対話音声における韻律予測のために

山下洋一　菅原孝夫　溝口理一郎
{yama,sugahara,miz}@ei.sanken.osaka-u.ac.jp

大阪大学 産業科学研究所
〒567 茨木市美穂ケ丘 8-1

自然な対話音声を合成するためには、先行発話との関係など対話に関する情報に基づいて韻律を決定する必要がある。しかしながら、対話の特徴は非常に多岐に広がり、対話のタスクや領域などによっても異なってくるため、様々な対話に共通した特徴を捉えるだけでは必ずしも十分ではない。本報告では、スケジュール調整を行う対話における固有名詞に注目し、アクセントごとの最大基本周波数を決定する規則の生成と評価を行った結果を示す。対話情報として7種の属性を用いることによって基本周波数の予測誤差を大きく減少させることができた。
キーワード：対話音声合成、韻律、対話コンテキスト、統計的手法

To Predict Prosodic Parameters in Spoken Dialogues

Yoichi Yamashita　Takao Sugahara　Riichiro Mizoguchi

I.S.I.R., Osaka University
8-1, Mihogaoka, Ibaraki, Osaka, 567 Japan

Dialogue features such as the relationship to the preceding utterance should be incorporated into the prosody prediction for synthesis of natural spoken dialogues. Characteristics of dialogues spread over a lot of aspects and are dependent on the task and the domain of the dialogue. It is not sufficient to model global characteristics of dialogues in order to predict prosody in a specific task. This paper describes generation and evaluation of rules which predict F0 maximum of accentual phrases, directing a special attention to the proper noun phrases in the schedule arrangement task. Seven features of the dialogue reduce the prediction errors of the F0 parameter.
keywords: spoken dialogue synthesis, prospdy, dialogue context, stochastic method

1　はじめに

自然な対話音声を合成するには、発話文の持つ統語的情報だけでなく対話に関する高次情報(対話情報)に基づいて基本周波数などの韻律パラメータを決定する必要がある。

対話情報を用いた韻律決定の研究としては、これまでに英文におけるピッチアクセントの予測に関する研究 [1, 2] が数多くなされている。これらの研究ではピッチアクセントの有無の決定が定性的に議論されており、日本語の対話音声合成における基本周波数の生成などには必ずしも十分ではない。日本語に対する研究では、藤崎らが先行する質問文を意図的に変えることにより回答発話中でのアクセント成分の変化を定量的に分析している [3]。この他にも孤立発話と文脈内での発話における基本周波数の違い [4] や、韻律パラメータと対話情報の関係の分析

[5] なども報告されているが、これらの研究ではモノローグの発話データが用いられている。実際の対話音声では、人工的に作った文脈やモノローグよりも韻律パラメータが多様に変化するため、対話音声データを対象とした研究が必要となる [6]。

本報告では、スケジューリング調整をタスクとした模擬対話データを用いて、アクセント句ごとの最大基本周波数を予測する規則の生成と評価について述べる。

2 手法

2.1 韻律の二段階予測

本手法では、対話音声中の韻律パラメータは、“S ルール (基本規則)” および “D ルール (対話規則)” の二種類の規則を順に適用することによって予測される [7]。対話音声の韻律に影響を与える要因は、品詞、係り受けやアクセント型などの文自体の基本的な素性と、先行発話との関係や焦点などの高次情報に大きく分けることができると考えられる。前者は、主に統語的な要因であり、対話音声に限らず朗読などの読み上げ音声の韻律にも大きく影響を与えるものである。一方後者は、対話中の発話が持つ要因で、対話との関連性が大きいことから、これを対話情報と呼ぶ。まず S ルールが前者の要因から基本的な韻律パラメータ値を決定し、次に D ルールが後者の要因に基づいて韻律パラメータを修正する。

これらの規則は、以下のような手順で二種類の音声データから生成される。

step1: 孤立発声された音声データから S ルールを生成する。
step2: S ルール を対話音声データに適用する。
step3: step2 の結果生じた誤差データに対話情報を追加し、D ルールを生成する。

