音声訂正：選択操作による効率的な誤り訂正が可能な音声入力インタフェース

緒方 淳[†] 後藤真孝[†]

本論文では，ユーザが認識誤りを選択操作により効率的に訂正することが可能な「音声訂正」という音声入力インタフェース機能を提案する．音声訂正では，ユーザが音声入力を開始すると，認識結果を単語ごとに区切った表示と，区切られた各区間に対する他候補（競合候補）が発話の最中から次々と画面に描画され，ユーザは競合候補の中から本来の正解を選択するだけで認識誤りを訂正することが可能となる．また，音声訂正では，ユーザが発声中であっても訂正処理が可能な「即時誤り訂正機能」と，ユーザが意図的に発声を休止し，認識処理を一時中断させることが可能な「発話中休止機能」を実現する．25 人の被験者による評価実験を行ったところ，音声訂正は使いやすく，効果的な音声入力インタフェースであることが確認された．

Speech Repair: Speech Input Interface Capable of Quick Error Correction by Using Selection Operation

Jun Ogata[†] and Masataka Goto[†]

In this paper, we propose a speech input interface function, called "*Speech Repair*", which enables a user to easily correct recognition errors by selecting candidates. During the speech input, this function displays not only the typical speech-recognition result but also other competitive candidates. Each word in the result is separated by line segments and accompanied by other word candidates. A user who finds a recognition error can simply select the correct word from the candidates for that temporal region. Furthermore, we introduce two additional functions: *immediate correction function* that enables the user to correct errors not only when the recognition process is complete but also whenever the user finds erroneous words, and *intentional suspension function* that enables the user to intentionally suspend and resume the recognition process. Experimental results with twenty-five subjects showed that the speech-repair function is easy to use and effective interface.

1. はじめに

音声認識技術を改良してどんなに認識率をあげていったとしても，人間にとって，つねに明瞭で曖昧性のない発声をし続けることはきわめて困難である以上，認識率はけっして 100%にはならない．したがって，音声認識を日常的に使えるインタフェースにするためには，必ずどこかで生じてしまう誤認識を容易に訂正できる音声入力インタフェースが不可欠となる．そのため，従来からそうした訂正のためのインタフェースが提案されてきた．たとえば，市販のディクテーションソフトでは，認識結果のテキスト表示をユーザが見て，誤認識を発見したら，その区間をマウス操作や音声入力で指定することができる．すると，その部分の他候補が表示されるので，ユーザは正しい候補を選択して訂正できる．文献 1) の研究ではこれを発展させ，発話の終了後にその認識結果を単語境界の線で区切った表示をし，かな漢字変換で単語の区切りを修正するように，その境界をマウスで移動できるようにした．この場合，正しい候補にたどり着ける可能性は高くなったものの，誤認識箇所の指定，単語境界の変更，候補の選択と，ユーザが訂正するための手間は増えてしまっていた．一方，文献 2) では，音声認識を利用したニュース字幕放送のために，実用的な認識誤り修正システムを実現している．しかし，2 人の分業を前提とし，1 人が誤認識箇所を発見してマーキングし，もう 1 人がその箇所の正解をタイピングする必要があったため，個人が自分の音声入力を訂正する目的では使えなかった．このようにいずれの従来手法も，まず最初に，ユーザが誤認識箇所を発見して指摘し，次に，その部分の他候補を判断して選択したり，タイピング

† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

して修正したりするといった手間を要していた．

本研究では，音声認識による認識誤りを，ユーザがより効率的で容易に訂正できる新たな音声入力インタフェース「音声訂正」を提案する．音声訂正では，ユーザが音声入力を開始すると，認識結果を単語ごとに区切った表示が発話の最中から次々と画面に描画される．同時に，区切られた各区間の他候補（競合候補）もつねに列挙されていく．ここで，競合候補の個数はその区間の曖昧さを反映しており，音声認識結果として信頼性が低い箇所ほど，多数の候補が表示される．ユーザはそれを見ながら，発話中あるいは発話終了後に正しい候補を選択するだけで訂正ができる．ここで重要なのは，わざわざユーザが誤認識箇所を発見して指摘しなくても，つねに競合候補がリアルタイムにフィードバックされ続けていることである．これにより，従来研究のように誤認識箇所の発見，指摘，提示された候補の判断，選択といった手間をかけずに，いきなり候補を見て選択するだけで，効率良く訂正できる．さらに，こうして発話の最中に候補を選べるようになると，選択操作の間，音声認識器に一時的に待ってほしくなることがある．そこで，単に発話中に有声休止（語中の任意の母音の引き延ばし）で言い淀むだけで発話を中断可能とし，その次の発話はあたかも中断前の発話が続いていたかのように入力できるようにした．

本インタフェースは，音声入力のみによる完全なハンズフリーの状況を想定したものではなく，音声を主とし，その他の入力デバイスも併用することによって，より快適で効率的な文章入力を行うことを目的としている．実際の利用状況としては，まず，キーボード，マウスなどのデバイスが自由に使える汎用コンピュータ（デスクトップ型，ラップトップ型）上でのドキュメント作成，文章入力があげられる．また，本インタフェースは，ペン操作を主とするタブレット PC や携帯型端末など，キーボードが利用できない環境において，より有効な文章入力手段になると考えられる．

以下，2 章において本研究で提案する「音声訂正」という新たな音声インタフェースについて述べ，3 章でその具体的な実現方法を説明する．次に，4 章において音声訂正インタフェースの実装の詳細について述べ，5 章で音声訂正の基本性能について実験的に評価する．最後に，6 章で音声訂正インタフェースに関するユーザビリティの評価を行い，7 章でまとめを述べる．

2. 音 声 訂 正

本研究で提案する「音声訂正」とは，音声認識器により引き起こされた誤認識を，ユーザとのインタラクションを介して訂正する機能である．通常の音声認識器では，確定した認識結果（単語列）をユーザに 1 つだけ提示していた．そのため，発話終了後にユーザが認識誤りを訂正するためには，以下の 2 つの手続きをとる必要があった．

