論　文

音声シフト：音高の意図的な変化を利用した音声入力インタフェース

尾本　幸宏[†]　　後藤　真孝[††]　　伊藤　克亘[††]　　小林　哲則[†]

Speech Shift: Speech Input Interface Using Intentional Control of Voice Pitch

Yukihiro OMOTO[†], Masataka GOTO[††], Katunobu ITOU[††],
and Tetsunori KOBAYASHI[†]

あらまし　本論文では，非言語情報の一つである音高を利用した，「音声シフト」という新たな音声入力インタフェース機能を提案する．従来の音声認識システムが主に言語情報だけを利用してきたのに対し，我々は非言語情報を積極的に活用することによって，音声のもつ潜在能力を引き出した使いやすいインタフェースを構築することを目指している．音声シフトでは，普通に発声した発話と故意に高く発声した発話を異なる入力モードに割り当てることで，音声のみでモード指定と情報入力とを同時に行うことを可能にする．例えば，音声ディクテーションにおいて，「改行」と普通に発声するとその文字が入力され（文字入力モード），それを高く発声すると行末が改行される（コマンドモード）機能が実現できる．こうした機能を実現するために，本研究では，故意に高い発声を識別する際に必要となる話者ごとの音高の基準を，有声休止区間の音高を用いて推定する手法も提案する．実際に，音声テキストエディタに応用し，理工系男性 20 人の被験者による評価実験をしたところ，音声シフトが使いやすく，効果的な入力方法であることが分かった．

キーワード　音声インタフェース，音声シフト，音声認識，音高，非言語情報

1. まえがき

従来の音声インタフェースにおいては，発話された単語あるいは単語列が運ぶ言語的情報（音韻的特徴）のみが，伝達されるべき情報として位置づけられてきた．構文情報や感情など，韻律的特徴が運ぶ様々な情報に注目した研究もあるが，それらにおいても，韻律的特徴は言語情報の円滑な伝達を支えるための補助情報としてしかとらえられてこなかった [1]〜[8]．本研究では，これまでとは異なる視点で韻律利用をとらえることで，音声のもつ潜在能力を引き出し，新たな音声インタフェースの可能性をひらくことを目指す．具体的には，意識的に韻律を制御することによって，積極的に情報を付与し，これによって様々なインタフェース機能を呼び出すことについて検討する．

このような立場に立つ研究として，非言語情報である有声休止（母音が引き延ばされるいいよどみ現象）を活用した後藤らの音声補完がある [9]〜[13]．音声補完では，ユーザがある単語の一部しか思い出せずに断片だけをいっていいよどむと，計算機側がその残りを補って入力することができる．従来，有声休止のようないいよどみは単に誤認識を招く一因と考えられていたが，音声補完ではそれを逆に活用し，有声休止を補完トリガーキー（UNIX 等では特殊キーの “Tab キー”）に位置づけることで，発声途中に故意にいいよどめば残りが補完され，手助けが受けられる使いやすい音声入力が実現された．

本論文では，非言語情報の一つである音高を利用し，声の高さで入力モードを切り換える「音声シフト」という新たな音声入力インタフェース機能を提案する．音声シフトでは，普通に発声した発話と故意に高く発声した発話を異なるモードに割り当てることで，独立したモード切換操作を必要とすることなく，音声のみでシームレスにモード切換を実現できる．

以下，**2.** で「音声シフト」のインタフェース機能を説明する．**3.** では，話者固有の音高の基準の推定法，及び，それに基づく音声シフト識別手法について述べる．**4.** では，音声シフトの音声テキストエディタへの応用を述べるとともに，言語情報を考慮した音声シ

† 早稲田大学理工学部，東京都
School of Science and Engineering, Waseda University, 3–4–1 Okubo, Shinjuku-ku, Tokyo, 169–8555 Japan
†† 産業技術総合研究所，つくば市
National Institute of Advanced Industrial Science and Technology, 1–1–1 Umezono, Tsukuba-shi, 305–8568 Japan

フト識別手法について説明する．**5.** では具体的な実装方法を述べ，**6.** で提案した識別手法の性能を実験的に評価する．**7.** では，音声シフトに関するユーザビリティの評価を行い，**8.** でまとめを述べる．

2. 音声シフト

「音声シフト」では，異なる高さの声で発声する行為を「モードの切換」ととらえることにより，特殊キーである “Shift キー”（あるいは “Alt キー”）の機能を実現する．従来の音声認識では，音声のもつ音韻的特徴のみに着目していたため，同じ単語を別の意味で扱うことは困難であった．本研究では，音声の韻律的特徴の一つである音高を利用することで，普通の高さで発声（通常発声）した発話と故意に高く発声（シフト発声）した発話を，異なるモードに割り当てることができる．

例えば，音声ディクテーションソフトでは，ユーザが「“保存” という文字列として入力したい」，「ファイルへの保存コマンドを実行したい」と思った際に，単に “保存” と発声したのでは区別がつかない問題があった．音声シフトを用いることにより，普通の高さで “保存” と発声するとその文字列が入力され，故意に高く “保存” と発声すると保存コマンドが実行されるといったように，声の高さによって，同じ単語を，異なる意味で扱うことで解決することができる．

従来，このような「文字入力モード」，「コマンドモード」といったモード切換を実現する方法として，以下の二つが用いられていた．一つ目は，キーボードやマウスなど，他の入力装置と併用して切り換える方法である．例えば，モードの切換という機能をショートカットキーとしてキーボードに割り当てたり，マウスでモード切換のボタン等を操作したりしていた．二つ目は，キーワード（予約語）を用いた切換方法である．例えば，まず「コマンドモード」と発声してシステムのモードを明示的に切り換えた後に，実行したいコマンドを「保存」と発声するなど，キーワードを発声してから実際に処理させたい内容を発声する方法があった．あるいは，「コマンドモードで保存」のように，キーワードを語頭に必ず付けて発声する方法もあったが，そのような言語的な方法を使う限り，その言葉自体を入力できないという問題があった（例えば，そうした方法の使用説明書を，その方法自身では音声入力困難であった）．

これらの方法と比較し，音声シフトには以下の利点がある．

- 音声のみで処理可能　マウス等を用いることなく，音声のみで多様な機能を呼び出せる．このため，操作手順が簡略化でき，操作性が向上する．
- 明示的なモード切換が不要　従来のモード切換が必要な方法では異なるモードにあった機能を，現在システムがどのモードであるのかを意識せずに，常にシームレスに呼び出すことができる．同時に，操作時間も短縮される．