2.2 韻律モデル

今回対象とした韻律パラメータはアクセント句における最大基本周波数である。このパラメータは、阿部らによって提案されている基本周波数に対する 2 階層制御方式において、グローバルモデルとして記述される韻律パラメータである [8]。また、S ルールとしても阿部らのグローバルモデルをそのまま用いた。

2.3 音声データ

二段階予測によって韻律パラメータを予測するには、二種類の音声データが必要になる。孤立発声データとしては男性話者一名が発声した ATR503 文を用い、対話音声データとしてはスケジュール調整を行う模擬対話を行い、問い合わせに答える側の発話の中で “人名”、“曜日”、“時刻” を含むアクセントだけを用いた。

模擬対話では、会議などを行うため数名の参加者の共通した空き時間を見つけることを対話のゴールとして設定した。回答者は、質問者以外の参加者の予定を全て把握しており、質問者と協調的に問題解決を図った。収録した対話の例を図 1 に示す。用いた対話数は 21 である。また、対話音声の話者は 7 名で、孤立発声データの話者とは異なる。

U001: あ、もしもし、田中ですけれども。
S001: はい。
U002: [えーと] 交通委員会の予定を決めたいので|はい|、[えーと] 山本先生、鈴木先生、山田先生との |はい| 時間の調整をしたいんですけれども。
S002: はい。
U003: えーとですね。私、月曜日出張なもんですから |はい|。[えーと] 火曜日はどなたか都合の悪い人いらっしゃいますかね。
S003: [えーと] 山田先生が、１０時から１２時まで、講義がありまして |ええ| １時から３時まで会議があるんですけど。
U004: あーそうですか。じゃ、水曜日はどうですか。
S004: 水曜日は、山田先生が１０時から１２時まで講義がありまして、昼からは３人とも空いてます。
U005: [えーと] 昼から皆さん予定がつまって [あの] (よて) |いや、空いてます|。空いてる。
S005: はい。
U006: ...

図 1: スケジュール調整対話の例

3 結果

3.1 S ルールの生成

まず、孤立発声データから S ルールを作成した。阿部らと同様に、7 個の属性 (質的説明要因) から数量化 I 類によってアクセント句の最大基本周波数を予測した。用いた質的説明要因も阿部らのものと同じもので、当該アクセント句の先行／後続境界、音節数、アクセント型、自立語の品詞、および先行／後続するアクセント句のアクセント型の 7 個である。

ここで、参考までに S ルールの評価を行った。対話コンテキストのない孤立発声データに S ルールを適用すると、平均誤差は 10.0Hz であった。

次にこの S ルールを対話コンテキストの影響を受けている対話データに適用した。ただし、孤立発声データと対話音声データの話者が異なるため、適用の際には話者の平均基本周波数の差を補正した。適用後の平均誤差は、表1のようになった。

表 1 に示したように、対話データに S ルールを適用した時の誤差は、話者によって若干の差はあるものの、先に示した対話コンテキストのない孤立発声データに適用した時の平均誤差 (10.0Hz) よりもかなり大きな値となっている。これが主に対話によって引き起こされたと考える。

表 1: 対話データへの S ルールの適用結果

話者	平均誤差 [Hz]		
	人名	曜日	時刻
KAW	6.9	10.5	9.8
KAT	14.9	28.4	21.7
UED	26.0	13.6	16.8
JUN	25.1	22.7	14.7
SET	15.5	24.9	14.3
NIS	19.8	14.6	15.9
SIM	17.2	16.7	14.8
平均	18.8	18.8	15.6

3.2 対話情報

“人名” および “曜日” に関する対話データに S ルールを適用した結果を定性的に分析することによって、D ルールで用いる対話コンテキストを反映する対話情報として以下の 7 個の属性を同定し、誤差データに与えた。[] 内は、それぞれの対話情報 (属性) が取り得る属性値を示している。各アクセント句に対する属性値は人手で決定した。

DF1: 前発話との関係 [話題の変更, 話題の復帰, 継続, 詳細化, 肯定の補足, 否定の補足, 説明要求の回答, W 質問の回答]

そのアクセント句を含む発話が、先行発話とどのような関係にあるかを分類した。

DF2: 同種の語に後続しているか [yes, no]

