ヒューマンインタフェース　97－14
音声言語情報処理　　40－14
（２００２．２．１）

擬人化音声対話エージェントのための表情合成技術

四倉　達夫　　森島　繁生

成蹊大学工学部

機械と人間とのコミュニケーション形態の１つとして擬人化エージェントが挙げられる。このエージェントがコンピュータディスプレイ上に表示し、言語情報やジェスチャ、表情等の非言語情報を理解・表出しあたかも人間同士が対面対話するようなリアルなコミュニケーション環境を構築可能なシステムが求められる。エージェントを構築するにあたり、最終的な目標として、いかにエージェント自体をリアルなものとし、コミュニケーションの際、現実世界との対話と遜色なくすることである。本稿ではこのシステムのエージェントの構築技術を紹介し、エージェントの顔モデル構築、表情合成、アニメーション手法について紹介する。

Facial Expression Synthesis for Life-like Spoken Communication Agent

Tatsuo YOTSUKURA　Shigeo MORISHIMA

Faculty of Engineering, Seikei University
{yotsu, shigeo}@ee.seikei.ac.jp

The multi-modal communication between man and machine style is to have a virtual human or an avatar appearing on the computer terminal that should be able to understand and express not only linguistic information but also non-verbal information. This is similar to human-to-human communication with a face-to-face style. Very important factor to make an avatar look believable or alive depends on how well an avatar can duplicate a real human's expression and impression on a face precisely. Especially in case of communication application using avatar, a real-time processing with low delay is inevitable. In this paper, we describe a current situation of our face image synthesis technology.

1. はじめに

人間と人間との対面対話は、言語情報のみならず手振りや首の動作等のジェスチャや顔表情等の非言語情報を含めた情報を用いてコミュニケーションを行う。擬人化エージェントを構築する際にもこれら情報は円滑なコミュニケーション時には欠かせないものとなっている。また、エージェント自身も実際の顔と遜色ないリアルな顔モデルを構築する必要があり、表情を付加させるためのルール付けもまた重要な要素となる。著者らの研究の最終目標は現実世界の臨場感を機械と人間とのインタフェースにおいても実現し、合成された顔によるノンバーバルなコミュニケーション環境の構築である。

このような目標に向けてのファーストステップとしてエージェントの顔モデルをある人物の顔のみならず表情や印象も含めコピーし、再現することである。そこで本稿ではこの目的を解決するためのエージェントの構築方法をを述べ、顔モデルの構築手法や表情・唇動作のルール化、アニメーション手法を述べていく。

本研究は擬人化音声対話エージェントの１つのモジュールとして機能する。音声対話エージェントは、音声認識、音声合成、対話制御、顔画像合成と統合・制御モジュールで構成されておるが、このいずれにおいても、クオリティという点で妥協は許されず、このすべてのバランスのよいクオリティが実現されてはじめてリアルな擬人化対話エージェントが実現される。

2. 顔モデルの生成

まず、エージェントの顔モデルの生成手法について述べる。本モジュールでは不特定多数の利用者を想定しているため、レンジスキャナ等の高価な装置を使用することなく、短時間で容易にモデルを作る方法を必要である。この方法としてさまざまな手法が存在する[1][2]が予め用意した標準顔モデルを変形し、顔画像にフィットさせて個々の幾何モデルを生成するのが一般的である。標準顔モデルは3次元で構成され、ワイヤフレームモデルと呼ばれる多面体近似モデルである。この3次元モデルは約800ポリゴンで形成され、さまざまな顔部位の細かい動きが再現されるようになっている。図1にワイヤフレームモデルを示す。

顔画像と標準顔モデルとを整合する方法としてまず、正面画像と側面画像の2つの顔画像を用意する。整合する際にはGUIツールを用いて、正面画像に関しては顔の部位毎に複数の格子点を一度に動かすことで顔のアウトラインや目、鼻、口などの顔を構成する部位を大まかに整合し、その後、格子点1つ1つを個々に動かして細部を整合する。側面画像はモデルの奥行き方向をセットするために用い、格子点を動かすことで整合を行う。GUIツールによるフィッティングの様子を図2に示す。なおこのツールの機能が限定されたものが現在フリーソフトとして利用可能でソースも公開されている[3]。