3. 実 現 方 法

音声シフトを実現するには，各発話区間が，通常発声（普通の高さで発声）とシフト発声（故意に高く発声）のどちらであるかを識別する必要がある．しかし，人が発話している際の声の高さ，すなわち基本周波数（以下，F_0）は大きく変動しており，話者によって声の高さには個人差がある．そのため，ある発話が故意に高く発声されたものかどうかを識別することは，難しい問題であった．

この問題を解決するためには，話者ごとに固有の音高の基準があるとよい．適切に話者固有の音高の基準を定めることができれば，発話区間中の声の高さを，F_0 という絶対的な尺度でとらえるのではなく，その話者にとって相対的にどれぐらい高く話しているかという尺度でとらえることができるからである．

本章では，話者固有の音高の基準となる基準基本周波数（以下，基準 F_0）を導入するとともに，有声休止区間を用いた基準 F_0 の推定法，及び，基準 F_0 をもとにした音声シフト識別手法について述べる．

3.1 基準 F_0 の導入

基準 F_0 は，話者にとってごく自然な，いわば地声の高さであると考える．藤崎モデル [14] では，基底基本周波数として基準 F_0 に相当する考え方が導入されているが，話者の基準 F_0 としての基底基本周波数を求めるには，ある程度長い安定した発話データを用意することが必要となる．そのため，長い発話をさせるという負担をユーザに与えてしまうことになる．

そこで本研究では，有声休止区間中の F_0 の平均を，基準 F_0 とみなす新たな手法を提案する．有声休止はいいよどみ現象の一つで，その発声中は調音器官の変化が小さくなるため，F_0 が安定し [15]，かつ，地声の F_0，すなわち基準 F_0 に近くなると仮定できる．また，有声休止は人間が自発的に発話する際には自然に現れるため，それを発声することがユーザの負担とはなら

ない，適切な手掛りといえる．更に音声入力中に何度も現れるため，有声休止の発声ごとに漸次的に推定値を更新することで，基準 F_0 の精度を高めることができる．本研究では，この基準 F_0 の更新を MAP 推定（最大事後確率推定）により実現する．なお，最初の有声休止が検出されるまでは基準 F_0 が求められないため，通常発声とシフト発声の識別をせずに，すべて通常発声とみなす．

上記で必要となる F_0 推定と有声休止検出には，文献 [15], [16] で後藤らが提案した F_0 推定手法，有声休止区間の検出手法を用いる．F_0 推定手法は，背景雑音等を伴う音響信号に対してもロバストに機能する特長をもち，コムフィルタの考え方に基づいて，入力信号中で最も優勢な高調波構造の F_0 を，音声の F_0 として推定する．一方，有声休止区間の検出手法は，任意の母音の引き延ばしを言語非依存に検出できる特長をもち，有声休止がもつ二つの音響的特徴（F_0 の変動が小さい，スペクトル包絡の変形が小さい）をボトムアップな信号処理によって検出する．これらの手法は，リアルタイムに動作させることができる．

3.2 有声休止区間の F_0 の分析

予備実験として，有声休止区間の F_0 がどの程度安定しているかを調べた．男性話者 6 人ごとの有声休止区間の F_0 の標準偏差は，平均 86.2[cent](注1)と小さかった．また，図 1 に，各話者ごとの有声休止の種類（「んー」「えー」「あのー」）による平均 F_0 を示す．図から，有声休止の種類が異なっても平均 F_0 はほぼ同一であるが，話者が異なると平均 F_0 が大きく異なることから，話者ごとに基準 F_0 を求める必要性があることが分かる．以上から有声休止区間の F_0 はほぼ一定で安定しており，話者の基準 F_0 として利用可能と判断した．

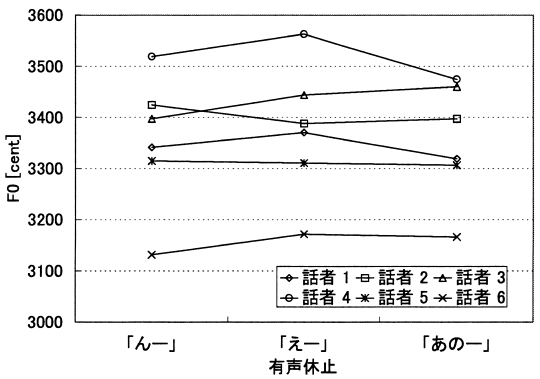


図 1　6 人の話者による 3 種類の有声休止での F_0 の平均
Fig. 1　Average F_0 for three typical Japanese fillers (with filled pauses) uttered by six speakers.

3.3 手法 1：基準 F_0 からのしきい値を用いた音声シフト識別手法

音声シフトを実現するために，各発話区間の音高の平均を求め，それが通常発声とシフト発声のいずれのカテゴリーに属するかを識別する．ただし，話者ごとの声の高さの違いを正規化するために，音高の平均から基準 F_0 を引いた値（音高の平均の基準 F_0 からの相対的な高さ）を用いる．以下，この値を相対発声音高と呼ぶ．

ここでは，通常発声とシフト発声を相対発声音高に基づいて識別する手法（手法 1）を提案する．本手法では，通常発声とシフト発声の二つのカテゴリーの相対発声音高の境界を，しきい値として明示的に求め，各発話の相対発声音高がそのしきい値より大きければシフト発声，小さければ通常発声と識別する．このしきい値は，基準 F_0 からの相対的な高さで表現され，通常発声とシフト発声の二つのカテゴリーが既知の学習データに対して，識別率を最大にするように定める．

本手法による識別の模式図を図 2 に示す．図の左側は，「えー・六年が経過・改行」と発話したときの F_0 パターンを表している（「改行」は故意に高く発声）．図は，有声休止区間 (A) の F_0 が基準 F_0 推定に用いられ，(B) の区間の平均 F_0 はしきい値（THLD: threshold）以下であるため「通常発声」と判定され，(C) の区間の平均 F_0 はしきい値以上であるため「シフト発声」と判定されることを示している．

