解剖学的アプローチによる
高精細・忠実な顔面筋モデルの作成と運動制御

（課題番号　13450161）

平成 13－16 年度科学研究費補助金　基盤研究（B）（2）
研究成果報告書

平成17年 3 月

研究代表者　森島　繁生
（早稲田大学理工学部）

平成 13－16 年度科学研究費補助金　基盤研究(B)(2)
研究成果報告書

課題番号

13450161

研究課題

解剖学的アプローチによる高精細・忠実な顔面筋モデルの作成と運動制御

研究組織

研究代表者　森島　繁生(早稲田大学理工学部応用物理学科　教授)
研究分担者　島田　和幸(鹿児島大学歯学部　教授)

研究経費

平成 13 年度	5,400 千円
平成 14 年度	3,800 千円
平成 15 年度	1,300 千円
平成 16 年度	1,300 千円
計	11, 800 千円

研究発表

[学会誌等論文]

2005

[1] 前島謙宣,四倉達夫,森島繁生,中村哲
"雑音環境下における音声の聞き取り実験による合成発話アニメーションの評価"
電子情報通信学会論文誌, Vol.J88-A,No.1, pp.71-82, January, 2005

[2] 森島繁生
"フューチャーキャストシステム『三井・東芝館』"
映像情報メディア学会誌,Vol.59, No.4, pp.522(36)-524(38), April, 2005

[3] 前島謙宣、森島繁生
"Future Cast System: 三井・東芝館"
3D 映像,Vol.19,No.2, pp.45-48, June, 2005

2004

[1] Shigeo Morishima , Satoshi Nakamura
"Multimodal Translation Using Texture Mapped Lip-Sync Images"
EURASIP Journal on Applied Signal Processing, pp.1637-1647, November, 2004

2003

[1] 森島繁生
" 擬人化インタフェース, 特集記事 GUIを越えて -Beyond Desktop 特集-"
ヒューマンインタフェース学会誌,Vol.5, No.2, pp.65-68, 2003

[2] 新田恒雄、西本卓也、川本真一、下平博、森島繁生、四倉達夫、山下洋一、小林隆夫、徳田恵一、広瀬啓吉、峯松信明、山田篤、伝康晴、宇津呂武仁、伊藤克亘、甲斐充彦、李晃伸、中村哲、嵯峨山茂樹
"Galatea: 音声対話擬人化エージェント開発キット"
日本顔学会誌,Vol.3, No.1, p.189, 2003

2002

[1] 四倉達夫、Kim Binsted, Frank Nielsen, Claudio Pinhanez, 森島繁生
"HyperMask:3次元顔モデルを用いた仮面の構築"
電子情報通信学会論文誌 D-2, Vol.J85-D-2, No.1, pp.36-45, January, 2002

[2] Tatsuo Yotsukura, Shigeo Morishima, Frank Nielsen et.al.
"HyperMask - projecting a talking head onto a real object"
The Visual Computer, Vol.18, No.2, pp.111-120, March, 2002

[3] 川本真一、森島繁生、四倉達夫、嵯峨山茂樹 他
"カスタマイズ性を考慮した擬人化音声対話ソフトウェアツールキットの設計"
情報処理学会論文誌, vol.43, No.7, pp.2249-2263, July, 2002

[4] 森島繁生
"エンタテインメントのための表情分析・合成技術"
日本バーチャルリアリティ学会論文誌, Vol.7, No.4, pp.533-542, 2002

[5] 森島繁生
"HAIにおけるエージェントのリアリティとコミュニケーションギャップ"
人工知能学会誌特集HAI:ヒューマンエージェントインタラクション, Vol.17, No.6, pp.687-692, November, 2002

2001

[1] Shigeo Morishima
"Face Analysis and Synthesis"
IEEE Signal Processing Magazine, Vol.18, No.3, pp.26-34, May, 2001

[Conference Proceedins]

2005

[1] Shigeo Morishima
"Face and Gesture Cloning for Life-like Agent"
HCI International 2005, AC.UAHCI.HCI.HIMI.OCSC.VR.U&I.EPCE, 2044.pdf, July, 2005

[2] Shigeo Morishima, Akinobu Maejima, Shuhei Wemlera, Tamotsu Machida, and Masao Takebayashi
"Future Cast System"
ACM SIGGRAPH 2005 Skech, ACM SIGGRAPH 2005 Full Conference DVD-ROM Disc 2,
020-morishima.pdf, SIGGRAPH 2005, August, 2005,

[3] Tatsuo Yotsukura, Shigeo Morishima and Satoshi Nakamura
"Speech to Talking Heads System Based on Hidden Markov Models"
ACM SIGGRAPH 2005 Poster, ACM SIGGRAPH 2005 Full Conference DVD-ROM Disc2,
028-yotsukura.pdf, ACM SIGGRAPH2005, August, 2005

[4] Shin-ichi KAWAMOTO, Tatsuo YOTSUKURA, Shigeo MORISHIMA, Satoshi NAKAMURA
"Automatic Head-Movement Control for Emotional Speech"
ACM SIGGRAPH 2005 Poster, ACM SIGGRAPH 2005 Full Conference DVD-ROM Disc2,
029-kawamoto.pdf, ACM SIGGRAPH2005, August, 2005

[5] Hiroaki Yanagisawa, Akinobu Maejima, Tatsuo Yotsukura, Shigeo Morishima
"Quantitative Representation of Face Expression Using Motion Capture System"
ACM SIGGRAPH 2005 Poster, ACM SIGGRAPH 2005 Full Conference DVD-ROM Disc2,
110-yanagisawa.pdf, ACM SIGGRAPH2005, August, 2005

[6] Shohei Nishimura, Shoichiro Iwasawa, Eiji Sugisaki, Shigeo Morishima
"Reconstructing Motion using a Human Structure Model"
ACM SIGGRAPH 2005 Poster, ACM SIGGRAPH 2005 Full Conference DVD-ROM Disc2,
111-nishimura.pdf, ACM SIGGRAPH2005, August, 2005

[7] 前島謙宣、森島繁生
"フューチャーキャストシステム"
日本顔学会誌, Vol.5, No.1, P182, September, 2005

[8] 柳澤博昭、前島謙宣、四倉達夫、森島繁生
"フェイスキャプチャによる顔表情合成及び顔表情の定量表現"
日本顔学会誌, Vol.5, No.1, P184, September, 2005

2004

[1] 西本 卓也, 荒木 雅弘, 伊藤 克亘, 宇津呂 武仁, 甲斐 充彦, 河口 信夫, 河原 達也, 桂田 浩一, 小林 隆夫, 嵯峨山 茂樹, 下平 博, 伝 康晴, 徳田 恵一, 中村 哲, 新田 恒雄, 坂野 秀樹, 広瀬 啓吉, 峯松 信明, 三村 正人, 森島 繁生, 山下 洋一, 山田 篤, 四倉 達夫, 李 晃伸
"Galatea: 音声対話擬人化エージェント開発キット"
インタラクション 2004 論文集, pp.27-28, March, 2004

[2] Tatsuo Yotsukura, Shigeo Morishima and Satoshi Nakamura
"Face Expression Synthesis Based on a Facial Motion Distribution Chart"
ACM SIGGRAPH 2004 Full Conference DVD-ROM, Disc 2: Posters - Web GraphicsAugust,
085-yotsukura.pdf, 2004

2003

[1] Tatsuo Yotsukura, Shigeo Morishima and Satoshi Nakamura
"Model-based Talking Face Synthesis for Anthropomorphic Spoken Dialog Agent System "
ACM Multimedia 2003, pp.351-354 (see mm2003-yotsukura.pdf), November, 2003

[2] 四倉達夫、祖川慎治、森島繁生、中村哲
"擬人化音声対話エージェントシステムにおける顔画像合成技術"
日本顔学会誌,Vol.3,No.1, p.145, 2003

2002

[1] Shigeo Morishima
"Multi-modal Translation System using Face Tracking and Lip Synchronization"
No.45, the CD-ROM Proceedings of the Third International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science and e-Medicine on the Internet, January, 2002

[2] S.Iwasawa,T.Yotsukura,S.Morishima
"Face Analysis and Synthesis for Interactive Entertainment"
Workshop note of International Workshop on Entertainment Computing (IWEC2002) pp.143-150, May, 2002

[3] Shigeo Morishima, Shin Ogata, Kazumasa Murai, Satoshi Nakamura
"Audio-Visual Speech Translation with Automatic Lip Synchronization and Face Tracking Based on 3-D Head Model"
Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.2117-2120, May, 2002

[4] Tatuo Yotsukura, Mitsunori Takahashi, Shigeo Morishima, Kazunori Nakamura, Hirokazu Kudoh
"Magical face: Integrated Tool for Muscle Based Facial Animation Conference Abstracts and Applications"
A Publication of ACM SIGGRAPH, p212, July, 2002

[5] Shigeo MORISHIMA and Satoshi NAKAMURA
"Multi-modal Translation and Evaluation of Lip-synchronization using Noise Added Voice"
The First International Joint Conference on Autonomous Agents & Multi-Agent Systems, July, 2002

[6] Shin-ichi Kawamoto, Shigeo Morishima, Tatsuo Yotsukura , Shigeki Sagayama et.al.
"Open-source software for developing anthropomorphic spoken dialog agent"
Proceedings of the International Workshop on Lifelike Animated Agents - Tools, Affective Functions, and Applications, held in conjunction with the Seventh Pacific Rim International Conference on Artificial Intelligence (PRICAI-02), pp.64-69, August, 2002

[7] Shigeo Morishima, Satoshi Nakamura
"Multi-modal Translation System and Its Evaluation"
Proceedings of Fourth IEEE International Conference on Multimodal Interface (ICMI2002), pp.241-246, October, 2002

[8] 四倉達夫、森島繁生
"リアルな顔合成による音声対話擬人化エージェントの開発"
JAWS 2002 (エージェント合同シンポジウム)講演資料―ソフトウェアエージェントとその応用 (SAA 2002) とマルチエージェントと協調計算 (MACC 2002)ー,pp.142-143,November, 2002

[9] Tatsuo Yotsukura and Shigeo Morishima,
"An Open Source Development Tool for Anthropomorphic Dialog Agent -Face Image Synthesis and Lip Synchronization-",
Proceedings of IEEE Fifth Workshop on Multimedia Signal Processing, 03_01_05.pdf, December, 2002

2001

[1] 緒方信、中村哲、森島繁生
"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 - "
情報処理学会シンポジウム,インタラクション2001論文集, Vol.2001, No.5, pp.203-210, March, 2001

[2] Shigeo Morishima, Tatsuo Yotsukura, Frank Nielsen, Kim Binsted et.al.
"HyperMask: Face Image Synthesis Driven by Natural Voice and Projected onto Real Objects"
Proceedings of the IASTED International Conference, Artificial Intelligence and Soft Computing, pp.342-347, May, 2001

[3] Shigeo Morishima, Shin Ogata, Satoshi Nakamura
"Video Translation System using Face Tracking and Lip Synchronization"
IEEE International Conference on Multimedia and Expo, pp.856-859, P515, August, 2001

[4] Shin Ogata, Kazumasa Murai, Satoshi Nakamura, Shigeo Morishima
"Model-Based Lip Synchronization with Automatically Translated Synthetic Voice Toward a Multi-Modal Translation System"
IEEE International Conference on Multimedia and Expo, pp.29-32, August, 2001

[5] Tatsuo Yotsukura, Hideko Uchida, Hiroshi Yamada, Nobuji Tetsutani, Shigeru Akamatsu, Shigeo Morishima
"Analysis and Simulation of Facial Movements in Elicited and Posed Expressions Using High-Speed Camera"
ACM SIGGRAPH2001 Conference Abstracts and Applications, Sketch and Applications, p146, August, 2001

[6] Shin Ogata, Takafumi Misawa, Satoshi Nakamura, Shigeo Morishima et.al.
"Multi-Modal Translation System by Using Automatic Facial Image Tracking and Model-Based Lip Synchronization"
ACM SIGGRAPH2001 Conference Abstracts and Applications, Sketch and Application, p.231, August, 2001

[7] Shigeo Morishima, Tatsuo Yotsukura, Hiroshi Yamada, Hideko Uchida, Nobuji Tetsutani and Shigeru Akamatsu
"Dynamic Micro Aspects of Facial Movements in Elicited and Posed Expressions Using High-Speed Camera"
10th IEEE International Workshop on Robot and Human Interactive Communication, pp.371-376, September, 2001

[8] Shigeo Morishima, Shin Ogata, Satoshi Nakamura
"Multimodal Translation"
Proceedings of Auditory-Visual Speech Processing (AVSP2001), pp.98-103, September, 2001

[9] Shigeo Morishima, Tatuso Yotsukura, Frank Nielsen, Kim Binsted, Claudio Pinhanez
"HYPER MASK- Projecting a Virtual Face onto a Moving Real Object"
Proceedings of Eurographics, 2001, pp.305-310, September, 2001

[10] Shigeo Morishima, Tatsuo Yotsukura
"HYPERMASK: Talking Head Projected onto Moving Surface"
Proceedings of International Conference on Image Processing, Vo.3, pp.947-950, October, 2001

[11] Shin-ichi Kawamoto, Shigeki Sagayama, Shigeo Morishima, Tatsuo Yotsukura, et.al.
"Developments of Anthropomorphic Dialog Agent: A Plan and Development and its Significance"
Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue(IPNMD-2001), pp.133-137, December, 2001

目次

第一章 概要

1．1 はじめに

人間の表情の分析・合成の研究が国内外で盛んに行われている。次世代計算機インタフェースの実現形態の一つとして、擬人化エージェントが注目されており、人間どうしが向かって対話するような自然さで、表情を介した計算機とのノンバーバルな対話を実現できる可能性が期待されている。その際、顔の立体形状をワイヤフレームで表現し、その格子点を移動して幾何変換によって表情の変化を表現するものが一般的であるが、モデル依存性が強く一般的でない。一方、表情筋をバネモデルで近似し少ないパラメータで表情を表現する試みがある。しかしあくまで線形スプリングによる近似であるため表現能力には限界がある。本研究では、人間の表情筋を実態に即して忠実にモデル化して、そのダイナミックな特性をルール化し、この新しい表情筋モデルの制御およびそのモデルから得られた知見に基づいてリアリティの高い表情合成を実現する試みを行った。

まず、いくつかの表情筋の解剖結果から、表情筋の形状、表情筋の配置、その発達の状況、個人差による違いの統計を明らかにし、表情筋と表出される表情との関係を忠実にモデル化して、個人差も考慮可能な表情顔モデルを構成することを目的とした。さらに個々の表情筋に運動特性を与え、脂肪層と皮膚層を付加することによって、リアリティが高い表情アニメーションの自動生成システムを実現し、プラグインソフトウェアとして実装を行った。

さらにこの表情筋モデル構成によってもたらされた様々な知見に基づき、一般化された表情合成ツールの構築、さらにこれを利用した表情合成アプリケーションの実現、また表情合成アルゴリズムの改善手法、そして表情合成の評価方法について検討を行ったので、ここに報告する。

