社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－SLP－55 (17)
2005／2／5

対話システムにおける言い直し・否定表現に着目した訂正発話の検出

矢野 浩利　北岡 教英　中川 聖一

豊橋技術科学大学　情報工学系
〒441-8580 愛知県 豊橋市 天伯町 雲雀ヶ丘 1-1
E-mail:hyano,kitaoka,nakagawa@slp.ics.tut.ac.jp

あらまし

近年、音声認識をベースとしたインターフェースを備えた製品の実用化が進んでいる。音声インタフェースの特有の問題として、コンピュータと人間が音声を通じてコミュニケーションを図る場合、誤認識は避けられない点がある。また、現在はインターフェースが未熟であるために、その回復が困難である。一般に誤認識が発生した場合のユーザの反応として、誤認識された部分を言い直したり、否定表現を用いることが挙げられる。したがって、それらを検出することにより誤認識からの回復が容易になると考えられる。

本稿では、言い直し判定をDPマッチングと重なり度、その組み合わせにより行った。否定表現の検出は、認識結果の事後確率と、単語終端におけるパワーの傾きに着目した方法を提案した。

パワーの傾きを用いることで判定性能の向上が得られることを示す。また、言い直し、否定表現の組み合わせにより訂正発話の検出性能が向上した。その結果、再現率0.864、適合率0.955でシステムが自身の誤りを検出できることを確認した。

キーワード　音声対話, 訂正発話, 言い直し発話検出, 否定表現検出

Detection of Correction Utterances Focused on Repetitions and Negative Expressions in a Dialog System

Hirotoshi Yano　Norihide Kitaoka　Seiichi Nakagawa

Department of Information and Computer Sciences, Toyohashi University of Technology
1-1 Hibarigaoka, Tenpaku-cho, Toyohashi, Aichi, 441-8580 Japan
E-mail:hyano,kitaoka,nakagawa@slp.ics.tut.ac.jp

Abstract

Recently, systems with interfaces based on speech recognition have been realized. When we communicate with computers through speech interface, misrecognition is inevitable. However, it is difficult to recover misrecognition because of the immaturity of the interface. Users often used repetitions of misrecognized parts and negative expression to deny the system's action according to the misrecognition. Thus, detection of such phenomena can make the recovery easier.

In this report, we propose to detect correction utterances detecting repetitions of previous utterances and negative expressions. We detect the repetitions using a Dynamic Time Warping-based method and a N-best hypotheses overlapping measure. We also propose a method to detect negative expressions using recognizer's intermediate hypotheses and the gradient of the power contour at the end of each word in the hypotheses. By combining these detection methods, we achieved 0.864 of detection rate and 0.955 of precision for correction utterance detection.

Keywords　Spoken dialog, correction utterance, repetition, negative expression

1 はじめに

近年、音声認識をベースとしたインターフェースを備えたシステムの実用化が進んでいる。しかしながらコンピュータと人間が音声を通じてコミュニケーションをはかる場合、誤認識は避けられない。また、現在はインターフェースが未熟であるために、誤認識からの回復が困難である。一般に、ユーザはシステムの誤認識に対して同じ内容の言い直し (繰り返し)、もしくは否定表現によって対処しようとする。しかし誤認識した発声内容をユーザが再発話しても一向に正しく認識されず何度も同じ誤認識を繰り返してしまうことが多い。またユーザが否定表現によりシステムに誤認識を伝えようとしても、その発話も他候補として認識されてしまうことがある。このような状況からなかなか抜け出せないとユーザの負担はかなり大きい。ここで、システムがユーザの言い直し、否定表現による訂正発話を検出できれば、誤認識からの回復が容易になると考えられる。