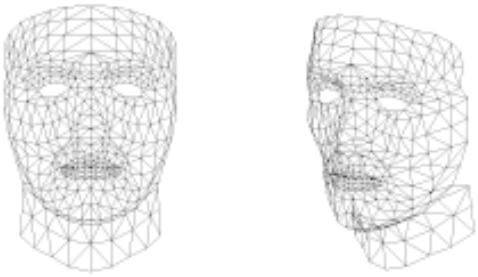


図1　標準ワイヤフレームモデル

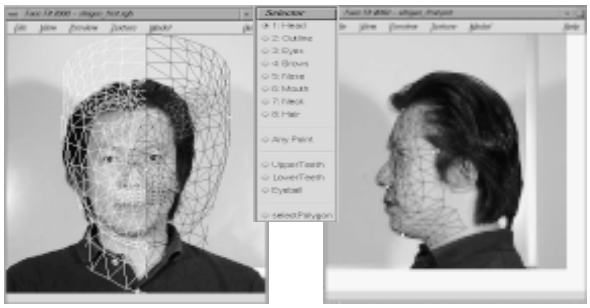


図2　フィッティングツール

表1　Action Unit　(一例)

AU No.		
AU1	Inner brow raiser	
AU2	Outer brow raiser	
AU4	Brow lower	
AU5	Upper lid raiser	
AU6	Cheek raiser	
AU7	Lid tightener	
AU8	Lips toward each other	
AU9	Nose wrinkler	
AU10	Upper lip raiser	
AU11	Nasolabial furrow deepener	
AU12	Lip corner puller	
AU13	Sharp lip puller	
AU14	Dimpler	
AU15	Lip corner depressor	
AU16	Lower lip depressor	

よりリアルなエージェントの生成を構築するため予め用意された歯および口内モデルを付加させることで口を開けた際リアルは表出が可能となる。また眼球については目を開けた際、フィッティングのみのテクスチャのみを使用すると不自然さが残る。従って整合したテクスチャモデルとシェーディングが施された球状のモデルとをブレンディングしたものを眼球モデルとした。このモデルは個人依存性が強いため当研究室で開発したツールを使用してマッチしたモデルを生成することが可能である。

3. 表情変形

人間の顔表情は顔の各部位の動きを組み合わせることにより表現することができる。人間の顔表情を画面内のモデルに表現させるためには顔の各部分の動きを定量的に与える表情記述規則が必要である。顔の表情変化を表現する方法としてFACS(Facial Action Coding System)[4]を導入している。これは、顔表面に現れる顔面筋の位置及び動きの方向を解剖学的に考慮した表情記述方法である。FACSは解剖学的に分類された44種類の運動単位AU(Action Unit)から成り立ちこのAUの組み合わせにより様々な表情を表現することが可能と

されている表1にAUの一例を示す。このAUの移動量および移動方向をパラメータとして3次元モデルを変形させ表情合成を行う。表情変化は3次元モデルの各格子点をAUの強さによって移動させる。

3.1. 表情の3次元計測

3次元モデルを変形するために実際の人物に対し各AUの表出を行ってもらう。計測の方法はCyberwareを用いる方法やモーションキャプチャを用いる方法があるが、撮影時間の長さや精度の問題から本稿ではレンジファインダを用いて表情編集を行う。

レンジファインダはNEC社製、Danae-Rを使用する。この機器は人物の正面からの撮影で2.5[sec]で正確な3次元形状が得ることができるため、Cyberware等の撮影と比べ被験者への負担を軽減することができる。撮影環境は成蹊大学内の研究室に暗幕を取り付け外部からの光を遮断できるようにした。被験者は1名でAUの表出経験の多い方であり、可能な限りFACSプロトコルに準拠できるよう配慮した。

レンジファインダから獲得した3次元座標と先述で述べた標準ワイヤフレームモデルとのフィッテングを行うことで格子点ごとのAUの移動量を求める。撮影したレンジファインダのデータは図3に示すとおりで、このデータをBitmap形式に変換させると同時に1画素ごとに3次元座標を対応付ける。変換した画像はフィッティングツールを用いて標準ワイヤーフレームを整合させ、ワイヤフレームの各点に重なる画素に対する座標値を用いることで計測を進める。図4左図にフィッテング結果を右図にレンジデータから生成した3次元頭部モデルを示す。