既にトロント大学のDr.Terzopoulosらのグループは、個々の表情筋を1本の線形スプリングで表現して、運動方式を導くことにより、表情の運動表現する手法を世界に先駆けて提案した。また、早稲田大学の橋本教授らのグループも同様のバネモデルによる表情制御法を提案しており、一般モデルを用意して個人モデルに適用させるアプローチを取る。筆者らのグループもこのバネモデルと頭蓋骨を考慮した表情筋のダイナミックなモデリングの研究を行い、表情合成を試みた。これらの研究はいずれも解剖学書の記述を参考にして、個々の表情筋に相当する 1 本のスプリングの配置を決定しており、個人差を考慮したモデル構成にはなっておらず、さらに合成される表情のリアリティにも限界が存在することが確認されている。

その原因の一つは、人間の表情筋が実際には面積と厚みを持っており、複雑に相互作用しながら表情が決まるが、1本のバネではこの微妙な変化を記述できない点にある。また表情筋の発達具合には個人差があり、特に笑筋については退化していく場合があることも明らかになっている。したがって、個々の表情筋のモデルは同等に扱うのではなく、その制御法を顔領域ごとに独自に定式化し、また発達の度合いによって制御方法を変化させることによって、個人差を考慮した表情合成を実現した点にある。

1．2　研究方法

平成13年度～平成14年度

(1)解剖による表情筋の構造の解析

人物の顔部分の解剖によって、表情筋の一つ一つを順に抽出して、その面積、容積、構造等を実際に明らかにし、静止画として、その3次元構造をコンピュータグラフィックスのモデリングツールを使ってデザインしてゆく。この処理を数名分実施して、表情筋のバリエーションについて解明してゆく。特にどの人物にも共通に存在する表情筋とその構造を明らかにしたり、また人それぞれによってバリエーションが存在したり退化したりしている表情筋はどれかを解明し、統計的に記述すること研究目的となる。

(2)表情筋の統計解析

すべての解剖データに基づいて個人個人の3次元表情筋モデルを作成する。さらに表情筋ごとにその面積や容積を計算し、個人による表情筋のバリエーションについて統計解析する。平均することによって、標準的な表情筋モデルを実現し、それぞれの表情筋によって個人のバリエーションを付加する手法について検討をおこなった。

(3)表情筋モデルの評価

表情筋モデルを3次元グラフィックスとして再現し、主観評価によってその妥当性を評価する。また、解剖前の顔の構造と比較して、表情筋の構造と顔の構造との対応関係を明らかにし、正面の顔写真から顔面筋の構造を予測する方式について検討する。また、脂肪層を表情筋モデルの上にかぶせ、さらに皮膚層を付加することによって、顔モデルを完成させた。

平成15年度～平成16年度

(1)表情変形ルールの定義

作成された表情筋モデルに関して、個々の表情筋を物理シミュレーションする方式について検討する。実際には、広がりのある表情筋はバネの集合体であると仮定し、運動方程式を解くことによって、表情筋が収縮する際の速度、加速度を計算して表情筋の運動を制御する。実際には、表情筋の体積保存則や、表情筋どうしの衝突の判定および反力の効果を考慮して、実態にできるだけ近い表情筋の運動モデルを構築した。

(2)表情合成の評価

個々の表情筋を実際に運動させて表出される表情の自然さについて主観評価を実施する。またEMGの測定データとその際の表出表情を記録し、実際に表情筋モデルをEMGで駆動して、合成される表情が観察されたものに近いかどうかを検討を行った。また、音声信号を利用した新たな客観評価手法を提案した。

(3)リアルな表情合成ツールの実現

平均的な表情筋モデルと、個々の表情筋へのバリエーションの付加を可能とする顔モデリングツールを作成する。これは撮影された正面顔画像に対してこのジェネリックな表情筋モデルをフィッティングさせ、近似的に人物の表情筋モデルを作成できるように配慮する。さらに実際に表情筋を動かしてみて、合成される表情と本人の表情を比較し、表情筋の強度の典型的な組み合わせを設定することによって、喜怒哀楽や微妙な表情の合成が可能となる。

簡単な作業で、標準顔ワイヤフレームを計測された3次元レンジデータにフィッティング可能な顔整合ツールを開発した。また、この整合された個人顔モデルに対して、表情合成したり、音声と同期してリップシンクを実現できる汎用の表情アニメーションツールを開発した。このツールは音声対話擬人化エージェント作成支援ツール(Galatea)に組み込まれている。

(4)表情合成アプリケーションの実現

まず表情合成のアプリケーションとしてビデオ翻訳システムを2001年に提案し、音声認識・翻訳・合成と、本表情合成アルゴリズムを結合して実現した。つぎに音声に同期して、仮面上に投影された表情合成画像がリアルタイムに表情変形を行うHyperMaskを実現し、舞台演技のツールとして提案した。また2005年の愛・地球博では、フューチャーキャストシステムとして、観客参加型のエンタテインメントシステムを提案し、そのキャストの表情合成手法に今回の表情合成アルゴリズムの一部が反映された。

第2章　研究成果

2．1　　表情筋モデルに基づく表情合成ツールの提案

図1に、トロント大学のDemetri Terzopoulosの表情筋モデルを示す。これは複数の線形スプリングによって表情変形を記述するモデルである。一方、我々の提案する表情筋モデルを図2に示す。これは個々の表情筋が、独自の形状とボリュームを持つもので、実体に即した表情合成を実現している。

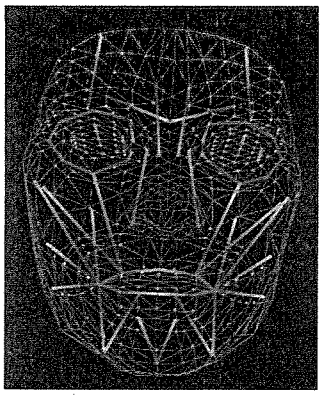


図1 トロント大学の表情筋モデル

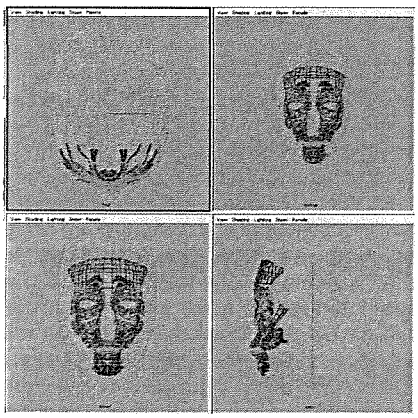


図2 今回提案する新しい表情筋モデル

この表情筋を変形することで、表情合成が可能となる。表情合成ルールはこの一般的に定義された「ジェネリックモデル」に対してルール化されているが、ユーザが任意に定義した顔モデルに対しては、このジェネリックモデルをターゲットに整合することで、ユーザ定義モデル用の表情変形ルールを近似的に生成する。

図3は、無表情の表情筋と怒りの表情筋変形の様子を示している。表情変形は、個々の表情筋の引っ張り強度をスライダーバーで調整することによって、特定の表情合成を行う。図4に表情合成ツールのGUI操作画面を示す。

図5は、ユーザ定義モデルへのジェネリックモデルの整合の様子を示す。この整合処理により、一般的に定義された表情筋モデルがユーザ定義モデル用にカスタマイズ可能となる。

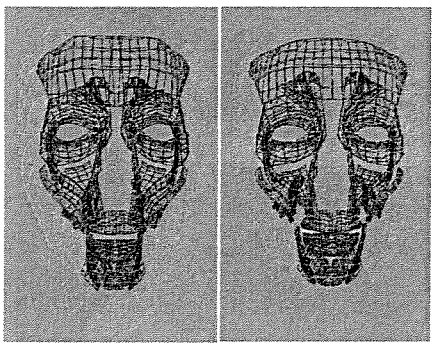


図3　表情筋モデルによる表情変形
無表情（左）　怒り（右）

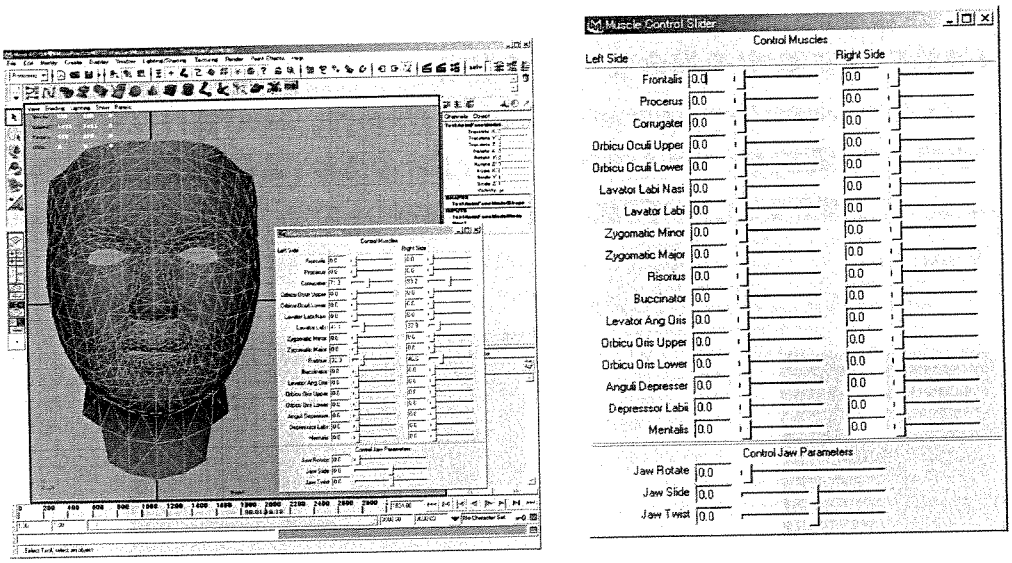


図4　表情合成ツールのGUI操作画面

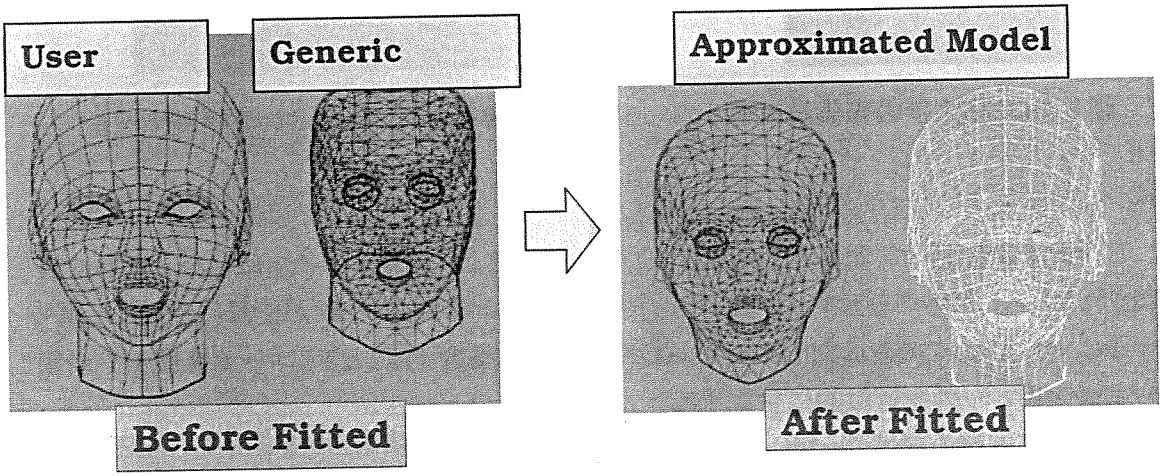


図5　ユーザ定義モデルへのジェネリックモデルの整合

Magical face: Integrated Tool for Muscle Based Facial Animation

Tatsuo Yotsukura
Seikei University/SEGA Corporation
3-3-1 Kichijo-ji Kitamachi, Musashino-shi
Tokyo, JAPAN
yotsu@ee.seikei.ac.jp

Mitsunori Takahashi
Shigeo Morishima
Seikei University
Kazunori Nakamura
Hirokazu Kudoh
SEGA Corporation

Introduction

In recent years, tremendous advances have been achieved in the 3D computer graphics used in the entertainment industry, and in the semiconductor technologies used to fabricate graphics chips and CPUs. However, although good reproduction of facial expressions is possible through 3D CG, the creation of realistic expressions and mouth motion is not a simple task.

In fact, producing such effects requires a tremendous amount of time and effort on the part of the CG creator because there has been no tool that allowed automation in the initial creative process or the reuse of high-quality facial expressions. As a result, the reproduction of facial expressions is an expensive process and the quality of the finished product depends on the skill level of the creators.

To solve these problems, we developed Magical Face, an integrated tool that can be used for facial animation by users of any level. This tool consists of three sub-tools - a facial expression editing tool that uses an anatomical muscle model, - a face model fitting tool, and lip-sync animation tool that is based on the user's voice. These operate through Maya plug-ins (Alias-Wavefront Inc.).

Facial Expression Editing Tool

This tool is based on a facial muscle model [Waters and Frisbie 1995; Lee et al. 1995] composed of facial tissue and the simulated muscles of a geometric model, which is arranged according to the anatomical configuration of the face. The muscular tissue is modeled as an aggregate of springs. In this model, forces affect a facial tissue element through the contraction of each muscle spring, so the combination of the contracting forces of various muscles produces a specific facial expression.

The control panel of the facial expression editing tool has slider bars for each muscle used to create a facial expression. When these bars are moved, each muscle detects the change in the applied force and the model is deformed. To realize motion that leads to a more natural facial expression, we have to adjust the target model and muscles by using a geometric model and move each slider bar to determine the appropriate muscle parameters for various expressions and mouth movements. The basic expression of Anger is shown in Figure 1.

Face Model Fitting Tool

When a specific expression of a facial model has been developed with the facial expression editing tool, we normally would still have to set up the geometry and muscle parameters again if we wanted to reuse that expression. However, the face model fitting tool allows us to avoid this process. Therefore, we propose the Multistage Ray-Cast Fitting plug-in software. The user selects a face model whose muscle parameters were defined using the tool (the source model), selects a different number of polygons and a different polygon geometry (the target model), and then executes the program that approximates the source model to the target model. Since the geometry of the deformed source model already has the muscle parameters, the specific expression and mouth shape of the target model can be easily formed. As well as a realistic human facial model, this model can use an animal's or a cartoon character's face as a target model. Figure 2 shows the adjustment result for several models.

Lip-sync Animation Tool

When determining the animation of the mouth shape to apply to the face model made by the face model fitting tool, the model would normally need to speak while synchronized with the voice of an actor or actress. Also, if this model has to closely resemble the real face, the number of mouth shapes that must be prepared will equal the number of voice visemes. To do this is very time consuming for the creator who must pay close attention to the voice actor/ actress whose voice must be synchronized with the mouth animation. The lip-sync animation tool makes this task easier by analyzing a sound and extracting the phoneme and phoneme length. This phoneme is then matched with the created mouth shape. To create a natural looking mouth shape animation, the phoneme is interpolated using the phoneme duration and the muscle's equation of motion. Currently, this tool can estimate each Japanese vowel. Therefore, the lip-sync animation tool significantly reduces the difficulty of applying a user's voice which previously required a series of complicated and time-consuming tasks.

References

Waters, K., and Frisbie, J. 1995. A coordinate muscle model for speech animation. Proc. Graphics INterface' 95, pp.163-170.

Lee, Y., Terzopoulos, D., and Waters, K. 1995. Realistic modeling for facial animation. Proceedings of SIGGRAPH' 95, pp. 55-62.

