社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

音声訂正：“CHOICE” on Speech

緒方　　淳†　　後藤　真孝†

† 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: †{jun.ogata,m.goto}@aist.go.jp

あらまし　本研究では，認識誤りを選択操作だけで訂正することが可能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る．競合候補の個数はその区間の曖昧さを反映し，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．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可能となる．実験の結果，音声訂正は誤りのほとんどを訂正することができ，本インタフェースの有効性を確認した．

キーワード　音声認識，音声入力インタフェース，誤り訂正，confusion network，有声休止

Speech Repair: “CHOICE” on Speech

Jun OGATA† and Masataka GOTO†

† National Institute of Advanced Industrial Science and Technology (AIST)
1–1–1, Umezono, Tsukuba, Ibaraki 305-8568, Japan
E-mail: †{jun.ogata,m.goto}@aist.go.jp

Abstract　In this paper, we propose a novel speech input interface function, called “*Speech Repair*”, where recognition errors can be easily corrected by selecting candidates. Along the speech input, this function shows not only the usual 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. To overcome the difficulty of generating appropriate candidates, we adopted a *confusion network* that can condense a huge internal word graph of a large vocaburaly continuous speech recognizer. In our experiments, almost all recognition errors were repaired and the effectiveness of Speech Repair was confirmed.

Key words　speech recognition, speech input interface, error correction, confusion network, filled pause

1. ま え が き

音声認識技術を改良してどんなに認識率を上げていったとしても，人間にとって，常に明瞭で曖昧性のない発声をし続けることは極めて困難である以上，認識率は決して 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．例えば，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，二人の分業を前提とし，一人が誤認識箇所を発見してマーキングし，もう一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，タイピングして修正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) も常に列挙されていく．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，常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欲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。

以下，2. 章において本研究にて提案する「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. 章で音声訂正の性能評価をして，その有用性を確認する．最後に 6. 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) をユーザに一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 つの手続きを取る必要があった．

（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，これらを一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 (音声入力開始と共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」のリストを常に表示する．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

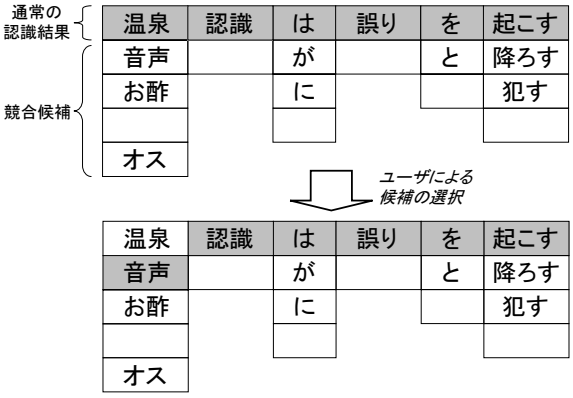

図 1 選択するだけで誤りの訂正ができる音声訂正インタフェース (「音声認識は誤りを起こす」という発声が誤認識された例)

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」という一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あまりに大量にあり過ぎて，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挙げられていた．

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，既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欲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) を，発話中休止機能のトリガーとして採用した．人間同士の対話においても，相手に少し待って欲しいときや，喋っている最中に考え事をするときなどに，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，単語グラフが挙げられる．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補提示は不可能である．

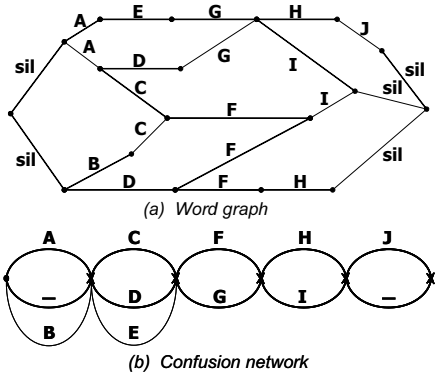

図 2 単語グラフと confusion network の模式図

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 [4]，語彙数が増えるにしたが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 [5] 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」[5] において導入された途中結果であり，過去に，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 2-a) を音響的なクラスタリングによりリニアな形式 (図 2-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 2-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 [5].

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ー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．

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

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，すなわち全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．例えば，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取れるように，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 (例えば図 4(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共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等) 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 [6] 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 4-(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．例えば，図 4-(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，既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，更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 [4]，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 [7] 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

3. 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 3 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 (LAN) 上の複数の計算機で

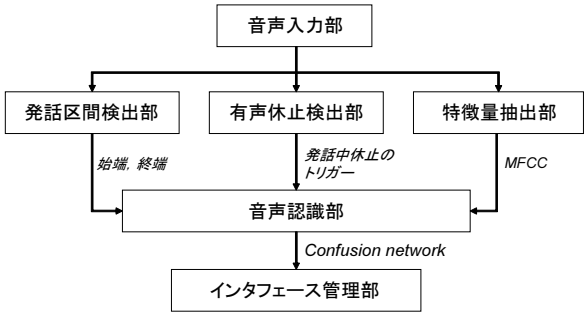

図 3　全体の処理の流れ

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) [8] を用いた．

処理の流れについて説明する．まず，マイクロフォン等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ーとして利用される．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．

4.2 実　行　例

図 4 に発話中休止機能を利用しない場合の表示画面を，図 5 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，さらに一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，文字の色を変えてわかりやすく表示している)．

5. 実　　験

音声訂正の基本性能を評価した結果を示し，実装したインタフェースの運用結果を述べる．

5.1 音声訂正の基本性能

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．ここでは，読み上げ音声，話し言葉音声それぞれ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) [9]，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 20K) [10] を用いた．評価用データは，