4. 音声シフト機能付き音声テキストエディタ

音声シフトは，様々なアプリケーションへ応用できるが，本研究ではその一例として，音声入力部に音声シフト機能をもった音声テキストエディタを実現した．この音声テキストエディタでは，「保存」等のコマンドをシフト発声することによって実現できる．このように文章入力中でシフト発声する場合，シフト発声される個所の前後関係には言語的な特徴があると考えられる．これは **3.3** の手法 1 では考慮されていない，言語

(注1)：cent の単位は，音高差（音程）を対数スケールの周波数で表す尺度で，半音差が 100[cent] に相当する．本論文では，Hz で表された周波数 f_{Hz} を cent で表された周波数 f_{cent} へ，$f_{\mathrm{cent}} = 1200 \log_2 \frac{f_{\mathrm{Hz}}}{440 \times 2^{\frac{3}{12}-5}}$ により変換する．

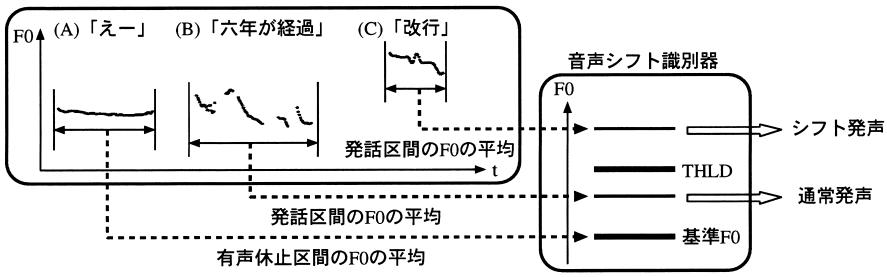


図 2 音声シフト識別手法（手法 1）の概要
Fig. 2 Overview of the speech-shift classification method (method 1).

的な事前知識である．ここでは，音声テキストエディタの概要を説明するとともに，そうした言語的な事前知識を考慮した音声シフト識別手法について述べる．

4.1 音声シフト機能付き音声テキストエディタの概要

従来の音声テキストエディタでは，**2.** で述べたように，ボタンまたはキーワードによって，文字入力モードとコマンドモードを切り換えながらテキスト入力していた．それに対して本研究では，通常発声を「文字入力モード」での入力，シフト発声を「コマンドモード」の機能呼出しに割り当てる．用意した後者の機能には，編集機能（削除，後退，太字，右寄せ，中寄せ，改行，やり直し，切取り，貼付け），文字列検索機能，ファイル操作機能（保存，開く，印刷，終了），文字入力対象切換機能（平仮名，片仮名，アルファベットの切換）がある．例えば，通常の文章を音声入力中に「改行」と高い声で発声すると，その発話は文字として入力されずに「改行」機能が呼び出され，効率良く文章入力できる．

また，一般的な音声エディタでは，ユーザが考えながら文章入力するときでもいいよどむことは許されないことが多いが，本研究で実装した音声エディタでは，基準 F_0 を推定するために，ユーザにいいよどむことを許容している．そのため，有声休止が検出された発話は，基準 F_0 の MAP 推定のみに用いられ，エディタへ入力されることがないようにした．

4.2 言語的な事前知識の利用

音声テキストエディタで，文章入力しながら音声シフトを利用した場合における発話の書き起こし例を図 3 に示す．図中，音声認識システムの出力に合わせて文章が単語に区切られており，silB, silE は発話の開始及び終了を表す記号，sp は発話中の小休止を表す記号となっている．また，山形括弧（< と >）で囲まれた単語は，シフト発声されて実際にコマンドが実行された単語である．

```
silB 改行 したい 場合 silE
silB 故意 に sp 高く 話す こと に よって silE
silB 実行 でき ます silE
silB <改行> silE
```

silB, silE: 発話の開始 (silB)，終了 (silE) を表す記号
sp: 発話中の小休止を表す記号
<XXX>: シフト発声により XXX コマンドが実行された単語

図 3 音声テキストエディタ利用時の書き起こし例
Fig. 3 An example of transcription using our voice-enabled word processor.

この例を人間が見て判断すれば，コンテクスト（単語の並び）から，第 1 行の「改行」は文字列としての入力，第 4 行の「改行」はコマンド入力であると推測できる．このように，あるコンテクストでは，コマンドか非コマンドかを，その前後の単語から判定できる場合がある．

シフト発声を識別する際にも，これら言語的な情報を事前知識として利用することで，より効果的な識別ができる可能性がある．例えば，コマンドでないものを少し高めに発声してしまった場合に，**3.3** の手法 1 だけでは，シフト発声と誤識別されてしまう可能性がある．このような場合でも，言語的な情報から，その発声がコマンドでない可能性が高ければ，適切に通常発声と判断されることが期待できる．

4.3 手法 2：言語的な事前知識を組み合わせた音声シフト識別手法

本手法では，音高情報と言語的な事前知識を組み合わせることで識別率の向上を目指す．各フレーム（10 ms シフト）ごとのスペクトルデータ列を $X = \{x_1, x_2, \cdots, x_N\}$（$N$ はフレーム数），音高列を $A = \{a_1, a_2, \cdots, a_N\}$，単語列を $W = \{w_1, w_2, \cdots, w_K\}$（$K$ は単語数）とし，各単語の発声がシフト発声かど

うかを表す指標の列を $C = \{c_1, c_2, \cdots, c_K\}$ とする．ここで，c_k はコマンド指標と呼んで，通常発声であれば $c_k = 0$，シフト発声であれば $c_k = 1$ とする．このとき，発話内容及び発話区間がシフト発声かどうかを同時推定することは，X, A が与えられたときの $P(W, C|X, A)$ を最大化する W, C を求めることにあたる．この推定問題は，次のように定式化される．

$$
\begin{aligned}
&\{\hat{W}, \hat{C}\} \\
&= \underset{W,C}{\operatorname{argmax}} P(W, C|X, A) && (1)\\
&= \underset{W,C}{\operatorname{argmax}} P(C|W, X, A) \cdot P(W|X, A) && (2)\\
&\cong \underset{W,C}{\operatorname{argmax}} P(C|W, A) \cdot P(W|X) && (3)\\
&= \underset{W,C}{\operatorname{argmax}} \frac{P(A|C, W) \cdot P(C|W)}{P(A|W)} \cdot P(W|X) && (4)\\
&\cong \underset{W,C}{\operatorname{argmax}} \frac{P(A|C) \cdot P(C|W)}{P(A)} \cdot P(W|X) && (5)\\
&= \underset{W,C}{\operatorname{argmax}} P(A|C) \cdot P(C|W) \cdot P(W|X) && (6)
\end{aligned}
$$