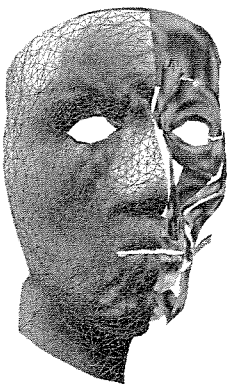
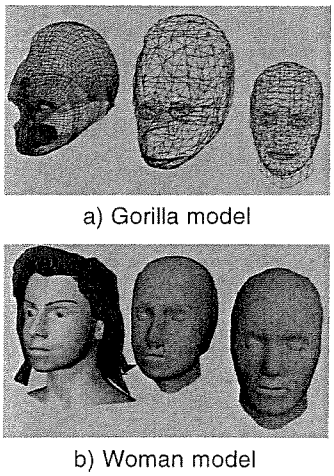


Fig.1 Facial expression of Anger



a) Gorilla model

b) Woman model

Fig.2 Face Model Fitting Tool
Left: Target model
Middle: Reconstracted model
Right: Source model

2．2　表情合成ツールの実現

2．1で得られた知見に基づいて、よりユーザフレンドーで使い勝手のよい表情合成ツールを構築した。これは毎回物理シミュレーションの計算を行うのではなく、基本動作として典型的な顔面動作を予め幾何変形ルールとしてモデル化し、この基本動作の組み合わせによって表情合成を支援するものである。また正面顔写真一枚からでも、簡易に3次元の表情合成が実現できるように工夫が行われている。この表情合成ツールは、音声対話擬人化エージェント作成支援ツール（Galatea）のサブプロセスとして組み込まれている。

Galatea: 音声対話擬人化エージェント開発キット

Galatea: An Anthropomorphic Spoken Dialogue Agent Toolkit

西本卓也 [1)] Takuya Nishimoto
荒木雅弘 [2)] Masahiro Araki
伊藤克亘 [3)] Katsunobu Itou
宇津呂武仁 [4)] Takehito Utsuro
甲斐充彦 [5)] Atsuhiko Kai
河口信夫 [3)] Nobuo Kawaguchi
河原達也 [4)] Tatsuya Kawahara
桂田浩一 [6)] Kouichi Katsurada
小林隆夫 [7)] Takao Kobayashi
嵯峨山茂樹 [1)] Shigeki Sagayama
下平 博 [8)] Hiroshi Shimodaira
伝 康晴 [9)] Yasuharu Den
徳田恵一 [10)] Keiichi Tokuda
中村 哲 [11)] Satoshi Nakamura
新田恒雄 [6)] Tsuneo Nitta
坂野秀樹 [12)] Hideki Banno
広瀬啓吉 [1)] Keikichi Hirose
峯松信明 [1)] Nobuaki Minematsu
三村正人 [13)] Masato Mimura
森島繁生 [14)] Shigeo Morishima
山下洋一 [15)] Yoichi Yamashita
山田 篤 [13)] Atsushi Yamada
四倉達夫 [11)] Tatsuo Yotsukura
李 晃伸 [16)] Akinobu Lee

(1) 東京大学大学院　情報理工学系研究科
（〒113-8656 東京都文京区本郷 7-3-1 E-mail: nishi @ hil.t.u-tokyo.ac.jp）
(2) 京都工繊大 (3) 名大 (4) 京大 (5) 静岡大 (6) 豊橋技科大 (7) 東工大 (8) 北陸先端大 (9) 千葉大
(10) 名工大 (11) ATR (12) 和歌山大 (13) ASTEM (14) 成蹊大 (15) 立命館大 (16) 奈良先端大

1　はじめに

著者らは，音声対話技術を活用した情報処理技術のいっそうの高度な利用を目指して，2003 年 11 月より 3 年間の予定で音声対話技術コンソーシアム (ISTC) を発足させた [1]．ISTC の主な活動は IPA（情報処理振興事業協会）のプロジェクト（Galatea プロジェクト）で開発された「音声対話擬人化エージェント基本ソフトウェア」の発展と拡充である．この開発キットはモジュール構成の柔軟さを考慮して設計され，マルチモーダル対話の研究開発支援プラットフォームを提供している [2]．本報告では，擬人化音声対話エージェントの開発キット Galatea の概要について述べる．なお，Galatea のダウンロード方法などの情報は下記を参照されたい．

http://hil.t.u-tokyo.ac.jp/~galatea/

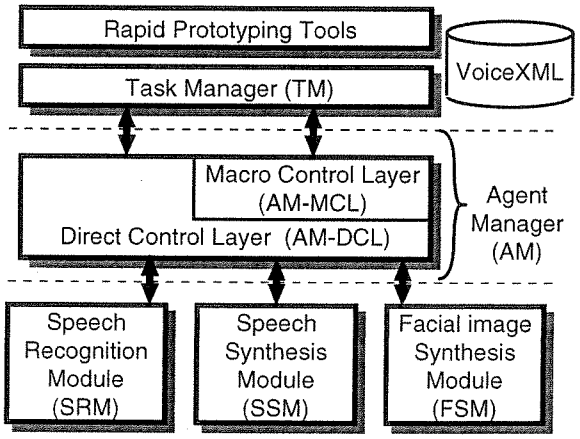


図 1: Galatea 開発キットの全体構成

2　Galatea の構成と特長

図 1 に開発キットの全体構成を示す．各モジュールは独立のプロセスとして設計されており，対話システムは単一 PC (Note PC (Mobile Pentium III 1.2GHz, 512MB) 上で動作確認済)，あるいは分散環境（複数 PC による並行動作）で使用することができる．以下に各モジュールを概説する．

(1) 音声認識：Julian[3] をベースに音声対話システムで要求される (a) 文法に基づく音声認識，(b) 発話中の逐次的な認識結果出力，(c) 認識処理の動的制御（中断，文法の切替等）の諸機能を提供している（図 2）．

(2) 音声合成：日本語テキスト音声合成に必要な基本機能（形態素解析（茶筌 [4]），読み・アクセント付与，韻律生成，合成波形生成）のほか，(a) 音素継続時間長を出力し顔画像の口唇との同期が可能，(b) テキスト埋め込みタグ（JEIDA 規格準拠）による韻律制御が可能，(c) 合成音を出力途中で中断可能（barge in 等）といった特長を持つ．合成器は HMM に基づく方式 [5] を採用し，男女各 1 名の話者モデルを提供している（図 3）．

(3) 顔画像合成：標準ワイヤーフレームモデル中の代表点と正面写真中の対応点を，短時間（5-10 分）のマウス操作で整合させるだけで，表情変化が可能である [6]．表情は怒り，喜び，悲しみ，驚き，嫌悪，恐れの 6 種を用意している（図 4）．音声対話のため，LipSync のほか，自律的な動作（うなづき，瞬き等）を提供している．

(4) エージェントマネジャ：対話部品が個々に規定するコマンドセットを使用して直接制御するレイヤと，対話管理に便利なマクロコマンドを利用して制御できるレイ

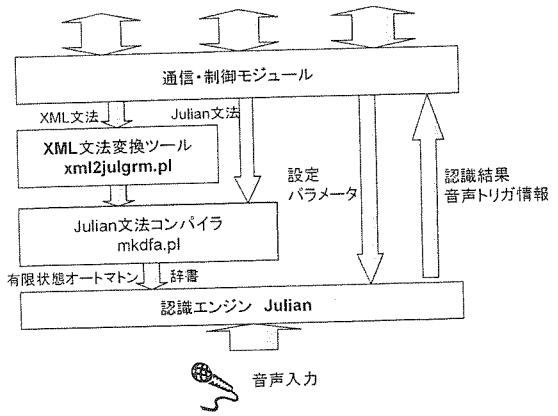


図 2: 音声認識モジュールの構成

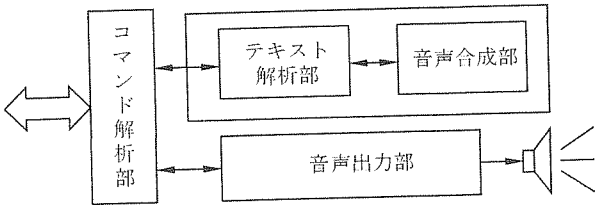


図 3: 音声合成モジュール (GalateaTalk) の構成

ヤの二つを提供している（Unix 版の場合 [6]. Windows 版では対話マネジャが各対話モジュールとソケット通信を行なう）.

3 開発支援ツール

対話記述言語として，(a) VoiceXML に GUI のためのタグを付加したもの（主に Linux 版で使用される [7], 図 5）と，(b) モダリティの追加が可能なマルチモーダル対話向け言語（XISL [7]; Windows 版で使用）の二つを提供している．現在，(a) では対話処理系および簡単な GUI ツールが，(b) ではラピッドプロトタイピングツール (Interaction Builder (IB) [8], 図 6) が開発キットに含まれている．

図 4: 表情合成結果の一例

```
<form id="main">
<field name="place">
<prompt> <emotion type="HAPPY">場所をどうぞ.
</emotion> </prompt>
<prompt count="3">
<emotion type="SAD">東京と京都のどちらですか？</emotion>
</prompt>
<grammar><rule><one-of>
<item><token sym="とうきょう">東京</token></item>
<item><token sym="きょうと">京都</token></item>
</one-of></rule></grammar>
</field>
</form>
```

図 5: VoiceXML による対話の記述例

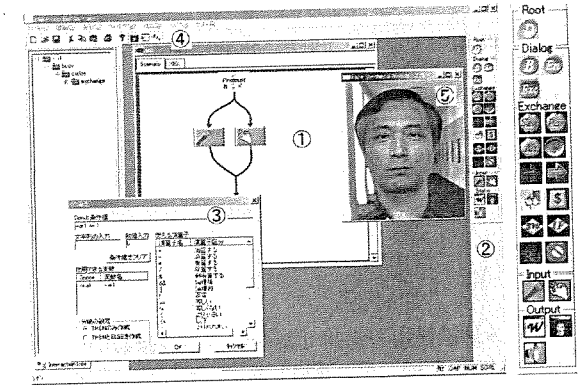


図 6: Galatea-IB の実行画面と対話部品バー

4 今後の予定

Galatea はオープンソース化を前提に開発された．音声対話技術コンソーシアム (ISTC) では今後，各サブモジュールの改良を行なっていく予定である．また，CD-ROM 配布，セミナー・講習会開催を通して関連研究と応用システム開発を支援する予定である．

参考文献

[1] http://www.lang.astem.or.jp/ISTC/index.html
[2] 嵯峨山ほか：情処研報，SLP-45-10, pp.57-64 (2003).
[3] 住吉ほか：情処研報，SLP-37-16, pp.91-96 (2001).
[4] http://chasen.aist-nara.ac.jp/
[5] http://hts.ics.nitech.ac.jp/
[6] 川本ほか：情処論誌，vol.43, no.7, pp.2249-2263 (2002).
[7] 西本ほか：人工知能学会全大，2C2-04 (2003).
[8] 足立ほか：情処研報，SLP-43-2, pp.7-12 (2002).

Open-source Software for Developing Anthropomorphic Spoken Dialog Agents

Shin-ichi Kawamoto[*1] **Hiroshi Shimodaira**[*1] **Tsuneo Nitta**[*3] **Takuya Nishimoto**[*4]
Satoshi Nakamura[*5] **Katsunobu Itou**[*6] **Shigeo Morishima**[*7] **Tatsuo Yotsukura**[*7]
Atsuhiko Kai[*8] **Akinobu Lee**[*9] **Yoichi Yamashita**[*10] **Takao Kobayashi**[*11]
Keiichi Tokuda[*12] **Keikichi Hirose**[*2] **Nobuaki Minematsu**[*2] **Atsushi Yamada**[*13]
Yasuharu Den[*14] **Takehito Utsuro**[*3] **Shigeki Sagayama**[*2]

*1 Japan Advanced Institute of Science and Technology, *2 The University of Tokyo, *3 Toyohashi University of Technology, *4 Kyoto Institute of Technology, *5 Advanced Telecommunications Research Institute International, *6 National Institute of Advanced Industrial Science and Technology, *7 Seikei University, *8 Shizuoka University, *9 Nara Institute of Science and Technology, *10 Ritsumeikan University, *11 Tokyo Institute of Technology, *12 Nagoya Institute of Technology, *13 The Advanced Software Technology and Mechatronics Research Institute of Kyoto, *14 Chiba University

Abstract

An architecture for highly-interactive human-like spoken-dialog agent is discussed in this paper. In order to easily integrate the modules of different characteristics including speech recognizer, speech synthesizer, facial-image synthesizer and dialog controller, each module is modeled as a virtual machine that has a simple common interface and is connected to each other through a broker (communication manager). The agent system under development is supported by the IPA and it will be publicly available as a software toolkit this year.

1. Introduction

Anthropomorphic spoken dialog agent (ASDA), behaving like humans with facial animation and gesture, and making speech conversations with humans, is one of the next-generation human-interface. Although a number of ASDA systems (Gustafson et al., 1999; Julia and Cheyer, 1999; Dohi and Ishizuka, 1997; Ushida et al., 1998; Sakamoto et al., 1997; Cassell et al., 1999) have been developed, communication between the ASDA system and humans is far from being natural, and developing high quality ASDA system is still challenging. In order to activate and progress the researches in this field, we believe that easy-to-use, easy-to-customize, and free software toolkit for building ASDA systems is indispensable.

We have been developing such an ASDA software toolkit since 2000, aiming to provide a platform to build next generation ASDA systems. The features of the toolkit are as follows: (1) basic functions to achieve incremental (on-the-fly) speech recognition, (2) mechanism for "lip synchronization"; synchronization between audio speech and lip image motion, (3) high customizability in text-to-speech synthesis, realistic face animation synthesis, and speech recognition, (4) "virtual machine" architecture to achieve transparency in module to module communication.

If compared to the related works such as CSLU toolkit (Sutton and Cole, 1998) and DARPA Communicator Program (DARPA, 1998), our toolkit is still germinal. However, it is compact, simple, easy-to-understand and thus suitable for developing ASDA systems for research purposes. At present, simple ASDA systems have been successfully build with the toolkit under UNIX/Linux operating systems, and the subset of the toolkit will be publicly available in the middle of the year 2002.

This paper is divided into six sections. Requirements for the ASDA software toolkit are discussed in section 2 followed by the discussion of system design in section 3. Implementation issue and evaluation are described in section 4. Finally the last section is devoted to conclusions.

This work is supported by IPA (Information-technology Promotion Agency, Japan).

2. Requirements for the toolkit

In this section, we discuss the requirements for the software toolkit to build ASDA systems which speak, listen, and behave like humans.

2.1. Key techniques for achieving natural spoken dialog

If compared to the keyboard-based conversation, typical phenomena are observed in speech-based conversation. These include the case that human listeners nod their heads or say "yes"during a conversation, and the case that the speakers control the prosody to indicate types of utterances such as questions, statements, and emotions. We regard it important for the toolkit to be a platform for human-like speech-based conversation for providing basic functions to achieve those phenomena.