(1) 認識結果の中から誤り箇所を探して指摘する．
(2) 指摘した誤り箇所を訂正する．

音声訂正では，これらを 1 度の操作で効率的に行うことで，認識誤りを訂正する際のユーザへの負担を減らすことを目的としている．

2.1 音声訂正の基本機能

図 1 に音声訂正インタフェースの画面表示の模式図を示す．音声訂正では，ユーザの発声が入力されると，図 1 上側に示すような結果が即座に提示される（音声入力開始とともに左から右へ順次表示されていく）．音声訂正では，従来の音声認識と異なり，最上段の通常の認識結果（単語列）に加えて，その下へ「競合候補」のリストをつねに表示する．競合候補とは，音声認識の認識処理過程において，通常の認識結果以外に可能性の高かった単語候補である．図 1 のように，通常の認識結果が各単語の区間ごとに区切られて，その単語に対する競合候補が整列して表示される．ここで，競合候補の個数はその区間の曖昧さを反映しており，音声認識結果として信頼性の低い箇所ほど，多数の候補が表示される．そのため，ユーザは候補が多いところに誤認識がありそうだと思って，注意深く見ることができる．逆に，認識結果として信頼性が高い区間は候補が少ないため，ユーザに余計な混乱を与えることがない．このように認識結果を提示することで，ユーザは競合候補の中から正解を「選択」する操作だけで，容易に認識誤りを訂正できる．

なお，図 1 のように，選択肢には必ず空白の候補が

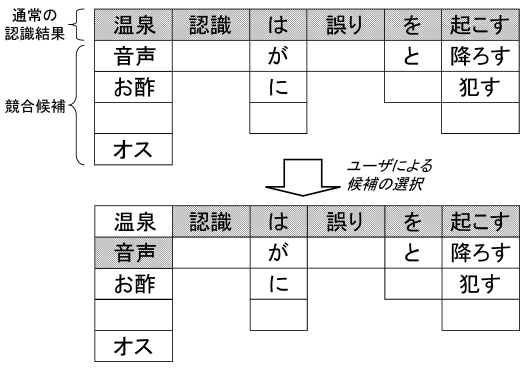

図 1 選択するだけで誤りの訂正ができる音声訂正インタフェース（「音声認識は誤りを起こす」という発声が誤認識された例）
Fig. 1 Overview of the speech-repair function.

含まれる．これを「スキップ候補」と呼び，その候補が属する区間の認識結果をないものとする役割を持つ．これにより，最上段の認識結果に湧き出し誤り（本来あるべきでない区間に余分な単語が挿入される誤り）が存在しても，ユーザはスキップ候補を選択するだけで容易に削除できる．つまり単語の置き換えと削除が，「選択」という 1 つの操作でシームレスに実行できる．また，各区間の競合候補は，上から可能性（存在確率）の高い順に並んでいる．つまり，上の方ほど音声認識結果として信頼性が高い候補であるので，通常はユーザが上から下へ候補を見ていくと，早く正解にたどり着けるようになっている．さらに，本インタフェースでは，発話中の認識結果として可能性のある単語候補が網羅的に列挙され，各区間にスキップ候補も持っているため，文献 1) で提案されているような認識結果の単語境界の変更も不要になるメリットがある．

以上の音声訂正の基本機能はシンプルだが，従来こうしたインタフェースは実現されていなかった．その理由としては，大語彙を対象とした連続音声認識では，競合候補を表示しようと思ってもその数が膨大なものとなってしまい，現実的な分量でユーザに提示することが困難だったからである．それに対し音声訂正では，効率的な中間表現形式である「confusion network」を，誤り訂正インタフェースへと応用することにより，大語彙，小語彙を問わず多様な入力音声に対して上述のような効果的な候補の提示，訂正を可能にした．

2.2 発話中における即時誤り訂正機能

使いやすいインタフェースを構築するには，ユーザの入力中に逐次現在の認識状態をフィードバックする必要がある．しかし，従来の一部の音声認識器では，発話が終了するまで認識結果が表示されないことがあった．仮に結果が表示されたとしても，競合候補のような他の可能性が示されることはなく，発話が終了してから結果を吟味するまで，誤りの訂正に移ることはできなかった．そのため，音声入力はキーボード入力と比べて，誤り訂正作業に多くの時間がかかる欠点があることが指摘されていた[3]．文献 3) によれば，その要因として，訂正自体の時間以外に，1) ユーザが誤り箇所を発見するための時間，2) 誤り箇所を指摘する（カーソル移動する）ための時間，が余計にかかる点があげられていた．

それに対して音声訂正では，発話中に認識の中間結果を競合候補付きでリアルタイムにフィードバックし続け，さらにユーザの選択も可能にすることで，発声の最中に誤りを即時に訂正可能な機能（即時誤り訂正機能）を実現する．これにより，上述の 2 点の作業時間が短縮される．また実際の訂正にかかる時間も，すでに表示されている候補を「選択」するだけなので早い．

2.3 発話中休止機能

前節の即時誤り訂正機能を使っていると，発話中に正しい候補を選択している間，音声認識器に一時的に続きを言うのを待ってほしくなる場面が出てくる．しかし，通常の音声認識器による認識単位は，無音で区切られた一息で言える区間なので，むやみに発声を中断するとうまく認識されないという問題があった．

そこで音声訂正では，発話中にユーザが意図した時点で，認識処理を一時停止させる新たな機能（発話中休止機能と呼ぶ）を実現する．そして次の発話が始まると，あたかも一時停止前の発話が続いていたかのように動作させる．このユーザの一時停止の意図を伝えるために，音声中の非言語情報の 1 つである有声休止（語中の任意の母音の引き延ばし，たとえば，図 6 中の「三度の飯よ**りー**」など）を，発話中休止機能のトリガとして採用した．人間同士の対話においても，相手に少し待ってほしいときや，喋っている最中に考えごとをするときなどに，このように有声休止によって言い淀むことが多い．そのため，ユーザは自然に一時停止をかけて，正しい候補を選択したり，続きの発話を考えたりできる．

3. 音声訂正の実現方法

提案する音声訂正インタフェースの具体的な実現方法を述べる．

3.1 音声認識における中間結果

音声訂正を実現するためには，図 1 のような効果的な競合候補の提示が不可欠である．競合候補は，音声認識のデコード処理の途中経過を表す中間的な表現形式（「中間結果」と呼ぶ）を用いて生成される．

