的とし、FRR 曲線のみに着目し、閾値を設定するなどの手法をとることで、すべての認証子を有効に活用することができると考える。

たとえば、第一段階において、他人受理率が 0% になるように閾値を設定し、他人の受理を完全に排除した後、第二段階で、誤って拒否された本人を救済するという考え方である。

次に、表 2 の等誤り率を考察すると、VoIP システムの第一段階での認証精度（誤り率 0.19%）が公衆回線環境下（誤り率 0.70 %）よりも良好であることがわかる。

以上の考察より、VoIP システム上での音声認証でも、従来の一般的な手法である、発話内容と声紋認証を利用した音声認証方式が有効であることを確認した。

VoIP システムでは、声紋認証のみでも十分な精度が得られることを確認したが、この結果は、被験者の規模が小さかったためによるものと思われ、今後は、提案方式の有効性を明確にするためにも、評価の母集団規模を大きくする必要があると考える。母集団を大きくすることで、声紋認証の低下が予想されるが、今回のわれわれが提案する手法を組み合わせることで、精度の低下を回避できる可能性もある。

7 まとめ

今後の幅広い展開が期待できる VoIP システム上で音声を利用した個人識別が可能かを検証したが、その結果、VoIP システム上で、従来の声紋や発話内容を利用した認証方式が有効であることを確認した。また、われわれが提案する複数の音声に含まれる認証要素を組み合わせる方式も、有効に働く可能性があることを確認した。

謝辞

本研究で検証したシステムは、情報処理推進機構より公募事業「音声を媒体とする複数の認証子を利用した重点領域情報技術開発事業」として援助を受け、構築しています。

参考文献

[1] 千村 保文, 村田 利文: SIP 教科書, 株式会社 IDG ジャパン (2003).

[2] 岩野 公司, 関 高浩, 古井 貞熙: 雑音に頑健な音声認識のための韻律情報の利用, 情報処理学会 研究 報 告 , 2003-SLP-46-10, vol.2003, no.58, pp.55-60 (2003).

[3] 河原 達也, 住吉 貴志, 李 晃伸, 坂野 秀樹, 武田 一哉, 三村 正人, 伊藤 克亘, 伊藤 彰則、鹿野 清宏: 連続音声認識コンソーシアム 2002 年度版ソフトウェアの概要 (2003).

[4] 鹿野 清宏, 伊藤 克亘, 河原 達也, 武田 一哉, 山本 幹雄: IT Text 音声認識システム, 株式会社 オーム社 (2001).