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

後藤 真孝　伊藤 克亘　秋葉 友良　速水 悟

本論文では，ユーザがある単語を思い出せずに断片だけを発声しても，システム側がその残りを補って入力することを可能にする「音声補完」という新しい音声インタフェース機能を提案する．既にテキストインタフェースでは，ファイル名の入力等で補完の概念が広く受け入れられているが，音声では効果的な補完機能は提案されていなかった．我々は，ユーザが単語発声途中に有声休止 (母音の引き延ばし) によって言い淀むと，それを含む補完候補の一覧を見ることができるインタフェースを構築し，労力をかけずに自発的に補完機能を呼び出しながら音声入力することを可能にする．実際に，有声休止検出機能と補完候補作成可能な音声認識機能を備えたシステムを実装して運用し，音声補完の有用性を確認した．本研究での有声休止は，従来の言語情報中心の音声入力インタフェースに導入された，新たな非言語情報のモダリティと捉えることができる．

1 はじめに

現在の音声入力インタフェースは，音声の持つ潜在能力を引き出していない．音声は，音韻や単語のような**言語情報**だけでなく，韻律や言い淀みのような**非言語情報**も含んでいるが，これまでの音声認識は，主に言語情報のモダリティしか利用していなかった．そのため，いわば認識誤りを起こすようなキーボードにしか過ぎず，いくら音声認識率を高くしたとしても，キーボードを越えるような使いやすいインタフェースは構築困難であった．音声ならではのメリットを引き出すためには，音声の持つ非言語情報が，人間同士のコミュニケーションでどのような役割を果たしているのかを問い直し，その役割を積極的に活用したインタフェースを構築する必要がある．

そこで本研究では，非言語情報の中でも特に話者の思考状態が現れやすい**有声休止** (filled pause) に着目する．有声休止は言い淀み現象の一つであり [12] [13] [14]，発話したい内容が断片的にしか思い出せないときや，何を発話していいのか判断に迷うときに，発声されることが多い．音響的には持続した有声音 (母音の引き延ばし) として現れ，例えば，話者が「音声補完」という単語を最後まで思い出せないときには，「おんせいー」と言い淀んだりする (「いー」が有声休止)．このとき，対話相手はしばしば話者の言いたいことを推測し，「音声補完？」のように候補を提示することで，話者が思い出すのを手助けしてくれる (場合によっては話者は最初から対話相手の助けを期待して有声休止を用いたりする)．このように，本来音声を使う場合には，いい加減で断片的な情報を伝えても，対話相手が様々な形で自分の発話や思考の手助けをしてくれることが期待でき，それが快適で優れた情報交換手段となっている一つの理由であると考える．これは，従来の音声認識には欠けていた視点である†1．

Speech Completion: Introducing New Modality into Speech Input Interface

Masataka Goto, Katunobu Itou, Tomoyosi Akiba, Satoru Hayamizu, 独立行政法人 産業技術総合研究所 [旧電子技術総合研究所], National Institute of Advanced Industrial Science and Technology (AIST) [former Electrotechnical Laboratory].

コンピュータソフトウェア, Vol.19, No.4(2002), pp.10–21.
[論文] 2001年4月30日受付.

†1 従来の音声認識では，話者は入力したいすべての音を丁

上記で例示した，対話相手による「音声補完？」という手助けは，発話された単語の断片の残りを補うことで，話者が述べようとしている単語全体の候補を提示している，つまり，単語を**補完**していると見なすことができる．この**補完** (completion) の概念は，テキストインタフェースでは既に広く受け入れられている．例えば，tcsh や bash などの UNIX シェルや，Emacs/Mule などのテキストエディタは，ファイル名やコマンド名の補完機能を提供している†2．こうした補完機能では，ユーザが補完機能を呼び出すキー (以下，**補完トリガーキー**と呼ぶ) を押したときに，途中までタイプされた単語の断片の続きが補われる．また，近年ペン入力でも，補完機能を持ったインタフェース [2] [9] [11]が提案されている．しかし，音声入力インタフェースでは，音声入力中に自然に補完機能を呼び出す手段がなかったこともあって，効果的な補完機能はこれまで提案されていなかった．

本研究は，このような補完による手助けという概念を音声入力の枠組みに導入することで，音声認識を中心とした音声インタフェースをより使いやすくすることを目的とする．本論文では，以下，2 章で「**音声補完** (speech completion)」という新しい音声インタフェース機能を提唱し，3 章でユーザが音声入力中に言い淀む (有声休止をおこなう) と計算機が補完候補を提示して手助けをしてくれるインタフェースを提案する．次に，4 章で具体的な実現手法について説明し，5 章でシステムの実装とその動作結果を示す．そして，6 章で音声補完の有効性を確認する評価実験の結果を述べる．最後に，7 章で関連研究やマルチモーダルインタフェースに関する議論をおこない，8 章でまとめを述べる．

寧に発声することが強いられていた．人間が自発的に話す場合には，有声休止や言い直し等の様々な言い淀み現象が自然に現れるが，そうした現象は誤認識の原因となるため，話者は書き言葉を読み上げたような発声をしなければならなかった．また，言い淀み現象がコミュニケーションにおいて持つ役割を積極的に活用しようとする研究も少なかった．

†2 特に，コマンド行の補完機能は，1970 年代初頭には既に複数のシステムに導入されていた [7].

2 音声補完

「音声補完」とは，音声入力中に，ユーザが補完機能を呼び出すことができるようにするための新たな音声インタフェース機能の総称である．ユーザが発話した断片をシステム側が補完してくれることで，以下のような利点が得られる．

- **記憶補助** 入力したい内容がうろ覚えでも，一部だけ思い出して発声すれば入力できる．
- **省力化** 入力内容が長くて複雑なときに，内容の特定に十分な部分まで発声すれば入力できる．
- **心理的抵抗の低減** 従来の音声インタフェースの多くが，すべての音を最後まで丁寧に発声することを強いていたのに対し，音声補完では思いついた断片だけを発声すればよく，心理的抵抗が少なく使いやすい．

音声補完を通常の音声入力に効果的に導入するには，ユーザが候補を見たいと思うタイミングで，自発的に補完機能を呼び出せることが重要となる．なぜなら，ユーザの望まないタイミングで次々と補完候補を出すような自動補完は，認識時に曖昧性が大きい音声では，煩わしく不適切な機能となりやすいからである (自動補完に関するより詳しい議論は 7.1 節でおこなう)．そこで，言い淀み現象の一つである**有声休止**に補完トリガーキーの役割を担わせることを提案する．補完トリガーキーとして音声入力中に有声休止をおこなうことは，1 章で述べたように人間にとって極めて自然であり，ユーザは自分の意志で，労力をかけずに補完機能を呼び出すことが可能になる．

