論　文

認識信頼度と対話履歴を用いた音声言語理解手法

藤原　敬記[†]　伊藤　敏彦[†]　荒木　健治[†]　甲斐　充彦[††]
小西　達裕[†††]　伊東　幸宏[†††]

Spoken Language Understanding Method Using Confidence Measure and Dialogue History

Noriki FUJIWARA[†], Toshihiko ITOH[†], Kenji ARAKI[†], Atsuhiko KAI[††], Tatsuhiro KONISHI[†††], and Yukihiro ITOH[†††]

あらまし　実環境での音声対話システムの使用において，誤認識を回避することは難しい．誤認識が起きると，システムはユーザの期待する応答とかけ離れた応答を行い，対話がスムーズに進まなくなることも多い．そこで本研究では，音声認識器が誤認識した場合でも，認識信頼度と対話履歴を用いることで正しくユーザの意図を推定することができる音声言語理解手法を提案する．これは，音声認識器が誤認識した場合でも多くの場合，複数候補（N-best）中に正解が含まれていること，システムが誤認識した場合にはユーザは大体訂正反応を示すこと，タスク指向対話には強い一貫性がありユーザは基本的に意味的・文脈的に関係した内容以外を発話しないことを利用する．また，提案手法ではあらかじめすべての認識可能単語を理解候補として保持し，言語理解部の対話戦略において音声認識結果中の単語との意味的関連性などを考慮している．これにより音声認識結果のN-best中に正解の一部が含まれていない場合でも，複数のユーザ発話の認識結果に基づくことで正しい意図を推定することが可能となっている．評価データにおいて，提案手法における対話単位での理解率は72.2%（21,430/29,670対話），単語単位での理解率は87.1%（77,544/89,010単語）であり，従来手法の最新認識結果の上位候補を優先するシステムの57.9%（17,178/29,670対話），75.4%（67,084/89,010単語）と比較しても有効である．

キーワード　音声言語理解，音声対話システム，誤認識，認識信頼度，対話履歴

1. まえがき

近年，音声研究の進歩とコンピュータの高速化を背景に音声対話システムの実用化が注目されている[1]～[4]．音声インタフェースは，特別な訓練なしに利用できることや目や手などを他のタスクに使用している状況でも利用できるといった利点がある．しかし一方で，誤認識や合成音声の自然性などの問題を抱えている．特に実環境下という状況では，ユーザの自由発声による影響や他話者の発話，周囲の雑音などにより，防音室内での読上げ音声に比べ誤認識は多く起きる．誤認識結果に基づいて言語理解を行えば，システムはユーザの意図を正しく推定することができないため，ユーザの予想とは大きくかけ離れた応答をすることになる．その結果，ユーザとシステムの間でのスムーズな対話が成り立たなくなることや，目的そのものの達成が困難になることがある．

誤認識を軽減することを目的とした研究は多く行われている．例えば音声入力時における他話者の存在や周囲雑音の影響を少なくするために，指向性マイクロホンを使用することや，発話に雑音が混ざることを考慮し，音声認識器の雑音重畳音声を用いた学習[5]といった研究が行われている．また音声対話システムにおいて，音声認識器から得られる認識結果がどの程度確からしいかを表す尺度である認識信頼度が使われることも多い[6]～[8]．しかし現在の技術では誤認識の発生を軽減することはできても，完全に取り除くことは難しい．

このため実環境下での音声対話システムの使用にお

[†] 北海道大学情報科学研究科，札幌市
Graduate School of Information Science and Technology, Hokkaido University, Sapporo-shi, 060–0814 Japan
[††] 静岡大学工学部，浜松市
Faculty of Engineering, Shizuoka University, Hamamatsu-shi, 432–8561 Japan
[†††] 静岡大学情報学部，浜松市
Faculty of Informatics, Shizuoka University, Hamamatsu-shi, 432–8011 Japan

いては，ある程度の誤認識は発生することを考慮した言語理解を行う必要がある．従来研究においても，システムの音声認識部による誤認識や言語理解部による誤理解を検知するために，ユーザ発話の音響的な特徴を用いる研究[9],[10]や，ユーザの繰返し発話を検知する研究[11]，質問–応答のようなやり取りからシステムの誤理解を発見する研究[12]などが行われている．またタスク指向の対話においては，対話には強い一貫性があるので，ユーザの意図推定に文脈情報を利用する[13]～[17]ことも行われている．更に，対話制御からのアプローチとして，疑わしい内容については応答しないといった対話戦略を用いる研究[18]もある．

本研究では，音声認識器が誤認識した場合でも多くの場合，複数候補（N-best）中に完全な正解，または部分的な正解が含まれていること[19]，システムが誤認識した場合にはユーザは大体訂正反応を示すこと[20]，タスク指向対話には一貫性があり基本的に意味的・文脈的に関係した内容以外を発話しないこと[21]を利用し，音声認識器が誤認識した場合でも，正しくユーザの意図を推定する音声言語理解手法を提案する．提案手法では，理解候補とした認識可能単語（キーワード）すべてに，それぞれの単語が，ある一つのゴール（本論文では一つの目的地設定）を達成するためにどの程度使われている可能性があるかを表すスコアを付与し，そのスコアにユーザが発話するたびに得られる最新発話の認識信頼度を反映させることで，対話の流れを考慮したユーザの意図理解を行う．

上述した多くの手法が音声認識結果のN-bestをリスコアリングすることによって誤認識から正しい認識・理解結果を推定している．そのため，N-best中に正しい認識結果が存在しない限り正しい理解をすることができない．それに比べ，提案手法ではあらかじめすべての認識可能単語を理解候補として保持し，音声認識結果のN-best中に完全な正解が含まれていない場合でも，理解候補と認識結果の意味的関連性やシステムとの対話を考慮する言語理解戦略により，N-best中に含まれない単語でも理解候補の上位にすることができ，その結果正しい意図を推定することが可能となっている．

本論文では音声認識結果に誤認識を含む場合でもユーザの意図を正しく推定するために，音声認識結果の認識信頼度と，ユーザの第1発話からの音声認識結果である認識履歴とシステム応答からなる文脈情報（以降，対話履歴）を用いて言語理解を行う手法を提案し，模擬的な対話データを用いた評価実験について述べる．

