TR－SLT－0069

言い直し音声の精度向上に関する検討

Study on Recognition for Re-speaking Speech

小窪 浩明　　山本 博史
Hiroaki KOKUBO　　Hirofumi Yamamoto
菊井 玄一郎
Genichiro Kikui

2004年4月28日

概要

音声認識を用いたシステムでは，認識誤りを避けて通ることは出来ない．一般に，認識誤りが生じると，発話者は正しい認識結果を得るために，同じ発話を繰り返す(言い直し)．本研究では，この言い直し発話に対する音声認識性能を向上させることを目的とした検討を行った．はじめに言い直し音声データの収集と収集した音声データの特徴について報告する．次に，言い直し音声の認識について，言い直し前の音声認識結果を導入すること効果について実験結果を示す．評価話者3名の言い直し音声に対するベースラインの文認識率 61.8%に対し，語彙制約によるアプローチでは 63.9%，ROVER 法に基づくアプローチでは 63.3%，音響スコアの平均によるアプローチでは 64.2%と，それぞれのアプローチについて認識性能の向上効果を確認した．

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

目次

1 はじめに

1.1 背景

音声認識において，認識誤りは不可避である．そのため多くの音声認識システムでは，認識結果をユーザーに提示して確認／修正を求める機能を持っている．音声認識誤りを修正する手段としては，複数の認識候補を提示して正解をユーザーに選択させたり，認識結果を取り消すコマンドを用意するなど多種のアイディアが提案されている．音声対話をモダリティーとするシステムでは，認識誤りに対する基本的な訂正手段は言い直しである．音声認識の結果に認識誤りが生じると，ユーザーは同じ発話を繰り返すことでその誤りの訂正を試みる．しかし，繰り返された発話が常に正しく認識されるとは限らない．言い直した発話に対して認識誤りが再び繰り返されることも少なくない．認識結果の訂正が精度よく機能しない認識システムは，アプリケーションの進行が途中で停滞してしまい，ユーザーのフラストレーションは増大する．そのため，認識誤りを言い直しによって修正する手段をもつシステムでは，言い直し音声を高精度で正しく認識できることが強く求められる．そこで本研究では，言い直し発話に対する認識向上について検討する．

1.2 先行研究

言い直し発話の認識に関しては，人名，地名，数字の認識など単語単位の言い直しを対象とした先行研究が報告されている．角谷ら [1][2][3] は，カーナビゲーションの目的地入力を対象とした地名認識において，再発声音声の分析と，その検出方法について報告している．彼女らの研究では，DP マッチングを用いたワードスポッティングによって言い直しであるか否かを判定し，言い直しと特定される部分に対して語彙や文法で制約をかけることにより認識性能の向上を図っている．

井ノ上ら [4] は，複数のキーワードを用いた情報検索システムの音声入力を対象とした，キーワード認識の言い直しについて報告している．彼らは，繰り返し音声 (キーワード) の検出のために，N ベスト候補の重複率，認識尤度差，DP スコアにもとづく三つの手法を提案し，それらを組み合わせることで繰り返し音声を高い精度で検出している．

伊藤ら [5] は，音素タイプライターを利用した未知語 (孤立単語地名) 推定を対象とした手法について報告している．複数音声サンプルの音韻系列を音素タイプライターによって推定し，得られた複数の音韻系列を複数のサンプルに当てはめ，それらの平均尤度が最大となる音韻系列を選択することで，音韻系列の推定精度を向上させている．

1.3 研究の目的

1.2 節で紹介した先行研究は，発話全体の言い直しを対象にしたものではなく，単語単位の言い直しを対象とした報告であった．本研究の目的は，言い直し発話に対する音声認識性能を向上することであり，特に文単位の言い直し音声の認識を対象とする．また，本研究で対象とする言い直し発話は，言い直し前の発話と同一の文を言い直す場合のみに限定し，言い換えなどは本研究の対象外とする．

2 提案手法

2.1 語彙制約によるビーム幅の拡大

音声認識誤りの原因には，音響モデル，言語モデルに起因するモデルのミスマッチによる誤りと，ビームサーチにより正解候補が枝刈りされる探索誤りとに分類される．探索誤りに関しては，ビーム幅を広げることで誤りを減少させることは可能である．しかし，仮説数の増加によりワークメモリの肥大化や探索処理時間の増加が生じるため，実用上設定可能なビーム幅には限界がある．

ところで，音声認識が誤った場合もNベスト候補(あるいはワードラティス)の中には正しい認識結果が含まれていることが多い．そこで，言い直し音声を認識する際は認識対象語彙を認識の際に生成されるワードラティスに含まれる単語に限定する．このように認識語彙を制約することによって，ビーム幅を拡大してもワークメモリの肥大化を防ぐことが可能となる．また，制約となるワードラティスに正しい認識結果が存在しているならば，語彙を制約しても正しい認識結果が外されてしまうことはない．

