

対話システムにおける 応答生成手法の改良とその実装

八木 裕司[†], 高田 靖也^{††}, 広瀬 啓吉[‡], 峯松 信明^{††},

[†] 東京大学大学院 工学系研究科, ^{††} 東京大学大学院 新領域創成科学研究科,

[‡] 東京大学大学院 情報理工学系研究科

Tel.: 03-5841-6767, Fax.: 03-5841-6648

{yagi, stakada, mine, hirose}@gavo.t.u-tokyo.ac.jp

あらまし 音声認識や自然言語処理、音声合成技術の進歩に伴い、数多くの音声対話システムが構築されてきている。しかしながら、音声対話システムの多くは、音声合成部に既存のテキスト音声合成器を用いているため、応答生成の過程で得られる言語情報を音声出力に反映させることができない。そこで我々は、概念音声合成 (CTS : Concept-To-Speech) を用いた音声合成のために必要な言語情報の取扱いについて検討し、従来より柔軟な応答生成手法を実現した。また、これらを応答音声に反映させるために必要な韻律制御手法についても改良を行なった。そして、これらを道案内システムへ実装した。

キーワード 音声対話システム、対話管理、応答生成、概念音声合成、言語情報

Improved concept-to-speech generation in a spoken dialogue system

Yuji Yagi[†], Seiya Takada^{††}, Keikichi Hirose[‡] and Nobuaki Minematsu^{††}

[†] Graduate School of Engineering, University of Tokyo,

^{††} Graduate School of Frontier Sciences, University of Tokyo,

[‡] Graduate School of Information Science and Technology, University of Tokyo

Tel.: 03-5841-6767, Fax.: 03-5841-6648

{yagi, stakada, mine, hirose}@gavo.t.u-tokyo.ac.jp

Abstract Use of text-to-speech conversion devices in spoken dialogue system makes it difficult to well reflect higher-level linguistic information obtainable during sentence generation process on reply speech. A method is necessary to directly converting content of reply into speech. This method, known as concept-to-speech conversion, was realized for the reply speech generation in our spoken dialogue system on road guidance. In order to realize the concept-to-speech conversion, syntactic structure of phrases in user's inputs is kept and is utilized for the sentence generation. Also, the concept is handled in phrase unit instead of sentence to increase the variety in generated sentences. Several improvements are also added to prosodic control in speech synthesis. In the spoken dialogue system, a user was guided to reach a place marked on a map through conversation.

Key words Spoken dialog system, Dialog management, Response Generation, CTS conversion, Linguistic information

1 はじめに

音声情報処理技術の進歩に伴い、音声対話システムの研究が盛んに行なわれるようになった。音声対話システムの究極の目標は、人間同士のような自然な対話を人間とシステムの間で行なうことである。そのため、音声対話システムにおける応答生成に関して、柔軟かつ複雑な応答文の生成と、その応答文を対話調音声で合成すること、が要求される。

我々は既に概念音声合成の枠組を実現し、それをエージェント対話システムに実装してこの手法が有効であることを示している [1, 2]。本発表では、概念音声合成手法のために、より汎用的な応答文生成を行なうための手法について紹介するとともに、韻律制御手法に関しても改良を行ない、それらを道案内対話システムに実装した結果についての報告を行なう。

2 応答文生成手法

2.1 対話システムにおける音声合成

現在の音声合成技術は、テキスト音声合成 (TTS : Text-To-Speech) が主流となっている。しかし、TTS はテキストから朗読音声を作成することを目的として作られたものであり、対話システムにおける音声合成を考えると、TTS の適用は不適切であると言える。より自然な対話調音声を作成するためには、統語構造や談話情報などの高次の言語情報に基づいて韻律を制御できる音声合成の枠組が必要である。そこで、概念音声合成 (CTS : Concept-To-Speech) により応答音声を作成することが提案されている [3]。TTS がテキストを入力するのに対し、CTS ではシステムの内部表現 (概念) を利用して音声を作成する。CTS では文の生成過程で正確な言語情報が得られるため、統語構造を韻律に反映させたり、談話情報で韻律の焦点を制御したりするということが容易に行なうことができる。

2.2 概念音声合成の実現 [1, 2]