一般的な音声認識の中間結果としては，N-best 文リスト[5]，単語グラフ[6] があげられる．N-best 文リストとは，認識結果に対する複数文候補を表す．N-best 文リストは簡易な表現形式ではあるが，1 単語のみ異なるような類似の文候補が大量に出現することになり，本来の正解を得るためには結局膨大な数の文候補を残す必要がある．単語グラフは，N-best 文リストを単語をリンクとするグラフにまとめたものであり，N-best 文リストをよりコンパクトにした形式となる．しかし，大語彙の連続音声認識になると，グラフ中の候補の数は膨大になり，候補間の競合関係が明示的に表現できていないため，音声訂正のような効果的な候

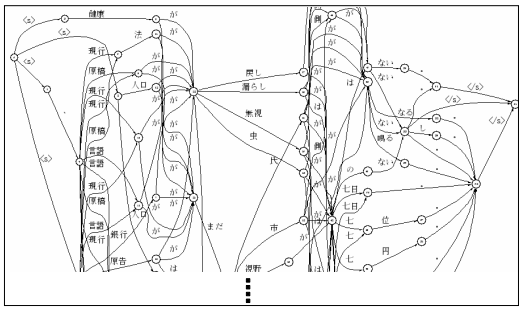

図 2 従来の中間結果（単語グラフ）の例
Fig. 2 Example of word graph.

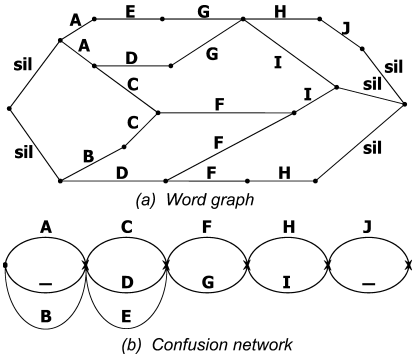

図 3 単語グラフと confusion network の模式図
Fig. 3 Word graph and confusion network.

補提示は不可能と考えられる．ここで，単語グラフの実際の一例を，**図 2** に示す（紙面の都合上，一部分だけ示した）．図 2 は 1 つの単語グラフ（ノード数 210，リンク数 810）の約半分程度を例示しただけだが，候補数が膨大であることが分かる．

なお，文献 1) では，単語グラフの前段階の中間結果である「単語トレリス」を直接用いて複数候補を提示する機能が提案されているが，単語トレリスは単語グラフ以上に候補の絞り込み能力を持たず[7]，語彙数が増えるに従って使用が難しくなる可能性があると考えられる．

3.2 confusion network

以上の問題を解決する新しい中間結果として，本研究では，音声認識器の内部状態をシンプルかつ高精度なネットワーク構造へ変換した confusion network[4] を導入する．confusion network は，音声認識における新たなデコーディング手法である「単語誤り最小化音声認識」[4] において導入された途中結果であり，これまでに，本研究のような誤り訂正インタフェースに応用しようという発想はなかった．

confusion network は，単語グラフ（**図 3** (a)）を音響的なクラスタリングによりリニアな形式（図 3 (b)）に圧縮することで求めることができる．ここで “sil”（silence）は発話開始，終了時の無音を表し，アルファベット 1 文字はグラフのリンク上の単語名を表している．また，図 3 (b) のネットワーク上の “-” はスキップ候補である．音響的クラスタリングは以下のステップにより行われる[4]．

クラスタの初期化（initialization）： 単語グラフ中のすべてのアークの中で，単語ラベルが同一で，始端ノードの時間と終端ノードの時間がそれぞれ同一のアークをクラスタリングする．

単語内クラスタリング（intra-word clustering）： 単語ラベルが同一で，時間的に重なりのあるアークをクラスタリングする．クラスタリングは以下のコスト関数を用いて，greedy アルゴリズムに基づき行われる．

$$S(E_1, E_2) = \max_{\substack{e_1 \in E_1 \\ e_2 \in E_2}} overlap(e_1, e_2) p(e_1) p(e_2)$$

ここで，E_1，E_2 はマージ対象のクラスタ，e_1，e_2 はそれぞれのクラスタ内のアークである．$p(\cdot)$ はそれぞれのアークの事後確率で，単語グラフ上で forward・backward アルゴリズムを用いて算出される．また，$overlap(\cdot, \cdot)$ は 2 つのアーク間の時間のオーバラップ率を表している．

単語間クラスタリング（inter-word clustering）： 単語ラベルの違うアークのクラスタリングを行う．基本的には，前ステップと同様の手順でクラスタリングが行われるが，コスト関数には以下を適用する．

$$S(F_1, F_2) = \mathrm{average}_{\substack{w_1 \in W(F_1) \\ w_2 \in W(F_2)}} s(w_1, w_2) p_{F1}(w_1) p_{F2}(w_2)$$

ここで，$W(F)$ はあるクラスタ F に含まれる全単語のリストを表しており，$p_F(w)$ はクラスタ F 中の単語 w の事後確率を表している．また，$s(\cdot, \cdot)$ は音響的な類似度を表しており，2 単語の音素列間の Levenshtein 距離によって計算される．クラスタリングは，マージ対象がなくなった時点，すなわちすべての候補のアラインメントが達成されたときに停止する．

confusion network の各リンクには，クラスタリングした各クラス（単語の区間）ごとに事後確率が付与され，それらの値は，各クラスでの存在確率，あるいはそのクラス内の他候補との競合確率を表す．各クラスのリンクは，存在確率の大きさでソートされ，認識結果として可能性の高いリンクほど上位に配置される．最終的に，各クラスから事後確率が最大となるリンクを選択すると，図 1 の最上段のような最終的な認識結果（最尤の候補）となる．また，各クラスで事後確率が高いリンクを取り出すと，図 1 の競合候補が得られる．

ただし confusion network では，クラス中の各候補は必ずしも時間的に同一区間の認識結果とは限らない．たとえば，時間的に 2 つのクラスをまたがった候補は，どちらか一方のクラスへ割り当てられる．我々の音声訂正では，そのような候補をユーザが選択すると，発声区間との時間的な整合性がとれるように，近隣でユーザが未選択なクラスの候補も自動的に選択され，訂正操作の回数を最小限にする（たとえば図 5 (1) で，「たちまち」を選択すると，その前の区間は自動的にスキップ候補が選択される）．