手法の概念図を図1に示す．言い直し前の発話と言い直し発話をそれぞれ認識器にかけると単語ラティスがそれぞれ生成される．この時生成された単語ラティスに含まれる単語を認識辞書から抽出して新たな認識辞書を生成する．この認識辞書を用いて，言い直し音声を再びデコードする．新たに生成した認識辞書はオリジナルの認識辞書に比べて語彙数は大幅に削減されているため，生成される仮説数も大幅に縮小され，ビームを広げた場合でもワークメモリの肥大化はおこらない．したがって，ビーム幅を拡大することにより探索誤りを軽減し，認識性能の向上が期待できる．

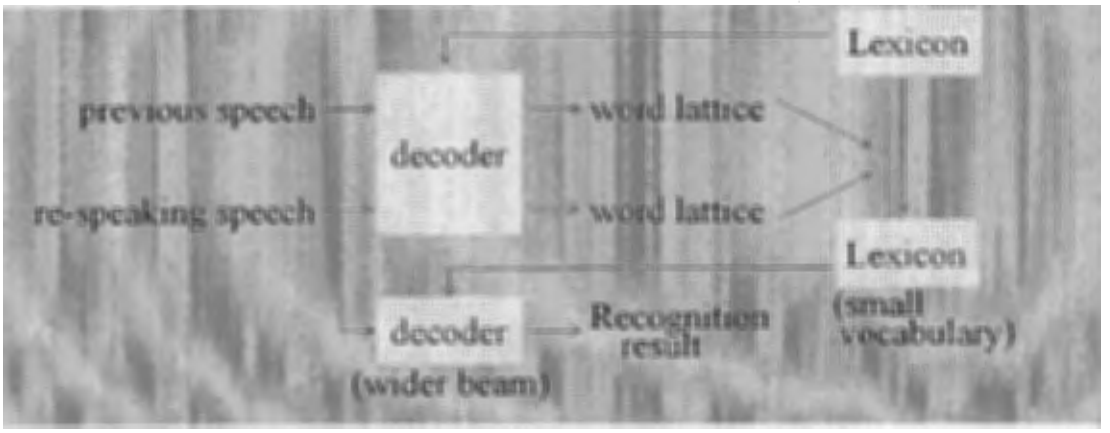

図 1: 語彙制約によるアプローチ

2.2 ROVER

ROVER(*Recognizer Output Voting Error Reduction*)法 [8] は，複数の認識器を用意し，それぞれの認識結果を多数決によって統合することで，単独の認識性能を上回る認識性能を獲得する手法である．複数の認識器による認識結果を用いる代わりに，言い直し前の発話と言い直

し発話の2つの発話の認識結果を入力とすることで，ROVER法の手法をそのまま言い直し音声の認識に適用させることができる．

例えば，言い直し前の発話の音声認識結果を $\mathbf{W_A} = \{w_1 w_2 w_3 w_4 w_5\}$，言い直し発話の音声認識結果を $\mathbf{W_B} = \{w_1 w_6 w_3 w_7 w_5\}$ とする．この場合，w_1, w_3, w_5 の認識結果は双方同じであり，w_2 と w_6，w_4 と w_7 の部分の認識結果は異なる．双方の認識結果が同じ部分の単語をマージすると図2のような単語ネットワークを作成することができる．ROVER法では，この単語ネットワークを通るパスのうちスコア最大のパスを選択することにほかならない．ここで使用するスコアは言語スコアのみを採用している [9].

また，語彙制約によるアプローチの拡張として，図2の単語ネットワークを制約としたFSA(Finite State Automaton)を用いてデコードすることも考えられる．言語制約として用いるFSAの例を図3に示す．FSA1はROVER法に基づいて生成された単語ネットワークをそのままFSAに変換したもの，FSA2は二つの認識結果が一致しない単語はanyクラスに置き換えたものである．ここで，anyクラスは辞書に登録されているすべての単語が当てはまるクラスである．したがって，図3のFSA2を用いて認識をおこなった場合，1番目の単語は w_1，3番目の単語は w_3 に制限されることになるが，2番目の単語と4番目の単語は，任意の単語で構わない．