上式の導出にあたっては，スペクトルデータ列 X と指標 C，音高列 A と単語列 W とは互いに独立としている．ここで更に，式 (6) の $P(A|C)$ を，

$$
P(A|C) \cong \prod_{k=2}^{K-1} P(\bar{a}_k|c_k) \tag{7}
$$

と近似することにする（$k = 1$ は発話の開始 silB に，$k = K$ は発話の終了 silE に対応するため除外する）．$\bar{a}_k$ は，単語 w_k の区間における平均音高と基準 F_0 との差であり，単語音高と呼ぶことにする．$P(\bar{a}_k|c_k)$ は，単語がシフト発声であるかないかが与えられたときに，どのような単語音高 $\bar{a}_k$ が出力されるかを表す確率であり，単語音高モデルと呼ぶ．単語音高モデルの具体的な構成法については **4. 3. 1** に述べる．$P(C|W)$ は，各単語がコマンドであるか非コマンドであるかを単語列から判断する事前確率であり，コマンド生起モデルと呼ぶ．コマンド生起モデルの具体的な構成法については **4. 3. 2** に述べる．$P(W|X)$ は従来の音声認識システムから出力される確率そのものである．

連続音声認識においても重みを介して言語モデルと音響モデルを結合するように，ここでも単語音高モデル，コマンド生起モデルは，以下のように重み α を介して結合することとする．

$$
\begin{aligned}
&\{\hat{W}, \hat{C}\} = \underset{W,C}{\operatorname{argmax}} \\
&\left(P(A|C)^{\alpha} \cdot P(C|W)^{(1-\alpha)}\right)^{\frac{1}{K-2}} \cdot P(W|X)
\end{aligned} \tag{8}
$$

式 (8) 中の $\frac{1}{K-2}$ 乗は，単語数での正規化を意味する．以上のように言語的な事前知識を利用してシフト発声の識別を行う方法を（手法 2）と呼ぶ．

式 (8) を解く場合，理想的には，式を最適化する単語列とコマンド指標列を，すべての単語境界仮説を網羅する形で求めることが望まれるが，この場合アルゴリズムは煩雑化する．そこで今回は，第 1 パスにおいて，言語モデルと音響モデルだけを使って単語列の N-best 候補を求めた上で，第 2 パスで，単語音高モデルとコマンド生起モデルによって，リスコアリングするというアプローチを採用する．

4. 3. 1 単語音高モデル

単語音高モデルは，あるコマンド指標（通常発声かシフト発声か）が与えられたときに，どのような単語音高が出力されやすいかを表したものである．これを求めるには，通常発声とシフト発声のそれぞれの単語音高の分布をモデル化する必要がある．単語音高の分布は，予備実験の結果から正規分布に近くなるため，ここでは正規分布でモデル化し，確率分布として表す．二つの正規分布の平均と分散は，カテゴリーが既知の学習データの単語音高から最ゆう推定する．

4. 3. 2 コマンド生起モデル

コマンド生起モデルは，ある単語列に対するコマンド指標列の生起確率を与えるものである．これを求めるには，単語列とコマンド指標列とが組になった多量の学習データが必要だが，一般に単語列の組合せは膨大な量になるため，そのような学習データを用意することは現実的でない．そこで，単語を少数のクラスに分類し，次の近似式を導入して，単語クラスの三つ組みと中央の単語の発話がコマンドかどうかの関係を調べることで，コマンド生起モデルを求める．

$$
\begin{aligned}
P(C|W) &\cong \prod_{k=2}^{K-1} P(c_k|w_{k-1}, w_k, w_{k+1}) && (9)\\
&\cong \prod_{k=2}^{K-1} P(c_k|v_{k-1}, v_k, v_{k+1}) && (10)
\end{aligned}
$$

ここで，v_k は単語 w_k が属する単語クラスを表す．本実験では，v_k として，以下の 3 種類の単語クラス（S, U, C）を用意した．

- S：無音区間（silB, silE, sp）
- U：コマンドとして使われない単語
- C：コマンドとして使われる単語

5. 実　　装

以上述べた音声シフト識別手法をもとに，**4.1** で述べた音声シフト機能付き音声テキストエディタを実装した．システム全体の処理の流れを図 4 に示す．本システムを構成する図 4 の七つの処理は，分散環境で動作する別々のプロセスとして実装した．これらの通信には，音声言語情報をネットワーク上で効率良く共有することを可能にするネットワークプロトコル RVCP（Remote Voice Control Protocol）[9]〜[12] を用いた．

音声認識部には，連続音声認識コンソーシアムが提供するツールキットを用いた [17]．このうち，認識エンジンは Julius 3.2，音響モデルは男性話者用 PTM，言語モデルは 2 万語彙のモデルを用いた．コマンドの語句は，言語モデルの学習時の 2 万語に含まれない語彙を一つの仮想的な語としてまとめた「未知語（UNK）カテゴリー」に追加した．手法 2 で音声認識結果の N-best 候補のスコアと，コマンド生起モデル，単語音高モデルの組合せ計算を行う必要がある．ここでは N-best 候補数を 5 とし，音声シフト識別部へと送信する．

音声シフト識別部では，基本周波数推定部，有声休止検出部，音声認識部の結果を受信し，**3.3** の手法 1，**4.3** の手法 2 のいずれかの手法により，各発話が通常発声かシフト発声かを識別する．有声休止が検出されていれば基準 F_0 の更新もここで行う．そして，識別結果，及び，算出された基準 F_0，発話区間の平均 F_0 をインタフェース管理部へと送信する．

インタフェース管理部（エディタ）では，音声シフト識別部からの結果を受信し，通常発声であれば「文字入力モード」で音声認識結果を文字列として入力する．また，シフト発声であれば「コマンドモード」としてとらえ，音声認識結果に対応するコマンドを実行する．

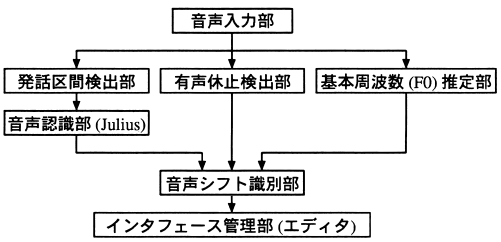


図 4　音声シフト機能付き音声テキストエディタの処理の流れ
Fig. 4　System overview of the speech-shift-enabled word processor.