3.3 即時誤り訂正機能の実現方法

即時誤り訂正機能では，いかに素早く中間結果を逐次提示できるかが重要となる．そのために本研究では，ある一定の時間（500 ms）ごとに，中間結果である confusion network を逐次生成できるよう，音声認識器を拡張した．具体的には，まず，ある時刻において生き残った単語候補の中から，尤度の大きさで上位 5 つを選択し，それぞれの候補から発話先頭に向かってバックトレースし，発話の先頭からその時刻までの単語グラフを生成する．上位 5 つに限定する理由としては，その時刻中に残ったすべての候補を用いると，不必要な（尤度が極端に低い）候補が多く含まれてしまい，また，単語グラフのサイズも不必要に大きくなり，リアルタイムの処理が困難になるためである．次に，前節で述べた音響的クラスタリングアルゴリズムにより，confusion network を生成する．このようにして，一定の時間ごとに競合候補とともに中間的な認識結果を生成し，ユーザ側に逐次提示することで，即時に誤りを訂正することを可能にした．

3.4 発話中休止機能の実現方法

発話中休止機能の具体的実現方法について説明する．発話中に有声休止（言い淀み）が検出され，その直後に一定の無音区間が検出されたら，音声認識器の動作を一時停止し，現時点の認識処理過程（それまでの仮説情報，探索空間での現在の位置情報など）を退避する．このとき，有声休止が発声され続けている区間は音声認識の対象とならず，スキップされる．再び発話の開始が検出されると（音声のパワーに基づいて検出），退避した認識処理過程から音声認識処理を再開し，発話終端が検出されるまで認識処理を続行する．

有声休止の検出には，文献 8) のリアルタイム有声休止検出手法を採用した．この手法は，有声休止（母音の引き延ばし）が持つ 2 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リアルタイムに検出する．そのため，任意の母音の引き延ばしを言語非依存に検出できるという特長を持っている．

3.5 未選択候補の自動訂正による訂正回数の最小化

音声認識では，ある単語を誤ると，その単語に引きずられる形で隣接する候補として言語的に誤った単語が認識されることがある（例，図 5 (1) 中，「音声 入力」⇒「温泉 入浴」）．このような「連続して発生する誤り」に対し，本研究では，ユーザがある候補を訂正すると，その隣接する候補も適切なものに自動的に訂正する機能を実現した[1)]．

本インタフェースでは，ユーザが選択した単語の前後それぞれの候補に対し，現在選択している候補との言語的接続確率（N-gram）を算出し，その値が最も大きい候補に自動的に修正する．たとえば，図 5 (1) において，ユーザが「温泉」を「音声」に訂正すると，「音声」との言語的接続確率が最も高い「入力」が自動的に選択され，「入浴」が「入力」へと訂正される．ただし，このとき，自動修正の対象となる区間が，すでにユーザにより訂正済みであるならば，自動修正は実行しない．このような機能により，ユーザの訂正操作の回数を最小限に抑えることができると考えられる．また，現段階では，ユーザが選択した単語の両隣の候補のみの自動修正を実装しているが，これの拡張として，さらに多くの候補に対して自動修正が行われるように，ユーザがある候補を訂正した情報を利用して，発話全体の候補を再度選択し直すという機能も考えられる．

3.6 音声訂正のための音声認識器

音声訂正インタフェースを実現するためには，音声認識器において，競合候補の作成をリアルタイムに行うことが不可欠である．しかし，高精度な単語グラフを生成するための N-best 探索アルゴリズムは，一般的に非常に大きな計算コストがかかり[7)]，認識結果の確定が遅くなり，その結果，効率的な訂正処理を行うことは困難となる．それに対し，本インタフェースでは，音声認識器の認識アルゴリズムとして，back-off 制約 N-best 探索手法[9),10)] を用いることで，リアルタイムに競合候補を生成，提示することが可能となっている．

4. 音声訂正インタフェースの実装

2 章で述べた各要素技術を用いて，提案した音声訂正インタフェースを実現するシステムを実装した．

4.1 システム構成

図 4 に，音声訂正インタフェースの各システム構成要素（プロセス）と，全体の処理の流れを示す．プロセスは図中の囲み字で示されており，ネットワーク

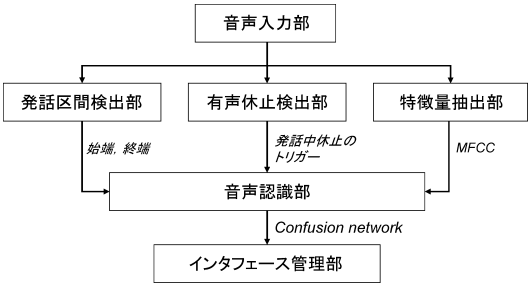

図 4 全体の処理の流れ
Fig. 4 System architecture.

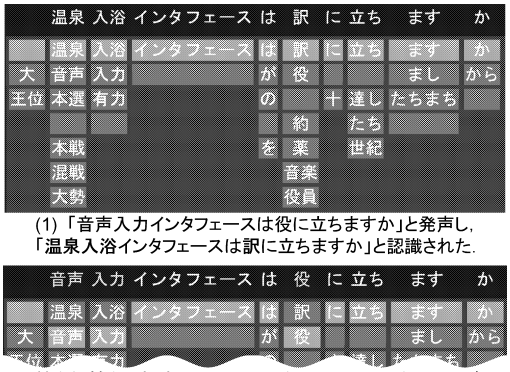

図 5 発話中休止機能を利用しない場合の画面表示例（「音声入力インタフェースは役に立ちますか」という文章を発声）
Fig. 5 Screen snapshots of speech repair.