2. システムのタスク

2.1 対話タスク

本論文では，対話システムの扱うタスクをカーナビゲーションシステムの目的地（以降，ランドマーク）設定とする．カーナビゲーションシステムを使用する状況では，目や手は運転という主たる動作に使っており，画面のメニュー操作やリモートコントローラーによる操作よりも音声インタフェースが有効である．また走行音や同乗者の発話などの定常的・非定常的な雑音により，誤認識が起きやすい．

今回，ランドマークとして使用できる単語は，インターチェンジ（以降，IC）名，駅名，市区町村名とした．更にこれらのランドマークには，県名，自動車道名，鉄道路線名などの属性を付与することができる．例えば「浜松西ICへ行く」や「東名自動車道の浜松西IC，静岡県の」などをユーザ発話として想定している．以降，ランドマークとして使用できる単語（IC名，駅名，市区町村名）と付与することができる属性（県名，自動車道名，鉄道路線名）を合わせて地名単語と呼ぶ．各地名単語はより抽象的な概念である“クラス”に属しており，更に各クラスは一つの“カテゴリー”に属しているものとする．これらは図1のように，3カテゴリー（Prefecture, Highway-Railway, Landmark）・6クラス（県，自動車道，鉄道路線，IC，市区町村，駅）に分類でき，地名単語は基本的に木構造をなしている．ユーザ発話としては，Prefecture, Highway-Railway, Landmarkカテゴリーの各単語を一括で発話する場合や，複数のターンに分割して発話する場合なども想定している．

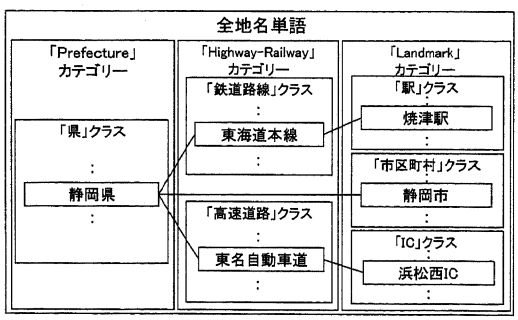


図1 発話内容の分類
Fig. 1 A classification of speech contents.

表 1　発話タイプ
Table 1　Speech type.

発話タイプ	解説と対話例
詳細化	システムの応答に対して新情報を追加する発話 U1：　静岡県の S1：　静岡県 U2：　東名自動車道の浜松西 IC へ行きます
訂正	システムの応答に対して訂正を行う発話 U1：　東名自動車道の浜松西 IC S1：　東名自動車道の浜松 IC を 設定してよろしいですか？ U2：　いいえ，浜松西 IC です
回答	システムの質問に対して回答を行う発話 U1：　東名自動車道の浜松西 IC に行きます S1：　東名自動車道の何 IC ですか？ U2：　浜松西 IC です
再入力	システムの応答で再入力要求をされた後の発話 U1：　浜松西 IC に行きます S1：　もう一度発話して下さい U2：　浜松西 IC

U：ユーザ発話，S：システム発話

2.2　発話タイプ

本タスクにおいては，すべてのユーザ発話は詳細化・訂正・回答・再入力発話の 4 種類に分類することができる．詳細化発話とはシステムの応答に対して新情報を追加する発話，訂正発話とはシステムの応答に対して訂正を行う発話，回答発話とはシステムの「何 IC ですか？」のような質問に対して回答を行う発話，再入力発話とは「もう一度発話して下さい」とシステムに再入力を要求された後の発話のことである．それぞれの発話タイプの例を表 1 に示す．

3.　システムの構成

本システムの概要を図 2 に示す．本システムは音声認識部・信頼度生成部・言語理解部・応答生成部・音声合成部・GUI 表示部からなる [22]．

音声認識部は，入力された音声から音響的ゆう度で順位付けされた複数の候補（N-best 候補）を出力する．音声認識部には SPOJUS [23] を用いる．

信頼度生成部では，音声認識部から出力されたゆう度付の N-best 候補から，各単語の単語認識信頼度と，その単語の属するクラスのクラス認識信頼度を求める．認識信頼度は，各候補 w のゆう度スコア $P(x|w)$ と N-best 中の出現頻度から事後確率に基づく尺度 $P(w|x)$ として計算される [6]．

言語理解部では認識された単語の各認識信頼度と対話履歴，更に対話理論，世界知識，周辺状況を用いて，ある一つのゴール（本論文では一つの目的地設定）を

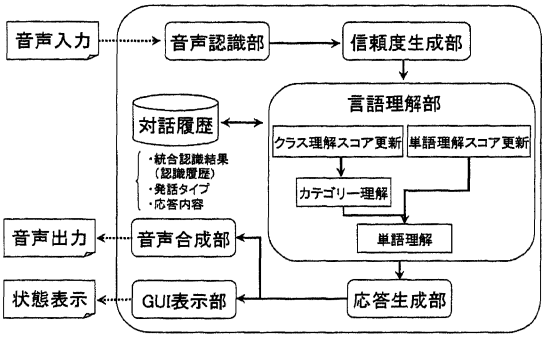


図 2　システムの概要
Fig. 2　Outline of speech dialogue system.

達成するまでにその単語やクラスが使用されている可能性の度合（以降，単語理解スコア・クラス理解スコア）を加算・減算し，1 ターン前の単語理解スコア・クラス理解スコアを更新する．次にクラス理解スコアからどのカテゴリーが発話されているかを推定するためのカテゴリー理解を行う．最後にカテゴリー理解の結果と単語理解スコアを用いて，ユーザがどのような内容を発話したかを推定し，理解結果として出力する．詳しくは次章で述べる．

応答生成部では言語理解部の理解内容と単語理解スコア・クラス理解スコアを用いて，ユーザへの確認や質問，次発話を促すための応答生成を行う．ここでは様々な対話戦略に基づき，最も適当だと思われる応答文を生成する．例えば，理解内容としては最も可能性が高い情報（ユーザが発話したと思われる情報）でも，信頼性が低い情報はユーザ側に伝達せず，とりあえず次発話を促す「分かったふり戦略」がある．この対話戦略により，システムが間違っているかもしれない情報はユーザには提示されず，ユーザから更に情報を引き出すことができる．この戦略は，常に入力要求や正誤確認を行うよりもさりげなく聞き返すことや，不自然に感じない他の情報を尋ねることを行い，総合的に判断する対話制御がユーザ満足度を高くする [18] という考えと同様のものである．