音声補完の対象には，単語や文節，文章など様々なレベルが考えられるが，本論文では以下，単語のみを取り上げて議論する．つまり，**単語補完機能**に論点を絞る．ただし，ここでの単語は，音声認識システムの単語辞書上 (言語モデル上) の 1 単語とする．したがって，例えば姓名が一つの単語として登録されているときには，姓だけのような部分的な発声から残りが補完される．

本論文では，ユーザが単語のどの部位を発声して入力したいか (補完したい方向) に応じて，以下の二種類の音声補完方式を提案する．ここでは，「宇多田ヒカル」という一単語を補完する場合を例に説明する．

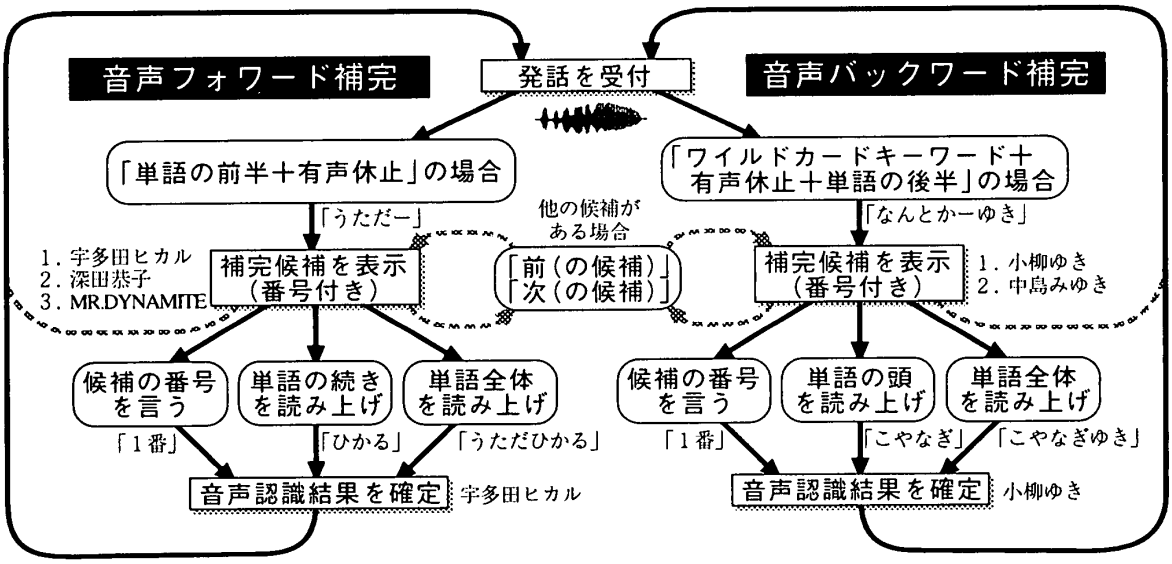


図1 音声補完の操作の流れ

1. **音声フォワード補完**

   単語の前半 (頭) がわかっているときに，その最後の音節で有声休止をおこなうことで，それに続く後半を補完する方式である．例えば，「うただー」と「だ」の音で有声休止をおこなって入力すると，「宇多田ヒカル」が補完候補の一つとして得られる．仮に「宇多田」という単語も単語辞書に登録されているときには，ユーザが，単に「宇多田」と入力したいのか，それとも補完候補を得たいのかをシステムが識別する上でも，有声休止を補完トリガーキーとして用いることが重要となる．

2. **音声バックワード補完** (別名，**音声ワイルドカード補完**)

   単語の後半 (末尾) がわかっているときに，事前に定めたキーワード (以下，**ワイルドカードキーワード**と呼ぶ) を言いながらその最後の音節で有声休止をおこない，続いて後半を発話することで，その前につながる前半を補完する方式である．例えば，「なんとか」をキーワードと定めた場合，「**なんとかーひかる**」と入力すると，「宇多田ヒカル」が補完候補の一つとして得られる．この「**なんとかー**」は，任意の文字列にマッチするワイルドカードに相当するため，これを音声ワイルドカード補完とも呼ぶ．入力したい単語辞書の中に，キーワードを部分文字列として含むような単語が仮にあったとしても，有声休止によってキーワードは識別可能であり，意図した箇所でのみ音声バックワード補完を呼び出すことができる．

3 音声補完機能付き音声入力インタフェース

構築した音声補完機能付き音声入力インタフェースの機能を説明する．ユーザは，以下のように有声休止を用いて音声補完しながら，単語を入力することができる (図 1)．

1. **[音声フォワード補完の場合]**

   単語の発声途中で母音を引き延ばすと，発声された断片から始まる補完候補 (単語) の一覧が，番号付きで即座に表示される．(ex.「うただー」と入力すると，「1. 宇多田ヒカル, 2. 深田恭子, 3. MR.DYNAMITE」のように補完候補が表示される†3.)

   [音声バックワード補完 (音声ワイルドカード補完) の場合]

   ワイルドカードキーワードの最後の母音を引き延ばし，続いて単語の後半を発声すると，その発声された断片が末尾に付く補完候補の一覧が，同様に番号付きで表示される．(ex.「**なんとかー**ゆき」と入力

†3 テキストの補完とは異なり，たとえ「うただ」から始まる単語が辞書中に一つしかなくても，音声の曖昧性から候補を一つに絞り込めないことが多い．上記の例では，/utadahikaru/, /fukadakyouko/, /misuta-dainamaito/ が音響的な類似度から補完候補として得られた．

すると，「1. 小柳ゆき, 2. 中島みゆき」のように補完候補が表示される.)

2. 候補が多くて画面に入りきらないときには，「前の候補」「次の候補」というマークが表示される．その場合，「前 (の候補)」や「次 (の候補)」と言えば他候補が見られる．候補が不適切なときや別の単語を入力したいときには，次の3.の選択をせずに別の発話に移ってもよい．
3. ユーザは 3 通りの方法で補完候補を選択できる．
   (a) 候補の番号を言う．(ex.「1 番」か「1」と言う.)
   (b) 単語の続きや単語の頭を読み上げる．(ex.「ひかる」，「こやなぎ」と言う.)
   (c) 単語全体を頭から読み上げる．(ex.「うただひかる」，「こやなぎゆき」と言う.)

