Vol.2009-NL-191 No.15
Vol.2009-SLP-76 No.15
2009/5/22

# 複数の音声対話システム併用のための発話識別

粟野 健太郎[†1] 伊藤 仁[†1]
伊藤 彰則[†1] 牧野 正三[†1]

本稿では複数の音声対話システムを併用することを目的とし，そのために必要な発話識別の方法を検討した．併用するシステムとして，確認応答型システムと一問一答型システムを用いた．識別の特徴量として発話の各タスクらしさを表すスコアと音声認識結果の尤度を用いた．発話識別は特徴量の大小比較とニューラルネットで行った．音声認識結果が 1-best 時と N-best 時の両方で識別実験を行ったところ，80%以上の正解率を得るとともに N-best 時の方が正解率が向上することが分かった．

# Utterance Discrimination for using Multiple Spoken Dialog Systems

Kentaro Awano,[†1] Masashi Ito,[†1] Akinori Ito[†1]
and Shozo Makino[†1]

We studied a method of utterance discrimination for a spoken dialog system that combines multiple dialog systems. Frame-based and example-based systems are used as systems for combination. We used similarities to tasks and likelihood obtained by a speech recognizer as features for the discrimination. A discrimination function is composed by a neural network. We conducted a discrimination experiment using 1-best and n-best recognition results of the speech recognizer. As a result, we obtained more than 80% accuracy, and the result by the n-best candidates was better than that by the 1-best candidate.

†1 東北大学大学院工学研究科
Graduate School of Engineering, Tohoku University

## 1. はじめに

近年の音声認識技術の発達により，音声でコンピュータを操作する音声対話システムが社会で用いられ始めている．一般に，音声対話システムは扱うタスクに特化して設計される．したがって開発者が既存のシステムに新タスクを追加したり，現タスクを削除しようとすると，システム全体の見直しが必要となる．ゆえに開発者は自由にタスクをカスタマイズしにくいという問題がある．

音声対話システムには，複数タスク間での複雑な制御を必要とするものもあるが，比較的単純なタスクを並列に扱えば十分である場合も多いと思われる．そこで本研究では，比較的単純なタスクを複数扱う対話システムを対象とし，このようなシステムにおいてカスタマイズが容易なシステム構成法を検討する．具体的なシステムの構成方法として，比較的単純な単一のタスクのみを対象とした対話システム（単機能システム）を複数並列に組み合わせることにより，それらの単機能システムすべてのタスクを扱うことができる対話システムを構築する．このようなシステムの例を**図 1** に示す．このようなシステムでは，各単機能システムは互いに独立である．したがって，タスクの追加はそのタスクに対応する単機能システムを追加するだけで行えるため，タスク追加が容易にでき，同様にタスク削除も容易にできる．

このシステムでは発話が入力されたとき，その発話は最も適切な単機能システムで処理される．そのためには，入力された発話をどの単機能システムで処理すべきかを判断する必要がある．そこで本稿では，発話を最適な単機能システムで処理するために必要な発話識別方法を提案し，識別精度を検討する．

## 2. 本検討で用いる単機能対話システム

提案システムを構成するためには，単機能システムとしてどのようなものを利用するかが問題となる．ここでは，比較的小規模なフレームベースの対話システム[1)] と，入力された質問発話に対して直ちに返答するタイプの対話システム[2)] を対象とする．ここでは，前者を「確認応答型システム」，後者を「一問一答型システム」と呼ぶことにする．

### 2.1 確認応答型システム

ここでいう確認応答型システムとは，比較的小規模のフレームに基づき，それを埋めるように対話を制御する音声対話システムである．ここでは，Konashi らによる対話システム[1)] をターゲットとする．このシステムでは，ユーザは**表 1** のようなタスク記述表を作成することで確認応答型システムを設計する．タスク記述表から，**図 2** に示す有限オートマトン

