ヒューマンインタフェース 97－3
音声言語情報処理 40－3
（2002．2．1）

音声シフト："SHIFT" on Speech

尾本 幸宏† 後藤 真孝†† 伊藤 克亘†† 小林 哲則†

†早稲田大学 理工学部 電気電子情報工学科
〒169-8555 東京都新宿区大久保 3-4-1
††産業技術総合研究所
〒305-8568 茨城県つくば市梅園 1-1-1

E-mail: †{omoto, koba}@tk.elec.waseda.ac.jp, ††{m.goto@, itou@ni.}aist.go.jp

あらまし 本稿では，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．我々は，従来の音声認識システムにおいて未使用だった非言語情報を活用することによって，音声の持つ潜在能力を引き出す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モードに割り当てることで，シームレスなモード切替えを音声のみで実現した．実際に，音声テキストエディタに応用したところ，効率よく文字入力できる，効果的なインタフェースを構築できた．

Speech Shift: "SHIFT" on Speech

Yukihiro Omoto† Masataka Goto†† Katsunobu Itou†† Tetsunori Kobayashi†

†Dept. EECE, Waseda University
††National Institute of Advanced Industrial Science Technology (AIST)

E-mail: †{omoto, koba}@tk.elec.waseda.ac.jp, ††{m.goto@, itou@ni.}aist.go.jp

Abstract This paper describes a new speech-input interface function called *speech shift*. While current speech-input interfaces have used only verbal infomation, we aim to build a more user-friendly speech interface by making use of nonverbal infomation. Our speech shift function enables a user to enter the same word in different meanings (functions) without changing a speech-input mode explicitly, using intentional F0 (fundamental frequency, perceived as pitch) control. For example, our speech shift interface can regard an utterance with a normal (low) F0 as a regular-mode input and an utterance with a high F0 as a command-mode input (such as word delete and file save commands). In our experience with a speech text editor with the speech shift function, the effectiveness of the speech shift was confirmed.

1 はじめに

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付けられてきた．そのため，韻律的特徴が，様々な付加的情報を伝達していることは知られてはいるものの，言語情報の円滑な伝達を支えるための補助的役割しか与えてこなかった．しかし，いままで以上に使いやすい音声インタフェースを構築するには，韻律的特徴を積極的に利用することが重要である．また，これまでとは異なる視点で韻律を捉えることにより，今までできなかった，もしくは，気付かなかった機能を，音声によって実現できる可能性がある．本稿では，非言語情報である韻律的特徴に着目することで，音声の持つ潜在能力を引き出した，新たな音声インタフェースの可能性について検討する．

我々はこれまでに音声補完[1, 2, 3]において，非言語情報にインタフェース機能を割り当てる研究アプローチを提案してきた．音声補完では，非言語情報である有声休止（母音が引き延ばされる言い淀み現象）を，補完トリガーキー（特殊キーの "Tab キー"）に位置付けることで，音声入力中に自然に補完機能を呼び出すことを可能にした．例えば，ユーザが「音声補完」という単語を最後まで思い出せないときは，断片的に「おんせいー」と言い淀むことにより，システムが「音声補完？」のように補完した候補を提示してくれる，効果的なインタフェースを構築することができた．

本稿では，非言語情報の一つである音高を利用し，声の高さで音声認識時の入力モードを切替える「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，音声のみでシームレスにモード切替えを実現できる．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，コンテキストに頼ることなく，異なる意味で扱うことで解決することができる．

以下，2章で「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．3章では，話者固有の音高の基準の推定法について述べ，4章で2つの音声シフト識別手法について述べる．5章では，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フト識別手法について説明する．6章では実装手法を述べ，7章で提案した識別手法の性能を実験的に評価する．最後に8章でまとめを述べる．

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「モードの切替え」と捉えることにより，特殊キーの “Shift キー” の機能を実現する [4]．従来の音声認識では，音声の持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．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ができる．このように音声に新たな役割を担わせることにより，より使いやすい音声インタフェースを構築する．

