就職模擬面接による身体引き込み観客キャラクタを用いた実映像対話システムの評価

國田　栞[*1]　　石井　裕[*2]　　渡辺　富夫[*2]

Evaluation of a Video communication System using Embodied Entrainment Audience Characters by a Simurated Job Interview

Shiori Kunita[*1], Yutaka Ishii[*2] and Tomio Watanabe[*2]

Abstract –
An embodied video communication system, called 'E-VChat', was developed for enhancing remote communication using a video image. This system is constructed by the partner's video image using interactive CG characters including user's own avatar and virtual audience characters. We have confirmed the effectiveness of the system by an experiment under free conversation. Furthermore, we developed an embodied entrainment audience characters system with partner's face to relate audience characters with the partner. This system could bring about more active interaction by virtual audience even if the remote partner does not respond to user's speech. In addition, we confirmed the effectiveness of the system in which face images of partner's face are superimposed on audience characters under free conversation with friends. However, this system is not always evaluated effectively in all situations including nervous situations such as a job interview. In this paper, a communication experiment by a simulated job interview is performed for ten student subjects and one interviewer who works for a company, and the effectiveness of the system is investigated by the sensory evaluation.

Keywords : Video Communication, Avatar Mediated Communication, Embodied Interaction, Virtual Audience, Job Interview

1. はじめに

遠く離れた家族や友人の顔を見ながらディスプレイを介して対話できる実映像対話システムは非常に有用であり，ネットワークの拡大とともに用途を増し，テレビ会議や WEB 面接などが普及しつつある．昨今の急速なスマートフォン端末の普及により誰でも手軽に動画撮影を行えるようになっており，Skype 等のビデオチャットアプリケーションをより手軽に扱うことを可能としている．

一方で遠隔コミュニケーションにおける従来の相手映像のみを使用したインタフェースは，相手のノンバーバル情報を映像から得ることができるという利点があるものの，対話相手との視線の不一致や背景の違いによる分離感などの空間的な問題によって，対話相手との身体的なかかわりを実感することが難しい場合があり，様々な取り組みがなされている [1],[2].

著者らは，話者の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 (Enhanced-Video Chat) を開発した [3]．また，E-VChat を用いた友人同士のコミュニケーション評価実験において，対話相手の顔画像を観客キャラクタに重畳合成したシステムの有効性が示されている [4].

前述のとおり，オンラインでの面接や会議など緊張感のある環境下での利用も想定されるため，友人や家族とのリラックスしたコミュニケーションと同様の効果が得られるのか検討が必要である．そこで本研究では，実際の遠隔面接のように，初対面同士で緊張を伴う状況下でのシステムの利用可能性を検討する．実際の就職活動における面接を想定した模擬面接による評価実験を行い，緊張時におけるシステムの有用性を評価している．

2. システム概要

2.1 コンセプト

システムのコンセプトを図 1 に示す．本研究ではビデオ映像に重畳合成した複数の観客キャラクタに，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．

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

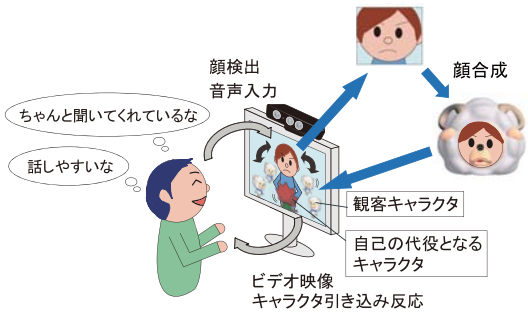

図 1　システムのコンセプト
Fig. 1　Concept.

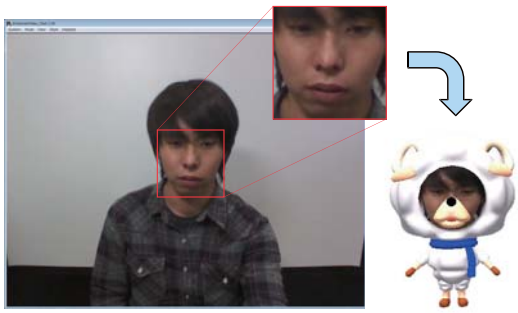

図 2　顔合成キャラクタ
Fig. 2　Audience character with the partner's face image.

