音声言語情報処理 45−14
（２００３．２．８）

“反省型” 信頼性尺度に基づく 書き起こしなしデータを用いた言語モデル学習

須藤 克仁　東中 竜一郎　中野 幹生　相川 清明

日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
〒 243-0198 神奈川県厚木市森の里若宮 3-1

{sudoh,rh,nakano}@atom.brl.ntt.co.jp, aik@idea.brl.ntt.co.jp

概要

本稿では，特定のドメインを対象とした音声対話システムのための言語モデルの学習手法として，書き起こしのない発話の音声認識結果を “反省型” 信頼性尺度に基づいて選別し，学習に用いる手法を提案する．“反省型” 信頼性尺度とは，対話が終了した後で対話中の各発話に対する音声認識結果の信頼性を推定するための尺度である．“反省型” 信頼性尺度では，従来用いられてきた，音響スコア・言語スコアなどの発話単位での局所的な特徴量に加え，対話全体から得られる特徴量を用いて，音声認識結果の信頼性を推定する．実験により，“反省型” 信頼性尺度による発話単位での認識結果の信頼性推定精度の向上と，およびそれに基づいて書き起こしなしの発話の音声認識結果を選別し，言語モデル学習に用いたときの，音声認識の単語正解精度・コンセプト正解精度がそれぞれ従来手法より若干向上することを確認した．

キーワード: 音声認識，信頼性尺度，言語モデル，音声対話システム，“反省型” 信頼性尺度

Utilizing Untranscribed Utterances for Language Model Training based on a *review-based* Confidence Measure

Katsuhito SUDOH, Ryuichiro HIGASHINAKA, Mikio NAKANO and Kiyoaki AIKAWA

NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
3-1, Morinosato-Wakamiya, Atsugi, Kanagawa 243-0198, JAPAN

Abstract

This paper proposes a method for language model training, utilizing untranscribed utterances as training sentences based on a “*review-based* confidence measure”. After a dialogue, the “*review-based* confidence measure” is used for scoring confidence of a speech recognition result for each user utterance in the dialogue. It uses dialogue-level features in addition to utterance-level and word-level features used for conventional confidence scoring methods, including acoustic and lexical score for the recognition result. Experimental results show that using the proposed confidence measure improves utterance-level correct/incorrect classification accuracy, and word/concept accuracy of speech recognition with the trained language model slightly.

Keywords: speech recognition, confidence measure, language model, spoken dialogue system, review-based confidence measure

1　はじめに

音声認識技術・音声合成技術の進歩に伴い，システムとの音声対話による商用の情報提供サービスが開始されるなど，音声対話システムに対する期待は高まっている．今後，音声対話システムが一般的なインタフェースとして普及し，活用されていくためには，音声対話システムを構成する個々のコンポーネントの性能向上だけでなく，目的とするドメインのタスクを扱う音声対話システムに必要な言語モデルや言語理解モデルなどを可能な限り容易に構築でき，またそれらに対する改良が効率的に行えることが求められる．

本稿では言語モデルに着目し，議論する．新規に音声対話システムを構築する際に必要となる言語モデルの作成に関しては，システム設計者が記述した文法に基づくアプローチと，実際のユーザとシステムとの対話音声を書き起こし，大量のコーパスを用いて統計的言語モデルを作成・改良するというアプローチがある．前者のアプローチはドメイン

に依存するような言語モデルを比較的短い時間で作成することができるが，自然言語処理に通じた設計者が必要であることと，文法に含まれないような言い回しに対応するために，頻繁に言語モデルを手動で修正しなくてはならないという問題がある．後者のアプローチのような，対象とするドメインに関する大量の対話コーパスに基づいた統計的モデルの性能が高いことはよく知られている．しかし，大量の書き起こしを作成するためには膨大なコストがかかり，新しいドメインを対象としたシステムを作成するたびに大量の書き起こしデータを用いて言語モデルを作成することは容易ではない．

そのため，書き起こしを利用せず，類似したドメインの対話コーパスを利用して作成した言語モデルを用いて，キーフレーズ抽出によって言い回しに柔軟な音声理解を実現する手法が提案されている [1]．本稿では，類似したドメインの大規模コーパスの存在を仮定していないため，こうした手法を直接用いることはできない．

