音声対話システムにおける挨拶発話の適切なタイミング生成

小林 弘幸*・大村 卓矢*・山本 知仁*

Proper Timing Generation of Greeting in Dialog System

Hiroyuki KOBAYASHI*, Takuya OMURA* and Tomohito YAMAMOTO*

Speech dialogue systems, such as Apple's "Siri," have gradually become more widespread, and in the near future, a greater number of general users will have the opportunity to communicate with such systems. To facilitate this, it is necessary for the system to communicate more naturally with users, and to realize that both verbal and nonverbal information must be taken into consideration. In this research, using an available speech recognition platform, we have developed simple dialogue system that consider utterance timing and evaluated it. As a result, it is clarified that the mean error of generated pause duration is controlled within 20 msec, and the proper fixed pause duration is turned to be 600 msec in this system. Moreover, the evaluation of generative models of pause duration showed that the model that synchronized the utterance duration weakly or fixed pause duration (600 msec) model tended to get best results.

Key Words: dialogue system, utterance timing, pause duration, utterance duration

1. はじめに

これまで音声対話システムは、カーナビゲーションシステムやコミュニケーションロボットなど、さまざまなシステムで利用されてきている。近年では、Appleの「Siri」のようなモバイル端末で動作するシステムが実用化されつつあり、その有用性が広く認知されてきている。また、「スマートメイちゃん」¹⁾のような音声認識システムに身体エージェントをもたせたものなど、このような対話システムはさまざまな広がりを見せており、これらを開発するようなプラットフォームも増えてきている。しかし、現在利用されている音声対話システムと人とのコミュニケーションにおいては、ユーザがシステムに対して違和感をもつことがある。これは、人間同士の対話には発話タイミングや韻律的な情報、ジェスチャーなどの非言語情報が存在しており、現在の音声対話システムはこれらを十分に考慮していないといったことが原因として考えられる。今後、音声対話システムは多様なユーザのインタフェースとなりうるため、さらに普及することが期待される。このことを考えると、人の対話における非言語的な要素を総合的な観点から解析してモデル化を行ない、音声対話システムへ導入していく必要があると考えられる。

これまでに、このような非言語情報に関する研究、特に本

研究で注目する発話タイミングのような対話の時間的構造に関する研究は、リズムの同調という観点から行なわれていることが多い。たとえば、Matarazzoらは面接時の対話において、面接官と面接者の間で発話長や反応潜時長が同調することを報告している^{2),3)}。また、長岡らや小磯らも発話速度や反応潜時長といった対話時の時間的特徴量が話者間で同調していくといったことを報告している^{4),5)}。現在では、このような現象を同調傾向⁶⁾と呼ぶことがあり、円滑なコミュニケーションを実現するための重要な要素であると考えられている。

これらの研究に加えて、近年ではコミュニケーションにみられるリズムの同調をアバターやロボットとの対話に応用しようという研究も行なわれている。たとえば北岡らは、談話コーパスより発話タイミングを決める決定木を求め、それを用いて自然な発話タイミングで対話することができるシステムの構築を行なっている^{7),8)}。さらに、渡辺らや山本らは発話とうなずきのリズムに観察される引き込み現象をアバター、ロボットなどさまざまな音声インタフェースに応用し、うなずきの反応がインタラクションの支援や促進に有効であることを示している^{9)~12)}。

特に、山本らは人の基本的なコミュニケーション行為として挨拶動作をとりあげ、その動作と発話タイミングの解析を行ない、結果として、挨拶動作が発話より先行することを明らかにしている。加えて、これらの特性を対話ロボットに実装して評価を行ない、発話に対して動作を先行させることが、システムに自然性を感じさせる要因になっていることも示している¹⁰⁾。

われわれの研究グループも、これまでに人の対話における

* 金沢工業大学大学院工学研究科 野々市市扇が丘 7-1

* Graduate School of Engineering, Kanazawa Institute of Technology, 7-1 Ohgigaoka, Nonoichi

(Received August 10, 2014)

(Revised December 3, 2014)