（LAN）上の複数の計算機で分散して実行することが可能である．プロセス間の通信には，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）[11] を用いた．

処理の流れについて説明する．まず，マイクロフォンなどから音声入力部に入力された音響信号は，ネットワーク上にパケットとして送信される．特徴量抽出部，有声休止検出部，発話区間検出部がそのパケットを同時に受信し，音響特徴量（MFCC）や有声休止，発話の始終端をそれぞれ求める．これらの情報は，パケットとして音声認識部に送信され，認識処理が実行される．このとき，有声休止は，発話中休止機能を呼び出すトリガとして利用される．音声認識部では，中間結果として confusion network が生成され，その情報はパケットとしてインタフェース管理部に送信される．インタフェース管理部では候補を表示し，マウスによるクリックや，パネル上をペンや指で触れる操作によってその選択を可能にする．

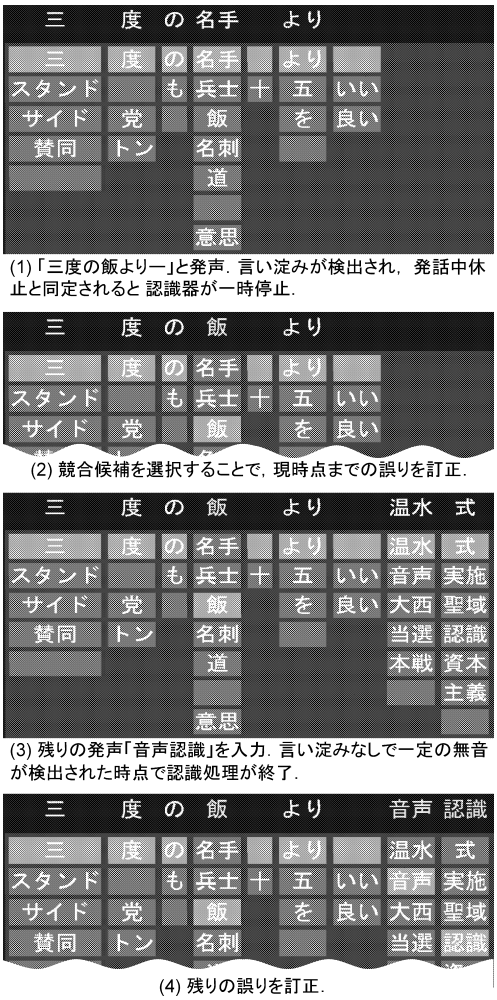

図 6 発話中休止機能を利用した場合の画面表示例（「三度の飯より音声認識」という文章を発声）
Fig. 6 Screen snapshots of intentional suspension function.

4.2 実 行 例

図 5 に発話中休止機能を利用しない場合の表示画面を，**図 6** に発話中休止機能を利用した場合の表示画面をそれぞれ示す．図 1 に相当する表示部分（「候補表示部」と呼ぶ）の上に，さらに 1 行追加されているが，これは，候補を選択して訂正した後の最終的な音声入力結果を表示している．候補表示部では，現在選択されている単語の背景が着色される．何も選択していない状態では，候補表示部の最上段の最尤単語列が選択されている．ユーザが他の候補をクリックして選択すると，その候補の背景が着色されるだけでなく，画面最上部の最終的な音声入力結果も書き換えられる（選択操作で訂正した箇所だけ，文字の色を変えて分かりやすく表示している）．

5. 音声データベースを用いた基本機能の評価

音声訂正が実用的に使えるかどうかを評価するには，認識誤りを訂正することがどの程度可能か，すなわち，表示される競合候補の中に本来の正解がどの程度含まれているか，を調査することが重要となる．そこで，ここでは，読み上げ音声，話し言葉音声の 2 種類の音声データベースを用いて，それぞれの場合に対しての評価を行った．

読み上げ音声に対する実験では，音響モデルとしては，新聞記事読み上げコーパス JNAS から学習した音節モデル（モデル数 244）[12]，言語モデルとしては，新聞記事テキストより学習された bigram（語彙数 20 K）[13] を用いた．評価用データは，JNAS 中の男性 25 人が発話した 100 発話である．一方，話し言葉音声に対する実験では，音響モデルとしては，CSJ（日本語話し言葉コーパス）中の男性話者 200 人の講演音声を用いて学習した音節モデル，言語モデルとしては，CSJ の「短単位データベース」中の 319 講演から学習した bigram（語彙数 14 K）を用いている．評価用データは，男性話者 4 人の学会講演中の 100 発話（各話者 25 発話）である．

実験では，上記のそれぞれの評価用データを対象に，候補を上位 N 個まで提示したときの訂正後の認識率（最終的な音声入力成功率）を，誤り訂正能力として評価した．つまりここでの認識率は，たとえば N = 5 の場合，上位 5 個以内に正解が含まれる割合で表される．通常の認識性能（N = 1 のときの認識率）は，読み上げ音声（JNAS）では 86.70%，話し言葉音声（CSJ）では 77.79%であった．また，それぞれの評価用データにおける未知語数は，JNAS では 4，CSJ では 25 であった．

図 7 に，N の値ごとの訂正後の認識率を示す．実験結果より，どちらのデータにおいても，提示する候補数を増やすと飛躍的に認識率が向上しており，JNAS では N = 11，CSJ では N = 14 で認識率は飽和状態となった．JNAS では最終的な認識率は 99.36%となり，約 95%の誤り（199/209）を訂正可能であることが分かった．一方，CSJ においても，通常の認識性能が JNAS に比べて約 10%低いにもかかわらず，最終的な認識率は 96.16%まで向上し，約 83%の誤り（324/391）を訂正可能であることが分かった．また，両データにおいて，N = 5 程度でもかなりの数の誤りを訂正できることも分かった．話し言葉音声である CSJ では，読み上げ音声である JNAS に比べて，訂正性能は低い値となったが，その原因としては，話し言葉音声特有の問題である，言い誤り，言い直し部分における誤りが多かったことがあげられる．未知語が多いことも含め，特に音声認識システムに用いる言語モデルの性能を向上させることで，さらなる訂正性能の改善が期待できる．

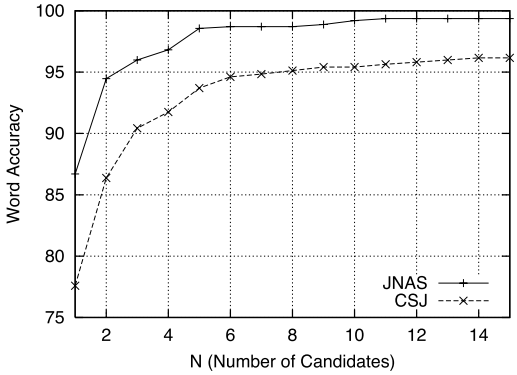

図 7 提示する候補数の上限を変えたときの訂正後の認識率（最終的な音声入力成功率）
Fig. 7 Word accuracy with different number of competitive candidates displayed.

