話者認証を用いたX Window施錠システムxvlock ——開発とその評価

山下昌毅† 杉山雅英†

音声波に含まれる個人性情報を用いて発話者の認識を行う話者認識技術は個人認証の手段として使用可能である．本論文では話者認証を用いたコンピュータアクセスコントロールを行うソフトウェアxvlockを開発し，その評価実験結果について述べる．xvlockは，UNIXのX Window Systemにおけるパスワードを鍵にした施錠システムの上に構築されており，パスワード認証が正常終了した後に，話者認証による個人認証を行う．SunOS Release 4.1.3_U1（S-4/IX）に実装し，さらに評価のための話者認証実験を行った．8 bit μ law 標本化周波数 8 kHz の低品質な入力音声に対して，93.9%の高い認証率を得た．環境の違いを吸収するための環境変数の導入によってxvlockは優れた汎用性を持ち，他のUNIX系のコンピュータシステムへの移植が可能である．

Speaker Verification Applied to xvlock in X Window Lock System —Development and Its Evaluation

MASAKI YAMASHITA† and MASAHIDE SUGIYAMA†

The speaker can be recognized using the individual features included in each voice wave. It is called the speaker recognition, which can be applied as a means of an individual verification. This paper develops a software system named "xvlock" which can manage computer access by the speaker recognition technique, and also describes the outline of xvlock and the performance evaluation. The implementation and the experiments did only one standard platform, but xvlock can be applied to the other platforms because of less platform dependency. The For low quality input voice (8 bit μ law, sampling rate: 8 kHz) implemented xvlock achieved 93.9% verification rate.

1. まえがき

音声波に含まれる個人性情報を用いて発話者の認識を行う話者認識技術[1),2)]は個人認証の手段として使用可能である．現在，コンピュータを使用する際のユーザ認証としては，IDとパスワードをキーボードから入力する方法が一般的である．しかし入力者がIDの正当な持ち主でない場合でも，入力したパスワードが正しければ本人として認証されてしまう．話者認識をパスワードによるセキュリティ保全と組み合わせて使用することで，より高いセキュリティを確保でき，部外者が容易に侵入できないシステムを構築できる．

UNIXにおけるX Window Systemには一時的に端末画面を施錠（lock）しユーザのパスワードの入力により解錠（unlock）を行うソフトウェアがあり，ユーザが短時間離席する場合などに用いられている．本論文ではその施錠/解錠ソフトウェアに，話者照合による使用者認証判定を付加したxvlockの開発とその評価について述べる．

話者認識に用いる音声内容の種類によって以下の3つに分類される：

(1) テキスト従属（text-dependent）
(2) テキスト独立（text-independent）
(3) テキスト指定（text-prompted）

第1のテキスト従属とは発話内容（テキスト）があらかじめ決められている．第2のテキスト独立では，発話内容を特に限定しない．テキスト従属に比べて発話内容を覚えておく必要がないので使用者への負担は少ないが，発話内容が限定されないので認識性能は低い．これらの2つの方法を用いた話者認識セキュリティシステムは，話者の録音音声を用いることによってシステムを破られてしまう危険性がある．第3のテキスト指定による方法[3)]では，システムの指定の発話内容を入力する．話者認識と不特定話者音声認識技術とを組み合わせる方式であり，話者性・入力音声の認識結

† 会津大学 コンピュータ理工学部
School of Computer Science and Engineering, The University of Aizu

表 1 platform による音声入力・施錠システムの違い
Table 1 Difference by various platforms.

	SunOS		HP-UX	SGI-IRIX	FreeBSD (PC/AT)
	標準	DAT-Link	標準	標準	標準
X lock コマンド	xlock	xlock	vuelock	xlock	xlock
音声録音コマンド	record	narecord	recorder	recordaiff	—
audio device	8 bit	16 bit	16 bit	16 bit	
符号化方式	μ-law	線形	線形	aiff	
標本化周波数	8 kHz	44 kHz	16 kHz	44 kHz	
符号化変換	必要	不必要	不必要	必要	不必要
integer (16 bit)	——	互換	互換	互換	swap
floating (32 bit)	——	互換	互換	互換	no

——：これを基準とする
SUN：S10，HP：HP9000/715/80，FreeBSD：GATEWAY2000 G6-266

果（指定のテキストであるかどうか）およびシステムの発話要求から話者が返答するまでの応答時間を総合的に評価して受理/棄却を決定する．

本論文ではシステムの実現の容易さやユーザの使用時における使いやすさを考えて，テキスト独立の話者認識方式を用いることにする[10),11)]．録音音声を用いることによってシステムを破られてしまう危険性があるが，話者認証に先だってパスワードによる認証を行うので，テキスト独立方式でもセキュリティ向上の点で問題にはならないと考えられる．

本論文は OS として UNIX を対象としているが，以下の URL に示すように，パーソナルコンピュータ（Windows 95 など）上で動作する話者性に基づくセキュリティソフトウェアがすでにいくつか販売されている．

- VoicEntry I（T-NETIX 社）
  http://www.T-NETIX.com/
  テキスト従属，スクリーンセイバ
- VoiceGuardian（Keyware 社）
  http://www.keyware.com/Products
  /VoiceGuardian/
  テキスト従属，話者モデルを暗号化，LAN 対応

2. xvlock の要求条件

従来の話者認識応用システムにおいては音声入力系や動作環境が固定されており，実現プログラムの移植可能性などは陽には要求条件とはならない．しかしながら，話者認証を各種の WS（Work Station）に実現するためにはプログラムは移植性，汎用性が高くなければならない．また platform の音声入力系の多様性に柔軟に対応できるものでなければならない．記憶容量の削減や管理の容易性のために platform ごとに認証用の話者モデルを持つのではなく，可能な限り共有化を図る必要がある．

WS に標準でサポートされる画面施錠コマンドと音声入力コマンドを使用することにより，システム開発のための不必要な労力を削減しかつ動作モジュールのコンパクト化を図ることとする．