また，大量の書き起こしなしのデータを用いて言語モデルを学習する手法 [2] が提案されている．この手法では，書き起こしなしのデータに対する音声認識結果の中から，正しいと推定されるものを抽出し，発話コーパスとして言語モデル学習に利用する．収録した対話データに対して書き起こし作業をすることなく言語モデルを学習できることは，新しいドメインを対象としたシステムの作成において有効である．そこで，本稿でもこのようなアプローチにより，言語モデルを学習する手法を検討する．

こうした手法において音声認識結果が正しいかどうか推定するために用いられる尺度を信頼性尺度 (confidence measure) [1] と呼ぶ．信頼性尺度の作成には学習データが必要であり，そのための書き起こしを作成しなくてはならない．だが，言語モデルの学習のたびに大量の書き起こしを作成することと比較すればその作業量は小さく，実際のシステム構築においては有効である．

しかし，従来の研究で用いられてきた信頼性尺度は，そのための十分な精度を持っているとは言い難く，信頼性尺度に基づく言語モデル学習 [2][4] よりも精度の高い信頼性尺度を用いることで，さらなる効果をあげることが可能である．従来の信頼性尺度は，音声認識を行った直後に信頼性を評価するために用いられ，その評価値によって認識結果を棄却するなど，対話管理のために多く利用されてきた．こうした尺度では，対象となる単一の発話の認識結果に関する情報のみから信頼性が計算されており，局所的な情報のみから信頼性を評価せざるを得ないために，精度が不足しているものと考えられる．

ところで，本稿で検討するような，書き起こしなしデータを用いた言語モデルの学習においては，対話が終了してから，対話中の個々の発話についての信頼性評価を行って問題はない．そこで本稿では，Hazen らの用いた信頼性尺度を拡張し，対話が終了した後で対話全体から得られる大局的な情報を活用して，対話中の発話に対する音声認識結果の正しさを評価するための，“反省型” 信頼性尺度を提案し，それに基づいて書き起こしなしのデータを言語モデルの学習に用いるための手法について述べる．

以下，2 節では従来用いられてきた信頼性尺度と，本稿で提案する “反省型” 信頼性尺度との違いと，“反省型” 信頼性尺度を用いた信頼性の評価方法について述べる．3 節では信頼性尺度に基づいて，言語モデル学習のための例文を書き起こしなしのデータから抽出する手法について説明する．4 節では本手法を用いた実験の結果を示す．5 節はまとめである．

2 信頼性尺度

2.1 従来手法

音声認識の信頼性尺度に関する研究 [3][5][6] では，個々の発話の認識時の音響尤度，言語尤度などの指標を用いて信頼性を評価している．このような信頼性尺度は，対話の各時点でのユーザ発話の認識結果の信頼性を求め，信頼性が低い認識結果を棄却するなど，対話管理のために活用されている [3][7]．こうした研究では，音声認識時に得られる音響スコア，言語スコア，N-best 内の単語一致率などの，さまざまな特徴量を用いて信頼性を評価している．

Hazen らの手法では，音声認識結果の信頼性を，発話単位・単語単位で評価するために，発話単位・単語単位それぞれの特徴量ベクトルから，識別器を用いて評価値を与え，正誤の識別を行う．識別器のパラメタは，少量の書き起こしに正誤識別の正解ラベルを付与したデータを用いて学習する．単語単位で正しいと評価されるのは，単語仮説と書き起こしの正解単語が一致した場合，発話単位で正しいと評価されるのは，1) 4-best 以内に書き起こしと完全に一致する仮説がある場合，2) 2/3 以上の単語仮説が正解である場合，である．

Hazen らの提案した信頼性尺度は，数多くの特徴量と正解との関係を求め，定式化することで，音声認識結果の信頼性を評価し，信頼性の低い認識結果の棄却といった対話管理に応用されている．

2.2 提案手法:“反省型” 信頼性尺度

言語モデルの学習に用いる文を必要とするのは対話終了後であり，Hazen らが用いているような，発話を認識する時点で得られる局所的な特徴に加えて，対話全体から得られる特徴を用いることが可能である．