GUI 表示部と音声合成部では，現在の理解状態の表示や合成音声によるユーザへの情報伝達を行う．

4.　言語理解部

ここでは言語理解部の詳細について述べる．言語理解部は前章で述べたように，まず音声認識結果の認識信頼度と対話履歴を用いて単語理解スコア・クラス理

解スコアを更新する．次に，クラス理解スコアからカテゴリー理解を行う．最後にカテゴリー理解結果と単語理解スコアを用いた単語理解が行われることで，言語理解内容が生成される．この単語理解スコア・クラス理解スコアは，言語理解の他にも応答生成においても使用される．

ユーザが新たな情報を追加した場合，システムはその情報だけでなく，以前に発話された情報との統合が必要になる．これに対し，訂正発話が行われた場合には，以前の理解内容を修正する枠組みが必要である．提案手法ではこれらの枠組みを単語理解スコア・クラス理解スコアの増減によって実現している．更新された単語理解スコア・クラス理解スコアは，履歴に基づく統合認識結果として対話履歴に残される．

4.1 クラス理解スコア更新

クラス理解スコアの更新では，ユーザが発話する（ターン t）たびに対話履歴と最新の認識結果から発話タイプ（表 1）を判定し，発話タイプごとに理解スコア更新式をすべてのクラスに対して適用する．発話タイプは，以前の情報に新しい情報を追加する働きが必要な詳細化・回答発話と，以前の情報を訂正する働きが必要な訂正・再入力発話の 2 種類に分類した．発話タイプの判定には，表 2 の四つの判定材料を用いている．これらの判定材料は発見的に求めたものであるため，これ以外の判定材料も存在する可能性はある．

4.1.1 詳細化・回答のクラス理解スコア更新式

詳細化・回答発話は前述のとおり，以前の情報に新しい情報を追加する必要がある．これらの発話タイプと判定された場合，クラス c の理解スコア更新式は以下のとおりである．なお，すべてのクラスとそのクラス理解スコアはあらかじめもっており，理解スコアにはある一つのゴール（本論文では一つの目的地設定）が始まるたびに初期値として 0 を与えている．またクラス理解スコアの上限や下限は設定していない．

$$Score_t(c) = Score_{t-1}(c) * weight_{na} + Conf_t(c) \quad (1)$$

表 2 発話タイプ判定
Table 2 Judgement of speech type.

判定材料	判定結果
応答が“もう一度発話してください”	訂正・再入力
認識結果に否定語が存在する	訂正・再入力
別のカテゴリーが発話された	詳細化・回答
それ以外	訂正・再入力

$Score_t$：対話履歴のクラス理解スコア

$Conf_t$：最新認識結果のクラス認識信頼度

$weight_{na}$：重み $(0.0 < weight_{na} < 1.0)$

c：スコアを更新するクラス

以前に発話された内容を考慮するために，対話履歴のクラス理解スコアと最新認識結果のクラス認識信頼度を足す．重み $weight_{na}$ により一定の割合で対話履歴のクラス理解スコアを下げているのは，“情報が古くなるごとに信頼性が低下する”という戦略を適用しているためである．この重み $weight_{na}$ は，次章で説明する対話データを用いて，カテゴリー理解精度が最も高くなる値を求めた．更新されたクラス理解スコアは統合認識結果として，対話履歴に残される．

4.1.2 訂正・再入力のクラス理解スコア更新式

訂正・再入力発話の更新式も，基本的には詳細化・回答発話と同じである．異なる点は，同カテゴリー異クラスの認識信頼度をマイナスしていることである．これにより，クラスを間違っていた場合に理解スコアが修正されやすくなる．

$$Score_t(c_a) = Score_{t-1}(c_a) * weight_{cr} - Conf_t(c_b) + Conf_t(c_a) \quad (2)$$

$Score_t$：対話履歴のクラス理解スコア

$Conf_t$：最新認識結果のクラス認識信頼度

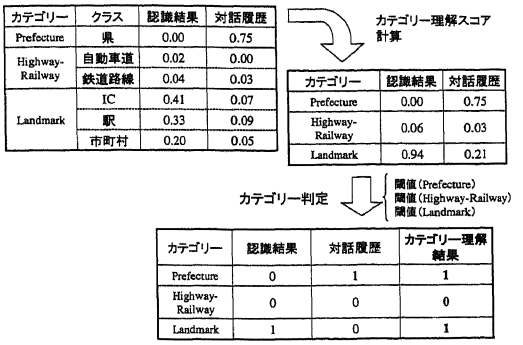
$weight_{cr}$：重み $(0.0 < weight_{cr} < 1.0)$

c_a：スコアを更新するクラス

c_b：c_a と同じカテゴリーで異なるクラス

4.2 カテゴリー理解

カテゴリー理解は，これまでの発話でどのような情報に関する発話がされたかを大まかに知るために，カテゴリーレベルでの理解を行うためのものである．図 3 にカテゴリー理解の例を示す．カテゴリー理解部では，対話履歴のクラス理解スコアと最新認識結果のクラス認識信頼度の両方に対して，カテゴリー理解スコアを計算する．対話履歴におけるカテゴリー理解スコアは，同じカテゴリーに属するすべてのクラス理解スコアを足したものであり，最新の認識結果においては同じカテゴリーに属するすべてのクラス認識信頼度を足したものである．それぞれのカテゴリー理解スコアはそれぞれのしきい値で判定され，Prefecture, Highway-Railway, Landmark の各カテゴリーに対し



カテゴリー	クラス	認識結果	対話履歴
Prefecture	県	0.00	0.75
Highway-Railway	自動車道	0.02	0.00
	鉄道路線	0.04	0.03
Landmark	IC	0.41	0.07
	駅	0.33	0.09
	市町村	0.20	0.05

カテゴリー	認識結果	対話履歴
Prefecture	0.00	0.75
Highway-Railway	0.06	0.03
Landmark	0.94	0.21

カテゴリー	認識結果	対話履歴	カテゴリー理解結果
Prefecture	0	1	1
Highway-Railway	0	0	0
Landmark	1	0	1

図 3　カテゴリー理解の例
Fig. 3　An example of category understanding.

