ヒューマンインタフェース 77－8
（1998. 3. 13）

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

会津大学 コンピュータ理工学部

あらまし　音声波に含まれる個人性情報を用いて、発話者の認識・認証を行な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。本報告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なうソフトウェア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password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なう。SunOS Release 4.1.3_U1(S-4/IX)における実装を行ない、さらに評価のための話者認証実験を行なった。8bit μ law標本化周波数8kHzという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系の環境でも実装、動作が可能である。

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

M. Yamashita, M. Sugiyama

The University of Aizu

Abstract　The speaker can be recognized using the individual features included in 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 control computer access by the speaker recognition technique, and describes the outline of XVLock and the performance evaluation. The implementation and the experiments did only one standard platform, but XVLock can be applied the other platforms because of less platform dependency. The implemented platform has only low quality voice input system, but the verification performance is 93.9%.

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・照合を行なう話者認識技術[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ー認証としては、IDとpassword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の正当な持ち主で無い場合でも、入力したpasswordが正しければ本人として認証されてしまう。話者認識をpassword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)しユーザーのpasswordで入力により解錠(unlock)を行なうソフトウェアがあり、ユーザーが短時間離席する場合などに用いられている。本報告ではその施錠/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を開発する。

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かなくて良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。これらの二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簡単にシステムを破られてしまうという。第3のテキスト指定による方法 [3] は、システムの指定の発話内容を、話者認識と音声認識技術とを組み合わせる方式であり、指定のテキストを話者が即座に応答できない場合には棄却する。

本報告ではシステムの実現のし易さやユーザーの使用における負担を考えて、テキスト独立の話者認識方式を用いることにする。

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個別の WS 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 の音声入力系の制約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 毎に認証用の話者モデルを持つのではなく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、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 仕様

XVLock では個々の環境に依存する部分に関しては、すべて環境変数で設定を行なうようにする。環境変数の設定をするだけで、すべての環境で XVLock を動作させることが可能である。

使用している環境変数とコマンドラインオプションの対応関係を表 2に示す。コマンドライン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。

以上を踏まえて、以下のような仕様とする。

1. 汎用性を与え、他の UNIX X Window platform に移植可能とする
2. 環境変数、コマンドラインオプションの導入により platform 依存部分を吸収する
3. 施錠を行なう前にマイクロホンの接続確認を行なう
4. 既存の画面施錠、音声入力コマンドの利用する
5. xvlock では password 関連処理を行なわない
6. 標準音声符号化方式: 線形 16bit, 音声特徴抽出: LPC Cepstrum 方式
7. 話者認証方式: VQ 符号帳によるテキスト独立話者認識方式
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

2.2 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 と unlock 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 speaker model 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。

2.3 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 上で開発を行なったが、source file の書換えを行なわずに HP-UX でも動作した。

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では、LPC Cepstrum 分析法を用いている。「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。通常、コンピュー

表 1: platform による音声入力・施錠システムの違い

	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	8bit	16bit	16bit	16bit	16bit
符号化方式	μ-law	線形	線形	aiff	線形
標本化周波数	8kHz	44kHz	16kHz	44kHz	44kHz
符号化変換	必要	不必要	不必要	必要	不必要
integer (16bit)	−−	互換	互換	互換	swap
floating (32bit)	−−	互換	互換	互換	no

−−: これを基準とする

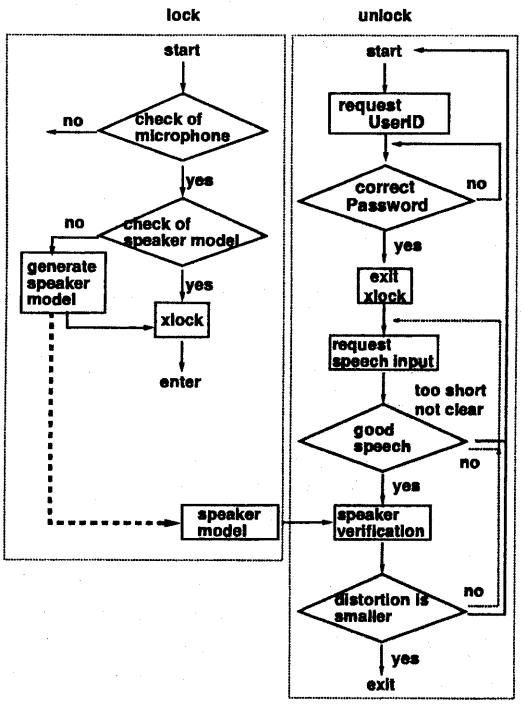

図 1: XVLock の処理の流れ

タに附属のマイクロホンを使用する。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。

音声符号化方式の変換 表 1に示す様に音声録音コマンド名と音声符号化方式 (ファイル形式) は、platform によって異なる。符号化方式を、線形 16bit monoral 形式の raw ファイル (ヘッダーなし) に変換した後、特徴抽出を行う。

音声区間切りだし 入力された音声には無音区間が含まれている為、XVLock において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行なう。

音声波形 (x_t) の F の点を分析単位 (1 frame) とし、各 frame 毎に音声パワー (u_n) を計算する。

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

ここで $F = 256$ であり、標本化周波数 8kHz の時は 32msec に対応する。時間軸に見て power (u_n) の最大値 u_{MAX} と最小値 u_{MIN} の値から SNR (R) を求める。