そこで本稿では，Hazen らの用いていた特徴に加え，対話全体から得られる特徴を用いた信頼性尺度を用いることにより，ユーザとシステムのやり取りの内容に基づいて，発話単位，また単語単位で音声認識結果の信頼性を評価することを考える．このような信頼性尺度を，システムが対話を終えた後，対話中の音声認識結果を省みることから，**“反省型” 信頼性尺度**と呼ぶことにする．

本稿で，“反省型” 信頼性尺度のために用いる特徴量を表 1 から表 4 に示す．表 1 および表 3 の特徴量は，各発話を認識した時点で計算可能な特徴量である．なお，U4,W8 は，Hazen らの手法では用いられていなかった局所的特徴量である．一方，表 2，表 4 の特徴量は，従来研究では用いられていなかった，すべて対話が終了した後で初めて計算できる特徴量である．U5,U6 の「対話終了時のユーザ要求」

[1] 信頼性尺度に基づく信頼性評価値を confidence score[3] と呼ぶ．

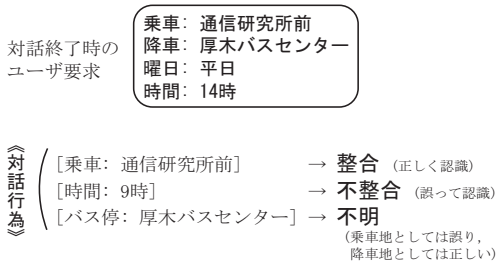

図 1: 対話終了時のユーザ要求と対話中の対話行為との整合・不整合

とは，対話が成功したときの，システムの理解状態である．それと整合するような対話行為はシステムがユーザ発話を正しく認識した結果得られたものであり，整合しないような対話行為はシステムがユーザ発話を誤って認識した結果得られたものであると考え (図 1)，認識結果に関する特徴量として，信頼性の評価のために用いる．

表 2 の特徴量は，ユーザの意図は対話を通じて一貫していて，ユーザは意図と反する内容を発話しないという仮定に基づいているが，仮定に反するような発話をユーザがした場合に，それが誤認識と同等に扱われて棄却されたとしても，十分な数の書き起こしなしデータを用意することでその影響を小さくすることが可能であるため，本稿では仮定に反する発話について特に考慮しない．

認識結果の信頼性評価には，Hazen らの手法と同じく識別器を用いる．その識別器のパラメタは，書き起こしを行った少量のデータを用いて学習する．識別の手法，パラメタの決定手法については，2.3 で述べる．なお，本稿では，認識結果が発話単位で正しいと評価される基準として，「認識結果から得られる対話行為が，書き起こしから得られる対話行為と等しい場合」を用いる．Hazen らの，「2/3 以上の単語仮説が正解である」という基準では，正しく内容を理解できないような認識結果であっても，正しいと識別してしまう可能性があるためである．

2.3 信頼性評価法

前節で述べたような特徴量ベクトルを用いて，発話単位・単語単位の信頼性は以下のように評価される [3].

まず，特徴量ベクトル f の各要素を，重みベクトル p を用いて線形結合し，スコア r を求める．

$$r = \vec{p}^T \vec{f} \tag{1}$$

認識結果が正しいという事前確率を $P(correct)$，認識結果が正しかったときに式 (1) から得られるスコアが r となる確率密度を $p(r|correct)$，認識結果が誤っているという事前確率を $P(incorrect)$，認識結果が誤っていたときに式 (1) から得られるスコアが r となる確率密度を $p(r|incorrect)$，とすると，しきい値 t に対し，認識結果の信頼性評価は，以下の識別式から得られる評価値 c として得られる．

$$c = \log\left(\frac{p(r|correct)P(correct)}{p(r|incorrect)P(incorrect)}\right) - t \tag{2}$$

c の値が正の場合，認識結果は正しく，c の値が負の場合，認識結果は誤っているという評価となる．

U1	最尤候補における 1 単語あたりの音響スコア平均
U2	最尤候補における各単語の、第 n 候補までの出現頻度の平均
U3	第 n 候補までの各単語の出現頻度の平均
U4	最尤候補における、正しく構文解析できなかった文節の割合