「山田先生と鈴木先生」における鈴木先生のように、先行するアクセント句の自立語が同じ種類の語 (人名なら人名、曜日なら曜日) の場合は yes、そうでない場合には no とする。この例における「山田先生と」では、前に人名のアクセント句がないため、no となる。

DF3: 句が前置されているか [yes, no]

そのアクセント句が発話内で前置されている場合には yes、そうでない場合には no とする。

DF4: 先行語 [単語, 不要語, ポーズ, 言い淀み, 文頭]

アクセント句に先行するカテゴリーを 5 種類に分類する。

DF5: 省略可能か [yes, no]

「山田先生は火曜日は空いてますか」に対して「火曜日は空いてます」と回答する発話における「火曜日」のように、省略しても意味が正しく伝わるような場合には yes、そうでない場合には no とする。

DF6: 対立する語 (概念) が示されているか [yes, no]

「私は 3 時から空いてるんですが、他の先生の都合はどうですか」に対する「鈴木先生が 3 時から 5 時まで講義です」のように、先に提示されたスケジュールと合わな

いような情報を提示する場合には yes、そうでない場合には no とする。

DF7: 先に同様の概念を提示しているか [yes, no]

「鈴木先生は 10 時から来客があります」に続けて「また、山田先生は午後から講義があります」を発話するように、いくつかの情報を順に提示する場合に、二つ目以降では yes、それ以外では no とする。

3.3 S ルールの評価

7 名の話者の誤差データに対話情報を与え、再び数量化 I 類でモデル化し、"人名"、"曜日" および "時刻" のデータに対してそれぞれ D ルールを得た。表 2に D ルールの評価結果を示す。なお、オープンな評価は 10 分割の Cross Validation によって行なった。D ルールは、人名、曜日のデータに対しては約 25% 程度誤差を減少させているが、時刻データに対しては誤差の現象は 15% 程度でやや効果が小さい。人名、曜日、時刻のデータに対する D ルールの重相関係数は、それぞれ 0.73, 0.66, 0.57 であった。時刻データに対してはかなり低い値となり、十分なモデル化ができていない。

表 2: D ルールの評価

(a) 人名

評価対象	D ルール適用前 [Hz]	適用後 [Hz]	
		クローズ	オープン
ALL	18.8	11.7	13.7
E20	33.9	15.9	18.6

(b) 曜日

評価対象	D ルール適用前 [Hz]	適用後 [Hz]	
		クローズ	オープン
ALL	18.8	12.5	14.1
E20	34.4	15.8	17.7

(c) 時刻

評価対象	D ルール適用前 [Hz]	適用後 [Hz]	
		クローズ	オープン
ALL	15.6	12.9	13.5
E20	30.5	20.4	21.3

(ALL): 全誤差データ
(E20): S ルールによる誤差が 20Hz 以上あったデータ

ここで、全ての発話、あるいは全てのクセント成分 (文節) が必ずしも対話コンテキストの影響を受けるわけではない。対話コンテキストの影響を大きく受けているところでは、S ルールを適用した時の誤差が大きいと考えられる。このため、誤差が 20Hz 以上あったデータのみを用いて D ルールを評価した。E20 はこの結果を示しており、どのデータにおいても大きく誤差が減少しているが、時刻データでは人名、曜日と比べるとやはり効果が小さい。なお、E20 に含まれる例題数は人名データが 60 個 (総数 156 個)、曜日データが 63 個 (総数 169 個)、時刻データが 121 個 (総数 390 個) である。

4 考察

4.1 対象データの制限

対話における韻律パラメータの変化は非常に多様であるため、全てのアクセント句をうまくモデル化する規則を得ることは容易ではない。著者らが先に行った研究 [7] では、観光案内をタスクとした対話において全てのアクセント句を対象として同様の手法でモデル化することを試みたが、重相関係数が 0.4 ～ 0.6 と非常に小さく、限られた対話情報ではうまくモデル化することができなかった。

今回の実験では、対象とするアクセント句を選択することによって (時刻データを除いて) うまくモデル化することができた。このように対象を選択することによって、対話から受ける影響が限定されることが期待できる。