音声訂正では，提示する候補数が多すぎるとユーザ側の混乱を招き，逆に少なすぎると誤りを訂正できなくなるが，confusion network を用いることにより，提示する競合候補数を抑えつつ，ほとんどの誤りを訂正することが可能であることが分かった．ただし，実験でも示されたように，音声認識システムが知らない未知語に関しては，現時点では，音声訂正を用いても訂正できない．この解決は今後の課題であり，ユーザとのさらなるインタラクションを介して未知語を解消する枠組みが必要になると考えている．

6. 被験者実験による音声訂正インタフェースの評価

ここでは，被験者実験による音声訂正インタフェースの評価について述べる．本実験では，コンピュータの利用状況として現状で最も一般的と考えられる，「デスクトップコンピュータ上での文章入力」をタスクとして，以下の 3 点について調査を行った．

- 音声訂正により，通常の音声入力（最尤単語列のみを出力する音声認識）に比べて，文入力がどの程度効率的になるか．
- 比較的長い文やうろ覚えの文を入力する際に，発話中休止機能が有効に働くかどうか．
- 音声訂正インタフェース全体に対してどのような印象を受けたか．

今回の実験では，次節で述べるように，評価の目的

表 1 被験者のキーボード入力リテラシ
Table 1 Keyboard literacy of subjects.

経験年数	被験者数
1〜3 年	7
3〜5 年	7
5〜10 年	11
10 年以上	2

に応じた 3 つの課題を被験者に対して行ったが，全体的な実験方法としては，提示した文を被験者に入力してもらい，1 字 1 句間違えずに入力できた時点で 1 つの文を入力完了とした．また，通常の音声入力の際に認識誤りが発生した場合や，音声訂正で競合候補中に本来の正解がなくて選択できない場合には，キーボード，マウスを利用してタイプ入力で訂正することとした．各課題の 3 つの文は，全被験者を通して共通だが，順番をランダムに変えたものを 6 通り用意して各被験者に割り当てた．

本実験で使用した音声認識システムは，音響モデルには，新聞記事読み上げコーパス JNAS から学習した triphone モデルを，言語モデルには，CSRC ソフトウェア 2000 年度版[13] の中から，新聞記事テキストより学習された 20,000 語の bigram をそれぞれ用いた．なお，音声を用いた実験では，音声認識単体の基本性能差による影響を排除するために，認識デコーダ，音響モデル，言語モデルに関しては，通常の音声入力，音声訂正の各実験ともに同一のものを用いた．すなわち，本実験における通常の音声入力による認識結果は，音声訂正の競合候補の中の最尤単語列（最上部に表示される単語列）のみを表示した場合に相当する．本実験には，音声入力ソフトウェアを普段利用していない，20 代の 25 人の被験者（男性 14 人，女性 11 人）が参加した．また，**表 1** に，被験者のキーボード入力リテラシを示す．ここで，キーボード入力リテラシは，「キーボードで文章を入力した経験のある年数」で表した．

6.1 実験方法

音声訂正の機能の有効性を確認するため，各被験者に対して，以下の 3 つの課題を順に実施した．

課題 1

まず，キーボードとマウスを利用して，以下の 3 つの文を入力した．

(1) これらの地域の中学では非行が少ない
(2) この病院のベッド数は十五床です
(3) お年寄りからの注文にも備え二千個分の材料を確保

なお，課題 1 は，以降の音声による文入力の際の訂正処理を，被験者がスムーズに行えるよう，本実験でのキーボード，マウスを利用した日本語入力環境（かな漢字変換）に慣れることを主な目的として実施した．

課題 2

次に，音声を利用した文入力として，上記の 3 文を以下の条件でそれぞれ入力した．

(1) 通常の音声入力を使用して文入力
(2) 音声訂正を使用して文入力

被験者は，課題を行う直前に，通常の音声入力，音声訂正（基本機能と即時誤り訂正機能）についての説明を受け，それぞれの入力手段について，上記の 3 文とは別の 2 文を用いて練習を行った．ただし，本課題では，音声訂正の文入力に対する基本性能，効率性を評価するため，音声訂正の発話中休止機能については被験者に対していっさい説明はしなかった（あくまでも被験者が発話中休止機能について知らされていないだけであり，機能自体はシステムに含まれていた）．また，課題 2 において，(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設定した．本課題の終了後には，音声訂正の基本機能，即時誤り訂正機能それぞれについてのアンケート（6 項目 7 段階評価）を実施した．

課題 3

最後に，発話中休止機能の評価を目的とした課題を行った．本研究で提案した発話中休止機能は，2.3 節で述べたように，ユーザが発声中に音声認識処理を意図的に中断させることが可能となるものである．このような機能は，たとえば，ユーザが一息では発声しきれないような比較的長い文を入力する際などに特に有効に働くと考えられる．また，我々は，本機能の実際の利用状況として，自発的に発声された音声に対する入力インタフェースを想定しており，そのため，休止状態にするためのトリガとしては有声休止を採用している．したがって，発話中休止機能の評価を目的とした本課題においては，被験者が自発的に発声する状況になるべく近くなるように，頭の中にある文を発声して入力している場面を想定し，以下のような手順で実験を行った．

まず，発話中休止機能についての説明を行い，その後被験者は実際に本機能を練習した．次に，本課題についての説明を行った後，発話中休止機能を使うかどうかは被験者の自由という条件で，以下の 3 つの文の入力を行った．

(1) これから先の高いレベルを目指すにはもう少し読みの裏付けが必要だ
(2) 九十四年度当初は予算と同規模で景気対策を理

由に二年連続で高水準となった
(3) 全国のコンビニエンスストア，書店，私鉄・地下鉄の駅などで販売年間講読者への宅配も行う予定

ただし，被験者には，各々の文を入力する前に，提示した入力対象文を記憶して（頭の中に入れて）もらった．このとき，記憶に要する時間は特に制限せず，被験者の自由とした．文を入力する際の制約として，発声している最中には，入力対象文は確認できないこととした．これは，先に述べた本課題の主旨とは外れた「読み上げ入力」を避けるためである．ただし，発話中休止機能を使って休止状態にし，発声がなされていないときは，「思考中」の状態に相当すると考え，入力対象文を確認することは可能とした．また，本課題の終了後には，発話中休止機能について，また，音声訂正全般についてのアンケート（5 項目 7 段階評価）を実施した．

