音声言語情報処理 43－1
（２００２．１０．２５）

音声対話システムにおける発話予測を利用した音声認識

玉井孝幸　堀内靖雄　市川熹

概要

本稿では、音声対話システムにおいて、次発話の予測情報を利用して自然発話の認識を行なう音声認識手法を提案する。本手法では、発話状態ごとに得られる次発話の予測文候補から、認識辞書と言語モデルを動的に生成し、それを用いて発話認識を行なう。このとき、自然発話の発声を許容するように認識辞書と言語モデルを生成する。さらに、大語彙音声認識用の言語モデルを用いた音声認識を並列に実行し、認識尤度を比較することによって、発話予測失敗の検出を可能とした。評価実験の結果、フィラー挿入文以外の自然発話に対して 100%の認識率が得られ、フィラー挿入文に対しても 97.4%という高い認識率が得られた。また、予測失敗時の検出率も 96.2%という高い数値が得られ、本手法の有効性が示された。

Speech Recognition Using Prediction for Spoken Dialogue System

Takayuki Tamai, Yasuo Horiuchi, Akira Ichikawa

Abstract

In this paper, we propose a method of speech recognition for spontaneous speech using prediction of the next user's utterance in a spoken dialogue system. A dictionary and a language model for speech recognition are generated dynamically based on a set of sentences which is predicted by the condition of the proceeding dialogue. The dictionary and the language model are modified so that the system can recognize spontaneous speech including inversion, ellipsis, fillers etc. Furthermore, the system can detect whether the prediction is correct by performing the usual speech recognition method in parallel and comparing the results by these two recognition methods. The result of the experiments shows the effectiveness of the recognition of predicted utterances and the detection of failures of prediction.

千葉大学
Chiba University

1　はじめに

近年、音声認識は統計的言語モデルを用いた手法が主流となっており、音声対話システムにおいても単語 N-gram などの統計モデルを用いて音声認識を行ない、得られた認識結果からキーワードを抽出し意味スロットを埋めていくという意味理解手法が一般的に用いられている。しかし、このように統計モデルを用いて音声認識を行なうとき、認識結果が受理不可能な認識結果となる場合がある。特に、フィラーや倒置などを含む自然発話に対してはそのような傾向がある。

また、音声対話システムにおいて、探索空間を小さくして認識性能を向上させることを目的として次発話の予測情報を用いる手法が提案されている。その手法には、発話状態遷移モデルを次発話の予測に利用しているもの[1]、次発話の発話内行為を予測しているもの[2]、発話から意味表現を統計的に推定するもの[3]などがあり、予測情報を使うことの有効性を示しているが、自然発話の認識を考慮に入れているものはほとんどない。しかし、発話予測を行ない探索空間を極端に小さくすることにより、自然発話にも対応するように言語モデルを構築することが可能であると考えられる。そこで本研究では、音声対話システムにおいて、次発話の予測情報を利用し、自然発話の認識を行なえる音声認識手法について提案する。

2　システム概要

現在、想定している音声対話システムのシステム概要を図 1 に示す。

本システムは、千葉大学構内の道案内をタスクとする音声対話システムである。対話制御部において、状態遷移モデルを用いて次発話の予測文候補列を作成する。そして、音声認識部がその予測文候補列を受け取り、自然発話の認識を可能とするように認識辞書および言語モデルを生成する。この認識辞書および言語モデルを用いてユーザの発話の認識を行なう。このとき、認識エンジンには Julius[4]を使用する。また、発話予測の成否判定のために、この認識と並行して大語彙連続音声認識用の言語モデルである毎日新聞記事データベースを用いて構築された単語辞書と言語モデル[4]を用いた認識も行なう。ここで、発話予測が成功と判定された場合、得られた認識結果を対話制御部へ送り、その認識結果に対応した応答を行なう。発話予測が失敗と判定された場合は、認識結果をユーザに提示し、実際に間違った認識結果であるかを確認し、判定ミスによるシステム側の誤解を避けるようにする。