4.2 対象データ間の比較

図 2は、D ルールにおける数量化 I 類のモデルパラメータの値 (属性値の係数) を示している。人名、曜日、時刻に対する規則を比較してみると、全体としては良く似た傾向を示しているが、一部傾向の違いが見られる。

・人名データに対する DF1 の「話題の変更／詳細化」の値が小さい。

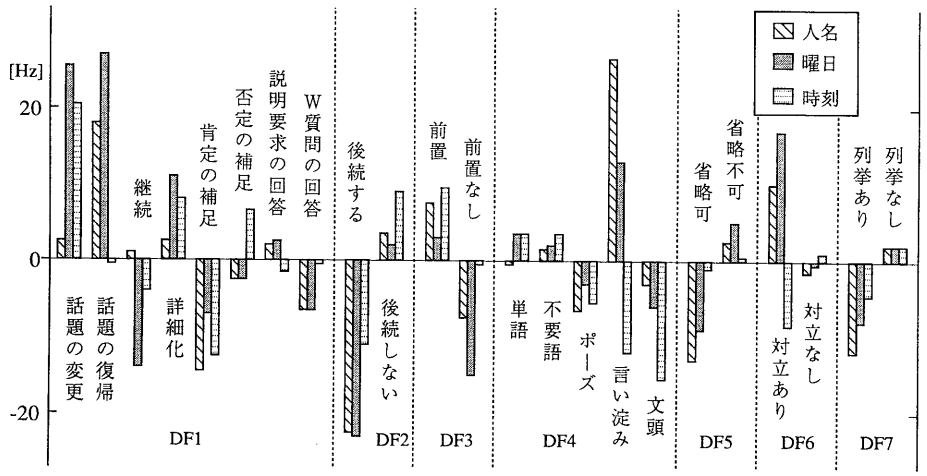

図 2: 対話情報に対する重み係数

今回用いたスケジューリングのタスクでは、「会議などの参加者の共通の空き時間を見つけること」が目的であるため、参加者 (人名) を固定し時間 (曜日, 時刻) を変えて相談することが主に対話で行われた。このため、前発話に対して話題が変更／詳細化される場合には、人名よりもむしろ曜日や時刻が問題とされたため、曜日や時刻に比べて人名では、これらの値が小さくなったと考えられる。

・曜日データに対する DF1 の「継続」の値が小さい。

「○曜日の○時」などのパタンが多く「○時」に意味の中心があるため、「○曜日」の基本周波数が相対的に抑えられた。

・時刻データに対する DF4 の「言い淀み」の値が小さい。

今回用いた時刻データでは、相手の相づちやポーズをはさんで句を繰り返すような場合があり、「言い淀み」となる例は数が少ないこともあり、これらの例題の DF4 の値を「言い淀み」とした。これまでの研究でも、言い淀みの後では基本周波数が増加することが指摘されている [9][10] が、このような値を割り付け方から、時間データに対しては異なった傾向となったと思われる。

・時刻データに対する DF6 の「対立あり」の値が小さい。

スケジュールが合わない場合には、時刻よりもむしろ「○先生がだめです」「○曜日はだめです」ということを主として伝えようとする場合が多く、時刻はそれほど強調されなかった。

4.3 二段階予測の有効性

二段階予測の有効性を確認するために、人名、曜日データに対して S ルールおよび D ルールで用いた属性を一度に同時に用いて、対話におけるアクセント句の最大基本周波数を予測する一段階のモデル化を行った。ただし、全ての例題に対して同じ値をとる属性は用いていないため、人名、曜日データに対して用いた属性数は

表 3: 一段階予測規則の評価

(a) 人名

評価対象	予測誤差 [Hz]	
	クローズ	オープン
ALL	11.7	14.6
E20	14.0	17.4

(b) 曜日

評価対象	予測誤差 [Hz]	
	クローズ	オープン
ALL	13.9	17.3
E20	17.1	21.2

それぞれ 11 個、12 個である。結果を表 3に示す。ALL、E20 の意味は表 2と同様である。

表2と表3を比較すると、一段階予測はクローズ評価では一部優っているものの、オープン評価では予測誤差が増加する傾向が見られ、クローズとオープンの評価の差が大きくなっている。このことから、二段階予測の方が学習例題への依存性の低い規則が生成されたと言える。