本研究では各AU単独の表情を1枚ずつレンジファインダを用いて撮影を行いその後、複数のAUを組み合わせた表情を10数パターン撮影した。そして取得した数と同数のワイヤフレームおよび3次元顔モデルを作成した。また撮影の際、ワイヤーフレームとデータとを正確に整合を行う必要があるため整合の補助としてマーカをアイライナーで図5のように描いた。撮影および計測結果を図6に示す。

図3　レンジファインダから計測されたデータ
左図：レンジデータ＋テクスチャマッピング
右図：レンジデータのみ

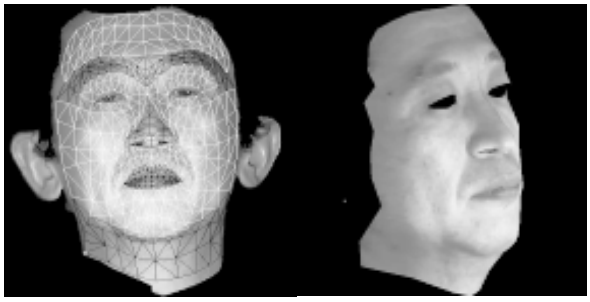


図4　整合結果

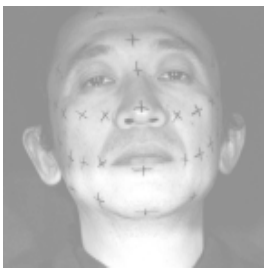
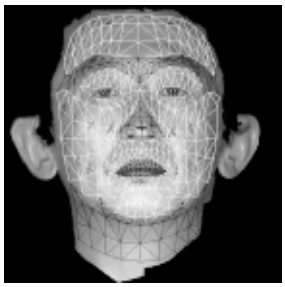
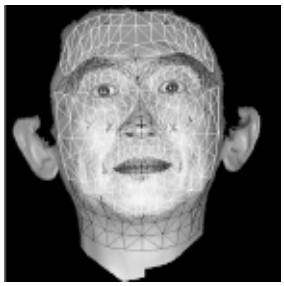


図5　被験者のマーカの位置

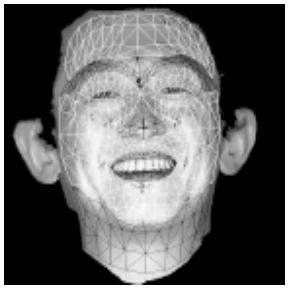


a) AU0: 無表情



b) AU5+20

b) AU1+2+4+5+20+25



c) AU6+21+26

図6 計測結果

3.2. 表情合成およびパラメータ化

作成した各AUの3次元顔モデルと無表情時のモデルの移動量の差分を求めることでAU毎のパラメータ化を図る。これらパラメータの強度や種類を組み合わせることでさまざまな表情を合成することが可能である。任意の表情を容易に製作を行えるように表情編集ツールを製作した。このツールは1つのAUにつき0%～100%まで操作可能なスライダーバーがついており、ユーザがマウスを使って操作することで簡単に表情が製作可能となっている。スライダーの100%時は対応したAUの移動差分量を割り当てている。

図8に表情編集ツールでの作成画像を2例示す。図左上が実画像、それ以外の図が合成画像である。図から分かるとおり各AUの合成画像の3次元頭部形状は実画像と比べても同様の印象が得られたと考えられる。

a)　AU17:オトガイを挙げる

b)　AU28:唇を吸い込む

図8 表情編集ツールによる表情合成例

4. 口形パラメータによる記述

発話時の口の形状を規定する口領域の変形パラメータ(以下、口形パラメータ)を表現するためにはAUとは異なる口領域の変形に限定したパラメータを用いる。パラメータ化の際、擬人化音声対エージェントの話音声認識モジュールで認識できる音素すべてをパラメータとした。これらの口形状を決定するため、図9に示すような口形編集ツールを用いてカスタマイズを行う。口形状は基本的に17個の唇の厚みと形状を表現するパラメータによって記述される。図9の個々のスライダーバーを制御することによって任意の3次元口形状を編集可能である。実際に合成される口形状が画面上からPreview可能で回転表示もできるため、奥行き方向の形状も含めて精密な編集をインタラクティブに実現できる。