2.1 共有化とカスタマイズ

多くの UNIX 環境では様々な種類の複数のコンピュータをネットワークで接続し，複数のコンピュータをあたかも同一のコンピュータのように使用している．またネットワークファイル共有のための NFS を用いて複数のコンピュータから同一のログインディレクトリなどのファイルシステムをアクセスすることができる．使用するコンピュータが異なれば，録音コマンドなどのおかれる directory，その名称や動作仕様も異なる．**表 1** にいくつかの platform による音声入力・施錠システムの違いを示す．同一の SunOS であっても異なる audio device を用いて音声を録音可能である．一方，環境が異なっていても同一名のコマンドが使用できる場合もある．

UNIX shell の環境変数や起動オプションで動作を制御し，動作環境の違いに対処することとする．たとえば，環境ごとに異なる音声録音コマンドなどは，その違いをユーザに設定してもらう方式をとり，環境の違いに左右されることなく音声の録音を行うことができる．

同一の platform であっても，マイクロホンの違いにより入力特性が異なる．入力特性の違いについては 3.4 節で述べるように特徴ベクトルの上での適応処理または正規化処理で対処することにする．

WS に音声を取り込む際に 16 bit 線形，さらに LPC 分析パラメータを表すために 32 bit 浮動小数点の符号化方式を用いる．各種の WS におけるこれらの符号の互換性について表 1 に示す．SUN は S10，HP-UX は HP9000/715/80，SGI は IRIX であり，FreeBSD は GATEWAY2000 社の G6-266 にインストールし

表2 オプションと環境変数の対応関係
Table 2 Relationship between command options and environment variables.

option	環境変数	内容	標準設定（SunOS の場合）
-U	XVLOCK_DEFAULT_CODEBOOK	VQ 符号帳ファイル名	CB.M32.N14.SunOS
-N	XVLOCK_NOISE_CODEBOOK	雑音ベクトルファイル名	NOISE.N14.SunOS
-X	XVLOCK_XLOCK	X Window System 施錠コマンド名	/usr/local/bin/xlock
-S	XVLOCK_S_RECORD	学習時の録音コマンド名	/usr/demo/SOUND/record
-L	XVLOCK_L_RECORD	認証時の録音コマンド名	/usr/demo/SOUND/record
-F	XVLOCK_CODEING	音声符号化方式変換プログラム名	sox -U -b input.au \ -s -w input.raw
-C	XVLOCK_WAVECEP	LPC ケプストラム分析プログラム名	~/bin/WaveCep

た FreeBSD を用いている．この表から SUN/HP-UX/SGI において音声波形のみならず LPC ケプストラム係数ベクトルの互換性があることが分かる．また GATEWAY2000（FreeBSD）に対しては音声波形のみは 8 bit 単位でのバイト交換を行えば，互換であることが分かる．したがって，音声入力に用いるマイクの特性が同一であれば SUN/HP-UX/SGI で作成する 32 bit 浮動小数点データは共有化できることになる．

xvlock では個々の platform の環境に依存する部分に関しては，環境変数（Environment Variable），もしくは起動オプションで設定を行うようにする．使用している環境変数と起動オプションの対応関係を**表 2** に示す．起動オプションで設定された値は環境変数による設定より優先するので，一時的な設定値の変更などに使用することができる．

2.2 仕　様

以上をふまえて，以下のような仕様とする．

(1) 汎用性を与え，他の UNIX X Window System の platform に移植可能とする．
(2) 環境変数，起動オプションの導入により platform 依存部分を吸収する．
(3) 施錠を行う前にマイクロホンの接続確認を行う．
(4) 既存の画面施錠，音声入力コマンドを利用する．
(5) xvlock ではパスワード認証処理を行わない．
(6) 音声符号化方式を線形 16 bit，音声特徴抽出を LPC ケプストラム方式[4] とする．
(7) 話者認証方式を VQ 符号帳によるテキスト独立話者認識方式[5),6)] とする．
(8) 話者認証判定の閾値をマイクロホン接続確認のための入力音声を用いて設定する．

2.3 動作の流れ

xvlock の動作の流れを**図 1** に示す．全体は大きく分けて lock 部，話者モデルの作成部，unlock 部から構成されている．図の左部分が施錠処理（lock）であり，右部分が開錠処理（unlock）である．話者学習（登録）部で作成された話者モデルは，unlock 部の話者認証に使用される．この話者モデルは 3.2 節で述べるベクトル量子化符号帳（VQ codebook）と呼ばれる話者の音声特徴ベクトルの集合である．テキスト独立話者照合方法を実現するためには，VQ 識別器による方法[6]，混合連続分布 HMM による方法[3] が知られている．前者を適用する場合には話者モデルとして VQ 符号帳（ベクトルの集合）となる．後者を適用する場合には話者モデルとして複数の連続分布（ガウス分布）を表現するための，平均ベクトル，共分散行列，分布の混合重み係数などを学習させまた照合のために保持する必要がある．したがって，前者に比べて，学習に要する音声の量や話者モデルに要する情報量が増大する危険性がある．したがって，本論文では前者を適用することとする．

2.4 実 現 例

現在，GNU C Compiler（2.6.3 以降），X Window System（X11R5 以降）の環境での動作を確認している．SunOS 上で開発を行ったが，source file の書き換えを行わずに HP-UX でも動作した．

以上の条件を満たすものとして，以下の WS での動作を確認している

- SunOS Release 4.1.3_U1 + X11R5（S-4/IX）
  gcc version 2.7.2.3.f.1
- HP-UX A.09.05 + X11R6（Model712/80）
  gcc version 2.6.3