   選択すると，その候補は強調表示され，音声認識結果として確定される．

音声補完は，一つの単語を入力中に，繰り返し呼び出すことが可能である．例えば，「サザンオールスターズ」を入力するときに，「さざんー」で候補一覧を見た後，「おーるー」でさらに絞り込まれた候補一覧を見て，最後に「すたーず」と言って確定できる．この例に示したように，単語中の長母音 (「おーる」の /o-/) では音声補完が呼び出されず，意図的に有声休止した箇所でのみ呼び出されるようにする必要がある．

なお，提案した二つの補完方式では，単語の頭か末尾がわかっていないと入力できないため，中央部分しかわからない場合には直接適用できない．その場合には，二つの補完方式を応用し，まず既知の中央部分までを音声バックワード補完で入力しつつ，その中央部分の最後の音節で有声休止をして音声フォワード補完を呼び出すことで，補完入力が可能となる (これを「音声ツーウェイ補完」と呼ぶ)．例えば，“Blankey jet city” を “jet” に基づいて補完したいときに「なんとかー じぇっとー」と入力する．その逆に，単語の頭と末尾が共にわかる場合 (単語の中央部分だけがわからない場合) には，頭か末尾のいずれか一方を用いて補完すればよい．

4 実現方法

3 章のインタフェースを構築するには，有声休止の区間を得る**有声休止検出部**と，単語の途中までの発声やキーワードに続く発声を認識して補完候補を作成する**音声認識部**を実現しなければならない．さらに，インタフェース全体の状態管理をする**インタフェース管理部**と，補完候補一覧や認識結果を提示する**画面表示部**も必要となる．以下，これらを順に説明する．

4.1 有声休止検出部

音声補完では，有声休止を高い精度でリアルタイムに検出することが重要である．しかも，任意の単語中の母音の引き延ばしを検出する必要があるため，トップダウン情報を使わない言語非依存な検出をしなければならない．そのような要件を満たす検出手法として，我々が文献 [3] [4] [6]で提案した，有声休止箇所のリアルタイム検出手法を用いる．本手法は，人手で正解を付与した自由発話音声の対話コーパスに対して，再現率 (正しく検出した有声休止数 / 正解の有声休止の総数) 0.75，適合率 (正しく検出した有声休止数 / 有声休止として検出した数) 0.70 の性能を持つ [4]．再現率と適合率のトレードオフは調整でき，今回の目的に合うように適合率を高く (誤検出を少なく) 設定することも可能である．例えば，適合率 0.93 となるように設定した場合，400 ms 以上の十分長い継続時間を持つ有声休止を検出すべき対象として評価すれば，再現率 0.91 の性能が得られた [4].

本手法は，有声休止が持つ二つの音響的特徴 (基本周波数の変動が小さい，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．検出結果の例を図 2に示す．以下では，音響的特徴の推定方法を示した後に，有声休止開始点を決定する方法を述べる．この結果は，次の音声認識部へと送信される．

4.1.1 基本周波数の推定

入力信号中で最も優勢な高調波構造の基本周波数を，音声の基本周波数として推定する．そのために，時刻 t において[†4]周波数 F が基本周波数となる可能

†4 現在の実装では，16 kHz / 16 bit で A/D 変換し，フレームシフト 10 msec (160 点) をすべての処理の時間単位とする．

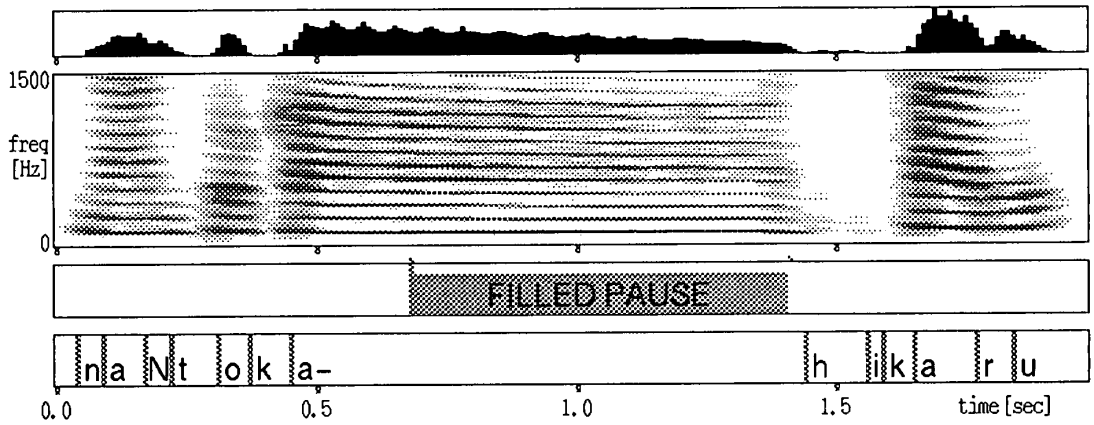


図 2 「なんとかーひかる」 /nantoka-hikaru/ に対する有声休止区間の検出例: パワーとスペクトルの時間変化 (上二段) とシステムが検出した有声休止区間 (中段), 有声休止区間での音素遷移を抑制した音声認識システムが出力した音素系列 (下段)

性 $P_{F0}(F,t) = \int_{-\infty}^{\infty} p(x;F)\ \Psi_p(x,t)\ dx$ を評価する. $p(x;F)$ は基本周波数が F の高調波成分を通過させるフィルタ関数, $\Psi_p(x,t)$ は周波数成分のパワー分布関数である. $P_{F0}(F,t)$ は各高調波構造が相対的にどれくらい優勢かを表すため, 基本周波数 $F_{F0}(t)$ は $F_{F0}(t) = \mathrm{argmax}_F\ P_{F0}(F,t)$ で求まる.

4.1.2 スペクトル包絡の推定

実環境でロバストに包絡を推定するために, $F_{F0}(t)$ の高調波構造上にある局所的な情報だけを利用する. $F_{F0}(t)$ の整数倍の周波数を中心とするガウス分布で重み付けしながら, その近傍の最大パワーを検出することで, 各高調波成分のパワーを求める. 次に, 隣接する成分のパワーの間を直線補間してスペクトル包絡を求める. 有声休止を検出するためには, 包絡の大局的な変形を捉えた方が良いため, 包絡を粗い周波数分解能でリサンプリングし, 低い方から n $(1 \leq n \leq \mathrm{N_{max}})$ 点目の周波数におけるスペクトル包絡 $Env(n,t)$ を求める. 最後に, 呼気による AM 変調の影響を除去するために $Env(n,t)$ を正規化する.