観客キャラクタは話者の発話音声を入力とする身体動作生成モデルに基づいてうなずきなどの聞き手動作を行う．また，自己の代役となるキャラクタが対話相手の目線の先に配置されていることで，お互いのインタラクションの把握を支援することができる．

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に出すことが困難な場合，本システムを対話相手に使用させることにより，自身の分身である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可能となる．

2.2　システム構成

対話者間は1Gbps 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) で撮影し，パソコンに USB 接続している．対話相手の顔検出は Kinect センサを用いて行い，検出した顔を画像として保存し観客キャラクタに合成することで，単なる話を聞く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（図 2）．

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，自己の発話に基づい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，自己の発話に基づいて聞き手動作を行う 4 体の観客キャラクタを対話相手の周りに配置している．

2.3　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 [4]．聞き手動作は，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を導入している．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，(1) 式）の線形結合で表される (2) 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 InterActor をうなずかせる．

$$R(i) = \frac{T(i)}{T(i) + S(i)} \tag{1}$$

$$M_u(i) = \sum_{j=1}^{J} a(j)\, R(i-j) + u(i) \tag{2}$$

$a(j)$: 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

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

$b(j)$: 予測係数
$V(i)$: 音声データ
$w(i)$: 雑音

3.　評価実験

3.1　実験概要

システムを用いて模擬面接形式で評価実験を行った．実験風景を図 3 に示す．実験では以下に示す A〜C の 3 つのモードを用意し，それぞれのモードで対話を行

図 3　実験風景
Fig. 3　Experimental scene.

図 4　3 種類のモード（応募者の使用画面）
Fig. 4　Three types of modes.

うことで評価を行った．A〜C それぞれのシステム画面を図 4 に示す．

A　相手映像のみ表示

B　相手映像に自己キャラクタと観客キャラクタを配置

C　相手映像に自己キャラクタと対話相手の顔画像を合成した観客キャラクタを配置

実験の手順を図 5 のタイムテーブルに示す．実験協力者には最初に各モードの違いの説明を行った．その後 9 分間の面接を行い，3 分ごとに室外から実験者がモードの切り替えを行った．モード切替時に面接の中断等は行っていない．

次に，3 つのモードについての一対比較実験を行った．まず，A〜C のモードのうち 2 つのモードを用いて各モードを使用させた後，一対比較によって総合的に良かった方を選択させた．モードの提示順序はカウンターバランスをとり，3 モードの比較で 3 ($=_3 C_2$) 回行った．

その後，身体的コミュニケーション支援の観点から，応募者役側に「楽しさ」,「対話しやすさ」,「一体感」,「安心感」,「場の盛り上がり」,「相手の反応の良さ」,「好み」,「システムを使用したいか」の 8 項目について 7 段階評価を行った．

本実験では，就職活動での緊張環境に近づけるため，面接官役は実際の企業の人事担当（40 代男性）の方が行った．応募者役は 19 〜22 歳の男女学生計 10 名である．

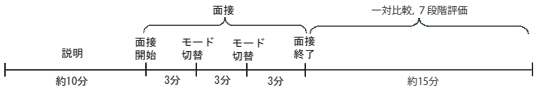

図 5　タイムテーブル
Fig. 5　Timetable.

表 1　一対比較の結果
Table 1　Result of paired comparison.

	A	B	C	total
A		5	6	11
B	5		8	13
C	4	2		6

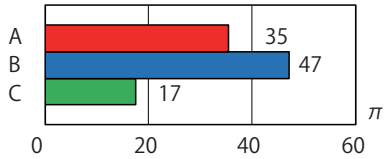

図 6　Bradley-terry モデルによる好ましさ π
Fig. 6　Preference π based on the Bradley-Terry model.