図10に典型的な母音の口形を示す。唇は厚みを持ち、さらに先述した口内のモデル持っているため、リアルかつ微妙な口の形状表現が可能となっている。

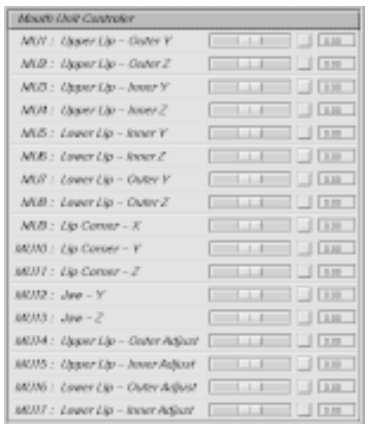
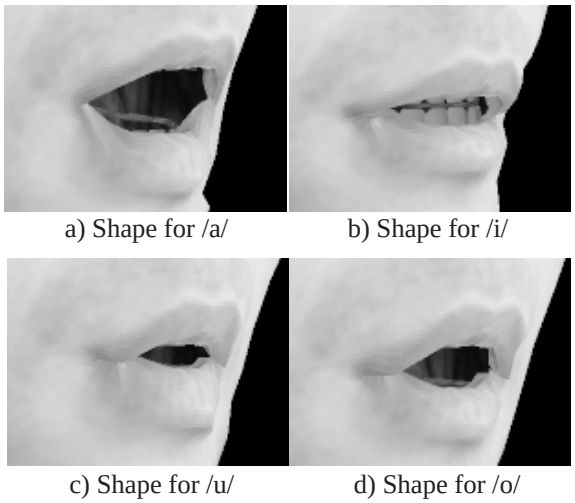


図9 口形エディタの制御パネル

5. 顔画像生成モジュール

先述で紹介した表情・口形パラメータを用いてエージェントの顔モデルの制御を行う。本章ではエージェントを円滑に動作させるための制御法や顔画像生成モジュールの機能について紹介する。



a) Shape for /a/　　b) Shape for /i/

c) Shape for /u/　　d) Shape for /o/

図10 典型的な母音口形状

5.1. 表情・口形状の制御

エージェントに対してアニメーションさせたい表情・口形パラメータは基本的に統合・制御モジュールから送信されてくる。このパラメータを用いて正確に表情をアニメーションさせる。表情に関して、現状のモジュール（以下β版モジュール）では表情の強度および表情の継続時間が処理可能となっている。表情の種類は怒り、喜び、悲しみ、嫌悪、恐れ、驚きの基本6感情が操作でき、これらの定義されたAUパラメータのデータは予め用意しておく。もちろん直接AU自体を制御することも可能である。

口形状に関しても同様に統合・制御モジュールから1文章分の音素（口形）パラメータと音素長が送信されるが、注意すべき点として合成音声との同期を考慮に入れる必要がある。この問題を解決するためにβ版モジュールでは統合・制御モジュールが音声合成、顔画像合成各モジュールに対し同時刻に相対時間での発話開始時間を送信することで解決している。対話実験の際、合成音声と口形状とが同期せず、ずれという問題は発生しなかった。

受信した口形状のパラメータをそのまま表示させると、ぎこちないアニメーションが生成させてしまう。口形状のリアルなアニメーションを表現させるため、受信した口形パラメータとその継続長に基づいて、キーフレーム位置の配置を行う。これらキーフレーム間を線形補間することで滑らかなアニメーションを生成させることが可能となる。

5.2. 頭部の制御

β版モジュールでは統合・制御モジュールから送信されたパラメータによって頭部の制御が可能となっている。また瞬きも制御でき、より自然なエージェント構築が可能となっている。

エージェントのキャラクタ変更・追加に関しては登場させたいエージェントの数だけデータを製作し、モジュール起動前にそれらデータを読み込み、要求に応じて切り替えを行う。データ生成は先述した顔モデル生成のGUIツールを用いることで簡単に構築可能である。