4.1.3 有声休止開始点の決定

有声休止を検出するための二つの特徴量として, 基本周波数の変動量 $A_f(t)$ とスペクトル包絡の変形量 $A_s(t)$ を求める. これらは, $F_{F0}(t)$ と $Env(n,t)$ の過去一定期間の対数スケール上での変化を, 最小自乗法で直線近似した直線の傾き $b_f(n)$, $b_s(n)$ と近似誤差 $err_f(n)$, $err_s(n)$ を用いて, $A_f(t) = |b_{F0}|$, $A_s(t) = \left(\frac{1}{\mathrm{N_{max}}}\sum_{n=1}^{\mathrm{N_{max}}} b_s(n)^2\right)\left(\frac{1}{\mathrm{N_{max}}}\sum_{n=1}^{\mathrm{N_{max}}} err_s(n)^2\right)$ のように定義される. そして, **有声休止らしさ** (有声休止と判定する信頼度) $P_{fp}(t)$ を, $A_f(t)$, $A_s(t)$ の短時間平均 $S_f(t)$, $S_s(t)$ に基づいて, $P_{fp}(t) = \exp\left(-\frac{(\mathrm{R}\ S_f(t) + (1-\mathrm{R})\ S_s(t))^2}{\mathrm{W}^2}\right)$ と定義する. R は特徴に対する重み付けを決める定数, W は変動・変形の考慮範囲を決める定数である.

最終的に, $P_{fp}(t)$ が十分高い値をとり続けるときに, 話者が有声休止をおこなっていると判定する. 具体的には, 条件 $P_{fp}(t) > e^{-1}$ を満たし続ける限り $P_{fp}(t)$ を累積加算して累積値を求め, それが一定の閾値より大きくなった時刻を有声休止開始点とする. 単語中の長母音は音高変化が大きいために通常は誤検出されないが, 仮に音高変化が小さい場合でも, この閾値を十分高く設定することで, 今回の目的に合った誤検出の少ない検出が実現できる.

4.2 音声認識部 (補完候補作成)

音声認識部では, 音声入力と有声休止検出部の結果を受信し, 音声認識結果 (尤度の高い順に上位 $\mathrm{N_{result}}$ 個) と音声補完候補をインタフェース管理部へと送信する. 補完候補一覧を作成する処理は, 不特定話者を対象とした連続音声認識システム niNja [8]を以下に述べるように拡張して実現する†5. その際, 有声休止を含まない通常の発話の認識には副作用のないようにする必要がある. 以下, 単語発声の補完を説明するが, 連続音声中の

†5 niNja は, 隠れマルコフモデルを用いた一般的な音声認識システムに準じた処理をしており, ここで新たに提案する補完候補の作成手法は, 他のシステムにも同様に適用可能である.

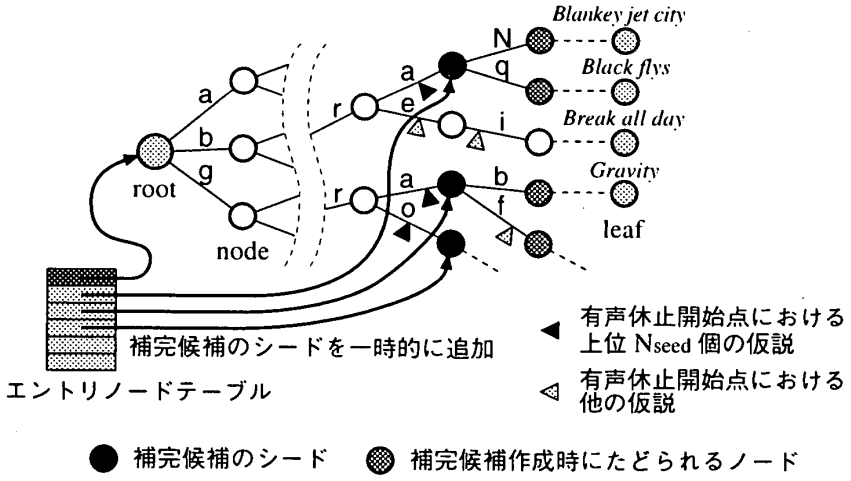


図 3 音声フォワード補完: 木構造の単語辞書における有声休止開始時点での仮説 (くさび形のマーク) と 音声補完候補の作成・エントリノードテーブルへの追加

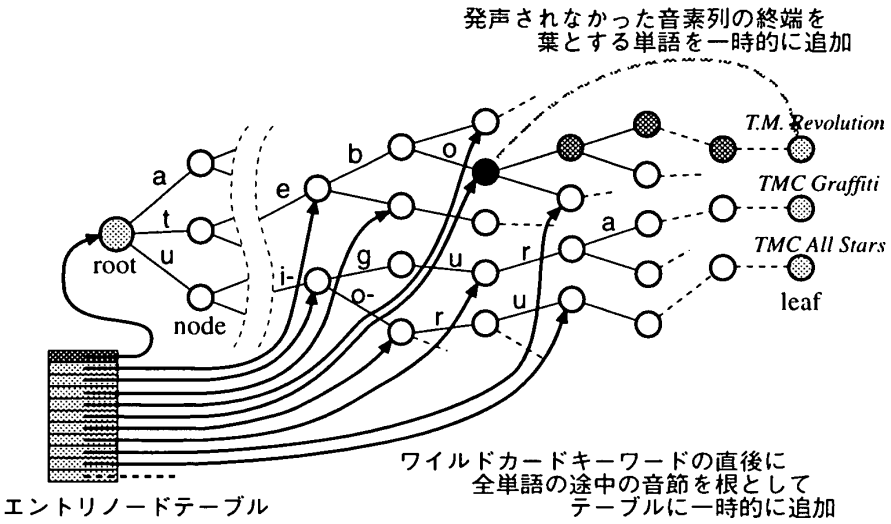


図 4 音声バックワード補完: ワイルドカードキーワードの直後だけ 全単語の途中の音節をエントリノードテーブルに一時的に追加

単語を補完することも同じ枠組で可能である.

本システムは単語辞書として, 入力対象の単語辞書 (人名等) 以外に, ワイルドカードキーワード辞書とインタフェース操作用語辞書 (候補番号や他候補の表示指示等) を使用する. 単語辞書は, 図 3のように木構造で保持される. この辞書を用いた認識処理では, 辞書の根から, フレーム同期で枝別れに応じて仮説を増やして, ノードを葉の方向へたどっていく†6. 図中のくさび形のマークが仮説をあらわす. 有声休止が検出されると, その時点で最も尤度の高い仮説がワイルドカードキーワードかどうかを判定し, 音声フォワード補完と音声バックワード補完のどちらを実行するかを決定する.