5.1 実 行 例

画面表示例を図 5 に示す．各図の右側は，基準 F_0（F_0 Base），しきい値（THLD），発話区間平均 F_0（F_0 Mean）のモニタであり，図 2 中の音声シフト識別器の枠内と対応している．通常発声とシフト発声の識別結果もモニタ上部に表示されるとともに，シフト発声の場合，画面最下部のステータスバーに，対応するコマンドが実行されたことを示すメッセージが表示される．

図 5 (a) は，通常発声した場面である．この時点では，過去に有声休止が検出されていないので，基準 F_0 は推定できていない．したがって，発話区間の平均 F_0 のみが表示されている．図 5 (b) は，有声休止が検出され，基準 F_0 の推定（更新）を行った場面である．図では基準 F_0 としきい値が表示されている．図 5 (c) は，「改行」をシフト発声した場面である．図から実際に改行コマンドが実行された様子が分かる．図 5 (d) は，「保存」とシフト発声した場面である．入力内容がファイルに保存され，図のステータスバーに保存コマンドを実行したメッセージが表示されている．

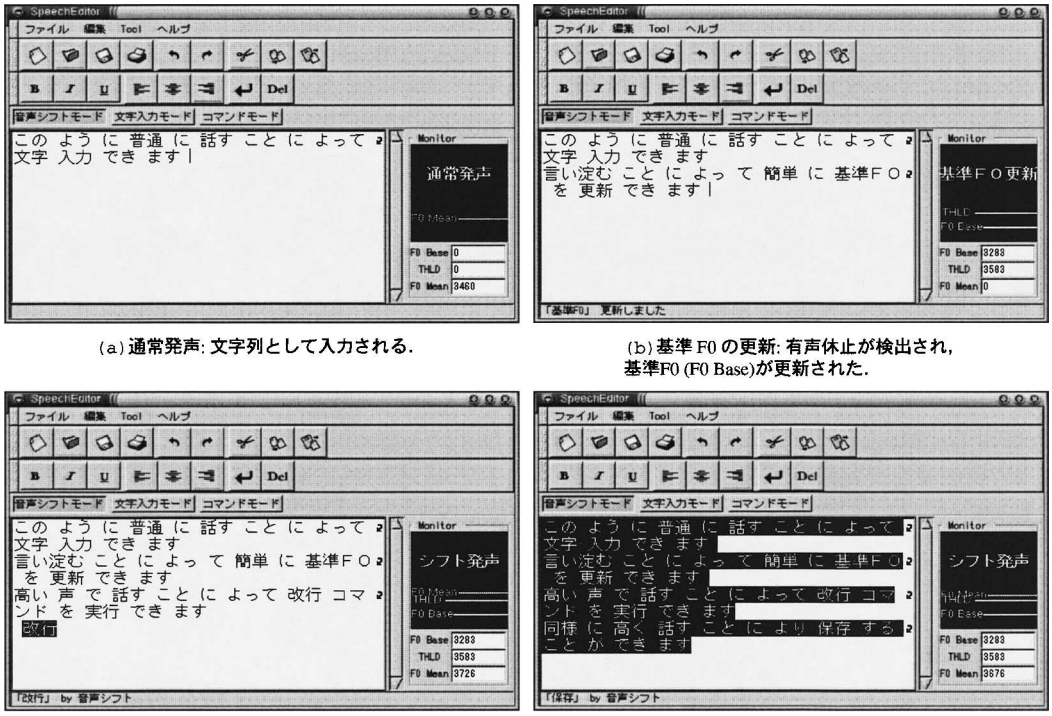
6. 音声シフト識別実験

提案した二つの識別手法の性能を実験的に評価する．

6.1 実 験 条 件

（a） 実験データ　　被験者は，男性話者 12 人である．ここでは，音声テキストエディタとして用いられる状況を想定し，通常発声で文字入力する場合とシフト発声でコマンド入力する場合を交互に収録した．前者の通常発声には様々な長さのタスクを用意した．その内容は，短い単語・複合語の計 30 発話，長いフレーズ・文章の計 30 発話であり，それぞれの半数の 15 発話にはコマンドとしても使われる単語を交ぜてある．通常発声とシフト発声を交互に収録したため，最終的な実験データは，通常発声（60 発話），シフト発声（60 発話）となる．音声データは，DAT に 48 kHz, 16 bit で録音したものを，16 kHz にダウンサンプリングして用いた．

（b） 実験方法　　通常発声とシフト発声の識別率を 3-fold cross validation 法で評価した．具体的には，



(a) 通常発声: 文字列として入力される.

(b) 基準 F0 の更新: 有声休止が検出され，基準F0 (F0 Base)が更新された.

(c) シフト発声 (改行): 平均F0 (F0 Mean)が閾値(THLD)を上回りシフト発声と識別され，実際に改行が行われている.

(d) シフト発声 (保存): シフト発声と識別され，保存コマンドが実行された.

図 5　音声シフト機能付き音声テキストエディタの画面表示例
Fig. 5　Screen snapshots of the speech-shift-enabled word processor.

通常発声，シフト発声をそれぞれ 20 発話（短い単語 10 発話と長いフレーズ 10 発話）ずつ 3 組に分け，2 組を各手法の学習データとし，残りの 1 組を評価用のデータとした．これを 3 通りすべての組合せで行い，その識別率の平均値を評価結果とした．また，手法 2 では単語ごとに識別結果が得られるが，すべての単語で正しく識別された場合のみ正解とした．

6.2 検討項目

3.3 の手法 1，**4.3** の手法 2 を比較評価する．

実験 1：(手法 1)　通常発声とシフト発声の二つのカテゴリー境界のしきい値を変化させ，学習データに対する識別率を最大にするしきい値を求めた．その際，話者ごとに別々にしきい値を設定する場合（話者依存）と，全話者共通のしきい値を設定する場合（話者非依存）の二つの条件を用意し，比較実験を行った．