6.2 実験結果

表 2 に，課題 2 における，通常音声入力を使用したときの認識率，音声訂正を使用したときの訂正前・訂正後の認識率をそれぞれ示す．ここで，「通常音声入力」と「音声訂正（訂正前）」の認識率の間には，わずかな差がみられるが，本実験では，認識器としてはいずれも同一のものを用いているため，この差は実験上の誤差であると考えられる．実際の被験者実験においても，音声訂正はほとんどの誤りを訂正可能とし，高い訂正能力を示した．また，課題 2 における各入力手段ごとの，文を入力し終えるまでの 1 文あたりの平均入力時間を**表 3** に示す．なお，課題 1 におけるキーボードでの平均入力時間は 23.65 秒であった（ただし，課題 1 ではタイプ入力する練習も兼ねているので，入力時間に関する厳密な比較対象にはならない）．音声訂正では，通常の音声入力に比べて約 31%の入力にかかる所要時間を削減できていた．以上の結果より，音声訂正は優れた訂正能力を持ち，より効率的に文入力が可能であることが分かる．また，課題 2 の実験に際し，音声訂正に関して，被験者はたかだか 2 つの文を練習として入力しただけであり，これより音声訂正の基本機能は，入念な練習は不要で，インタフェースとしても直観的であったといえる．課題 3 の実験では，本機能が使用されたのは，全被験者の全発声のうちで約 61%（46/75）であった．また，課題 3 の 3 文を入力するにあたり，発話中休止機能をまったく使用しなかったのは全 25 人のうち 4 人であり，今回実験に参加した多くの被験者が本機能を利用していたことが分かった．

表 2 課題 2 における認識率
Table 2 Word accuracy in task 2.

	認識率 (%)
通常音声入力	86.12
音声訂正（訂正前）	85.03
音声訂正（訂正後）	97.70

表 3 各入力手段の平均入力時間
Table 3 Average time of each input methods.

	平均入力時間 (sec.)
通常音声入力	14.41
音声訂正	9.94

図 8 に音声訂正使用後のアンケート結果を示す．図中，(A) は音声訂正の基本機能について，(B) は即時誤り訂正機能について，(C) は発話中休止機能について，図 (D) は音声訂正システム全般についてのアンケート結果となっており，各項目とも −3〜+3 の 7 段階尺度での評定を行った．評定値が全体的に高い傾向にあったのは，項目 1.，4.，7. で，音声訂正，各機能は直観的で分かりやすいインタフェースであったことが分かる．(A) の音声訂正の基本機能については，項目 1.，2.，3. ともに評定値が高く，候補を選択するだけで誤り訂正することの有効性，効率性が示された．(B) の即時誤り訂正機能については，項目 5. が他に比べると低い評定値となった．これは，発声途中から逐次表示される候補をチェックし，訂正処理を行うことに対して，被験者が一定の負担や慌ただしさを感じたためである．しかしながら，この点に関しては，ある程度慣れることにより有効に使用できそうであるという意見は多かった（項目 6.）．また，今回の実験においては，即時誤り訂正機能により，多くの被験者は発声をしながらの訂正処理は難しいながらも，発声が終了した瞬間に，認識処理が実行中であっても即座に訂正処理に移行でき，最終的な文の入力を比較的早く完了させることができていた．(C) の発話中休止機能についても，(B) と同様に機能を有効に使えたか（項目 8.），という点については，低い評定値の被験者が他の項目と比べて多かった．その原因の 1 つとして，有声休止の発声をどの程度継続すればよいかが把握できず，不必要に長く言い淀んで戸惑ってしまう被験者がいたことがあげられる．現状のシステムでも，有声休止が検出されて発話中休止のモードに移ると，システム画面の右上に「発話中休止」というメッセージを表示していたが，初めのうちはこれを見逃す被験者がいた．教示やユーザへのフィードバックの仕方を工夫することで，より有効性が高くなると考えられる．有効に使えた被験者においては，本実験のような比較的

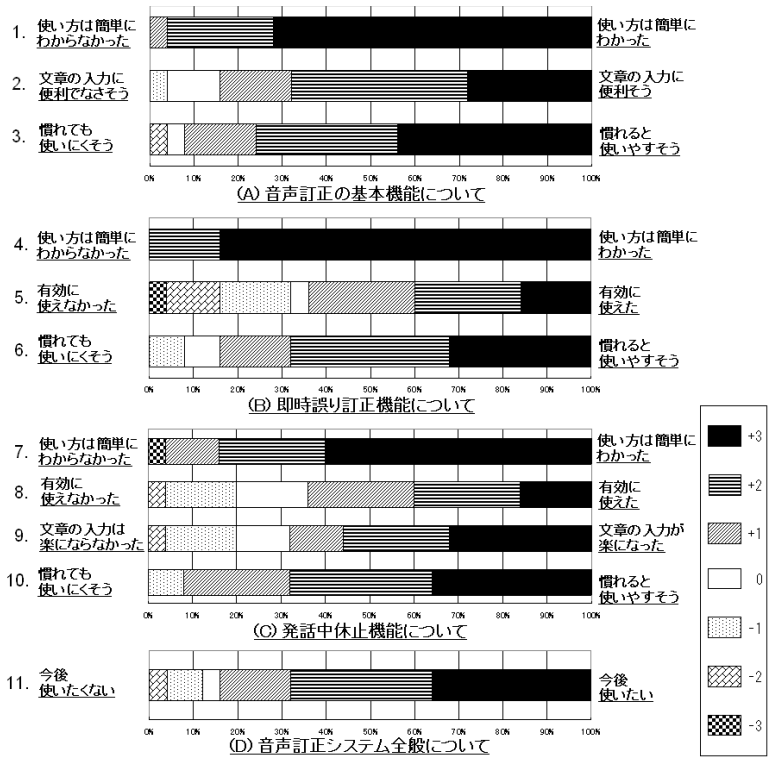

図 8 アンケートの集計結果
Fig. 8 Results of the questionnaire.

