J-001

音声駆動型身体的引き込み観客キャラクタに対話相手顔画像を合成した実映像対話システムの開発

Development of a Video Communication System with Speech-Driven Embodied Entrainment Audience Characters with Partner's Face

中山 志穂† 渡辺 富夫‡ 石井 裕‡
Shiho Nakayama Tomio Watanabe Yutaka Ishii

1. はじめに

近年，ビデオチャットなどの実映像を用いた遠隔コミュニケーションが盛んに行われている．実映像を用いた対話において，対話相手のみの表示や Picture-in-Picture を用いた表示手法では空間的に分断され，話者同士のインタラクション把握が困難になるといった問題がある．人は対面コミュニケーションにおいて，身振り手振りやうなずきといったノンバーバル情報を用いて円滑にコミュニケーションを行っており，これらノンバーバル情報に着目した様々なコミュニケーション支援の方法が検討されている[1]-[4]．著者らは，相手との身体的リズムの共有を支援するために，発話音声に基づいてうなずきなどの身体的引き込み動作を自動生成する音声駆動型身体的引き込みキャラクタ InterActor を自己の代役としてビデオ映像に重畳合成した実映像対話システム E-VChat[5]を開発してきた．ここで使用する InterActor は話し手動作，聞き手動作の両機能を有しており，画像処理により話者の頭部動作を反映させている．また，E-VChat の画面に，自己の代役だけでなく，話者に対して聞き手動作を行う複数の観客キャラクタを配置したシステムを開発し，緊張場面における対話実験により有効性を示している[6]．

本研究では，対話相手の顔画像を観客キャラクタに合成することで相手の分身であることを明確にし，話しやすい場を提供する実映像対話システムを開発している．本システムは，画像処理によるフェイストラッキングを用いて対話相手の顔の位置を検出し，顔周辺の画像を保存することでキャラクタに反映させることができる．

2. E-VChat システム

2.1 システム概要

実映像を用いたコミュニケーションにおいて，著者らはこれまでに E-VChat システムを開発している[5]．E-VChat システムでは，対話者を撮影する Web カメラはモニタ上部に設置されると考え，視線のずれを考慮して対話相手の正面映像に対して自己の代役となるキャラクタを対話相手の目線の先に存在するように重畳合成している．仮想的に対面コミュニケーション場を形成することで，話者同士のインタラクション把握を円滑にしている（図 1）．自己の代役となるキャラクタが，音声のリズムに基づいて身振り手振りやうなずきなどの話し手・聞き手動作を行うことで身体的リズムの共有を助け，自己に対して共感反応をフィードバックすることで，自らに会話意欲を促進させる．

図 1 E-VChat 使用画面

図 2 観客キャラクタ配置

また，システムのプロトタイプでは話者の頭部動作をキャラクタに連動させるため，ジャイロセンサと加速度センサを用いた頭部動作計測デバイスを使用していた．しかし，実用性や装着の煩わしさなどの問題があったため，画像処理によるフェイストラッキングを用いて顔の角度情報を検出し，自己キャラクタに話者の頭部動作を連動させることを可能にしている．頭部動作を連動させることで話者の肯定・否定といった意思を自己キャラクタに反映することができ，自由対話を想定したコミュニケーション実験によって有効性が示されている[5]．さらに，自己の代役だけでなく，話者の話に対してうなずきなどの聞き手動作を行う観客キャラクタを複数配置したシステムを開発している（図 2）．ビデオチャットを利用した面接などの緊張を要する場面において，共感反応を行う観客キャラクタが存在することにより，緊張感を緩和する効果が得られることが確認されている[6]．

2.2 音声に基づくキャラクタ動作モデル

E-VChat で使用するキャラクタには，入力された音声の ON-OFF パターンから，うなずきやコミュニケーション動作のタイミングを推定し，話し手動作や聞き手動作を行わせる．音声データは 16bit 22.050kHz でサンプリングし，閾値で二値化するとともに，音節間の短時間の無音区間