て判定結果の論理和を計算する．そこで得られた結果が，対話開始から現在までに発話されたカテゴリー内容として理解される．

4.3　単語理解スコア更新

言語理解部は，システムが理解すべき全単語と各単語に対する単語理解スコアをあらかじめもっている．この理解スコアには，ある一つのゴール（本論文では一つの目的地設定）が始まるたびに初期値として 0 を与えている．こうすることで，たとえ誤認識によりユーザの発話した情報の一部が認識結果に含まれない場合でも，以下に述べる戦略により関連単語としてその単語の単語理解スコアが上昇し，ユーザ発話を正しく推測できる場合がある．本論文におけるシステムが理解すべき全単語とは，すべての地名単語である．言語理解部は最新の認識結果を獲得するたびに，単語理解スコア更新を行う．システムの応答内容とユーザ発話タイプ（詳細化，訂正，回答，再入力）から，既存の単語理解スコアを上下させて，新しい単語理解スコアを更新する．単語理解スコアの更新には，以下の 10 種類の戦略を用いて行う．なお，単語理解スコアの上限値や下限値は設定していない．

- 戦略 1：古い情報は信頼性が低くなるという仮定のもとに，新しい認識結果が入力されるたびに，対話履歴中のすべての単語理解スコアを下げる．戦略 1 が適用される場合の単語 w_A の単語理解スコア更新式は以下のとおりである．

$$Score_t(w_A) = Score_{t-1}(w_A) - weight_1 \quad (3)$$

$Score_t$：対話履歴の単語理解スコア

$weight_1$：重み $(0.0 < weight_1 < 1.0)$

w_A：単語 A

- 戦略 2：対話履歴中の単語 A と最新認識結果の単語 B が意味的（本論文では地理的）に関連がある場合，単語 A の単語理解スコアを上げる．戦略 2 が適用される場合の単語 w_A の単語理解スコア更新式は以下のとおりである．

$$Score_t(w_A) = Score_{t-1}(w_A) + weight_2 * Conf_t(w_B) \quad (4)$$

$Score_t$：対話履歴の単語理解スコア

$Conf_t$：最新認識結果の単語認識信頼度

$weight_2$：重み $(0.0 < weight_2 < 1.0)$

w_A：単語 A

w_B：単語 B

この戦略は単語 A と単語 B が同一単語である場合にも適用されるが，単語 A と単語 B が同一単語である場合は，他の場合と比べて意味的な関連の強さは異なる．そこで，単語 A と単語 B が同一単語である場合だけは $weight_2$ の代わりに専用の重み $weight'_2$ を使う．

- 戦略 3：対話履歴中の単語 A と認識結果の単語 B が，意味的（地理的）に関連がない場合，単語 A の単語理解スコアを下げる．戦略 3 が適用される場合の単語 w_A の単語理解スコア更新式は以下のとおりである．

$$Score_t(w_A) = Score_{t-1}(w_A) - weight_3 * Conf_t(w_B) \quad (5)$$

$weight_3$：重み $(0.0 < weight_3 < 1.0)$

- 戦略 4：認識結果に肯定語（「はい」,「うん」など）が含まれていた場合，システム応答に含まれていた単語 A の単語理解スコアを上げる．戦略 4 が適用される場合の単語 w_A の単語理解スコア更新式は以下のとおりである．

$$Score_t(w_A) = Score_{t-1}(w_A) + weight_4 * Conf_t(w_{yes}) \quad (6)$$

$weight_4$：重み $(0.0 < weight_4 < 1.0)$

w_{yes}：肯定語

- 戦略 5：認識結果に否定語（「いいえ」,「違う」など）が含まれていた場合，システム応答に含まれていた単語 A の単語理解スコアを下げる．戦略 5 が適用される場合の単語 w_A の単語理解スコア更新式は以

下のとおりである．

$$Score_t(w_A) = Score_{t-1}(w_A) - weight_5 * Conf_t(w_{no}) \quad (7)$$

$weight_5$：重み $(0.0 < weight_5 < 1.0)$

w_{no}：否定語

- 戦略 6：認識結果の単語 A とシステム応答に含まれる単語 B が意味的（地理的）に関連がある場合，単語 B の単語理解スコアを上げる．戦略 6 が適用される場合の単語 w_B の単語理解スコア更新式は以下のとおりである．

$$Score_t(w_B) = Score_{t-1}(w_B) + weight_6 * Conf_t(w_A) \quad (8)$$

$weight_6$：重み $(0.0 < weight_6 < 1.0)$

- 戦略 7：認識結果の単語 A とシステム応答に含まれる単語 B が意味的（地理的）に関連がない場合，単語 B の単語理解スコアを下げる．戦略 7 が適用される場合の単語 w_B の単語理解スコア更新式は以下のとおりである．

$$Score_t(w_B) = Score_{t-1}(w_B) - weight_7 * Conf_t(w_A) \quad (9)$$

$weight_7$：重み $(0.0 < weight_7 < 1.0)$

- 戦略 8：システム応答が質問で（例，“何 IC ですか？”），認識結果の単語 A が回答の関係にある（例では，IC 名）場合，単語 A の単語理解スコアを上げる．戦略 8 が適用される場合の単語 w_A の単語理解スコア更新式は以下のとおりである．

$$Score_t(w_A) = Score_{t-1}(w_A) + weight_8 * Conf_t(w_A) \quad (10)$$

$weight_8$： 重み $(0.0 < weight_8 < 1.0)$

- 戦略 9：認識結果の上位は正解である可能性が高いため，上位に含まれる単語の単語理解スコアを上げる．戦略 9 が適用される場合の単語 w_A の単語理解スコア更新式は以下のとおりである．

$$Score_t(w_A) = Score_{t-1}(w_A) + weight_9 * Conf_t(w_A) \quad (11)$$

$weight_9$：重み $(0.0 < weight_9 < 1.0)$

ただし $weight_9$ は，単語 A の N-best 中の順位により重みが異なる．

- 戦略 10：音声認識器の特性上，発話長が長い発話は認識されやすく，逆に短い発話は認識されにくい．つまり一つのカテゴリーのみを発話された場合よりも，二つまたは三つのカテゴリーを同時に発話された場合の方が正しく認識されやすい．このため，単語 A の含まれる認識結果が 2 カテゴリー以上からなる場合と，認識結果が 1 カテゴリーしかない場合では，単語 A の単語理解スコアの上げる幅を変える．戦略 10 が適用される場合の単語 w_A の単語理解スコア更新式は以下のとおりである．

$$Score_t(w_A) = Score_{t-1}(w_A) + weight_{10} * Conf_t(w_A) \quad (12)$$

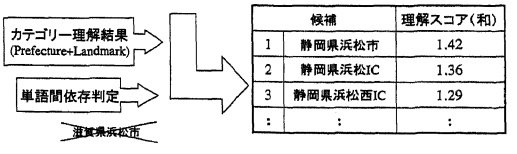
$weight_{10}$：重み $(0.0 < weight_{10} < 1.0)$