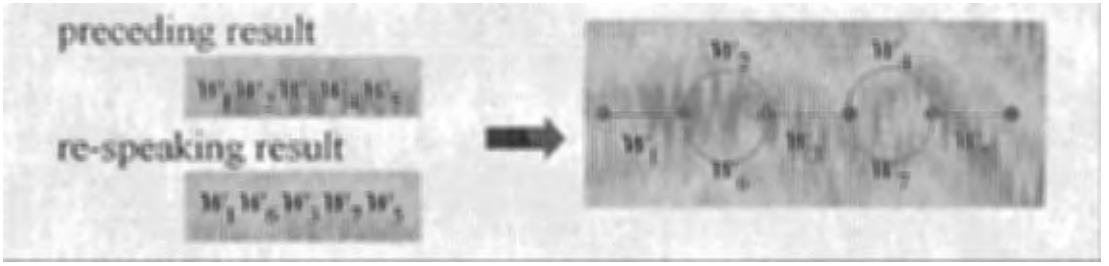

図 2: ROVER法にもとづく単語ネットワーク

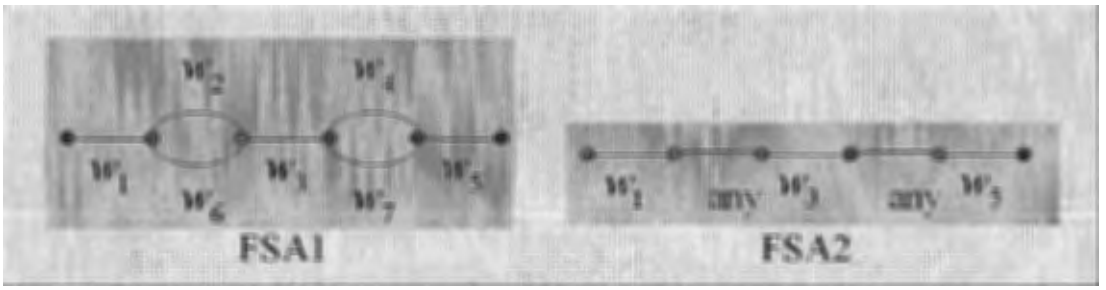

図 3: ROVER法にもとづくFSAの制約

2.3 音響スコアの平均

伊藤ら [5] は，音素タイプライターを利用した未知語 (孤立単語地名) 推定に複数の音声サンプルを利用し，上位N個の音韻系列のうち複数サンプルの平均尤度が最大となる音韻系列を選

択することで推定精度を向上させている．図 4に彼らのアプローチの概念図を示す．図の例では，２回の単語発声に対して，音素タイプライターが生成した単語 (音素列) の N ベスト候補がそれぞれ生成される．例えば，一回目の認識単語 $word_a$(音響スコア s_{a1}) についてみると，二回目の N ベスト候補にも $word_a$(音響スコア s_{a2}) が存在する．そこで二つの音響スコアの平均をとる．

$$s_a = (s_{a1} + s_{a2})/2 \tag{1}$$

この平均スコアを計算をそれぞれの N ベスト候補に現れる認識単語すべてのペアついて行い，スコアの平均値が最大となる単語

$$w_i = \underset{w_i}{\operatorname{argmax}}(s_i) \tag{2}$$

を認識結果とする．

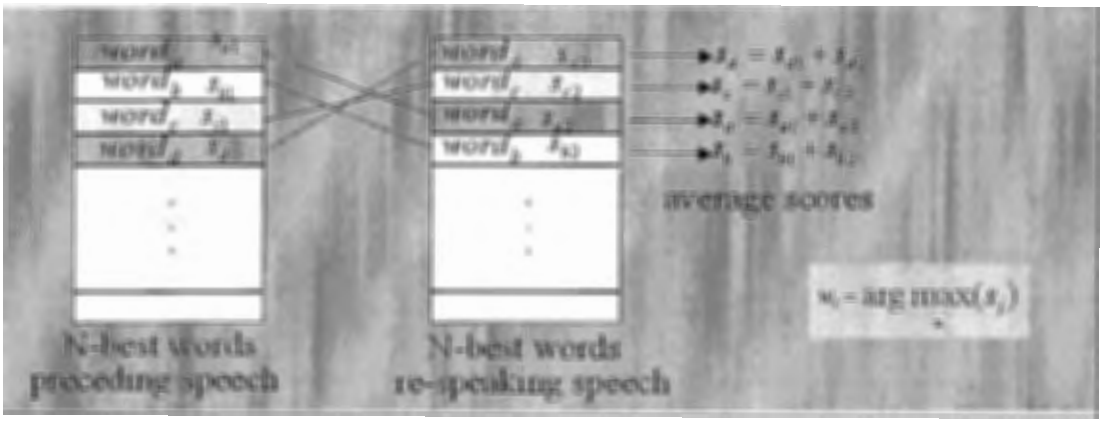

図 4: 複数音声サンプルを用いた音響スコアの平均

同様のアプローチは連続音声認識に対しても応用可能である．ただし音韻タイプライターの場合では，音韻の種類が高々 150 程度であるのに対して，連続音声認識の場合は語彙数が 10,000 を越える．その結果，一方の認識結果に現れた文 (単語列) が他方の認識結果の N ベスト (もしくはワードラティス) の中に必ずしも存在するとは限らない．そこで，一方の認識結果として現れた単語列から FSA(Finite State Automaton) を作成し，その FSA を制約条件として再デコードすることで所望の文 (単語列) に対する認識スコアを獲得する．