また xvlock では以下のソフトウェアを使用する．

- `sox`：音声符号化方式変換プログラム
- `xlock`, `vuelock`：パスワードによる X Window System 施錠プログラム

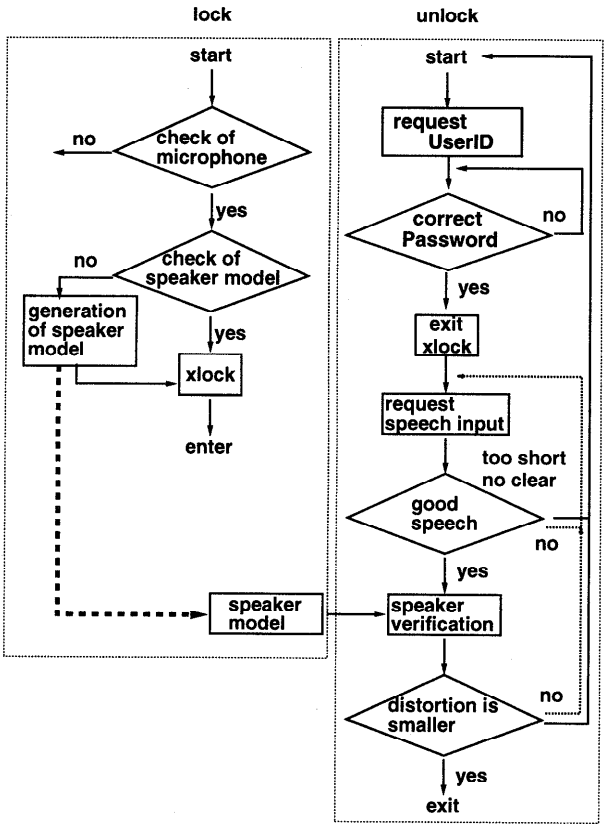


図 1 xvlock の処理の流れ
Fig. 1 Flow chart of xvlock.

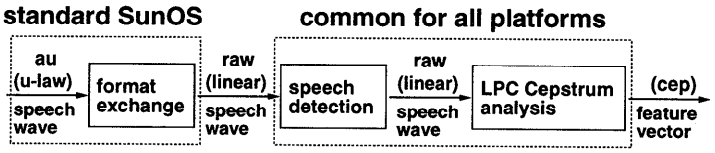


図 2 音声特徴抽出（SunOS 版）
Fig. 2 Voice feature extraction (for SunOS).

3. 話者認識の方法

3.1 音声特徴抽出

話者登録と照合判定で用いられる音声特徴抽出部は，「音声の入力と変換」「音声区間切り出し」「特徴抽出」の 3 つで構成されている．音声特徴抽出法として LPC ケプストラム分析法を用いている．

SunOS の場合の処理の流れを図 2 に示す．

3.1.1 音 声 入 力

xvlock は環境変数で指定された音声録音用コマンドを用いて音声入力する．コンピュータに付属のマイクロホンなどを使用する．マイクロホンの形態としてスタンドマイクやヘッドセットマイクがあるが，マイクロホンと発話者の口唇との距離や角度が話者登録時と認証時で同一に保たれる必要があるので，ヘッドセットマイクが望ましい．

3.1.2 音声符号化方式の変換

表 1 に示すように音声録音コマンド名と音声符号化方式は，platform によって異なる．符号化方式を，線形 16 bit monoral 形式の raw ファイル（ヘッダーなし）に変換する．

3.1.3 音声区間切り出し

入力された音声には無音区間が含まれているため，以下に述べるアルゴリズムを用いて音声区間の切り出しを行う．

音声波形 (x_t) の F 個の点を分析単位（1 フレーム）

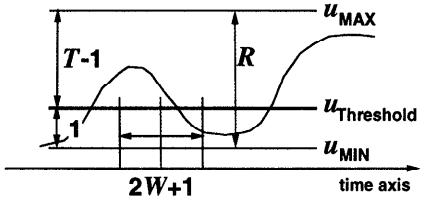


図 3 音声区間検出の概念
Fig. 3 Procedure of voice interval detection.

とし，n 番目のフレームの音声パワー (u_n) を以下の式で計算する．

$$u_n = 10\log_{10}\left(\frac{1}{F}\sum_{t=1}^{F} x_{t+nF}^2\right)$$

ここで $F = 256$ であり，標本化周波数 8 kHz のときは 32 msec に対応する．音声パワー (u_n) の最大値 u_{MAX} と最小値 u_{MIN} の値から SNR (R) を求める．

$$R = u_{\mathrm{MAX}} - u_{\mathrm{MIN}}.$$

図 3 に示すように指定のフレームの音声/雑音の判定を以下のように行う．そのフレームを中心として前後 W フレーム分（合計 $2W+1$ フレーム分）の音声パワーの平均 $(\hat{u}_n)$

$$\hat{u}_n = \frac{1}{2W+1}\sum_{t=n-W}^{n+W} u_t$$

を計算し，以下の不等式を満たすとき，音声区間とする．

$$\hat{u}_n > u_{\mathrm{Threshold}} = u_{\mathrm{MIN}} + \frac{R}{T} = \frac{u_{\mathrm{MAX}} + (T-1)u_{\mathrm{MIN}}}{T} \quad (1)$$

本論文では，経験的に $W = 3$, $T = 3$ としている．

3.1.4 音声特徴抽出

音声区間を切り出した後，LPC ケプストラム分析を用いてフレームごとに LPC ケプストラム係数ベクトルに変換を行う．ここで，LPC 分析次数，LPC ケプストラム打切り次数は 14 としている．LPC ケプストラム係数は話者認識だけでなく音声認識などを含む多くの音声処理応用システムにおいて使用されている標準的な特徴量である．

3.2 話者登録（話者モデルの作成）

話者認識のための話者モデルである VQ 符号帳（codebook）$\boldsymbol{V} = \{v_m\}(m = 1,\ldots,M)$ を，入力音声から LBG アルゴリズム [7] を用いて作成する．入力される音声のテキストがあらかじめ定められていないので，任意の音素が入力される可能性がある．したがって，音声には照合時に使用される音素が偏りなく含まれていることが望ましい．実現システムでは日本語の 50 音表を順に読み上げるようにした．音声入力終了後，録音された音声の特徴抽出を行い，VQ codebook を自動的に作成する．

3.3 認証判定

作成された $\boldsymbol{V}$ と入力音声を用いて照合を行う．認証用の入力音声の特徴ベクトルの列，$\boldsymbol{x}_l\ (l = 1,2,\ldots,L)$ に対して以下の式で VQ 歪み（LPC ケプストラム距離）を計算する．

$$D = \frac{1}{L}\sum_{l=1}^{L}\min_{1\le m\le M} d(\boldsymbol{x}_l, \boldsymbol{v}_m) \quad (2)$$