3.2　実験結果

一対比較の結果を表 1 に示す．また，表 1 から，Bradley-Terry モデルを想定して好ましさ π を最ゆう推定した結果を図 6 に示す．観客キャラクタを複数体配置する B が最も高く評価され，顔合成したキャラクタを複数体配置する C が最も低く評価された．

7 段階評価の結果を図 7 に示す．Friedman 検定において,「楽しさ」「一体感」「システムを使用したいか」の項目に有意水準 5 %で有意差が認められた．さらに多重比較として Wilcoxon 符号順位検定を行った結果,「楽しさ」の項目に AB，AC 間に有意水準 5 %で有意差が認められ，A に対して B，C が高く評価された．「一体感」の項目に BC 間に有意水準 5 %で有意差が認められ，B が高く評価された．また,「システムを使用したいか」の項目に AB 間，BC 間に有意水準 5 %で有意差が認められ，いずれも B が高く評価された．

4.　考察

一対比較において，B の相手映像に自己キャラクタと観客キャラクタを配置するモードが高く評価され，C の相手映像に自己キャラクタと対話相手の顔画像を合成した観客キャラクタを配置するモードが最も低く評価された．また，7 段階評価においても，B の相手映像に自己キャラクタと観客キャラクタを配置するモードが最も高く評価された．B モードが高く評価されたことは，自由記述においても「面接という緊張し

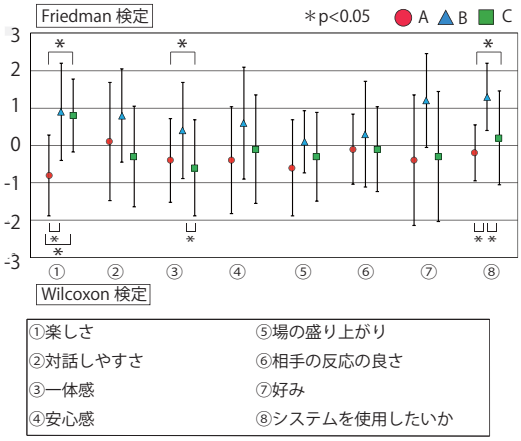

図 7　7 段階評価の結果
Fig. 7　Result of seven points bipolar rating.

た場面でキャラクタが目に入ると，安心して，少し和む．かわいい.」という意見が得られ，対話相手の周りに聞き込み動作を行うキャラクタを配置することで，安心感が出て話しやすくなることが示された．それに対し C モードについては,「楽しさ」の項目で B モードと同程度に高く評価されているものの，他の項目では A モードと同等の評価となっており,「一体感」および「システムを使用したいか」の項目では B モードに対して有意差が認められるなど，友人同士を対象とした先行研究における実験結果とは異なる結果が得られた．また自由記述においても「相手の顔が多くなり，緊張が強くなった」という意見が得られた．よって，C モードは観客キャラクタに相手顔画像が合成されることにより面接官の圧迫感が増し，重圧を感じたため評価を下げたと考えられる．

5.　おわりに

本研究では，実際の遠隔面接のように，初対面同士で緊張を伴う状況下でのシステムの利用可能性を検討した．実際の就職活動における面接を想定した模擬面接による評価実験を 3 つのモードを用意して行い，緊張時におけるシステムの有用性を評価した．評価実験の結果，キャラクタを複数体配置することで，緊張緩和の効果が見られたが，対話相手の顔画像を合成した場合には楽しさは感じるものの，十分に観客キャラクタの効果を生かせず，評価が下がることが確認された．

参考文献

[1] Sellen,A.J.:Assessing Video-Mediated Communication: A Comparion of Different Analytic Approaches; Video-Mediated Communication,LEA, pp.95–106(1997).
[2] Kawanobe, H., Aosaki, Y, Kuzuoka, H., and Suzuki, Y.: iRIS: A Remote Surrogate for Mutual Reference, Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction, p.403 (2013).
[3] 石井，渡辺: 相手顔画像を合成した身体的引き込み観客キャラクタを用いた実映像対話システム，ヒューマンインタフェース学会論文誌，Vol.17,No.3,pp265–273(2015).
[4] 石井，中山，渡辺: 非接触計測による自己キャラクタ対面合成型実映像対話システム E-VChat の評価，ヒューマンインタフェース学会論文誌，Vol.16,No.3, pp.179–186(2014).