Vol.2009-NL-191 No.15
Vol.2009-SLP-76 No.15
2009/5/22

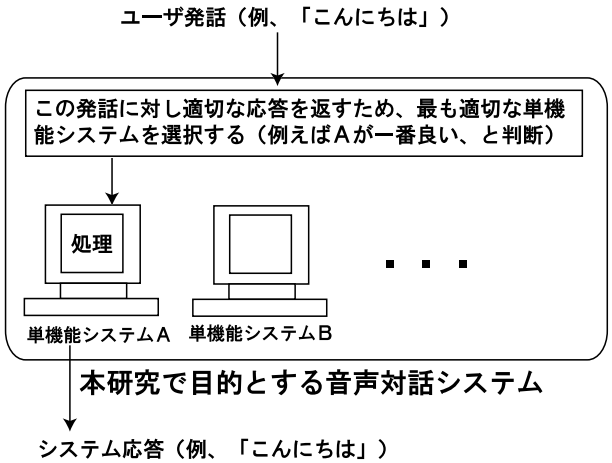


**図 1** 目的とする音声対話システムの概要
Fig. 1 intended Spoken Dialog System

**表 1** タスク記述表 (タスク：仙台駅での切符販売)
Table 1 Task Description Table

| スロット名 | 品詞 | 種類 | 単語 | 初期値 | 助詞 |
|---|---|---|---|---|---|
| 駅名 | 名詞 | 必須 | 東京，福島，… | $\phi$ | まで |
| 切符 | 名詞 | 任意 | 普通，新幹線，… | 普通 | $\phi$ |
| 枚数 | 名詞 | 任意 | 一枚，二枚，… | 一枚 | $\phi$ |
| 動作 | 動詞 | 任意 | 下さい, お願いします, … | 下さい | $\phi$ |

と等価な認識文法が生成され，これを用いて音声認識を行う．生成されたオートマトンは，文節単位での倒置や省略，フィラー挿入などを受理できるように設計される．

このシステムの主な特徴としては，タスク内の発話は高精度で認識できるが，タスク外の発話や未知語に弱いということが挙げられる．確認応答型システムは限定された使用環境下でより確実な対話が望ましいタスク，例えば何かを依頼する質問に適した対話システムと言える．

### 2.2 一問一答型システム

一問一答型システムとはユーザの発話に対するシステムの応答があらかじめ決まっているシステムである．対話の際には，発話と用例テキストとのマッチングを行い，最大スコアに

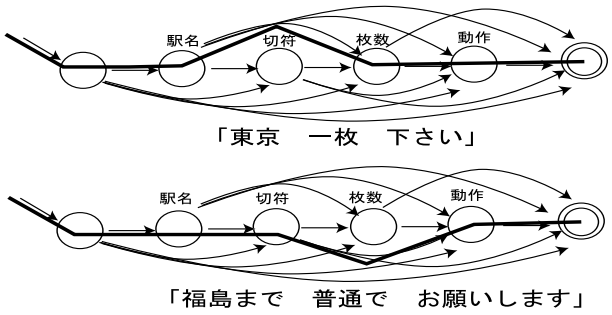


**図 2** 記述文法の例 (タスク：仙台駅での切符販売)
Fig. 2 Example of Grammar

**表 2** 用例テキストの例
Table 2 Sample of Example Text

| |
|---|
| #101 おはよう＋オハヨー＋感動詞 。＋。＋記号 |
| #102 こんにちは＋コンニチワ＋感動詞 。＋。＋記号 |

**表 3** 応答候補文の例
Table 3 Sample of Answer Text

| |
|---|
| #101 おはよう。 |
| #102 こんにちは。 |

なった用例テキストに対応した応答候補文を応答として返す[2)]．言語モデルには N-gram を用いる．ユーザは**表 2** のような発話に当たる用例テキストと，**表 3** のようなシステムの応答に当たる応答候補文を作成することで一問一答型システムを設計する．