ここで L は入力音声のベクトルの数（入力音声長に対応する）であり，$d(\boldsymbol{x},\boldsymbol{v})$ は以下で定義される LPC ケプストラム距離である．

$$d(\boldsymbol{x},\boldsymbol{v}) = \sum_{n=1}^{N}(c_n^x - c_n^v)^2 \quad (3)$$

ここで，$\boldsymbol{x} = (c_n^x)$, $\boldsymbol{v} = (c_n^v)$ は LPC ケプストラムベクトルである．歪みの値 D が閾値 D_T よりも低い値の場合，VQ codebook の話者と同一人物であると受理し，閾値よりも高い値の場合，本人ではないとして棄却する．D_T の設定法に関しては 4.3.1 項の式 (8) で述べる．

3.4 入力系の音響特性の違いの正規化処理

入力音声を標本化して音声特徴量に変換するまでに，入力系の違いにより音声は変形を受ける．マイクロホンの音響特性の違いや伝達特性の違いは線形フィルタで近似できる．したがって，その違いの周波数特性を $h(\lambda)$ とすると入力音声の短時間スペクトル（ピリオドグラム），$P(\lambda)$ は線形フィルタリングされ，$P^*(\lambda) = P(\lambda)h(\lambda)$ となる．したがって，それに対応する LPC スペクトル $f^*(\lambda)$ も近似的に線形フィルタリングされる．

$$f^*(\lambda) = f(\lambda)h(\lambda)$$

LPC ケプストラム係数は対数スペクトル $\log f(\lambda)$ のフーリエ係数として定義されるので，以下の関係式を得る．

$$\begin{aligned} c_n^* &= \int_{-\pi}^{\pi}\log f^*(\lambda)e^{-jn\lambda}\frac{d\lambda}{2\pi} \\ &= \int_{-\pi}^{\pi}(\log f(\lambda) + \log h(\lambda))e^{-jn\lambda}\frac{d\lambda}{2\pi} \\ &= \int_{-\pi}^{\pi}\log f(\lambda)e^{-jn\lambda}\frac{d\lambda}{2\pi} \end{aligned}$$

$$+\int_{-\pi}^{\pi}\log h(\lambda)e^{-jn\lambda}\frac{d\lambda}{2\pi}=c_n+h_n$$

ここで h_n は $h(\lambda)$ のフーリエ係数である．

$$h_n=\int_{-\pi}^{\pi}\log h(\lambda)e^{-jn\lambda}\frac{d\lambda}{2\pi} \quad (4)$$

したがって，LPC ケプストラムベクトル $\boldsymbol{c}$ は以下のように $\boldsymbol{h}=(h_n)$ を用いて平行移動で得られることになる．

$$\boldsymbol{c}^*=\boldsymbol{c}+\boldsymbol{h} \quad (5)$$

学習用の音声の集合を $\boldsymbol{X}$, $\boldsymbol{X}^*$ とするとき，式 (5) より以下の式で与えられる．

$$\boldsymbol{X}^*=\boldsymbol{X}+(\boldsymbol{h})=\{\boldsymbol{x}+\boldsymbol{h}\ |\ \forall\boldsymbol{x}\in\boldsymbol{X}\} \quad (6)$$

たとえば，SUN において線形 16 bit の学習音声から作成した VQ 符号帳 $\boldsymbol{V}$ と，他の WS (HP, SGI) での学習音声から作成した VQ 符号帳 $\boldsymbol{V}^*$ との関係を求める．表 1 に述べたように，符号帳の浮動小数点形式が互換であるので，付録 A.1 で述べるように VQ 符号帳作成アルゴリズムが一定の条件を満たす場合には，入力するためのマイクロホンなどの音響特性の違い $\boldsymbol{h}$ を用いて以下のようにかける．

$$\boldsymbol{V}^*=\boldsymbol{V}+(\boldsymbol{h})=\{\boldsymbol{v}+\boldsymbol{h}\ |\ \forall\boldsymbol{v}\in\boldsymbol{V}\} \quad (7)$$

ここでベクトル $\boldsymbol{h}$ は式 (4) に与えられる 2 つの入力系の音響特性の違いに対応する補正のためのケプストラム係数ベクトルである．WS ごとに VQ 符号帳を持つのではなく，標準の VQ 符号帳 $\boldsymbol{V}_{\mathrm{SUN}}$ とマイクロホンなどの入力音響特性の違いを補正する特徴ベクトル $\boldsymbol{h}$ を持てばよいことになる．$\boldsymbol{h}$ の推定法については付録 A.2 で述べる．

4. xvlock の使用手順

4.1 installation

xvlock のパッケージの置かれている directory から ftp などで入手し☆，以下の手順で install を行う．使用方法などの詳細情報については README を参照する．

```
% gzip -d XVlock.tar.gz
% tar xvf XVlock.tar
% edit Makefile
% make install
```

Makefile では作成コマンドを置く directory を指定する．default では `~/bin` となる．`make install` で話者モデル，およびモデル作成のための入力音声をおくための directory (`~/.xvlock/{CB,Voice}`) を作成する．

`~/.xvlock/CB` 話者モデル

`~/.xvlock/Voice` 入力音声

話者モデルの置かれている directory を command search path に登録する．次に，必要に応じて環境変数（表 2 を参照）を `~/.cshrc` などの中で設定する．複数の WS を使用し，同一の login directory を用いる場合は，WS に対応して設定を行う．詳しくは xvlock のパッケージ中の `.cshrc.SunOS` などを参考にする．

4.2 話者登録（話者モデルの作成）

認証に先だって認証用の話者モデルを以下のように作成する．

```
% xvlock -M
```

呼び出される `MakeCB` は `c shell` で記述された簡単な shell script である．画面の指示に従い，画面に表示されるテキストを発声し録音を行う．1 単語 3 秒間で 15 単語の録音を行う．録音は 50 秒ほどで終了する．SunOS 4.1.3_U1 (S-4/IX) を用いた話者モデル作成処理に要する時間は**表 3** に示すように作成する話者モデルの codebook のベクトル数に線形に比例する．VQ 符号帳の大きさが 32 のとき，99 秒（音声切り出しなし）および 46 秒（音声切り出しあり）となる．音声切り出し処理を行うことで，学習用のベクトル数が減少するので codebook の作成時間を短縮できることになる．作成された話者モデルは directory (`~/.xvlock/CB`) に置かれる．