表 1: 各発話の認識時に得られる特徴量 (発話単位)

U5	最尤候補における、対話終了時のユーザ要求と整合する対話行為の割合
U6	最尤候補における、対話終了時のユーザ要求と矛盾する対話行為の割合
U7	最尤候補における各対話行為の、対話全体での出現頻度の平均

表 2: 対話終了後に得られる特徴量 (発話単位)

W1	単語内での音響スコアの最小値 (対数尤度)
W2	単語内での音響スコアの標準偏差 (対数尤度)
W3	単語内での最大の音響スコアからの差の平均 (対数尤度)
W4	単語内の音響信号観測サンプル数
W5	認識結果候補の数
W6	他認識結果候補での単語出現頻度
W7	(言語スコア＋音響スコア) の音響信号観測時間平均
W8	正しく構文解析できなかった文節においてこの単語が占める割合 (正しく解析された文節に含まれる単語の場合は 0)

表 3: 各発話の認識時に得られる特徴量 (単語単位)

W9	発話単位でのスコア (発話単位の特徴量がすべて揃うのが対話終了後であるため，対話終了後に得られる特徴量となる．)

表 4: 対話終了後に得られる特徴量 (単語単位)

ここで，確率密度 $p(r|correct)$，$p(r|incorrect)$ はそれぞれガウス密度関数で以下のように表現する．

$$p(r|correct) = \frac{1}{\sqrt{2\pi}\sigma_{correct}} exp\left(-\frac{(r-\mu_{correct})^2}{2\sigma^2_{correct}}\right) \tag{3}$$

$$p(r|incorrect) = \frac{1}{\sqrt{2\pi}\sigma_{incorrect}} exp\left(-\frac{(r-\mu_{incorrect})^2}{2\sigma^2_{incorrect}}\right) \tag{4}$$

重みベクトル p の各要素および事前確率 $P(correct)$，$P(incorrect)$，確率密度 $p(r|correct)$，$p(r|incorrect)$，しきい値 t はそれぞれ，書き起こしを用いて正誤のラベルが付加した認識結果を用いて，計算・学習する．p の初期値は Fisher の線形判別分析法より求め，識別誤差が最小となるように，p の各要素およびしきい値 t を更新するという繰り返し演算によって得る．

3 言語モデルの学習

前節で説明したような信頼性尺度に基づいて，対話中のユーザ音声に対する書き起こしのない音声認識結果から，統計的言語モデルの学習に用いる例文として適切なものを選ぶことを考える．

音声認識結果の信頼性は，発話単位，単語単位の双方について評価されるため，

1. 発話単位での信頼性評価値が 0 未満の認識結果を棄却する

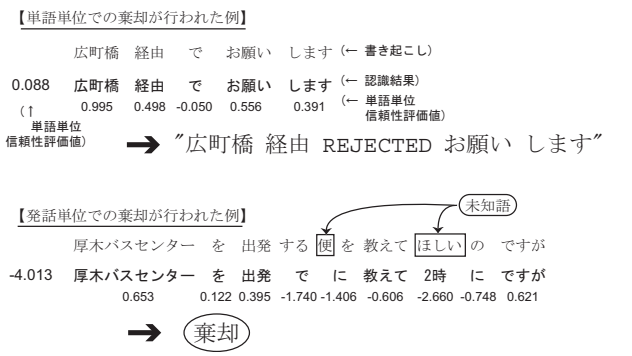

図 2: 信頼性評価値による認識結果の棄却

2. 単語単位での信頼性評価値が 0 未満の単語仮説は棄却し，シンボル REJECTED に置き換える

3. 複数個連接したシンボル REJECTED は，ひとつのシンボル REJECTED に置き換える

という形で，正しく認識された可能性の高い音声認識結果を，言語モデルのトレーニング文として抽出する (図 2).

