音声言語情報処理 38－5
（2001．10．19）

認識信頼度を用いた誤認識修正支援エディタの検討

遠藤 拓、ナイジェル ワード、寺田 実
東京大学大学院 情報理工学系研究科
知能機械情報学専攻
{entaku,nigel,terada}@sanpo.t.u-tokyo.ac.jp

あらまし： 音声認識技術を用いたテキストエディタ (音声入力エディタ) は入力速度が速いものの、誤認識の発見と修正に時間がかかってしまう。このためにキーボード入力に比べて、文書作成に要する時間に差がなくなっている。本研究では誤認識修正の効率を向上させるために、Confidence Measure を用いて誤認識である可能性の高い単語をハイライト表示してユーザに示す機能とそのハイライト間を TAB キーを押して簡単にジャンプできる機能を備えたエディタを提案する。またこれらの機能が、音声認識エンジンの認識率・Confidence Measure の質・文書中に残るエラーに対するユーザの寛容性に関して、どの程度有効か予測するモデルを提案する。

キーワード： Confidence Measure, ディクテーション, エディタ, 誤認識, 修正, 支援

Displaying Speech Recognizer Confidence Information to Support User Correction of Misrecognitions

Taku Endo, Nigel Ward, Minoru Terada
Department of Mechano-Informatics,
School of Information Science and Technology,
University of Tokyo

Abstract: When dictating with speech recognition, most of the time is spent correcting errors. Since it is possible to predict which words are probably in error, using various confidence measures, we propose an editor with two new functions: display probable errors in red, and enable jumping to the next probable error with the tab key. Simple experiments suggest that these functions can be valuable. We also propose a model of editing costs and give experimental support for its validity, then use this model to enable prediction of the conditions under which these functions are useful, depending on three parameters: speech recognition rate, confidence measure quality, and user tolerance for uncorrected errors.

Keyword: Confidence Measure, Dictation, Editor, Error Correction, Time Cost, Cognitive Model

1　はじめに

1.1　問題

近年、音声ディクテーションソフトが普及しつつあるが、市販のソフトを購入した者の約半数が使わなくなってしまうと言われている。何故だろうか。そもそも、発声という行為は人間にとって日常的に行われているものなので、音声入力はキーボード入力に比べて疲労が少ないはずである。また、入力速度という観点から見てもキーボード入力より速い。しかし、ディクテーションによる文書作成では入力よりも誤認識の修正に多くの時間がかかってしまうので、さほど速くはないのが実情である (表１)。

入力方法	時間 (min)
キーボード入力	6’09〜8’25
音声入力 (修正有)	4’35〜5’12
音声入力 (修正無)	1’55〜2’05

表 1: 約 700 字の日本語の文章入力における音声入力とキーボード入力の比較

Karat ら [1] は文書作成タスクにおいて、単純な音声入力とキーボード入力の比較を行っている。そこで、誤認識の修正に時間がかかる理由として訂正そのもの以外に次のことを挙げている。

- ユーザが誤認識箇所を発見するのに時間がかかる。
- 発見した誤認識箇所へカーソルを移動させるのに時間がかかる。

1.2　低信頼度単語表示機能の提案

誤認識箇所をユーザに示すことと誤認識箇所へのカーソル移動を容易にすることで修正速度が向上するはずだと考える。

ところが、現状では誤認識箇所を完璧に示すことは不可能である。しかしながら、近年研究されている Confidence Measure の手法を用いれば、誤認識である可能性の高い箇所を示すことは可能である。

本研究では、入力に対する音声認識の信頼度（Confidence Score）が与えられたと仮定し、

- 誤認識である可能性の高い単語を入力者に発見しやすくするためにハイライト表示する機能
- ハイライト間を簡単にジャンプできるコマンド機能。

この２つの機能を実装したテキストエディタを提案する。また、この機能が認識システムの認識率、Confidence Measure の精度との関係の中で、どの場合にどの程度有効であるか検証する。

そのために、誤認識単語の修正をモデル化して低信頼度単語強調表示機能の有効性を判別する式を立て、実験によってその式の妥当性を確認した。また、現在の Confidence Measure を用いれば有効であることも調べた。