リズムの同調に観点を置いて解析を行なっている。具体的には、指示発話「積み木を取ってください」と応答発話「はい」、もしくは「はいわかりました」、それぞれ1回からなる簡単な対話を設定し、被指示者の応答発話の違いによる発話、反応潜時長間の相関関係を解析した¹³⁾。また、これらの実験から得られた結果を基にタイミング制御モデルを構築し、それをアバターに実装して評価する実験¹⁴⁾、およびコミュニケーションロボットに実装して評価する実験を行なった¹⁵⁾。結果として、タイミング制御の有無によって対話の印象評価に差が現れ、制御モデルを導入したシステムとのコミュニケーションがより好ましい印象を与えることを明らかにしてきた。

このように、対話におけるリズムの同調の解析や、リズムを制御するコミュニケーションシステムに関する研究はこれまでも行なわれてきている。一方で、現在実現されつつあるより高度な音声対話システムのプラットフォームにおいて、このようなリズムの制御ができるかどうかの検証や、リズムを生成するモデルを導入した場合のユーザの評価に関してはまだ十分に研究が行なわれていない。今後、音声対話システムはより複雑な対話を実現することが期待され、その際には、より複雑なタイミング制御が求められることが考えられる。これを実現するためには、まずは挨拶のような簡単な対話における基礎的な知見を得て、この結果を基に複雑な対話にアプローチしていくことが重要であると考えられる。よって本研究では、現在利用できる音声対話システムのプラットフォームを利用し、反応潜時長を調整することで挨拶の発話タイミングを制御する、簡易的な音声対話システムの構築をまず試みる。つぎにそのシステムを用いてユーザが対話を行ない、印象評価を行なうことで、このような音声対話システムの実現可能性について議論する。

2. 発話タイミングを考慮する音声対話システム

本研究では、まず現在利用できる音声認識システムのプラットフォームを用いて反応潜時長を調整し、発話タイミングを制御する簡易的な音声対話システムを構築する。本研究における、反応潜時長とは **Fig. 1** に示すようにユーザの発話終了から、システムが応答を開始するまでの時間とする。

現在、オープンに利用できるこのようなプラットフォームとしては、ローカルなパーソナルコンピュータ (PC) 環境で動作するシステムと、クラウド技術を使用したシステムの2種類が考えられる。このうち、ローカルなPCで動作するシステムは発話タイミングなどの時間的な制御が比較的容易であるが、クラウド技術を利用するシステムに比べアプリケーションの拡張性が乏しいという側面がある。一方で、クラウド技術を利用したシステムは、ネットワークのリソースを利用することができ、さまざまなアプリケーションへの応用がしやすいが、ネットワークトラフィックの影響などを受けるため、発話タイミングなどの時間制御が難しいという側面がある。これら2種類のプラットフォームのうち、本研究ではまずローカルなPC環境で動作するシステムに着目し、その

Fig. 1 Temporal durations of dialogue

Fig. 2 Configuration of simple utterance dialogue system

知見を述べる。具体的には **Fig. 2** に示すような、音声認識エンジン「Julius¹⁶⁾」を用いた簡易的な音声対話システムを構築し、発話タイミングの制御可能性について検討する。

本研究で構築したローカルな PC (クロック周波数: 2.60 GHz, メモリ容量: 4 GB) で動作する音声対話システムは大きく分けて、音声認識部、発話タイミング制御部、応答部の3つに分けられ、音声認識部は Julius で行なう。Julius には認識結果やシステムの状態を知るためのコールバック関数が備わっており、本システムにおいて発話タイミングの制御は、このコールバック関数 (status_rec.begin, status_rec.stop, output_result)¹⁶⁾を用いて行なう。これにより、発話区間の計測開始と終了時に割り込み処理を行なうことができ、取得した時間の差分を求めることで発話区間をリアルタイムで検出できる。つぎに反応潜時長を固定長、もしくは生成モデルを用いて算出し、その時間だけ待機した後、応答発話を行なうことで発話タイミングを制御する。本システムにおける応答は、アクエスト社の音声合成ライブラリ「Aques Talk2」によって生成された wav ファイルを再生することで行なう。