提案方式の概念図を図 5に示す．ここで，単語列 $\mathbf{W}_i = \{w_{i1}w_{i2}, ..w_{in}\}$ の認識スコアを $score(\mathbf{W}_i)$，その時の音響スコアを $Ascore(\mathbf{W}_i)$，言語スコアを $Lscore(\mathbf{W}_i)$ とすると，

$$score(\mathbf{W}_i) = Ascore(\mathbf{W}_i) + \alpha \cdot Lscore(\mathbf{W}_i) \tag{3}$$

となる (α は言語重み)．

ここで，言い直し前の発話の認識結果 $\mathbf{W}_i$ に対するスコアを $score^b(\mathbf{W}_i)$，言い直し発話の認識結果 $\mathbf{W}_i$ に対するスコアを $score^r(\mathbf{W}_i)$ とすると，二つのスコアの平均は，

$$\begin{aligned} score^{average}(\mathbf{W}_i) &= (score^b(\mathbf{W}_i) + score^r(\mathbf{W}_i))/2 \\ &= (Ascore^b(\mathbf{W}_i) + Ascore^r(\mathbf{W}_i))/2 + (Lscore^b(\mathbf{W}_i) + Lscore^r(\mathbf{W}_i))/2 \end{aligned} \tag{4}$$

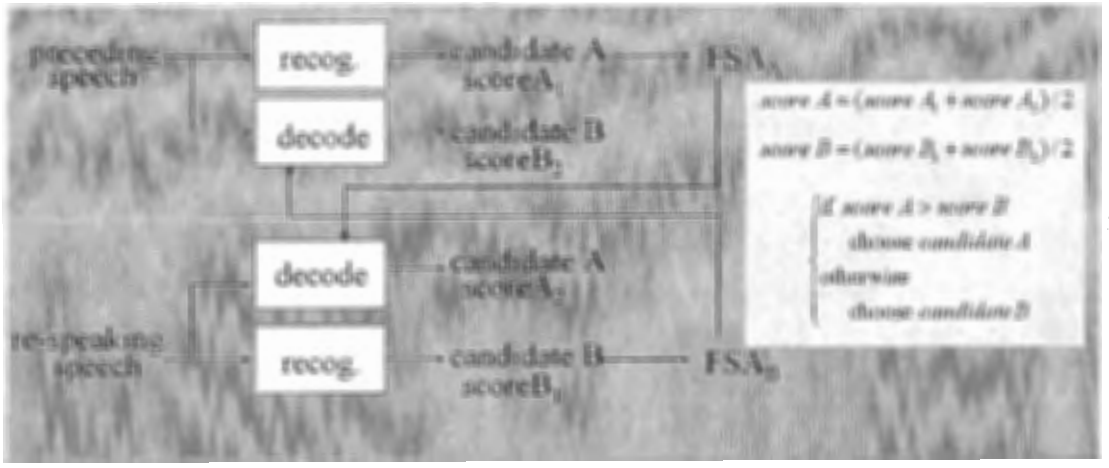

図 5: 言い直し音声ペアに対する認識スコアの平均

単語系列が共通であればそれぞれの言語スコアは等しいため，

$$score^{average}(\mathbf{W}_i) = (Ascore^{b}(\mathbf{W}_i) + Ascore^{r}(\mathbf{W}_i))/2 + Lscore(\mathbf{W}_i) \tag{5}$$

となる．従って，認識スコアの平均は，音響スコアの平均をとることによって，音響スコアに対する信頼度を向上させていることに他ならない．

3 言い直し音声データ

3.1 音声収録

評価用音声データの収録について述べる．収録条件を表1にまとめた．発話者は女性２名，男性１名の計３名．発話内容はBTECコーパス[7]のset01(510文)である．発話文はPCに接続されているディスプレー上に１文ずつ表示され，発話者はその文を読み上げる．ただし，発話者には「朗読口調にはならずに，できるだけ会話口調になるように」との教唆がなされている．一文の発話が完了すると，次の一文がディスプレーに表示される．この時，１／２の確率で，図6のような言い直しを要求する表示を出す．この表示が現れると，発話者は直前に発話した内容を繰り返し発話する．理想的には，ディスプレーに発話内容を表示せずに言い直しを行う方法をとることが望ましい．しかし，長い発話文では繰り返すべき発話内容を覚えることが困難であるという理由から，言い直しの発話を行う際にもディスプレーに発話内容を表示させたままにした．この手順の収録を一人の発話者について２セット繰り返す．ただし，１セット目の収録で言い直した文章は，２セット目の収録ではいい直しはせず，１セット目でいい直しを行わない文章は２セット目でいい直しを行うように，言い直しの指示は予めコントロールされている．この結果，評価音声２セットの収録が完了すると，510文章すべてに対する言い直し音声の収録も完了する．

