音声駆動型身体引き込みキャラクタによる対話エージェントの開発

Development of a Communication Agent by a Speech-Driven Embodied Entrainment Character

吉田 実央[1] 渡辺 富夫[2] 石井 裕[2]
Mio Yoshida[1], Tomio Watanabe[2], Yutaka Ishii[2]

[1] 岡山県立大学大学院 情報系工学研究科
[1] Graduate School of Systems Engineering, Okayama Prefectural University
[2] 岡山県立大学 情報工学部
[2] Faculty of Computer Science and Systems Engineering, Okayama Prefectural University

Abstract: In face-to-face communication, humans communicate smoothly by sharing each embodiment such as paralanguage, nodding, eye blinking, and body motions. Recently, applications which utilize speech recognition technology through a communication agent are popular. However, such a communication agent cannot react appropriately to inputted voice by a user, so that it is difficult for the user to talk with the agent. On the other hand, the systems which answer utterance contents by natural language processing are developed. We have developed a speech-driven embodied entrainment CG character called “InterActor” which automatically generates communicative motions and actions such as nods for entrained interaction from voice rhythm based on only speech input. In this paper, we develop a communication agent that generates smooth communication behaviors by InterActor for promoting interaction between a speaker and a voice communication agent.

1 はじめに

対面コミュニケーションにおいては，単に言葉によるバーバル情報だけでなく，音声の周辺言語やうなずき，まばたき，表情，身振り・手振りなどの身体動作といった言葉によらないノンバーバル情報が話し手と聞き手で相互に引き込み，対話者相互に関係を成立させ，円滑にコミュニケーションしている[1]. 今日，音声認識技術を用いた対話エージェントが一般的に利用されている．しかし，現在使用されている音声認識アプリの多くは，テキスト情報と音声情報でコミュニケーションを行うため，身振りやうなずきといった身体性が活かされておらず，発話者と対話エージェントのかかわりを視覚的に把握しづらく，話しかけづらい．

一方で，自然言語処理によって発話内容に適切な応答を行うシステム開発も進められ[2]，会話応答システムが円滑なコミュニケーション動作を行うことで，より自然なインタラクションが実現できると考えられる．

著者らは，これまでに会話音声と身体動作の引き込み効果に着目し，発話音声からコミュニケーション動作を自動生成する音声駆動型身体引き込みキャラクタ InterActor を開発している[3]．また，音声合成を用いてテキストを読み上げ，テキスト情報から言葉の意味に対応した情動表現を自動生成するメッセンジャーシステムを開発し，コミュニケーション支援に有効であることを示してきた[4].

本研究では，音声認識により発話音声をテキスト化し，そのテキストに対して雑談対話を生成する機能を用いて，音声合成による音声対話エージェントを構築する．この音声対話エージェントとのインタラクションを促進させることを目的とし，音声から円滑なコミュニケーション動作を自動生成する対話エージェントシステムを開発している．

2 音声対話エージェント

2.1 コンセプト

本システムのコンセプトを図 1 に示す．本システムでは，対話エージェントがノンバーバル動作を提示し，話者に話を聞いているという安心感を与えることで会話が弾み，さらに発話者と対話エージェン

トが身体性を共有することでインタラクションが促進され，自然な会話を可能にする．対話エージェントの利用場面としては，パートナーロボットのような話し相手や生活支援システム，受付ロボットなどの案内役など様々な状況が考えられる．ユーザが求める要求に対してより的確な情報提供を行うためには，ネットワーク接続されたシステムやデータベースなどから情報を取得し，ユーザへ提供する方法がある．その際，システムとの接続時のネットワーク遅延や応答遅延によって，ユーザとエージェントのインタラクションが損なわれ，自然な会話が成り立たない．本システムでは，システムおよびユーザの状態に基づいて音声対話エージェントが身体的リズム同調による動作生成により，円滑なインタラクションを実現する．

