対話における韻律変化・タイミングのモデル化と音声対話システムへの適用

Response Timing and Prosody Change Modeling in Conversations and Their Application to a Spoken Dialog System

西村良太† 北岡教英‡ 中川聖一†

Ryota NISHIMURA† Norihide KITAOKA‡ Seiichi NAKAGAWA†

†豊橋技術科学大学情報工学系 ‡名古屋大学大学院情報科学研究科

†Department of Information and Computer Sciences, Toyohashi University of Technology,

‡Graduate School of Information Science, Nagoya University

†{nishimura, nakagawa }@slp.ics.tut.ac.jp, ‡kitaoka@sp.m.is.nagoya-u.ac.jp

Abstract

If a dialog system can respond to a user as natural as a human, the interaction will be smoother. It is important to generate natural timing of responses to user's utterances to realize a natural conversation like human to human. And, it is also important for the system to control the prosodic information to generate response speech. In this paper, to develop the cooperative speech dialog system, we analyzed the correlation between prosodic synchronous and "warm up", "familiarity" or "informality" in human-human dialogs. And we try to model these features. And then, we installed the model to the system. The dialog system could generate the timing of *aizuchi*, turn-taking and response timing based on a decision tree in real time, to realize chat-like conversation. The system also uses the prosody controll model to controll the prosodic change for output speech.

1 はじめに

近年，計算機の性能向上と音声認識技術の発達に伴い，音声認識技術が多くの場面で使われるようになってきた．そして，音声認識技術を用いたインターフェースも発展してきている．近年では，観光案内や情報検索，カーナビゲーションシステムへの応用など，様々な対話システムが検討・実用化されている．しかし，一般にこれらのシステムにおいては，ユーザがシステムに話しかけた際に，途中にシステムからの反応が全くなく，ユーザ発話終了後にシステムが応答を返すまでユーザ発話を聞いているのか分からないといった問題があり，これが音声認識を利用した音声対話システムに壁を感じる一因となっている．今後，音声対話システムがより身近なものになり，生活の中に入り込んでくることが予想されるが，その際には，より自然な対話を実現する必要がある．人間同士の雑談のような対話は，自然な対話のひとつの理想の形であると考えられる．我々はこのような対話に現れる様々な現象をシステムで実現し，自然な対話システムを構築することを目指している．

例えば，人間同士の会話においては，話者は互いにうなづきやあいづちによって相手の発話を理解していることを明示しており，それにより会話がスムーズに進行する．これを音声対話システムにも応用し，ユーザの発話に対して，システムがあいづちなどの反応を返す事が出来ればユーザは人間と対話している場合と同じように自然に対話が行えるのではないかと考えられる．また，話者交替のタイミングも，対話の自然性を考える上では重要である．相手の発話が終わったのか，まだ続くのかを的確に把握し，適切なタイミングで応答を返すことができれば，円滑に対話を進めることが可能になる．

このように，あいづちや話者交替を行う場合には，相手の発話に応じて適切なタイミングで応答を返し，時にはそれらをオーバーラップさせることによって，スムー

ズに会話が進行していく．そこで我々は，音声対話システムにおいてあいづちや，システムからユーザへの割り込み発話など，種々の現象を考慮しそれらを適切なタイミングで行う天気予報を話題とする雑談システムを構築した [1].

文献 [1] では，円滑に対話を実現するために，タイミングのみに重点を置いていた．しかし，実際の人間同士の対話においては，対話が進むにつれて，韻律が同調して進んでいる．垣田 [6] は，簡単な質問応答形式にて話者の基本周波数に関して話者間で関係があるかを実験により調査しており，ほとんどの話者で，一方の話者の基本周波数が高ければ，もう一方の話者の基本周波数も高くなることを指摘している．また，長岡ら [7] は，交替潜時 (Switching pause) が，2 話者間で有意な正の相関を示したと報告している．これらのことから，人間同士の対話の韻律情報には，何らかの関係がある事が予想される．人間同士の対話のように，より円滑に対話を進める為には，音声対話システムにおいても，韻律情報を制御することが必要である．