ただし $weight_{10}$ は，単語 A が含まれる認識結果のカテゴリーの数により重みが異なる．

4.4 単語理解

単語理解は，ある一つのゴール（本論文では一つの目的地設定）において，ユーザが発話した可能性が最も高い内容を理解するためのものである．図 4 に単語理解の例を示す．カテゴリー理解結果をもとに，対話履歴の中から単語理解スコアの和が最も高い単語の組合せを決定する．図 4 の例ではカテゴリー理解結果として，Prefecture カテゴリーと Landmark カテゴリーが発話されたと推定されているので，Prefecture カテゴリーと Landmark カテゴリーに属する単語の単語理解スコアの和を求めている．このとき，“滋賀県浜松市” のような実際にはあり得ない組合せを生成しないように，単語間の意味的制約（本論文では地理的制約）を考慮している．

Prefecture	理解スコア	Highway-Railway	理解スコア	Landmark	理解スコア
静岡	0.88	東名高速	0.25	浜松市	0.54
福岡	0.13	名神自動車道	0.01	浜松IC	0.48
滋賀	0.39	東海道本線	0.17	浜松西IC	0.41
:	:	:	:	:	:



	候補	理解スコア(和)
1	静岡県浜松市	1.42
2	静岡県浜松IC	1.36
3	静岡県浜松西IC	1.29
:	:	:

図 4 単語理解の例
Fig. 4 An example of word understanding.

5. 評価実験

5.1 対話データの作成

本システムでは，理解スコア更新式やカテゴリー理解などにおいて複数の重みやしきい値が存在する．これらパラメータの最適な値を求めるには，多種の発話パターンからなる対話データが必要となる．そこで対話データを作成するために以下のような方法を用いた．まず対話として，U1–S1–U2 の対話を想定する．この対話によるランドマーク設定の可能・不可能は考えないこととする．次にある一つのランドマークを設定する上で一般的に使われる複数の発話パターンを発話者に個別に発話してもらい音声を収録する．このようにして個別に収録した音声を以降，発話データと呼ぶ．収集した発話データ中からユーザの第 1 発話となり得るものを U1 として実際のシステムに入力し，システム応答 S1 を求める．この S1 に続く自然なユーザ発話をすべて発話データから選択し，U1–S1–U2 の模擬的な対話データを生成する．このようにして作成した対話データを以降，模擬対話と呼ぶ．

本実験では設定するランドマークを“浜松西インター”とし，このランドマークを設定する際に発話されるすべての発話パターンを用意した．用意した発話パターンは表 3 に示す 21 通りである．これらの発話パターンを実際に情報系学部・大学院生 5 名に，一つの発話パターンにつき 3 回，計 63 回の発話をしてもらった．このようにして，合計で 315 発話の発話データを収集した．更に実環境下での音声認識率の低下を考慮し，この発話データに車の走行雑音を重畳した．この雑音を重畳した発話データの N-best の第 1 候補におけるキーワード認識率を調べた結果，56.8%であった．本実験において N-best は 100-best とした．なお音響モデル・言語モデルに基づき求めた音響ゆう度が小さく信頼性が低い候補については，認識結果として出力されないため，認識結果の候補数が必ずしも 100 ではない．本実験で用いた発話データでは，認識結果として出力された平均候補数は約 30 であった．また認識結果の第 1 候補について「静岡」,「東名」,「浜松西」の各単語が別の特定の単語へ誤認識される傾向を調べたところ，極端に偏ることはなかった．例えば,「静岡」が最も誤りやすい単語は「三重」（三重県の「三重」）であるが,「静岡」が発話された 180 発話のうち「静岡」を誤認識した発話は 73 発話あり，その誤認識した発話の中で「三重」へ誤認識した割合は 20.5%（15/73 発話）であった．これらの発話データを用いてパラメータ推定用の模擬対話（以降，学習データ）を作成した．作成した模擬対話の数は 7,530 対話である．

表 3 発話パターン
Table 3 Speech pattern.

1	静岡県
2	東名自動車道
3	浜松西インター
4	静岡県の浜松西インター
5	静岡県の東名自動車道
6	東名自動車道の浜松西インター
7	静岡県の東名自動車道の浜松西インター
8	はい，静岡県
9	はい，東名自動車道
10	はい，浜松西インター
11	はい，静岡県の浜松西インター
12	はい，静岡県の東名自動車道
13	はい，東名自動車道の浜松西インター
14	はい，静岡県の東名自動車道の浜松西インター
15	いいえ，静岡県
16	いいえ，東名自動車道
17	いいえ，浜松西インター
18	いいえ，静岡県の浜松西インター
19	いいえ，静岡県の東名自動車道
20	いいえ，東名自動車道の浜松西インター
21	いいえ，静岡県の東名自動車道の浜松西インター

この学習データを使い，すべてのカテゴリー（Prefecture・Highway-Railway・Landmark カテゴリー）の単語が，発話者の発話した単語と過不足なく一致する対話数の割合（以降，完全一致率）が最も高くなるようにパラメータを推定した．すべてのカテゴリーの単語が過不足なく一致するとは，例えば模擬対話において 2 カテゴリーしか発話されてないときに，その 2 カテゴリーについては発話された単語とシステムの推定した単語が一致し，なおかつ残りの 1 カテゴリーは発話されてないとシステムが推定した場合である．学習データにおける最適なパラメータを用いた場合の完全一致率は 71.5%（5,386/7,530 対話）であった．また対話が始まってからユーザが発話したカテゴリーを過不足なく推定できた対話数の割合（以降，カテゴリー理解率）は，74.7%（5,632/7,530 対話）であった．このカテゴリー理解率と完全一致率はともに対話ごとに求められる割合であるが，カテゴリー理解率は，カテゴリーレベルでのユーザ発話と一致する割合であり，完全一致率は，単語レベルでのユーザ発話と一致する割合である．例えばユーザが対話において「静岡県浜松西 IC」と発話したとすると，カテゴリー理解はクラス理解スコアを用いて,「Prefecture カテゴリーと Landmark カテゴリーが発話された」と