こうして得られたトレーニング文を用い，統計的言語モデル (本稿では単語 N-gram) を学習する．実際には，トレーニング文中の各単語を，認識語彙中で同じ属性 (クラス) に属する単語と置き換えた文もトレーニング文として活用できるので，1 つのトレーニング文に対し，文中の単語を同じクラスに属する単語に置き換えた n 文を生成し，言語モデル学習に用いる．トレーニング文は，信頼性尺度が用意されていれば，対話収録後に自動的に計算することが可能であるため，収録した大量のデータも，書き起こしなしで言語モデルの学習に利用することができる．

また，信頼性尺度の学習のために用いた発話に対する書き起こしは，そのままトレーニング文として言語モデル学習に利用することができる．

4 実験

4.1 実験の内容

ユーザと音声対話システムとの対話収録データを用いて，実験を行った．収録に用いた音声対話システムは，音声対話システム作成ツールキット WIT[8] を用いて構築されたバス時刻表案内システムである．音声認識エンジンには Julius3.3 を用いた．認識語彙数は 145，初期言語モデルはネットワーク文法からランダムに生成された例文を用いて作成した．総収録対話数は 722(84 名，7341 発話) である．

このデータの中から，

(A) 信頼性尺度 (識別器) のパラメタ学習に用いる書き起こしありの学習データセット (CM-trainset)

(B) 言語モデルの学習に用いる書き起こしなしデータセット (LM-trainset)

(C) 評価用データセット (testset)

を重なりがないように選び，実験を行った．実験の内容は，以下の 2 種類である．

信頼性尺度の精度評価 (A) の書き起こしを用いてパラメタを学習した，発話単位・単語単位それぞれの信頼性尺度を用いて (C) の認識結果を評価し，(C) の書き起こしから得られる発話単位・単語単位での正誤とどの程度一致するか，Hazen らの信頼性尺度 [3] と比較する．

学習した言語モデルの性能評価 (A) の書き起こしを用いてパラメタを学習した信頼性尺度に基づいて，3 節の手法で (B) から得られるトレーニング文と，(A) の書き起こしから学習した言語モデルを用いて音声認識を行ってその認識精度を評価し，『(A) の書き起こしのみを用いて言語モデルを学習した場合』，『Hazen らの信頼性尺度に基づく言語モデル学習を行った場合』，『(A)(B) に対する書き起こしを用いた場合』と比較する．

4.2 信頼性尺度の精度評価

図 3 に，式 2 のしきい値 t を変化させ，“反省型” 信頼性尺度，および Hazen らの信頼性尺度に基づいて，(C) に対する音声認識結果を発話単位・単語単位で正誤分類したときの，正分類の適合率 (precision)，再現率 (recall) を示す．

また，図 4 に，(A) のデータ量を変化させ，発話単位・単語単位でそれぞれ正誤分類したときの正解率 (classification accuracy) を示す．

図 3, 図 4 から分かるように，本稿で提案する “反省型” 信頼性尺度を用いた場合，Hazen らの信頼性尺度と比較して，発話単位では平均で約 4.5% 分類精度が向上している．一方，単語単位での精度の向上はほとんど見られない．これは，単語単位の信頼性尺度で用いられる特徴量は，表 4 の W9(発話単位でのスコア) の値の違いのみであること，さらに，発話単位での正誤が分かっていたとしても，本稿のシステムの言語理解部が助詞の挿入や脱落に対して敏感でなく，挿入や脱落のある文でも正しい理解結果を出力してしまうことがあるために，理解結果に影響を与えないような助詞などの正誤分類は非常に難しいこと，が原因と考えられる．

また，(A) の学習データ量が増加しても，精度改善の度合いはそれほど大きくないことも確認できる．500 発話程度の学習データ量でおおむね問題ないと考えられるため，次節の実験では 500 発話程度のデータを (A) として用いることにする．

4.3 言語モデルの性能評価

(A) として 6 名分の対話データ (約 500 発話) を 4 セット，(B) として 12 名分の対話データ (約 1000 発話) を 4 セット，(C) として 12 名分の対話データからランダムに選んだテストセット (約 200 発話) を 3 セット，それぞれ用意し，言語モデルの評価を行った．以下の図では，(A) のあるセットに対して，それぞれ (B)(C) の組み合わせで 12 通りの実験を行った結果の平均を示し，(A) のセットの違いによる言語モデルの性能を比較している．

言語モデルは，トレーニング文 1 文に対し，文中の単語を同クラスの単語に置き換えた 50 文を生成したテキストに対し，CMU-Cam SLM Toolkit を用いて作成した．