反応潜時長の調整は、基本的には先に述べたように発話終了後に時間長分だけシステムが待機し、発話を開始すれば実現できる。しかし、Julius を含め音声認識エンジンを使用する場合には、いくつかの要因からの時間遅れを考慮する必要がある。その要因の1つは、認識システムが発話の終了を判断する際に必要とする時間的なマージンである。システムの音声終了判定は、多くの場合、ある時間入力がないことをもって行なわれる。この判定にどの程度の時間的なマージンをとるかで、時間遅れが変化する。本研究ではこの時間遅れを決定するために、あるマージン値 (150 msec) をシステム側で設定した。

また、発話内容によって、システムが判定する発話終了タイミングは変動する。そこで、実験で使用する挨拶発話「こんにちは」に対する発話終了タイミングの推定精度を向上させるため、被験者1名に発話速度を「ゆっくり」、「ふつうに」、「はやく」の3段階で「こんにちは」という発話を各3回ずつ、シス

テムに対して行なった。このときの人の音声データを録音し、この実測値とシステム側で計算される値を比較して発話終了タイミングを推定することで、遅れ時間（約 250 msec）を算出した。もう 1 つの時間遅れの要因としては、wav ファイルを再生する際に生じる遅延がある。本システムでは wav ファイルの再生に Windows のライブラリ関数である PlaySound() を用いているが、この関数の実行には一定の遅延（約 30 msec）が生じることが知られている。本研究では音声認識後、反応潜時長を遅延分補正することでこの時間遅れに対応した。

また、Julius において多くの単語の認識を行なうためには、あらかじめ辞書に多くの単語を記憶させておく必要がある。一方で、この単語数が多いと認識率が下がってしまい、正確に回答することができない。本研究では、「こんにちは」という単語のみを登録することで認識率を上げ、できるだけ正確に回答できるように設定した。

3. 実験手法

3.1 評価する固定長の発話タイミングと生成モデルについて

本研究では、まず先に述べたローカル環境の音声対話システムが固定長の発話タイミングで制御できるかどうかを検証すると同時に、どの程度の固定長がユーザにとって最も適切かを明らかにする実験を行なう。具体的には、複数の固定の反応潜時長をシステムに設定してユーザと対話させ、これらを一対比較（中屋の変法を用いた）することによって、適切な発話タイミングを評価した。対話の内容としては、被験者が「こんにちは」とシステムに話し、システムが「こんにちは」と答えるものとした。

本研究では、まず実験 1 として 5 つの固定の反応潜時長（300, 500, 700, 900, 1100 msec）を用意し、適切な固定長の絞り込みを行なった。その後、時間の刻み幅を狭めた 5 つの固定の反応潜時長（400, 500, 600, 700, 800 msec）の中から、適切な固定長をさらに絞り込む実験 2 を行なった。続いて、先行研究で明らかになった、発話タイミング生成モデルの評価を実験 3 として行なった。このとき、先行研究の実験^{13)~15), 17)}から、人間同士の簡単な発話応答において、相関分析を行なった結果、指示発話長が長くなると反応潜時長も長くなることが示唆された。この結果から Fig. 3 および以下に示す 5 つのモデルを用意し、これらを一対比較することで、適切なモデルの評価を行なった。各式中、定数は実験 1, 2 から得られた結果より求められ、PD は反応潜時長を、UD は発話長を示している。

(a) Model 1: $PD = -(1.0 \times UD) + Const.1$

(b) Model 2: $PD = Const.2$

(c) Model 3: $PD = (0.5 \times UD) + Const.3$

(d) Model 4: $PD = (1.0 \times UD) + Const.4$

(e) Model 5: $PD = (2.0 \times UD) + Const.5$

これらのモデルのうち、Model 4 は、先行研究を基に構築した、発話長から反応潜時長を算出する線形のモデルである（傾きの係数は簡単のため 1.0 とした）。Model 3 は、Model 4