1章で述べたように，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」と思った場合，それらを区別するためには，通常は「文字入力モード」,「コマンドモード」といったモード切替えが必要となる．この切替えを実現するための方法は様々なものが考えられる．以下に例をあげると，第一に，キーボードやマウスなど，他の入力装置と併用して切替える方法がある．これはモードの切替え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（MS-IME の「Alt+半角」で「かな入力」と「英数字入力」を切替えられるように），マウスでモード切替えのボタン等を操作することによって実現する方法である．第二には，キーワード（予約語）を用いた切替え方法である．例えば，まず「コマンドモード」と発声しシステムのモードを切替えた後に，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場合や,「コマンドモードで保存」のように，キーワードを語頭に付けて発声することで実現する方法である．これらの手法は，一般的によく使われているが，操作が煩雑になる場合が多い．また，後者では，キーワードそのものを文字入力することが困難となってしまう．これに対し，音声シフトでは，声の高さによってモードの切替えがスムーズに実現できる．

音声シフトを用いることで以下の利点が得られる．

- **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の呼び出しができる．このため，操作手順が簡略化でき，操作性が向上する．
- **明示的なモード切替えが不要** 従来の音声インタフェース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．このため，操作時間の短縮につながる．

3 話者固有の音高の基準を表す基準基本周波数

音声シフトを実現するには，各発話区間が，通常発声とシフト発声のどちらであるかを識別する必要がある．ここで，通常発声とは普通の高さで発声すること，シフト発声とは故意に高く発声することである．

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，F0) 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，話者の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，難しい問題といえる．

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，単に F0 という絶対的な尺度で捉えるのではなく，話者が自分自身にとって相対的にどれくらい高く話しているかという尺度で捉えることができるからである．また，この音高の基準とシフト発声の音高との相対的な関係が話者によらず一定なら，話者共通なモデルを作ることも可能となる．

そこで，4章で述べる音声シフト実現手法に用いるための，話者固有の音高の基準となる基準基本周波数 (以下，**基準 F0**) を新たに導入し，有声休止区間を用いた基準 F0 の推定法について述べる．

3.1 基準 F0 の導入

話者固有の音高の基準を表すための基準基本周波数 (基準 F0) 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 [8] では，基底基本周波数として基準 F0 に相当する考え方が導入されているが，話者の基準 F0 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区間を用いることが望ましい．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 F0 の平均を，話者固有の基準 F0 とみなすことで推定する手法を提案する．有声休止は言い淀み現象の一つで，その発声

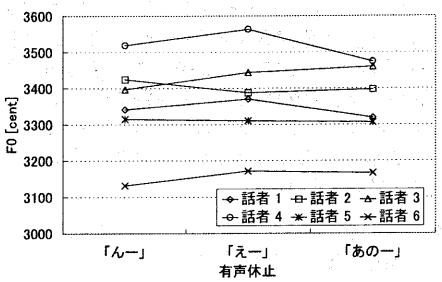

図 1: 話者ごとの有声休止の種類による F0 の平均

中は思考中のために調音器官の変化が小さくなるため，F0 が安定し [5, 6, 7]，かつ，地声の F0 (すなわち，基準 F0) 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，適切な手がかりと言える．さらに音声入力中には頻繁に現れるため，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 F0 の精度を高めることができる．本研究では，基準 F0 の更新を MAP 推定 (最大事後確率推定) により実現する．なお，上記で必要となる F0 推定と有声休止検出には，文献 [5, 6, 7] で後藤らが提案した F0 推定手法，有声休止区間の検出手法を用いる．

3.2 有声休止区間の F0 の分析

予備実験として，有声休止区間の F0 がどの程度安定しているかを調べた．男性話者 6 人毎の有声休止区間の F0 の標準偏差は，平均 86.2[cent][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 (「えー」「んー」「あのー」) による平均 F0 を示した．図から，有声休止の種類が異なっても平均 F0 はほぼ同一であるが，話者が異なると平均 F0 が大きく異なることから，話者ごとに基準 F0 を求める必要性があることがわかる．以上から有声休止区間の F0 はほぼ一定で安定しており，話者の基準 F0 として利用可能と判断した．