5.3. リアルなエージェント構築のための機能付加

β版モジュールでは先に述べた機能が含まれているが、今後の正式版モジュールは以下に述べる機能を付加させることで現状よりリアルなエージェントを構築することができると考えられる。

まず第一に現在、エージェントのモデルは顔と首のモデルのみで構成されており頭髪は考慮されていない。そのためβ版では背景に顔モデルとの整合の際に使用した正面画像を配置することで頭髪モデルを補っている。当研究室では頭髪のモデル化および運動制御に関する研究が行われている[5]。この研究を用いればエージェントの頭髪やその運動に関する問題は解決するが、演算量の問題からリアルタイム上かつ高リフレッシュレートでの運用には残念ながら難しい。したがって正式版に関しては標準ワイヤーフレームモデルに頭髪の形状のモデルを付加させ、顔整合の際と同様に頭髪もまた整合を行うことで問題を解決してゆく。

つぎに顔モデル自身のクオリティ向上を図る必要がある。例として表情変化時の皺の表現が挙げられる。β版では皺の表現は難しいが今後、モデルに皺ができる部分をMicroPolygonなどを用いてシェーディングを行う手法やPer Pixel Shadingを用いたバンプマッピングを利用する方法などを用いて動的かつリアルな皺の合成を目指す。

最後に胴体のモデルの付加を正式版モジュールに導入を検討している。エージェントを人間らしくかつ信頼性の高いものとするためにエージェント自身が自由度の高い動きを表現していく必要がある。より自然なエージェントを構築するためにも上記した機能は必要であり、随時追加してゆく。

6. まとめ

本稿では、音声対話擬人化エージェントの実現に向けた顔画像合成に関する技術について紹介した。これら紹介した技術は本システムのみならず多方面での運用が考えられる。例えば使用した口形・表情パラメータのみを使用した顔画像通信システムである。通常、動画像通信は画像自身を圧縮しそれを相手先に伝送するが、本手法を用いることで情報圧縮の限界を目指すことも可能である。この応用例として当研究室では多人数コミュニケーションシステムを構築し超低ビットレートでの運用が可能となった[6]。

実写との融合を行う研究も進めており、例えば図11に示すような主人公の顔部分を置換して表情合成を行う手法[7]やオリジナルの音声を認識し、機械翻訳を行い表情合成した声と再度リップシンクさせるビデオ翻訳の実現のため、口周辺部のみを実画像と置換させ合成する手法[8]についても検討している。

図11　顔の置き換えと表情合成

参考文献

[1] 伊藤, 三澤, 武藤, 森島, “複数アングル画像からの3次元頭部モデルの作成と表情合成”, 信学技報 , Vol.99, No582, pp7-12,2000

[2] T.kuratate,H.Yehia,E.Vatikiotis-Bateson, *Kinematics-Based Syntheses of Realistic Tracking Face*, International Conference on Auditory-Visual Speech Processing , pp.185-190, 1998

[3] 森島, 八木,”顔の認識・合成のための標準ツール”, 信学技報 , Vol.44, No.3, pp.119-126, 2000.

[4] Ekman, P. and Friesen, W.V. , *Facial Action Coding System*. Consulting Psychologists Press Inc., 1978.

[5] Kishi,K. and Morishima, S., *Dynamic Modeling of Human Hair and GUI Based Hair Style Designing System*, SIGGRAPH ‘98, Conference Abstracts and Applications, Sketches 255, 1998.

[6] 四倉, 藤井, 森島,”サイバースペース上の仮想人物による実時間対話システムの構築”, 情報処理学会論文誌 , 第40巻 , 第2号 , pp.677-686, 1999.

[7] 森島,”The Fifteen Seconds of Fame　－視聴者参加型インタラクティブ映画の提案－”, フォーラム顔学・ 8, 第3回日本顔学会大会予稿集, 1998.

[8] 緒方, 中村, 森島, “ビデオ翻訳システム - 自動翻訳合成音声とのモデルベースリップシンクの実現 -”, インタラクション’2001, pp203-210, 2001