Fig. 3 Generation models of utterance timing

Fig. 4 Experiment system

の傾きを緩やかにしたものであり、逆に Model 5 は傾きが急なモデルとなっている。Model 1 は、Model 3 と傾きが逆になるモデルである。Model 2 は、実験 1, 2 より算出される固定長である。これらのモデルは、動的に反応潜時長を生成するため、発話長が極端に短い場合（Model 1 の場合には長い場合）に、非常に小さな値を取りうる可能性がある。しかしながら、本システムで安定して実現できる最も短い反応潜時長が 300 msec 程度であるため、これより短い反応潜時長が算出された場合には、300 msec の反応潜時長を使用することにした。本研究では、以上の発話タイミング生成モデルを用いて実験を行ない、モデルの特性によってどのように被験者の評価が変化するかを明らかにする。

3.2 被験者と実験手順

実験 1~3 には、健康な 20 代男子大学生 15 名（平均年齢 22.4 歳）が参加した。各実験において、被験者は、Fig. 4 に示すようにモニターと向き合うように椅子に座った。その後、ワイヤレスヘッドフォンを装着し、モニターに表示される指示に従って「こんにちは」と発話を行なった。このとき、発話を行なう速度は実験者がディスプレイを通じて指示し、速度には「ゆっくり」、「ふつうに」、「はやく」の 3 つを用意した。「ふつうに」の速度では、被験者が普段挨拶をするような速度で話すように指示した。実験中、視覚情報が実験結果に影響することを考慮し、実験者と被験者の間にカーテンを引いた。また、システムが発話タイミングを正確に制御できているかを確認するために、ボイスレコーダ（Roland R-05）を使用して実験中の音声すべてを録音した。各実験は以下の手順で

行なった。

- 手順1：被験者に実験目的と、発話速度の指示、評価用紙記入方法の説明を行なう。
- 手順2：被験者の音声に合わせてマイクレベルの調整を行なう。
- 手順3：被験者は実験者の指示に従って、発話速度を3段階で変化させて発話を行なう。このとき、各発話速度において、2つの固定長もしくはモデルを交互に2回ずつ、4回発話を行なう。
- 手順4：各発話速度の試行結果を総合的に判断し、評価用紙の「練習」の記入欄に印象評価の結果を記入してもらう。
- 手順5：被験者は実験者の指示に従って、発話速度を3段階で変化させて発話を行なう。このとき、各発話速度において、2つの固定長もしくはモデルを交互に2回ずつ、4回発話を行なう。
- 手順6：各発話速度の試行結果を総合的に判断し、評価用紙に記入をしてもらう。
- 手順7：手順5から6までの流れを1セットとし、これを10回繰り返す。

固定長もしくはモデルの選択は、順序効果が出ないようにランダムに行ない、被験者に指示する発話の速度「ゆっくり」、「ふつうに」、「はやく」もランダムに提示した。音声認識に失敗した場合、先行固定長（もしくはモデル）、後行固定長（もしくはモデル）の組み合わせをもう一度行なうようにした。

実験中のシステム側のデータ（発話長、固定の反応潜時長の設定値もしくは生成モデル、対話の合計時間、合計時間から逆算した反応潜時長）は、ログとしてシステム側に保存した。このデータと被験者に記入をもらった評価用紙、レコーダによる録音データを基に解析を行なった。対話システムの主観的な評価には「好きか」、「自然か」、「丁寧か」、「親しみやすいか」、「速いか」の5項目を用意し、被験者に5段階で評価させた。最初の4つの項目は先行研究を基に¹⁵⁾被験者の印象を調査する目的で、残りの1つは被験者が発話タイミングの違いを正しく認識しているかを調査する目的で用意した。