4 音声シフト実現方法

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 F0 を引いた値 (音高の平均の基準 F0 からの相対的な高さ) を用いる．以下，この値を発声音高と呼ぶ．ここでは，通常発声とシフト発声を発声音高に基づいて識別する 2 つの手法を提案する．

- **手法 1: 識別率を最大化する音声シフト識別手法**

  通常発声とシフト発声の二つのカテゴリの発声音高の境界を，明示的に求める手法である．この境界は，学習データに対して，二つのカテゴリの識別率を最大にするように求める．

- **手法 2: 確率分布モデルによる音声シフト識別手法**

  通常発声とシフト発声のそれぞれの発声音高の分布を，確率分布によって表す手法である．5.3 節で述べるように，他の確率等と組み合わせることができる．

4.1 手法 1: 識別率を最大化する音声シフト識別手法

本手法では，各発話の発声音高が，事前に定められたある閾値 (境界) 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．この閾値は，基準 F0 からの相対的な高さで表現され，カテゴリが既知の学習データに対して，通常発声とシフト発声の二つのカテゴリの識別率を最大にするように定める．

4.2 手法 2: 確率分布モデルによる音声シフト識別手法

本手法では，通常発声とシフト発声のそれぞれの発声音高の分布を事前に求め，各発話の発声音高がいずれの分布に属するのが尤もらしいかに基づいて識別する．発声音高の分布は，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する．以下，このモデルを発声音高モデルと呼ぶ．二つの正規分布の平均と分散は，カテゴリが既知の学習データの発声音高から最尤推定する．識別時には，両カテゴリの分布に対して，識別対象の発声音高の尤度が高いカテゴリを求める．

5 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持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，シフト発声される箇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 4 章の手法 1，手法 2 では考慮されていない，言語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共に，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．

5.1 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 章で述べたように，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替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」の機能呼び出しに割り当てた．用意した後者の機能には，編集機能 (「Back space」「Delete」「ポールド」「右