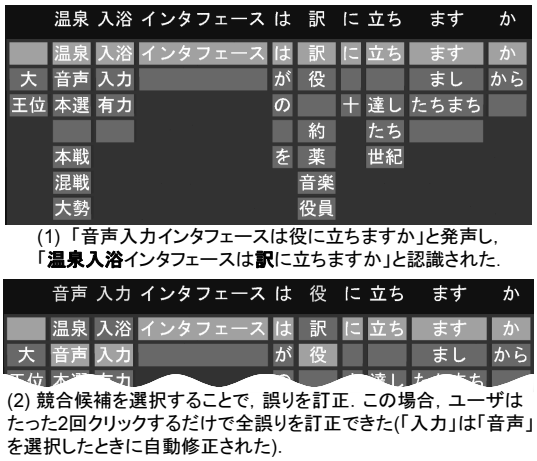

図 4　発話中休止機能を利用しない場合の画面表示例 (「音声入力インタフェースは役に立ちますか」という文章を発声)

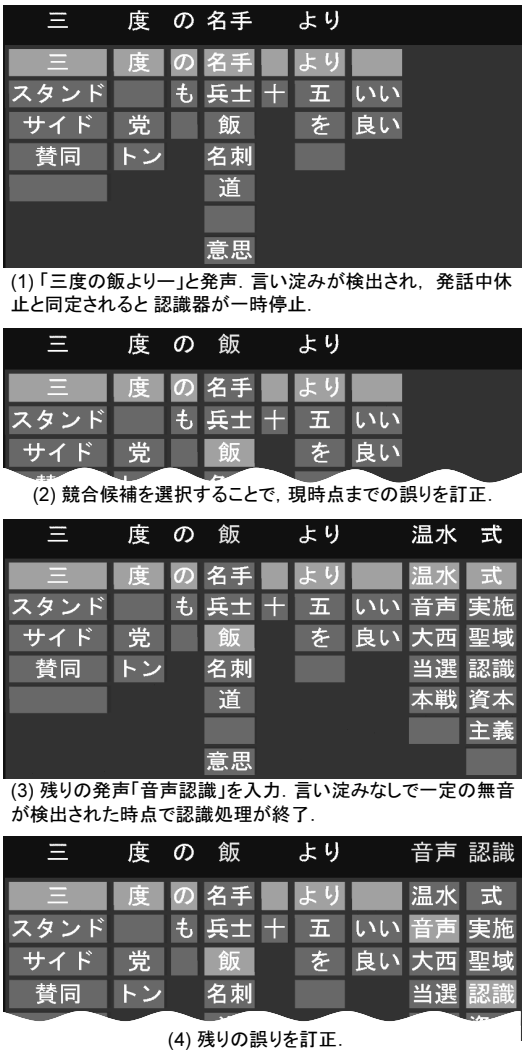

図 5　発話中休止機能を利用した場合の画面表示例 (「三度の飯より音声認識」という文章を発声)

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 名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 14K) を用いている．評価用データは，男性話者 4 名の学会講演中の 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，例えば 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 6 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わかった．一方，CSJ においても，通常の認識性能が JNAS に比べて約 10%低いにも関わらず，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わかった．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わかった．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挙げられる．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，更なる訂正性能の改善が期待できる．

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わかった．ただし，実験でも示されたように，音声認識システムの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．

5.2 音声訂正の運用結果

実際に，4 人のユーザに新聞記事の文章を読み上げてもらい，本インタフェースにより訂正処理を行ってもらった．どのユーザも，提示される競合候補に混乱されることなく，適切に訂正処理が行えることを確認した．言い淀みによる発話中休止機能も適切に使用され，特に長い文章を入力する場合は，本機能を使用すれば入力の際の労力が軽減されたとの感想を得た．また，使用方法も選択のみの操作で単純であり，GUI も直感的でわかりやすいと評価された．実際に，他人が使用している様子を見たユーザが，訓練せずに即座に使用できることがわかった．

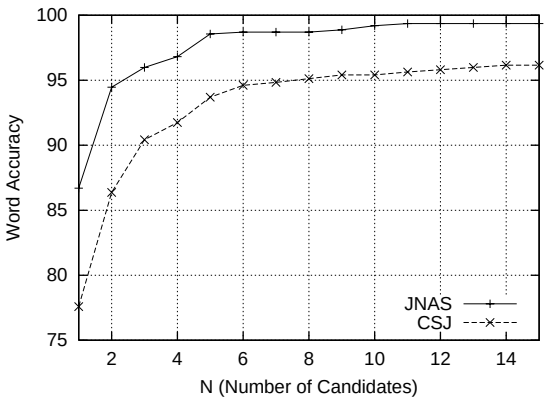

図 6　提示する候補数の上限を変えたときの訂正後の認識率 (最終的な音声入力成功率)

6. ま と め

本稿では，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，発話中でもリアルタイムに選択訂正が可能であり，音声訂正が使いやすく効果的であることがわかった．

今後は，訂正に要する作業負荷や作業速度などに関する定量的な評価，未知語への対処を行っていく予定である．また，言い淀み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．

文　献

[1] 遠藤 他: “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 他: “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] C-M.Karat, et al: “Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems”, Proc. CHI’99, pp.568-575, 1999.

[4] 李 他: “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, 1, pp.1-9, 1999.

[5] L.Mangu, et al: “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.

[6] 後藤 他 : “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.

[7] 緒方 他: “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.

[8] 後藤 他. “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.

[9] 緒方 他: “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.

[10] 河原 他: “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”, 情処研報, 2001-SLP-38-6, 2001.