そこで，本研究では，人間同士の対話音声から，対話者間でどのように韻律情報に相互作用があるかを分析する．そこから，どのような動きがあるかを見出し，モデル化を行う．このモデルを，対話を円滑に進めるための韻律情報制御モデルとして，音声対話システムに組み込んだ．この際，人間同士の対話音声からの情報が，対話を進める上での理想的なものであると仮定し，各種特徴を調べた．

2 人間同士の対話における話者間の韻律の関係

2.1 対話コーパス

応答タイミング・韻律変化のモデル化に際して，人間同士の対話を調査した．調査に用いたコーパスは，国立国語研究所から提供されている「日本語話し言葉コーパス」(Corpus of Spontaneous Japanese;CSJ) である [2]. CSJ コーパスは，現代日本語の自発音声を種々の研究用付加情報とともに大量に格納したデータベースであり，語数にして 750 万語，時間にして 660 時間の音声が含まれている．CSJ コーパスは，自動音声認識の研究リソースとして活用することが想定されているため，ほとんどが学会講演のような独話（モノローグ）音声であるが，対話音声としては，表 1 に示す内容・時間数が存在する．

今回の対話コーパスの調査は，これらの対話音声に対して行った．対話は，１つ 10 分〜20 分程の長さであり，全体で 58 対話，12 時間である．

表 1: CSJ の対話コーパス

種類	内容	ファイル数	時間
D01	模擬講演インタビュー	16	3.4
D02	課題指向対話	16	3.1
D03	自由対話	16	3.6
D04	学会講演インタビュー	10	2.1

模擬講演インタビューと学会講演インタビューは，16 名により事前に行われた学会講演ないし模擬講演（10 名は両方，6 名は模擬講演のみ）に関してインタビュワーが様々な質問を発し，講演者がこれに答える形式の対話である．発話の大半は，質問に対する回答によって占められている．インタビュワーは 20 代と 30 代の女性各 1 名のいずれかが勤めている．

課題指向対話では，インタビューとの対比のため，参加者 2 名（上記インタビューと同一ペア）の発話量が等しくなりやすい課題が選定されている．具体的には，実在の芸能人に講演を依頼した場合の謝礼（ギャラ）の額を想像し，その額が高い順に，芸能人 9 ないし 10 名をソートするタスク（ギャラ・タスク）である．対話開始時点で各話者に手渡されている人名リストは，わざと一致しないように作成してあるので，謝礼額の推定に先立って（あるいは同時に），推定対象となる芸能人の完全なリストを作成するための対話も必要とされる．

自由対話では，課題の制約なしに，10 分程度，自由に対話を行っている．これらの 4 種類の対話音声は，同一の話者ペア（講演者とインタビュアー）によって発話されている．

全対話には女性対女性，女性対男性の対話があり，男性対男性対話はない．いずれも女性（インタビュアー）が，男性または女性と対話をするものである．音声収録には，対面ブースを用いており，話者はそれぞれ独立した防音ブースに入るが，ガラス越しにお互いの顔が見えるようになっている．お互いの音声は，イヤホンを通して聞くようになっている．

2.2 対話中の 2 話者の基本周波数 (F0) の変動

コーパス中の実際の対話の音声の基本周波数（F0）をプロットしたものを図 1 に示す．線の違いは話者の違いを示している．値は対数値 (log F0) で，各話者の全体の平均値を揃えてある．

この図中の対話の話題は，前半が「子供に将来の夢を聞いたら “子猫ちゃん” だった」というもので，後半が「子供の言語獲得について」である．前半は，面白いエピソードであるため，二人から「笑い」も起きており，