[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．本稿では，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．

```
silB 改行 したい 場合 silE
silB 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) を表す単語
<XXX>: シフト発声により XXX コマンドが実行された単語

図 2: 音声テキストエディタ利用時の書き起こし例

寄せ」「センタリング」「改行」等), ファイル操作機能 (「保存」「開く」等), 文字入力対象切替え機能 (「ひらがな」「カタカナ」「アルファベット」の切替え等) 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, 効率よく文章入力できる.

また, 一般的な音声エディタでは, ユーザが言い淀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 F0 を推定するために, ユーザに言い淀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 F0 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.

5.2 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 2 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単語となっている. また, "<",">" 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.

この例を人間が見て判断すれば, コンテキスト (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, あるコンテキストでは, 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多い.

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と考えられる. 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, 4 章の手法 1 や手法 2 では, シフト発声と誤識別されてしまうことがある. 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いことがわかれば, 適切に通常発声と判断できることが期待できる.

5.3 手法 3: 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 (10ms シフト) 毎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.

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

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. ここでさらに, 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. $\bar{a}_k$ は, 単語 c_k の区間における平均音高と基準 F0 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, 音高の平均をとる区間が単語であることを除いて, 4.2 節で述べた発声音高モデルに相当する. $P(C|W)$ 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 5.3.1 節に述べる. $P(W|X)$ 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の結合するように, ここでも発声音高モデル, コマンド正規モデルなどは重みを介して結合することとする.

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

式 (8) 中の $\frac{1}{K-2}$ 乗は, 単語数での正規化を意味する. また, 以下の実験では, 重みを $\alpha + \beta = 1, \gamma = 1$ と拘束することにする. 以上の確率によってシフト発声の識別を行う方法を**(手法 3)** とする.

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, 全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，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．

なお，今回の実験では，一発話内のデータは全てコマンドか全て非コマンドかに限った．このため，式(7)は，さらに発声音高の生起確率で近似した．

5.3.1 コマンド生起モデル

本節では，単語列からコマンド指標の生起確率を与えるコマンド生起モデルの構成法について述べる．

ここで一般にコマンド生起モデルの学習データを数多く集めることは困難であり，単純に単語列とコマンド指標列との関係をモデル化することは難しい．そこで，単語をいくつかのクラスに分類し，単語クラスの3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た上で，次の近似式を導入した．

$$P(C|W) \cong \prod_{k=2}^{K-1} P(c_i|w_{i-1}, w_i, w_{i+1}) \quad (9)$$

$$\cong \prod_{k=2}^{K-1} P(c_i|v_{i-1}, v_i, v_{i+1}) \quad (10)$$

ここで，v_i は単語 w_i が属する単語クラスである．本実験では，クラス v_i としては，w_i が silB, silE, sp のときの S，コマンドとして使われない単語のときの U，コマンドとして使われる単語のときの C，の3種類を採用した．

6 実装

以上述べた音声シフト識別手法を元に，5章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3に示す．本システムを構成する図3の7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よ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) [1, 2, 3] を用いた．

音声認識部には，CSRCツールキットのうち，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 [9]．手法3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，発声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，4.1節の手法1，4.2節の手法2，5.3節の手法3そ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F0の更新もここで行なう．そして，識別結果，および，算出された