推定できればよい．一方，完全一致率はカテゴリー理解結果と単語理解スコアを用いて，「Prefecture カテゴリーの単語として静岡県，Landmark カテゴリーの単語として浜松西 IC が発話された」と推定できなければならない．このカテゴリーレベルでの一致で正解とみなすか単語レベルでの一致で正解とみなすかが，カテゴリー理解率と完全一致率の違いである．更に，Prefecture・Highway-Railway・Landmark カテゴリーの各単語が発話者の発話した単語と一致する割合を単語ごとに計算したところ（以降，単語一致率）87.0%（19,655/22,590 単語）となった．単語一致率の分母となる単語数が完全一致率の分母となる対話数の 3 倍となっているのは，カテゴリーは三つあり，各カテゴリーの単語がユーザの発話した単語と一致するかを調べているためである．ユーザが対話中で 2 カテゴリーしか発話していない場合でも，システムは発話されていないカテゴリーを特定し，そのカテゴリーについては地名単語を出力しないようにしなければならない．そこで地名単語を出力していないことを調べるために，1 対話につき三つのカテゴリーの単語を調べている．

学習データと同様の方法で，評価用の模擬対話の発話データを収集した．評価用の発話データは，学習データとは異なる情報系学部・大学院生 10 名に発話してもらい，収集した発話データは 630 発話となった．走行雑音を重畳した発話データの N-best の第 1 候補におけるキーワード認識率は 67.7%である．これらの発話データを用いて評価用の模擬対話（以降，評価データ）を作成したところ，作成した模擬対話の数は 29,670 対話となった．なお，ユーザの発話した単語が音声認識結果の N-best 中に存在しない対話は，29,670 対話中 3,305 対話あった．

5.2 実験方法

最新の音声認識結果（N-best）の上位候補を最優先する言語理解手法と，本論文で述べた言語理解手法の性能を比較するための評価実験を行った．この実験のために 3 種類のシステムを用意した．一つ目は，最新の認識結果の N-best の上位候補を最優先する言語理解手法を採用したシステム（以降，SYS-A）であり，二つ目は，本論文で示した言語理解手法を採用したシステム（以降，SYS-B）である．三つ目は SYS-B とほぼ同じであるが，カテゴリー理解結果に正解を与えた場合のシステム（以降，SYS-C）である．SYS-B では，カテゴリー理解を誤ると必ず誤理解を起こしてしまうので，カテゴリー理解の精度が理解精度を大きく左右する．そこでカテゴリー理解率を 100%にした場合，どの程度の理解性能になるのかを調べた．これらのシステムの言語理解部以外の性能はすべて同等である．

5.3 結果と考察

評価データについて発話タイプごとの完全一致率と単語一致率を求めた．評価データの全模擬対話における完全一致率は，SYS-A が 57.9%，SYS-B が 72.2%，SYS-C が 89.2%であった．SYS-A に比べ，SYS-B の完全一致率は約 15 ポイント上回っており，提案手法が有効であることが分かる．更に SYS-C の完全一致率は SYS-B に比べて 17 ポイント高い．この結果はカテゴリー理解の精度を上げることにより，更なる理解率の向上が可能であることを示している．なお，SYS-B のカテゴリー理解率は，78.8%（23,408/29,670 対話）であった．

評価用の模擬対話を発話タイプ別に分け，更に各発話タイプ中において，「U1 の認識結果の第 1 候補が正解であった場合/なかった場合」（グラフ中では “U1:OK”/“U1:NG” として表記），「U2 の認識結果の第 1 候補が正解であった場合/なかった場合」（グラフ中では “U2:OK”/“U2:NG” として表記）のそれぞれの組合せについて「システムが対話ごとに出力した理解結果とユーザ発話が一致した場合/しなかった場合」（グラフ中では “SB:OK”/“SB:NG” と表記）を求めた．これを図 5 に示す．グラフ上部の数値は発話タイプごとの対話数を表し，グラフ中の数値は割合を表している．このグラフより，訂正発話と再入力発話の一部を除き，SYS-C，B，A の順で理解率が高い．このことから訂正発話以外においては，誤認識がどのタイミングで発生しても提案手法が有効であることが確認された．

評価データの単語一致率は，SYS-A が 75.4%，SYS-B が 87.1%，SYS-C が 95.5%であった．単語レベルにおいても提案手法が有効であることが分かる．発話タイプ別の単語一致率（図 6）においても同様に，訂正発話と再入力発話の一部を除き，SYS-C，B，A の順で単語一致率が高い．なお，図 6 における「“SB:OK”/“SB:NG”」は「システムがカテゴリーごとに出力した理解結果とユーザ発話が一致した場合/しなかった場合」であり，グラフ上部の数値は発話タイプごとの単語数を表す．完全一致率よりも単語一致率の方が高いことから，システムの誤理解には，ユーザ

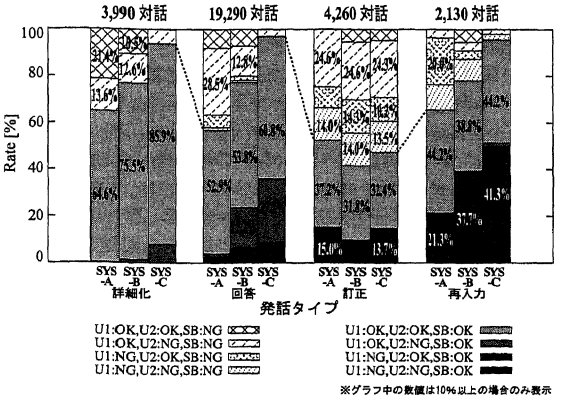


図 5　発話タイプごとの完全一致率
Fig. 5　A full match rate of each speech type.