音声フォワード補完の場合, その時点で有効な仮説 (尤度の高い順に上位 N_{seed} 個) から葉の方向へたどることで, 補完候補の生成を実現する. それらを尤度の高い順に番号付けして, 上位 N_{choice} 個をインタフェース管理部へ送信する. 生成する際に用いた仮説に対応するノードを**補完候補のシード**と呼ぶ. 例えば, 図 3の一番上の黒い丸がシードであるとすると, 補完候補は "Blankey jet city" と "Black flys" になる. 同時に, そこまでに認識した音素列を求めることにより, 各候補

†6 現在の音声認識では, 音素単独での認識精度が不十分なので, 単語の音素列を一音素ずつ順に確定するのではなく, このように複数の仮説によって次に続く音素を予測しながら最終的に最も尤もらしい仮説を求める. ただし本研究では, 認識精度を高める工夫として, 有声休止区間での音素遷移を抑制している.

においてどこまで発声されたかを調べ，候補と併せて送信する．

ユーザが補完候補を見た後に，単語の続きを言っても選択できるように，認識を開始する根を登録する**エントリノードテーブル**を導入し，単語の途中からの認識を可能にする．通常の単語の頭からの認識では，このテーブルには辞書の根だけが登録されている．単語の途中から認識を開始したい場合には，図 3のように補完候補のシードを根として一時的に (有声休止を伴う発話の次の発話だけ) 追加する．追加エントリは，有声休止後の続きの音素列だけを言えば認識されるが，認識結果としては，その単語全体を送信する．

一方，音声バックワード補完の場合，有声休止終了時点以降に発声された単語の後半部分を認識し，補完候補を生成する必要がある．この単語の途中からの認識は，辞書中の全単語の途中の音節を，エントリノードテーブルに一時的に (ワイルドカードキーワードの直後だけ) 追加することで実現する (図 4)．そして葉に到達した仮説から尤度の高い順に番号付けして，上位 N_{choice} 個を送信する．その後，単語の頭を言っても選択できるようにするために，各候補で発声されなかった音素列の部分の終端を葉とする単語を一時的に登録する．例えば，「小柳ゆき」を「**なんとかー**ゆき」で入力した場合，/koyanagi/ の末尾を一時的に葉とする単語を追加する．

4.3 インタフェース管理部・画面表示部

補完候補の選択等のインタフェース全体としての機能を提供する．まず，有声休止を含まない発話の場合には，単に認識結果を受け取って表示する．一方，有声休止を含む発話の場合には，音声補完候補を受け取った時点でポップアップ式の補完候補ウィンドウを出現させ，その中に候補一覧を表示する (音素列上で既に発声された部分は違う色で表示する)．ただし，音声バックワード補完の場合には，ユーザが処理の進行状態を把握できるよう，ワイルドカードキーワードの部分を認識した直後に一旦そのキーワードを表示し，続く発話の認識後に候補一覧を表示する．そして，図 1の操作の流れに従いながら，次の発話の認識結果に応じて表示・選択・確定等の処理をおこなう．その際，常に 1 位の認識結果を用いるのではなく，選択操作に該当する結果が上位 $N_{priority}$ 個以内にあれば，それを優先させて用いる．これは，単語の続きを言って選択する場合等に，その発声に近い単語辞書上の別の単語の尤度が高くなり，適切に選択できない事態を回避するためである．

5 実装

以上述べてきた音声補完機能付き音声入力インタフェースのプロトタイプシステムを実装した．日本のポピュラー音楽のヒットチャート (2000 年度のすべての週間ランキングのシングル上位 20 曲) から，曲名 (342 語) とアーティスト名 (179 語) のデータベースを作成し，音声補完対象の単語辞書 (計 521 語) とした．本実装では，4.2 節の各定数を $N_{result} = 5$，$N_{choice} = 20$，$N_{seed} = 15$，$N_{priority} = 3$ と設定した．これらは単語辞書の内容や規模，インタフェースの用途等に応じて調整する必要がある．

効果的に負荷分散が可能で，拡張性が高くなるように，本システムを構成する図 5の 8 つの機能を，分散環境で動作する別々のプロセスとして実装した．そのために，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) を設計し，それに基づいて実装した．RVCP は，RMCP (Remote Music Control Protocol) [5]を音声言語情報の伝送用に拡張したプロトコルである．

本システムを運用したところ，提案したインタフェースが機能し，ユーザが音声補完機能を呼び出しながら，インタラクティブに単語入力することができた．音声補完中の画面表示例を図 6，図 7に示す．