2　修正支援のモデル

2.1　閾値による分類

Lee [2] によると、音声認識システムで認識された単語の Confidence Score が図 1 のように分布するのが一般的である。これらの単語を図 2 のように、ある一定の閾値を決めることで以下の４つに分類することができる。
本エディタでは、閾値よりスコアが低い単語をハイライト表示することにする。

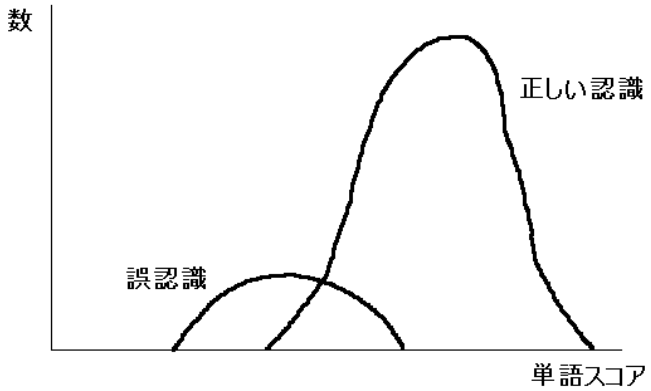

図 2: Confidence Measure 後の単語スコア分布

分類 A　：閾値よりスコアが低く、かつ誤認識の単語
分類 B　：閾値よりスコアが低いけれども正しい認識の単語
分類 C　：閾値よりスコアが高いけれども誤認識の単語
分類 D　：閾値よりスコアが高い、正しい認識の単語

理想的には全ての認識単語がＡとＤに分類されることが望ましい。

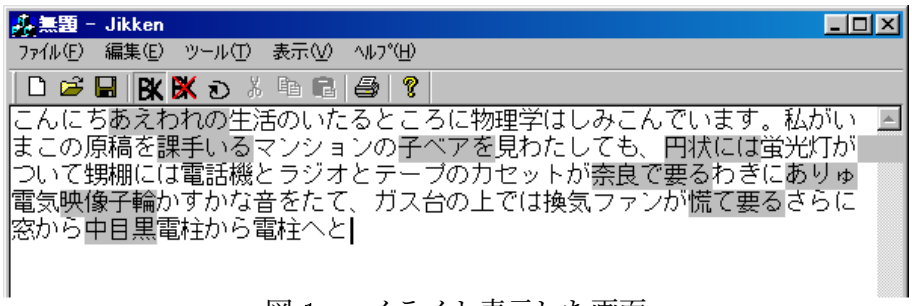

図 1: ハイライト表示した画面

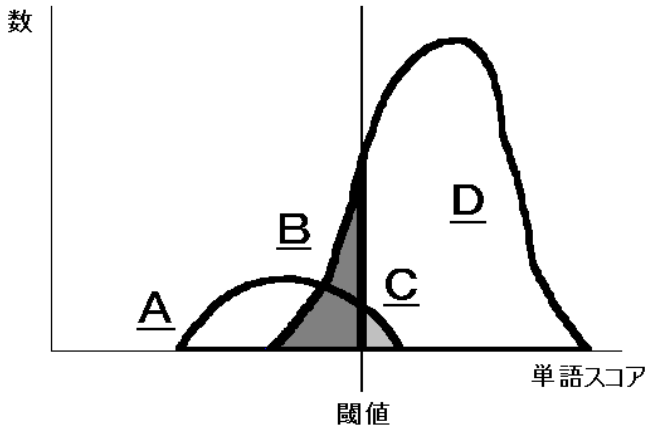

図 3: 認識単語の分類

2.2 誤認識修正のコスト

Karat[1] らは、文書作成の際には、少しずつ入力してはその都度修正を繰り返すやり方よりも、全文入力後に一気に修正を加えた方が効率がよいと述べている。そして、ディクテーションソフトの利用者も後者の方法を使用しているようです。

本エディタ使用者の目的は短時間で質の良い（誤りの少ない）文書を作成することにある。ハイライト表示機能を用いて修正時間を大幅に短縮するために、ハイライト表示機能を用いて修正時間を大幅に短縮するために、ハイライトのかからない誤認識単語（Cに分類される単語）には修正を加えないものとして誤認識修正のコストを考えてみる。誤認識単語の修正は以下の手順で行われる。