このシステムの長所は，用例テキストと応答候補文の数を増やすことにより対話の幅を広げられることである．しかし，認識のための言語モデルとして N-gram を用いているため，認識率や応答の精度が記述文法よりも低いことが欠点である．一問一答型システムはユーザに情報を提供するようなタスクに適した対話システムと言える．

## 3. 発話識別の方法

### 3.1 識別アルゴリズム

発話識別のアルゴリズムは以下の通りである

(1) 発話を音声認識する．ただし，N-gram と記述文法を並列に用いて認識する．また，認識結果は 1-best とする．
(2) 音声認識結果から発話の特徴量（音声認識スコア，用例スコア，記述表スコア）を抽出する．
(3) 予め学習により求めておく識別関数により，発話が確認応答かどうかを識別する．

Vol.2009-NL-191 No.15
Vol.2009-SLP-76 No.15
2009/5/22

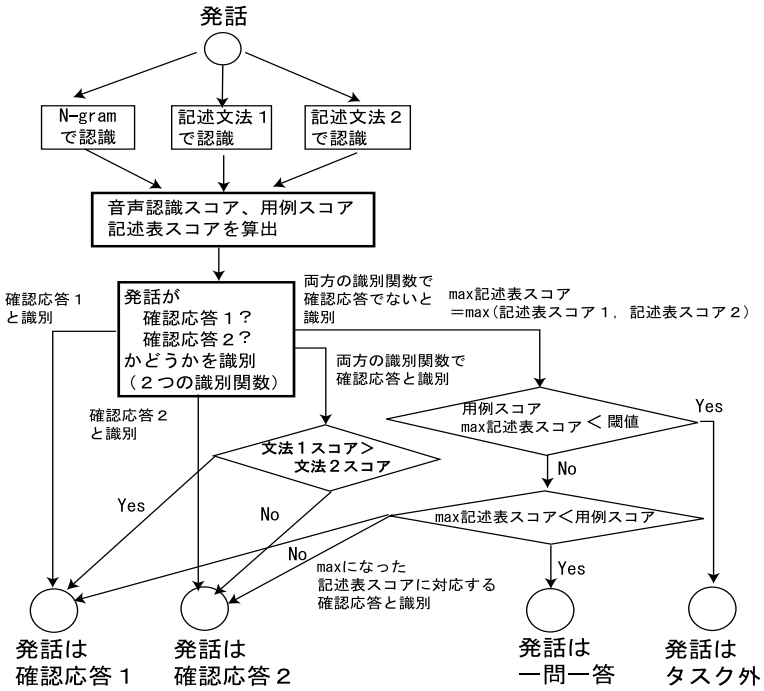


**図 3** 本検討における発話識別のフローチャート
Fig. 3 Method of Utterance Discrimination in this study

（4） 識別関数で確認応答ではないと識別された発話を，用例スコア・記述表スコア・閾値の大小比較により，確認応答・一問一答・タスク外のいずれかに識別する．

なお，本検討ではそれぞれタスクが異なる確認応答型システムを 2 つ，一問一答型システムを 1 つ併用した．この場合，一問一答型システムは N-gram を用いた認識を行い，2 つの確認応答型システムはそれぞれ独立の記述文法を用いて認識を行う．このときの識別のフローチャートを**図 3** に示す．ここで，記述文法 1・記述文法 2 は，それぞれ 2 つの確認応答型システムの言語モデルである．

### 3.2 特 徴 量

発話識別に用いる特徴量は発話の音声認識結果より求める．

（1） **記述表スコア**

記述表スコアは N-gram による認識結果文の確認応答型タスクらしさを表す特徴量であり，N-gram 認識結果文から求める．ここで，確認応答型システムでの音声理解には，記述文法の認識結果が用いられる．しかし，記述文法を用いると，たとえ現在の発話が確認応答型システムへの入力でなかったとしても，認識結果は必ず確認応答型システムで理解できる単語列となる．そこで，ここでは記述文法ではなく N-gram の認識結果を利用し，N-gram の認識結果がどれだけ確認応答型で理解される内容なのかを「確認応答型らしさ」とする．このとき，タスク記述表のスロットに入りうる単語が文中に多く含まれる程，その文は確認応答型タスクらしいと考えられる．また，必須スロットの単語は話者の意図を理解するのに重要な単語であるので，マッチングにおいて重みを付与する．以上より，文 $x$ の記述表スコアは式 (1) で表される．

