音声対話システムにおける対話中の韻律変化のモデル化と適用*

◎西村良太（豊橋技科大）　北岡教英（名古屋大）　中川聖一（豊橋技科大）

1　はじめに

人間と機械が対話を行う際に，機械が人間同士の会話と同じように，話者交代，割り込み，あいづちなどを自然に返すことが出来れば，より円滑な対話を行うことが期待できる．そのためには応答を返すタイミングや，出力音声の韻律情報を，実際の人間同士の対話のように制御する必要がある．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応答タイミングや韻律的な同調と，対話としての盛り上がり・意見の相違などとの関連を分析し，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実装した．

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･分析した．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) の中の対話コーパスである [2].

コーパスには，4 つのタスクがあり，D01〜D04 という名前が付いている．D01 は模擬講演インタビュー，D02 は課題指向対話，D03 は自由対話，D04 は学会講演インタビューである．

2.2　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のコーパスで調べた．

対話中の基本周波数は，各発話中の logF0 の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．

Table 1　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

相関値は表 1 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．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．（注：CSJ に男性同士の対話はない）

3　対話の印象と対話現象の関係

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 名の被験者（男性 1 名，女性 3 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（親しみがある 5-1 親しみがない）
- 盛り上がり（良い 5-1 盛り上がっていない）
- 相手の意見に（同意 5-1 意見を戦わす）
- 立場の違い（目上 5-1 目下）
- L 話者（インタビュアー）のフランクさ（気を使う 5-1 くだけている）
- R 話者（インタビュイー）のフランクさ
- L 話者の表現（敬語ばかり 5-1 敬語を使っていない）
- R 話者の表現

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の平均値は，0.470 であった.

その他のアンケート結果については，表 2 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，各対話音声の情報との相関を表 3 に示す.

Table 2　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

Table 3　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
年の差	-0.483	0.068	-0.411
実際の年の差	-0.301	-0.171	-0.002

表 3 中の ‘年の差’ は，‘立場’ の目上・目下の評価値を，目上か目下かに関係なく基準（レベル 3）からどれだけ離れているかという値にしたものである（$|\ Ans. - 3\ |$）．‘実際の年の差’ は，実際の話者の年齢差の値である．

---

*Modeling prosodic change in conversations and its application for a spoken dialog system. By Ryota NISHIMURA (Toyohashi University of Technology), Norihide KITAOKA (Nagoya University) and Seiichi NAKAGAWA (Toyohashi University of Technology)

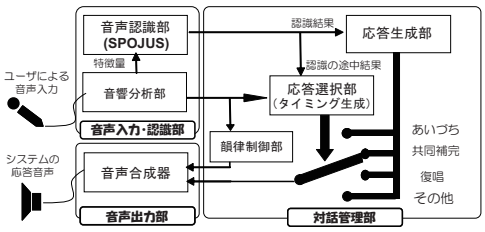

Fig. 1　システムの構成図

表 3 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，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場合にオーバーラップがたくさん起こる傾向を示している．「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 相関値に関しても，オーバーラップ頻度と同じような相関値を示している．

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．

‘年の差’ の項目を見てみると，‘F0 相関値’‘フィラー頻度’ のそれぞれについて，高い相関が出ている．これは「年齢差があると，F0 の同調性が無くなり，またフィラーの回数が少なくなる」ということである．また，‘実際の年の差’ は，‘フィラー頻度’ の相関がなく，アンケート結果と違っている．

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．

システムは，図 1 に示すような機構になっており，リアルタイムに処理を行っている．このことにより，システムはユーザへの割り込み（オーバーラップ）応答が可能である．また，韻律の制御も実装しており．ユーザの韻律変化に従って，システムの出力音声の韻律を変化させることが出来る．

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．決定木学習には，RWC 音声対話データベースを用いた [3]．素性は以下のとおりである．

- 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」「復唱」「共同補完」「話者交替 (システム応答)」「話者継続（待ち）」の 5 クラスを用意した．

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）のモデル化を考察した．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．

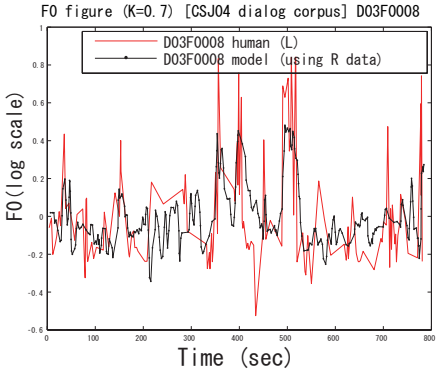

Fig. 2　対話音声の F0 とモデル出力値

モデルは，目標値に向かって時定数を持って追従するように（1）式を用いた．

$$M(t) = \mu_{sys} + \alpha_{sys}(t)$$
$$\alpha_{sys}(t) = \alpha_{sys}(t-1) + K(\alpha_{usrN\mu} - \alpha_{sys}(t-1)) \quad (1)$$

ここで，$M(t)$ は対話ターン t におけるモデルの logF0 値で，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），$\alpha_{sys}(t)$ 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，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 (1) 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 2 に示す．

CSJ コーパスの対話音声とモデル値の相関は，表 4 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 1）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．

Table 4　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　おわりに

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，決定木による応答タイミング生成モデルと共に，音声対話システムに実装した．今後はシステムの主観評価を行っていく予定である．

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] 田中，速水，山下，鹿野，板橋，岡.: “RWC 計画における音声対話データベースの構築” 情報処理学会音声言語処理 1 1 − 7，1996.