4.3 施錠および開錠のための認証処理

```
% xvlock
```

と入力することにより施錠を行うことになる．通常の話者認識システムではマイクロホンの接続は管理者の責任で事前に行われていると仮定して動作してよいが，一般的な使用者に対してはマイクロホンが必ずしも接続されているとは限らない．

4.3.1 マイクロホン接続確認

マイクロホン接続確認のため初期録音が行われる．ここで使用者が実際に開錠（unlock）時と同様の発声を要求する．音声が正常に入力されなかった場合，すなわち，取り込まれた音声パワーが 0 であるときには正常にマイクロホン接続されていないことを警告し終了する．起動時に毎回音声入力するのが面倒である場

表 3 話者モデル（codebook）作成時間（秒）
Table 3 Processing time for codebook generation.

VQ 符号帳の大きさ	音声切り出し	
	なし	あり
16	45	20
32	99	46
64	173	92
128	297	128
フレーム数（フレーム）	2820	1257

☆ 現在，一般公開の準備中である．

合には，起動時にマイクロホン接続確認の処理を省くことも可能である．

```
% xvlock -noMicCheck
```

3.3 節で述べたように，unlock の際に使用する閾値を入力音声 $\boldsymbol{x}_l$（L_0 は入力音声の長さに対応する）に対する歪み，D_0，

$$D_0 = \frac{1}{L_0}\sum_{l=1}^{L_0} \min_{1\le m\le M} d(\boldsymbol{x}_l, \boldsymbol{v}_m)$$

をもとにして式 (8) で設定する．

$$D_T = \alpha D_0 \tag{8}$$

ここで経験的に $\alpha = 1.05$ としている．

xvlock の起動時のオプション -d を用いて明示的に閾値を設定することも可能である．

```
% xvlock -d 0.5
```

この数値は 3.3 節の入力音声の話者モデルに対する LPC ケプストラム距離による歪みに対する閾値 D_T であり，大きい値に設定すると開錠しやすくなり，他人の入力音声をも本人として受理する危険性がある．一方，小さい値に設定すると本人であるのに棄却される可能性がある．風邪をひいたり喉を痛めたりして，話者モデル作成時と声が変化している場合や入力背景音が変化している場合などは，lock したものの unlock できないという事態が発生する可能性がある．施錠時の初期録音を行うことで，このような事態を回避できる．

D_T の動的な設定方法に関しては cohort 法が検討されているが，実現するには計算量が増加する．

4.3.2 lock/unlock

マイクロホン接続確認処理および閾値設定が正常に終了した後，環境変数で指定された X Window System の画面施錠コマンド（xlock）を呼び，lock した後に xvlock の待機画面になる．何らかのキー入力で xlock のパスワードによる認証を開始し，パスワードによる認証が完了しない場合は，音声入力の画面は現れない．パスワードによる認証が正常終了した場合，音声入力画面に移行する．ここで画面の指示に従い，マイクロホンから 5 秒程度の音声を入力する．ただし SNR が 30 dB 以上の音声区間が 2 秒以下である場合，入力音声による認証を行わず xlock を再度実行する．SNR が低い場合はマイクのスイッチが入っていない可能性がある．短時間の入力音声に対しては認証精度の低下の危険性がある．良好な音声が入力された場合，音声による認証処理を行い，受理されれば xvlock を終了する．棄却された場合には，もう一度 xlock プログラムを実行する．

入力音響特性が異なることによる認証性能の劣化を防ぐためには，環境ごとに話者モデルを作成し，起動 option で使用するモデルを指定することも可能である．

```
% xvlock -U ~/.xvlock/CB/CB.SunOS.DAT-Link
```

5. xvlock の評価

5.1 話者認証実験

話者認証実験の実験条件を**表 4** に示す．評価実験の条件については SunOS での使用を想定して設定した．実験には 14 人の話者を用い，学習には 50 音表を各々の段ごとに 1 単語として 5 音節をまとめて発話させた．

表 4 話者認証の実験条件
Table 4 Experimental conditions.

話者数	男性 13 人 女性 1 人 ともに 20 歳前後
学習用音声	「あ」行,「か」行,「さ」行 「た」行,「な」行,「は」行 「ま」行,「や」行,「ら」行 「わをん」 「が」行,「ざ」行,「だ」行 「ば」行,「ぱ」行
認証用音声 各話者 1 回発声	音声 1「本日は晴天なり」 音声 2「青い屋根の家」
符号化方式	8 bit μ-law
標本化周波数	8 kHz
フレーム長	32 ms（256 点）
フレーム周期	16 ms（128 点）
窓関数	ハミング窓
高域強調	$(1-0.97z^{-1})$
特徴抽出	LPC ケプストラム分析（次数：14）
VQ 符号帳作成法	LBG algorithm
VQ 歪み尺度	LPC ケプストラム距離 ユークリッド距離
WS	SunOS 4.1.3_U1 (S-4/IX)
マイクロホン	HP 社 Headset Microphone

表 5 話者認証実験結果
Table 5 Results of speaker verification experiments.

音声切り出しあり

	認識率（%）			
VQ 符号帳サイズ	16	32	64	128
音声 1	88.7	94.8	91.2	91.8
音声 2	90.4	92.9	92.6	93.1
平均	89.6	93.9	91.9	92.5

音声切り出しなし

	認識率（%）			
VQ 符号帳サイズ	16	32	64	128
音声 1	85.2	84.9	88.2	87.4
音声 2	76.9	84.1	84.6	89.3
平均	81.1	84.5	86.4	88.4

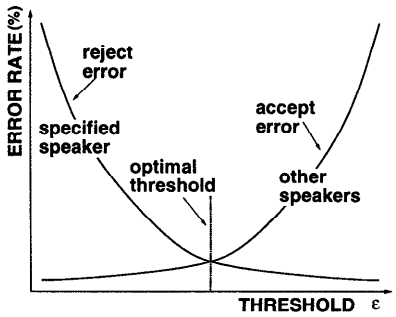


図4 話者認識率の定義
Fig. 4 Definition of verification error rate.