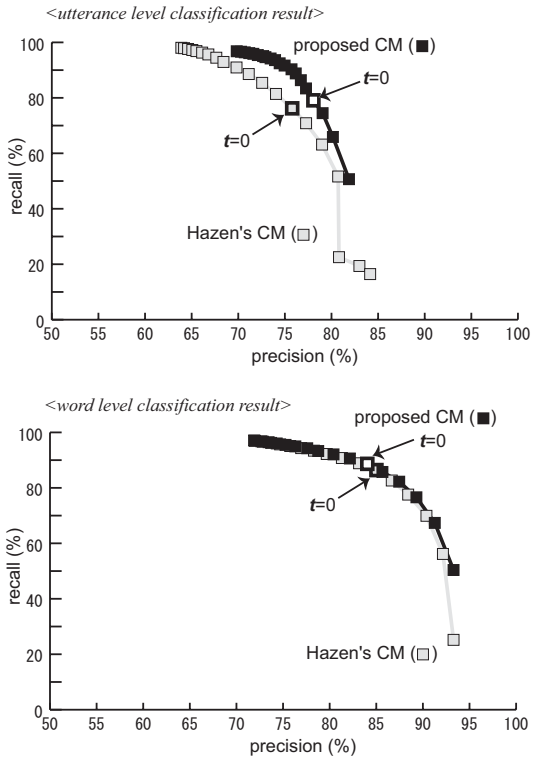

図 3: 信頼性尺度に基づく正誤分類の適合率・再現率 (上: 発話単位 / 下:単語単位)

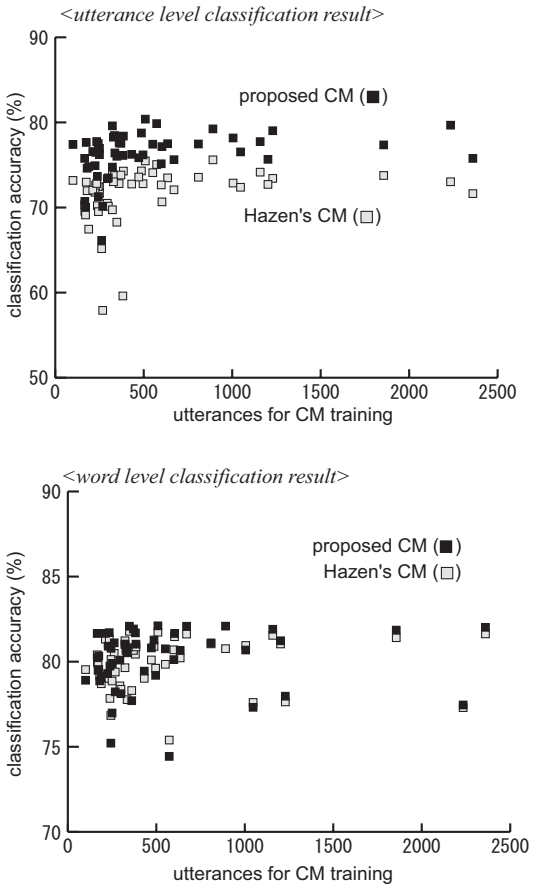

図 4: 信頼性尺度学習データ量と信頼性尺度の精度の関係 (上:発話単位 / 下:単語単位)

CM-trainset	トレーニング文数	適合率
CM-trainset1	649	74.87%
CM-trainset2	708	75.26%
CM-trainset3	715	75.15%
CM-trainset4	683	76.91%

表 5: 抽出されたトレーニング文数と適合率の平均 (Hazen の信頼性尺度)

CM-trainset	トレーニング文数	適合率
CM-trainset1	657	80.17%
CM-trainset2	731	77.30%
CM-trainset3	741	79.51%
CM-trainset4	604	82.55%

表 6: 抽出されたトレーニング文数と適合率の平均 ("反省型" 信頼性尺度)