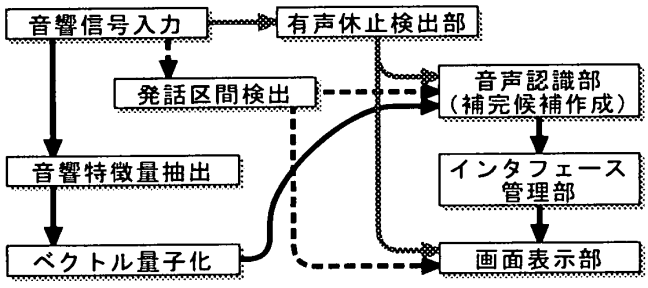


図 5 システムを構成する 8 つのプロセス

(1) 「うただー」と入力

(2) 「だー」の有声休止中

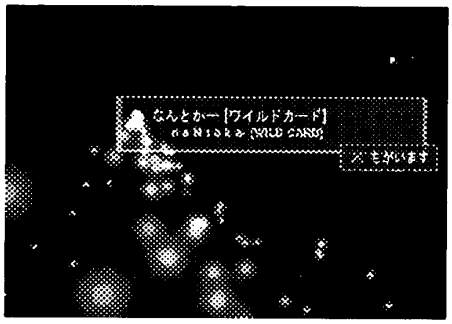
(3) 補完候補ウィンドウが回転しながら登場

(4) 「1番」と入力した直後

(5) 1番の候補が輝いて跳躍

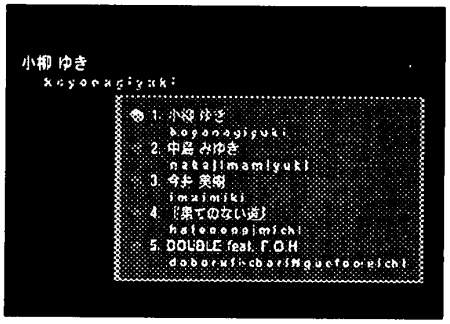
(6) 1番の候補「宇多田ヒカル」が認識結果として確定

図6 音声フォワード補完中の画面表示例



(a) 「なんとかー」と入力した直後

(b) 「ゆき」と入力した直後に補完候補ウィンドウが登場



(c) 「1番」と入力して「小柳ゆき」が認識結果として確定

図7 音声バックワード補完中の画面表示例

6 評価実験

音声補完の有効性を確認するために，5章のシステムを用いて，以下の点を調査する実験をおこなった．

- 音声補完の使用に慣れた後，音声補完を使用するか使用しないかを自由に選んで入力するときに，どのように入力するか．
- うろ覚えの単語を入力するときに，音声補完を使用するか使用しないか．
- 音声補完の使用後にどのような印象を受けたか．

実験には，20～50代の45名の被験者 (男性24名，女性21名) が参加した．

6.1 実験方法

被験者に，音声フォワード補完と音声バックワード補完の入力方法を説明した後[†7]，紙面に記載された単語を異なる条件で音声入力させた．被験者は，まず練習として，ある1単語 (「ボクの背中には羽根がある」) につい

†7 我々は，音声補完を初めて使用するユーザには，新しい音声インタフェース機能として具体的な入力方法を教示することを想定している．そのような教示をしない場合にユーザがどのように振る舞うかの調査は，今後の課題である．

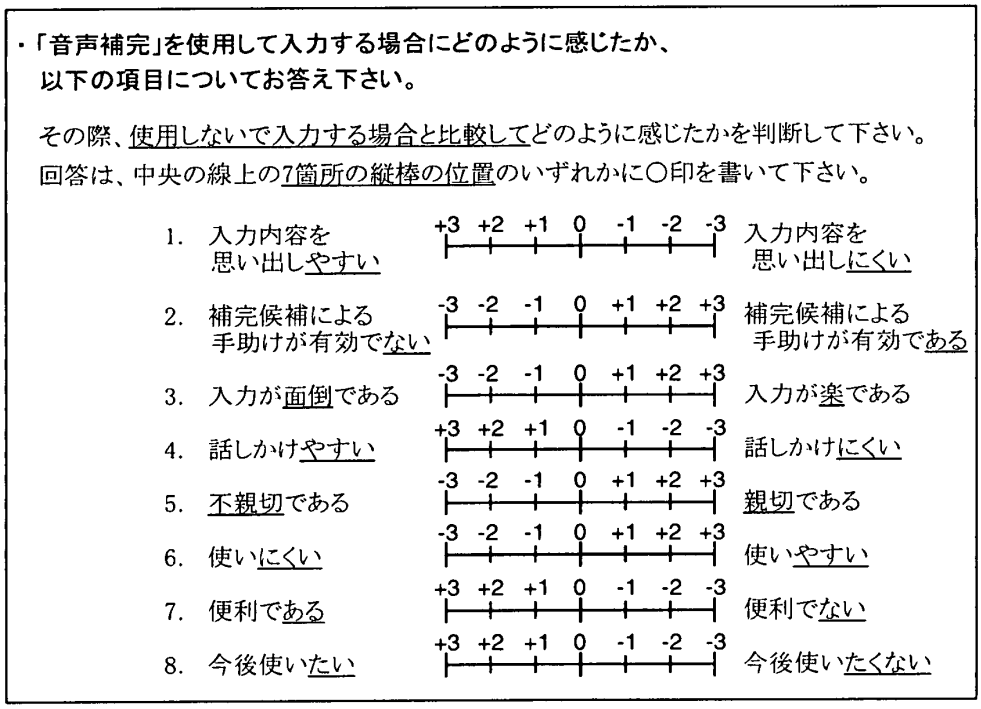


図8 被験者に対する音声補完使用後のアンケート (−3 ～ +3 の数字は実際の用紙には記載されていない)

て，「ぼくのー」や「なんとかー，はねがある」のように指示された通り読み上げて発声し，音声補完を体験した．次に，あらかじめ音声認識システムの単語辞書の中から決められた5単語 (曲名もしくはアーティスト名)

1. yaen front 4 men feat. saki
   (ヤエン フロント フォー メン フィーチャリング サキ)
2. 水・陸・そら，無限大
   (みず りく そら，むげんだい)
3. 恋はスリル，ショック，サスペンス
   (こいは スリル，ショック，サスペンス)
4. 神経がワレル暑い夜
   (しんけい が ワレル あつい よる)
5. 関東裸会 三羽烏
   (かんとう はだかかい さんばからす)

を1セットとして，そのセットを以下の3つの条件で入力した．

(1) 「音声補完」を使用しないで入力
(2) 「音声補完」を使用して入力
(3) 「音声補完」を使用するか使用しないかを自由に選んで入力

(2) の条件で，音声フォワード補完，音声バックワード補完のどちらを使用するか (単語中のどの箇所を発声するか) は，単語ごとに自由とした．5単語は全被験者を通じて共通だが，順番をほぼランダムに変えたものを5通り用意して被験者に均等に割り当てた．被験者は紙面に記載された単語セット (読みも記載) を見ながら，まず一番上の条件で5単語を入力し，次に一つ下の条件で同じ5単語を入力していく．ただし，(1) と (2) の条件の順番は，被験者ごとに交互に変わるようにした．

それから，単語セットの紙が取り去られ，被験者は何も単語を見ない状態で，再び同じ5単語をそれ以上思い出せなくなるまで入力した．その際，単語の入力順序は自由とした．これを入力条件 (4) とする．

(4) 何も単語を見ない状態で思い出しながら自由に入力

単語を意図的に記憶しないように，被験者は，入力条件 (1) ～ (3) の実験中には，後で思い出しながら入力する実験があることは知らされていない．

最後に，被験者にはアンケートとして，図8の8項目について7段階で評価させた．

6.2 実験結果

実験の結果，被験者全員がすべての入力を完了することができ，音声補完を使用するのに特別な訓練は不要であった．被験者一人が練習も含めて全条件で入力するのに平均で約7分間を要した．