実験 2：(手法 2)　学習データから，二つのカテゴリーのそれぞれの単語音高モデルを求めた．コマンド生起モデルの学習コーパスには，音声シフト機能付き音声テキストエディタを実際に数分間使用した履歴（1 人の話者による 240 単語の入力）を用いた．単語音高モデルとコマンド生起モデルとの重み α を 0.0 から 1.0 まで変化させて実験し，実験 1 と同様に，話者ごと（話者依存）のモデルと，全話者共通（話者非依存）のモデルを使用する二つの条件で比較実験を行った．

6.3 実験結果

各手法の実験結果を図 6 に示す．どちらの手法においても，話者依存の場合の方が高い識別率となった．同時に，話者依存，話者非依存のいずれの場合も，手法 2 の識別率の方が高く，言語的な事前知識の導入が有効であったことが分かる．手法 2 の重み α に関しては，話者依存のモデルと話者非依存のモデルのいずれの場合も，$\alpha = 0.6$ のときに最も識別率が高く，図 6 はその結果を示している．

7. 評価実験

音声シフトのユーザビリティの評価実験として，音声テキストエディタで「文字入力」と「コマンド入力」を切り換える以下の四つの方法について比較検討する．

(A) マウス操作を用いたボタンによるモード切換
(B) 音声（キーワード）によるモード切換
(C) シフトキーによるコマンド入力
(D) 音声シフトによるコマンド入力

方法 (A)，(B) は，モードを明示的に切り換えながらテキスト入力する方法である．方法 (C) は，音声シフトで音高を高くする代わりにキーボードのシフトキーを押す方法で，シフトキーを押しながら発声するとコマンドモードでの入力となる．方法 (D) は，**3.3** の手法 1 を用い，話者ごとに別々にしきい値を設定した．

7.1 実験方法

被験者はテキストエディタを使い慣れている理工系の男性話者 20 人である．まず，被験者に切換方法 (A)〜(D) についての説明をした後，各方法で練習をさせた．その後，紙面に記載されたテキストを各切換方法を用いて入力させた．入力対象テキストの文字数は 62 文字で，紙面には，「改行」「中寄せ」などコマンド入力する個所（19 個所）も表記されている．テキストは四つの方法とも同じものを使用し，方法 (A)〜(D) の順番は話者ごとにランダムに変えた．その際，課題を完了するまでの操作時間を記録した（実験 1）．

次に，入力対象テキストを変更し，コマンド入力のたびに方法 (A)〜(D) の使いやすいものを選びながら入力させた．文字数が 73 文字，コマンド入力個所が 18 個所のテキストを用い，被験者が選んだ切換方法の種類及び回数を記録した（実験 2）．

最後に，被験者にはアンケートとして，表 1 の各項目について，5 段階評価（5：肯定的，3：どちらともいえない，1：否定的）をさせた．また，切換方法 (A)〜(D) のすべての対比較により，どちらを好んで使いたいかを 5 段階評価させた．

7.2 実験結果

実験 1 の四つの切換方法において，課題を完成させるまでの操作時間の平均と標準偏差を図 7 に示す．この操作時間は，被験者がコマンド入力によって音声認識誤りを修正する時間も含んでおり，修正が容易にできるかどうかも結果に反映されている．

平均操作時間によると，方法 (A) のボタンによるモード切換と方法 (D) の音声シフトが最も操作時間が短かった．次に方法 (C) のシフトキーが短く，最も長かったのは方法 (B) の音声（キーワード）によるモード切換であった．方法 (A) が短かったのは，被験者がこのようなテキストエディタにおける操作に慣れているからと考えられ，方法 (B) が最も長かったのは，モードを切り換える際にキーワードを発話する必要があり，他の方法よりも時間がかかったためである．ここで，音声シフトは方法 (A) とほぼ同様の操作時間となっており，スムーズな入力が可能であったと考えられる．しかし，音声シフトの場合の標準偏差は，ほかよりも大きく，被験者による操作時間のばらつきがあったことが分かる．

次に実験 2 において，被験者が選んだ切換方法の平均回数と割合を表 2 に示す．この結果から，音声シフトが高い割合で使用されたことが分かる．音声シフトを使用しなかった被験者は 1 人もおらず，20 人中 14 人の被験者は，テキスト入力中に音声シフトによるコマンド入力のみを使っていた．

なお，本評価実験の方法 (D) での通常発声とシフト

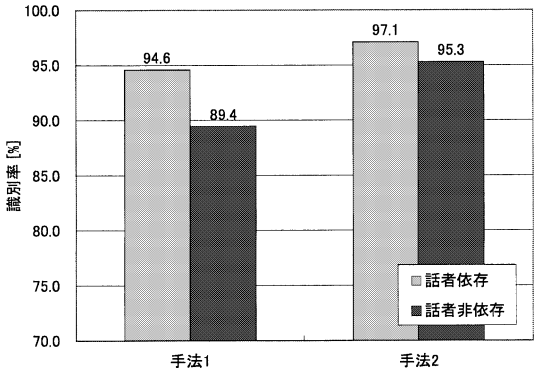


図 6 通常発声とシフト発声を識別する性能の評価結果
Fig. 6 Evaluation results regarding the ability to distinguish between normal and shift utterances.

表 1 アンケート項目
Table 1 Questionnaire items.

番号	質問
1	操作に手間はかからなかったか.
2	使い方は簡単に分かったか.
3	予測しない動作が多く生じなかったか.
4	誤入力の修正はしやすかったか.
5	すぐに慣れたか.
6	慣れると使いやすかったか.
7	操作が軽快で，楽だったか.
8	今後，使いたいか.

表 2 四つの切換方法が使用された割合
Table 2 Usage ratio of the four different mode-switching methods.

	(A)	(B)	(C)	(D)
平均使用回数	2.5	0	3.9	25.3
割合 [%]	7.9	0.0	12.3	79.8

発声の識別率は 97.5%で，図 6 の手法 1 で話者依存のときの識別率より高かった．これは，図 5 の各画面の右側の音高表示が有効に機能し，各発話の音高がしきい値よりもどれぐらい上下しているかがフィードバックされるため，被験者が自分で音高を調節できたからだと考えられる．

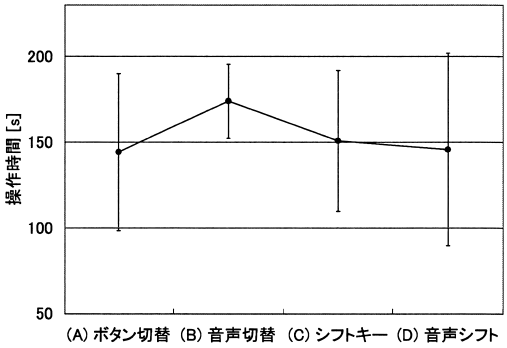


図 7　操作時間の平均と標準偏差の測定結果
Fig. 7　Results of measuring the average and the standard deviation of operation time.