In addition, speed, quality and balance are also important factors for the toolkit. For example, if the system fails to respond to the user quickly, it loses the naturalness and efficiency of conversation. If the agent's face, voice and behavior are artificial and far from natural, or if the agent looks very similar to humans apart from the point that the voice is synthesized, then the users feel something strange and it prevents them to communicate with the system naturally.

2.2. Configuration for the easy-to-customize

As a common basic toolkit for research and development, the toolkit should not be designed for a specific purpose, but it should be used for multi-purpose. The agent's face, voice, and tasks must be customizable so that the users

of the toolkit can customize the agents easily depending on the purposes and applications. The customizability includes that the agent characters should be replaced easily by changing the face and voice of a person to those of an another person.

2.3. Modularity of functional units

In some situations, system creators or toolkit users will not be satisfied with the performance of the original modules in the toolkit and they would like to replace them with the new ones or add new ones to the system. In such cases, it would be desired that each functional unit (module) is well modularized so that the users can develop, improve, debug and use each unit independently from the other modules. This would help to improve the efficiency of software development.

Moreover, modularizing the functional units enables the system to work in parallel,

2.4. Open-source free software

The technology used for creating the toolkit is still not enough to achieve human-like conversation. Therefore it is desired that not only the creators of the toolkit but also the researchers and developers who use the toolkit would contribute to improve the toolkit further. In that sense, the toolkit should be released as a free software along with the program source codes.

There have been no existing ASDA softwares so far satisfying all of the requirements described above.

3. Toolkit design and outline

In this section, we discuss the design of the toolkit and its module functionality to achieve the requirements given in the previous section.

First of all, to fulfill the requirements of modularity and customizability, the toolkit must have at least three functional units (speech recognition, speech synthesis, and facial animation synthesis) for task customization, and a unit for integrating those units, which we name as "agent manager".

3.1. Speech recognition module (SRM)

The authors have been developing the Japanese large vocabulary continuous speech recognition (LVCSR) engines, Julius (Kawahara et al., 1998; Lee et al., 2001) and SPOJUS (Nakagawa and Kai, 1994). Julius employs N-grams as a statistical language model (LM), though, as a toolkit for various tasks, grammar-based LM is suitable for small tasks, where easy-to-use and easy-to-customize LMs are preferable. In order to provide such a grammar-based recognition engine as a functional module of the toolkit, "Julian" (Fig. 2) has been developed. Julian can change more than one grammar sets on the instant, and it can output incremental speech recognition results.

3.2. Speech synthesis module (SSM)

To achieve customizable speech synthesis module (SSM), the module has to accept arbitrary Japanese texts including both of "Kanji" (Chinese) and "Kana" characters, and synthesize speech with a human voice clearly in a

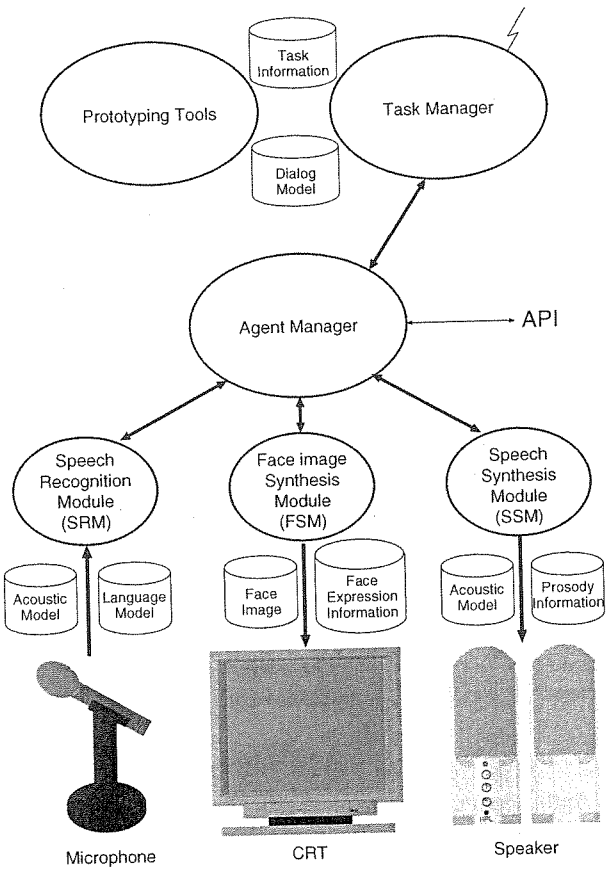


Figure 1: ASDA platform

specified style. For this purpose, HMM-based speech synthesis method is employed in which spectrum, pitch and duration are modeled simultaneously in a unified framework of HMM (Yoshimura et al., 1999). Lexical and syntactic analyzer is developed as well.

Another important function of this module is to implement a mechanism for synchronizing the lip movement with speech, which is called "lip-sync". The employed mechanism is based on the sharing of each timing and duration information of phoneme in the speech that is going to be uttered between the SSM and the FSM (facial image synthesis module).

3.3. Facial image synthesis module (FSM)

The basic software of synthesizing human facial images can synthesize human facial animations of any existing person if a single photo image of the person is given and the image is fitted manually to a standard 3D wire-frame model (Morishima et al., 1995). The software including a model fitting tool is publicly available as a result of the former IPA project (facetool, 1998). Under the current ASDA toolkit project, we are enhancing the former software package to support higher quality and controllability of agent facial image, and precise lip-sync with synthetic speech. Fig. 3 shows the process of fitting a 3D wire-frame model to a real human face, and the examples of the synthesized human facial images.

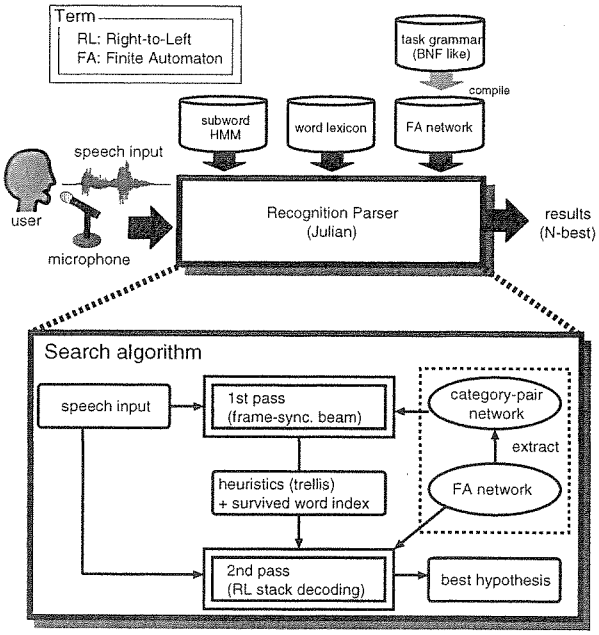


Figure 2: Speech recognition module

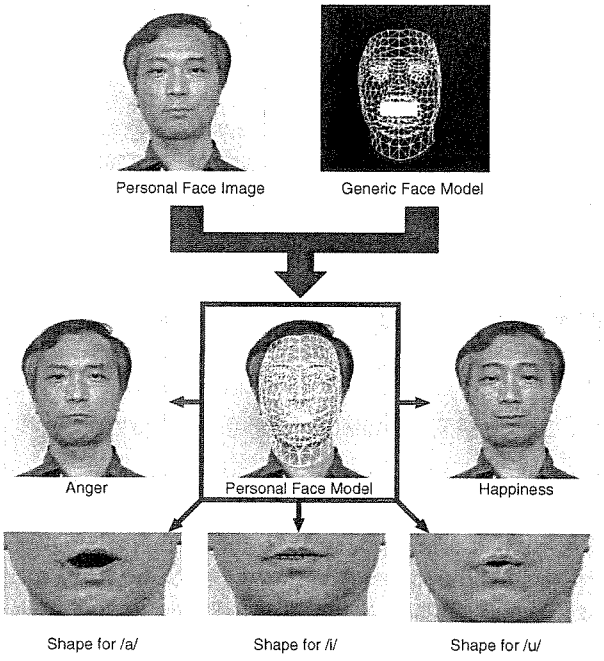


Figure 3: Facial image synthesis module

3.4. Module integration and customization tools

3.4.1. Agent manager

The Agent Manager (AM) serves as an integrator of all the modules of the ASDA system. One of its main functions is to play a central role of communication where every message from a module is sent to another module with the help of the AM. Here, the AM works like a hub in the Galaxy-II system (Seneff et al., 1998). Another essential function of the AM is to work as a synchronization manager between speech synthesis and facial image animation to achieve the precise lip-sync.

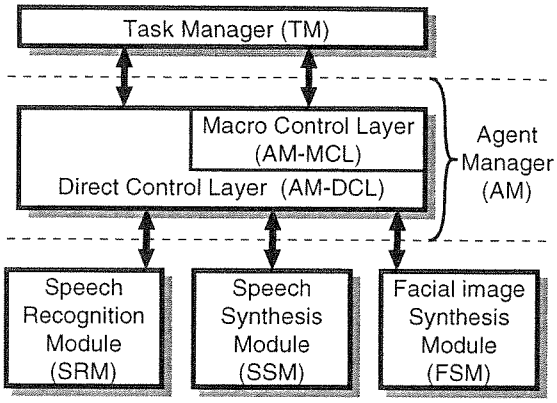


Figure 4: Basic configuration of the AM and Modules

The AM consists of two functional layers: the Direct Control Layer (AM-DCL) and the Macro Control Layer (AM-MCL). Fig. 4 illustrates a schematic representation of the relationship between the AM and the various modules. The AM-DCL works as a dispatcher receiving commands from a module and forwarding them to the designated module. On the other hand the AM-MCL is a macro-command interpreter processing the macro commands mainly issued by the Task Manager (TM). There are mainly two functions for the AM-MCL. The first one is to simply expand each received macro-command in a sequence of commands and send them sequentially to the designated module. The second function is to process macro-commands that require more complicated processing than just expanding the commands. This happens in the case where more than one modules are involved. Currently, the lip synchronization process is realized by a macro command and an example is given in section 4.

3.4.2. Virtual Machine model

As is previously described, the AM works as a hub through which every module communicates with each other. It is desired that every module has a common communication interface so that the AM can make connection with each module regardless of the interface used in the module. Furthermore, having a common interface reduces the effort of understanding and developing module dependent interfaces. For this purpose a virtual machine (VM) model is employed, where module interface is modeled as a machine with slots, each of which has a value and attribute controlled by a common command set. Each slot can be regarded as a switch or dial to control the operation or a meter to indicate machine status. Fig. 5 illustrates the communication between the AM and a virtual machine model. Changing the slot values by a command corresponds to check or control the running status of the module or the function. For example, following command to the speech synthesis module means starting voice synthesis of a given text right now.

```
set Speak = Now
```

3.4.3. Task manager (TM)

As a software toolkit, it is crucial to define the complexity of tasks that the software can deal with. In that sense, VoiceXML (VoiceXML, 2000) has been employed

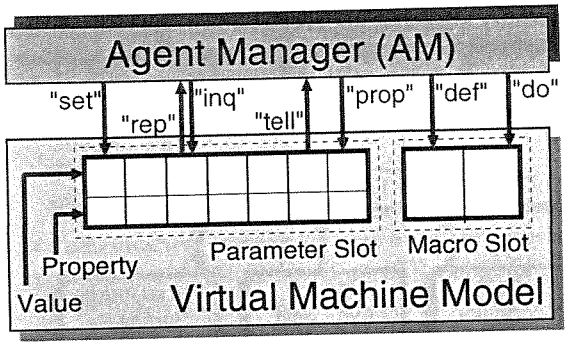


Figure 5: Relationship between the AM and a virtual machine model

Figure 6: Screenshot of ASDA

as a basic description language to describe the tasks. Since VoiceXML is originally designed for voice communication over the telephone, and difficulties arise when it is applied to other applications such as anthropomorphic dialogue agents, extensions to its specification are being made in this project. For example, the original specification of VoiceXML does not include any functions to control facial expressions of anthropomorphic dialogue agents.

3.4.4. Prototyping tools

For achieving an easy to customize toolkit, we have a plan to provide prototyping tools. These tools manage some agent customization features. For example, dialog scenario, and related parameters.

4. Experimental Systems

Using the software toolkit, we have built several experimental ASDA systems to evaluate the toolkit. A screenshot of the system and an example of a user-system interaction are shown in Fig. 6 and Fig. 7 respectively.

All the tasks employed were very basic, small vocabulary where the number of uttered word is less than 100 and the perplexity is less than 10. The tasks include (1) echo-back task which repeats what it heard using speech recognition and synthesis, (2) simple appoint-arranging task which changes facial expressions as the conversation goes on, (3) fresh food ordering task that takes orders from customers and responses with "yes" and nodding on the fly.

Those systems consist of the SRM (Kawahara et al., 1998), the SSM (Yoshimura et al., 1999), the FSM (Morishima, 2001), the AM, and a simple task-specific TM

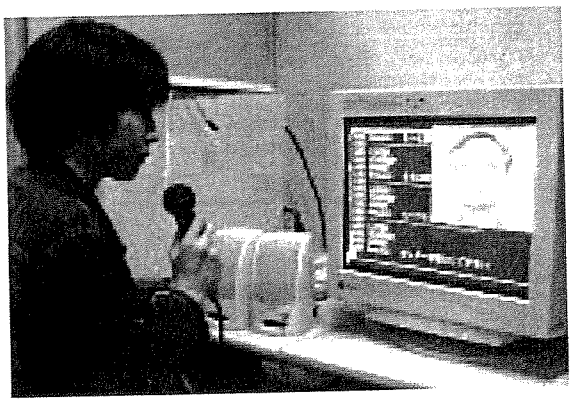


Figure 7: An example of user-system interaction

SHORT TITLE

SRM: Speech recognition module
SSM: Speech synthesis module
FSM: Facial image synthesis module
AM: Agent manager
TM: Task manager
AUTO: Autonomous head-moving module

COMPUTER SPEC.

PC #1 ... CPU: Pentium III Xeon 1GHz x 2, MEMORY: 512MB
PC #2 ... CPU: Pentium III 600MHz x 2, MEMORY: 512MB
PC #3 ... CPU: Mobile Pentium III 1.2GHz, MEMORY: 512MB

SYSTEM ENVIRONMENT

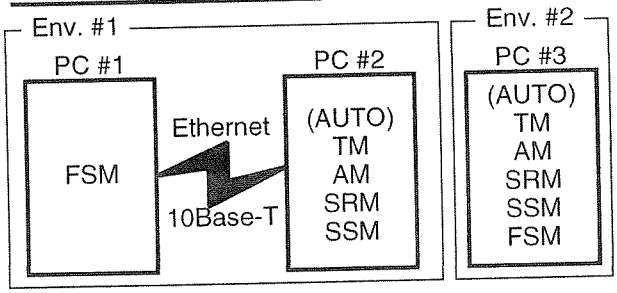


Figure 8: Hardware configuration of the ASDA

which was programmed directly with the command set of the toolkit. We implemented the systems on several platforms with different configurations. Fig. 8 shows the hardware configurations. Some of the demonstration movies (in Japanese, unfortunately) are available in our web site (`http://iipl.jaist.ac.jp/IPA/`).

Fig. 9 shows an example of how the AM and related modules work in the echo-back task. However, the FSM and lip-synchronization mechanism have been omitted in the figure for brevity. Here, the macro commands, which is introduced in 3.4.1., are used in the procedure 3 and 4 to achieve lip-synchronization between the speech and animation. Fig. 10 shows the sequence of commands involved in this lip-synchronization process.