概念音声合成を用いることで、統語構造や談話情報などを応答文・応答音声に反映させることができる。そのため我々は [1] において、言語情報を一貫して構文木構造を保持したまま扱い、また、言

語情報の統一的な処理のためにタグを付与し、言語情報の取り扱いを容易にする、という手法を提案し、その有効性を確認している。図 1 は、[1, 2] において我々が構築したシステムにおけるタグと構文木構造の例である。

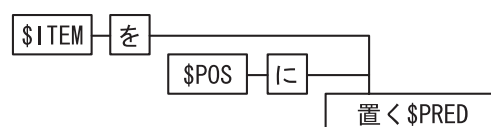


図 1. 「\$ITEM を \$POS に置く」という文章のタグと構文木構造

システム内部では、このタグと構文木構造を保持した言語情報を LISP 形式で表現している。例えば、図 1 を LISP 形式で表現すると、「(置く \$PRED (を (\$ITEM)) (に (\$POS)))」となる。このようにタグを埋め込んで表しておくことで、タグの部分に単語や句を接続する等を容易に行なうことができる。

応答文を作成する際には、文の部品を統語構造に従って接続し、タグを参照することで単語の重要度を与える。この時、活用形は構文木構造のかけ受けから一意に決定される。その後、韻律を制御するフレーズ指令・アクセント指令のパラメータ等をこの構文木構造やアクセント結合規則 [4] を基に決定し、音声として出力する [2]。

2.3 提案応答文生成手法

従来手法においては、\$ITEM タグや \$POS タグを用いてアイテムや場所の情報を表現していたが、この手法では定型文の \$ITEM タグや \$POS タグに単語を埋め込むといういわゆる「穴埋め問題」になっており、必ずしも柔軟な応答生成とは言えない、という問題があった。例えば図 1 は、従来構築していたシステムにおける「置く」という動作の定義そのものである。つまり、一つの定型文からは一通りの応答文しか生成することができず、少しでも異なる応答文を生成しようと思えば、異なる定型文を容易する必要があった。

この問題を解決し、より汎用的な応答生成を行なうために、従来のタグつき LISP を拡張する、ということを行なう。

今回提案する手法は、上記のような単語単位の

タグのみではなく、連文節を複数接続して応答文を生成する、というものである。本手法は従来の「穴埋め問題」とは異なり、定型文として与えるのは必要最小限（つまり連文節単位）に止める。この連文節の接続によって、より柔軟な、より汎用的な応答生成を行なう。

2.3.1 連文節

連文節単位で応答文の構成要素を生成する。例としては、「(に(\$NOUN))」や「(まで(\$NOUN(\$ADJ)))」のようなものが挙げられる。この段階で、既に統語構造や単語の重要度等を保持する。

2.3.2 文生成

前節で得られた連文節を、統語構造を考慮しながら接続することで応答文全体を生成する。接続方法によって、単文のみならず重文や複文といった種々多様な応答文が生成できると期待できる。

2.4 提案手法における応答文生成の例

例として、「右に曲がって駅まで行ってください」という応答文を生成する手順を以下に示す。

1. 「(に(\$DIRECTION))」というフレームから、「右に」という名詞句を生成する。
2. 「(まで(\$LANDMARK))」というフレームから、「駅まで」という名詞句を生成する。
3. 「(て(\$VERB(\$NOUN_PHRASE)))」というフレームから、「右に曲がって」・「駅まで行って」という動詞句を生成する。
4. 2つの動詞句を連結することで「右に曲がって駅まで行って」という連文節を生成する。
5. 「(ください(\$VERB_PHRASE))」というフレームから、「右に曲がって駅まで行ってください」という応答を生成する。

上記の例は重文であるが、同様の手法によって「右に曲がってください。そして駅まで行ってください」という2つの単文を生成することもできる。その際には、「そして」等の接続詞を適宜挿入する。

3 韻律制御手法

3.1 音声合成における韻律規則

本対話システムにおける音声合成には、対話音声の分析に基づいて韻律規則を規定したもの[5]が用いられている。この韻律規則は、フレーズ指令とアクセント指令の2種類の指令に対する規則からなる。これらの規則は、朗読調の音声に対して構築された規則[6]を、対話音声の分析結果に基づいて対話調音声向けに変換したものである。各指令の大きさは数量化分析によって決められており、数字は指令決定の際に考慮される各項目（パラメータ）のどのカテゴリに分類されるかの値を示す。