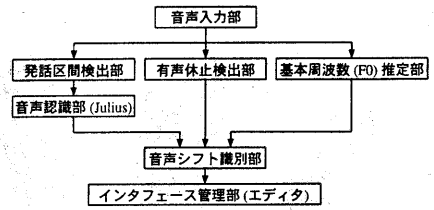

図 3: 全体の処理の流れ

基準F0，発話区間の平均F0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」として捉え，音声認識結果に対応するコマンドを実行する．

7 音声シフト識別実験

提案した3つの識別手法の性能を評価するため，音声シフト識別実験を行う．

7.1 実験条件

様々な長さの単語を用いて通常発声とシフト発声の識別を行う．

(a) 実験データ

被験者は，男性話者5人である．短い単語 (15語)，複合語 (15語)，長いフレーズ (15語)，コマンド (15語) の計60語を1セットとし，それを通常発声とシフト発声の両方で発声したものを収録した．音声データは，DATに 48kHz，16bitで録音したものを，16kHzにダウンサンプリングして用いた．

(b) 実験方法

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，コマンド以外の単語をシフト発声したものは除外して評価する．そのような発声は，本エディタの通常の使用範囲ではありえないからである．つまり，評価用のデータは，識別時のデータは，通常発声の単語全て (60語) と，シフト発声のコマンド (15語) の計75語となる．また，単語の種類ごとに5語ずつ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通り全ての組合せで行い，その識別率の平均値を評価結果とした．

7.2 検討項目

4.1節の手法1，4.2節の手法2，5.3節の手法3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境界の閾値を1[cent]ずつ変化させ，学習データに対する識別率を最大にする閾値を求めた．その際，話者ごとに別々に閾値を設定する場合 (1-a) と，全話者共通の閾値を設定する場合 (1-b) の

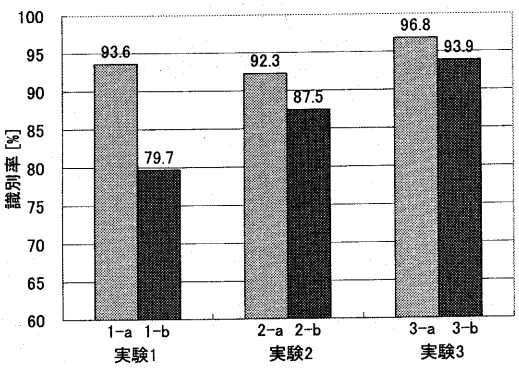

図 4: 実験結果

二つの条件を用意し，比較実験した．

実験 2: (手法 2) 学習データから，二つのカテゴリのそれぞれの発声音高モデルを求めた．その際，話者ごとにモデルを求める場合 (2-a) と，全話者共通のモデルを求める場合 (2-b) の二つの条件を用意し，比較実験した．

実験 3: (手法 3) 発声音高モデルには，実験 2 と同じモデルを使用した．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を用いた．また，重み α は，0.0 から 1.0 まで 0.1 間隔で変えて実験し，実験 2 同様，話者ごとのモデル (3-a) と，全話者共通のモデル (3-b) を使用する二つの条件で比較実験した．

7.3 実験結果

各実験結果を図 4 に示す．どの手法においても，話者ごとに最適化したモデルを用いた場合の方が高い識別率となった．手法 1 と手法 2 を比較すると，話者ごとのモデルを用いた場合には，両者に大きな差は見られないが，話者共通のモデルを用いた場合には，手法 2 の方が識別率が高いことがわかる．これは，本実験のようなオープンな実験では，手法 2 の方がロバストに識別できたことを示している．一方，手法 2 と手法 3 を比較すると，話者ごとのモデルと話者共通のモデルのいずれの場合も，手法 3 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わかる．

実験 3 において，重みを変化させたときの識別率の変化を図 5 に示す．話者ごとのモデルと話者共通のモデルのいずれの場合も，発声音高モデルの重みが高い方が識別率が良い傾向にある．これは，発声音高モデルの重みが低いと，コマンドのシフト発声を通常発声と誤識別しやすくなるためである．このように言語的な情報を多少考慮するだけでも，特に話者共通モデルの場合に，システムの性能が大きく向上したことがわかる．

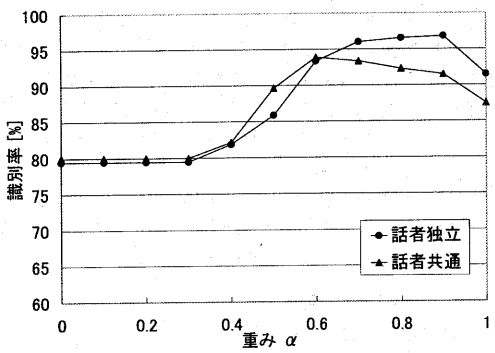

図 5: 重みを変化させたときの識別率の変化

8 おわりに

本稿では，音高を利用して，通常の発話時と故意に高く発話した時のモードの切替えを実現した，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 F0 を用いた，話者ごとの基準 F0 の推定法，および，それに基づくシフト発声の識別手法を提案した．さらに，音声シフトを音声テキストエディタに応用すると共に，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戸惑うユーザもいたが，慣れてくると有用な機能として問題なく利用することができた．

「音声補完」「音声シフト」の一連の研究は，キーボードと対比するならば，Tab キーや Shift キーなど，特殊キーの役割を実現したものと捉えることができる．文献 [2] でも述べたように，音声による様々な特殊キー機能の実現によって，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，今後幅広く検討を行っていきたい．

参考文献

[1] 後藤 真孝, 伊藤 克亘, 速水 悟: 音声補完: "TAB" on Speech, 情処研報, 2000-SLP-32-16, pp.81-86, 2000.

[2] 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, WISS2000, 近代科学社, pp.153-162, 2000.

[3] 後藤 真孝, 伊藤 克亘, 速水 悟: 音声補完の評価, 情処研報 2002-SLP-40-4, 2002

[4] 尾本 幸宏, 後藤 真孝, 伊藤 克亘, 小林 哲則: 音声シフト:音高を利用した新たな音声入力インタフェース, WISS2001, 近代科学社, pp.17-26, 2000.

[5] 後藤 真孝, 伊藤 克亘, 速水 悟: 自然発話中の言い淀み箇所のリアルタイム検出システム, 情処研報, 99-SLP-27-2, pp.9-16, 1999.

[6] Goto, M., Itou, K., Hayamizu, S.: A Real-time Filled Pause Detection System for Spontaneous Speech Recognition, *Proc. of Eurospeech '99*, pp.227-230, 1999.

[7] 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, VOL.J83-D-II NO.11, pp.2330-2340, 2000.

[8] Fujisaki, H., Hirose, K.: Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *J. Acoust. Soc. Jpn. (E) 5*, pp.233-242, 1984.

[9] 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, オーム社, 2001.