4. 実験結果

4.1 固定長の発話タイミングの結果

まず、固定長の発話タイミングの結果について示す。**Table 1**、**Table 2**は実験1、2において実現された反応潜時長の平均値と設定値との誤差の平均を示している。各値は、レコーダで記録された15被験者×1条件（「ふつうに」の条件のみ）×8試行＝120データ分から計算された。結果より、各時間長において20 msec程度（平均で20.8 msec、300 msec時に誤差が大きくなるのは、300 msecまでにシステムが応答の準備をできない場合があるためである）の誤差があることがわかるが、この誤差の値は、中間的な値である600 msecのときに3%程度である。誤差はできるだけ小さいほうが好ましいが、本実験において、発話-応答までの全体の

Table 1 Set pause duration and its generated value (Experiment 1)

PD	300	500	700	900	1100
Mean	325	509	701	903	1099
Error	30.2	23.1	20.8	20.5	22.0

Table 2 Set pause duration and its generated value (Experiment 2)

PD	400	500	600	700	800
Mean	405	500	601	699	795
Error	22.1	20.8	20.4	20.7	19.9

Fig. 5 Results of pairwise comparison in Experiment 1

Table 3 Mean value in Experiment 1

	Natural	Like	Polite	Familiar	Fast
300	-0.0800	-0.173	-0.613	-0.253	1.16
500	0.573	0.520	0.173	0.627	0.507
700	0.307	0.360	0.240	0.400	-0.120
900	-0.160	-0.0933	0.320	-0.120	-0.613
1100	-0.640	-0.613	-0.120	-0.653	-0.933

Table 4 Standard deviation in Experiment 1

	Natural	Like	Polite	Familiar	Fast
300	0.412	0.495	0.487	0.503	0.433
500	0.593	0.619	0.610	0.495	0.262
700	0.437	0.427	0.502	0.432	0.261
900	0.439	0.425	0.526	0.356	0.225
1100	0.457	0.338	0.661	0.470	0.454

長さが約2 secであることを考えると、現段階でも十分に固定長で発話タイミングを制御できているのではないかと考えられる。

このように固定長の発話タイミングが実現された対話において、被験者の主観的な評価がどのようなものであったかをつぎに示す。**Fig. 5**に実験1における各時間長に対する一対比較の結果を示し、**Table 3**、**Table 4**に具体的な平均値と標準偏差の値を示す。この実験結果において、5つの評価項目すべてにおいて、主効果に統計的な有意差がみられた($p < .05$)。これらの項目のうち、まず「速いか」の結果をみると被験者が適切に発話のタイミングを認識しているのがわかる。この結果のもと、評価項目「好きか」、「自然か」、「親しみやすいか」をみるとすべて同じ傾向にあることがわかり、500 msecが最も良い結果を得ていることが図よりわかる。一方、「丁寧か」については900 msecが最も良い評価を得てお

Fig. 6 Results of pairwise comparison in Experiment 2

Table 5 Mean value in Experiment 2

	Natural	Like	Polite	Familiar	Fast
400	-0.0133	-0.0267	-0.373	-0.107	0.907
500	0.280	0.227	-0.0667	0.293	0.360
600	0.293	0.373	0.0667	0.307	-0.0400
700	0.000	0.000	0.240	-0.0267	-0.413
800	-0.560	-0.573	0.133	-0.467	-0.813

Table 6 Standard deviation in Experiment 2

	Natural	Like	Polite	Familiar	Fast
400	0.573	0.619	0.661	0.473	0.431
500	0.510	0.341	0.289	0.449	0.427
600	0.406	0.489	0.391	0.449	0.408
700	0.575	0.589	0.359	0.546	0.225
800	0.427	0.431	0.530	0.454	0.539

り、やや遅い反応のほうが良い評価を得ていることがわかる。

この実験 1 を踏まえて、われわれは構築した対話システムの適切な固定長が 400~800 msec の中にあると考え、時間の刻み幅を小さくした実験 2 を続いて行なった。Fig. 6 にその結果を示し、Table 5, Table 6 に具体的な平均値と標準偏差の値を示す。まず、実験 1 と同様に 5 つの評価項目すべてにおいて、主効果に統計的な有意差がみられ ($p < .05$)、「速いか」の項目をみると、被験者が適切に発話のタイミングを認識しているのがわかる。また「好きか」、「自然か」、「親しみやすいか」の評価項目はすべて同じ傾向にあり、実験 2 では 600 msec が最も良い評価を得る結果となった。「丁寧か」に関しては、700 msec が最も良い評価を得ており、実験 1 と同様にやや遅い反応のほうが良い評価になることがわかった。これらの実験結果より、本研究では 600 msec 前後が構築したシステムにとって適切な固定長であると判断した。