$$Tscore = \sum_{i=1}^{N} \frac{\alpha_i \times S_i(x)}{x \text{ 中の単語数}} \tag{1}$$

$$S_i(x) = \begin{cases} 1 & \text{if } x \text{ 中にスロット } i \text{ の単語が 1 個以上含まれている} \\ 0 & \text{else} \end{cases} \tag{2}$$

ただし，$N$ は全スロット数，$\alpha_i$ はスロットの種類の重み (必須ならば 1 以上，任意ならば 1) である．

（2） **用例スコア**

用例スコアは一問一答型タスクらしさを表す特徴量であり，N-gram 認識結果文から求める．文中にマッチングする用例テキストの自立語形態素が多く含まれる程，その文は用例テキストに似ている．そこで，この考え方を基に早川らの応答候補文選択スコアリング方法[3] で用例スコアを算出する．用例テキスト集合を $A$ とすると，文 $x$ の用例スコアは式 (3) で表される．

$$Escore = \max_{a \in A} \frac{M_{ax}}{\max(M_a, M_x)} \tag{3}$$

ただし，$M_a$，$M_x$ はそれぞれ用例テキスト $a$，認識結果文 $x$ 中の自立語形態素数で，$M_{xa}$ は $a$ と $x$ でマッチした自立語形態素数である．

（3） **音声認識スコア**

音声認識結果の尤度を表す音声認識スコアを用いる．

言語モデルに N-gram を用いたときのスコアを $Nscore$ とすると，式 (4) で表される[4]．

$$Nscore = \log p(X|W) + \alpha \times \log p(W) + \beta w \tag{4}$$

Vol.2009-NL-191 No.15
Vol.2009-SLP-76 No.15
2009/5/22

ただし，$\alpha$ は言語重み，$\beta$ は挿入ペナルティ，$w$ は単語数である．実験では，$\alpha = 8.0$, $\beta = -2.0$ とした．

言語モデルに記述文法を用いたときのスコアを $Gscore$ とすると，式 (5) で表される[4]．

$$Gscore = \log p(X|W) \tag{5}$$

### 3.3 識別関数

153 発話（確認応答 1 が 48 発話，一問一答・タスク外が 105 発話）について，各発話を認識スコア差分（$Nscore - Gscore$），および用例スコアと記述表スコアの差分の 2 次元でプロットしたものを**図 4** に示す．図 4 より，各グループの発話はまとまって分布していることが分かる．したがって，この 2 次元を特徴ベクトルとして，発話として確認応答型タスクかどうかを識別する識別関数を設計することができる．なお，識別関数はニューラルネットを用いる．

識別関数を設計するに当たっては，1 つの確認応答型システムとそれ以外を識別するための関数を設計することとする．したがってシステム内に複数の確認応答型システムがある場合，複数の確認応答型タスクと判断されることがある．この場合，選ばれた複数の確認応答型システムに対応する記述文法での音声認識の尤度を比較し，尤も尤度の高いシステムの発話であると判定する．**図 5** は確認応答型タスク 98 発話（確認応答 1 が 48 発話，確認応答 2 が 50 発話）をプロットした図である．横軸が確認応答 1 で用いる文法の認識スコア，縦軸が確認応答 2 で用いる文法の認識スコアである．図 5 より，各確認応答発話はその確認応答用文法スコアのほうが他の文法スコアよりも大きいことが分かる．これにより，上記の識別方法が有効であることがわかる．