現在も言い直し (繰り返し) 音声の検出に関する研究はいくつかなされている。今井らは、未知語処理のための孤立単語の繰り返し音声検出の手法として、1. 認識候補の重なり度による識別手法、2. 認識尤度差による識別手法、3. パワーの時系列ベクトル間の距離による識別手法、の3通りの手法を提案しており、手法1と手法3を組合せることによって、recall・precision ともに約 90%の識別性能を得ている [5]。また、言い直しを検出するにあたって、ユーザが誤認識時にどのような言い直しをするかを知ることも必須である。平沢らは、システム確認が誤解を含む場合のユーザ応答については、もう一度同じことを繰り返す場合が多く、繰り返しのユーザ発話は、元のユーザ発話に比べてピッチ・継続時間長が大きく、発話速度の低下が見られると報告している [6],[7]。Oviatt らは、訂正発話では継続時間が長くなるが、パワーやピッチについては、大きな差は見られないと報告している [8]。また、Levow は、認識誤りの訂正と棄却誤りの訂正について言い直し発話の分析を行ない、認識誤り訂正の方が継続時間がより長くなると報告している [9]。また、Levow は韻律情報を利用した訂正発話における訂正箇所の特定を行っており、85%の判定精度を得ている [10]。Swerts らは、訂正連鎖において、エラーからより遠い訂正は近い訂正よりピッチ・パワーが大きく、継続長が長くゆっくりで、先行ポーズが長いと報告している [11]。また、山肩らは、F0 とパワーを用いて訂正発話の特徴を分析したところ、誤認識された発話と初回訂正発話の変化は有意ではなかったと報告している。しかし、変化のタイプによって被験者を分類し、再分析したところ、有意水準1%で変化が有意であったと報告している [12]。角谷らも訂正発話の韻律情報の分析を行っており、言い直し発話では、1回目の発話よりもユーザの声量、声の高さによる抑揚が大きくなっているが、声の高さ、発話速度については明らかな変化は見られなかったと報告している [1]。また、角谷らは言い直し判定の効果として、システム主導型対話システムにおいて、質問ごとに語彙・文法を用意し、言い直し判定結果により語彙・文法を設定することで、パープレキシティの減少を得ている [2],[3]。

本稿では、さらに否定表現を用いた訂正発話の検出を行うことを考える。ここでは、音声認識の事後確率とパワーの変化に着目した正確な否定表現の検出法を提案する。さらに言い直し (繰り返し) 発話検出と組み合わせることで、システムの自己誤り検出精度の向上をはかる。

System：目的地を設定してください
User：えーと、静岡県の**浜松インター**です
System：静岡県浜松駅に設定します
User(Repaired)：東名高速道路の**浜松インター**

図 1: 文発声における言い直しの例

2 訂正発話の分析

対話システムを使用して収集された、話者 10 名による全 512 発話を用いる [14]。タスクは図1に示したような自然な発話による目的地設定である。目的地を発話してもらい、システムが正しく認識するまで訂正入力してもらうものである。512 発話のうち 220 発話が訂正発話であった。その内訳を表1に示す。

表 1: 訂正発話の内訳

言い直し	否定表現	出現数	
一部の言い直し	なし	37	48
	あり	11	
全区間の言い直し	なし	113	144
	あり	31	
なし	あり		28

言い直し発話は 192 発話、否定表現が含まれる発話が 70 であった。認識を誤った部分のみの言い直しも可能としているが、話者の傾向としては全区間を言い直すものが 141 発話と多くを占めた。否定表現としては、「いいえ」「違います」と、これに似た表現を対象としている。図1のようにシステムの応答は「はい」「いいえ」といった発話を促すものではないが、ユーザには否定表現を使う傾向が見られた。言い直しや否定表現による訂正発話は 220 発話

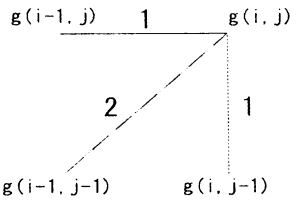

図 2: DP パス

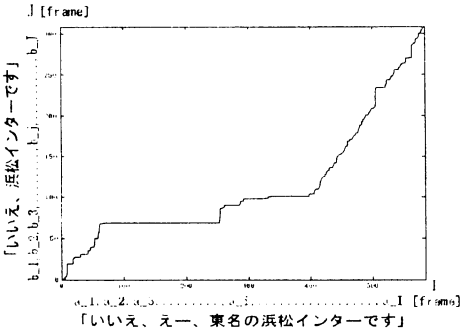

図 3: DP マッチングの例

(192+28) となった。

3 言い直し検出方法 [3]

3.1 DP マッチングによる判定

言い直した音声 (例：「東名高速道路の浜松インター」) が直前に発声した音声 (例：「えーと、静岡県の浜松インターです」) に含まれているかどうかを調べ、言い直しかどうかを判定する。