認証には2つの異なるテキストの音声を用いた．入力音声の符号化方式は8bit μ-law 8kHzであり，線形16bitに変換した後，14次のLPC分析，LPCケプストラム分析を用いて特徴抽出を行った．認識結果を**表5**に示す．ここで認証率は**図4**に示すように「本人に対する拒否誤り率」と「詐称者に対する受理誤り率」の平均が最小になる最適な閾値を設定し求めた．VQ符号帳の大きさに関しては音声切り出しありのとき，32に対して最良の認証率93.9%であり，符号帳をそれより大きくしても改善されない．一方，16に減少しても大きな劣化はないことが分かる．音声の種類によって若干性能は異なるが，その差は2%以内である．簡単な音声区間切り出し手法ではあるが，5〜10%程度の認証率の向上を達成している．音声が低品質であることを考慮すると，ここで得られた認証率は妥当な性能であると考えられるが，指紋を用いた場合と比較して認証性能が十分とはいえない．したがって，パスワードによる認証と組み合わせることが必要であると考えられる．実際に用いる場合にはマイクロホン接続確認の音声を用いて，式(8)で決定される閾値 D_T の係数 α をより大きく設定し本人の拒否誤りを減少させ，使い勝手を向上させることが可能である．

8bit μ-law 8kHzという低品質の音声にもかかわらず高い認証率を得た．16bit線形などのより高品質の音声に対して，より高い認証性能の実現が期待できる．ただし実験で用いた学習用の音声と認識用の音声は同日に連続して録音したので，経時変化や入力環境の違い（マイクロホンと話者との距離の変化，話者の体調の変化）などの影響に関しては今後の検討課題である．

5.2 認証処理速度および必要記憶容量

認証のための距離計算の処理量 S は以下の式で与えられる．ここで，M は話者モデルのベクトルの数，L は入力音声の長さ，N はLPCケプストラムベクトルの次数である．

$$S \propto M \times L \times N$$

一般に L が大きくなるほど認証性能は向上し安定するが，使用者にとっては短時間入力での動作が望ましい．M は5.1節で述べたように32程度あれば十分であり，平均処理時間は1秒以下であり，十分高速に動作することが分かる．一方，話者モデルを表現するための必要記憶容量は以下で与えられる．

$$M \times N \times 4 \quad \text{(byte)}$$

$M = 32$，$N = 14$ の場合には1792byteとなる．

異なる入力音響特性を指定するベクトル $\boldsymbol{h}$ を用いれば $N \times 4$ byte増加するだけである．3.4節で述べた話者モデル共有化法の入力環境の変化への適応の有効性については今後の検討課題である．

6. むすび

話者認証を用いてコンピュータへのアクセスコントロールを行うソフトウェアxvlockを開発しその評価について報告した．xvlockの評価実験において8bit μ lawの低品質な入力音声に対して93.9%という高い認証性能を実現した．実験における学習用音声と評価用音声を，同一の日に連続して収録しているので，話者の経時変化による認証性能の劣化が予想される．またマイクロホンの接続確認の必要性，ユーザに対する使い勝手の評価，雑音環境下における性能評価および音響特性の正規化法の有効性については今後の検討課題である．また本論文ではテキスト独立話者認証の手法を用いたが，テキスト従属やテキスト指定の手法を用いた実現法についても検討する．また一般公開し多数のユーザに使用してもらうことにより，セキュリティやユーザインタフェースに関しての改善を検討する．

謝辞 xvlockの評価実験のために協力をいただいた学内の方々および論文を執筆するにあたり有益な助言をいただいた研究室の皆様に深謝いたします．

参考文献

1) O'Shaughnessy, D.: Speaker Recognition, *IEEE ASSP Magazine*, pp.4–17 (1986).
2) Mammone, R.J., Zhang, X. and Ramachandran, R.P.: Robust Speaker Recognition, *IEEE Signal Processing*, Vol.13, No.5, pp.58–71 (1996).
3) 松井，古井：テキスト指定形話者認識，電子情報通信学会論文誌，J79-D-II, No.5, pp.647–656 (1996).
4) Atal, B.S.: Effectiveness of Linear Prediction Characteristic of the Speech Wave for Auto-

matic Speaker Identification and Verification, *JASA*, No.55, pp.1304–1312 (1974).
5) 杉山，鹿野，相川：母音標準パタンの教師なし学習法，音響学会講演論文集，1-1-7, pp.13–14 (Oct. 1983).
6) Soong, F.K., Rosenberg, A.E., Rabiner, L.R. and Juang, B.H.: A Vector Quantization Approach to Speaker Recognition, *Proc. ICASSP85*, Vol.11, No.4, pp.387–390 (1985).
7) Linde, Y., Buzo, S. and Gray, R.M.: An Algorithm for Vector Quantizer Design, *IEEE Trans. COM-28*, pp.84–95 (Jan. 1980).
8) Shikano, K.: Spoken Word Recognition Based Upon Vector Quantization of Input Speech, Technical Report of ASJ Speech Committee, SP82-60, pp.473–480 (Nov. 1982) (in Japanese).
9) Katagiri, S., Lee, C.H. and Juang, B.H.: A Generalized Probabilistic Descent Method, *Proc. Acoust. Soc. of Japan*, 2-p-6, pp.141–142 (Sep. 1990).
10) 山下，杉山：話者認証を用いたX Window施錠システムxvlock開発とその評価，Vol.98, No.22, 98-HI-77-8, pp.43–48 (1998).
11) Yamashita, M. and Sugiyama, M.: Speaker Verification Applied to Display Lock System –Development and Its Evaluation, *AVIOS98*, pp.41–45 (Sep. 1998).