手順 a ：誤認識の発見

手順 b ：カーソルの移動

手順 c ：削除と入力

参考文献 [3] にはテキストエディタでの各操作にかかる時間を調べた実験の結果が記載されている。書く手順に要する時間はその結果から引用することにする。

2.2.1 修正支援機能のないエディタの修正コスト

誤認識修正を支援する機能（誤認識単語ハイライト表示機能とハイライト間のジャンプ機能）を備えていない通常のテキストエディタで誤認識を修正する場合、1 単語の修正にかかるコストの平均は以下の式で表すことができる。(H:Highlight, N:No highlight)

$$Cost_N = T_{Na} + T_{Nb}(A + C) + T_{Nc}(A + C) \quad (1)$$

ここで、T_{Na},T_{Nb},T_{Nc} は 1 単語の修正に対してそれぞれの手順 a,b,c で要する時間 (sec) である。先述のように参考文献 [3] から値を引用すると、

$$T_{Na} = 0.8 \quad (sec) \quad (2)$$

$$T_{Nb} = 1.0 \quad (sec) \quad (3)$$

$$T_{Nc} = 4.0 \quad (sec) \quad (4)$$

となる。

2.2.2 修正支援機能のあるエディタの修正コスト

本研究で作成する、ハイライト表示とハイライト間ジャンプ機能を備えたテキストエディタを用いた場合の 1 誤認識単語にかかる平均の修正コストは、

$$Cost_H = T_{Ha}(A + B) + T_{Hb}(A + B) + T_{Hc}A + lC \quad (5)$$

と書くことができる。ハイライトがない場合と同様、T_{Ha},T_{Hb},T_{Hc} は a,b,c 各手順に要する時間 (sec) である。また、最後の項は、ハイライトがないために残ってしまう誤認識単語をエラーコストとして加えたものである。

変数 l は誤認識のままエラーとして残る 1 単語をに対するコストである。言い換えるなら、誤認識

単語１語を修正するのに費やしてもよい時間、修正時に１単語をエラーとして残すことで短縮される時間である。友人への e-mail などといった、多少のミスがあっても許される文書、ミスの少なさよりも入力速度が重視される場合には l の値を小さく設定し、逆にレポートや論文のような、入力時間が長くなったとしてもミスが少ないことが望ましい文書の場合には l の値を大きく設定する。

ディスプレイを見て１つの単語が誤認識かどうか判断する時間と誤認識の単語に修正を加える時間はハイライトがない場合の修正とほぼ同じであるから、T_{Ha},T_{Hc} については

$$T_{Ha} = T_{Na} = 0.8 \quad (sec) \tag{6}$$

$$T_{Hc} = T_{Nc} = 4.0 \quad (sec) \tag{7}$$

が成り立つ。

しかし、ハイライト間のジャンプ機能を付加したことで T_{Hb} は T_{Nb} よりも短縮される。これも参考文献 [2] からおよその値を求めると、

$$T_{Hb} = 0.2 \quad (sec) \tag{8}$$

となる。

ユーザにとって、これらのコストが

$$Cost_H \leq Cost_N \tag{9}$$

となる場合に本エディタの低信頼度単語強調表示機能が有効であると言えるはずである。

2.3 閾値の設定

本節では 2.2 節で述べた、単語を分類するための閾値の設定方法について論じる。

2.3.1 ハイライトの信頼性

Confidence Score を用いて単語を分類するので、実際には正しく認識された単語をハイライト表示してしまう (B) ことや、逆に、誤認識単語であるにもかかわらずハイライト表示がされない場合 (C) がある。そこで、ハイライトの信頼性 R を以下のように定義する。

$$R = 1 - kC - B \tag{10}$$

k はＢとＣに関する重み付けの定数である。k の値が大きいほどＣに重点がおかれる。
この R が最大となるように閾値を設定する。

2.3.2 k の算出

本研究で作成するエディタでは 2.2 節で論じた修正コスト $Cost_H$ を最小とし、前節で論じたハイライトの信頼性 R を最大としたい。したがって (5) 式、(10) 式より