Note that the modules operate in parallel and thus the speech recognition process is active while the agent is speaking. As a result, we confirmed that the system responded to the users quickly, at the same time face animation and synthesized voice were synchronized. However, in this case, we assumed that the ideal environment that the results of speech recognition are not influenced by the output

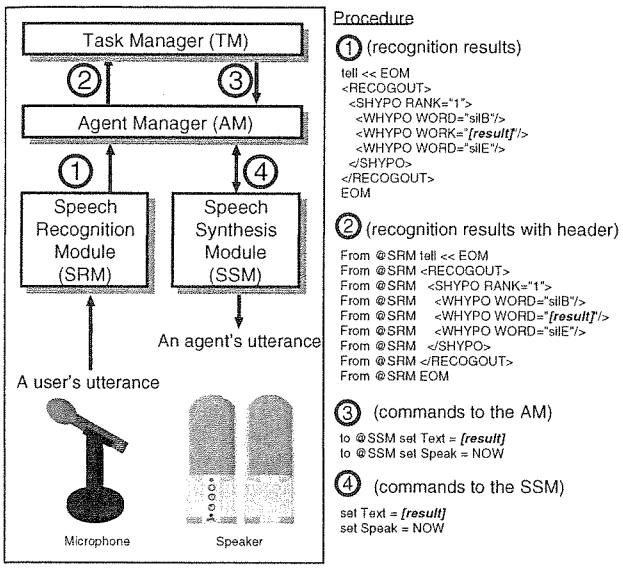


Figure 9: An example of echo-back processing task

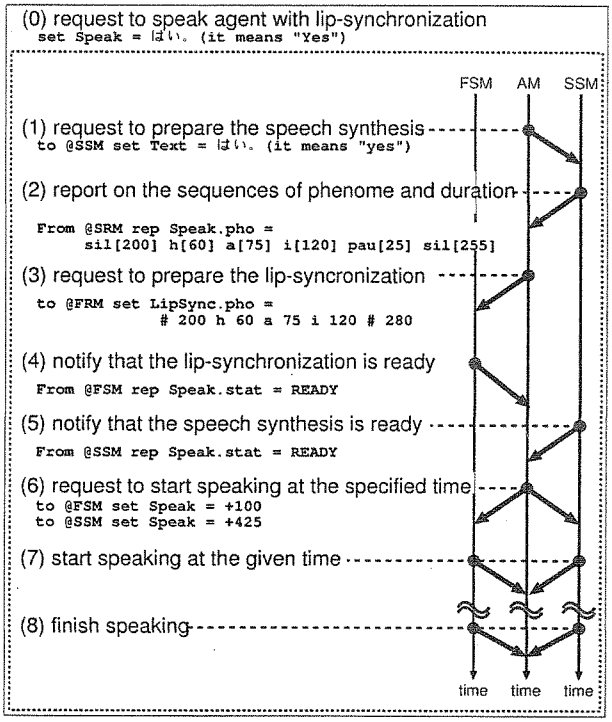


Figure 10: Processing flow among the AM, the SSM, and the FSM when agent speaks (an example of processing in the AM)

of speech synthesis.

5. Discussion

This section describes the current developing status of the software toolkit and discusses further improvement.

5.1. Customization features

In SRM, multi-grammar supporting has been realized where grammars can be changed instantly, and those grammars are easy to customize by means of a supporting software tool.

The SSM can synthesize speech from arbitrary text sentences of mixed kanji and kana (Chinese characters and phonetic script), with customizable prosody. Though speaker adaptation has not been implemented, the employed HMM-based approach is promising in case of speaker adaptation (Tamura et al., 2001b; Tamura et al., 2001a).

The FSM synthesizes 3D realistic facial animations from a single snapshot of a person's face by fitting a wire-frame model to a 2D picture. A software tool is provided to help fitting a standard wire-frame model to the input picture, whose manually fitting operation takes normally 10 minutes. Once the fitting is completed, one can get realistic 3D facial animation of the person whose motion, including blinking and facial expression, is easily and precisely controllable by commands in real time. Comparing to the cartoon based existing approaches where the number of characters is very limited, the proposed framework enables to generate facial animations of almost unlimited number of characters as far as facial pictures are provided.

5.2. Software Modularity of functional units

As is described in the previous section, the virtual machine model enables highly modularity of each functional units such as SRM, SSM and FSM. Furthermore, communication interface based on the UNIX standard I/O stream helps to develop and debug software modules easily.

5.3. Achievement of natural spoken dialog

Although the implemented mechanism for lip-sync contributes to enhance the naturalness of the synthetic facial animation, number of issues are yet to be implemented to make the agent behave like humans. For example, humans move their heads while they are speaking. Besides the facial animation, realtimeness of conversation is another crucial factor for the agent's naturalness as is described in section 2.1. A simple mechanism for incremental speech recognition has been implemented in the SRM. The mechanism provides frame-synchronous temporal candidates giving maximum scores at the moment before observing the end of utterance. These incremental recognition results will help to achieve interactive spoken dialog including nodding.

6. Related Works

Several attempts have been made to develop ASDA toolkits. Among them, the CSLU toolkit (Sutton and Cole, 1998) is similar to our toolkit. The CSLU toolkit provides a modular, open architecture supporting distributed, cross-platform, client/server-based networking. It includes interfaces for standard telephony, audio devices, and software interfaces for speech recognition. It also includes text-to-speech and animation components. This flexible environment makes it possible to easily integrate new components and to develop scalable, portable speech-related applications. Although the target of both of the toolkits are similar, function wise and implementation wise they are different. Compared to the speech recognizer and speech

synthesizer of the CSLU toolkit that support several European languages, our toolkit supports Japanese language. The TTS in the CSLU toolkit is based on "unit selection and concatenation synthesis" from natural speech. It is a data-driven and *non* model-based approach. However, the TTS in our toolkit employs the HMM-based synthesis that is a data-driven and model-based approach. The different approaches give different characteristics to TTS. Generally speaking, the model-based TTS requires less training samples and it can control speech more easily than the non model-based TTS at the expense of speech quality.

Similar system architectures for distributed computing environment are employed in the Galaxy-II (Seneff et al., 1998) of DARPA Communicator (DARPA, 1998), the SRI Open Agent Architecture (OAA) (OAA, 2001), and our toolkit. Each of them have a central module called "Hub", "facilitator" and Agent Manager (AM) respectively. If compared to the existing systems which employs a large number of commands, our toolkit is more compact and simpler and it has only 8 commands and 2 identifiers so that the programmers can understand and use the toolkit easily.

7. Conclusions

The design and architecture of a software toolkit for building an easy to customize anthropomorphic spoken dialog agent (ASDA) has be presented in this paper. Human-like spoken dialog agent is one of the promising man-machine interfaces for the next generation. The beta-version of the software toolkit described in this paper will be released publicly in the middle of 2002. However, a number of factors are to be improved. Because of the high modularity and simple communication architecture employed in the toolkit, we hope that it would speed up the researches and application development based on ASDA, and as a result the toolkit would be upgraded.

8. References

J. Cassell, T. Bickmore, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. 1999. Requirements for an architecture for embodied conversational characters. In D. Thalmann and N. Thalmann, editors, *Proceedings of Computer Animation and Simulation '99 (Eurographics Series)*, pages 109–122.

DARPA. 1998. DARPA Communicator Program. http://fofoca.mitre.org/.

Hiroshi Dohi and Mitsuru Ishizuka. 1997. Visual Software Agent: A Realistic Face-to-Face Style Interface connected with WWW/Netscape. In *IJCAI Workshop on Intelligent Multimodal Systems*, pages 17–22.

facetool. 1998. Facial Image Processing System for Human-like "Kansei" Agent. http://www.tokyo.image-lab.or.jp/aa/ipa/.

Joakim Gustafson, Nikolaj Lindberg, and Magnus Lundeberg. 1999. The August Spoken Dialogue System. In *EuroSpeech*, pages 1151–1154.

Luc Julia and Adam Cheyer. 1999. Is Talking To Virtual More Realistic? In *EuroSpeech*, pages 1719–1722.

T. Kawahara, T. Kobayashi, T. Takeda, N. Minematsu, K. Itou, M. Yamamoto, T. Utsuro, and K. Shikano. 1998. Sharable software repository for Japanese large vocabulary continuous speech recognition. In *ICSLP-98*, pages 3257–3260.

A. Lee, T. Kawahara, and K. Shikano. 2001. Julius — an open source real-time large vocabulary recognition engine. In *European Conf. on Speech Communication and Technology*, pages 1691–1694.

S. Morishima, S. Iwasawa, T. Sakaguchi, F. Kawakami, and M. Ando. 1995. Better Face Communication. In *Visual Proceedings of ACM SIGGRAPH'95*, page 117.

Shigeo Morishima. 2001. Face Analysis and Synthesis. *IEEE Siginal Processing Magizine*, 18(3):26–34, may.

Seiichi Nakagawa and Atsuhiko Kai. 1994. A Context-Free Grammar-Driven, One-Pass HMM-Based Continuous Speech Recognition Method. *Systems and Computers in Japan*, 25(4):92–102, September.

2001. OAA (The Open Agent Architecture). http://www.ai.sri.com/~oaa/.

Kenji Sakamoto, Haruo Hinode, Keiko Watanuki, Susumu Seki, Jiro Kiyama, and Fumio Togawa. 1997. A Responce Model for a CG Character Based on Timing of Interactions in a Multimodal Human Interface. In *IUI-97*, pages 257–260.

Stephenie Seneff, Ed Hurley, Raymond Lau, Christine Pao, Philipp Schmid, and Victor Zue. 1998. GALAXY-II: A Referece Architecture for Conversational System Development. In *ICSLP-1998*, pages 931–934.

S. Sutton and R. Cole. 1998. Universal speech tools: the cslu toolkit. In *Proceedings of the International Conference on Spoken Language Processing(ICSLP)*, pages 3221–3224.

Masatsune Tamura, Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi. 2001a. Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 805–808, May.

Masatsune Tamura, Takashi Masuko, Keiichi Tokuda, and Takao Kobayashi. 2001b. Text-to-speech synthesis with arbitrary speaker's voice from average voice. In *Proceedings of European Conference on Speech Communication and Technology*, volume 1, pages 345–348, September.

H. Ushida, Y. Hirayama, and H. Nakajima. 1998. Emotion Model for Life-like Agent and its Evaluation. In *AAAI-98*, pages 62–69.

VoiceXML. 2000. Voice eXtensible Markup Language VoiceXML Ver1.0. http://www.voicexml.org.

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *EuroSpeech*, volume 5, pages 2347–2350.

An Open Source Development Tool for Anthropomorphic Dialog Agent -Face Image Synthesis and Lip Synchronization-

Tatsuo YOTSUKURA and Shigeo MORISHIMA
Faculty of Engineering, Seikei University
3-3-1 Kichijoji Kitamachi, Musashino-shi
Tokyo, JAPAN
{yotsu, shigeo}@ee.seikei.ac.jp

Abstract—**We describe the design and report the development of an open source ware toolkit for building an easily customizable anthropomorphic dialog agent. This toolkit consist of four modules for multi-modal dialog integration, speech recognition, speech synthesis, and face image synthesis. In this paper, we focus on the construction of an agent's face image synthesis.**

Keywords—face image synthesis; anthropomorphic dialog agent; lip synchronization; facial animation;

I. INTRODUCTION

The development of human interface technology is one of the final goals to attain toward building a computer agent that can talk, ask questions, and exhibit behaviors that mimic humans. The past few years have witnessed the development technologies to make this dream possible. However compared with the interactions of our fellow human being, it is still at the elementary.

This paper describes the basic design and reports a demonstration of the project software. The software is intuitive, easy to understand, and ensures fully interactive dialog with the agent. The basic software consists of four modules: speech recognition, speech synthesis, facial image synthesis, and multi-modal dialog integration. System control and data management capabilities in a dispersed environment are essential for these various modules to interoperate smoothly as a single-dialog system. Several systems exhibiting these capabilities have been developed, which include DARPA's Communicator Program [1] that is based on MIT's Galaxy-II [2], and the Open Agent Architecture (OAA) developed by SRI [3].

We focus our description in this paper on the face image synthesis module, and discuss how to generate a synthetic agent's face in a real-time processing condition by exactly copying a real person's face. A very important factor in making an agent look believable or alive is how precisely an agent can duplicate a real human facial expression.

II. SOFTWARE CONFIGRATION

The anthropomorphic dialog agent that is now under development consists of four basic software modules, all of which will be made available in the form of freeware. By implementing the software as separate modules, this is not only an effective tool for assessing the various constituent technologies, it also provides a versatile R&D platform making it easy to build original dialog systems by simply plugging in different software modules developed independently by the different R&D institutes involved in the project as required. Figure 1 shows the basic configuration of the agent manager and the modules

A. *Integrating Anthropomorphic Dialog agents Module*

New basic software is being developed to integrate and control the dialog component modules and to manage the dialog. Some of the specific projects that are currently under way include (a) an Agent Manager (AM) providing low-level control of the speech recognition, speech synthesis, facial image synthesis, and other modules; (b) the capability to interpret Voice XML [4] based high-level dialog descriptions; (c) a Task Manager (TM) for controlling dialog using the functions provided by the AM; and (d) a prototyping tool to provide a GUI environment supporting the setting of parameters and the description and control of scenarios, all things that are necessary to construct dialog systems.

B. *Speech Synthesis Module: SSM[5]*

New basic speech synthesis software is being developed that not only clearly reads sentences of mixed kanji and kana (Chinese characters and phonetic script), but also shares data to enable synchronization with a facial image. This enables lip-sync, the synchronization of sound and motion so the facial movements of speech coincide with the sounds.

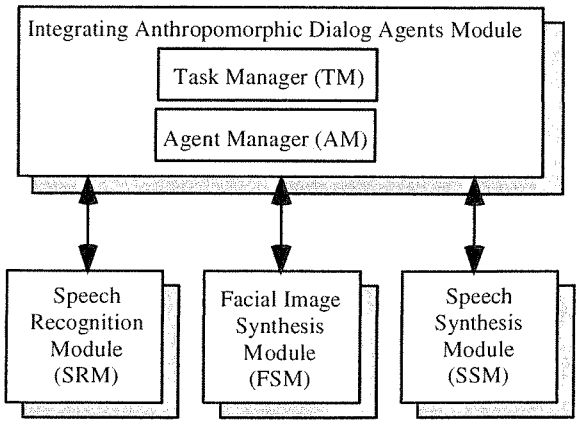


Figure 1. Basic Configuration of Agent Manager and Modules.

Furthermore, anticipating changes in the nature of the speech to reflect different circumstances or the intent of the speaker, we are also seeking ways to control a range of different emotions and speech rhythms.

C. *Speech Recognition Module: SRM[6]*

It should be easy enough to extend the capabilities of that software package to accommodate dialog processing and implement flexible control. Specifically, we are doing away with grammar-based recognition and recognition results, and developing functions that can deal with unnecessary words and poses, and can provide dynamic control of recognition processing.