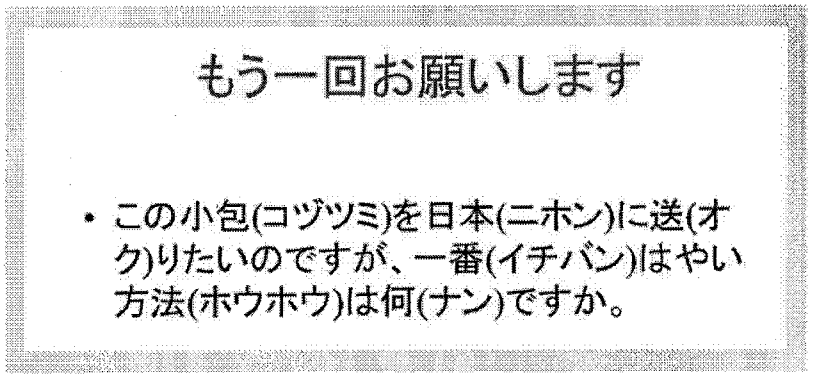

図 6: 言い直しのプロンプト例

表 1: 言い直し音声データの収録条件

発話者	女性２名 (f01,f02)，男性１名 (m01)
発話内容	BTEC[7] set01 (510文)
発話スタイル	会話文の読み上げ
発話セット数	通常２セット＋言い直し１セット
A/D	16kHz,16bit PCM

3.2 言い直し音声の特徴

音声認識システムを使用するユーザは，何度も言い直しを要求されるとしだいにストレスやいらだちをおぼえるようになり，発話スタイル (声の大きさ，スピード，ピッチなど) にも変化が現れる．

角谷らの調査 [1] によると，音声のパワーについては，３回以内の繰り返しでは大きな変化は見られないものの，３回を越えて繰り返すと回数を増すごとにパワーは大きくなっていく．また，ピッチや発話スピードに関しては，顕著な変化は観測されなかったことが報告されている．この報告をふまえた上で，今回作成した評価音声データに関して，言い直し音声の特徴について調査した．

はじめに，音声のパワーについて言い直し前の発話と言い直し発話を比較した．図 7は，１文目から５１０文目までを横軸にとり，言い直し発話の音声パワーを言い直し前の発話のパワーで正規化した値をプロットしたグラフである．

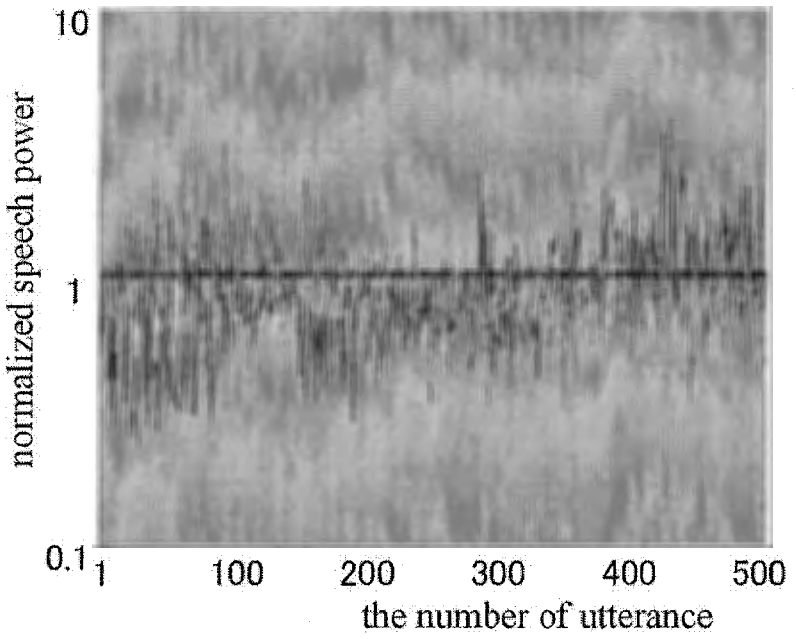

図 7: 言い直し発話の正規化音声パワー

表 2はこの正規化したパワーの値に関して５１０文の平均値と分散を計算した値である．分散は大きく発話毎のばらつきは大きいものの，正規化した音声パワーの平均値は 0.93 から 1.02 となり，言い直し発話の音声パワーは言い直し前の音声パワーから大きな変化は生じていないことが確認された．

同様に，発話スピードについても調査した．図 8は，１文目から５１０文目までを横軸にとり，言い直し発話の継続時間を言い直し前の発話の継続で正規化した値をプロットしたグラフである．表 3はこの正規化した値に関する５１０文の平均値と分散である．正規化された発話継続長の平均値は，1.02 から 1.13 とやや言い直し音声の長さが長くなる傾向が見られるものの大きな変化はみられなかった．また，分散も小さい．