$$B : C = l : T_{Ha} + T_{Hb} \tag{11}$$

$$B : C = k : 1 \tag{12}$$

である。この２式から k は

$$k = \frac{l}{T_{Ha} + T_{Hb}} \tag{13}$$

と求まる。つまり、閾値の設定はユーザがどのような文書を作成するかによって決めることになる。

2.3.3 低信頼度単語強調表示機能の有効性の予測

前述したが、本エディタが有効となるのは、

$$Cost_H \leq Cost_N \tag{14}$$

となる場合である。
式 (1)〜(9) を代入して整理すると、

$$-lC - B + 5C + 0.8 \geq 0 \tag{15}$$

この式では [2] から引用した数値のために、式から A が消えてしまった。

次に、A,B,C,D の数が予め既知である例文を用いて実験を行い、この式がユーザの評価と合致するか調べた。

3 実験

3.1 誤認識表示の有効性の検証

低信頼度単語強調表示機能を付加すると修正時間が短縮される可能性があることを確認するために、以下の実験を行った。誤りを含む例文（日本語 120 単語程度）を用意し、その誤り箇所を 100 %の信頼性でハイライト表示したものとハイライト表示しないものをそれぞれ被験者５人に修正させ、その修正時間を比較した。

なお、この時のエディタはハイライト間のジャンプ機能は備えなかった。

被験者	ハイライト有の修正時間 (sec)	ハイライト無の修正時間 (sec)
A	378	386
B	335	377
C	371	387
D	375	411
E	324	262

表 2: 120 単語程度の例文の修正にかかる時間

この結果から、誤認識箇所をハイライト表示することが大方の人にとって有効であると確認できた。

3.2 低信頼度表示の有効性の検証

次に、誤り箇所のハイライト表示の信頼性が 100 %ではない場合の有効性についての実験を行った。この実験では、本研究のエディタと普通のテキストエディタ、すなわち、「低信頼度認識単語強調表示機能」と「ハイライト間ジャンプ機能」がある場合とない場合の誤認識修正速度の比とハイライトの信頼性 R との関係を調べた。

3.2.1 方法

手順は以下の通りである。

70 語程度の例文 4 つに誤りを含ませ、強調表示する単語をランダムに選ぶことで全単語を A,B,C,D に分類した。A,B の単語をハイライト表示する場合としない場合のそれぞれで 10 人の被験者に修正させ、その速度を計測。

ハイライト表示する場合、ハイライト表示のみの場合と更にハイライト間のジャンプ機能を使える場合の 3 通り行った。l の値は例文の内容が e-mail 並であったことから 5(sec) に決めた。l と (13) 式から定数 k を得て、k と A,B,C,D の単語数から R を求めて速度の比と R の関係を調べた。

被験者 10 人には合計 44 回の修正を行ってもらった。同じ誤り方をした同じ文章を 3 通りの方法で行い、更に誤り箇所を覚えてしまわぬように別な誤り文章についても同じ同様の修正を行ってもらった。

3.2.2 結果

以下に修正速度比とハイライトの信頼性 R のグラフ (図 3、図 4) を示す。R と修正速度比（ハイライト有の場合の修正速度／ハイライト無の場合の修正速度）の相関係数はそれぞれ 0.195, 0.488 であった。

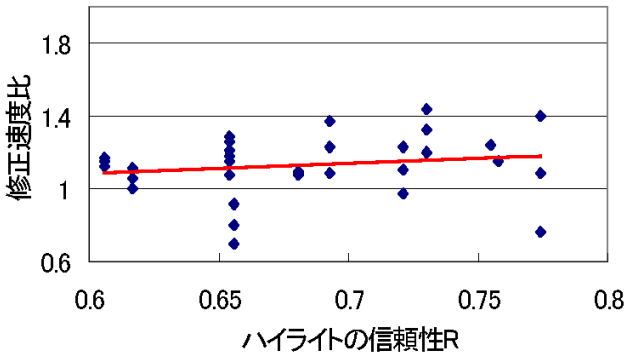

図 4: ハイライトの信頼性と修正速度比

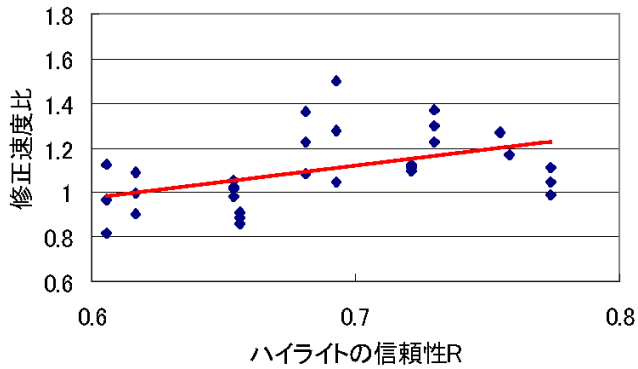

図 5: ハイライトの信頼性と修正速度比 (ハイライト間ジャンプ機能有)

