Vol.2014-NL-216 No.5
Vol.2014-SLP-101 No.5
2014/5/22

統計的音声対話システムにおける音素系列を用いた頑健な応答選択

佐伯 昌幸[1,a)]　李 晃伸[1,b)]

概要：一問一答形式の音声対話システムにおいては，実際の質問と応答を対応付けたコーパスに基づいた統計的な応答選択を行うことで，高い応答精度が期待できる．しかし，個々のタスクごとに大量の質問応答データを収集するのは多大な労力を要する．我々は，システム設計者が質問応答内容を表すキーワードと応答文の組をタスク知識としてあらかじめ用意することで，そこから頑健な応答選択を行う統計的音声対話システムを構築する手法について研究している．本稿では，CRF に基づく応答選択において，音素系列情報を考慮したモデル化を行う手法について提案し，実験を行った結果，応答正答率が 1.90%向上した．

1. はじめに

近年，音声認識や音声合成の技術発展, 計算機性能の向上により音声対話システムが注目され始めている．我々は音声対話システムの中でもユーザの一つの質問に対して適切な一つの応答を返す一問一答形式の音声対話システムを研究している．

音声対話システムを高い精度で実現する手法として統計的手法が主に用いられる．統計的手法において音声認識ではシステムのタスクに沿ったデータから統計モデルの学習を行い，応答選択では，質問文と応答文を対応付けた対話データから統計モデルを学習する．このとき，入力であるユーザの発話には，独特な言い回しや表現方法などのユーザの個人性が存在するため，全ての個人性に対応できるコーパスを構築することは難しく、殆どの多くのユーザの個人性に対応できるコーパスを構築するにもコストがかかる．また，対話データはタスク知識に強く依存するため，あらゆるタスクにおいて，高精度なシステムを構築するために必要な発話データを大量に収集することは困難である．

我々はタスクごとの対話データの大規模収集を行わずに高精度な音声対話を実現する手法の一つとして，タスク知識としてキーワードと応答文のみが与えられた条件での音声対話システムの構築について研究している．ここで，キーワードとは応答選択においてタスク知識となりうる単語を指しており，例えば，「図書館はどこですか？」という質問文では「図書館」と「どこ」の単語がキーワードとなる．このキーワードを設計者側から与えることにより少量の記述でタスク知識を与えることができ, 容易に音声対話システムの構築が可能となる．また，キーワードのみを抽出するのではなく，キーワード以外の単語系列を応答選択に利用する手法 [1][2] が研究されている．これにより少ないタスク知識における音声対話システムの応答精度を改善できることが示されている．

本研究では，認識誤りに頑健にするために，応答選択において利用する情報に音素系列に関する情報を追加することを提案する．音素系列を応答選択に利用することで, 音声認識誤りで出力された単語が正解の単語と比べて音素系列的に似通った単語であれば，応答選択で正しい応答文が選択される可能性が高まる．

以下, 第 2 節では一問一答形式の音声対話システム, 第 3 節ではキーワード以外の単語系列を利用する手法について述べ, 第 4 節では提案法である音素系列を用いた応答選択, 第 5 節では提案法の評価実験, 第 6 節では本研究についてまとめる．

2. 一問一答形式の統計的音声対話システム

一問一答形式の音声対話システムとは，ユーザの発話一つに対しシステムが一つ応答を返すシステムである．複数回ではなく一度の対話でタスクを終了するため，複雑な処理を行う必要がない．一問一答形式の音声対話システムは入力発話を認識する音声認識部，認識結果から応答文を選択する応答選択部，選択された応答文の音声を合成する音声合成部から構成される．音声認識部では統計に基づく手法が主流になっており，また応答選択部でも音声認識部と

[1] 名古屋工業大学大学院 工学研究科
Graduate School of Engineering, Nagoya Institute of Technology
[a)] saeki@slp.nitech.ac.jp
[b)] ri@nitech.ac.jp