図 8 に表 1 のアンケート結果を示した．多くの項目において，モードを明示的に切り換える方法 (A), (B) とシームレスなモード切換方法 (C), (D) ではそれぞれ同じ傾向があり，後者の方がより評価が高い傾向にあることが分かる．項目 1, 4, 6, 7, 8 は方法 (C), (D) が高く，方法 (A), (B) が低いという同じ傾向が見られ，シフトキーや音声シフトは，慣れると使いやすく，操作に手間がかからず楽であり，被験者は今後も使いたいと思ったことが分かる．また，それにより修正操作も行いやすかったと考えられる．項目 5 では方法 (A), (B) も約半数の支持を得ており，方法 (C), (D) だけでなく方法 (A), (B) においてもすぐに慣れることができたと分かる．項目 2 はすべての方法で高い支持を得ていることから，どの方法も使い方は簡単

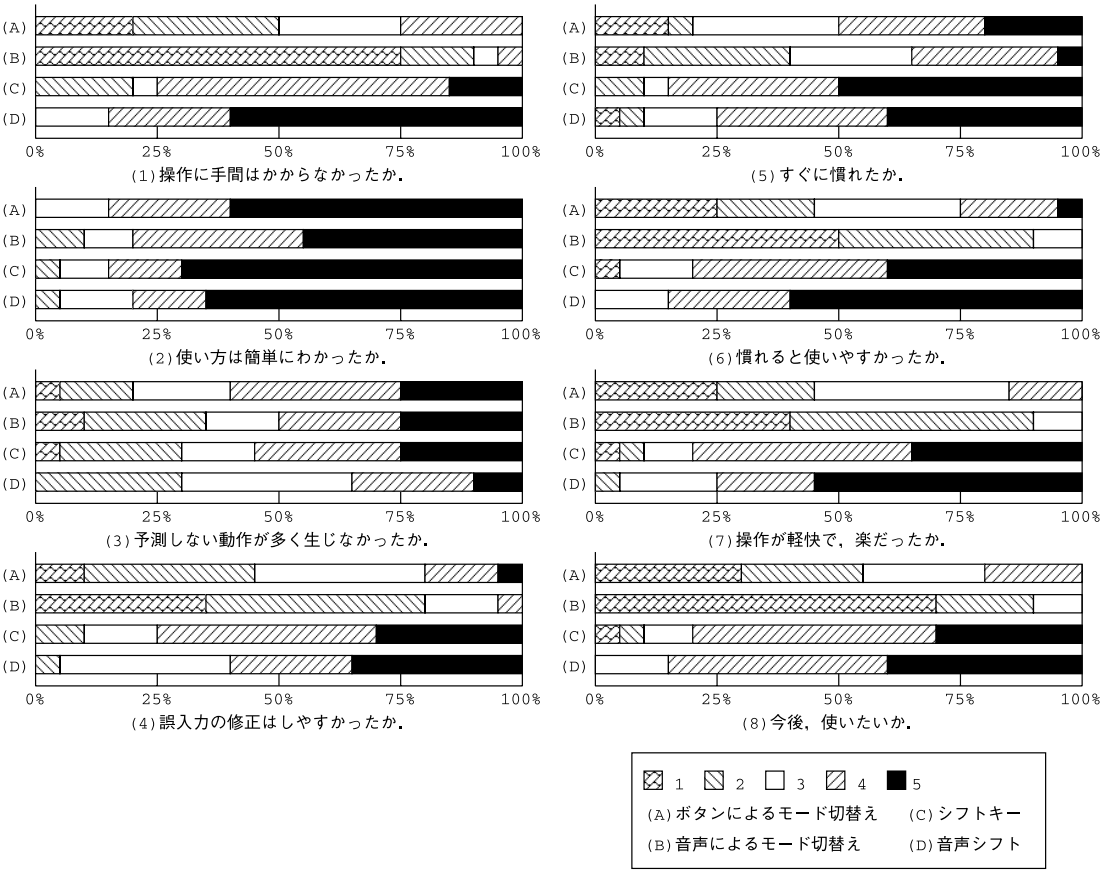


図 8　アンケート結果（8 項目の評価）
Fig. 8　Results of the questionnaire (evaluating 8 items).

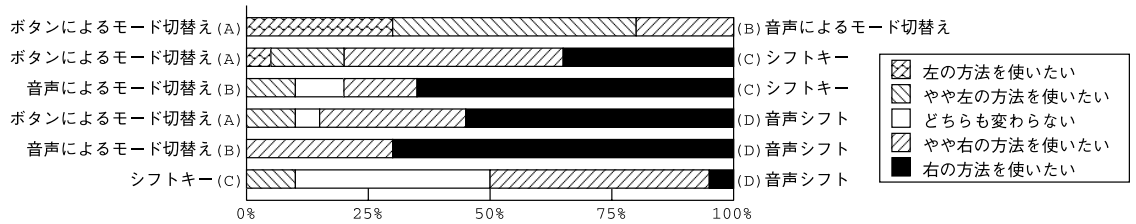


図 9　アンケート結果（対比較）
Fig. 9　Results of the questionnaire (pairwise comparisons).