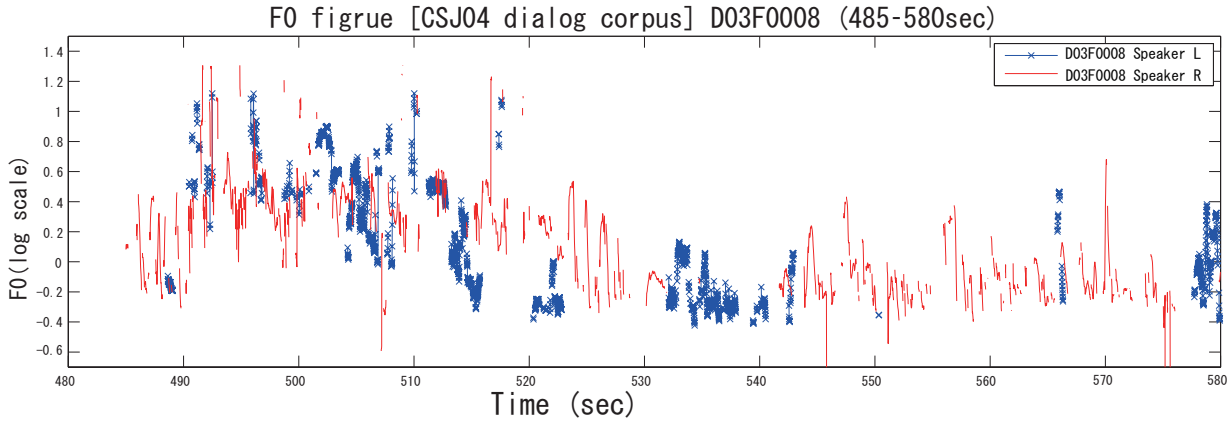

図 1: CSJ 対話音声の例

双方がよく発言し，盛り上がっている．後半は，まじめな話に入っていき，二人とも落ち着いて話している．

図を見ると，盛り上がっているところでは，F0 が高くなっており，ダイナミックレンジも大きくなっている．それに対して，落ち着いている部分では声の高さも，それぞれの平均値からあまり変化しておらず，ダイナミックレンジも小さい．

このように，対話のはずみ度合い，盛り上がりと，基本周波数をはじめとする韻律には強い関係があることが予想される．それは 2 話者が互いに影響しあって強調的に変動しているものだと考えられる．

2.3 対話中の 2 話者の基本周波数の相関

対話をしている 2 人の基本周波数の相関をこれらのコーパスで調べた．

対話中の基本周波数は，各発話中の基本周波数を発話時間で割った平均値を代表値として，1 つの発話ごとに 1 つの値で表した．また，その点は，発話開始時刻と発話終了時刻の間（中央）にとった．2 話者の基本周波数の相関を求める際には，データ数を揃えなければならないので，一方の話者の各発話の代表点の時刻に対して，もう一方の話者の二つの代表点から補間値を求めて，データ数を揃えた．

表 2: 各対話の F0 の相関

種類	最大値	最小値	平均
D01	0.382	0.070	0.195
D02	0.477	-0.002	0.222
D03	0.521	0.012	0.234
D04	0.206	-0.005	0.085
平均	0.397	0.019	0.184

2 話者間の対数基本周波数 logF0 の相関値は表 2 のようになった．一般に相関値は大きく，対話中での 2 話者のお互いの F0 値には関連があると言える．また，58 対話中 4 対話を除いて正の相関を示しており，対話において，声の高さは相手に合わせて変化していくと考えられる．そして，対話の内容によって，相関値に違いが見られた．比較的自由な形式の対話（D2，D3）は，インタビュー形式のもの（D1，D4）に比べて相関が高い．つまり，自由な形式の対話では基本周波数が同調する傾向が高いということを示している．また，性別による違いもあり，相関値が高いものには，女性同士の対話が多かった．一方，女性と男性の対話は，相関値の低いものが多かった．

実際に音声を聞いて比べてみると，相関が高い対話は，盛り上がる場所や落ち着く場所が良く一致している．あいづちも，相手の声の調子に応じてうまく変化させており，それによって対話がスムーズになっている様子が感じとれる．

3 対話の印象と対話現象との関係

CSJ コーパスの対話音声を実際に人間が聞いた場合の各対話の印象と，コーパス中の現象（2 話者の logF0 相関値，オーバーラップ頻度，フィラー頻度）との関係を調べた．4 名に対話音声を聞いてもらい，各対話について，以下の各項目について 5 段階のアンケートをとった．