評価セット (C) に含まれる語彙外単語 (OOV) の割合は，平均で 7.3%である．また，Julius での認識時の音響モデルは Julius3.3 付属の JNAS 性別非依存 PTM トライフォンモデル (64 混合，3000 状態) を用い，言語重みと挿入ペナルティは，第 1 パスで 9.0,-6.0，第 2 パスで 10.0,-6.0 とした．なお，対話収録時に用いた言語モデルを用いて (C) に対する音声認識を行ったときの単語正解精度は 48.60%，コンセプト正解精度は 60.26%であった．

表 5，表 6 に，Hazen の信頼性尺度および "反省型" 信頼性尺度によって，書き起こしなしデータ (B) から抽出された言語モデルトレーニング文の数と，その中で実際に正しい発話認識結果であったものの割合の，4 セットの (B) についての平均をそれぞれ示す．

図 5, 図 6 に，信頼性尺度に基づいて学習された言語モデルを用いて，テストセットに対する音声認識を行ったときの，単語正解精度，コンセプト正解精度をそれぞれ示す．比較対象として，(A) に対する書き起こしのみで作成した言語モデル，(A)(B) に対する書き起こしをすべて用いて作成した言語モデルを用いた場合の結果も合わせて示す．図から分かるように，書き起こしなしのデータを言語モデル学習に用いることで，単語正解精度，コンセプト正解精度ともに向上が見られた．また，"反省型" 信頼性尺度を用いた場合，Hazen の信頼性尺度を用いた場合より，若干だが改善が見られた．

4.4 考察

"反省型" 信頼性尺度を用いた場合，発話単位での正誤判定の精度が約 4.5%向上することを確認した．したがって，"反省型" 信頼性尺度は対話終了後に対話中の音声理解結果を評価する尺度として有効である．

単語単位での信頼性尺度の精度はあまり変化しないことも確認できた．"反省型" 信頼性尺度のために新たに提案した特徴量は，理解単位である文節の定義との関係が強く,「平日 に」というユーザ発話が「平日」「平日 で」「平日 に は」のように認識されても理解結果に影響を与えないといったように，助詞の挿入や脱落に対してあまり敏感でないためと考えられる．

また，Hazen らの信頼性尺度に基づいて作成した言語モデルを用いた場合と，"反省型" 信頼性尺度に基づいて作成

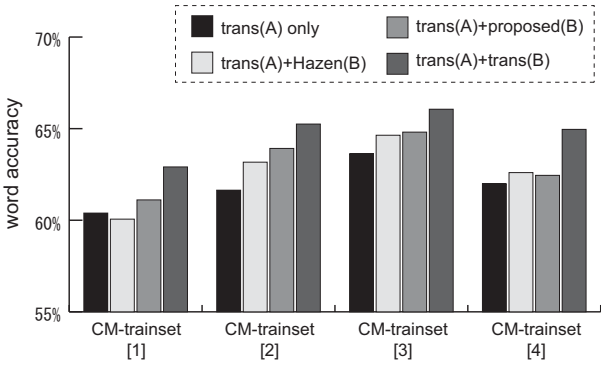

図 5: 音声認識時の単語正解精度

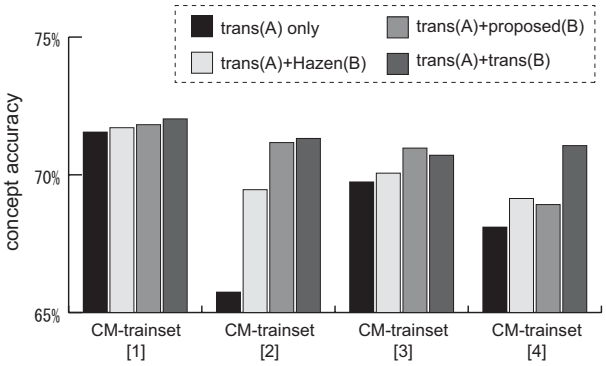

図 6: 音声認識時のコンセプト正解精度

した言語モデルを用いた場合の音声認識結果の差は，用いた信頼性尺度の学習データに依存して変化している．表 5, 表 6 に示したように，信頼性尺度に基づいて抽出されたトレーニング文の適合率は “反省型” 信頼性尺度を用いたもののほうが高いため，その差を音声認識結果の差として反映させられるよう，抽出されたトレーニング文の言語モデル学習への適用方法に検討を加える必要がある．