## 4. 発話識別実験

確認応答型システム 2 つと一問一答型システム 1 つを併用したときの発話識別精度を調べた．発話識別実験条件を**表 4** に示す．確認応答型 1 として仙台駅での切符販売，2 として仙台のおみやげ販売，一問一答型として仙台の観光案内をタスクとして設定した．各タスクの発話例を**表 5** に示す．また，識別関数は中間層が 1 層で，ノード数が入力 2，中間 3，出力 2 のニューラルネットで学習した．実験として図 3 にある閾値を変化させたときの識別正解率の変化を調べた．

まず，学習により得られた識別関数で評価データを識別したときの正解率を**表 6** に示す．切符販売文法用では切符販売発話と（仙台観光案内発話+タスク外発話）の識別を，おみやげ販売文法用ではおみやげ販売発話と（仙台観光案内発話+タスク外発話）の識別をしている．いずれの正解率も 94%以上と，高精度な識別関数が得られた．

次に，発話識別実験の結果を**図 6** に示す．図 6 は全体の正解率と一問一答側，確認応答側，タスク外の各カテゴリにおける正解率の閾値の値による変化を示している．まず，全体正解率の最大値は 82.5%であり，最大時はタスク外正解率が低いことが見られた．また，閾値を大きくするにつれてタスク外正解率が上がり，一問一答側正解率が下がる傾向が見られた．その理由は，図 3 より閾値以上の用例スコアを持つ発話が減るためである．なお，確認応答側正解率が高かった理由は識別関数による識別が高精度であるためである．

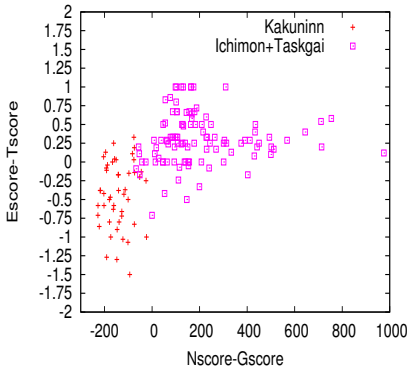


**図 4** 認識スコア差分および用例スコアと記述表スコアの差分の関係
Fig. 4 Relation between "$Nscore - Gscore$ "and "$Escore - Tscore$ "

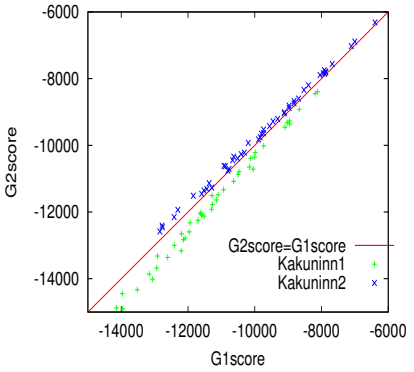


**図 5** 各確認応答で用いる文法の認識スコアの関係
Fig. 5 Relation between $Gscores$

**表 4** 実験条件
Table 4 Experiment condition

| | |
|---|---|
| 学習データ | 男性話者 1 名の読み上げ音声 203 発話 |
| 評価データ | 男性話者 5 名の読み上げ音声 120 発話 |
| 用例テキスト | 485 文 |
| タスク記述表 | 切符：97 単語，おみやげ：30 単語 |
| スロット重み $\alpha_i$ | 必須：4，任意：1 |
| 音響モデル | CSRC 標準成人モデル |
| N-gram | 毎日新聞 11 年分で作成 |
| 記述文法 | 切符販売用とおみやげ販売用 |
| 認識エンジン | Julius-4.0.1 |

Vol.2009-NL-191 No.15
Vol.2009-SLP-76 No.15
2009/5/22

**表 5** 各タスクの発話例
Table 5 Sample utterances of each task

| タスク | 発話例 |
|---|---|
| 仙台駅での切符販売 | 東京まで一枚下さい |
| 仙台のおみやげ販売 | 萩の月一箱下さい |
| 仙台観光案内 | 東北大学はどこにありますか |