3.2.3 考察

キーボードでのタイピング入力速度には個人差がある。本研究では誤認識の修正にキーボードを用いるので、ハイライトの有無による修正速度の変化を見るのに、単純な速度ではなくハイライトがない場合を基準とした各個人の修正速度の比を使うことにした。しかしそれでも、環境の影響、被験者の体調、偶然等による実験値のばらつきは大きなものであったことが結果のグラフからもうかがえる。

直観的には、ハイライトの信頼性 R と修正速度比は比例関係にあると考えられる。しかし、ハイライト表示機能だけが実装されている場合の R と修正速度比の相関係数は小さかった。図3、図4からわかるように被験者ごとの実験値のばらつきが大きかったのが原因であると思われる。

被験者全体の修正速度比の平均は表3のようになった。ハイライト表示によって修正速度が向上することがわかる。ハイライト間ジャンプ機能がある場合はハイライト表示のみの場合よりも遅くなってしまった。本来、ジャンプ機能によってカーソル移動の手間が少なくなるので修正速度が速くなるはずである。これは被験者がハイライト間ジャンプという、日ごろ使い慣れていない操作をすることになるので逆に時間がかかってしまったことが原因と考えられる。

エディタの機能	修正速度比
ハイライト機能のみ	1.13
ハイライト間ジャンプ機能も有	1.10

表 3: 修正速度比の平均値

3.3 有効性判別式の検証

この実験では2.3節で述べた有効性の判別式(15)が実際にユーザの評価と合致するか調べた。

3.3.1 方法

手順は以下の通りである。

手順 1 実験1と同様に誤りを含んだ例文を用意し、各単語を A,B,C,D に分類。

手順 2 A,B の単語をハイライト表示し、ハイライトされている部分だけを被験者に修正させた。

手順 3 誤認識が残る (C) 修正後の文を被験者に「e-mail として」、「論文として」それぞれの評価を5段階でさせた。

手順 4 有効性の判別式(15)とユーザの評価が一致するか確認した。

この実験では R の決定のために、l を e-mail: 3(sec)、論文: 20(sec) とした。この値は実験前の事前調査で大方の予想を立てて計算した値である。実験は 10 人の被験者に合計 44 回の修正を行ってもらった。

3.3.2 結果

この実験の結果は以下の表のようになった。表中の有効性はYが本エディタが有効であること、Nが有効でないことを示している。

l	R	認識率	有効性	評価
3	0.658	0.740	Y	3
20	0.303	0.740	N	2
3	0.716	0.712	Y	4
20	0.410	0.712	N	3
3	0.733	0.797	Y	3
20	0.621	0.797	Y	2
3	0.658	0.740	Y	3
20	0.303	0.740	N	1

表 4: 予測した有効性と l、R、認識率の表 (一部)

また、被験者の評価は（ 全然ダメ・ダメ・どちらでもない・まあよい・とてもよい）の5段階でとった。被験者の評価をどこで2つに分けるか考える上で、$Cost_H$ と $Cost_N$ の比較の際には両者の値が等しい場合もハイライトは「有効である」としたので、評価3の「どちらでもない」も「よい」の方で判断することにする。したがって、評価1と2は「よくない」、3, 4, 5は「よい」とする。その結果、全体の 72 %で式 (15) の有効性と被験者の評価が合致した。

被験者の評価 式による判別	1	2	3	4	5
Y (個)	6	9	13	33	3
N (個)	4	11	9	0	0

表 5: 式による有効性の判別とユーザ評価

3.3.3 考察

72 %の割合で本エディタの有効性の式(15)がユーザの評価と合致したので、この式がエディタの有効

性を測るための式として、ほぼ利用可能であると言える。ただし、T_{Ha},T_{Hb},T_{Hc},T_{Na},T_{Nb},T_{Nc} の値を正確に測定すればこの割合はもっと向上すると思われる。また、文章全体の残留エラーが少なくても、キーワードがエラーになるとユーザの評価が低くなってしまったことも式と評価の合致率の低下に結びついたといえる。