入力条件 (3) で，音声補完を使用するか使用しないか

表 1 単語を見ながら入力する際に 音声補完を使用するか使用しないかを調査した結果

	使用	不使用
一人当りの平均回数	3.71 回	1.29 回
割合	74.2%	25.8%

表 2 うろ覚えの単語を思い出しながら入力する際に 音声補完を使用するか使用しないかを調査した結果

	使用	不使用
一人当りの平均回数	2.56 回	0.62 回
割合	80.4%	19.6%

を自由に選べるときに，どのように入力したかを調査した結果を表 1に示す．この結果から，入力条件 (1), (2) で音声入力に慣れた被験者が，74.2% の割合で音声補完を使用して単語を入力したことがわかる．そのうち，音声フォワード補完と音声バックワード補完の使用割合は，前者が 61.7%，後者が 38.3% であった．また，39 名の被験者が 5 回中 3 回以上音声補完を使用しており，1 回も使用しなかった被験者はいなかった．

表 2は，入力条件 (4) で，被験者が何も単語を見ない状態で入力するときに (うろ覚えの単語を入力するときに)，音声補完を使用するか使用しないかを調査した結果である．被験者は平均 3.18 単語を思い出しながら入力できたが，その 80.4% で音声補完を使用していた．そのうち，音声フォワード補完と音声バックワード補完の使用割合は，前者が 66.1%，後者が 33.9% であった．また，被験者が事前に入力する単語を知っていたかどうかをアンケート時に報告させたところ，平均 1.67 単語を既に知っていた (ただし，事前に知っていた単語を必ずしも入力できていなかった)．一つも事前に知らなかった被験者は 11 名いたが，その全員が思い出して入力する際には音声補完を必ず使用していた．

図 8のアンケートを集計した結果，8 項目のそれぞれについて −3 〜 +3 の 7 段階尺度で評定した値の平均は表 3となった．評定値の比率を帯グラフで図 9に示す．最も評定値の平均が高かったのは 2. と 7. の項目で，実際に，補完候補による手助けが有効であり，音声補完が便利であったことがわかる．次に，1., 5., 8. の項目で評定値が高く，入力内容を思い出しやすく親切であり，被験者は今後も使いたいと思っていることがわかる．3. と 6. の項目も，約 3 分の 2 の被験者が，入力が楽で使いやすいと支持していた．一方，4. については他と異なり，被験者の反応の大半は −1 〜 +1 にあった．その理由として，本実験では何を入力するかが明確に指示されていたため，従来の音声認識に対する話しかけにくさを感じたり，音声補完が話しかけやすいかどうかを判断できる状況とはならず，適切に評価されなかった可能性がある．これについては，今後検討の余地が残された．

以上から，45 名の全被験者が音声補完を使いこなすことができ，音声補完を使用するか使用しないかが自由な条件でも，使用されることが多かったことがわかる．特に，思い出しながら入力する際には，入力単語を一つも事前に知らなかった被験者は必ず音声補完を用いていた．また，アンケートの結果からも，音声補完の有効性が確認できた．

7 議論

音声補完は，音声認識をインタフェースとして使いやすくするにはどうすべきかという観点から生まれた研究であり，今後様々な方向への発展が考えられる．以下では，そのような方向性も含めて議論する．

7.1 関連研究

テキスト (キーボード) 入力で広く受け入れられている補完機能として，1 章では補完トリガーキーによる手動補完に言及したが，WWW ブラウザの URL 入力や，Reactive Keyboard [1]では，自動補完機能が導入されている．これは，ユーザがタイプしている最中に，システム側が補完候補一覧を次々と提示していく機能である．また，ペン入力に関しても，自動補完機能を持った予測ペン入力インタフェース [2]や POBox [9] [11]が提案されている．前者では升目にペンで手書きした文字から，後者ではソフトキーボードで入力した文字から，続きの文字列が辞書や履歴等に基づいて予測・提示される．これらは，予測インタフェース [10]とも呼ばれ，有効性が示されてきた．

しかし，音声入力の場合には，上記のような自動補完は不適切な機能となりやすい．キーボード入力や，ソフ

表 3 音声補完使用後のアンケートの集計結果 (評定値の平均)

項目	評定値の平均	+1 ～ +3 の占める比率
1. 入力内容を思い出しやすい	+1.56	77.8%
2. 補完候補による手助けが有効である	+2.02	91.1%
3. 入力が楽である	+1.18	66.7%
4. 話しかけやすい	+0.09	40.0%
5. 親切である	+1.49	77.8%
6. 使いやすい	+0.96	66.7%
7. 便利である	+2.13	93.3%
8. 今後使いたい	+1.40	80.0%

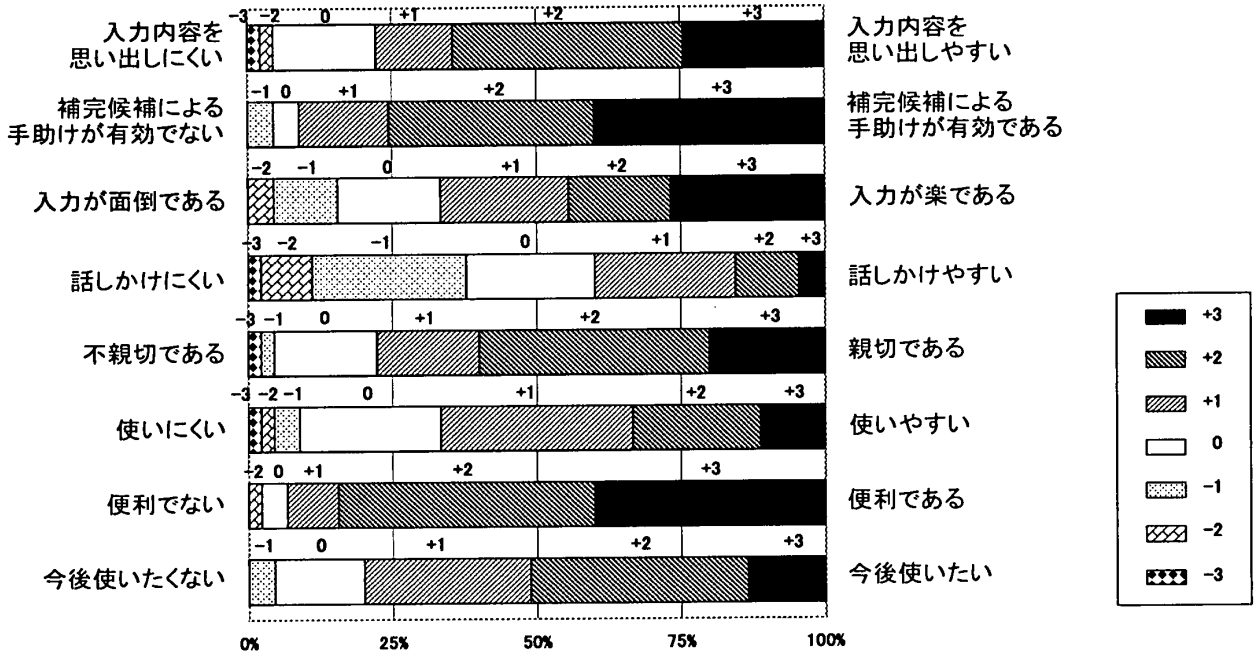


図 9 音声補完使用後のアンケートの集計結果 (帯グラフ)