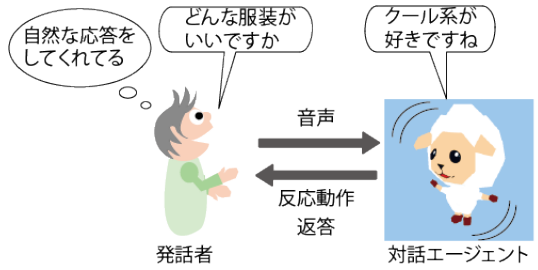


図 1：コンセプト

2.2 音声対話システム

本研究では，音声対話システム構築のために，(株) NTT ドコモ社が提供している API (Application Programming Interface) [5]を使用する．使用した API は音声認識 API ，雑談対話 API ，音声合成 API の 3 種である．

システム内部の API の動作を図 2 に示す．システムを起動し，発話者が通信端末に向けて発話した際，その発話音声を音声認識 API により音声 PCM データとして音声認識サーバに送信する．音声認識サーバを通して音声データがテキストデータとして端末へ送信されるので，通信端末画面上に発話者の発話内容をテキスト表示する．次に，通信端末へ送信されたテキストデータを雑談対話 API へ送ることにより，雑談対話サーバから送信されたテキストデータに対応した返答内容のテキストデータを通信端末が受信する．通信端末へ送信されたテキストデータが音声合成 API を介することにより，テキストデータから生成された音声 PCM データを受信した後，合成音声を通信端末から出力する．

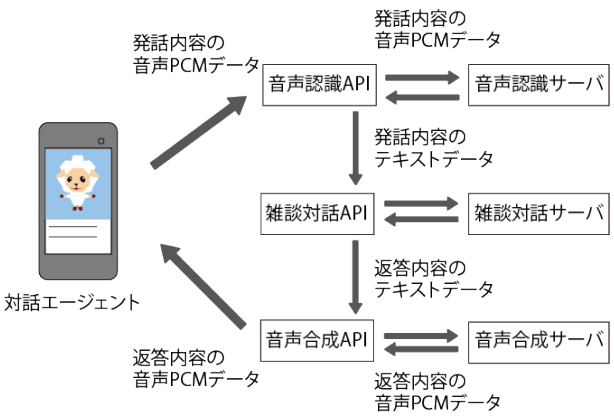


図 2：API によるシステムの動作

2.3 対話エージェント動作

対話エージェント動作の生成には，著者らがこれまでに人の対面コミュニケーション時の身体的リズムの引き込み現象に着目して開発した，会話音声を入力としてキャラクタの豊かなコミュニケーション動作を自動生成するインタロボット技術 iRT（InterRobotTechnology）[3]を使用する．iRT はコミュニケーション時の発話音声と身体動作との関係をモデル化することで発話音声からコミュニケーション動作を自動生成し，身体リズムの引き込みによりインタラクションを円滑にして，コミュニケーションを支援する技術である．

聞き手動作の場合，音声の ON-OFF パターンに基づくうなずき反応モデルと，腕部および上体部に対してうなずきの予測値に基づく身体動作モデルが導入されている．聞き手動作の予測モデルを図 3 に示す．うなずきの予測モデルはマクロ層とミクロ層からなる階層モデルである．マクロ層では音声の呼気段落区分での ON-OFF 区間からなるユニット区間にうなずきの開始が存在するかを[i - 1]ユニット以前のユニット時間率$R(i)$（ユニット時間区間での ON 区間の占める割合，式（2））の線形結合で表される式（1）の MA（Moving-Average）モデルを用いる．予測値$M_u(i)$がある閾値を超えて，うなずきが存在すると予測された場合は，処理はミクロ層に移る．ミクロ層では音声の ON-OFF データを入力とし，式（3）MA のモデルでうなずきの開始時点を推定する．

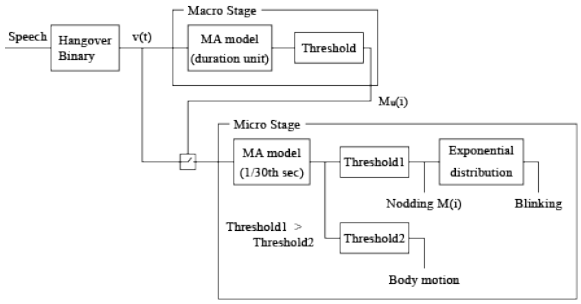
$$M_u(i) = \sum_{j=1}^{J} a(j)R(i-j) + u(i) \quad (1)$$

