人間同士の対話の印象と韻律変化との関係の分析とそのモデル化*

◎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，相手に同調を示すことができれば，より円滑な対話を行うことが期待できる．そのためには，実際の人間同士の対話の印象がどのような要因で決められるのかを把握し，その情報を用いて相手に良い印象を与えていくようにする必要がある．本研究では，韻律情報に着目し，人間同士の対話の印象と韻律変化との間にどのような関係性があるのかを分析した．そして，音声対話システムへの実装を目指した円滑に対話を行うための韻律制御モデルの構築を試みた．

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

今回の分析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 [1]．コーパスには，4 つのタスクがあり，D01〜D04 という名前が付けられている．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．1 つの対話は 10 分〜20 分程度であり，計 58 対話である．

このコーパスについては，我々の先行研究にて大まかな対話単位の分析を行った [2]．より詳細な分析を行うために，本稿ではコーパスをトピックごとに 1 分程度の長さで分割したものも用意し，各トピックごとに対する分析も行った．

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．軌跡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．

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，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，トピックごとに分割したコーパスで調査した．全 58 対話を 1 分程度のトピックで分割した．分割したトピック数の合計は，709 である．ここで，F0 相関値を出す際に，2 話者間の発話数が少ないものは正しい結果が得られないため，2 話者の内，少なくとも片方の話者が 10 発話以下の発話数である場合には，その対話は分析対象外とした．分析対象となったデータ数は，566 トピックである．

Table 1　各対話の F0 の相関値

種類	最大値	平均	標準偏差
D01	0.716	0.145	0.247
D02	0.758	0.202	0.293
D03	0.710	0.166	0.265
D04	0.771	0.047	0.288
全体	0.771	0.150	0.276

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）にとり，次の同じ話者の発話の平均値と直線で結んだ．

2 話者間の対数基本周波数 logF0 の相関値を表 1 に示す，一般に相関値は大きく，対話中での 2 話者のお互いの F0 値には相関があると言える．また，566 トピック中，389 トピック (68.7%) が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．

詳細に分析すると，対話の内容によって，相関値に違いがみられた．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) に比べて相関が高くなっている（表 1 の平均）．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，F0 相関値が高かった上位 100 トピックのうち，67%が女性同士の対話であった．一方，女性と男性の対話は，相関値が低いものが多かった．但し，コーパスに含まれる対話は，片方が必ず女性であるので，男性同士の対話は収録されておらず，男性同士の対話での F0 相関がどのようになるかは，別のコーパスを用いて調査する必要がある．

図 2 に，F0 相関値のヒストグラムを示す．対話者間の F0 相関値は 0.0〜0.4 が多く見られることが分かる．

3　対話の印象と対話現象の関係

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，log パワー相関値，話速相関値）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」,「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.「年齢差」については，インタビュアー（L 話者）と相手の年齢差を予想してもらい評価値を付けた．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．

---

*Analysis of relationship between impression of human-human conversations and prosodic change and its modeling . By Ryota NISHIMURA (Toyohashi University of Technology), Norihide KITAOKA (Nagoya University) and Seiichi NAKAGAWA (Toyohashi University of Technology)