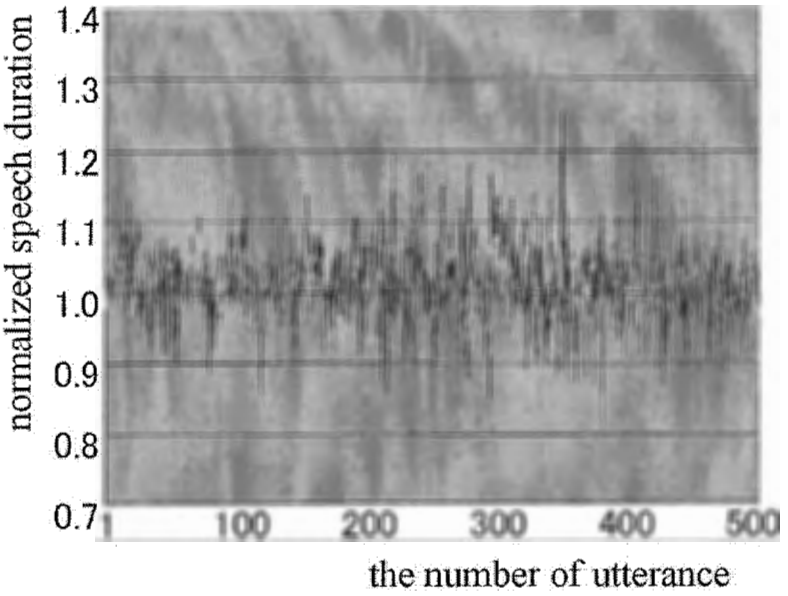

図 8: 言い直し発話の正規化音声継続長

表 2: 言い直し発話の正規化音声パワーの平均値と分散

	Average	Standard deviation
f01	0.93	0.46
f02	1.02	0.37
m01	0.96	0.51

表 3: 言い直し発話の正規化音声継続長の平均値と分散

	Average	Standard deviation
f01	1.02	0.05
f02	1.13	0.04
m01	1.10	0.07

4 評価実験

4.1 実験条件

提案方式を評価するための認識実験を行った．実験条件を表 4に示す．音響モデルは，音素環境依存 HMnet を男性用，女性用それぞれ 1 つずつ用意した．言語モデルは，デコード時に単語 bigram を用い，生成された単語ラティスを単語 trigram でリスコアリングする．評価用音声は，f01 と f02 が女性話者，m01 が男性話者の音声である．認識時のビーム幅は 110 であり，メモリを 2GB 搭載した Linux マシンを用いた場合，これ以上ビーム幅を広げると極端に探索時間が遅くなるか，メモリ不足のため仮説の展開が途中で中断されてしまう．

表 4: 認識実験条件

デコーダ	ATRSPREC[10]
音響モデル	音素環境依存 HMnet[6] ·1400 状態 5 混合 (男性用モデル) ·1400 状態 15 混合 (女性用モデル)
辞書	エントリー数 約 36,000
言語モデル	単語 bigram(decode)，単語 trigram(rescore)
評価音声	女性話者 2(f01,f02)，男性話者 1(m01)
ビーム幅	110

4.2 ベースラインの評価

提案手法を評価する前に，ベースラインの認識率を求めた．音声認識結果を表 5に示す．表中，preceding speech が言い直し前の発話に対する認識結果であり，re-speaking speech が言い直し発話の認識結果である．また，1-best は一位候補の正解率，network は生成された単語ラティス内での正解率を示す．

1-best の結果を見ると，3 名の評価音声とも言い直し前の発話よりも言い直し発話の方が単語正解精度 (word accuracy)，文正解率 (word correct) とも向上している．言い直し発話の方が認識率が良くなる理由としては，特に長い文章の場合，同じ発話を繰り返すことでリハーサル効果が生じ，言い直し発話の方がよりなめらかに話すことができたためと考えられる．また，3 名の評価話者のうち，f01 は単語正解精度が 90% 以上，文正解率が 70% 以上であるのに対し，他の 2 名は，単語正解精度で 80% 前後，文正解率で 50% 前後と大きな開きがある．

ネットワーク正解率についても同様のことが言える．また，文正解率では一位候補の正解率に比べて 10 ～ 15% 高い値を示しており，一位候補の認識結果が誤っていたとしても，何らかの方法を用いて単語ラティスの中から正しい認識結果を見つけ出せる可能性を示している．

4.3 語彙の制約

ビーム幅拡大による探索エラーを減少させることを目的とした認識語彙の制約に関する評価実験を行った．実験では，言い直し前の発話を認識した結果得られたワードラティスと言い直

表 5: ベースラインの音声認識結果

		1-best		network	
		preceding speech	re-speaking speech	preceding speech	re-spreaking speech
word accuracy(%)	f01	90.9	93.1	95.7	96.6
	f02	78.6	81.5	88.7	89.7
	m01	83.4	85.2	91.7	93.6
sentence correct(%)	f01	71.4	75.9	83.3	84.9
	f02	45.7	51.8	57.3	61.6
	m01	54.1	57.8	69.4	73.1