D. *Facial image Synthesis Module: FSM[7]*

We are enhancing the software package to support higher quality agent facial image synthesis, animation control, and precise lip-sync with synthetic and natural speech. Some of the specific enhancements include a GUI able to map standard wire frames to images of heads shot from different angles to easily generate 3D models of human heads, sharing of data with the speech synthesis module, more precise lip-sync, the ability to add any facial expressions, and the ability to control nodding and blinking.

III. FACE OF AGNET MODELING

To generate a realistic agent, a generic face model is manually adjusted to the user's frontal face image to produce a personal face model and all of the control rules for facial expressions are defined as a movement of grid points in a generic face model (Figure 2). This model has an added oral and teeth model in the mouth, in order to generate a real looking facial expression. A synthesized face results from the texture mapping of the user's frontal image onto the modified personal face model.

Figure 3 shows a personal model both before and after the fitting process for a front-view image obtained by using our original GUI based face-fitting tool. The front view and profile images are put into the system and then corresponding control points are manually moved to a reasonable position by mouse operation. In the case of an expert user, the time required for adjustment of a model is about 5-10 minutes.

IV. CUSTOMIZING THE FACE MODEL

To change the agent's mouth shape and facial expressions, we need to define these models, so we built system to edit these models.

A. *Designing the Mouth Shape*

The set of mouth shapes can be easily edited by the mouth-shape-editing tool in the Face Synthesis Module (Figure 4). We can change each mouth parameter to determine a specific mouth shape, which can be seen in the preview window. Typical vowel mouth shapes are shown in Figure 5. A tongue model is now under construction. The kind of defined mouth shape is 14 shown in Table 1.

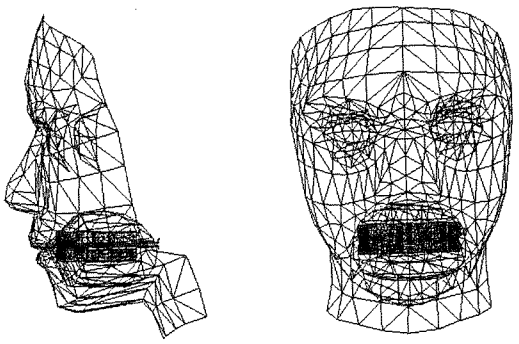


Figure 2. Generic Face Model (with Teeth and Oral Model)

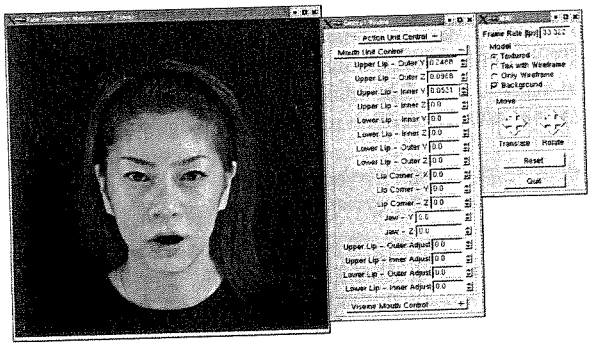


Figure 3. Model Fitting by GUI Tool

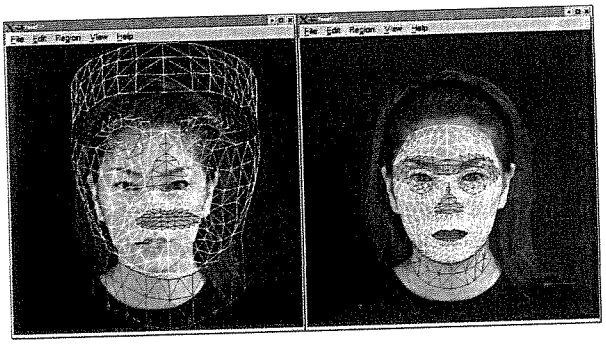


Figure 4. Control Panel for Mouth Shape Editor

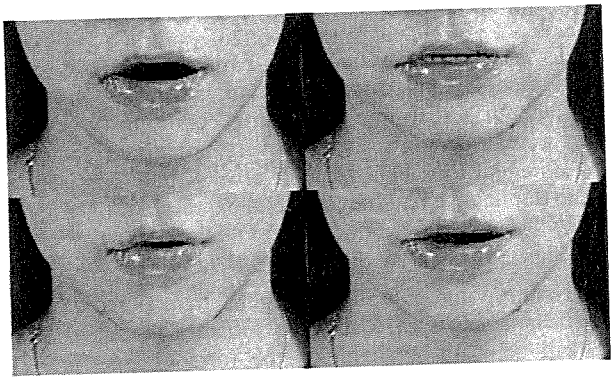


Figure 5. Typical Mouth Shapes

TABLE I. CONVERSION TABLE FROM PHONEME TO MOUTH SHAPE

Mouth shape No.	Phoneme (defined by SRM)
0	/h/, /y/, /cl/, /pau/, /sil/
1	/r/, /ry/
2	/m/, /b/, /p/, /my/, /by/, py/
3	/t/
4	/n/, /d/, /ny/
5	/k/, /g/, /ky/, /hy/, /gy/, /N/
6	/f/
7	/s/, /sh/, /ch/, /ts/, /z/, /j/, /dy/
8	/w/
9	/a/, /A/
10	/i/, /I/
11	/u/, /U/
12	/e/, /E/
13	/o/, /O/

TABLE II. EXAMPLE OF AUS

AU No.	AU name
AU 1	Inner brow raiser
AU 2	Outer brow raiser
AU 4	Brow lower
AU 5	Upper lid raiser
AU 6	Cheek raiser
AU 7	Lid tightener
AU 8	Lips toward each other
AU 9	Nose wrinkler
AU 10	Upper lip raiser
AU 12	Nasolabial furrow deepener

Figure 6. Typical Facial Expressions

B. *Designing the Facial Expression*

Expression modification is made by combining a basic expression (for example: inner brow raiser, cheek raiser, etc.) Therefore, we employed the facial action coding system (FACS)[8], which is an objective method for quantifying facial movement. FACS is an anatomically based coding scheme that codes the facial muscular movements in terms of 44 action units (AUs) or action unit combinations (AU combinations). Table 2 shows the an example of AUs.

The set of an expression can be edited by using the mouth AU editing tool in the Face Synthesis Module. By using this tool, the agent can transform by default defining six basic expressions (happiness, anger, disgust, fear, surprise and sadness) and user defining unique expression. Figure 6 shows examples of typical expressions.

V. LIP-SYNC ANIMATION

The agent's facial animation is realized by combining the mouth shapes and the expressions that the preceding chapter defined. In this method of major importance, the mouth animation made from the Face Synthesis Module and agent's voice made form the Voice Synthesis Module must be synchronize. In this chapter, we introduce the technique used for the synchronization that provides the realistic animation.

Management of the synchronization between the two modules might be implemented by a higher-level module that is separate and distinct from the Agent Manager. We are also considering defining and implementing a new type of module that is dedicated exclusively to synchronization. However, considering the importance of the lip-sync capability for agents and how frequently such capability is used in spoken dialog, we have currently implemented this function using a macro command provided by the Agent Manager.

The essential data that is needed for lip-sync when an agent speaks is the durations of each phoneme making up the speech. This information is obtained by interrogating the speech synthesis module. One might be able to think of other kinds of information that would be useful in this context, but for the time being we only use the duration of each phoneme.

It is also necessary to verify that the two modules are ready to speak before speaking can actually begin. This information is obtained using the following procedure. The speech process is divided into two parts: prepare to speak and begin to speak. The modules are designed to automatically generate a message indicating that they are ready to speak. As soon as the Agent Manager detects this information from the two modules, the Agent Manager directs that they can actually begin speaking. Figure 7 shows the sequence of commands involved in this process.

VI. PROTOTYPE SYSTEM

The prototype of the system was used in a demonstration at Japan Advanced Institute of Science and Technology. The system screen and the dialog scenery with the system are shown in Figures 8. We arranged two systems; one is comprised of a dual Intel 1GHz Pentium III Xeon computer to process FSM and an Intel 800MHz Pentium III computer that

processes all other modules, except for the FSM (environment #1). The other system is comprised of an Intel 1.2GHz Mobile Pentium III laptop computer that processes all modules (environment #2). An evaluation of the performance of the face synthesis module was performed using this system.

Drawing frame rates ware an average of 20[frame/sec] in environment #1, and 15[frame/sec] in environment #2 because of #1 operating one module as against #2 operating all modules on a PC. In addition, to use the phoneme, the phoneme duration and speaking start time of the agent supplied from Agent Manager, was checked visually by outputting result synchronizing with a synthetic voice.

VII. Conclusion

This paper described the design and reported the development of the basic software for an anthropomorphic dialog agent. As the project unfolds, we will further expand and enhance the functions of the dialog component modules and the multi-modal dialog integration module, and explore the feasibility of incorporating standard distributed object environment architecture such as CORBA[9].

In addition, we presented a technique of for use with a Facial Image Synthesis module. An agent duplicated original facial expressions and behavior. And the agent's synthesis voice was used to realize lip-sync. Finally, The validity of this module was confirmed by making a prototype system.

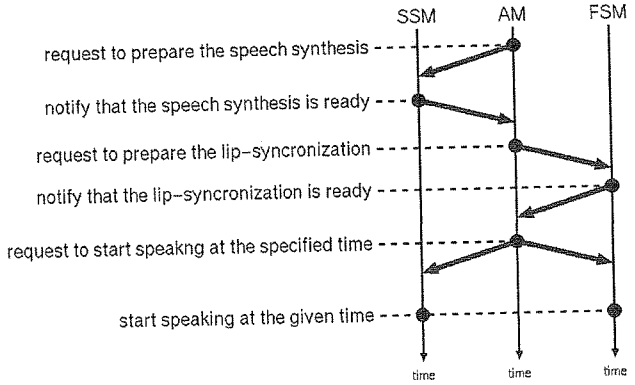


Figure 7. Processing Flow between AM, SSM, and FSM When an Agent speaks

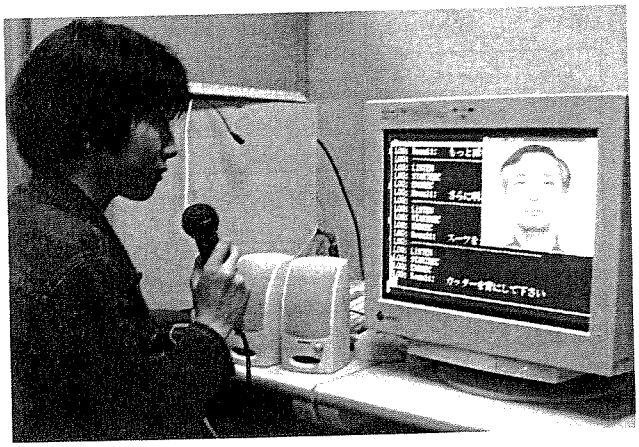


Figure 8. Screenshot of Anthropomorphic Dialog Agent

Acknowledgment

Part of this work was supported by the Information-technology Promotion Agency's program to support original information technology.

References

[1] DARPA: Communicator Program (1998). http://fofoca.mitre.org/.

[2] Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P. and Zue, V.: GALAXY-II: A Referece Architecture for Conversational System Development, ICSLP-1998, pp. 931–934 (1998).

[3] OAA: (The Open Agent Architecture). http://www.ai.sri.com/~oaa/.

[4] VoiceXML: (Voice eXtensible Markup Language Ver1.0) (2000). http://www.voicexml.org.

[5] Yoshimura, T., Tokuda, K., Masuko, T.,Kobayashi, T. and Kitamura, T.: Speaker Interpolation for HMM-based Speech Synthesis System, J Acoust. Soc. Jpn. (E), Vol. 21, No. 4, pp. 199–206 (2000).

[6] Itou, K., Hayamizu, S., Tanaka, K., Tanaka, H.: Sysstem design data collection and evaluation of a speech dialogue system, IEICE Trans. Inf. And Syst., Vol.36, No.1, pp.121-127 (1993)

[7] Morishima, S.: Face-to-face Communication in Cyberspace using Analysis and Synthesis of Facial Expression, Proceedings of '99 International Workshop on Advanced Image Technology(IWAIT99), pp.111-118 (1999)

[8] Ekman, P., Friesen, W. V.: Manual for the Facial Action Coding System and Action Unit Photographs. Palo Alto, CA: Consulting Psychological Press. (1978)

[9] CORBA: (The Common Object Request Broker Architecture). http://www.corba.org/.

リアルな顔合成による音声対話擬人化エージェントの開発

四倉　達夫　　森島　繁生

成蹊大学工学部

{yotsu, shigeo}@ee.seikei.ac.jp

1. はじめに

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。エージェントを構築するにあたり、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同等の環境を構築できるかがキーポイントとなる。本稿ではエージェントシステム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。

本研究は擬人化音声対話エージェントシステムの1つのモジュールとして機能する。このシステムは、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いるが、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。図1に実際本システムを使用し、ユーザとエージェントの対話風景を示す。

2. エージェント顔モデルの生成

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。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。顔画像と標準顔モデルとを整合する方法と

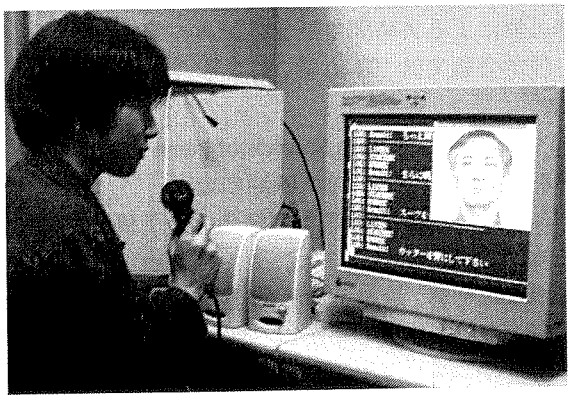


図1　音声対話擬人化エージェントシステム

図2　フィッティングツール

してまず、正面画像を用意する。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。GUIツールによるフィッティングの様子を図2に示す。

3. 顔モデルのカスタマイズ

3.1. 表情の設計

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)[1]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。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。

図3に複数のAUを組み合わせて制作した合成画像を示す。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。

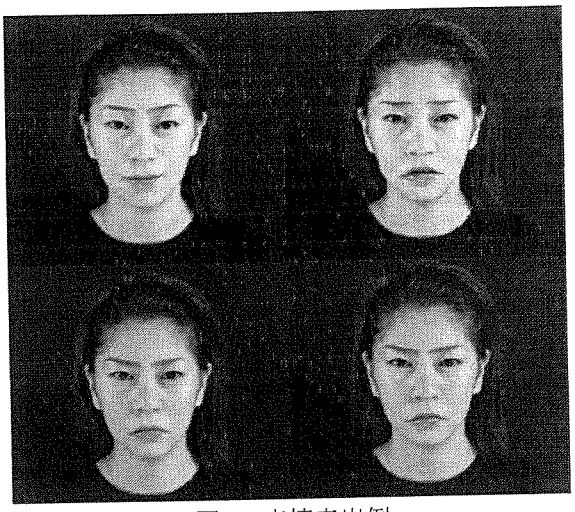


図3　表情表出例

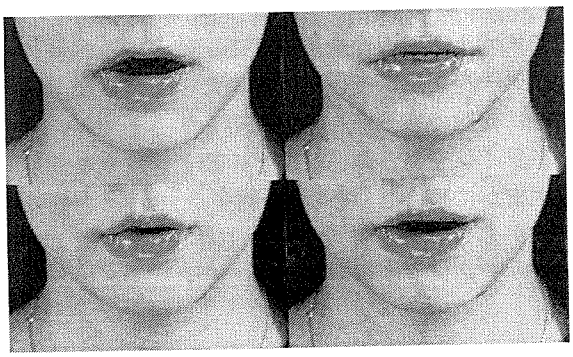


図4 典型的な母音口形状

3.2. 口形状の設計

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。図4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。

4. アニメーション

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。

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。

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しないという問題は生しなかった。

5. 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[2]。一方で実写との融合を行う研究も進めており、例えば主人公の顔部分を置換して表情合成を行う手法[3]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[4]についても検討している。

参考文献

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

[2] 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.