5 おわりに

本稿では，対話が終了した後に対話全体を省みて得られる情報を用いて音声認識結果の信頼性を評価するための，“反省型” 信頼性尺度を提案し，それに基づいて書き起こしなしのデータから信頼性の高い認識結果を抽出し，言語モデルの学習に利用する手法について述べた．

実験結果から，“反省型” 信頼性尺度は発話単位での音声認識結果の正誤判定において，従来手法よりも高い性能を持つことを確認した．しかし，単語単位での音声認識結果の正誤判定では，あまり差が見られなかった．

また，言語モデル学習については，“反省型” 信頼性尺度を用いることで，従来手法よりも信頼性の高い音声認識結果をより多くトレーニング文として抽出できることを確認した．しかし，それを言語モデル学習に用いたときの性能差は小さく，さらなる改善の余地がある．

“反省型” 信頼性尺度では，音声認識結果の信頼性を評価するための特徴量として，対話終了時のユーザ意図と，対話中の音声理解結果の整合性のような，音声対話システムの言語理解部に依存するものを用いている．文法や理解規則などの違いにより，こうした特徴量の有効性は変化すると考えられるので，言語理解部の特性と，“反省型” 信頼性尺度の性能との関係について検討する必要がある．

今後の課題としては，信頼性評価に用いる特徴量を新たに加える，逆に寄与の小さい特徴量を除くことで，信頼性尺度の精度をさらに向上させること，こうした手法において有効な言語モデル学習法のさらなる検討を加え，言語モデルの性能と，音声理解精度を向上させることが挙げられる．また，能動学習による言語モデル学習 [4][9] において，“反省型” 信頼性尺度を学習サンプル選択基準として用いて学習効率を向上させることも有効である．さらに，“反省型” 信頼性尺度を用いて，言語モデル以外のものの学習を実現することも考えられる．

謝辞

日頃よりご指導いただく村瀬洋メディア情報研究部長，有益なアドバイスをいただくマルチモーダル対話研究グループの皆様に感謝いたします．また，信頼性尺度作成ツールを提供していただいた，MIT 音声言語システムグループの皆様に感謝いたします．

参考文献

[1] Kazunori Komatani, Katsuaki Tanaka, Hiroaki Kashima, and Tatsuya Kawahara. Domain-independent spoken dialogue platform using key-phrase spotting on combined language model. In *Eurospeech-2001*, pp. 1319–1322, 2001.

[2] Roberto Gretter and Giuseppe Riccardi. On-line learning of language models with word error probability distributions. In *ICASSP-01*, Vol. I, 2001.

[3] Timothy J. Hazen, Stephanie Seneff, and Joseph Polifroni. Recognition confidence scoring and its use in speech understanding systems. *Computer Speech and Language*, Vol. 16, pp. 49–67, January 2002.

[4] Mikio Nakano and Timothy J. Hazen. Utilizing untranscribed user utterances for improving language models based on confidence scoring. 言語処理学会年次大会, 2003, to appear.

[5] Thomas Kemp and Thomas Schaaf. Estimating confidence using word lattices. In *Eurospeech-97*, Vol. 2, pp. 827–830, 1997.

[6] A. Wendemuth, G. Rose, and J.G.A. Dolfing. Advances in confidence measures for large vocabulary. pp. 705–708, 1999.

[7] 駒谷和範, 河原達也. 音声認識結果の信頼度を用いた効率的な確認・誘導を行う対話管理. 情報処理学会論文誌, Vol. 43, No. 10, pp. 3078–3086, 2002.

[8] Mikio Nakano, Noboru Miyazaki, Norihito Yasuda, Akira Sugiyama, Jun-ichi Hirasawa, Kohji Dohsaka, and Kiyoaki Aikawa. WIT: A toolkit for building robust and real-time spoken dialogue systems. In *First SIGdial Workshop on Discourse and Dialogue*, pp. 150–159, 2000.

[9] Dilek Hakkani-Tür, Giuseppe Riccardi, and Allen Gorin. Active learning for automatic speech recognition. In *ICASSP-02*, Vol. IV, pp. 3904–3907, 2002.