4.2 発話タイミング生成モデルの結果

本研究では先の実験 1, 2 に続いて、先行研究で得られている発話タイミングを生成するモデルを評価する実験 3 を行なった。このとき、3 章で示した各モデルの定数は実験 1, 2 で得られた結果より決めたが、まず最適な固定長は実験 2 より 600 msec とした。その上で、平均的な被験者の発話長 (= 661 msec) を実験 1, 2 の「ふつうに」の条件より算出し、各モデルにおいて、この発話長が入力されたときに 600 msec の反応潜時長を返すように定数を決定した。

Fig. 7 (a)~(e) に本実験で使用された各モデルが生成し、

Fig. 7 Result of generated pause duration in each model

システム側で記録された反応潜時長の分布を示す。用いたモデルのうち、Model 2 から Model 5 は傾きがだんだんと大きくなるモデルであり、反応潜時長の分布がだんだんと広がることが想定される。実際に図をみると、そのように分布が変化していることから、モデルが想定どおりに動作していることがわかる。Model 1 に関しては Model 4 と線対称な結果が得られるはずであるが、実際に結果をみるとそのように

Fig. 8 Result of pairwise comparison in Experiment 3

Table 7 Mean value in Experiment 3

	Natural	Like	Polite	Familiar	Fast
Model 1	-0.160	-0.307	-0.173	-0.107	0.213
Model 2	0.173	0.240	0.0667	0.0933	0.000
Model 3	0.160	0.227	0.160	0.200	-0.173
Model 4	0.0133	0.0933	0.0667	0.0267	-0.0800
Model 5	-0.187	-0.253	-0.120	-0.213	0.0400

Table 8 Standard deviation in Experiment 3

	Natural	Like	Polite	Familiar	Fast
Model 1	0.445	0.443	0.455	0.505	0.524
Model 2	0.449	0.427	0.363	0.484	0.393
Model 3	0.433	0.526	0.381	0.438	0.425
Model 4	0.476	0.399	0.448	0.521	0.412
Model 5	0.591	0.591	0.560	0.568	0.603

なっていることから、モデルが想定どおりに動作したことが確認できる。

このようなモデル動作の結果のもと、被験者が各モデルに対してどのような主観的な評価を行なったかについて示す。Fig. 8に、実験3における一対比較の結果を示し、Table 7、Table 8にこのときの平均値と標準偏差の値を示す。これらの結果のうち、「好きか」の評価項目において主効果に統計的な有意差がみられ($p < .05$),「自然か」「親しみやすいか」に有意傾向($p = 0.068$, $p = 0.055$)がみられた。これらの3つの評価項目においては、発話タイミング固定のModel 2とタイミングが変化するModel 3に統計的に有意な差はみられなかったが、「好きか」「自然か」ではModel 2とModel 3が同程度の評価を得ているのに対し、「親しみやすいか」では、Model 3のほうが良い評価を得ている傾向にあるのが図よりわかる。一方、大きく変化するModel 5や逆に変化するModel 1などは、評価が低いことがわかる。これらの結果より、今回の実験で用いている挨拶のような対話では、発話長に対して緩やかに同調するような発話タイミングを生成するモデルか、600 msec程度で固定のモデルが良い印象を与えることが示唆される。

5. 考 察

本研究では、現在利用されている音声対話システムのプラットフォームを利用して、反応潜時長を調整することで発話タイミングを制御する簡単な音声対話システムを構築した。また、