方法としては、音声特徴量における直前の発話音声 A の i フレーム目 a_i と現発話音声 B の j フレーム目 b_j とのケプストラム距離 $g(i,j)$ に基づく DP マッチングを行う。a_i と b_j のユークリッド距離 (局所距離) $d(i,j)$ を求め、パスが進むごとに局所距離 $d(i,j)$ が加算される。図 3 のような DP パスを用いて (I,J) までの最小累積距離 $g(I,J)$ を求める。その結果得られたパスの線形な区間における距離とスコアによって判定を行う。助詞などの短いマッチングを回避するために最小区間長を設定し、線形区間がこれより長く、スコアが閾値よりも小さい場合その発話には言い直しが存在すると判定し、大きい場合は言い直しでないと判定する。

3.2 認識候補の重なり度による判定

直前の発話 A、現発話 B の音声認識結果の N-best 候補を求める。B が A の言い直しであるなら、認識結果の N-best には同一の候補が数多く得られると予想できる。同一の候補が含まれている数を"重なり候補数"とし、重なり度としては以下のように定義する。

$$\text{重なり度} = \frac{2 \times \text{重なり候補数} \times \text{発話の長さ}}{Nbest \text{ 候補数} + \text{直前の発話の } Nbest \text{ 候補数}} \quad (1)$$

ここで、現在の発話が誤認識した部分のみを言い直しであり構成単語数が少ない場合、重なり度は小さくなってしまうので、発話の長さを掛けることにより正規化している。言い直し発話の場合は重なり度が大きく、言い直しでない発話の場合は重なり度が小さくなるということが予想される。DP マッチングと同じように、スコアに閾値を設け、重なり度が閾値よりも大きい場合には、その発話は言い直しであると判定する。

```
Frame 28
No. 1 / P=-1948.5(-69.6) / 紀伊
No. 2 / P=-1951.8(-69.7) / いいえ
No. 3 / P=-1953.6(-69.8) / 伊江
No. 4 / P=-1958.3(-69.9) / 久井
No. 5 / P=-1958.3(-69.9) / 土居

Frame 29
No. 1 /  P=-1996.9(-68.9) / いえ
No. 2 /  P=-2000.2(-69.0) / 伊江
No. 3 /  P=-2002.0(-69.0) / 紀伊
No. 4 /  P=-2006.7(-69.2) / 久井
No. 5 /  P=-2011.7(-69.4) / 土居

Frame 30
No. 1 / P=-2044.5(-68.2) / いえ
No. 2 / P=-2047.9(-68.3) / 伊江
No. 3 / P=-2049.7(-68.3) / 紀伊
No. 4 / P=-2054.4(-68.5) / 久井
No. 5 / P=-2054.4(-68.5) / 土居
```

図 4: 認識途中結果の例

4 否定表現の検出法

ユーザの否定表現は、前発話のシステムの認識結果が誤っていることを表す。一般的には音声認識器の結果から否定表現の検出を行っているが、否定表現自体を誤認識することも少なくなく、対話を混乱させる。そのため、音声認識結果に含まれる否定表現をより正確に検出する方法を考える。否定表現は発話文頭に見られることが多い。そこで、文頭における否定表現を検出するため、発話の先頭から 100 フレーム (1 秒) までの認識途中結果を出力する。途中結果を用いることで検出率の向上が期待できるが、誤認識により否定語を出力する例も発生する。そこで数多く得られた否定表現に対して認識結果の事後確率と否定表現の単語における音響的特徴に注目し、沸き出し誤りを防ぐ。途中結果は、図 4 のように出力され、ランク、音響尤度、認識単語名などが含まれている。音節単位の認識も並行して行い、途中結果を出力し、否定表現である単語区間における事後確率を以下の式により求める。

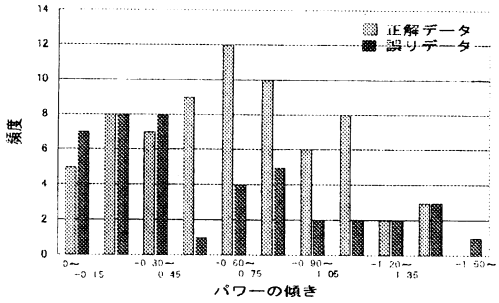

図 5: 否定表現として判定された発話のパワーの傾きのヒストグラム

事後確率 = 単語認識の音響尤度 − 音節列認識の音響尤度 (2)