同様に統計的手法を用いる研究が行われ，有効性が示されている．

一問一答形式の統計的音声対話システムは，ユーザが発話した質問発話から抽出した特徴量に対して出力確率が最大となる応答を選択するシステムとして考えられる．質問発話の音声信号系列を入力 O，それに対応する応答文を出力 A とするとき，出力確率が最大となる応答文 $\hat{A}$ を定式化すると式 (1) のように表現される．

$$\hat{A} = \arg\max_{A} P(A|O) \tag{1}$$

2.1 ディクテーションに基づく音声対話システム

質問発話の音声から直接応答文を選択することは困難であるため，質問発話の音声と応答文の間の中間表現として質問発話の音声認識結果の単語列 W を定義し，ディクテーションを行う．このとき式 (1) は，以下のように置き換えられる．

$$\hat{A} = \arg\max_{A} \sum_{W} P(A|W)P(W|O) \tag{2}$$

一般的な音声対話システムにおいて式 (2) の $P(A|W)$ は応答選択部，$P(W|O)$ は音声認識部を表している．このシステムは，質問発話の音声信号から応答文を選択する一般的な音声対話システムの枠組みであり，応答選択部と音声認識部それぞれ適切に学習された統計モデルを組み込むことが重要である．この手法を用いて高い精度を実現するためには大量の学習データが必要である．応答選択部 $P(A|W)$ では質問文と応答文が対応付けられた対話データから学習を行い，音声認識部 $P(W|O)$ では大量の発話データとテキストコーパスから学習を行う．また，応答選択部 $P(A|W)$ のモデル化に用いられる情報はタスク特有の情報であることが多く，タスクごとのデータ収集が必要である．

2.2 キーワードに基づく音声対話システム

文ではなくキーワードと応答文を対応付けることで，より頑健な応答選択を行う手法がある．代表的な実装方法としてディクテーションした単語列からキーワードを抽出する手法とキーワードのワードスポッティング手法がある．

ディクテーションを用いた手法は，キーワード列を K とすると，式 (3) と表すことが出来る．

$$\hat{A} = \arg\max_{A} \sum_{K,W} P(A|K)P(K|W)P(W|O) \tag{3}$$

式 (2) のシステムと式 (3) のシステムを比較すると，応答選択部 $P(A|K)$ において必要となるデータは発話文 O ではなくキーワード K のみである．キーワード抽出にはあらかじめ登録された単語を正確に抽出できるテキストマッチングなどが考えられる．また，応答選択部 $P(A|K)$ において実際使用されるデータはキーワードのみであるため，音声認識部 $P(W|O)$ では文の認識精度よりキーワードの認識精度が主要となる．

一方，ワードスポッティングという手法は以下の式のように表せる．

$$\hat{A} = \arg\max_{A} \sum_{K} P(A|K)P(K|O) \tag{4}$$

式 (4) の音声認識部 $P(K|O)$ では，入力発話から直接キーワードを認識している．ワードスポッティング手法では，ガーベージモデルを用いることで発話音声中からキーワードを抽出する手法が主流である [3]．ガーベージモデルとは任意の発話にマッチングさせるモデルで，キーワード以外の単語列の認識にこのモデルを適用する．一般的には音素連結モデルや N-gram モデルがガーベージモデルとして与えられる．

2.3 現状の課題

これらのシステムを高い精度で構築しようとする際の問題をまとめる．まず一つ目に，音声対話システムに使用される統計モデルの学習に膨大な学習データが必要となる点である．統計モデルの学習データには，音声認識部では大量の発話音声とテキストコーパス，応答選択部には質問文と応答文の組み合わせを書き起こした対話データが必要となる．また，ユーザの発話にはそのユーザごとの言い回しや表現が存在する．それら全てに対応するようなテキストコーパスや対話データを構築するのは困難である．

二つ目に，質問文と応答文の内容がタスクへの依存が高い点である．音声認識部で使用される統計モデルは比較的タスクへの依存が低いため，ある程度タスク間での共有が可能であるが，応答選択部において使用される質問と応答の内容はタスク特有の内容が多く含まれる．よって，タスクごとに質問と応答の内容の対話データを収集するには高いコストが必要となる．