そのシステムの動作に関する評価を行なった。結果として、ローカルなPC環境で動作する音声対話システムにおいて、固定長の発話タイミングで発話を行なう場合、誤差20 msec程度でシステムがタイミングを制御できることが明らかになった。この結果のもと、複数の固定長の発話タイミングを被験者に示し、主観的な評価を行なった結果、600 msecが「好きか」「自然か」「親しみやすいか」の評価項目において、最も良い結果を得ることが明らかになった。また、この時間長とこれまで先行研究で得られてきた発話タイミング生成モデルを基に複数のモデルの評価を行なった結果、緩やかに発話長に同調するモデルか、もしくは600 msecで固定の発話タイミングが「好きか」「自然か」「親しみやすいか」の評価項目において、最も良い評価を得る傾向があることが明らかになった。

まず、本研究ではJuliusを用いて発話タイミングを制御する音声対話システムを構築したが、固定長の発話タイミングを提示する実験1, 2において、その誤差が20 msec程度に収まった。これは、このようなローカルなPCを用いた音声対話システムにおいては、すでに十分に発話タイミングを制御してユーザに発話を提示することが可能であることを意味している。

その上で、ユーザに良い主観的な評価を与える固定の発話タイミングについては、本実験に関しては600 msec程度となった。この時間長に関しては、発話内容¹³⁾やアバターなどをを用いた場合¹⁴⁾、さらには身体動作が加わった場合¹⁰⁾などによって適切な値が変わることがあるが、視覚的な情報がない挨拶のような単純な音声対話においては、今回の実験で得られた結果付近が適切であることが推測される。

これに加えて本研究では、これまでの先行研究で得られた発話タイミング生成モデルと実験1, 2で得られた結果を踏まえ、複数のモデルを構築し、その評価を行なった。結果として、緩やかに話者の発話長に同調するモデルか、もしくは固定長の発話タイミングが良い評価を得ることが明らかになった。この結果は、これまで明らかにされてきた発話長-反応潜時長間でみられる同調傾向が、対話の印象を良くするという結果を支持するものであり、われわれのモデルが有効であることを示唆している。

一方、一般的な対話は発話長が長く、さまざまな文脈を含むため、反応潜時長が決まる過程はより複雑であることが推測される。また、これらの要因のほかに、地域や文化、年齢の違いによっても適切な反応潜時長は変化することが考えられる。たとえば、先行研究では、反応潜時長の制御が、若年者より高齢者で効果的であることを示しており¹⁷⁾、高齢者には今回提示した制御モデルがより良く評価されることも考えられる。よって、今後は、今回の実験のような単純な対話を基礎としてさらに複数の発話からなる対話や、さまざまな「文脈」を含む実験を行ない、そこでみられる発話のリズム構造をボトムアップ的に明らかにしていくことを考えている。また同時に、音声対話システムに発話リズムに関するモデルを

導入して、その評価を行なっていくことも考えている。

6. ま と め

本研究では、現在利用されている音声対話システムのプラットフォームを利用して、発話タイミングを制御する簡単な音声対話システムを構築し、その評価を行なった。結果として、発話タイミングを誤差 20 msec 程度で制御できることが明らかになり、固定の発話タイミングに関する評価を行なった結果、600 msec が最も良い結果を得ることが明らかになった。加えて、複数の発話タイミング生成モデルの評価を行なった結果、緩やかに発話長に同調するモデルもしくは 600 msec で固定の発話タイミングが、良い評価を得る傾向があることが明らかになった。

今後は、複数の発話からなる対話や、さまざまな「文脈」を含むより一般的な対話に近い実験を行ない、発話のリズム構造を体系的に明らかにしていくことを考えている。また同時に、音声対話システムに発話リズムに関するモデルを導入して、より実用的な音声対話システムの実現を考えている。