音声認識部における処理は、次章でさらに詳しく説明する。

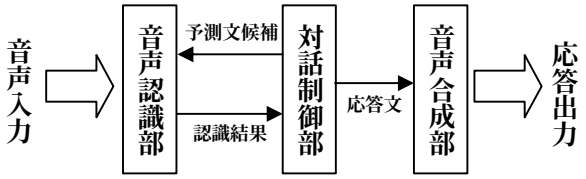

図 1　システム概要

3　音声認識部における処理

本手法では、音声認識というタスクを、予測文を認識するというタスクに置き換えることにより、探索空間を大幅に小さくする手法を検討する。予測文を認識するためには、発話ごとに予測文認識のための言語モデルを動的に作成する必要がある。

3.1　言語モデルの作成

対話制御部で作成された次発話の予測文候補列からそれらの予測文を認識可能な認識用辞書と言語モデル（文節 bigram）を作成する。そのとき、自然発話の認識を可能とするように認識辞書と言語モデルを修正する。具体的には、倒置、文節の省略、助詞の欠落、フィラー・言い淀みの挿入に対する対応を行なった。

（a）倒置

与えられた予測文候補の倒置文を作成し、その倒置文も認識できるように認識辞書と言語モデルを修正する。

仮に「階段が 見えます」という 1 文が予測文候補として与えられたときの認識辞書と文節 bigram をそれぞれ図 2、図 3 に示す。そして、倒置への対応を施した後の認識辞書と文節 bigram をそれぞれ図 4、図 5 に示す。文節 bigram で、文節の右側の番号は文節番号を表している。この文節番号を付与することにより、言語モデルを修正する際に、「見えます階段が見えます」のような受理不可能な認識結果を許容しないようにしている。

<s> [] silB
</s> [] silE
階段が 0 [かいだんが] k a i d a N g a
見えます 1 [みえます] m i e m a s u

図 2　認識辞書の例

<s> 階段が 0
階段が 0 見えます 1
見えます 1 </s>

図 3　bigram の例

<s> [] silB
</s> [] silE
階段が 0 [かいだんが] k a i d a N g a
見えます 1 [みえます] m i e m a s u
見えます 2 [みえます] m i e m a s u
階段が 3 [かいだんが] k a i d a N g a

図 4　倒置への対応を施した認識辞書の例

<s> 階段が 0
<s> 見えます 2
階段が 0 見えます 1
見えます 1 </s>
見えます 2 階段が 3
階段が 3 </s>

図 5　倒置への対応を施した bigram の例

(b) 文節の省略

発話の際に省略可能な文節に対しては、次発話予測部において、「+はい +分かりました」のように予め人手により印（+）をつけておくことにし、予測文候補の中にその印のついた文節があった場合、倒置への対応と同様に、印のついた文節を省略した省略文を作成し、それらの文も認識できるように認識辞書と言語モデルを修正する。

(c) 助詞の欠落

発話の際に欠落する可能性のある助詞(「が」や「は」など)が予測文候補の文節末に出現している場合には、図 6 のように、認識辞書においてその助詞を省略した発音系列を省略していない発音系列と並べて書き加えるようにする。これにより、助詞が欠落して発話された場合にも助詞がそのまま発話された場合にも同様に認識することができる。

<s> [] silB
</s> [] silE
階段が 0 [かいだんが] k a i d a N g a
階段が 0 [かいだんが] k a i d a N
見えます 1 [みえます] m i e m a s u

図 6　助詞欠落への対応を施した認識辞書の例

(d) フィラー・言い淀みの挿入

フィラーや言い淀みは、ほとんど文頭または文節間に現れる。そこでまず、フィラーを認識可能とするために、フィラークラス<filler>を作成し、辞書に様々なフィラーを登録しておく。そして、フィラークラスを挿入した文節 bigram を、文頭にフィラーが挿入されても認識可能となるように作成する。このとき、文節間にもフィラークラスを挿入するという考えもあるが、その場合、辞書と言語モデルのサイズが膨大になってくるという問題が出てくるのと、本手法による言語モデルを用いた場合、文節間で現れるフィラーは文頭で現れるフィラーに比べ認識に与える影響は小さいと考えられるため、文頭への挿入のみとした。