† 岡山県立大学大学院 情報系工学研究科
‡ 岡山県立大学 情報工学部

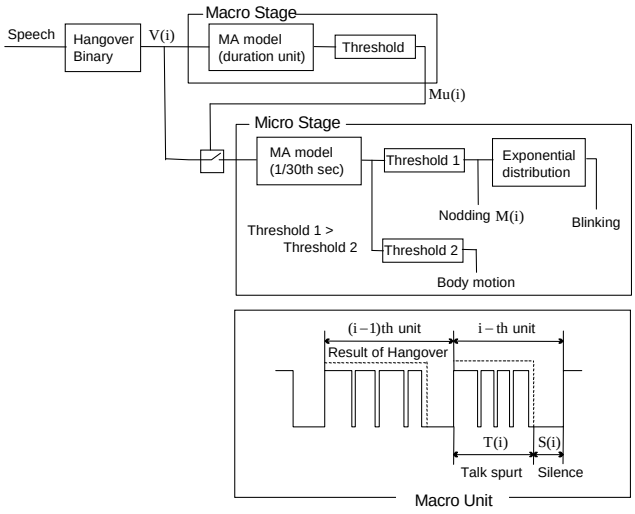

図 3　発話音声に基づく動作生成モデル

による発話の断片化を除去するために 133msec でハングオーバ処理を施している．

聞き手動作においては，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（図 3）を導入している[7]．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)（ユニット区間での ON 区間の占める割合，(2)式）の線形合成で表される(1)式の MA(Moving Average)モデルを用いて予測する．予測値 Mu(i)がある閾値を越えて，うなずきが存在すると予測された場合には，処理はミクロ層に移る．ミクロ層では音声の ON-OFF データ（30Hz，60 個）を入力とし，(3)式を用いて MA モデルでうなずきの開始時点を推定する．予測値が閾値を越えた場合にはキャラクタをうなずかせる．身体動作についてもこの予測値を用い，うなずきよりも低い閾値で各部位（頭部，胴部，右肘，左肘）のうち，いずれかを選択して動作させることでうなずきと関連付けている．

また，話し手動作においては，話者自身の発話の ON-OFF パターンに基づいて，頭部動作およびうなずき，身振り手振りといった身体動作を行わせることで，発話音声と関係付けた．

これらによる対話リズムの生成により，話者自身に会話意欲の促進を働きかけ，コミュニケーション場の生成を支援する．

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

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

$T(i)$：i 番目のユニットでの ON 区間

$S(i)$：i 番目のユニットでの OFF 区間

$u(i)$：ノイズ

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

$b(j)$：予測係数

$V(i)$：音声データ

$w(i)$：ノイズ

2.3　キャラクタ重畳合成手法

実映像を用いた対話において，Web カメラをモニタ上部に設置する場合，カメラは話者を見下ろす形となり，画面を注視している話者と視線のずれが生じてしまう．そこで，画面下部中央に自己の代役となるキャラクタを配置することで，正面から撮影される対話相手が自己の代役と対面しているように見える．対話相手と自己の代役を仮想的に対面合成させることで，対話相手は自己の代役を見ながら対話しているように見え，視線のずれによって生じる違和感を軽減することができる．また，相手の視線の先に自己の代役を配置することで，対話相手と話者の関係性を視覚的に把握でき，お互いのインタラクション把握が容易になり円滑なコミュニケーションを行うことができる．

3.　対話相手顔画像合成型観客キャラクタによる E-VChat システム

3.1　コンセプト

初対面同士のような緊張を要する場面では，相手の実映像に対して話者の話に聞き手動作を行う複数の観客キャラクタを配置することで緊張が緩和され，円滑なコミュニケーション支援に有効である[6]．

システムのコンセプトを図 4 に示す．本研究では，映像から顔の位置を検出し，周りに配置した観客キャラクタに対話相手の顔画像を合成した実映像対話システムを提案する．顔画像を合成することにより，観客キャラクタが対話相手の分身であると知覚でき，単なるキャラクタではなく，相手の反応の一部と捉えることができる．観客キャラクタは話者の発話音声と 2.2 節の身体動作モデルに基づいて，うなずきや身振り手振りといった「聞き手」の身体的引き込み動作を行う．また，自己の代役となるキャラクタが配置されていることで，対話相手とのインタラクション把握を容易にすることができる．