参 考 文 献

- 1) 山本, 大浦, 西村, 打矢, 内匠, 李, 徳田: スマートフォン単体で動作する音声対話 3D エージェント「スマートメイちゃん」の開発, 情報処理学会インタラクション 2013 講演会予稿集, 675/680 (2013)
- 2) J.D. Matarazzo and A.N. Wiens: Interviewer influence on durations of interviewee silence, *Journal of Experimental Research in Personality*, **2**-1, 56/69 (1967)
- 3) J.D. Matarazzo, M. Weitman, G. Saslow and A.N. Wiens: Interviewer influence on durations of interviewee speech, *Journal of Verbal Learning and Verbal Behavior*, **1**-6, 451/458 (1963)
- 4) 長岡千賀: 対人コミュニケーションにおける非言語行動の 2 者間相互影響, 対人社会心理学研究, **6**, 101/112 (2006)
- 5) 小磯, 下嶋, 岡田, 片桐: 会話相互作用におけるリズム, 人工知能研究会資料, SIG-SLUD-9602, 19/24 (1996)
- 6) 大坊郁夫: 同調傾向, 中島, 安藤, 子安, 坂野, 繁樹, 立花, 箱田 (編), 心理学辞典, 有斐閣 (1999)
- 7) N. Kitaoka, M. Takeuchi, R. Nishimura and S. Nakagawa: Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems, 人工知能学会論文誌, **20**-3, 220/228 (2005)
- 8) 伊藤, 北岡, 西村: 音声対話における発話タイミングの影響, 信学技報, NLC2008-20, 7/12 (2008)
- 9) 渡辺, 大久保, 中茂, 檀原: InterActor を用いた発話音声に基づく身体的インタラクションシステム, ヒューマンインタフェース学会論文誌, **2**-2, 21/29 (2000)
- 10) 山本, 渡辺: ロボットとのあいさつインタラクションにおける動作に対する発声遅延の効果, ヒューマンインタフェース学会論文誌, **6**-3, 87/94 (2004)
- 11) 山本, 渡辺: 身体的エージェントの情報提示インタラクションにおける動作に対する発声タイミング制御の効果, ヒューマンインタフェース学会論文誌, **10**-2, 135/143 (2008)
- 12) 長井, 渡辺, 山本: 聞き手のうなずき反応を視触覚提示する音声駆動型身体的引き込みシステム, 日本機械学会論文集 C 編, **75**-755, 2059/2067 (2009)
- 13) 山本, 阿部, 武藤, 三宅: 対話コミュニケーションにおける 2 種類の発話タイミング構造, 計測自動制御学会論文集, **45**-10, 522/529 (2009)
- 14) 武藤, 高野, 大良, 小林, 山本, 三宅: 音声インタフェースにお

ける発話タイミング制御とその評価, ヒューマンインタフェースシンポジウム 2007 講演会予稿集, 639/642 (2007)

- 15) 高杉, 吉田, 中津, 横山, 山本, 三宅: コミュニケーションロボットとの対話における反応潜時長と顔き先行時間長の影響評価, 計測自動制御学会論文集, **46**-1, 72/81 (2010)
- 16) Julius: <http://julius.sourceforge.jp/>
- 17) 高杉, 山本, 武藤, 阿部, 三宅: コミュニケーションロボットとの対話を用いた発話と身振りのタイミング機構の分析, 計測自動制御学会論文集, **45**-4, 215/223 (2009)

[著 者 紹 介]

小 林 弘 幸 (学生会員)

2013 年金沢工業大学情報学部情報工学科卒業。同年金沢工業大学大学院工学研究科情報工学専攻博士前期課程に進学し、現在に至る。主としてヒューマンインタフェース, ヒューマンコミュニケーションの研究に従事。

大 村 卓 矢

2013 年金沢工業大学情報学部情報工学科卒業。同年金沢工業大学大学院工学研究科情報工学専攻博士前期課程に進学し、現在に至る。主としてヒューマンインタフェース, ヒューマンコミュニケーションの研究に従事。

山 本 知 仁 (正会員)

2004 年東京工業大学大学院総合理工学研究科知能システム科学専攻博士課程修了 (博士 (工学))。同年, 金沢工業大学工学部情報工学科講師。2011 年より, 金沢工業大学情報学部情報工学科准教授。主としてヒューマンインタフェース, ヒューマンコミュニケーションの研究に従事。電子情報通信学会, ヒューマンインタフェース学会などの会員。