さらに、文節間の言い淀みに対応させるために、文節末に無音モデルを付与するという処置を施した。この処置により、文節間で発生する冗長区間がこの無音モデルと対応が取られるようになり、認識率が上がると期待される。今回使用した音響モデルは Julius とセットで公開されている triphone モデルであるが、無音モデル『sp』を含む triphone が存在していないため、無音モデルと同等の性質を持つ促音モデル『q』を文節末に付与した。

フィラー、言い淀みへの対応を施した認識辞書と文節 bigram の例をそれぞれ図 7、図 8 に示す。この例では、フィラーに「えっと」と「えー」を登録している。

最後に、「階段が +見えます」という 1 文が予測文候補として与えられた場合における(a)～(d)全ての対応を施した認識辞書と文節 bigram の例をそれぞれ図 9、図 10 に示す。この例は、予測文候補が 1 文のときの例だが、実際には予測文候補は複数である。

<s> [] silB
</s> [] silE
<filler>0 [] e q t o
<filler>0 [] e q t o q
<filler>0 [] e:
<filler>0 [] e: q
階段が 1 [かいだんが] k a i d a N g a
階段が 1 [かいだんが] k a i d a N g a q
見えます 2 [みえます] m i e m a s u
見えます 2 [みえます] m i e m a s u q

図 7　フィラーへの対応を施した認識辞書の例

<s> <filler>0
<s> 階段が 1
<filler>0 階段が 1
階段が 1 見えます 2
見えます 2 </s>

図 8　フィラーへの対応を施した bigram の例

<s> [] silB
</s> [] silE
<filler>0 [] e q t o
<filler>0 [] e q t o q
<filler>0 [] e:
<filler>0 [] e: q
階段が 1 [かいだんが] k a i d a N g a
階段が 1 [かいだんが] k a i d a N g a q
階段が 1 [かいだんが] k a i d a N
階段が 1 [かいだんが] k a i d a N q
階段が 2 [かいだんが] k a i d a N g a
階段が 2 [かいだんが] k a i d a N g a q
階段が 2 [かいだんが] k a i d a N
階段が 2 [かいだんが] k a i d a N q
見えます 3 [みえます] m i e m a s u
見えます 3 [みえます] m i e m a s u q
見えます 4 [みえます] m i e m a s u
見えます 4 [みえます] m i e m a s u q
階段が 5 [かいだんが] k a i d a N g a
階段が 5 [かいだんが] k a i d a N g a q
階段が 5 [かいだんが] k a i d a N
階段が 5 [かいだんが] k a i d a N q

図 9　自然発話への対応を施した認識辞書の例

<s> <filler>0
<s> 階段が 1
<s> 階段が 2
<s> 見えます 4
<filler>0 階段が 1
<filler>0 階段が 2
<filler>0 見えます 4
階段が 1 </s>
階段が 2 見えます 3
見えます 3 </s>
見えます 4 階段が 5
階段が 5 </s>

図 10　自然発話への対応を施した bigram の例

3.2　発話予測の成否判定

これまで説明してきた音声認識手法では、ユーザの発話が予め予測していた予測文候補に含まれていることが前提で認識を行なっており、予測に失敗した場合を考慮に入れていない。しかし、発話予測が毎回必ず当たるとは限らない。発話予測が外れた場合には外れたということを知る事が対話制御を行なう上で必要不可欠である。また、発話予測に成功した場合、その情報をシステム側が得ることができれば、認識結果に対する確認応答をする必要がなくなるため、対話のやり取りをよりスムーズに行なうことができる。

そこで本研究では、発話予測の成否を判定するために、予測に基づく認識と並行して大語彙連続音声認識用の言語モデルである毎日新聞記事データベースを用いて構築された単語辞書と言語モデルを用いた認識も行ない、得られた認識結果から認識尤度を抽出する。複数の認識器を用いる手法には、発話内容ごとに言語モデルを用意する手法があるが[5]、本手法は、予測による小語彙モデルと、大語彙モデルを用意し、大語彙モデルによる認識では、主に認識尤度のみを抽出する。