トキーボードを用いたペン入力では，各キーを押したことを認識する段階で曖昧性はなく，各文字の境界は明らか (明確に分節可能) である．手書き文字認識を用いたペン入力でも，文字の認識時に曖昧性はあるものの 1 文字の単位は音素より大きく，各文字は分節可能な条件で入力される．一方，音声入力では，音素の認識時に曖昧性が高い上に，音素の境界を決定することも難しい (分節が困難である)[†8]．そのため，そもそもどの時点で補完候補を提示するかが一意に決まらず，仮に一定間隔で提示したとしても，キーボード入力やペン入力のように高精度で適切な候補を提示し続けることは困難である．「予測を使わない場合に比べて少しでも不都合がある場合には予測インタフェースは使われない傾向がある」[10]ことから考えても，音声の自動補完は煩わしく，実用的でない可能性が高い．それに対して本研究では，有声休止によってユーザが意図した箇所で明示的に補完機能を呼び出せるようにしたことで，補完して欲しくないときには一切干渉することがない実用的なインタフェースが実現できた．

7.2 音声中の複数のモダリティを活用したマルチ

†8 いわば楷書でなく草書で書かれた文字列のようなものである．

モーダルインタフェース

従来の音声インタフェースでは，1章でも述べたように，音声認識を中心とした言語情報のモダリティが主に利用されてきた．それに対して本研究の音声補完では，有声休止のような，音声中に含まれる非言語情報のモダリティを積極的に利用して，より使いやすいインタフェースを実現した．これは，音声音響信号が持つ複数のモダリティを活用した，一種のマルチモーダルインタフェースであると我々は捉えている．

そして，今後他の非言語情報のモダリティも導入していくことで，さらに使いやすい音声インタフェースが構築できる可能性がある．キーボード[†9]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，本研究での有声休止の位置付けは，いわば特殊キーのTab (UNIXシェルやEmacsエディタの補完トリガーキー) に相当する．これを第一歩として，音声の音高や話速等の他の非言語情報を特殊キーとして活用するような研究が，今後発展していく余地は大きい．しかも，キーボードの機能の範囲に留まる必要はない．音声フォワード補完の有声休止が音韻情報を同時に伝えていたことからもわかるように，多くの非言語情報のモダリティは，言語情報と同時に伝えられるようなメタな情報伝達手段である．このように高い潜在能力を持つ音声のメリットをさらに引き出せば，従来から議論されている音声の利点 (ハンズフリー，速い入力速度，等) とも相まって，優れたインタフェースを生み出していけるはずである．

8 おわりに

本論文では，発話された単語断片の残りを補うことでユーザの音声入力を手助けする「音声補完」という新しい音声インタフェース機能を提唱し，「音声フォワード補完」と「音声バックワード補完 (音声ワイルドカード補完)」の二つの補完方式を提案した．実際に，インタラクティブに音声入力可能なシステムを実装し，曲名とアーティスト名の入力で有用性を確認したが，これは住所入力や各種固有名詞の入力等の様々な局面にもすぐに適用できる．音声補完による手助けは，ひとたび使用し始めると，手助けがない状態を不便に感じるほど便利なものであり，今後，音声入力インタフェースを構築する上で，不可欠な機能の一つになることが予想される．

今後は，補完候補の選択操作の自由度を高めたり (タッチパネルとの併用や候補が一つに絞れる場合の自動確定等)，補完対象を単語よりも長い単位に拡張したりしていく予定である．また，音声補完を発端とした新たな音声インターフェース研究の方向性も探求していきたい．

†9 本論文ではキーボードが，通常キー (英数字等の文字がそのまま入るキー) と特殊キー (Tab や Shift 等の特別な働きを持つキー) の二種類で構成されると考える．

参考文献

[1] Darragh, J. J., Witten, I. H. and James, M. L. : The Reactive Keyboard: A Predictive Typing Aid, *IEEE Computer*, Vol. 23, No. 11 (1990), pp. 41–49.

[2] 福島俊一, 山田洋志 : 予測ペン入力インタフェースとその手書き操作削減効果, 情処学論, Vol. 37, No. 1 (1996), pp. 23–30.

[3] 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 音声言語情報処理 *99-SLP-27-2*, 1999, pp. 9–16.

[4] 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 (2000), pp. 2330–2340.

[5] 後藤真孝, 根山亮, 村岡洋一 : RMCP: 遠隔音楽制御用プロトコルを中心とした音楽情報処理, 情処学論, Vol. 40, No. 3 (1999), pp. 1335–1345.

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

[7] 井田昌之, 亀井信義 : Emacs 解剖学 入力の補完, *bit*, Vol. 29, No. 2 (1997), pp. 85–95.

[8] 伊藤克亘, 速水悟, 田中穂積 : 音素文脈依存モデルと高速な探索手法を用いた連続音声認識, 信学論 *(D-II)*, Vol. J75-D-II, No. 6 (1992), pp. 1023–1030.

[9] 増井俊之 : ペンを用いた高速文書入力手法, インタラクティブシステムとソフトウェア *IV*, 近代科学社, 1996, pp. 51–60.

[10] 増井俊之 : 予測 / 例示インタフェースの研究動向, コンピュータソフトウェア, Vol. 14, No. 3 (1997), pp. 4–19.

[11] Masui, T. : An Efficient Text Input Method for Pen-based Computers, *Proceedings of CHI '98*, 1998, pp. 328–335.

[12] Rose, R. L. : The communicative value of filled pauses in spontaneous speech, Master's thesis, University of Birmingham, 1998.

[13] 田窪行則 : 音声言語の言語学的モデルをめざして — 音声対話管理標識を中心に —, 情報処理, Vol. 36, No. 11 (1995), pp. 1020–1026.

[14] 田中敏 : 「休止」の意味論, 言語, Vol. 22, No. 8 (1993), pp. 20–27.