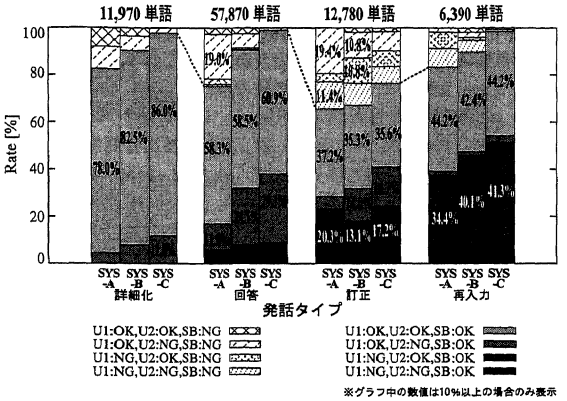


図 6　発話タイプごとの単語一致率
Fig. 6　A word match rate of each speech type.

発話の一部だけを誤る場合も多いことが分かる．この特徴と提案手法をうまく利用すれば，システムが内部的には誤理解していたとしても，部分的に正しい（信頼性が高い）理解内容をもとに，ユーザに誤理解を悟られないように対話を進めることで，新たに入力された関連情報から迅速にシステムの誤理解部分を修正することが可能である．そのため，誤理解した部分の情報を明示的にユーザに再入力させることなく，ユーザの発話内容を完全に理解することが可能となる．

今回の評価実験において，提案手法の訂正発話の精度が低い原因として，学習データにおいて訂正発話の対話数が少ないことが考えられる．学習データの訂正発話数が少ないため，それを用いて最適化したパラメータが，訂正発話に対する理解精度を下げる傾向になったと考えている．訂正発話が少ないのは，本システムでは誤った情報がシステム応答に含まれるとユーザの満足度が大幅に低下すると考え，発話された可能性が低い単語については，応答内容として出力しない「分かったふり戦略」のためである．また今回，発話データは防音室で音声を収録し，後処理で走行雑音を重畳しているので，実環境下での発話タイプの割合と必ずしも一致するとは限らない．このため，今後実環境下で収録した発話データによるパラメータの推定・性能評価を行う必要がある．

また，単語理解スコア更新における戦略 1〜10 の効果を調べるため，最適化したパラメータのうち，各戦略の重み $weight_n$（$n = 1, \cdots, 10$）の大きさと適用回数を比較した．戦略 1，2，3 は，重みの値は小さいが適用回数は他の戦略に比べ非常に多く，理解スコアの更新に強い影響をもつ戦略であるといえる．また戦略 4，7，8 は，適用回数は少ないが重みの値が大きく，適用される際には理解スコアを大きく変動させるので，これらの戦略も強い影響があると考えられる．しかし上述以外の戦略も含めた各戦略同士の相互作用によって正しく理解できた対話も多く，学習データの誤り傾向やタスクによって各戦略の重みも変化するため，今回影響が弱いとみなした戦略が，必ずしも不要な戦略であるとは本実験だけではいえず，戦略ごとの定性的な有効性については今後の課題である．

なお，提案手法の適応範囲に関してだが，提案手法をすべての音声対話タスクに対して適応することは難しい．しかしながら，近年の音声対話システムの主流タスクであるスロットフィリングや検索タスクなどの情報入力タスクでは，入力される情報（キーワード）は木構造のような階層的で依存関係がある形で表現できる場合が多い．そのような場合においては，本論文の提案手法は有効であり適応範囲は広いと考える．ただし，単語理解スコア更新戦略に関しては，タスクやドメインによって若干の修正・追加は必要になる場合が考えられる．

6. むすび

本論文では音声認識結果に誤認識を含む場合でもユーザの発話を正しく推定するために，音声認識結果の認識信頼度と対話履歴を用いて言語理解を行う方法について述べた．模擬対話を用いた評価実験では，最新の認識結果の上位候補を優先する従来手法に比べ，理解率（完全一致率）において約 15 ポイントの性能向上が見られ，提案手法が有効であることが示された．また誤理解が起きている場合でも，ユーザ発話の一部

は正しく推定できていることが分かった．更にカテゴリー理解精度を上げることで，理解率を向上させることも可能である．

今後の課題としては，まず発話データに関する課題がある．本論文で扱った発話データは，防音室で収録し後処理で走行雑音を重畳した音声である．これが必ずしも実環境下での発話に近いとは限らないので，より実環境下に近い状態で評価を行うためには，実際に自動車を運転している状態での車内音声を収録し，発話データとする必要がある．次に対話の長さに関してだが，今回の評価はU1-S1-U2の2ターン限定であったため，ランドマークの設定が終了したものもあったが，設定途中のものもあった．そこでより長いターンの対話での理解率の評価も必要である．更に実際の運転状況に近い状態（ドライブシミュレータなど）でシステムを使用した場合の理解率やユーザの満足度など[24]を調べる必要もある．言語理解部の戦略に関する課題としては，各戦略の有効性の調査がある．本論文で述べた単語理解スコア更新のための各戦略は，学習データ，タスクなどにより理解スコアの更新における影響は大きく変化する．このため，各戦略が定性的にどの程度有効なのかを調べる必要がある．最後にシステムの拡張に関してだが，今後はより複雑な状況に対応できるようシステムを拡張することを考えている．今回のシステムが扱ったのは，ユーザにとってランドマークが既知の地名であった．しかしランドマークの名前を知らない場合や，検索を行った後にランドマークに設定する場合なども考えられるので，これらの状況にも対応できるような拡張を考えている．

文　献

[1] 原　直，白勢彩子，宮島千代美，伊藤克亘，武田一哉，“音声対話による楽曲検索システム，” 情処学研報，SLP-53, pp.31-36, Oct. 2004.

[2] 渡辺裕太，関口芳廣，鈴木良弥，“ビデオ装置を例とした家電品の音声対話機能について，” 情処学論，vol.44, no.11, pp.2690-2698, Nov. 2003.

[3] 河口信夫，牛窪誠一，松原茂樹，岩　博之，梶田将司，武田一哉，板倉文忠，“走行車室内音声対話収録システムの開発，” 信学論（D-II），vol.J84-D-II, no.6, pp.909-917, June 2001.

[4] T. Itoh, A. Kai, T. Konishi, and Y. Itoh, “An understanding strategy based on plausibility score in recognition history using CSR confidence measure,” Proc. ICSLP '04, pp.2133-2136, Jeju Island, Korea, Oct. 2004.

[5] 小窪浩明，天野明雄，畑岡信夫，“車載用音声認識における騒音対策とその評価，” 信学論（D-II），vol.J83-D-II, no.11, pp.2190-2197, Nov. 2000.