本実験の結果 (表 4) から、ハイライトの信頼性 R が低い場合には被験者の評価も低くなっていることがわかる。したがって、本エディタの有効性を計る指標としてこの R の値も利用できると思われる。

4 Confidence Measure の適用可能性

4.1 Confidence Measure について

Wessel ら [4] は音声認識結果の Confidence Score をワードグラフ、N-best リスト、音響的な安定性、仮説の密度のそれぞれ 4 つから求め、それらの比較を行っている。Wessel らの Confidence Measure の評価方法は次のようである。認識結果の Confidence をそれぞれの方法から算出し、ある閾値を基にして単語ごとに正しいか誤りかのタグをつけていく。次にそれらを実際の正誤と比較し、誤認識単語が正しいと判別された割合（false acceptance rate）と正しい認識単語が誤りと判別された割合（false rejection rate）をグラフにしている。閾値を変化させるとこの 2 つの値はトレードオフに変化し、曲線を描く（Detection-Error Tradeoff Curves）。図 8 の曲線はそのグラフの 1 つである。

4.2 有効性の検証

有効性の判別式 FAR(false acceptance rate), FRR(false rejection rate) は 2.2 節の分類で表すとそれぞれ、

$$F_{rr} = \frac{B}{B+D} \tag{16}$$

$$F_{ar} = \frac{C}{A+C} \tag{17}$$

となる。また、認識システムの認識率を r とするならば $r = B + D$, $1 - r = A + C$ である。これらを用いて (15) 式を書きかえると、

$$F_{rr} \leq \frac{1-r}{r}(5-l)F_{ar} + \frac{0.8}{r} \tag{18}$$

となる。図 6・図 7 は認識システムの認識率がそれぞれ 80 %・90 %の時の各 l によってどの F_{rr},F_{ar} なら本機能が有効か予測するグラフである。DET 曲線が線の左側に入るならば、低信頼度単語強調表示機能が有効であるといえる。当然ながら、ユーザがエラーに対して厳しいほど、高性能の Confidence Measure が要求される。

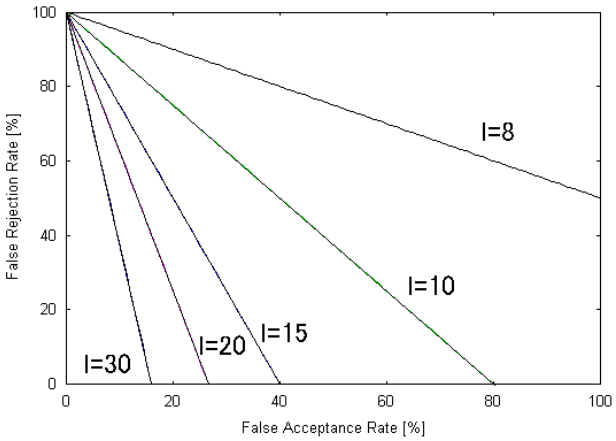

図 6: 本機能の有効性の予測 (認識率 80 %)

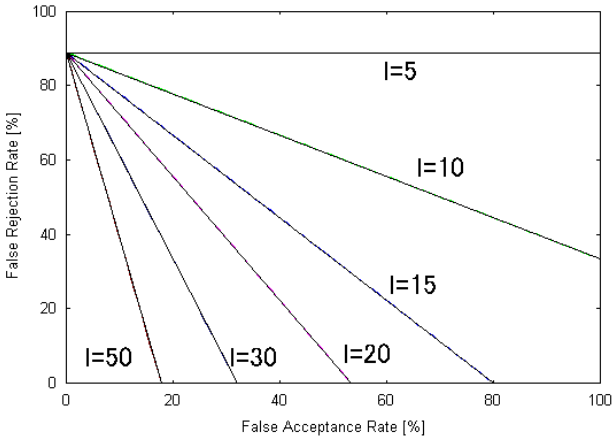

図 7: 本機能の有効性の予測 (認識率 90 %)