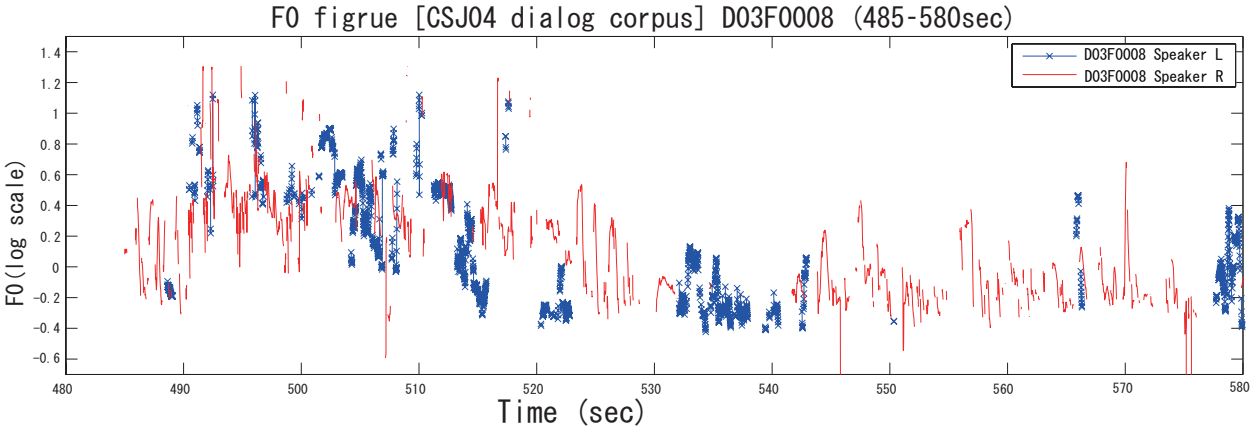

Fig. 1　CSJ 対話音声の例

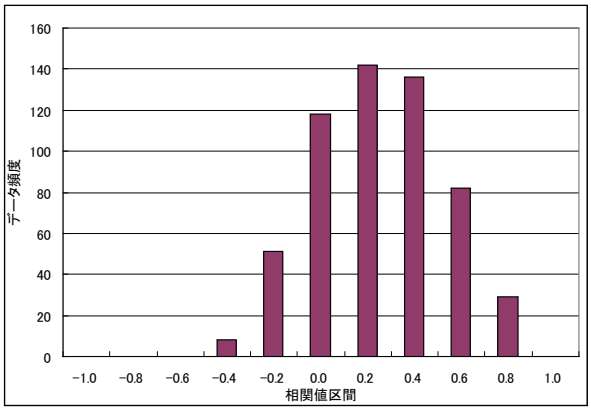

Fig. 2　トピック毎の F0 相関値のヒストグラム

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	
	対話毎評価	トピック毎評価
親しさ	0.444	0.361
盛り上がり	0.470	0.446
同意・否定	0.387	0.241
年齢差	0.478	0.376
かみ合い	結果無し	0.337
L 話者のフランクさ	0.399	0.128
R 話者のフランクさ	0.384	0.178
L 話者の表現	0.300	0.134
R 話者の表現	0.262	0.159

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，回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.

- 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）
- 年齢差（差が無い 5-1 差がある）
- L 話者（インタビュアー）のフランクさ（くだけている 5-1 気を使っている）
- R 話者（インタビュイー）のフランクさ
- L 話者の表現（敬語を使っていない 5-1 敬語ばかり）
- R 話者の表現

さらに詳細な分析を行うためのトピックごとに分割したデータに対しては，1 対話から 2 トピック，合計 116 トピックについて評価を行った．こちらについては 6 名の被験者（男性 4 名，女性 2 名）に対話音声を聞いてもらい，各対話について，前述の各項目に「かみ合い」の指標も追加してアンケートをとった.

- かみ合い（良い 5-1 ぎこちない）

ここで，アンケート結果の被験者間での違いを見るために，アンケート結果の被験者間の相関値を調べた．各アンケート項目に対する相関値の平均値は表 2 のようになった．対話毎評価はトピック毎に分割していない 10 分程度の対話音声に対する評価であり，トピック毎評価は，トピックごとに 1 分程度に分割した対話を用いて評価している.

この結果から,「親しさ，盛り上がり，年齢差，かみ合い」については，各被験者間で相関が高く，これらの指標については，安定して回答できたと言える．その他の項目については相関値が低くなっており，特に「フランクさ，表現」について相関値が低いことから，これらの項目については，回答にバラツキがあり，安定して回答できなかったと言える.

また，対話毎評価とトピック毎評価を比較すると，トピック毎評価の方が，被験者間の相関値が低くなっている．これは，被験者間で対話毎評価と比較して，安定して評価をつけることが難しかったということである．トピック毎のデータの場合，一つの対話を 1 分程度としてあるので，そこから評価をつけることが難しいと被験者から報告があり，表 2 の結果と一致している.

アンケート結果の評価値と，各対話音声の情報「F0 平均・分散の相関値」「パワー平均・分散の相関値」「話速平均・分散の相関値」との相関を表 3 に示す．結果は，あいづちやフィラーなどによって影響を受けていることが考えられるので，フィラーを除いた結果も示す．フィラーを除く際に CSJ 対話コーパスの中の詳細にタグ付けが行われている「コア」と呼ばれるコーパスを用いた．このコアは，全 58 対話コーパス中 18 対話にのみにしか付与されていないため，ここでの分析には，18 対話のみを用いている．また，ここでの結果は「対話毎評価」を用いた結果であり，トピック毎評価による結果ではない.

表 3 を見てみると，F0 平均は，各指標と相関がある．特に,「年齢差，フランクさ，表現」において，F0 平均と評価との相関が高かった.