[6] 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，” 情処学論，vol.43, no.10, pp.3078-3086, Oct. 2002.

[7] C. Raymond, Y. Esteve, F. Bechet, R. De Mori, and G. Damnati, “Belief confirmation in spoken dialog systems using confidence measures,” Proc. ASRU 2003, pp.150-155, St. Thomas, U.S. Virgin Islands, Nov. 2003.

[8] 堤　修一，磯部俊洋，森島昌俊，“複数の正規化尤度を複合的に用いた音声認識結果の信頼度算出法，” 情処学研報，SLP-57, pp.31-36, July 2005.

[9] D.J. Litman, J.B. Hirschberg, and M. Swerts, “Predicting automatic speech recognition performance using prosodic cues,” Proc. 6th Applied Natural Language, Processing Conference (NALP-NAACL00), pp.218-225, Seattle, USA, April 2000.

[10] 甲斐充彦，石丸明子，伊藤敏彦，小西達裕，伊東幸宏，“目的地設定タスクにおける訂正発話の特徴分析と検出への応用，” 日本音響学会全国大会論文集，2-1-8, pp.63-64, Oct. 2001.

[11] 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返し訂正発話の検出と認識，” 信学論（D-II），vol.J87-D-II, no.7, pp.1441-1450, July 2004.

[12] 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，” 情処学研報，SLP-34, pp.239-244, Dec. 2000.

[13] 神田直之，駒谷和範，尾形哲也，奥乃　博，“データベース検索タスクの文脈的制約を用いた音声対話システムの実験的評価，” 情処学研報，SLP-55, pp.107-112, Feb. 2005.

[14] 山本博史，谷垣宏一，匂坂芳典，“対話者の前発話を利用した統計的言語モデル，” 信学論（D-II），vol.J84-D-II, no.12, pp.2507-2514, Dec. 2001.

[15] C. Bousquet-Vernhettes and N. Vigouroux, “Context use to improve the speech understanding processing,” Proc. SPECOM 2001, pp.89-92, Moscow, Russia, Sept. 2001.

[16] ウッティウィワッチャイチャイ，古井貞熙，“談話理解における対話コンテキストに基づく非線形リスコアリング，” 情処学研報，SLP-51, pp.37-42, May 2004.

[17] R. Higashinaka, K. Sudoh, and M. Nakano, “Incorporating discourse features into confidence scoring of intention recognition results in spoken dialogue systems,” ICASSP2005, vol.1, pp.25-28, Philadelphia, USA, March 2005.

[18] 大森久美子，東田正信，“効率的な音声対話制御方式に関する一考察，” 情処学研報，SLP-32, pp.45-50, July 2000.

[19] 趙　國，宮山章子，山下洋一，“N-best音声認識における認識スコアを利用した候補提示数の決定，” 信学論（D-II），vol.J88-D-II, no.6, pp.1003-1011, June 2005.

[20] 平沢純一，宮崎　昇，中野幹生，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，” 日本音響学会全国大会論文集，3-8-10, pp.85-86, March 2000.

[21] H.P. Grice, “Logic and conversation,” in Speech Acts, Syntax and Semantics, ed. P. Cole and J. Morgan,

vol.3, pp.41–58, Academic Press, New York, 1975.
[22] 水谷　誠，伊藤敏彦，甲斐充彦，小西達裕，伊東幸宏，“音声認識の信頼度と対話履歴を利用した最尤推定型言語理解，” 情処学研報，SLP-45-19, pp.113–118, Feb. 2003.
[23] 中川聖一，甲斐充彦，“文脈自由文法制御による One Pass 型 HMM 連続音声認識法，” 信学論（D-II），vol.J76-D-II, no.7, pp.1337–1345, July 1993.
[24] 石川　泰，澤田久美子，城戸恵美子，“音声インタフェースの評価，” 音響誌，vol.61, no.2, pp.79–84, Feb. 2005.

（平成 17 年 9 月 5 日受付，18 年 1 月 10 日再受付）

藤原　敬記　（学生員）

2004 北海道大学大学院工学研究科修士課程了．現在，同大学院情報科学研究科博士後期課程在学中．音声言語処理，音声対話システムに興味をもつ．日本音響学会，情報処理学会各会員．

伊藤　敏彦　（正員）

1999 豊橋技術科学大学大学院工学研究科博士後期課程電子・情報工学専攻了．同年静岡大学情報学部情報科学科助手．2004 北海道大学情報科学研究科メディアネットワーク専攻助教授．博士（工学）．音声言語情報処理研究に従事．情報処理学会，人工知能学会，日本音響学会，ヒューマンインターフェース学会各会員．

荒木　健治　（正員）

1982 北大・工・電子卒．1988 同大大学院博士課程了．工博．同年，北海学園大学工学部電子情報工学科助手．1989 同講師．1991 同助教授．1998 同教授．1998 北海道大学大学院工学研究科電子情報工学専攻助教授．2002 同教授．現在，北海道大学大学院情報科学研究科メディアネットワーク専攻教授．自然言語処理，特に，機械翻訳，音声対話処理などの自然言語処理の研究に従事．情報処理学会，言語処理学会，人工知能学会，認知科学会，ACL，IEEE，AAAI 各会員．

甲斐　充彦　（正員）

1991 豊橋技科大・情報工学卒．1996 同大大学院博士後期課程了．同年豊橋技術科学大学工学部助手．1999 静岡大学工学部システム工学科講師．2000 同助教授．音声認識を中心とした音声言語処理と対話処理に興味をもつ．博士（工学）．日本音響学会，情報処理学会，人工知能学会各会員．

小西　達裕　（正員）

1987 早大・理工・電子通信卒．1992 同大大学院博士後期課程了．1991 早稲田大学理工学部情報学科助手．1992 静岡大学工学部情報知識工学科助手．現在，同大学情報学部情報科学科助教授．博士（工学）．知的教育システム，知的対話システムなどに興味をもつ．情報処理学会，人工知能学会，教育システム情報学会，日本認知科学会各会員．

伊東　幸宏　（正員）

1987 早稲田大学大学院博士後期課程了．同年，同大学理工学部電子通信学科助手．現在，同大学情報学部情報科学科教授．工学博士．自然言語処理，対話システム，知的教育システム等に興味をもつ．情報処理学会，人工知能学会，言語処理学会，教育システム情報学会，日本認知科学会各会員．