であったことが分かる．また，項目 3 で方法 (D) の音声シフトが他の方法よりも若干評価が低い．これは，シフト発声と通常発声の誤識別が影響している可能性がある．しかし，他のメリットが上回っていたために，総合的には音声シフトは高く評価されていたと考えられる．

最後に，図 9 に方法 (A)〜(D) の各切換方法すべての対比較により，使うとすればどちらを使いたいかというアンケート結果を示した．この結果からも，方法 (A)，(B) のモードを切り換えるインタフェースより方法 (C)，(D) のシームレスなモード切換の方が好まれることが分かる．また，方法 (C) と (D) の比較では，50%の被験者が音声シフトを支持しており，シフトキーを支持する被験者は 10%であったことから，音声シフトが多くの被験者に使いやすいインタフェース機能として受け入れられたことが分かる．

8. むすび

本論文では，ユーザが意図的に制御する音高を利用して，普通の高さで発声したときと故意に高く発声したときのモード切換を実現する，新たな音声入力インタフェース機能「音声シフト」を提案した．また，有声休止区間の F_0 を用いた，話者ごとの基準 F_0 の推定法，及び，それに基づくシフト発声の識別手法を提案した．更に，音声シフトを音声テキストエディタに応用するとともに，言語的な情報を効果的に組み合わせるシフト発声の識別手法を提案した．音声シフトは通常のコミュニケーションでは明示的には用いられないため，本エディタを初めて使用する際には少々とまどうことも考えられる．評価実験の結果では，理工系男性 20 人の被験者の多くはすぐに操作に慣れ，慣れると使いやすく今後も使いたいと思われるインタフェースであることが分かった．

「音声補完」「音声シフト」の一連の研究は，非言語情報を導入することで使いやすい音声インタフェースを構築していこうというメッセージをもっている．キーボードとの対比で考えれば，従来の音声認識が扱ってきたのは，通常キーの一部に過ぎない．それに対して，「音声補完」「音声シフト」はいわば特殊キーの Tab キーや Shift キーを実現したものととらえることができる．今後，更に他の非言語情報によって音声の潜在能力を引き出すことにより，音声インタフェースの可能性がどのように広がるかについて，幅広く検討していきたい．

文　献

[1] A. Waibel, “Prosodic knowledge sources for word hypothesization in a continuous speech recognition system,” Proc. ICASSP 87, pp.856–859, 1987.
[2] 小松昭男，大平栄二，市川　熹，“韻律情報を利用した構文推定およびワードスポッティングによる会話音声理解方式,” 信学論（D），vol.J71-D, no.7, pp.1218–1228, July 1988.
[3] 高橋　敏，松永昭一，嵯峨山茂樹，“ピッチパタン情報を考慮した単語音声認識,” 信学技報，SP90-17, 1990.
[4] 江口徳博，尾関和彦，“韻律情報を利用した係り受け解析,” 音響誌，vol.52, no.12, pp.973–978, 1996.
[5] 前川喜久雄，北川智利，“パラ言語情報の生成と知覚,” 信学技報，SP99-10, 1999.
[6] A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür, “Modeling the prosody of hidden events for improved word recognition,” Proc. Eurospeech ’99, pp.311–314, 1999.
[7] K. Hirose and K. Iwano, “Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition,” Proc. ICASSP 2000, pp.1763–1766, 2000.
[8] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, “VERBMOBIL: The use of prosody in the linguistic components of a speech understanding system,” IEEE Trans. Speech Audio Process., vol.8, no.5, pp.519–532, 2000.
[9] 後藤真孝，伊藤克亘，速水　悟，“音声補完：‘TAB’ on Speech,” 情処学音声言語情報処理研報，2000-SLP-32-16,

pp.81–86, 2000.
[10] M. Goto, K. Itou, T. Akiba, and S. Hayamizu, “Speech completion: New speech interface with on-demand completion assistance,” Proc. HCI International 2001, vol.1, pp.198–202, 2001.
[11] 後藤真孝，伊藤克亘，秋葉友良，速水　悟，“音声補完：音声入力インタフェースへの新しいモダリティの導入，” コンピュータソフトウェア，vol.19, no.4, pp.10–21, 2002.
[12] M. Goto, K. Itou, and S. Hayamizu, “Speech completion: On-demand completion assistance using filled pauses for speech input interfaces,” Proc. ICSLP-2002, pp.1489–1492, 2002.
[13] 後藤真孝，“解説　音声補完：言い淀むと助けてくれる音声インタフェース，” 情報処理，vol.43, no.11, pp.1210–1216, 2002.
[14] H. Fujisaki and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” J. Acoust. Soc. Jpn. (E), vol.5, pp.233–242, 1984.
[15] 後藤真孝，伊藤克亘，速水　悟，“自然発話中の有声休止箇所のリアルタイム検出システム，” 信学論（D-II），vol.J83-D-II, no.11, pp.2330–2340, Nov. 2000.
[16] M. Goto, K. Itou, and S. Hayamizu, “A real-time filled pause detection system for spontaneous speech recognition,” Proc. Eurospeech ’99, pp.227–230, 1999.
[17] 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄，音声認識システム，オーム社，2001.

（平成 16 年 2 月 6 日受付，7 月 12 日再受付）

尾本　幸宏

2000 早大・理工・電気電子情報卒．2002 同大大学院修士課程了．同年松下電器産業（株）入社．在学中，音声認識，ヒューマンインタフェースに関する研究に従事．日本音響学会会員．

後藤　真孝　（正員）

1993 早大・理工・電子通信卒．1998 同大大学院博士課程了．同年，電子技術総合研究所（2001 に産業技術総合研究所に改組）に入所し，現在に至る．2000 から 2003 まで科学技術振興事業団さきがけ研究 21「情報と知」領域研究員を兼任．博士（工学）．音楽情報処理，音声言語情報処理等に興味をもつ．1992 jus 設立 10 周年記念 UNIX 国際シンポジウム論文賞，1993 NICOGRAPH’93 CG 教育シンポジウム最優秀賞，1997 情報処理学会山下記念研究賞（音楽情報科学研究会），1999 平成 10 年電気関係学会関西支部連合大会奨励賞，2000 WISS2000 論文賞・発表賞，2001 日本音響学会第 18 回粟屋潔学術奨励賞・第 5 回ポスター賞，2002 情報処理学会山下記念研究賞（音声言語情報処理研究会），2002 日本音楽知覚認知学会研究選奨，2003 インタラクション 2003 ベストペーパー賞各受賞．情報処理学会，日本音響学会，日本音楽知覚認知学会，ISCA 各会員．

伊藤　克亘

1988 東工大・工・情報工卒．1993 同大大学院博士課程了．博士（工学）．1993 電子技術総合研究所入所．2003 名古屋大学大学院情報科学研究科助教授，現在に至る．音声を主とした自然言語全般に興味をもつ．情報処理学会，日本音響学会各会員．

小林　哲則　（正員）

1980 早大・理工・電気卒．1985 同大大学院博士課程了．1997 より早大・教授．音声・画像処理を中心とする知覚情報システムの研究に従事．2001 年度本会論文賞受賞．情報処理学会，日本音響学会，IEEE 等各会員．