表 4 キーボード入力リテラシと音声訂正評価との関係
Table 4 Relationship between keyboard literacy and evaluation of speech repair.

経験年数	項目 11. の平均評定値
1〜3 年	2.43
3〜5 年	1.86
5〜10 年	0.91
10 年以上	−0.50

長い文に対する入力が格段に楽になったとの意見を得た（項目 9.）．項目 11. の音声訂正システム全般については，8 割以上の被験者が今後使いたいとの意見を得られた．また，被験者のキーボード入力リテラシごとの，項目 11. の平均評定値を**表 4** に示す．表 4 より，キーボード入力の経験年数が少ない被験者ほど高い評定値，すなわち音声訂正を今後とも使いたいとの印象を持っていたことが分かる．一方，キーボード入力リテラシの高い被験者に関しては，本実験では比較的低い評定値となっているが，タブレット PC や携帯型端末など，キーボードを使用できない状況においては，そのような被験者に対しても高い評定値が得られる可能性があると考えられる．

以上のアンケート結果からも本インタフェースの有効性が確認できた．

7. ま と め

本論文では，音声認識による認識誤りをユーザによって効率的かつ容易に訂正できる「音声訂正」という新たな音声入力インタフェースを提案した．本研究では音声認識における中間結果として confusion network を用いることにより，ユーザ側に認識結果の競合候補を効果的に提示でき，誤りのほとんどを訂正可能であることを示した．また，25 人の被験者による評価実験を行った結果，音声訂正により文入力が効率化され，被験者にとって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．

今後は，未知語への対処を行い，より効率的な誤り訂正処理について検討する予定である．本研究は「音声補完シリーズ」[14] の第 5 弾に位置付けられるが，これから有声休止以外の非言語情報も積極的に取り入れ，音声ならではの機能を持った，さらに使いやすい音声入力インタフェースを実現していきたいと考えている．

参 考 文 献

1) 遠藤 拓，寺田 実：音声入力における対話的

候補選択手法，インタラクション 2003 論文集，pp.195–196 (2003).
2) 安藤彰男，今井 亨，小林彰夫，本間真一，後藤淳，清山信正，三島 剛，小早川健，佐藤庄衛，尾上和穂，世木寛之，今井 篤，松井 淳，中村章，田中英輝，都木 徹，宮坂栄一，磯野春雄：音声認識を利用した放送用ニュース字幕制作システム，信学論，Vol.J84-D-II, No.6, pp.877–887 (2001).
3) Karat, C.-M., Halverson, C., Horn, D. and Karat, J.: Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems, *Proc. CHI'99*, pp.568–575 (1999).
4) Mangu, L., Brill, E. and Stolcke, A.: Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Network, *Computer Speech and Language*, Vol.14, No.4, pp.373–400 (2000).
5) Schwartz, R. and Chow, Y.L.: The *N*-Best Algorithm: An Efficent and Exact Procedure for Finding *N* Most Likely Sentence Hypotheses, *Proc. ICASSP'90*, pp.81–84 (1990).
6) Ortmanns, S., Ney, H. and Aubert, X.: A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition, *Computer Speech and Language*, Vol.11, No.1, pp.43–72 (1997).
7) 李 晃伸，河原達也，堂下修司：単語トレリスインデックスを用いた段階的探索による大語彙連続音声認識，信学論，J82-D-II, No.1, pp.1–9 (1999).
8) 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).
9) 緒方 淳，有木康雄：大語彙連続音声認識における最ゆう単語 back-off 接続を用いた効率的な *N*-best 探索法，信学論，Vol.84-D-II, No.12, pp.2489–2500 (2001).
10) 緒方 淳，後藤真孝：単語誤り最小化音声認識のための N-best 探索手法，日本音響学会 2004 年秋季研究発表会講演論文集，1-P-17, pp.195–196 (2004).
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) 緒方 淳，有木康雄：日本語話し言葉音声認識のための音節に基づく音響モデリング，信学論，Vol.J86-D-II, No.11, pp.1523–1530 (2003).
13) 河原達也，住吉貴志，李 晃伸，武田一哉，三村正人，伊藤彰則，伊藤克亘，鹿野清宏：連続音声認識コンソーシアム 2000 年度版ソフトウエアの概要と評価，情報処理学会研究報告，2001-SLP-38-6 (2001).
14) 後藤真孝：非言語情報を活用した音声インタフェース，情報処理学会研究報告，2004-SLP-52-7, pp.41–46 (2004).

(平成 18 年 3 月 7 日受付)
(平成 18 年 10 月 3 日採録)

緒方 淳（正会員）

1998 年龍谷大学理工学部電子情報学科卒業．2000 年同大学大学院修士課程修了．2003 年同大学院博士後期課程修了．同年産業技術総合研究所入所し，現在に至る．博士（工学）．音声認識，音声インタフェースに関する研究に従事．2000 年度日本音響学会粟屋潔学術奨励賞，2001 年度電子情報通信学会学術奨励賞，WISS2004 ベストペーパー賞，2006 年度情報処理学会山下記念研究賞，WISS2006 ベストペーパー賞各受賞．日本音響学会，電子情報通信学会各会員．

後藤 真孝（正会員）

1993 年早稲田大学理工学部電子通信学科卒業．1998 年同大学大学院理工学研究科博士後期課程修了．同年電子技術総合研究所（2001 年に独立行政法人産業技術総合研究所に改組）に入所し，現在に至る．2000 年から 2003 年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，2005 年から筑波大学大学院システム情報工学研究科助教授（連携大学院）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持つ．1997 年情報処理学会山下記念研究賞（音楽情報科学研究会），2000 年 WISS2000 論文賞・発表賞，2001 年日本音響学会粟屋潔学術奨励賞・ポスター賞，2002 年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 年日本音楽知覚認知学会研究選奨，2003 年インタラクション 2003 ベストペーパー賞，2005 年情報処理学会論文賞等 18 件受賞．電子情報通信学会，日本音響学会，日本音楽知覚認知学会各会員．