$$R(i) = \frac{T(i)}{T(i)+S(i)} \quad (2)$$

$a(j)$: 予測係数
$T(i)$: i 番目ユニットでの ON 区間
$S(i)$: i 番目ユニットでの OFF 区間
$u(i)$: 雑音

$$M(i) = \sum_{j=1}^{K} b(j)V(i-j) + w(i) \qquad (3)$$

$b(j)$: 予測係数
$V(i)$: 音声データ
$w(i)$: 雑音



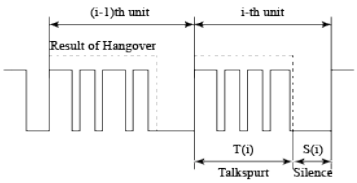


図 3：聞き手のインタラクションモデル

話し手動作の場合も同様に，音声の ON-OFF パターンに基づく身体全体の動作を予測するモデルと音声の振幅に基づく腕部動作モデルが導入されている．

2.4 プロトタイプ構築

音声対話システムおよび CG キャラクタを用いたシステムのプロトタイプ画面を図 4 に示す．システムの構成には通信端末（スマートフォン，タブレット），Wi-Fi ネットワーク環境を使用する．発話者の音声の大きさを青い波形で，音声認識を判定する閾値を上下の赤いラインで示している（図 4A）．赤いラインはオプションから呼び出せるポインタを動かすことにより上下の間隔を調整でき，音声入力を認識する閾値を変化させることができる．発話者が通信端末に向けて発話した際，通信端末に発話者の発話内容をテキスト表示する．その後，対話エージェントの返答内容をテキスト表示し，それと同時に返答内容を読み上げる合成音声を通信端末から出力する．通信端末の画面上部には音声対話エージェントを配置している．エージェントは横方向に 360° 回転させることができる．対話エージェントは発話者の発言に対する聞き手動作や合成音声の再生に合わせ話し手動作を行う．iRT を用いたコミュニケーション動作により発話者とのインタラクション，および発話意欲の促進が期待される．

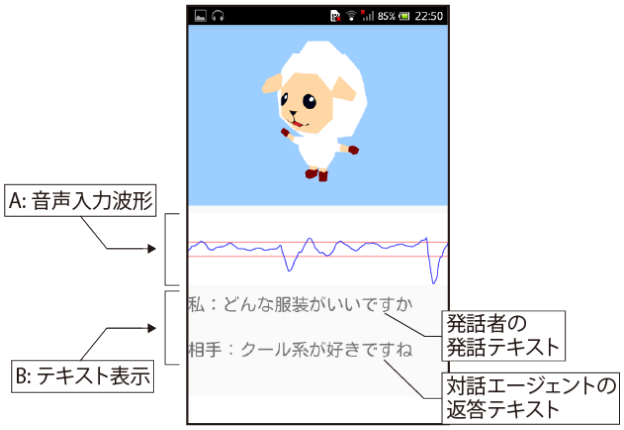


図 4：システムの使用画面

3 まとめ

本研究では，音声対話エージェントとのインタラクションを促進させることを目的とし，音声から円滑なコミュニケーション動作を自動生成する対話エージェントを開発した．

今後は，発話タイミングを把握するため，システム処理状況に応じた動作の検討，および発話動作提示タイミングの検討を行う予定である．

参考文献

[1] 渡辺富夫: コミュニケーションにおける身体性，ヒューマンインタフェース学会誌, Vol. 1, No. 2, pp. 14-18, (1999)

[2] 東中竜一郎，船越孝太郎，荒木雅弘[他]，塚原裕史，小林優佳，水上雅博: テキストチャットを用いた雑談対話コーパスの構築と対話破談の分析，自然言語処理, Vol. 23, No. 1, pp. 59-86, (2016)

[3] 渡辺富夫，大久保雅史，中茂睦裕，檀原龍正: 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, Vol. 2, No. 2, pp. 21-19, (2000)

[4] Mizuki Kohara, Hiraku Shikata, Tomio Watanabe and Yutaka Ishii: Speech-Driven Embodied Entrainment Character System with Emotional Expressions and Motions by Speech Recognition, Proc. of the 2014 IEEE/SICE International Symposium on System Integration (SII2014), pp. 431-435, 2014-12.

[5] NTT ドコモ：
https://dev.smt.docomo.ne.jp/?p=docs.api.index
（2017/11/14）