以下、フレーズ指令とアクセント指令のそれぞれについて、各々の指令を配置する位置を決定する指令の生成規則と、指令を構成するパラメータの意味について説明する。

3.1.1 フレーズ指令

フレーズ指令には、文頭・文中・文末の3種類の指令があり、文頭フレーズ指令(Ph)は $PI_1I_2I_3I_4I_5I_6$ 、文中フレーズ指令(Pm)は PI_2I_5 、文末フレーズ指令はP0と表される。フレーズ指令におけるパラメータの意味とその値を表1に示す。ただし、FRDとはFundamental Routine of Dialogueのことで、質問や要求の発話とその応答の組からなる対話の基本単位のことを指す。

表 1. フレーズ指令のパラメータ

	値	意味
I_1	1	FRDを開く
	2	FRDを閉じる
I_2	1	フレーズ中に重要語を含む
	2	フレーズ中に重要語を含まない
I_3	1	話題を変更している
	2	話題を変更していない
I_4	1	接続詞に従属する
	2	接続詞に従属しない
I_5	1	対応成分のモーラ数が7以下
	2	対応成分のモーラ数が8以上
I_6	1	疑問の終助詞「か」で終わる
	2	疑問の終助詞「か」で終わらない

3.1.2 アクセント指令

アクセント指令には、 $DI_1I_2I_3 \cdot FI_1I_2I_3 \cdot A0$ の 3 種類があり、 $DI_1I_2I_3$ は起伏型または頭高型のアクセントの立ち上げ、 $FI_1I_2I_3$ は平板型のアクセントの立ち上げ、 $A0$ は両者の立ち下げを示す指令である。アクセント立ち上げ指令におけるパラメータの意味とその値を表 2 に示す。

表 2. アクセント指令のパラメータ

	値	意味
I_1	1	重要で新しい
	2	重要でないが新しい
	3	すでに現れているが重要
	4	すでに現れていて重要でない
I_2	1	フレーズ先頭
	2	フレーズ途中
I_3	1	名詞
	2	動詞
	3	形容詞・副詞
	4	指示語・疑問詞
	5	接続詞

3.2 フレーズ指令挿入位置

フレーズ指令挿入位置決定規則は、以下のようになっている。

- 文の先頭には Ph を挿入し、文末には P0 を挿入する。文の境界には S1 を挿入する。
- ICRLB 境界には Pm を挿入する。ただし、直前の Ph/Pm からの距離が L1 モーラ以下であるときは、Pm を省略する。
- ICRLB 境界から文頭、文末、他の ICRLB 境界までのモーラ数が L2 より大きければ、単語間の結合の強さに従って、当該モーラ数の L2 の剰余個以下の Pm を挿入できる。

S1 は休止記号を表す。他に S2、S3 があり、数字が小さいほど休止の長さが長い。変数として、 $L1=5$ 、 $L2=12$ を用いている。ICRLB 境界とは Immediate Constituent with Recursively Left-Branching structure の略であり、文の構文木において、右枝分れ境界で前後を区切られ、かつ左枝分れのみを含む単語連鎖のことである。

単語間の結合強度については、まず毎日新聞 97 年度版 1 年分からデータベースを構築し、bigram を作成する。作成した bigram から、文節境界前後の単語についての結合強度確率 P を次式で求める。

$$P = \frac{f(w_2, w_1)}{f(w_1)}$$

文節境界前後の単語間の結合確率が高い箇所は言語的な結合が強いため、フレーズ指令を挿入しない。

3.3 アクセント指令挿入位置

日本語において、単語と単語が結合して文節や複合単語ができるとき、そのアクセントは構成要素それぞれを単独に発声したときのものとは異なるものになり、アクセント核の移動・生起・消失が起こる。この現象をアクセント結合と呼ぶ。音声対話システムにおいては、より自然な応答音声の生成が求められるため、個々の単語のアクセント型に加えて、このアクセント結合についても考慮する必要がある。[4] は、勾坂らによる一連の結合規則 [7] を聴取実験によって見直したものである。本システムにおいては、このうちの付属語アクセント規則及び文節間結合規則 [8] を用いている。