**表 6** 識別関数の識別精度
Table 6 Accuracy by Discriminant function

| 識別関数 | 正解率 (%) |
|---|---|
| 切符販売文法用 | 94.4 |
| おみやげ販売文法用 | 100 |

今回識別誤りをした発話の特徴は 3 つあった．まずは音声認識誤りをした発話であり，特に仙台の地名などの固有名詞を認識できなかった発話が多かった．次に用例テキストと文体が似たタスク外発話（例：「東京**はどこにありますか**」）が一問一答側と誤識別されることがあった．そしてタスク記述表の単語（例：「東京」「萩の月」）を含む発話が確認応答側と誤識別されることがあった．

## 5. N-best の利用

ここまでは音声認識結果が 1-best であるとして検討した．しかし，1-best よりも N-best を用いた方が音声認識誤りを吸収でき，より正確に発話の特徴量を求められると考えられる．そこで，本節では N-best を用いた時の発話識別について検討する．ただし特徴量の多くを N-gram 認識結果から算出することから，今回は N-gram による認識結果のみを N-best とした．

### 5.1 特徴量の算出

N-best を用いた場合，候補が複数あるためそれに伴い各特徴量も複数個得られる．したがって，そこからどのようにして発話の特徴量として各特徴量を一つずつ得るかが問題となる．そこで，本稿では次の方法を検討した．まず，$Nscore$ および $Gscore$ については，従来と同じく最尤候補のみから計算する．一方，用例スコアと記述表スコアについては，次の (a)（b）2 通りの方法を検討する．

(a) 各候補のスコアのうち最も大きいものをその発話のスコアとして用いる
(b) 各候補のスコアの平均をその発話のスコアとして用いる

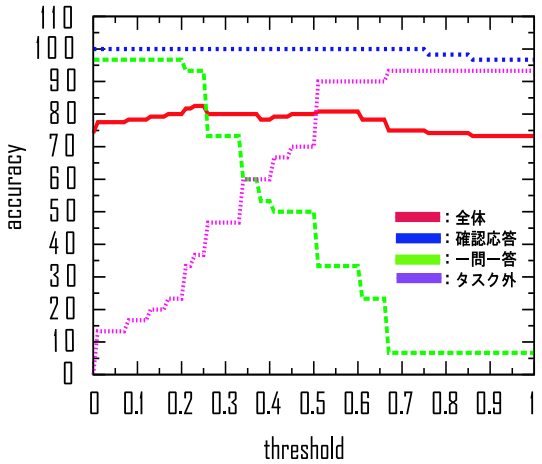


**図 6** 閾値による識別正解率の変化
Fig. 6 Accuracy of utterances discrimination changed by threshold

**表 7** 実験条件で表 4 とは異なる部分
Table 7 Part of experiment condition being different from Table.4

| | |
|---|---|
| 発話の用例スコアと記述表スコアの算出法 | 5.1 節の（a），(b）2 種類 |
| 識別関数の学習データ | 表 4 の学習データの 5-best 認識結果 |
| 評価データ | 表 4 の評価データの 1〜100-best 認識結果 |

### 5.2 実　　験

5.1 節に示した方法で発話の特徴量を抽出し，N-best の N によって正解率がどう変わるかを調べた．実験条件は基本的に表 4 と同じだが，表 4 と異なる部分を**表 7** に示す．なお，特徴量の算出方法は，学習時と評価時で同じになるように合わせてある．

まず，発話の特徴量算出方法を（a）としたときの結果を**図 7** に示す．図 7 の凡例はそれぞれ，A-acc が全体正解率，E-acc が一問一答正解率，T-acc が確認応答正解率，NTE-acc がタスク外正解率，threshold が各正解率が図中の値になったときの図 3 にある閾値である．結果から，10-best 前後で全体正解率が最大になることが見られた．この理由としては用例スコアと記述表スコアが最大になる候補が 10-best くらいまでに現れたからだと考えられる．つまり，その後に出てくる候補は識別に影響を及ぼさなかったと言える．また，最大正解率は 85.0%であり，1-best の時よりも 2.5%高くなった．