付　　録

A.1 VQ符号帳作成と空間の変換 (ϕ, f) の可換性について

N 次元ベクトル空間 R^N の部分集合 $\boldsymbol{X}$, $\boldsymbol{V}$ を学習用ベクトルの集合，VQ符号帳とし，$\boldsymbol{X}$ からLBGアルゴリズムを用いて作成するVQ符号帳を $\boldsymbol{V}=\phi(\boldsymbol{X})$ と表すことにする．ベクトル空間上の変換を $f: R^N \to R^N$ とし，$f(\boldsymbol{X})=\{f(x) \mid x \in \boldsymbol{X}\}$ と表すこととする．

> f がアファイン変換 $f(\boldsymbol{x})=\boldsymbol{F}\boldsymbol{x}+\boldsymbol{b}$ で，行列 F がユニタリであるとき，f,ϕ は可換である．
>
> $$\phi(f(\boldsymbol{X}))=f(\phi(\boldsymbol{X}))$$
>
> ただし，LBGアルゴリズムにおいて，ベクトル間の距離はユークリッド距離とし，集合の2分割代表点の計算には内分法[8]を用いるものとする．

VQ符号帳の作成アルゴリズムであるLBGアルゴリズムはLloydアルゴリズムと2分割処理から成り立っており，Lloydアルゴリズムは以下の2つの手順（区分化および中心計算）から成り立っている．

(1) Lloydアルゴリズム

(a) 区分化：$\boldsymbol{V}$ による集合 $\boldsymbol{X}$ の区分化

$$\wp_m=\{\boldsymbol{x} \in \boldsymbol{X} \mid d(\boldsymbol{x},\boldsymbol{v}_m)<d(\boldsymbol{x},\boldsymbol{v}_n), \quad \forall n \neq m\}$$

(b) 中心計算：集合 $\boldsymbol{A}=\{\boldsymbol{a}_i\}$ の中心 (centroid)

$$c(\boldsymbol{A})=\left\{\boldsymbol{x} \in R^N \mid \forall \boldsymbol{y}(\neq \boldsymbol{x}) \in R^N, \sum_i d(\boldsymbol{a}_i,\boldsymbol{x})<\sum_i d(\boldsymbol{a}_i,\boldsymbol{y})\right\}$$

(2) 2分割処理

区分化では，距離による最近傍へのラベリング，中心計算では，内分点計算，2分割処理では距離による最近点，最遠点の探索と内分点計算から成り立っている．したがって，変換により距離が不変であること，変換と内分点計算が可換であることに帰着される．

$\boldsymbol{V}$ による集合 $\boldsymbol{X}$ の区分化には距離が用いられており，補題1に示すように $\boldsymbol{F}$ がunitaryであれば距離は不変である．したがって，補題3から区分化に関する可換性が成り立つ．また補題2より中心計算も可換性が成り立つ．一方，領域の2分割処理に対しても領域の中心 v_0 から最遠点 v_1 を算出し，v_1 からの最遠点 v_2 を算出し，v_0 と v_1, v_2 との内分ベクトルを領域の新たな代表点とする．このとき，距離は不変であるので最遠点の算出と変換は可換であり，領域の新たな代表点算出と変換は可換となる（証明終了）．

補題1：距離の不変性

$f(\boldsymbol{x})=\boldsymbol{F}\boldsymbol{x}+\boldsymbol{b}$ で $\boldsymbol{F}$ はunitary行列であり，$d(\boldsymbol{x},\boldsymbol{y})$ がユークリッド距離であるとき，以下の式が成り立つ．

$$d(f(\boldsymbol{x}),f(\boldsymbol{y}))=d(\boldsymbol{x},\boldsymbol{y}),$$
$$(||f(\boldsymbol{x})-f(\boldsymbol{y})||^2=||\boldsymbol{x}-\boldsymbol{y}||^2)$$

証明：$f(\boldsymbol{x})-f(\boldsymbol{y})=\boldsymbol{F}(\boldsymbol{x}-\boldsymbol{y})$ であるので，

$$||f(\boldsymbol{x})-f(\boldsymbol{y})||^2=(\boldsymbol{x}-\boldsymbol{y})^t\boldsymbol{F}^t\boldsymbol{F}(\boldsymbol{x}-\boldsymbol{y})$$

ここで，$\boldsymbol{F}$ がunitaryであるので $\boldsymbol{F}^t\boldsymbol{F}=\boldsymbol{I}$ を代入すれば以下の式を得る．

$$=(\boldsymbol{x}-\boldsymbol{y})^t(\boldsymbol{x}-\boldsymbol{y})=||\boldsymbol{x}-\boldsymbol{y}||^2$$

補題2：c, f の可換性

アファイン変換 $f(\boldsymbol{x})=\boldsymbol{F}\boldsymbol{x}+\boldsymbol{b}$ に対して $d(\boldsymbol{x},\boldsymbol{y})$ がユークリッド距離であるとき，以下の式が成り立つ．

$$c(f(\boldsymbol{A}))=f(c(\boldsymbol{A}))$$

証明：$d(\boldsymbol{x},\boldsymbol{y})$ がユークリッド距離であるとき，c は以下のように計算される．

$$c(\boldsymbol{A}) = \frac{1}{|A|}\sum_{\boldsymbol{a}\in\boldsymbol{A}} \boldsymbol{a}$$

したがって，左辺は以下のように変形される．

$$\begin{aligned} c(f(\boldsymbol{A})) &= \frac{1}{|f(\boldsymbol{A})|}\sum_{f(\boldsymbol{a})\in f(\boldsymbol{A})} f(\boldsymbol{a}) \\ &= \frac{1}{|\boldsymbol{A}|}\sum_{\boldsymbol{a}\in\boldsymbol{A}} (\boldsymbol{F}\boldsymbol{a}+\boldsymbol{b}) \\ &= \frac{1}{|\boldsymbol{A}|}\sum_{\boldsymbol{a}\in\boldsymbol{A}} \boldsymbol{F}\boldsymbol{a}+\boldsymbol{b} \\ &= \boldsymbol{F}\left(\frac{1}{|\boldsymbol{A}|}\sum_{\boldsymbol{a}\in\boldsymbol{A}} \boldsymbol{a}\right)+\boldsymbol{b} \\ &= f(c(\boldsymbol{A})) \end{aligned}$$

単調増加関数 ψ を用いて以下の関係式が成り立つとき，f が距離 d に関して単調増加性を満たすと呼ぶ．

$$d(f(\boldsymbol{x}), f(\boldsymbol{y})) = \psi(d(\boldsymbol{x}, \boldsymbol{y}))$$