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

指定の frame の音声/雑音の判定を以下のように行なう。その frame を中心として前後 W frame 分 (合計 $2W+1$ frame 分) の power 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時、音声区間とする。

$$\hat{u}_n > u_{\mathrm{MIN}} + \frac{R}{T}$$

本報告では経験的に $W=3, T=3$ としている。

音声特徴抽出　音声区間として切り出した後、LPC Cepstrum 分析を用いて frame 毎に LPC Cepstrum 係数ベクトルに変換を行なう。ここで、LPC 分析次数、LPC Cepstrum 打ち切り次数は 14 としている。LPC Cepstrum 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 アルゴリズムを用いて作成する。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。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ない、 VQ codebook を自動的に作成する。

3.3 認証判定

作成された $\boldsymbol{V}$ を用いて照合を行なう。認証用の入力音声の特徴ベクトルの列 (x_l) に対して以下の式で VQ 歪み (LPC Cepstrum 距離) を計算する。

$$D=\frac{1}{L}\sum_{l=1}^{L}\min_{1\le m\le M} d(x_l,v_m)$$

ここで L は入力音声のベクトルの数 (入力音声長) であり、$d(x,v)$ は LPC Cepstrum 距離である。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で述べる。

3.4 入力系の音響特性の違いと LPC ケプストラム係数との関係

入力音声を標本化してディジタル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ーで近似できる。従って、入力音声の短時間スペクトル (ピリオドグラム) は線形フィルタリングされ、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。

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

ここで h_n は $h(\lambda)$ のフーリエ係数である。従って、LPC ケプストラムベクトル $\boldsymbol{c}$ は以下のように平行移動することになる。

$$\boldsymbol{c}^*=\boldsymbol{c}+\boldsymbol{h}$$

例えば SUN において 線形 16bit 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、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 h を用いて以下のようにかける。

$$\boldsymbol{V}^*=H(V)=\boldsymbol{V}+(\boldsymbol{h})$$

4 XVLock の使用手順

4.1 installation

XVLock のパッケージのおかれている directory から ftp などで入手し、以下の手順で install を行なう。使用方法などの詳細情報については README を参照する。

```
% tar xvf XVlock.tar
% edit Makefile
% make install
```

Makefile では作成コマンドをおく 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。

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) の設定を~/.cshrc などの中で行なう。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なう。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作成しなければならない。作成のために以下のように行なう。

```
% 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なう。1 単語 3 秒間で 15 単語の録音を行なう。録音は 50 秒程で終了する。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時間は作成する話者モデルの codebook のベクトル数に線形に比例し、VQ 符号帳の大きさが 32 の時、46 秒 (音声切り出しあり)、99 秒 (音声切り出しなし) となる。また 音声切り出し処理を行なうことで、 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) におかれる。

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なうことになる。通常の話者認識システムではマイクロホンの接続は使用者の責任で事前に行なわれていると仮定して動作して良いが、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なわれる。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、正常には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合には、 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事も可能である。

```
% xvlock -d 1.0
```

この数値は 3.3で算出する入力音声の話者モデルに対する LPC cepstrum 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易くなってしまうので、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。また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で述べたように入力音声に対する歪み (D_0) をもとにして、unlock の際に使用する閾値を ($D_T = \alpha D_0$) 設定する。ここで経験的に $\alpha = 1.05$ としている。

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 の音声入力待機画面になる。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、音声入力待機画面に移行する。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 30dB 以下であったり音声区間が 2 秒以下である場合、入力音声による認証を行なわず xlock が再度実行される。 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可能性がある。良好な音声が入力された場合、認証を処理を行ない、受理されれば unlock を行なう。棄却された場合には、もう一度 xlock プログラムが実行される。

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 3に示す。認証には 14 人の話者からの 2 種類の音声を用いた。評価実験の設定条件は SunOS 使用を想定している。認識結果を表 4に示す。ここで認証率は「本人に対する受理率」と「詐称者に対する拒否率」の平均が最大になる最適な閾値を設定し求めた。予備的な実験の結果、音声区間切り出しを行った場合の認証率の方が 5% - 10% 程度高い。VQ 符号帳の大きさに関しては音声切り出しありの時 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。

表 2: オプションと環境変数の対応関係

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	登録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 inpur.raw
-C	XVLOCK_WAVECEP	LPC Cepstrum 分析プログラム名	~/bin/WaveCep

5.2 認証処理速度および必要記憶容量

認証のための処理速度 S は $S \propto M \times L \times N$ で与えられる。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 $M \times N \times 4$ (byte) である。$M = 32, N = 14$ の場合には 1792 byte となる。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毎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。異なる入力音響特性を指定する h を用いれば $N \times 4$ byte 増加するだけである。3.4で述べた話者モデルを共有化法を用いて、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なうソフトウェア 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 8 bit μ law の低品質な入力音声に対して 93.9% の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は、同じ日に連続して収録しているので、話者の体調の変化や経時変化による性能の劣化が予想される。また雑音環境下における性能評価を含め、今後の課題である。本報告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、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ーに使用してもらうことにより、セキュリティやユーザーインターフェースに関しての改善を検討する。

参考文献

[1] D.O'Shaughnessy, "Speaker Recognition", IEEE ASSP Magazine, pp.4-17 (1986). [2] R.J.Mammone, X.Zhang, R.P.Ramachandran, "Robust Speaker Recognition", IEEE Signal Processing, Vol.13, No.5, pp.58-71 (Sep. 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, 5, pp. 647-656, 1996.

表 3: 話者認証の実験条件

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	8bit μ-law, 8kHz
特徴抽出	LPC Cepstrum 分析 (次数: 14)
VQ 符号帳作成法	LBG algorithm
VQ 歪み尺度	LPC Cepstrum 距離
WS	SunOS 4.1.3_U1(S-4/IX)
マイクロホン	HP 社 Headset Microphone

表 4: 話者認証実験結果 (音声切り出しあり)

	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