さらに否定表現における音響的特徴に注目する。言い直しと否定表現を含んでいる発話においては、否定表現を発話した後、話者により長さは異なるが無音区間が生じる (使用したコーパスでは全例)。そこで、否定表現の終端では音声にパワーが低下する傾向があると考え、単語の終端のパワーの傾きを否定表現の検出に用いることにする。パワーの傾きは、否定表現の単語終端から遡った 5 フレーム (0.05 秒) までの対数パワーを最小二乗法で近似することで求める。否定表現が含まれている 70 発話より求めた否定表現の終端におけるパワーの傾きと、認識器により否定表現に誤認識された単語終端におけるパワーの傾きを図 5 に示す。図 5 より、正解データでは、パワーが落ちていることがわかり、誤認識されたものはパワーの傾きが小さく、パワーが継続して高くなっている傾向を見ることができる。

判定に用いる場合には、正解データより平均と分散を求め正規分布で近似したものを事前に求めておく。そして、テストデータのパワーの傾きから尤度を求める。

図 4 のように途中結果から得られた全ての否定表現における、フレーム i において認識した否定表現 w について事後確率 $P_{i(w)}$ とパワーの傾き $D_{i(w)}$ を求める。2 つの重みつき和をとり、最大値をとるものをその発話における否定表現が含まれることに対するスコアとする。

$$\text{否定表現のスコア} = \max_{i} \quad W_p P_{i(w)} + W_d D_{i(w)} \qquad (3)$$

5 評価実験

5.1 実験条件

2 節で示した目的地設定タスクにおいて、提案手法による判定を行う。音声分析条件を表 2 に示す。

表 2: 音声分析条件など

音声特徴量	38 次元 MFCC12 次元+Δ+ΔΔ Δパワー+ΔΔ
サンプリング周波数	16 kHz
分析窓	ハミング窓
フレーム長	25 ms
フレーム周期	10 ms
音響モデル	114 音節 HMM

音声認識は CFG 駆動の音声認識システム SPOJUS[15] を用いた。システムの認識語彙数は約 10500 であり、全国の地名、交通機関名などが含まれている。ユーザの発話した全単語のうち、約 10%は未知語であった。

5.2 評価尺度

評価尺度として、訂正発話を正しく訂正発話と判定した割合 (再現率) と言い直しと判定した発話が正しく判定された割合 (適合率)、2 つの値により求められる F 値を以下のように定義する。なお、閾値の決定は、DP、重なり度、組みあわせとも、10 人の話者で cross-validation を行いトレーニングデータで得られた閾値をテストデータに適用している。

$$\text{再現率} = \frac{\text{正しく訂正発話と判定された数}}{\text{訂正発話の数}} \qquad (4)$$

$$\text{適合率} = \frac{\text{正しく訂正発話と判定された数}}{\text{訂正発話と判定した発話の数}} \qquad (5)$$

$$\text{F 値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}} \qquad (6)$$

5.3 言い直し検出の判定結果

5.3.1 DP マッチングによる判定結果

得られた DP パスの共通部分より判定を行う。線形区間の最小距離は 25 フレーム (250ms) とし、それより短いものは言い直しとしない。閾値を変化させたときの結果を図 6 に示す。横軸は再現率、縦軸は適合率で、右上に近づく程、判定性能が高いことを示している。F 値が最も高くなるように閾値を設定すると、再現率 0.891、適合率 0.900、F 値 0.895 が得られた。

5.3.2 認識候補の重なり度による判定結果

得られた重なり度に閾値を設けて判定を行う。結果は、F 値が最も高くなるように閾値を設定すると、

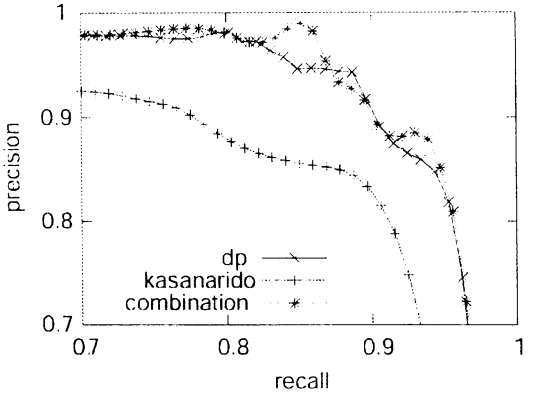

図 6: 言い直し検出の判定性能

再現率 0.807、適合率 0.783、F 値 0.795 であった。重なり度の判定性能は DP と比べて非常に低くなってしまっている。これは認識結果の単語正解精度が 55%と低いため、同じ単語も異なる候補として認識してしまっていることが原因と考えられる。認識精度の向上があれば、重なり度はより有効になると考えられる。