補題 3：区分化の可換性

$\wp(\boldsymbol{X} : \boldsymbol{v_m}) = \{\boldsymbol{x} \in \boldsymbol{X} \mid d(\boldsymbol{x}, \boldsymbol{v_m}) < d(\boldsymbol{x}, \boldsymbol{v_n}), \forall n \neq m\}$ と表す．f が単調増加性を満たすとき，以下の式が成り立つ．

$$f(\wp(\boldsymbol{X} : \boldsymbol{v}_m)) = \wp(f(\boldsymbol{X}) : f(\boldsymbol{v}_m))$$

証明：$\forall f(\boldsymbol{x}) \in f(\wp(\boldsymbol{X} : \boldsymbol{v}_m)) \Leftrightarrow \boldsymbol{x} \in \wp(\boldsymbol{X} : \boldsymbol{v_m})$

$\Leftrightarrow \forall n \neq m, d(\boldsymbol{x}, \boldsymbol{v}_m) < d(\boldsymbol{x}, \boldsymbol{v}_n)$

$\Leftrightarrow$ f が単調増加写像であるので，単調増加関数 ψ があり $\forall n \neq m, d(f(\boldsymbol{x}), f(\boldsymbol{v}_m)) = \psi(d(\boldsymbol{x}, \boldsymbol{v}_m)) < \psi(d(\boldsymbol{x}, \boldsymbol{v}_n)) = d(f(\boldsymbol{x}), f(\boldsymbol{v}_n))$

$\Leftrightarrow f(\boldsymbol{x}) \in \wp(f(\boldsymbol{X}) : f(\boldsymbol{v}_m))$

A.2 音響特性の差異の補正ベクトル h_X の推定法

3.4 節に述べたマイクロホンなどの入力音響特性の違いを補正するベクトル $\boldsymbol{h}_X$ を用いれば音響特性の違いに対応する VQ 符号帳を推定することが可能である．そこで，音響特性の差異の補正ベクトル $\boldsymbol{h}_X$ の推定について述べる．$\boldsymbol{h}_X$ は 2 つの入力系の音響特性の違いに対応する補正ベクトルであるので，音響的に同一の単語などの音声を 2 つの入力系を通して得られる LPC ケプストラム係数ベクトルの差の平均として求める．

$$\boldsymbol{h}_X = \frac{1}{L}\sum_{\boldsymbol{c}} \left(\boldsymbol{c}^X - \boldsymbol{c}^{SUN}\right)$$

ここで L は対応付けされるベクトルの個数である．同一の単語を 2 つの系に入力することは実際的には困難であるので，指定された単語を WS:X において入力し，あらかじめ蓄えられている同一の単語音声との間の動的計画法を用いてベクトル間の対応を求め，それにより上の式から $\boldsymbol{h}_X$ を算出することになる．話者照合率を評価基準として識別学習（誤り訂正学習）[9] を行う方法も可能である．

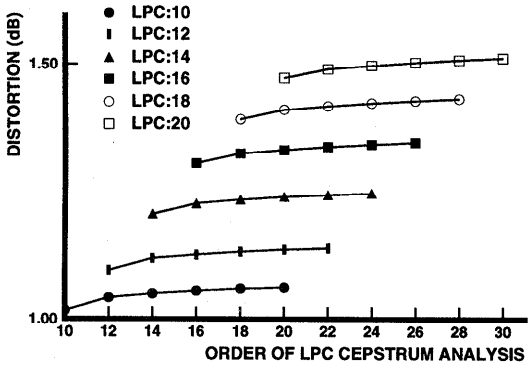


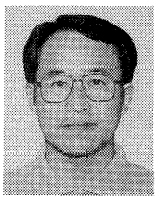
図 5 8 bit μ law と 16 bit 線形音声の品質の違い（パワーなし）
Fig. 5 Speech quality difference between 8 bit μ law and 16 bit linear.

A.3 符号化方式の変換によるスペクトル歪みについて

SUN は 8 bit μ law 音声符号化方式を用いており，HP-UX などでは 16 bit 線形 を用いている．そこで，8 bit μ law と 16 bit 線形との間の音声品質の違いを，種々の LPC 分析次数，LPC ケプストラム分析次数に対する LPC ケプストラム距離（dB 換算）を算出し図 5 に示す．評価に用いる音声は ATR 音声データベース（216 音素バランス単語）の男性話者（MAU）の最初の 20 単語である．LPC 分析次数および LPC ケプストラム分析次数が高くなるにつれてスペクトル歪みは大きくなっている．この図から 8 bit μ law 入力音声は 16 bit 線形入力音声に対して 1.2 dB 程度の雑音が重畳していることになる．ここでこの算出には音声パワー項を考慮していない．

(平成 10 年 6 月 16 日受付)
(平成 10 年 9 月 7 日採録)

山下 昌毅
1974 年生．1993 年会津大学コンピュータ理工学部コンピュータソフトウェア学科入学．1999 年同大学卒業，大学院入学予定．同大学入学以来，音声認識の研究に従事．



杉山 雅英（正会員）

1954年生．1977年東北大学理学部数学科卒業．1979年同大学院理学研究科数学専攻修士課程修了．同年日本電信電話公社武蔵野電気通信研究所（現在NTT武蔵野研究センター）入所．1985年東北大学より工学博士号を取得．1986年から米国AT&T Bell研究所滞在研究員，1987年からNTT基礎研究所主任研究員，1990年からATR自動翻訳電話研究所主幹研究員の後，1993年から会津大学コンピューター理工学部ヒューマンインタフェース学講座教授．現在まで，LPCスペクトル距離尺度（歪み尺度），ベクトル量子化による音声認識，特徴ベースによる音声認識，教師なし話者適応，テキスト独立話者認識，音声スペクトル推定，情報幾何学（微分幾何学）による音声分析，音響特徴量による言語識別，音声特徴キーによる音声検索などの音声認識処理の研究に従事．日本音響学会，電子情報通信学会，情報処理学会，人工知能学会，IEEE各会員．