4.4 今後に残された課題

(1) モデル化に用いる対話情報の同定

本報告で用いた 7 個の対話情報は、主に "人名"、"曜日" のデータにおける S ルール適用誤差の分析をもとにして決定された。このため、"時刻" データに対しては対話による影響を必ずしも十分に説明できていない。対象を限定してモデル化を行う場合には、用いる対話情報を対象毎に洗練する必要がある。

(2) 対話情報に対する値の割り付け

用いる対話情報によって、例題にどの値を割り付けるかという tagging の問題は (特に、前発話との関係などに関して)、値の分類と共に非常に難しい問題である。著者らの実験においては数名の研究者で議論の結果決定していったが、意見の別れる例もあり、ガイドラインのようなものを作成することによって「ゆれ」をおさえることが必要である。

(3) 対話時における対話情報の獲得

本報告では、対話音声合成のための韻律予測において、対話情報を利用するための手法とその効果に研究の主眼が置かれているため、対話を進めている状況での対話情報の獲得に関しては全く言及していない。実際の対話音声合成では、対話を通じての対話情報を獲得が非常に重要な問題であり、対話のモデル化、対話管理と合わせて今後の大きな課題である。

5 おわりに

対話音声合成のために、スケジューリング調整タスクにおける人名、曜日、時刻に関するアクセント句を対象として、最大基本周波数を予測する規則の生成を行った。対話における現象は非常に多岐にわたり、それぞれの現象も多様であることから、対象を限定し知見を積み重ねていくことが重要であるように思われる。

謝辞 音声データ収集、実験に協力してくれた大阪大学大学院生 (現シャープ (株)) 作田瑞 君に感謝します。本研究では、日本音響学会の研究用連続音声データベースの一部を使用した。また、本研究の一部は、文部省科研費 (重点領域研究「音声対話」, No.05241105) の援助を受けた。

参考文献

[1] J. Hirschberg, "Using Discourse Context to Guide Pitch Accent Decisions in Synthetic Speech," Proc. of ESCA Workshop on Speech Synthesis, Autrans, 181-184 (1990).

[2] A.I.C. Monaghan, "Intonation Accent Placement in a Concept-to-Dialogue System," Proc. of 2nd ESCA/IEEE Workshop on Speech Synthesis, New York, 171-174 (1994).

[3] 藤崎博也, 広瀬啓吉, 高橋登, 横尾真, "連続音声中におけるアクセント成分の実現," 日本音響学会音声研究会資料 S84-36, 279-286 (1984).

[4] J.M. Garrido, J. Listerri, C. de la Mota and A. Rios, "Prosodic Differencies in Reading Style:Isolated vs. Contextualized Sentences," Proc. of Eurospeech '93, 573-576 (1993).

[5] J. Hirshberg and B. Grosz, "Intonation Features of Local and Global Discourse Structure," Proc. of DARPA Speech and Natural Language Workshop, 441-446 (1992).

[6] 阪田真弓, 広瀬啓吉, "対話音声の韻律的持徴の分析と合成," 信学技報 **95**, SP95-17, 55-62 (1995).

[7] 宮原進, 山下洋一, 溝口理一郎, "対話情報に基づいた韻律生成," 信学技報 **94**, SP94-76, 49-56 (1994).

[8] 阿部匡伸, 佐藤大和, "音節区分化モデルに基づく基本周波数の 2 階層制御方式," 音響学会誌 **49**, 682-690 (1993).

[9] D. O'Shaughnessy, "Analysis and Automatic Recognition of False Starts in Spontaneous Speech," Proc. of ICASSP '93, **II**, 724-727 (1993).

[10] C.H. Nakatani and J. Hirschberg, "A Corpus-Based Study of Repair Cues in Spontaneous Speech," J. Acoust. Soc. Am. **95**, 1603-1616 (1994).