そして、得られた 2 つの認識尤度を比較すると、予測に成功していたとき、つまり、発話内容が予測文候補に含まれていたときには予測に基づく認識により得られた認識尤度の方が高くなり、予測に失敗していた時、つまり、発話内容が予測文候補に含まれていなかったときは大語彙認識用の言語モデルを用いた認識により得られた認識尤度の方が高くなると考えられる。よって、2 つの認識結果から得られる認識尤度を比較するこ

とにより、発話予測の成否判定をすることができると考えられる。

4　評価実験

4.1　提案手法の性能評価

まず、本手法により次発話の予測情報を利用して生成された言語モデルの性能を評価するための実験を行なった。

言語モデルの生成に用いた予測文候補は、それぞれ 5〜12 文ずつの予測文候補 6 パターンである。そして、認識対象とした文は、予測文候補そのままの文(予測基本形文)42 文、倒置文 46 文、助詞欠落文 24 文、文節省略文 13 文、フィラー挿入文 58 文の計 183 文である。フィラー挿入文には、予測基本形文の文頭および文節間に「えっと」という言葉を挿入した文を用いた。この計 183 文の認識対象文を 8 名の被験者にそれぞれ発話してもらい、データを収録した。この発話の際に、フィラー挿入文における「えっと」の発話は、なるべく自然な形で発話するよう指示した。そして、収録データを、上述の手法を用いて認識させた。

その結果、表 1 に示すように、予測基本形文、倒置文、助詞欠落文、文節省略文では認識率 100%という最高の値を得ることが出来た。さらに、フィラー挿入文においても 97.4%という非常に高い認識率を得ることが出来た。

表 1　提案手法言語モデルの性能評価

	認識率
予測基本形文	100%
倒置文	100%
助詞欠落文	100%
文節省略文	100%
フィラー挿入文	97.4%

さらに、フィラー・言い淀みが挿入された場合の対応として、フィラークラスの挿入と、促音モデル『q』の付与を行なったが、この対応が実際にフィラーの挿入に有効であったかを確認するための評価実験を行なった。

フィラー・言い淀みへの対応を施さない言語モデルを用いた認識を行ない、先ほど得られたそれぞれの文の認識率と比較する。

得られた結果を表 2 に示す。この結果を見ると、フィラー挿入文以外の文には、フィラー・言い淀みへの対応を施さない場合においても認識率に変化は見られなかったが、フィラー挿入文においては認識率が 57.6%となり、フィラー・言い淀みへの対応を施した場合に比べて極端に認識率が落ちてしまっていることが分かる。

この結果より、フィラークラスの挿入および促音モデル『q』の付与が、フィラーが挿入された場合の認識に対して有効であるということが言える。

以上の実験結果から、本研究の提案手法である自然発話の認識を目的とした、次発話の予測情報を利用して言語モデルを作成し音声認識に利用する手法が、実際に自然発話認識に有効であると結論付けることができる。

表 2　フィラー・言い淀み非対応時の性能評価

	認識率
予測基本形文	100%
倒置文	100%
助詞欠落文	100%
文節省略文	100%
フィラー挿入文	57.6%

4.2　発話予測成否判定の性能評価

発話予測の成否判定がどの程度正確であるかを評価するために、発話予測成功時と発話予測失敗時に分けて、本研究で提案した言語モデル(提案モデル)と大語彙認識用の言語モデル(大語彙モデル)で同じように発話を認識させ、認識尤度の大小を比較した。認識対象文は、先ほどの実験で用いた計 183×8 文である。

発話予測成功時および発話予測失敗時における認識尤度の比較結果を表 3 に示す。ここで、発話予測成功時とは、発話内容が予測文候補の中に含まれていた場合のことを指し、発話予測失敗時とは、発話内容が予測文候補の中に含まれていなかった場合のことを指す。また、認識エラーとは、得られた認識結果がエラーとなり認識尤度が得られなかったものである。認識エラーとなった場合は、発話予測に失敗したものとしてみなすことができる。そして、提案モデルの方が認識尤度が高ければ発話予測成功と判定し、大語彙モデルの方が認識尤度が高ければ発話予測失敗と判定す