- 相手との親しさ
  - 親しみを持っている (5 4 3 2 1) 遠慮している
- 盛り上がり
  - 良い (5 4 3 2 1) 盛り上がってない
- 相手の意見に同意しながら対話
  - 同意しながら対話 (5 4 3 2 1) 意見を戦わせながら対話
- 立場の違い
  - 目上 (5 4 3 2 1) 目下
- L 話者 (インタビュアー) のフランクさ
  - 気を使っている (5 4 3 2 1) くだけている
- R 話者 (インタビュイー) のフランクさ
  - 気を使っている (5 4 3 2 1) くだけている
- L 話者 (インタビュアー) の表現
  - 敬語ばかり (5 4 3 2 1) 敬語をつかっていない

- R 話者 (インタビュイー) の表現
  - 敬語ばかり (5 4 3 2 1) 敬語を使っていない

アンケートは，この形式のアンケート用紙に丸を付けて回答する形式にした．

「相手との親しさ」,「盛り上がり」に関しては，対話中の 2 話者からの印象の平均を考えて評価値を付けるようにした.「立場の違い」については，インタビュアー（L 話者）側から見て相手が目上か目下かというところを評価して値を付けてもらった．また,「フランクさ」は言い方の違いを示すものであり，声の調子などから受ける印象を評価してもらった.「表現」は文字上でどのようになっているかを示しており，語彙的に敬語を用いているかどうかを評価してもらった．

被験者は 4 名（男性 1 名，女性 3 名）であり，実験に際して,「正式な聞き取り実験の前に，10 分間程度評価サンプルからランダムに対話サンプルを聞いて雰囲気をつかむこと」と「各対話について，解答する際に全体をかいつまんで 5 分以上は聞くこと」を注意事項として伝えた．

ここで，アンケート結果の被験者間での違いを見る為に，被験者間での結果の相関を調べた.「盛り上がり」に対するアンケート結果の相関値は表 3 のようになった．平均値は，0.470 であった．

表 3: アンケート被験者同士の「盛り上がり」の相関

被験者	相関値
被験者 1 - 被験者 2	0.523
被験者 1 - 被験者 3	0.538
被験者 1 - 被験者 4	0.630
被験者 2 - 被験者 3	0.314
被験者 2 - 被験者 4	0.342
被験者 3 - 被験者 4	0.476
平均	0.470

この結果から,「盛り上がり」に関しては各被験者間で相関があることが分かる．このことから，各対話の盛り上がりの指標として，このアンケート結果を用いることとする．また，その他のアンケート結果については，表 4 のようになった.「話者の表現」への回答が比較的バラツキがあったようである．

アンケート結果の評価値と，各対話音声の情報「対話中のオーバーラップ頻度」,「対話中の 2 人の F0 相関値」,「対話中のフィラー頻度」との相関を表 5 に示す．ここで,「対話中のオーバーラップ頻度」とは，対話中に話者交替がオーバーラップして起こった回数を，対話中の話者交替の総数で割ったものである ((1) 式)．また,「対話中のフィラー頻度」は，対話中にフィラーが発話された回数を，全発話数で割ったものである ((2) 式)．

表 4: 各アンケート項目の被験者間の相関の平均値

アンケート項目	相関値の平均
親しさ	0.444
盛り上がり	0.470
同意・否定	0.387
立場	0.478
L 話者のフランクさ	0.399
R 話者のフランクさ	0.384
L 話者の表現	0.300
R 話者の表現	0.262

$$\text{対話中のオーバーラップ頻度} = \frac{\text{オーバーラップした話者交替数}}{\text{対話中の話者交替の総数}} \quad (1)$$

$$\text{対話中のフィラー頻度} = \frac{\text{フィラーの発話数}}{\text{対話中の発話の総数}} \quad (2)$$

表 5: 被験者評価値と対話現象との相関

	F0 相関値	オーバーラップ頻度	フィラー頻度
親しさ	0.348	0.627	0.072
盛り上がり	0.350	0.718	0.127
同意･否定	0.279	0.638	0.090
立場	-0.282	-0.098	0.267
L 話者フランクさ	-0.283	-0.417	0.108
R 話者フランクさ	-0.266	-0.637	0.047
L 話者の表現	-0.340	-0.238	0.265
R 話者の表現	-0.182	-0.404	0.289