3. キーワードとガーベージに基づく音声対話システム

本研究はタスク知識としてあらかじめキーワード K と応答文 A の組の集合が与えられている条件下で，キーワードとガーベージを用いた音声対話システム [4] をベースとする．これは，なるべく少ないタスク知識を人手で与えるだけで高精度な統計的音声対話システムを実現するような応答選択の枠組みである．システムの全体構成を図 1 に示す．以下，本システムについて概説する．

3.1 システム構成

ガーベージとは，音声認識によって得た単語列においてキーワード以外の単語列のことを指す．本システムではこのガーベージを応答選択の補助をする情報として利用する．ガーベージを G とすると，キーワードとガーベージに基づく統計的音声対話システムは，式 (5) で表せる．

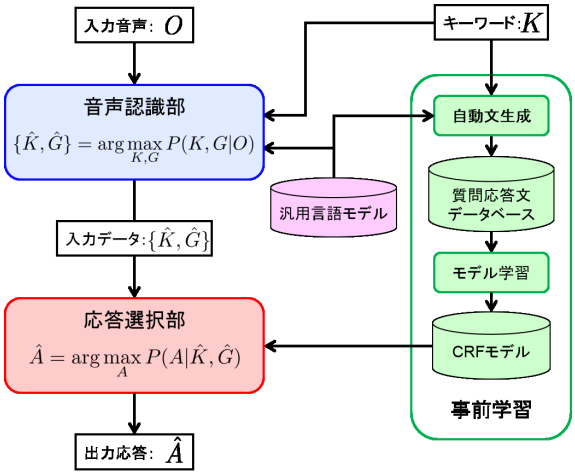

図 1 文生成に基づく音声対話システムの全体構成

$$\hat{A} = \arg\max_{A} \sum_{K,G} P(A|K,G)P(K,G|O) \quad (5)$$

式 (5) の $\{K,G\}$ は入力発話の認識結果の単語列を表しており，$\{K,G\}$ はキーワードとガーベージが区別されてデータ内に存在していることを示している．また，$P(K,G|O)$ と $P(A|K,G)$ はそれぞれ音声認識部と応答選択部を表している．

音声認識部では，ユーザが発話した音声から $P(K,G|O)$ が最大となる $\{\hat{K},\hat{G}\}$ を出力する．ここでは，ワードスポッティング手法 [5] を用いており，入力発話からタスクごとに与えられたキーワード集合を抽出すると同時に，キーワード以外の発話部分 G については，ガーベージモデルとして用いる汎用 N-gram 言語モデルの認識結果を用いる．これにより，出力された単語列を形態素解析後，応答選択部へ渡される．

応答選択部は，音声認識部から与えられる $\{\hat{K},\hat{G}\}$ に対して出力確率が最大となる $\hat{A}$ を出力する．モデルの学習には，後述する条件付確率場 (Conditional Random Fields; CRF) [6] に基づく対応付けを行なっている．CRF の学習データとして応答文 A に対応する $\{K,G\}$ の大量の組み合わせが必要であるが、これは認識時に用いるものと同じ汎用 N-gram 言語モデルを用いて K に接続しやすい G を生成し，K を含む仮想発話文集合を作成する．

このように，キーワードのみを用いる手法に比べ，キーワード以外の部分についてもタスク非依存の汎用言語モデルを用いて認識結果と応答選択を照合することで，頑健な応答選択が行える．

3.2 汎用 N-gram 言語モデルに基づく自動文生成

キーワードを用いた自動文生成では，N-gram に基づく文生成手法を行う．この手法は，まず，キーワードまたは文開始記号 $< S >$ · 文終了記号 $< /S >$ に挟まれた区間ごとに単語を探索し，尤度が高い順に並べる．その後，区間ごとに探索した単語列の結合を行うことで文を大量に生成する．最後に，生成された文のうち尤度の上位 N 文を選択し，単語列に分割後，対応する応答文と組み合わせることで，CRF の学習に用いるデータを作成する．

表 1 CRF の学習データの例

Input data	Output label
今日	今日の天気は、晴れです。
の	今日の天気は、晴れです。
天気	今日の天気は、晴れです。
を	今日の天気は、晴れです。
教え	今日の天気は、晴れです。
て	今日の天気は、晴れです。