表 3　提案モデルと大語彙モデルを用いた認識尤度の比較

		発話予測成功時	発話予測失敗時
提案モデル尤度 ＞ 大語彙モデル尤度	(予測成功と判定)	90.4%	3.8%
提案モデル尤度 ＜ 大語彙モデル尤度	(予測失敗と判定)	9.6%	85.7%
認識エラー	(予測失敗と判定)	0%	10.5%

るため、発話予測成功時においては 90.4%の割合で判定成功、発話予測失敗時においては 96.2%の割合で判定成功であることが分かる。発話予測成功時における 9.6%の割合での判定ミスというのは少し多いが、発話予測に成功しているにも関わらず、発話予測失敗と判定されたときには、そこで得られた認識結果をユーザに提示し確認することにより、その判定はミスであり実際は発話予測に成功していて認識結果も正しいものであったということをすぐに判断することができる。問題となるのは、発話予測失敗時における検出率であるが、96.2%という高い検出率が得られており、かなり良い性能であると言うことができる。

また、実際に発話予測に成功しているかどうかという観点から、発話予測性能に対する発話予測成否判定率を求めた結果を図 11 に示す。

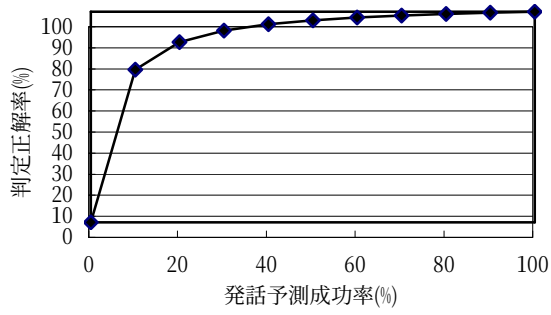

図 11　予測成功時における発話予測成功率に対する判定正解率

この図を見ると、発話予測性能が高ければ高いほど判定正解率は高くなっていることが分かり、予測成功率が 50%以上になると、95%以上の判定正解率を得ることができている。すなわち、発話予測性能が仮に 50%程度でも発話予測判定の失敗率は 5%以下であり、予測の精度が上がれば上がるほどその割合は小さくなる。

よって、これらの結果から、本手法が発話予測の成否判定手法として有効であるということが言える。

5　まとめ

本稿では、音声対話システムにおける音声認識手法として、発話予測を利用して自然発話の認識を行なう音声認識手法、および発話予測の成否判定手法を提案した。その結果、本手法が実際に自然発話認識に有効であり、さらに発話予測の成否判定手法も有効であることが確認できた。

問題点を挙げると、今回の評価実験に用いた発話が強制的な自然発話だったということが挙げられる。つまり、発話内容をあらかじめ指定しての発話だったため、完全な自然発話にはなっていない。よって、今後の課題として、本手法を実際に音声対話システムへ実装し、そのシステムを用いた評価実験を行ない、本手法の有効性を確認する必要がある。

また、本手法は発話予測の精度が高ければ高いほど有効であると考えられるため、本手法の音声認識がより効果的となるような精度の良い発話予測手法を検討していくことが必要である。

参考文献

[1] 鈴木雅実，松崎克郎，井ノ上直己，谷戸文廣．発話状態の予測に基づく対話音声認識手法とその効果．情報処理学会研究報告，96-SLP-12，pp.1-6，1996.

[2] M. Nagata, T. Morimoto. An Information- Theoretic Model of Discourse for Next Utterance Type Prediction．情報処理学会論文誌，Vol.35，No.6，pp.1050-1061．1994.

[3] 甘粕哲郎，村上仁一，小原永．音声対話システムにおける次発話予測の 1 手法．日本音響学会 2001 年春季講演論文集，pp.45-46，2001.

[4] 河原，李，小林，武田，峰松，嵯峨山，伊藤，伊藤，山本，山田，宇津呂，鹿野．日本語ディクテーション基本ソフトウェア（99 年度版）．日本音響学会誌，Vol.57，No.3，pp.210-214，2001.

[5] 田熊竜太，岩野公司，古井貞熙．並列処理型計算機を用いた音声対話システムの検討．人工知能学会研究会資料，SIG-SLUD-A201，pp.21-26，2002.