表 5 を見てみると，オーバーラップ頻度は，全体的に高い相関値を示している.「親しさ」「盛り上がり」「同意・否定」は，オーバーラップの頻度が高いと，評価値も高くなっている．つまり，オーバーラップがたくさん起こっている対話では，親しさがあったり，盛り上がっていたり，同意して対話が進んでいるということである.「フランクさ」に対しては，負の相関が高いので，くだけていると感じた対話にて，より多くオーバーラップが起こったということである．

F0 相関値に関しては,「親しさ」「盛り上がり」「同意・否定」と正の相関関係がある．つまり，基本周波数が 2 話者間で同調して進んでいる対話は，親しさがあったり，盛り上がっていたり，同意して対話が進んでいるということである．また，F0 相関値が高いとくだけた対話の印象になると言える．

フィラー頻度に関しては,「立場」と「表現」にて正の相関がある．これは，相手が目上であったり，敬語を使っていたりする場合には，フィラーが起こりやすいことを示している．

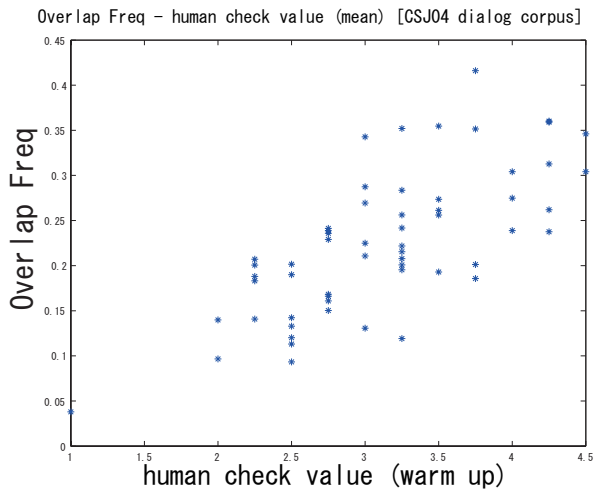

図 2: オーバーラップ頻度と被験者評価値 (盛り上がり)

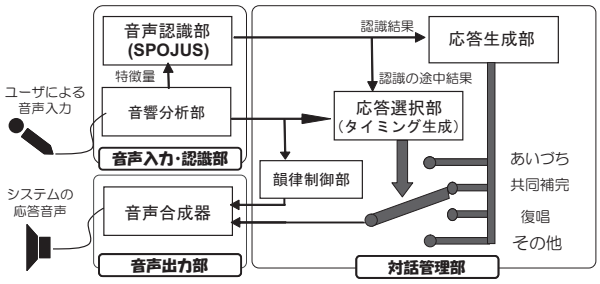

図 3: システムの構成図

図 2 は，オーバーラップ頻度と被験者の評価値（盛り上がり）の関係をプロットしたものである．この図を見ても，オーバーラップ頻度と盛り上がりとの間に，相関があることが分かる．

4 対話システム

4.1 対話システムの構成

3 節より，話者の韻律（基本周波数）やオーバラップ発話は対話のスムーズさや盛り上がりと強い関係があることが分かった．そこで，韻律の変動やオーバラップといった現象を対話システムが実現するための仕組みを考察し，実装して構築した我々の音声対話システムの概略を説明する．

システムの構成は，図 3 に示すように，大きく分けて 3 つの部分からなっている．音声入力・認識部は，入力された音声を解析し，音声認識とピッチ（基本周波数）・パワーの計算をする．対話管理部は，認識結果と韻律情報から応答文と応答タイミングを生成する．音声出力部は，応答文を音声にて出力する．この流れに沿って，リアルタイムに音声認識と韻律解析を行う．ポーズの検出をしてから応答をする従来のシステムと違い，ポーズを検出する必要がないため，応答の遅延の問題を解消している．また，これにより，システムはユーザ発話への割り込み（オーバーラップ）応答が可能になっている．また，韻律の制御も実装した．図中の韻律制御部は，人間（ユーザ）の韻律の変動にしたがって，システムの韻律を制御して出力することができる．すなわち，3 節で分かったような対話者間での同調が実現可能となっている．