対話相手の反応が薄い場合においても，分身であるキャラクタが身体的引き込み動作を行うことで対話相手が同調しているように感じることができ，話しやすい場が提供される．また，様々な理由により自身の反応を表現できない場合，本システムにより，相手の分身としての

観客キャラクタが聞き上手な引き込み反応を行うことで，しっかり話を聞いているという印象を与えることができる．

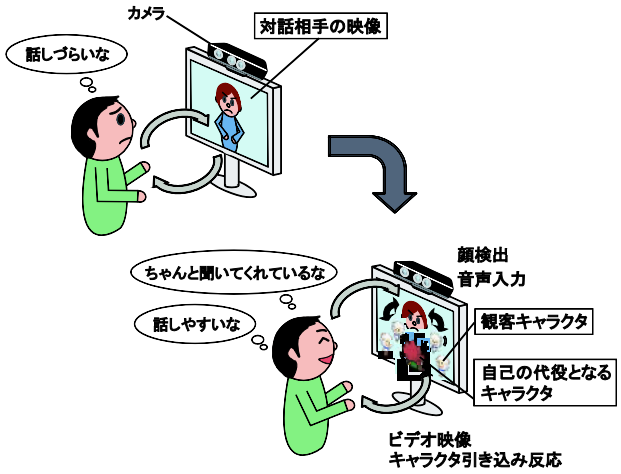

図 4　コンセプト

3.2　システム構成

システム構成における対話者間は 1GB/s のイーサネットで接続されており，音声通信を行う．映像は Kinect for Windows(L6M-00005)[8]で撮影し，パソコンに USB 接続している．また，対話相手の顔検出も Kinect センサを用いて行う．

モニタ画面上には対話相手の映像とともに，自己の代役となるキャラクタを対話相手の目線の先に配置する．自己キャラクタは話者の頭部動作を画像処理により連動させるとともに，自己の発話によって話し手動作を行い，相手の発話に対して聞き手動作を行う．また，自己の話に聞き手動作を行う 4 体の観客キャラクタを対話相手の周りに重畳合成している．さらに，対話相手の顔の位置を検出し，顔周辺を画像として保存し観客キャラクタに合成することで，単なる話を聞いてくれるキャラクタではなく，対話相手の分身であることを明確にすることができる．

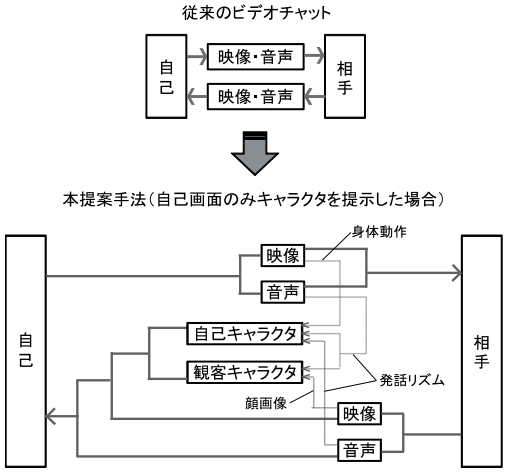

図 5　システム構成の模式図

図 5 のように，従来のビデオチャットを用いた対話の場合，お互いに音声と映像のみしか伝わらず，その 2 つの情報では，対話相手とのインタラクションを把握することが困難な場合がある．本システムを用いて対話を行う場合，音声と映像に加え，話者音声に基づいて動作するキャラクタが介在し反応することでコミュニケーション場が生成され，話しやすい場の提供や会話意欲の促進を行うことが可能となる．

3.3　対話相手顔画像合成手法

対話相手の顔検出は，Kinect のフェイストラッキングを用いて取得した顔の位置情報を用いる．フェイストラッキングにより取得した顔周りの座標の位置を利用し，キー入力により対話相手の顔周辺を画像として保存し，キャラクタの顔にテクスチャとして合成する．