表 6: 語彙制約の評価結果 (文正解率)

	baseline	Limited vocabulary		
beam	110	110	120	130
f01	75.9%	72.7%	75.7%	77.3%
f02	51.8%	49.6%	52.4%	54.9%
m01	57.8%	56.3%	58.8%	59.6%
average	61.8%	59.5%	62.3%	63.9%

し後の発話を認識した際に得られたワードラティスとからそれぞれに含まれる単語を取り出し，新たな認識辞書を生成した．この語彙を限定した辞書を用いて，言い直し音声を再び認識する．

語彙を制約した実験では，ベースライン評価時のビーム幅 110 を基準に，ビーム幅を 110，120，130 と広げていった．実験結果を表 6に示す．ベースラインは，表 5に示した言い直し音声の文正解率である．ベースラインのビーム幅 110 は，語彙制約を行わない場合，2MB のメモリを搭載した PC 上で実用上動作するビーム幅の上限である．ビーム幅 110 で比較すると，語彙を制約することでベースラインよりも文正解率が低下している．これは，語彙が少ないほどフレーム毎の最尤言語スコアの値が大きく変動してしまうため，ベースラインと同じビーム幅では正解候補が枝刈りされる危険性が増えてしまったと考えられる．一方，ビーム幅を大きくしていくと文正解率は上昇していく．ビーム幅を 130 にしたときには，3 名の平均文正解率は 63.9% となり，ベースラインに対して 2.1% 向上した．

4.4 ROVER

ROVER 法にもとづくアプローチについて評価した．認識結果を表 7に示す．言い直し前の発話に対する認識結果と言い直し発話に対する認識結果とを二つの認識結果と見なして ROVER で統合した結果 (ROVER) は，3 名の平均文正解率で 63.3% とベースラインの 61.8% を上回った．また，ROVER 法で認識結果を統合する際に生成される単語ネットワークから FSA を作成した．この FSA を言語制約として用いた認識結果が FSA1 と FSA2 である．FSA1 は単語ネットワークをそのまま FSA に変換した場合，FSA2 は単語ネットワークで二つの発話の認

表 7: ROVERに基づくアプローチの評価結果 (文正解率)

	baseline	ROVER	FSA 1	FSA 2
f01	75.9%	74.5%	75.5%	76.2%
f02	51.8%	53.3%	52.0%	52.0%
m01	57.8%	59.5%	59.5%	59.4%
average	61.8%	63.3%	62.3%	61.8%

識結果が一致しない部分をanyクラスに置換してFSAを作成した場合の結果である (図3参照). anyクラスとは，辞書内のすべての単語が受理可能なクラスである．評価者3名の平均文正解率は，FSA1で62.3%，FSA2で61.8%であった．FSA1，FSA2の評価で期待していた効果が得れなかったのは，4.3節でも指摘したようにビーム幅は110のままにしたことが原因と思われる．したがって，ビーム幅を拡大することで文正解率を向上させることは可能である．

4.5 音響スコアの平均

言い直し前の発話の認識結果と言い直し発話の認識結果との認識スコアを平均することで，音響スコアの信頼性の向上を試みた．実験では，一方の認識結果として現れた単語列からFSA(Finite State Automaton)を作成し，そのFSAを制約条件とすることで所望の単語列のみを受理する音声認識を行った．

認識結果を表8に示す．N-bestは，言い直し前の発話の認識結果と言い直し発話の認識結果のそれぞれのN-best(N=100)に対して，共通して現れる認識候補文に関してスコアの平均値をとりその平均値が最大となる認識候補文を選択した場合の評価結果である．連続音声認識では，同じ文章を発話したとしても，一方の発話の認識候補文が他方の認識候補のN-bestに存在するとは限らない．したがって，いずれか一方のN-bestに正解認識結果が含まれない場合には，誤った認識結果が選択されてしまう．そのため，3名の平均文正解率は59.5%とベースラインの結果よりも悪くなっている．

一方，表8のre-decodingは，一方の発話の認識候補文が他方の認識候補のN-bestに存在するように，FSAによる言語制約を用いて再デコードし，スコアの平均をとった場合の評価結果である．1-bestは，認識結果の一位候補の単語系列からFSAを作成した場合，word latticeは，認識の結果生成された単語ラティスからFSAを作成した場合である．1-bestの条件では，言い直し前発話か言い直し発話のどちらか一方の一位候補に正解文が存在しない限り，スコアーの平均をとっても正解が現れないのに対し，word latticeの場合は，ラティス内に正解文が存在すればどちらの発話の一位候補に正解がなかったとしてもスコアーを平均することで正しい認識結果が得られる可能性がある．しかし，図8の結果では，1-best，word latticeともベースラインの文正解率を上回っているものの，1-bestの条件の方が僅かながら良い結果となった．