4.2 応答タイミングのモデル化

応答のタイミングは，決定木を用いることでモデル化している．オーバーラップをしない部分に関しては，機械学習によって学習した決定木 [8] を用いている．ただし，オーバーラップをする部分に関しては，十分な学習データが用意できていないため，学習することが出来ない．したがって，先行研究 [3–5] を元に人手で作成した決定木を用いている．

決定木の学習には，人工知能学会が提供している「談話タグつき音声対話コーパス」の全 28 対話から 11 対話を用い，それらの書き起こし結果に，学習に必要なタグを付けたデータを用いている．機械学習には，C4.5 を用いている．決定木に用いている素性は以下のものである．決定木学習に用いた素性の数は，19 種類である [8].

- 直前のユーザ発話開始から現在までの時間
- ポーズ長
- 前のシステム発話終了時刻からの時間
- ピッチ・パワーの 100ms 区間の傾き（3 つずつ）
- ピッチ・パワーの 500ms 区間の傾き（5 つずつ）

決定木の出力として,「あいづち」「話者交替 (システム応答)」「話者継続（待ち）」の 3 クラスを用意した．学習した決定木の root からの一部を図 4 に示す．

```
前のシステム発話終了時刻からの時間 <= 820 :
|   直前のユーザ発話開始から現在までの時間 <= 720 :
|   |   パワー 500ms 区間の傾き (3) <= 2.93346 :
|   |
|   直前のユーザ発話開始から現在までの時間 > 720 :
|   |   ピッチ 100ms 区間の傾き (1) <= 3.24639 :
|   |
|
前のシステム発話終了時刻からの時間 > 820 :
|   ポーズ長 <= 650 :
|   |   直前のユーザ発話開始時刻からの時間 <= 714 :
|   |
|   ポーズ長 > 650 :
|   |   パワー 500ms 区間の傾き (4) <= 0.997341 :
```
図 4: 決定木の例

しかし，今回作成した決定木の性能は，まだ不十分である．これに関しては，今後，決定木に用いる素性の見直しや，十分な学習データを用意して，全決定木を機械学習で獲得していく予定である．

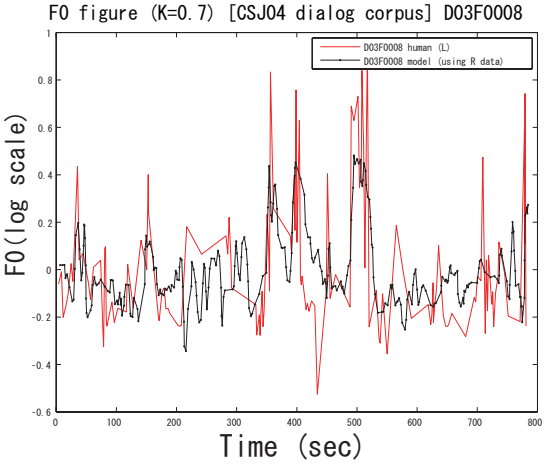

図 5: 対話音声の F0 とモデル出力値

4.3 韻律情報変化のモデル化

韻律情報として，基本周波数（F0）のモデル化を考える．実際の人間同士の対話では，お互いの F0 の変化には相関があることが分かったので，そのことを踏まえ，相手の F0 の変化をみて，その変化に合わせてシステム側の F0 を変動させるようにモデル化を行った．

モデルは，目標値に対して時定数を持って変動するように（3）式を用いた．

$$\begin{aligned} M(t) &= \mu_{sys} + \alpha_{sys}(t) \\ \alpha_{sys}(t) &= \alpha_{sys}(t-1) + K(\alpha_{usr3\mu} - \alpha_{sys}(t-1)) \end{aligned} \tag{3}$$

ここで，$M(t)$ は対話ターン t におけるモデルの値である．μ_{sys} は，時間によって変化しない，システムの標準値（平均値）を表す．$\alpha_{sys}(t)$ は，対話ターン t での，システムのオフセット値である．$\alpha_{usrN\mu}$ は，ユーザの直前 N 発話のオフセット値の平均である．これが目標値である．$\alpha_{sys}(t-1)$ は，t の 1 ターン前のシステムのオフセット値である．K は，時定数を表す．