5.3.3 DP と重なり度の組み合わせによる判定結果

DP マッチングと重なり度によって得られたスコアの重みつき和により言い直し確率を求める。上の例と同じように閾値による判定を行った。F 値が最も高くなるように設定すると、再現率 0.865、適合率 0.949、F 値 0.905 であった。重なり度を組み合わせることで、DP 単独の手法よりわずかではあるが判定性能の向上が見られた。

5.4 否定表現検出の判定結果

得られた否定表現のスコアに対し、閾値を用いることで判定を行う。事後確率のみで判定した結果と、パワーの傾きも含めて判定した結果それぞれにおいて、F 値が最も大きくなるときの結果を表 3 に示す。

表 3: 否定表現の判定性能

手法	再現率	適合率	F 値
認識結果の *1-best(baseline)*	0.614	0.86	0.716
事後確率による判定	0.743	0.922	0.819
事後確率＋傾きによる判定	0.771	0.931	0.844

表 3 より、パワー傾きを用いることで正しくない候補の棄却に成功していることがわかる。棄却に成功した例としては、地名発話の一部が否定表現に置き換わってしまったもの (例:「名神高速道路…」における「めいし」区間を「いいえ」に誤認識してしまう) がある。単語途中であり、パワーが低下しないため否定表現のパターンとは異なることになる。

User:愛知県豊橋市
User(repaired) いいえ,豊橋市です
言い直しでないと誤判定

User:愛知県豊橋市
User(repaired):愛知県、豊橋市です
否定表現と誤判定

図 7: 判定誤りの例

言い直しと否定表現を含む発話は、言い直し検出もしくは否定表現検出のいずれかで検出できればよいが、否定表現のみの発話は、否定表現検出でしか検出ができない。言い直しを含む否定表現の発話は全て検出できたが、否定表現のみの発話は 28 発話中 20 発話の検出にとどまっている。否定表現のみの発話を検出するほかの手法が必要となる。

5.5 言い直し・否定表現の組み合わせ

前述した言い直し検出と否定表現検出の組み合わせにより、ユーザの訂正発話の検出を行う。システムは 2 つの手法の判定結果より、ユーザの訂正発話を以下のように判断する。

表 4: 2 つの判定結果によるシステムの判定

言い直し	否定表現	ユーザ発話
なし	なし	訂正でない
なし	あり	訂正
あり	なし	訂正
あり	あり	訂正

2 つの判定を用いているが、どちらか、あるいは両方の判定が誤ることがある。表 4 における、「言い直しなし、否定表現あり」の例については組み合わせによりそれぞれの判定結果を見直すことが可能である。否定表現を含み、前発話とは無関係の内容を発話することは通常考えにくい (使用したコーパスでは皆無)。誤って判定される例としては、図 7 のようなものが考えられる。

否定表現のみの発話でないもの、すなわち一定発話長以上のもので、このケースに割り当てられた発話はどちらかの判定が間違っているものとする。ここで発話が複数の単語より構成されているものは、否定表現のみの発話でないとし、他のケースに割り当てられるよう再判定をする。その際、式 (3) における言い直しスコアと、否定表現のスコアを異なる重みで組み合わせる。実験結果より、否定表現のスコアを重要視したほうがよい結果となった。最終的な組み合わせによる言い直しと否定表現の判定性能

を表5に示す。

表 5: 組み合わせ手法による判定性能

手法	再現率	適合率	F 値
言い直し判定	0.896	0.983	0.937
否定表現の判定	0.800	0.918	0.855

発話長を用いることで図7のような例を防ぐことができ、判定性能の向上が得られていることがわかる。システムは自身の誤りを検出率 0.864、適合率 0.955 の判定性能を得ることができた。

6 訂正発話検出による効果

言い直し判定を行うことで、認識を誤った前発話の情報を用いて認識を行うことが可能となる。これにより、認識率の向上が得られると予想できる。

角谷ら [2] は、再認識を行う方法として、一度認識した結果の認識候補において重なり度を判定し、判定結果より辞書を絞り込む方法を行った。この方法は重なり度を用いるため、2回の認識処理を要しリアルタイム処理が不可能であるが、認識率の向上を得ている。また、対話ターンによって、システムの質問に対してユーザの発話する語彙が異なるようなタスクでは、質問ごとに言語モデルを別々に用意しておくことで語彙を限定することができ、パープレキシティが減少できたとしている。ユーザの発話が訂正発話であれば前の質問に対応する言語モデルで認識させることになる。また訂正発話の判定に迷うような場合は、両方のケースを想定して認識を行うことで判定誤りによる認識の誤りを防ぐことも可能である。