このグラフと Confidence Measure の DET Curve のグラフを重ね合わせることでその Confidence Measure のスコアを用いたときの各 l 毎の有効性が判別できる。図 8 は [3] に記載されている DET Curve と有効性判別のグラフを重ね合わせたものである。認識率は [3] に記述のある、NAB 64K コーパ

スの 88.9 %という値を用いた。よって、現在の Confidence Measure でも、e-mail はもちろん、フォーマルな文書に対しても低信頼度単語表示機能は役に立ちそうである。

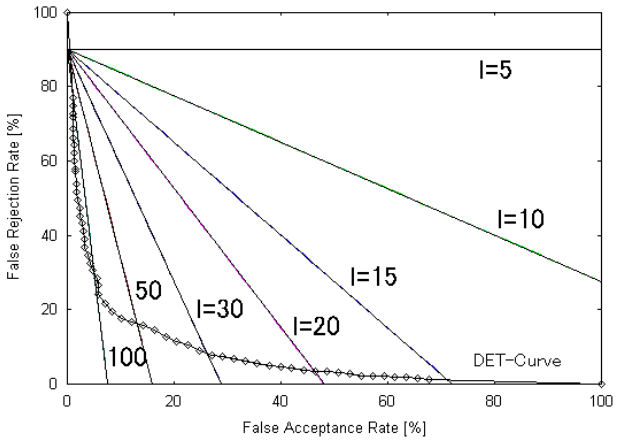

図 8: エディタの有効性と DET-Curve (認識率 88.9 %)

5 既知の問題点

§ 2 の実験では、ハイライト表示があることで修正速度が向上するとわかった。しかし実際には文中のハイライトの割合が過剰に多くなると、必ずしも修正速度は向上しないだろうと考えられる。事実、実験のデータから、ハイライトの割合と修正速度比のグラフは図 9 のようになった。

式 (5) のコスト式にはハイライト過剰時の修正速度減少に関するエラーコストの項が含まれていないため、図 8 でも FRR が大きい場合には l の値が大きい場合にも有効だと判別されてしまうが、FRR が 35 %以上だとこのモデルが使えない可能性もある。今後は修正速度も考慮した本エディタの有効性の判別式を考える必要があるだろう。

6 まとめ

本研究では誤認識修正の効率を向上させるために、誤認識である可能性の高い部分をユーザに示すこと機能（低信頼度単語表示機能）とそのハイライト間をジャンプできる機能を備えたエディタを提案し、その有効性を検証した。ユーザの誤認識の修正をモデル化し、本機能の有効性を判別する式を得た。実験の結果、ハイライト表示すること

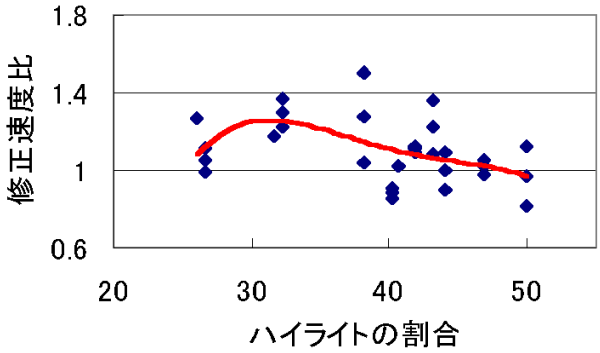

図 9: ハイライトの割合と修正速度比

で修正速度が向上すること、モデルから求めた判別式がユーザの評価と一致することが確認できた。さらに、この判別式を使えば、既存の Confidence Measure を用いた場合に、本機能を実装するエディタがどの程度有効であるかがわかるようになった。

参考文献

[1] C-M. Karat, C. Halverson, D. Horn and J.Karat: Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems. In *Proc. of CHI '99*, pp.568-575, 1999.

[2] Chin-Hui Lee: Statical Confidence Measures and Their Applications. In *Proc. of ICSP 2001*, pp.1021-1028, 2001.

[3] Stuart K. Card, Thomas P. Moran, Allen Newell: *The Psychology of Human-Computer Interaction*, Lawrence Erlbaum Associates, 1983

[4] Frank Wessel, et al : Confidence Measure for Large Vocabulary Continuous Speech Recognision. In *Proc. of IEEE Transactions on Speech and Audio Processing Vol,9*, pp.288-298, 2001