アクセント指令挿入位置の決定については、以下の手順で行なう。

1. 付属語アクセント規則に従い、各文節内の仮アクセント核を決定する。
2. 文節間結合規則に従い、文節間のアクセント結合を文頭から巡回評価する。ただし、フレーズ指令を挟んだ結合は行なわない。また、焦点の当てられている文節についても結合は行なわない。

3.4 音素・韻律記号列生成手法

これらの規則や構文木構造等の情報を用いて韻律制御記号を含む音素記号列を生成する手法は、以下の手順で行なう。

1. 構文木の接続に従って単語の活用形を決定
2. 構文木構造やモーラ数、単語間結合強度に従って、フレーズ指令挿入位置を決定
3. モーラ数や単語の重要度・新規性に従って、フレーズ指令のパラメータを決定

4. 品詞や単語の重要度・新規性、フレーズ指令挿入位置に従って、アクセント指令のパラメータを決定
5. アクセント結合規則に従って、アクセント指令挿入位置を決定

この結果、

P111212 hi F311 da ri e ma ga sx te A0
 P11 D311 zi A0 n zja ma de P21 i F413
 sx te ku da sa A0 i P0 S1

というような音素記号列が生成され、音声合成器に送られる。そして、音声合成器がこれを音声として出力する。

4 道案内対話システム概要

本システムのタスクは、仮想地図を用意し、システムがユーザに対して指示を行なうことで、ある地点から目標地点まで移動する、というものである。図2は本システムにおける仮想地図である。仮想地図において提示される情報としては、目標物の場所・名前、経路の距離があり、システム側はこれらの情報を全て保持している。図3は、実際にユーザに提示されるインターフェースであり、円はユーザの視界に相当する。つまり、ユーザは現在地にある目標物と道が出ている方向しかわからない。システムはあらかじめ目標地点までの最短経路を算出しておき、その経路をユーザに指示することでタスクの完了を目指す。

本システムは、音声認識部・構文解析部・対話管理部・音声合成部及びユーザ用 GUI (図3) から構成される。音声認識部には、Julian[9]を用いている。認識結果が構文解析部に渡され、構文木構造を持った言語情報が対話管理部に渡される。対話管理部はまず入力からユーザの位置情報を抽出し、ユーザに指示を行なうための応答文を、韻律制御記号を含んだ音素記号列として生成し、それが音声合成部において音声として出力される。音声合成器としては、波形接続方式の合成器が用いられている。

図3中における「工事中」等の一時的な情報はシステムは保持していない。その結果、システム

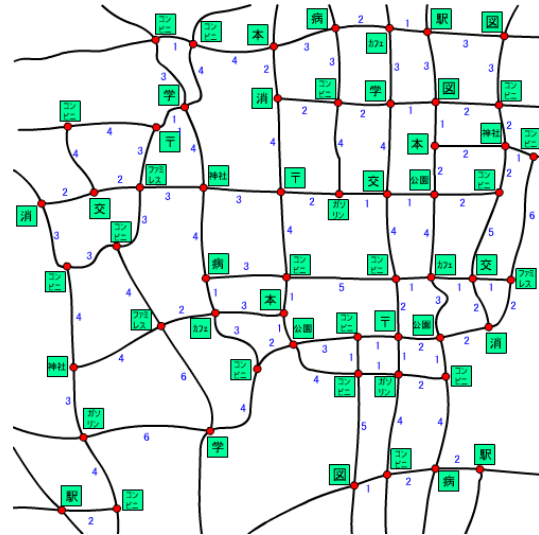


図 2. 仮想地図

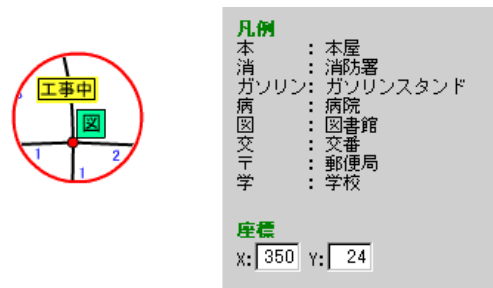


図 3. インタフェース

とユーザの間で誤解が生じうる。そのような状況においてユーザが正解経路から外れたことが判明したら、まず以下の手順でユーザの現在地を推測する。そして、その地点からの最短経路を再計算し、ユーザに再指示を行なう。