3.3 CRF に基づく応答選択

CRF に基づく応答選択 [7] では，認識結果の単語列とそれに対応する応答文との間に存在する特徴を表す素性関数，およびその重みを用いて対応をモデル化する．

学習データは，表 1 のように入力質問文の形態素解析した結果の単語列に，対応した応答文を割り当てる．応答文の選択では，出力される識別候補は第 5 位まで使用し，出力される応答文が全て一致する候補の中で最尤の応答文を出力する．

また，使用する素性テンプレートは，$< m,r >$ と $< r >$，$< r,r' >$ である．m は単語，r は応答文，r' は 1 つ前の応答文を表している．

4. 音素系列を付加した頑健な応答選択の検討

本システムは文生成を用いない手法に比べて高い応答精度が得られることが示されているが，キーワードやガーベージ部分の認識誤りを考慮していない．そこで，本研究では，文生成と認識結果においてガーベージに加えて音素系列も用いる応答選択手法を提案する．

本節では，音素系列付与による応答選択の頑健化，モデルの学習方法，および音素系列を追加したシステムの全体構成について述べる．

4.1 音素系列付加による応答選択の頑健化

発話した音声に認識誤りが発生した場合でも，発話したユーザからすると意図的に異なる単語を認識させようとした訳ではない．よって，認識誤りが発生した単語は，元々認識させようとした単語と比較しても，音素列の観点で見たとき，似ている単語であると考えられる．例えば，表 2 のようにユーザが「高校」と発話し，「方向」と認識されてしまった場合，単語として「高校」と「方向」は全く異なる単語である．しかし，これらの単語を音素系列に変換すると「高校」は”k o: k o:” と，「方向」は”h o: k o:” となり，音素系列の視点で見ると”o: k o:” が共通しており，非常に似ている単語であることが分かる．これを利用し，応答選択部で利用されるタスク知識に音素レベルの情報を組

表 2　単語の音素系列

単語	高校	方向
単語の音素系列	k o: k o:	h o: k o:
音素系列の共通部分	o: k o:	o: k o:

み込むことで，認識誤りが発生した単語が入力された場合でも，正しい応答文を選択する可能性が高まることが期待できる．

4.2　音素系列の学習

音素系列を応答選択の CRF に組み込んだ場合の入力系列と出力系列を述べる．このシステムにおいて，入力系列は質問文の単語列と質問文の音素系列，出力系列は応答文である．このとき，質問文は形態素解析後の単語列を系列として入力している．質問文の単語列を $\boldsymbol{x}^{(1)} = \{x_1^{(1)}, x_2^{(1)}, \cdots, x_n^{(1)}\}$，質問文の音素系列を $\boldsymbol{x}^{(2)} = \{x_1^{(2)}, x_2^{(2)}, \cdots, x_m^{(2)}\}$，応答文を $\boldsymbol{y} = \{y_1^{(1)}, y_2^{(1)}, \cdots, y_n^{(1)}\}, \{y_1^{(2)}, y_2^{(2)}, \cdots, y_m^{(2)}\}$ とした時，入力系列が与えられた時の出力系列が出力される確率 $P(\boldsymbol{y}|\boldsymbol{x})$ は以下のようになる．

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp\left\{\left(\sum_{i=1}^{n}\sum_{k=1}^{K^{(1)}} \lambda_k^{(1)} f_k^{(1)}(x_i^{(1)}, y_i^{(1)})\right) + \left(\sum_{i=1}^{m}\sum_{k=1}^{K^{(2)}} \lambda_k^{(2)} f_k^{(2)}(x_i^{(2)}, y_i^{(2)})\right)\right\} \quad (6)$$

式 (6) において，$K^{(1)}, K^{(2)}$ は素性数，$f_k^{(1)}, f_k^{(2)}$ は単語列と音素系列の学習する特徴を表す素性関数である．また，それぞれの素性関数を以下のように表せる．

$$f^{(1)}(x_i, y_i) = \begin{cases} 1 & \text{if } (x_i^{(1)}, y_i^{(1)}) \text{exists} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$f^{(2)}(x_{i-1}, x_i, x_{i+1}, y_i) = \begin{cases} 1 & \text{if } (x_{i-1}^{(2)}, x_i^{(2)}, x_{i+1}^{(2)}, y_i^{(2)}) \\ & \text{exists} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