直前の認識結果と、現在の認識結果において、重なり度が高ければ2つの認識結果には同じような候補が得られていることになる。そこで、認識結果として、単純に言い直し発話の認識結果の 1-best を用いず、2つの認識結果の N-best で共通して存在する候補を用いた効果的な認識結果の獲得方法を今後の課題としている。

7 まとめ

本稿では、システムの誤認識に対するユーザの訂正を検出する方法として、言い直し検出と否定表現の検出により行った。

言い直し検出方法として、DP マッチングと重なり度を用いた判定方法を行い、DP マッチングによる判定では検出率 0.891、適合率 0.900 の性能が得られ、重なり度では検出率 0.807、適合率 0.783 が得られた。さらに組み合わせにより再現率 0.865、適合率 0.949 が得られた。

否定表現の検出方法では、認識結果より事後確率とパワーを用いて正しいものの検出を行い、パワーの傾きを用いて再現率 0.771、適合率 0.931 が得られ、パワーの特徴を用いることで判定性能の向上が見られた。

言い直しと否定表現の検出による方法では、それぞれを補うことにより判定性能の向上が得られ、システム自身の誤りも再現率 0.864、適合率 0.955 の判定性能が得られた。

言い直し判定を用いることで、前回の発話の結果を利用することが可能となり、語彙の限定や、認識結果のリスコアリングを行うことができ、認識率の向上が期待できるのではないかと考えている。

参考文献

[1] 角谷 直子, 北岡 教英, 中川 聖一 :「カーナビの地名入力における誤認識時の訂正発話の分析と検出」, 情処研報,2001-SLP-37-11(2001.7)

[2] 角谷 直子, 北岡 教英, 中川 聖一 :「カーナビの地名入力における誤認識時の訂正発話の検出と認識」, 電気学会東海支部連合大会,751,pp.376(2001.11)

[3] 北岡 教英, 角谷直子, 中川 聖一 :「音声対話システムの誤認識に対するユーザの繰り返し訂正発話の検出と認識」, 電子情報通信学会論文誌,Vol.J87-D-II No.7(2004.7)

[4] 井ノ上 直己, 今井 裕志, 橋本 和夫, 米山 正秀 :「誤認識訂正のための繰り返し音声検出手法」, 電子情報通信学会論文誌,Vol.J84-D-II No.9(2001.9)

[5] 今井 裕志, 井ノ上 直己, 橋本 和夫, 米山 正秀 :「未知語処理のための繰り返し音声検出手法」, 電子情報通信学会,SP99-26,pp.1-6(1999.6)

[6] 平沢 純一, 宮崎 昇, 相川 清明 :「質問-応答連鎖からの音声対話システムの誤解の検出」, 電子情報通信学会, SP2000-115,pp.34-41(2000.12)

[7] 平沢 純一, 宮崎 昇, 相川 清明 :「音声対話システムの誤解に対するユーザ応答の分析」,H12 年度春季日本音響学会講演論文集, 3-8-10,pp85-86(2000.3)

[8] Sharon Oviatt, Margaret MacEachern and Gina-Anne Levow : 「Predicting hyperarticulate speech during human-computer error resolution」, Speech-Communication,Vol.24,pp.87-110(1998)

[9] Gina-Anne Levow : 「Adaptation in spoken corrections: Implications for models of conversational speech」,Speech-Communication,Vol.36,pp.147-163(2002)

[10] Gina-Anne Levow : 「Identifying Local Corrections in Human-Computer Dialogue」,ICSLP2004,pp.313-316,2004

[11] Marc Swerts,Diane Litman and Julia Hirschberg :「Correction in Spoken Dialogue System」, ICSLP2000, Vol.2,pp.615-618(2000)

[12] 山肩 洋子, 河原 達也 :「音声対話システムにおける訂正発話の韻律的特徴の分析」, 人工知能学会研究会,SIG-SLUD-A101-3(2001.6)

[13] 伊藤敏彦, 岩本義行, 水谷誠, 湯浅裕規, 甲斐充彦, 小西達裕, 伊藤幸弘 :「目的地設定タスクにおける対話状況の違いによる言語的特徴の分析」, H13 年度秋季日本音響学会講演論文集,2-1-9,pp65-66(2001.10)

[14] 日本語連続音声認識システム SPOJUS-SYNO,http://www.slp.ics.tut.ac.jp/SPOJUS/