フィラーの有無に関わらず,「パワー平均」が指標全体において高い相関を示している．特に「親しさ，盛り上がり，同意・反発，年齢差」の指標と相関が高かった．また,「話速平均」においても,「親しさ，盛

Table 3　被験者評価と対話現象との相関
(a) フィラーあり

フィラーあり	親しさ	盛り上がり	同意・反発	年齢差	フランク L	フランク R	表現 L	表現 R
F0 平均	0.394	0.368	0.253	0.598	0.400	0.344	0.627	0.553
F0 分散	0.132	0.022	-0.038	-0.101	0.173	0.082	0.226	0.248
パワー平均	0.510	0.547	0.701	0.342	0.193	0.320	0.257	0.231
パワー分散	-0.020	0.054	-0.080	-0.331	0.010	0.234	-0.139	0.201
話速平均	0.429	0.361	0.253	-0.219	0.340	0.529	0.045	0.608
話速分散	0.173	0.212	0.203	0.325	0.437	0.251	0.401	0.188

(b) フィラー抜き

フィラー除去	親しさ	盛り上がり	同意・反発	年齢差	フランク L	フランク R	表現 L	表現 R
F0 平均	0.277	0.238	0.138	0.395	0.379	0.387	0.502	0.522
F0 分散	0.334	0.231	0.243	-0.050	0.213	0.157	0.273	0.212
パワー平均	0.563	0.623	0.607	0.569	0.350	0.382	0.486	0.419
パワー分散	-0.117	-0.249	-0.125	-0.121	0.185	0.040	-0.148	-0.148
話速平均	0.272	0.268	0.226	0.051	0.463	0.450	0.196	0.259
話速分散	0.462	0.513	0.357	0.076	0.435	0.589	0.263	0.605

Table 4　オーバーラップ頻度，フィラー頻度と被験者評価の相関

	対話毎		トピック毎	
	overlap	filler	overlap	filler
親しさ	0.627	0.072	0.483	-0.266
盛り上り	0.718	0.127	0.580	-0.178
同意	0.638	0.090	0.568	-0.112
年齢差	0.068	-0.411	0.349	-0.299
かみ合い	値無し	値無し	0.641	-0.132
L フランク	0.417	-0.108	0.327	-0.381
R フランク	0.637	-0.047	0.511	-0.172
L 表現	0.238	-0.265	0.240	-0.353
R 表現	0.404	-0.289	0.193	-0.226

り上がり」の指標において高い相関が見られた．

表 4 に，対話現象として，オーバーラップ頻度とフィラー頻度について，被験者評価との相関を求めて示した．表 4 を見ると，オーバーラップ頻度は，各印象評価と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沢山起こる傾向を示している．「フランクさ」に対しても高い相関があるので，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．

フィラー頻度に関しては，「年齢差」「表現」で大きな負の値となっていることから，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．また，「親しさ」「盛り上がり」「同意・否定」に関しては，トピック毎に分けた対話においては，全て負の相関があった．「親しさ」につていは，対話毎に評価をした場合には，相関が見られなかったが，トピック毎に評価することで，負の相関が見られ，フィラーが沢山起こる対話は，あまり親しさが感じられないということである．これは，直感とも合うので，トピック毎に評価をしたことで，結果が安定したものと思われる．

4　韻律変化のモデル化

3 節において，F0 平均，パワー平均，話速平均について，親しさや盛り上がりなどと相関があることが分かった．そこで，これらについてモデル化を行い，そのモデルの評価を行う．最終的には，これらのモデルを音声対話システムに実装することが目的である．

実際の人間同士の対話では，お互いの韻律情報が同調するという実験結果が，相関などを求めることにより得られた．そのことを踏まえ，相手の韻律情報の変化をみて，その変化に合わせてシステム側の韻律情報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（1）式を用いた．

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

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．

このモデルの検証のため，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 3 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 L 側の推定値が示されており，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（ターゲット）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を，表 5 に示す．対話毎に評価を行った結果 (a) とトピック毎に評価を行った結果 (b) を示す．

モデルの出力値と実際の値との相関は，基本周波数 (F0)，パワーにおいて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 (F0) の相関値（表

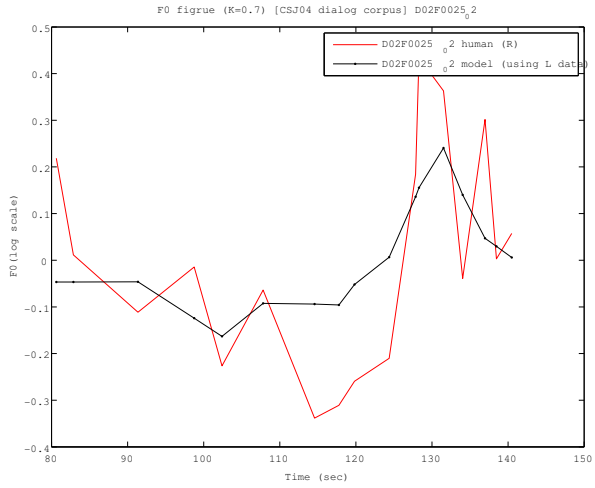

Fig. 3　対話音声の F0 とモデル出力値

1）と同程度になっていることから，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．

また，対話毎の評価に比べ，トピック毎の評価の方が最大値が大きくなっていることから，トピック毎のデータの方が，対話内での印象のばらつきが少なく，各状況をより細かく評価できたものであると考えられる．