素性テンプレートは，$< m, r >$ と $< r >$，$< r, r' >$，$< p', p, p'', r >$ を使用する．m は単語，r は応答文，r' は 1 つ前の応答文，p は当該音素，p' は 1 つ前の音素，p'' は 1 つ後の音素を表している．この時，式 (7) は $< m, r >$，式 (8) は $< p', p, p'', r >$ を示している．音素系列の学習では，出現する音素と応答文を直接対応付けても，応答選択としての情報が少ないと考えられる．したがって，式 (8) のように，音素の前後情報を考慮した素性の学習を行う．

4.3　システムの全体構成

提案システムの全体構成を図 2 に示す．赤の点線の枠が提案法で追加した部分であり，応答選択で利用される情報に音素系列を追加している．応答選択部へ入力する音素系列は，音声認識結果の音素系列を用いる．また，CRF モデルの学習に用いる音素系列は，汎用 N-gram 言語モデルの認識用辞書から生成文に対応する音素系列に変換することで生成する．

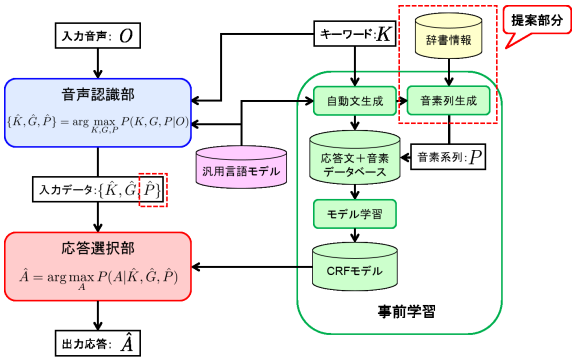

図 2　提案システムの全体構成

表 3　評価用データベースの詳細

	学習データ	テストデータ
データ数	12664	792
キーワードセット種類	499	173
キーワード種類	378	148
キーワード述べ	13779	803
応答種類	168	85

表 4　データセットの例

キーワードセット	応答文
たけまる君　おはよう	おはようございます
また来ます	お元気で
友達	友達になってください
すごい	すごいでしょ
英語	その件にかんしては、勉強中です

5.　評価実験

一問一答形式の音声対話システム「たけまるくん」[8] で収集されたデータベースを用いて提案法を評価した．

5.1　実験条件

評価データベースの詳細を表 3 に示す．学習セットは，2002 年 11 月から 2004 年 10 月の期間（2003 年 8 月は除く）において収録された発話のうち，書き起こし文の出現頻度が 2 回以上の有効発話である．また，テストセットは 2003 年 8 月に収録された有効発話である．有効発話とは，「たけまるくん」に向けての質問発話として，人手で分類されたユーザ発話を指す．この有効発話と応答文の組を正解データとして、発話内の自立語をキーワードとして本実験タスク用の質問応答文を構築した．実験で利用したキーワードと応答文の例を表 4 に示す．

汎用言語モデルには CSRC で配布されている Web 言語モデル (WebLM) [9] を使用した．WebLM は，タスクを限定しない Web から収集したテキストコーパスで学習した

表 5 システムの応答正答率

手法	応答正答率 (%)
従来法	81.31
提案法	83.21
発話音声の書き起こしを用いたシステム	92.30
書き起こしに音素系列を付与したシステム	95.45

単語 3-gram モデルで，語彙数 60,250 語である．音響モデルには，CSRC 最新版 [9] に含まれる全世代話者用音響モデルのうち，総状態数 3000，コードブック数 129，1 コードブックあたり 128 混合分布を持つ PTM モデルを使用した．

音声認識部のキーワードスポッティングは大語彙連続音声認識エンジン Julius[10] に実装した．応答選択部の質問文生成の実装には，Julius rev4.1.2 のライブラリを使用した．また，CRF に基づく応答選択には，CRF++[11] を使用した．