顔画像合成イメージを図 6 に示す．Kinect センサを用いて取得した顔の位置情報を元に，ビデオ映像を画像として保存する．フェイストラッキングが可能な範囲には限界があり，両目・鼻・口などの特徴点をトラッキングするため，片目が隠れる角度などは，顔検出を正常に行うことができない場合がある．正常に顔検出ができなかった場合，顔画像をキャラクタに合成する際にずれが生じてしまう可能性があるため，対話相手の顔画像を取得する際には正面を向き，静止している状態である必要がある．

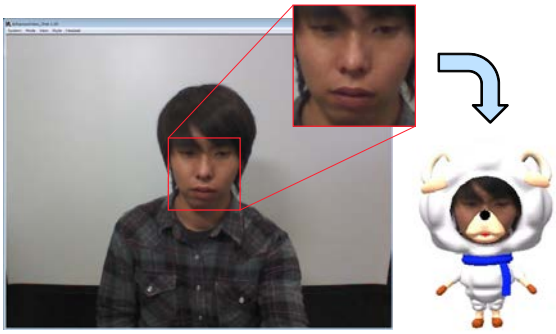

図 6　顔画像合成イメージ

図 7　システム使用画面

システムの使用画面を図 7 に示す．対話相手の目線の先に自己の代役となるキャラクタ，周りに観客キャラクタを配置する．対話相手が上司など目上の人の場合，緊張を和らげるだけでなく，対話相手の分身である観客キャラクタがうなずくことで，話しやすい場を提供することができる．また，親しい間柄の対話において，話者の話に反応する観客キャラクタが会話意欲の促進を働きかけることで対話の盛り上げ役となることが期待される．

4. おわりに

本論文では，話者の話に共感反応を行う観客キャラクタに対話相手の顔画像を合成し，話しやすい場を提供する実映像対話システムを開発した．本システムと提案手法が実映像対話に及ぼす影響などについての検討は今後の課題である．

謝辞

本研究は科学研究費（22300045，24118707）の助成を受けたものである．

参考文献

[1] 渡辺 富夫, “身体的コミュニケーション技術とその応用”, システム/制御/情報, Vol.49, No.11, pp.431-436(2005)

[2] 森川 治, 橋本 佐由理, 前迫 孝憲, “仮想的な抱擁を取り入れた遠隔カウンセリングシステム”, 日本バーチャルリアリティ学会論文誌, Vol.14, No.1, pp.3-10(2009)

[3] 守屋 悠理英, 田中 貴紘, 藤田 欣也, “ボイスチャット中の音声情報に基づく会話活性度推定方法の検討”, ヒューマンインタフェース学会論文誌, Vol.14, No.3, pp.283-292(2012)

[4] Otsuka, K., “Multimodal Conversation Scene Analysis for Understanding People’s Communicative Behaviors in Face-to-Face Meetings”, Human Interface, Part II, HCII 2011, LNCS 6772, pp.171-179(2011)

[5] Takada, T., Ishii, Y., Watanabe, T., “Development of an Embodied Video Communication System with a Superimposed Entrainment Character Driven by Voice and Head Motion Inputs”, Proc.of First International Symposium on Socially and Technically Symbiotic Systems, No.40, pp.1-4(2012)

[6] 高田 友寛, 中山 志穂, 渡辺 富夫, 石井 裕, “複数の身体的引き込みキャラクタを重畳合成した実映像対話システム”, 第 14 回 IEEE 中国支部学生シンポジウム論文集, pp.515-516(2012)

[7] Watanabe, T., Okubo, M., Nakashige, M., and Dan- bara, R., “InterActor: Speech-Driven Embodied Inter- active Actor”, International Journal of Human-Computer Interaction, Vol.17, No.1, pp.43-60 (2004)

[8] Kinect for windows SDK from Microsoft Reserch. http://kinectforwindows.org/