[3] 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.

[4] 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

2．3　表情合成アプリケーションの研究

表情合成のアプリケーションとしてビデオ翻訳システムを2001年に提案し、音声認識・翻訳・合成と、本表情合成アルゴリズムを結合して実現した。つぎに音声に同期して、仮面上に投影された表情合成画像がリアルタイムに表情変形を行うHyperMaskを実現し、舞台演技のツールとして提案した。また2005年の愛・地球博では、フューチャーキャストシステムとして、観客参加型のエンタテインメントシステムを提案し、そのキャストの表情合成手法に今回の表情合成アルゴリズムの一部が反映され、多くの観客に感動をもたらした。

Special Issue 特集 「愛・地球博」における最新映像技術

6. フューチャーキャストシステム

『三井・東芝館』

正会員 森島繁生†

キーワード CGキャラクター，実時間表情アニメーション，没入型シアター，3次元顔形状計測，顔特徴認識，ブレンドシェープ

1．まえがき

映画の登場人物に扮して宇宙を駆け巡る体験をしてみたい，あるいは，正義の味方の役で物語の主人公を演じてみたい，そういう希望を描いた経験が誰しも一度や二度はあると思う．ここで述べる『フューチャーキャストシステム』とは，そのような希望を容易に叶えてくれるまったく新しいエンタテイメントシステムである．

三井・東芝館では，この『フューチャーキャストシステム』を世界で初めて実現し，シアターが収容できる定員である240名の来場者全員が，映像の中に出演できるアトラクションとして具現化している．本稿では，この『フューチャーキャストシステム』の概要について述べる．

2．フューチャーキャストシステムの特徴

従来，イベント会場や遊園地などに設置されている各種の映像アトラクションは，映像を来場者の視覚に一方的に提供するものであり，来場者がその映像中に没入する感覚を高め，臨場感を増す工夫がなされている．『フューチャーキャストシステム』は，実際に来場者を映像中に登場させ，さらに映像中で出演者として演技させることで，ストーリーへの没入感を増大させている点に特徴がある．いわば，物語の中における自分の分身の姿を，観客の側から客観的に見つめさせることで，感動を呼び起こさせるまったく新しい発想のアトラクションである．

来場者は，まず20人ずつのグループに分割され，プレショールームに誘導される．ここでまず，顔の3次元スキャニングを行って顔の立体形状が計測される．この立体形状は，鼻の高さや頬の膨らみなど個人の特徴を反映するものであり，本人そっくりのCGキャラクターを生成する際に重要な情報となる．また同時に撮影される正面画像は，皮膚の色や目の色，男女の特徴，年齢特徴等の個人情報を表現するものとして重要である．

プレショールームで来場者が説明を聞いている間に，本人のCGモデルが自動的に計算され，スタンバイされる．また男女判定および年齢推定が自動的に行われ，ラベルとして情報が付与される．

メインシアターが開場となるや来場者は着席し，メインショーが開始される．ここで自分とそっくりのCG化された登場人物が活躍するストーリー映像を体験することになる．あらかじめ緻密に制作されたストーリー映像と，CG合成された登場人物の映像がブレンドされスクリーンに映し出される．登場人物の映像はリアルタイムで合成されており，キャストシナリオデータにしたがって演技する．また環境の変化も環境シナリオデータに基づいてタイムリーに映像に反映されるため，ストーリーに融合した違和感のない人物映像が合成される．観客そっくりの登場人物は，時にセリフを喋り，時には感情を露わにして，ストーリー映像の中で演技する．

最初20人ずつであった小規模のシアターは，途中で仕切りが取り除かれ80人シアターとなり，スクリーンサイズが4倍となって，さらに物語は展開していく．フィナーレでは，正面のスクリーンが下降して空間が広がり，三つの80人シアターが地球を取り囲んで合体する．一度のショーで240人の観客全員に感動をもたらすことができる．

3．キーとなる映像技術

『フューチャーキャストシステム』は，視聴者参加型のイベントシステムであるが，大人数の参加を実現可能とするためにオペレータの介入が必要ない点が特徴である．すなわちすべての処理が基本的に全自動で実行される．

3.1 3次元スキャニング

半円上に配置された7台のディジタルカメラで構成される3次元スキャナを利用し，映像処理によって顔の3次元計測を実施する．3次元スキャナの外観を図1に示す．

処理時間の制約から，画像サイズは現時点では縦800ピクセル，横600ピクセルに限定されているため奥行き情報の精度に限界はあるが，現時点のスクリーンサイズでは充分な精度と言える．プロセッサの高速化によって処理時間の短縮化が図れれば，さらに画像解像度を高めることで，奥行き精度を格段に向上させることも可能である．またカメラ台数を追加し，さらに360°方向にカメラを配置することで，原理的には頭部全周囲の3次元計測も可能である．

†早稲田大学 理工学部 応用物理学科
"Future Cast System/Mitsui Toshiba Pavilion" by Shigeo Morishima (Department of Applied Physics, Waseda University, Tokyo)

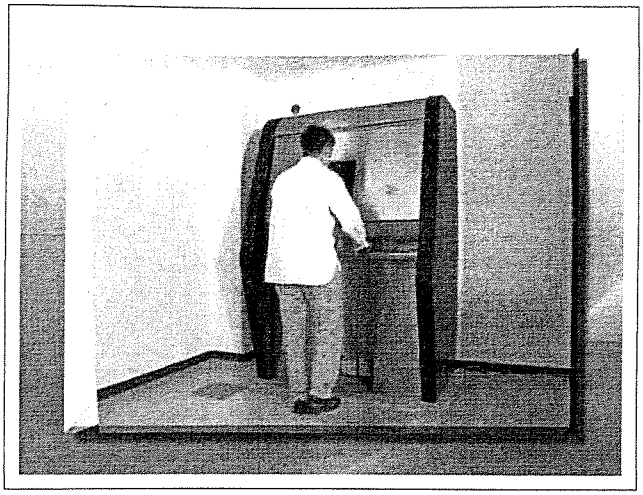


図1　3次元スキャナの外観（口絵カラー参照）

図3　特徴点抽出とワイヤフレームの整合結果（口絵カラー参照）

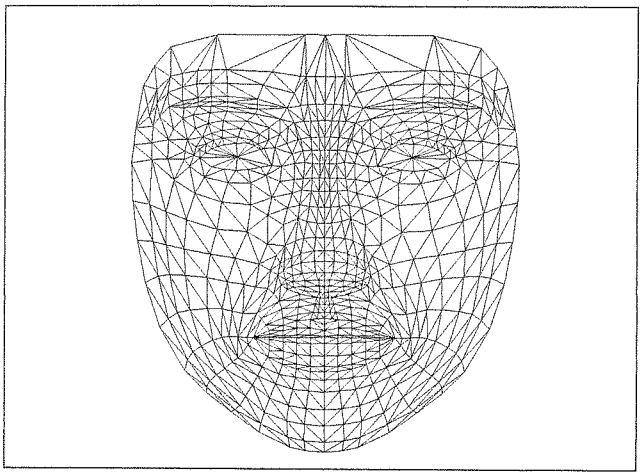
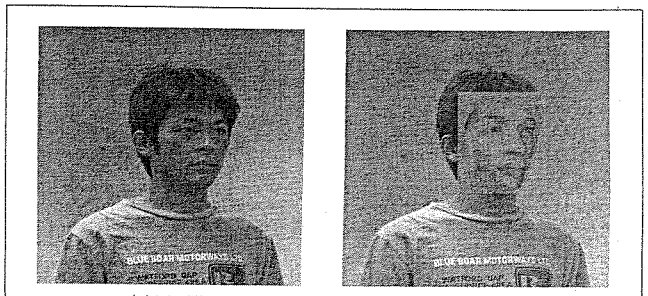


図2　標準ワイヤフレームモデル



(a)原画像　(b)個人CGモデル

図4　原画像のスキャンによる個人CGモデル生成（口絵カラー参照）

また正面のカメラで撮影された顔の画像データは，1画素ごとに奥行きデータとの対応が取られる．この各画素位置に配置された奥行きのデータおよび色のデータは，次の個人CGモデル生成プロセスに受け渡される．

3.2　個人CGモデル生成

奥行きデータから個人顔モデルを生成するために，標準となる顔ワイヤフレームを正面顔画像に整合する．図2に顔のワイヤフレームモデルを示す．

この際，画像処理によって，目，眉，唇，顎などの特徴点が自動計算される．奥行き情報は対応する画素位置から決定される．特徴点以外の顔ワイヤフレームの格子点については，補間ルールを適用して位置を決定する．このプロセスを経て個人向けの3次元の顔CGモデルが生成される．

図3は，特徴点抽出結果とワイヤフレームの整合結果を示す．またこの段階で正面画像の処理によって，男女判定および年齢推定が行われる．

図4に，来場者の原画像とテクスチャマッピングを施した個人CGモデルを示す．

3.3　表情合成

怒りや悲しみの表情，発話時の口形状など任意の表情の表現を行うために，基本となる表情パターンを予め基本形状として準備しておき，この基本形状のブレンドによって目的となる表情を表現する．基本表情パターンは，顔の標準ワイヤフレームの格子点の移動量を標準値として定義されており，標準ワイヤフレームを顔画像に整合させて個人適応させ変形した際には，その移動量も各個人の動きとして正規化が行われる．図5に表情合成結果を示す．

口の形状とともに，口内部の歯の位置も制御される．今回のアプリケーションでは，ストーリーは予め固定されているので，任意のシナリオに対応する必要はないが，インタラクティブなアプリケーションにも対応可能な構成になっている．

3.4　リアルタイム表情合成

今回の『グランドオデッセイ』というストーリーに基づき，各シーン毎の照明環境をはじめとする，レンダリングに必要な環境パラメータは，時系列として環境シナリオデータとして，予め映像作成時に記憶され保存されている．また，各映像フレームで，セリフに基づく口の動きのパラメータおよび表情のパラメータも，時系列としてキャストシナリ

図5　表情合成結果（口絵カラー参照）

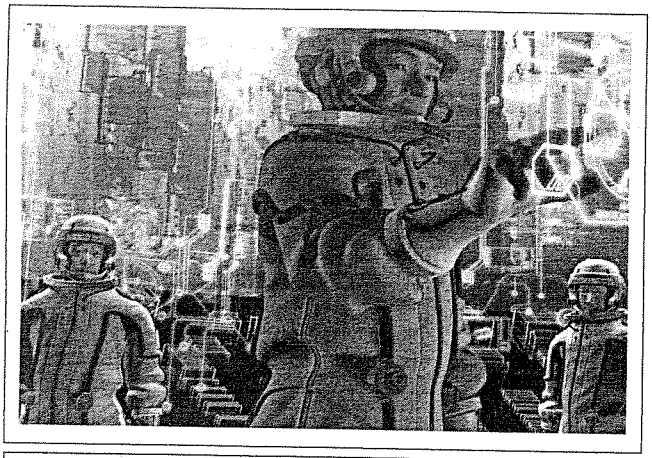


図6　ストーリーのひとこま（口絵カラー参照）

オデータとして保存されている．この口形状パラメータおよび表情パラメータは，先の基本表情パターンのそれぞれをどれだけの比率で混合するかというブレンド率のパラメータによって定義している．よってメインストーリーが開始されると，このレンダリングのための環境シナリオデータと，基本表情ブレンド率および顔の位置と向きの時系列であるキャストシナリオデータに基づいて，毎フレーム同期して，表情変形およびレンダリングが実行される．

また先の男女判別結果に基づいて，予め男女別の声優がそれぞれ発声し記録されている音声ストリームが選択される．年齢推定結果は，ソーティングされて配役の際に参考データとされる．よって，どの役になるかは，その都度来場者の年齢構成によって変化し，着席位置や，スキャニングの順番には左右されない．

4．演出上の工夫

三井・東芝館では一度の上映で，240人の観客を収容する必要があるが，240人全員を同一のスクリーン上に登場させることは，映像面積的にも，演技時間的にも不可能に近い．そこで，ストーリー序盤では，20人毎に仕切られた200インチ程度の分割スクリーン上で映像が展開し，このスクリーン上で20人の登場人物を，本人に自分自身がよく理解できるように大写しにされてストーリーはすすんでゆく．よって最初は12面のスクリーンが別々に出演者を変えて，同一のストーリー映像を流していることになる．

ストーリー途中で四つの分割シアターは，一つの合体したスクリーンとなって，よりダイナミックに映像を展開する．よってここで映像サイズは4倍となり，三つのストーリー映像が同時に進行するように切替わる．

最後の場面では，正面のスクリーンが下降して消滅し，その後ろから地球を囲んで他の二つのグループと対面してフィナーレを迎えるという演出を行っている．

図6は，今回のストーリーである『グランドオデッセイ』のひとこまである．登場人物の顔が，すべて来場者の顔に置き換わっており，照明条件も考慮され，顔のレンダリング結果は，シーンによくマッチしていることが見て取れる．

5．むすび

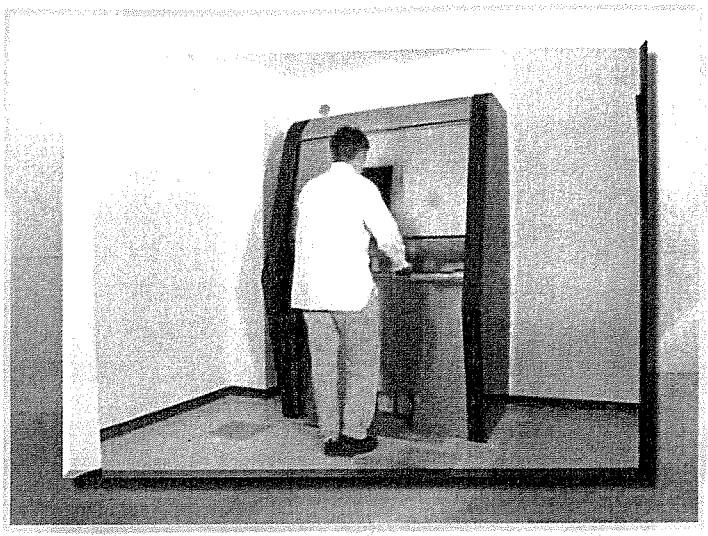
世界初のエンタテイメントシステムである，『フューチャーキャストシステム』の概要について述べた．これはアトラクションのストーリーに同期して登場人物の動きを定めたキャストシナリオデータを実時間で処理し，来場者の分身であるCGキャラクタを実時間制御して，表情や動きをつける点に特徴がある．よって来場者のキャラクタを即座にアトラクション映像に登場させ演技させることが可能となり，この結果，来場者へのストーリーへの没入感を増大させ，臨場感を飛躍的に増大させ，感動を呼び起こすことができる．