Vol.2009-NL-191 No.15
Vol.2009-SLP-76 No.15
2009/5/22

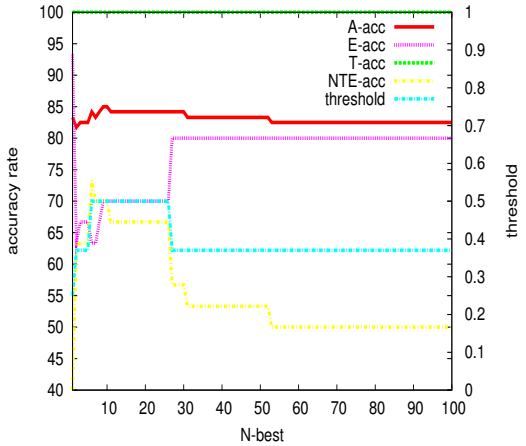


**図 7** 候補数による識別正解率の変化（方法（a）でスコア算出時）
Fig. 7 Accuracy of utterances discrimination changed by Candidate-num (N) on using (a)

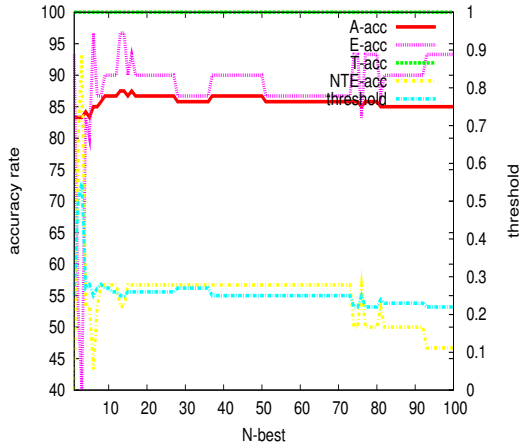


**図 8** 候補数による識別正解率の変化（方法（b）でスコア算出時）
Fig. 8 Accuracy of utterances discrimination changed by Candidate-num (N) on using (a)

次に，発話の特徴量算出方法を（b）としたときの結果を**図 8** に示す．図 8 の凡例も図 7 と同様のものを示す．結果から，15-best 前後で正解率が最大になることが見られた．また，最大正解率は 87.5%であり，1-best 時よりも 5%，方法（a）よりも 2.5%高くなった．つまり，発話の用例スコアと記述表スコアは各候補のスコアの平均とする，という方法の方が有効であることが分かった．

## 6. ま と め

本稿では単機能音声対話システムを複数併用することで，複数のタスクに対応しつつカスタマイズも容易にできる音声対話システムの構築を目的とし，それに必要な発話識別方法を検討した．まず，音声認識結果を 1-best としたときの正解率は全体で 82.5%となった．また，N-best にすることで正解率が最大 5%向上することが分かった．

今後はタスクをさらに追加して，検討した識別方法の性能を調べたい．また，現在は単純な特徴量の大小比較でタスク外かどうかを識別しているので，パターン認識などの手法を取り入れることを検討して高精度化を目指したい．

## 参 考 文 献

1) T.Konashi *et al.*, “A spoken dialog system based on automatic grammar generation and templete-based weighting for autonomous mobile robots, ” Proc.ICSLP,vol.I,pp.189-192,2004.
2) 西村竜一 他, “実環境研究プラットホームとしての音声情報案内システムの運用 ”, 信学論, Vol.J87-D-II, No.3, pp.789-798, 2004
3) 早川直樹, 他, “音声情報案内システムの応答文選択におけるスコアリング手法の改善 ”, 日本音響学会秋期講演論文集, 3-2-8, pp.87-88, 2006
4) 目黒豊美, 他, “音声対話システムにおけるタスク外発話判定法の検討 ”, 日本音響学会春季講演論文集, 1-P-27, pp.177-178, 2007