音声認識時の単語列探索時のビーム幅は 2500，言語重みは 8.0，挿入ペナルティは-2.0 である．自動生成時の生成文数は，尤度の上位 N 文を利用するかで決まり，499〜4990 文までである．CRF 学習のハイパーパラメータは $C = 10, 20, 30, 35, 40, 45, 50$ の中から最大となる応答正答率を用いる．また，キーワードスポッティング時のキーワードへの遷移確率を-1.0〜-4.0 まで 0.1 刻みずつ変動させた中から最大となったものを用いている．

システムの評価は応答正答率を比較することで行う．応答正答率は，入力発話文に対して選択される応答文があらかじめ付与されている正解ラベルと一致するかどうかで算出する．また，評価実験において以下の 4 つのシステムを比較する．書き起こしのデータは，音声認識誤りがない場合の応答選択の結果を比較するために使用した．

- 従来法
  キーワードとガーベージを用いた文生成に基づく統計的音声対話システム
- 提案法
  上記の手法に音素系列の情報を付与したシステム
- 発話音声の書き起こしを用いたシステム
  入力するテストデータにユーザが発話した音声を正しく書き起こしたデータを用いたシステム
- 書き起こしに音素系列を付与したシステム
  上記の書き起こしのデータを用いたシステムに音素系列を付与したシステム

5.2 提案システムの評価

各手法の応答正答率を表 5 に示す．提案法である音素情報を追加した統計的音声対話システムでは，従来法と比較して 1.90 %の応答正答率の向上が見られた．実際の実験結果から抽出した例を表 6 に示す．例 1 は，提案手法によって応答が正解となった例である．ユーザが「さようなら」と発話し，認識結果が「おはよう奈良」となり，従来法では，「おはよう」と「奈良」の 2 つの単語が入力され，「おはよう」と言う単語から「おはようございます」と言う応答文が誤って選択された．しかし，提案システムでは，「おはよう奈良」の音素系列である”ａｈａｙｏ: ｎａｒａ” の音素系列を組み入れた結果，正しい応答文の「さようなら」が選択された．このことから，前後情報を考慮した音素系列の学習によって単語の認識誤りに強い応答選択が可能となることが分かる．また，質問文が正しく認識された場合でも，従来法では間違った応答文を，提案法では正しい応答文を返した事例が存在した．これは，応答選択において単語の学習では補えきれない範囲を音素系列の学習によって補っていると考えられる．

次に，書き起こしデータとそれに音素系列を追加したデータとの結果を比較すると，従来法と提案法の応答正答率の向上率の 1.90 %より高い 3.15 %の向上が見られた．例としては，表 6 の例 2 のように，「今日は暑いね」という入力文の応答文として，書き起こしデータを用いたシステムは，「奈良県の天気は〜です」を選択したが，書き起こしデータに音素系列を付加したシステムは，正解の応答文である「暑いですね」を選択した．この提案法の目的として，音素レベルの情報を用いることで認識誤りの補正をすることであったが，認識誤りのない状態に音素情報を追加することによっても，応答正答率が向上した．

提案法では応答正答率は向上したが，全ての入力において正しく応答文を導くことが出来た訳ではなく，逆に元々正しい応答文が選択されていたにも関わらず，音素系列の学習によって間違った応答文が選択された場合が存在した．また，音素系列の前後情報の学習では補えない認識誤りが存在した．この認識誤りの例としては，表 6 の「何時で歌」という入力文の例である．この入力文の正しい認識は「何時ですか」であり，これらの文を音素系列に変換すると，それぞれ”ｎａＮｊｉｄｅｕｔａ” と”ｎａＮｊｉｄｅｓｕｋａ” となる．音素系列を比較すると”ｄｅｕｔａ” と”ｄｅｓｕｋａ” の部分が似ていることが分かる．しかし，この学習ではこのような音素系列の繋がりを学習することができず，従来法では正しい選択が行われたにも関わらず，提案法ではこの入力文に対する応答文は間違った選択がなされていた．前後情報以外の情報も活用する学習や，テストデータから，認識誤りのパターンとしては，母音の認識は比較的安定しているが，子音の認識で誤りが発生し，入力する音素系列に影響を及ぼしているパターンが多く存在していたため，母音をより利用する学習が有効的ではないかと考えられる．なお，モノフォンで音素系列の学習を行った場合では従来法と比べ，応答精度が 0.88%低下した．したがって，コンテキスト情報が必要であることが示された．