Table 5　モデルと実際の値との相関

(a) 対話毎の評価

	最大値	平均値	標準偏差
F0	0.550	0.116	0.128
パワー	0.440	0.121	0.125
話速	0.185	-0.121	0.135

(b) トピック毎の評価

	最大値	平均値	標準偏差
F0	0.714	0.087	0.302
パワー	0.765	0.104	0.289
話速	0.884	-0.066	0.329

5　人間は韻律情報のみで音声を評価できるか

対話中の話者間の基本周波数 (F0) に相関があることや，対話中の印象と韻律情報との間に相関があることが分かった．そして，4 節で韻律情報を用いて実際にそれらをモデル化し妥当性を示したが，実際の人間が盛り上がりについて韻律情報 (ピッチ (F0)，パワーなど) のみで評価を行うことができるかどうかについては分からない．そこで，韻律情報のみを残した音声（以降，ハミング音）を用いて，その音声に対して評価を行う被験者実験を行った．

被験者は，通常の対話音声に対しても評価実験を行った 3 名（男性 1 名，女性 2 名）である．評価項目については，3 節のトピック毎音声の評価に用いたものと同じものを用いるが，「表現」は言語情報に対する評価であったため，ここでは省略した．

ハミング音評価の被験者間の相関値の平均を，表 6 に示す．ここで，評価項目が 3 つになっているが，これは被験者のうちの 1 人が，これ以外の項目（同意，年齢差，フランクさ）について評価することが不可能であった為である．残りの被験者 2 人は，これ以外の項目を評価はしたものの，対話毎評価においては 58 対話中 50 対話程度に「3」が付与されており，トピック毎対話においては相関値が非常に低くなっていた (0.1 程度であった)．「同意，年齢差，フランクさ」については，評価が困難であったと考えられる．

Table 6　ハミング音評価の被験者間の相関値の平均

	対話毎評価	トピック毎評価
親しさ	0.347	0.505
盛り上がり	0.425	0.584
かみ合い	データ無し	0.446

Table 7　通常音声とハミング音との間の評価の相関 (3 人中 2 人のデータ)

	対話毎評価	トピック毎評価
親しさ	0.398	0.492
盛り上がり	0.457	0.400
かみ合い	データ無し	0.357

表 2 と表 6 を比較すると，表 2 では評価対象のデータ時間が短くなると（対話毎 v.s. トピック毎）被験者間の相関が低くなっているのに対し，表 6 では相関が高くなっている．これは，前者は短くすることで言語情報にあまり頼れなくなり評価が難しくなっていたが，後者は韻律情報しかないため，短くなったことで被験者間の評価のずれが小さくなったのではないかと考えられる．

表 7 に，通常音声とハミング音との間の評価の相関を示す．この表によって，通常の音声を聞いた場合の評価値と，ハミング音を聞いた場合の評価値との間の違いを見ることが出来る．ここで，1 人の被験者について，全項目について非常に相関が低かった為，表 7 には含んでいない．残りの 2 人の被験者については，表にある項目については 0.3〜0.5 の相関が見られた．このことから，韻律のみの情報から，ある程度は，「親しさ，盛り上がり，かみ合い」といったことが判定できることが示された．一方，言語情報のみによってどの程度評価できるかについても調査を行い，これらの間の関係性について分析を行いたい．

6　おわりに

円滑な対話を実現するために，人間同士の対話の印象と韻律変化との間にどのような関係性があるのかを分析した．そして，音声対話システムへの実装を目指した，円滑に対話を行うための韻律制御モデルの構築を試みた．今後は，このモデル化を音声対話システムに実装して，被験者実験にて評価する予定である．

韻律のみの音を聞いて評価をすることが出来るかということについては，ある程度は判定が可能であるのではないかという結論に至った．一方，人間は言語情報のみを与えられた場合に，そこから韻律情報を復元して発生している可能性もあり，そのことを調査する為に，言語情報から想定される韻律とは異なる韻律情報を持つ音声に対する評価実験やテキスト情報のみからの評価実験を行う予定である．

参考文献

[1] 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.

[2] 西村良太，北岡教英，中川聖一.: “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” 日本音響学会 2007 年春季研究発表会，社団法人 日本音響学会，1-9-3，pp.5-6，2007.