1. 直前のユーザ位置の周辺を探索
2. システムが想定していた現在地の周辺を探索
3. どこを通過したかをユーザに尋ね、直前位置・通過地点・ユーザ現在地が一直線に並ぶ箇所を網羅的に探索

以下に対話例の一部を示す。S:がシステム、U:がユーザの発話を表す。システム発話のうち下線を引いてある単語(方向・場所を表す単語)は重要であるとみなし、太字で表されている単語(対話中において初出の場所を表す語)は新しいとみなす語である。

S: 東へ行って コンビニ まで行ってください

U: コンビニに着きました

S: 左 へ曲がって 神社 まで行ってください

U: ガソリンスタンドに着きました

S: 神社 はまだ先ですね

S: まっすぐ 神社 まで行ってください

U: まっすぐ進めません

S: 道はありますか

U: 工事中です

S: では 右 に曲がって 公園 まで行ってください

一方、2.3 節で示した提案手法を用いることで、上記対話例とは異なる発話を行なうことも可能となる。以下に、上記対話例における最初の FRD の異なる発話例を示す。

S: 東 へ行ってください

U: はい

S: そして コンビニ まで行ってください

U: コンビニに着きました

上記のように単文に区切ることで FRD の数は増加するが、ユーザの知識等によってはこちらの応答生成が望ましいことも考えられる。

5 まとめ

より柔軟でより複雑な応答音声の生成のために、従来手法を改良することで新たな応答文生成手法を提案した。また、韻律制御手法に関しても新たな手法を提案し、それらを道案内音声対話システム上に構築した。

今後は、本手法の評価として、提案する応答生成手法の妥当性について検討していく。具体的には、ユーザによって応答生成の様式を変え、様々な応答に対する妥当性やシステム使用感を探る。また、韻律制御に関しても、聴取実験による評価を行なうことを考えている。

参考文献

- [1] 八木 裕司, 多胡 順司, 峯松信明, 広瀬 啓吉: “エージェント対話システムにおける対話管理と応答生成,” 情報処理学会研究報告, 2003-SLP-47-13, pp.65-70, 2003.
- [2] 八木 裕司, 広瀬 啓吉, 峯松信明: “韻律に着目した対話システムにおける応答生成の改良,” 日本音響学会 2004 年春季研究発表会講演論文集, vol.1, 3-8-7, pp.135-136, 2004.
- [3] 桐山 伸也, 広瀬 啓吉: “応答生成に着目した学術文献検索音声対話システムの構築とその評価,” 電子情報通信学会論文誌, D-II, vol.J83-D-II, no.11, pp.2318-2329, 2000.
- [4] N. Minematsu, R. Kita and K. Hirose: “Automatic estimation of accentual attribute values of words to realize accent sandhi in Japanese text-to-speech conversion,” Proc. IEEE 2002 Workshop on Speech Synthesis, Santa Monica, 2002.
- [5] K. Hirose, M. Sakata and H. Kawanami: “Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features,” Proc. ICSLP96, pp.378-381, 1996.
- [6] 河合 剛, 広瀬 啓吉, 藤崎 博也: “日本語文章音声合成のための韻律規則,” 日本音響学会誌, vol.50, no.6, pp.432-442, 1994.
- [7] 匂坂 芳典, 佐藤 大和: “日本語単語連鎖のアクセント規則,” 電子情報通信学会論文誌, vol.J66-D, no.7, pp.849-856, 1983.
- [8] 広瀬 啓吉, 藤崎 博也: “音声合成とアクセント・イントネーション,” 電子情報通信学会誌, vol.70, no.4, pp.378-385, 1987.
- [9] 河原 達也, 住吉 貴志, 李 晃伸, 坂野 秀樹, 武田 一哉, 三村 正人, 山田 武志, 西浦 敬信, 伊藤 克亘, 伊藤 彰則, 鹿野 清宏: “連続音声認識コンソーシアム 2001 年度版ソフトウェアの概要,” 情報処理学会研究報告, 2002-SLP-43-3, pp.13-18, 2002.