6. むすび

本研究では，キーワードと応答文のみの少量のタスク知

表 6 応答選択結果

	例 1	例 2	例 3
ユーザ発話	さようなら	今日は暑いね	何時ですか
発話の音素列	s a y o: n a r a	ky o: w a a ts u i n e	n a N j i d e s u k a
認識結果	おはよう奈良	今日は暑いね	何時で歌
認識結果の音素系列	o h a y o: n a r a	ky o: w a a ts u i n e	n a N j i d e u t a
従来法の選択文	おはよう	奈良県の天気は〜です	今は，〜時〜分です
提案法の選択文	さようなら	暑いですね	さようなら
正しい応答文	さようなら	暑いですね	今は，〜時〜分です

識しか与えられない条件下で，認識結果の音素系列の情報を付与した応答選択部による音声対話システムを提案した．音素系列の情報を加えたタスク知識を利用することで，認識誤りに頑健な応答選択が可能な音声対話システムを構築した．

評価実験では，従来法であるキーワードとガーベージを用いた文生成に基づく音声対話システムと比較して，応答正答率が 1.90 %の向上した．音素系列の有効性が確かめられた．今後の課題として，応答選択のモデル学習に最適な音素の特徴を学習することで，更なる精度の向上が期待できる．また，音素系列以外の情報を応答選択に用いることを検討する必要がある．

参考文献

[1] 吉見 孔孝，南角 吉彦，李 晃伸，徳田 恵一，“音声対話システムのための N-gram に基づくキーワードからの文生成”，電子情報通信学会技術研究報告，SP2009-83, PP.71-76, Dec.2009.
[2] 平野 隆司，南角 吉彦，李 晃伸，徳田 恵一，“双方探索に基づく N-gram に基づくキーワードからの文生成”，日本音響学会 2011 年春季研究発表会，2-P-40(b), Mar. 2011.
[3] 河原 達也，宗続 敏彦，三木 清一，堂下 修司，“会話音声中の単語スポッティングのための言語モデルの検討”，電子情報通信学会技術研究報告, SP94-28, pp.41–48, June. 1994.
[4] 平野 隆司, 加藤 杏樹, 南角 吉彦, 李 晃伸, 徳田 恵一, “登録キーワードと汎用言語モデルを用いた音声認識部・応答選択部の密結合に基づく統計的音声対話システム”, 情報処理学会研究報告. SLP, 音声言語情報処理 2012-SLP-92(3), 1-6, 2012-07-12.
[5] 加藤 杏樹，南角 吉彦，李 晃伸，徳田 恵一，“音声対話システムのためのキーワードの共起制約に基づくスポッティングアルゴリズムの評価”，信学技報，vol.110, no.357, pp.25-30, Dec. 2010.
[6] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” Proc.of ICML, pp.282-289, 2001.
[7] Y. Yoshimi, R. Kakitsuba, Y. Nankaku, A. Lee, and K. Tokuda, “Probabilistic Answer Selection Based on Conditional Random Fields for Spoken Dialog System,” Proc. of ICSLP,pp.215–218, 2008.
[8] 西村 竜一，西原 洋平，鶴見 玲典，李 晃伸，猿渡 洋，鹿野 清宏，“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”，電子情報通信学会論文誌，Vol.J87-D2, No.3, pp.789-798, 2004.
[9] 連続音声認識コンソーシアム (CSRC) http://www.lang.astem.or.jp/CSRC/
[10] Akinobu Lee and Tatsuya Kawahara,“Recent Development of Open-Source Speech Recognition Engine Julius”,Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp.131–137, Oct. 2009, Sapporo, Japan.
[11] CRF++ http://crfpp.googlecode.com/svn/trunk/doc/index.html