表 8: 認識スコアの平均を用いた評価結果 (文正解率)

	baseline	N-best	re-decording (1-best)	re-decording (word lattice)
f01	75.9%	74.7%	76.5%	76.1%
f02	51.8%	48.0%	54.5%	54.1%
m01	57.8%	55.7%	61.6%	60.8%
average	61.8%	59.5%	64.2%	63.7%

5 まとめ

言い直し音声の認識性能を向上させることを目的とした諸検討について報告した．はじめに，言い直し音声に関して評価用音声データの作成方法を述べ，次いで作成した言い直し音声データの特性について調査した．調査の結果，言い直し音声は，言い直し前の音声に対して，パワー，発話長とも大きな変化はないことが明らかとなった．この結果は，カーナビゲーションの地名入力を対象とした先行研究 [1] の分析結果とも一致した．

言い直し音声の認識に関して３つの方式を提案し，収録した音声データを用いてそれぞれの提案方式について評価を行った．

- 言い直し前の発話と言い直し発話の認識結果の単語ラティスに含まれる単語に認識対象語彙を絞り，ビーム幅を拡張した．その結果，ベースラインと同じビーム幅では，語彙を少なくすることで，ビームの変動が不安定になるため，ベースラインの文認識率よりも悪くなったものの，ビーム幅を大きくしていくほど文認識率は向上した．ビーム幅 130 の時，３名の平均文正解率は 63.9% であり，ベースラインの 61.8% よりも 2.1% の向上が見られた．

- ROVER 法に基づくアプローチについても評価実験を行った．言い直し発話に対する認識結果とを二つの認識結果と見なして ROVER で統合した．その結果，３名の平均文正解率で 63.3% とベースラインの 61.8% を上回った．また，ROVER 法で認識結果を統合する際に生成される単語ネットワークに基づき作成した FSA を言語制約に用いた評価実験では，性能改善の効果はほとんどみられなかった．ただし，今回の評価ではビーム幅に関して考慮していないので，ビーム幅を広げることによる効果については検討に値する．

- スコア平均のアプローチでは，言い直し前と言い直し発話の２つの認識結果のスコアを平均することで，音響スコアの信頼度の向上を図った．２つの認識結果の平均をとる際 N-best(N=100) を用いると，一方の N-best に現れた認識結果が，他方の N-best に存在しないことも多く，評価話者３名の平均文正解率は，％と，ベースラインよりも悪くなってしまった．そのため，一方の発話で現れた認識結果が，他方の N-best にも現れるように FSA で制約をかけた．この結果，平均文正解率は 64.2% と，ベースラインよりも 2.4% の向上がみられた．

本研究では，言い直し発話の認識に関して，３つのアプローチを提案した．評価実験では，それぞれのアプローチともベースラインを上回る文正解率が達成できたものの，その改善率はそれほど大きいものではない．今後は，事後確率などのコンフィデンススコアの導入なども必要であろう．さらに，正しい認識結果が復元できない場合には，誤り個所を特定するアプローチなども検討していかなければならない．

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， SLP37-11,pp.61-66,(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言い直し発話の検出手法，音講論 2-5-17,pp.107-108,(2002.3)

[3] 北岡 教英，角谷 直子，中川 聖一：対話音声中の言い直し発話の検出と認識，情処研報 SLP-46,pp.101-106,(2003.5)

[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，信学論 D-II Vol.J84-D-II No.9 pp.1950-1959,2001.9

[5] 伊藤 克宣，速水 悟，田中 和世：単語発声の複数サンプルを利用した未知語の音素系列の推定，信学会論文誌 D-II vol.J83-D-II No.11 pp.2152-2159,2000.11

[6] 内藤 正樹, Harald Singer, 山本 博史, 中嶋 秀治, 松井知子， 塚田 元， 中村 篤, 匂坂 芳典, “旅行会話タスクにおける ATRSPREC の性能評価”, 音講論,3-1-9,pp.113-114,Oct. 1999.

[7] T.Takezawa, E.Sumita, F.Sugaya, H.Yamamoto and S.Yamamoto, “Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world,” Proc. 3rd International Conference on Language Resources and Evaluation, pp.147-152, 2002.

[8] J.G. Fiscus, “A post-processing system to yield reduced word error rates: recognizer output voting error reduction(ROVER)”, ASRU 347-354,1997.

[9] 小窪 浩明，山本 博史，菊井 玄一郎，”ROVER 法を用いた音声認識結果の統合と誤認識文のリジェクション”，日本音響学会 2003 年度春季大会,2003.3.

[10] 清水 徹, 山本 博史, 政瀧 浩和, 松永 昭一, 匂坂 芳典, “大語彙音声認識のための単語仮説数削減”, 信学論 (D-II),vol.J79-D-II,No.12,pp.1355-1358,Dec. 1996.