ここで，現在は $K = 0.7$，$N = 3$ を用いている．最適値の設定は今後の課題である．

このモデルの確認のため，人間同士の対話のうちの一方の話者の韻律を (3) 式に当てはめて得られた値と実際のもう一方の話者の韻律とを比較した．これを図 5 に示す．両方の値は対数値であり，それぞれ平均値を引くことで正規化している．この図の場合には，話者 R の値をモデルに入力して得られた値が示されており，この結果は，システムが話者 L 側になったことを想定して出力した結果であるので，正解として実際の話者 L の値を描画してある．

CSJ コーパスの対話音声とモデル値の相関は，表 6 のようになった．モデルの出力値と実際の値との相関は，正の相関を示しており，また，実際の 2 話者間の基本周波数の相関値（表 2）と同程度になっており，対話中の韻律の変動をモデル化することが出来ていると考えられる．

表 6: モデルと実際の値との相関

種類	最大値	最小値	平均
D01	0.427	-0.105	0.136
D02	0.478	-0.025	0.205
D03	0.553	-0.016	0.208
D04	0.257	-0.212	0.080
平均	0.429	-0.090	0.157

5 まとめと今後

本研究では，協調的な音声対話システムを実現するために，人間同士の対話における韻律的な同調と対話としての盛り上がりとの関連を分析し，そのモデル化を試みた．また，そのモデルを音声対話システムに搭載した．音声対話システムは，リアルタイムにあいづち，話者交替などの応答タイミングを検出し種々の雑談現象を扱い応答することが出来る雑談に向けた対話システムであり，タイミングの検出と応答の種類の決定には決定木を用い，応答を出力する際の韻律情報は，モデルを用いて制御している．

今後は，このシステムを用いて被験者実験を行い，システムの評価を行う予定である．また，韻律制御については，現在は単にユーザに対して追従するだけであるが，対話の内容によっては能動的に韻律を変化させるようにしたい．そして，現在は出力音声として合成音声を用いているが，これを人間の声の編集合成音に変更し，使用感がどのように変化するかも調査する予定である．

参考文献

[1] 西村良太，北岡教英，中川聖一.: “応答タイミングを考慮した雑談音声対話システム” 人工知能学会研究会資料，SIG-SLUD-A503-05 (2006).

[2] Maekawa K., Koiso H., Furui S., Isahara H.: “Spontaneous speech corpus of Japanese”, *Proceedings of the Second International Conference of Language Resource and Evaluation*, 2, pp.947-952 (2000).

[3] 野口広彰，片桐恭弘，伝康晴.: “心理実験を用いたあいづち応答の手がかり特徴の検証” 人工知能学会研究会資料，SIG-SLUD-A002-13(2000).

[4] 大須賀智子, 堀内靖雄, 西田昌史, 市川熹.: “音声対話での話者交替/継続の予測における韻律情報の有効性”. 人工知能学会誌 Vol. 21 No. 1, pp.1-8, (2006).

[5] 小磯花絵，堀内靖雄，土屋俊，市川熹.: “先行発話断片の終端部分に存在する次発話者に関する言語的･韻律的要素について”, 電子情報通信学会技術報告，NLC95-72，pp.25-30(1996).

[6] 垣田邦子，“簡単な”質問-答”形式の対話における F0 の話者間相互作用”，日本音響学会講演論文集, 2-P-2, pp.305-306 (1995).

[7] 長岡千賀，小森政嗣，Draguna Raluca Maria，中村敏枝.: “2者対話における好意の表出～交替潜時を分析指標として～”，日本心理学会第 65 回大会発表論文集, 0341 (2001).

[8] Kitaoka, N., Takeuchi, M., Nishimura R., Nakagawa S.: “Response Timing Detection Using Prosodic and Linguistic Information for Human-friendly Spoken Dialog Systems”, 人工知能学会論文誌, Vol. 20, No. 3, pp.220-228 (2005) .