この技術は，エンタテイメントの一つの実現形態であると同時に，コンピュータビジョン，コンピュータグラフィックスなどの要素技術の集大成によって，初めて実現できるものである．個々の要素技術は，まだ進化を遂げている最中であるが，今後その進化の過程で，よりリアリティの高い，さらに大きな感動を生むエンタテイメントの創造に発展していくことは，理解に難くない．

是非とも，愛・地球博の三井・東芝館を訪れて頂き，多くの方々にこのまったく新しいエンタテイメント『フューチャーキャストシステム』を体験して，感動して頂くことを希望してやまない．

（2005年1月17日受付）

森島　繁生（もりしま　しげお）　1982年，東京大学工学部電子工学科卒業．1987年，同大学大学院博士課程修了．現在，早稲田大学理工学部応用物理学科教授．映像処理，信号処理，ヒューマンコンピュータインタラクションなどの研究に従事．CREST「コンテンツ制作の高能率化のための要素技術研究」プロジェクトリーダ．工学博士．正会員．



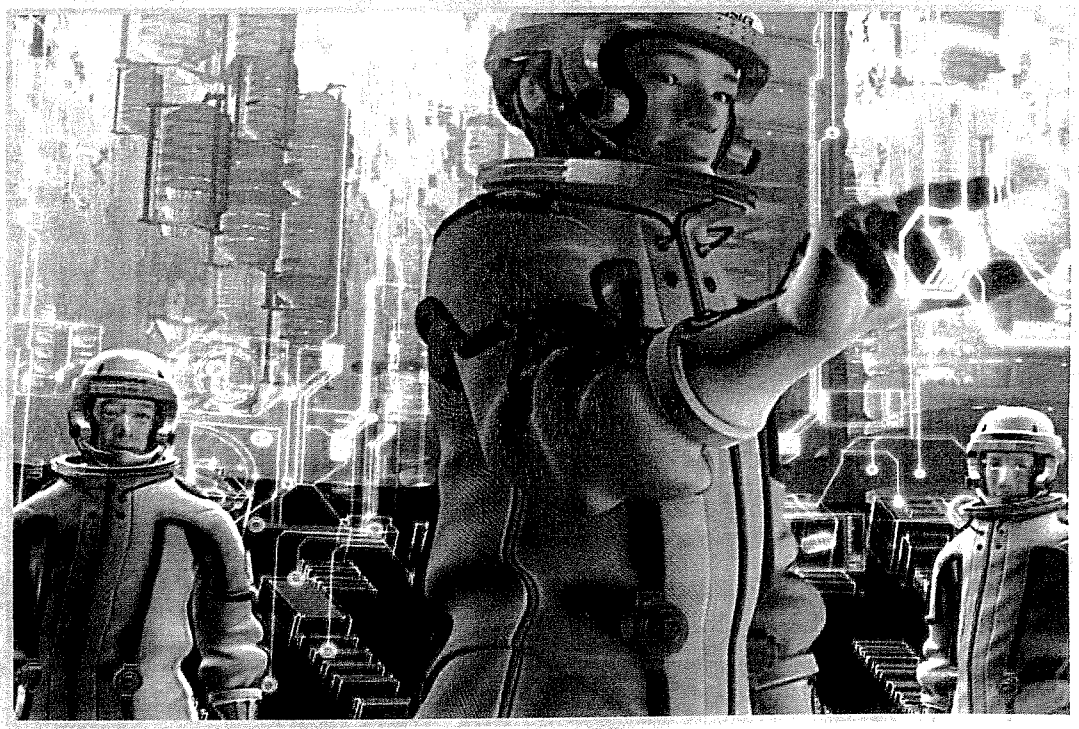
6章　図1　3次元スキャナの外観

6章　図3　特徴点抽出とワイヤフレームの整合結果

(a)原画像　　(b)個人CGモデル

6章　図4　原画像のスキャンによる個人CGモデル生成

6章　図5　表情合成結果



6章　図6　ストーリーのひとこま

Future Cast System ： 三井・東芝館

前島 謙宣† 森島 繁生†
†： 早稲田大学大学院 理工学研究科 物理学及応用物理学専攻
〒169-8555 東京都新宿区大久保3-4-1
Email ： {akinobu@toki. , shigeo@} waseda.jp

あらまし 映画の登場人物に扮して、宇宙を駆け巡る体験をしてみたい。あるいは正義の味方の役で物語りの主人公を演じてみたい。そういう希望は誰しも描いた経験が１度や２度はあると思う。本稿で紹介するフューチャーキャストシステムは、そのような希望を容易に叶える全く新しいエンターテインメントシステムである。三井・東芝館では、世界で初めてこれを実装し、シアターへの来場者全員である２４０名が出演できるアトラクションとして具現化している。従来、イベント会場や遊園地などに設置されている各種の映像アトラクションは、映像を来場者の視覚に一方的に提供するものであり、来場者がその映像中に没入する感覚を高め、臨場感を増す工夫がなされている。フューチャーキャストシステムは、実際に来場者を映像中に登場させ、さらに映像中で出演者として演技させることで、ストーリーへの没入感を増大させている点に特徴がある。

キーワード CGキャラクター、実時間表情アニメーション、没入間シアター、顔特徴認識、ブレンドシェープ

Future Cast System ： Mitsui-Toshiba pavilion

Akinobu MAEJIMA† Shigeo MORISHIMA†
†： Major in Pure and Applied Physics, Science and Engneering
Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
Email ： {akinobu@toki. , shigeo@} waseda.jp

abstract Most likely, everyone has gone through dreaming of acting a hero in the movie. The system called "Future Case System"(FCS) is a new entertainment system that is easily able to accomplish that dream. FCS has been implemented in MITSUI TOSHIBA pavilion at Expo 2005 Aichi Japan first in the world. This attraction is to allow 240 theater attendances to have a role of the movie they are watching. Comparing FCS with Previous Attraction Systems (PASs), PASs provide the visual information to visitors one-sidedly. On the other hand, since attendances' faces come out on the screen, they can feel as if they are acting in the movie. This is the main idea of FCS and makes attendances' feeling of immersion increase.

Keywords CG Character, Realtime Facial Animation, Immersive Theater, 3D Facial Range Scan ning, Facial Feature Recognition, Blend Shape

1．はじめに

映画の登場人物に扮して、宇宙を駆け巡る体験をしてみたい。あるいは正義の味方の役で物語りの主人公を演じてみたい。そういう願望は誰しも描いた経験が１度や２度はあると思う。ここで述べるフューチャーキャストシステムとは、そのような希望を叶えてくれる全く新しいエンターテインメントシステムである。三井・東芝館では、このフューチャーキャストシステムを世界で初めて実現し、シアターへの来場者全員である２４０名が出演できるアトラクションとして具現化している。本稿では、このフューチャーキャストシステムの概要について紹介する。

2．Future Cast System の特徴

従来、イベント会場や遊園地などに設置されている各種の映像アトラクションは、映像を来場者の視覚に一方的に提供するものであり、来場者がその映像中に没入する感覚を高め、臨場感を増す工夫がなされている。フューチャーキャストシステムは、実際に来場者を映像中に登場させ、さらに映像中で出演者として演技させることで、ストーリーへの没入感を増大させている点に特徴がある。いわば、物語の中における自分の分身の姿を、観客の側から客観的に見つめさせることで、感動を与える全く新しい発想のアトラクションである。

来場者は、まず２０人づつのグループに分割され、プレショールームに誘導される。ここでまず、顔の３次元スキャニングを行って、顔の立体形状が計測される。この立体形状は、鼻の高さや頬の膨らみなど個人の特徴を反映するものであり、本人のそっくりのＣＧキャラクターを生成する際に重要な情報となる。また同時に撮影される正面画像は、皮膚の色や目の色、男女の特徴、年齢特徴等を表現する情報として重要である。

プレショールームで来場者が説明を聞いている間に、本人のＣＧモデルが自動的に計算されスタンバイされる。また男女判定および年齢推定が自動的に行われ、ラベルとして情報が付与される。

メインシアターが開場となるや来場者は着席し、メインショーが開始される。ここで自分とそっくりのＣＧ化された登場人物が活躍するストーリー映像を体験することになる。あらかじめ制作されたストーリー映像と、ＣＧ合成された登場人物の映像がブレンドされスクリーンに映し出される。登場人物の映像はリアルタイムで合成されており、シナリオにしたがって演技する。また環境の変化もタイムリーに反映されるため、ストーリーに融合した違和感のない映像が影響される。観客そっくりの登場人物は、時にセリフセリフを喋り、時には感情を露わにして演技する。

途中、２０人づつであった小規模のシアターは、途中で仕切りが取り除かれ８０人シアターとなり、さらに物語は展開してゆく。フィナーレでは、正面のスクリーンが下降して空間が広がり、３つの８０人シアターが地球を囲んで合体する。１度のショーで２４０人の観客に感動をもたらすことが可能である。

3．キーとなる映像技術

フューチャーキャストシステムは、視聴者参加型のイベントシステムであるが、大人数の参加を可能とするためにオペレータの介入をほとんど必要としない点が特徴である。つまり全ての処理が全自動で実行される。

3.1　3次元スキャニング

半円上に配置された７台のディジタルカメラで構成される３次元スキャナを利用し、映像処理によって顔の３次元計測を実施する。３次元スキャナの外観を図１に示す。処理時間の制約から、画像サイズは現時点では縦640ピクセル、横480ピクセルに限定されているため奥行き情報の精度に限界はあるが、プロセッサーの高速化によって処理時間の短縮が図れれば、さらに画像解像度を高めることで、奥行き精度を現在以上に向上させることも可能である。またカメラ台数を追加し、さらに３６０度方向にカメラを配置することで、原理的に頭部全周囲の３次元計測も可能である。また正面のカメラで撮影された顔の画像データは、１画素ごとに奥行きデータとの対応が取られている。この各画素位置に配置された奥行きのデータおよび色のデータは、次の個人ＣＧモデル生成プロセスに受け渡される。

3.2　個人 CG モデルの生成

奥行きデータから個人顔モデルを生成するために、標準となる顔ワイヤフレームを正面顔画像に整合する。図２に顔の標準ワイヤフレームモデルを示す。

この際、画像処理によって、目、眉、唇、顎などの特

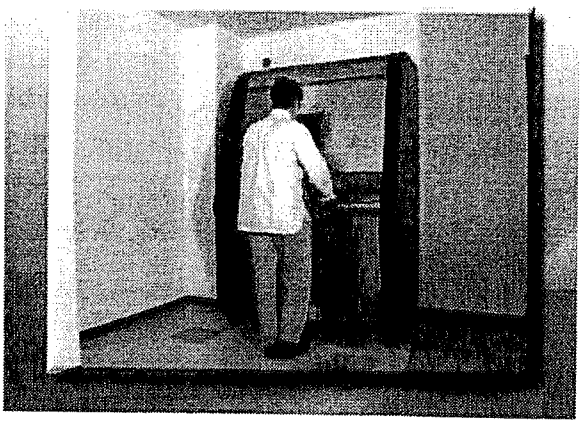


図１　３次元スキャナーの概観

徴点が自動計算される。特徴点の奥行きは、その対応するデータを読み取り、特徴点以外の顔ワイヤフレームの格子点については、補間ルールを適用して位置を決定する。このプロセスを経て個人向けの顔ＣＧモデルが生成される。

図３は、特徴点抽出結果とワイヤフレームの整合結果を示す。また正面画像の処理によって、男女判定および年齢推定が行われる。

図４に、来場者の原画像とテクスチャマッピングを施した個人ＣＧモデルをを示す。

3.3 表情合成

怒りや悲しみの表情、発話時の口形状など任意の表情の表現を行うために、基本となる表情パターンを予め基本形状として準備しておき、この基本形状のブレンドによって目的となる表情を表現する。基本表情パターンは、顔の標準ワイヤフレームの格子点の移動量を標準値として定義されており、標準ワイヤフレームを顔画像に整合させて個人適応させて変形した際には、その移動量も各個人の動きとして正規化が行われる。図５に表情合成結果を示す。

口の形状と共に、口内部の歯の位置も制御される。今回のアプリケーションでは、ストーリーは固定されているので、任意のシナリオに対応する必要はないが、インタラクティブなアプリケーションにも対応可能な構成になっている。

3.4 リアルタイム表情合成

今回の『グランドオデッセイ』というストーリーに基づき、各シーン毎の照明環境をはじめとするレンダリングに必要な環境パラメータは時系列として環境シナリオデータとして、予め映像作成時に記憶され保存されている。また各映像フレームで、セリフに基づく口の動きのパラメータおよび表情のパラメータも時系列としてキャストシナリオデータとして保存されている。この口形状パラメータおよび表情パラメータは先の基本表情パターンのそれぞれをどれだけの比率で混合するかというブレンド率によって定義している。よってメインストーリーが開始されると、このレンダリングのための環境シナリオデータと、基本表情ブレンド率の時系列であるキャストシナリオデータに基づいて、毎フレーム同期して、表情変形およびレンダリングが実行される。

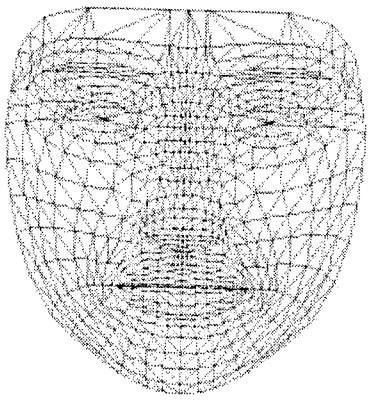


図２　標準顔ワイヤーフレームモデル

また先の男女判別結果に基づいて、予め男女の声優がそれぞれ発声した音声ストリームが選択される。年齢推定結果は、ソーティングされて配役の際に参考データとされる。よって、どの役になるかは、その都度、来場者の年齢構成によって変化し、着席位置や、スキャニングの順番には左右されない。

4．演出上の工夫

１度の上映で、２４０人の観客を収容する必要があるが、２４０人全員を同一のスクリーン上に登場させることは、映像面積的にも、時間的にも不可能に近い。そこ

図３　特徴点の抽出とワイヤーフレームの整合結果

で、ストーリー序盤では、２０人毎に仕切られた２００インチ程度の分割スクリーン上で映像が展開し、このスクリーン上で２０人の登場人物を、本人に自分自身がよく理解できるように大写しにされてストーリーはすすむ。よって最初は１２面のスクリーンが別々に出演者を変えて、同一の映像ストリームを流していることになる。ストーリー途中で４つの分割シアターは、１つの合体したスクリーンとなって、よりダイナミックに映像を展開する。よってここで映像サイズは４倍となり、３つの映像ストリームが同時に進行することになる。最後の場面では、正面のスクリーンが下降して消滅し、その後ろから地球を囲んで他の２つのグループと対面してフィナーレを迎えるという演出を行っている。

図６は、今回のグランドオデッセイのひとこまである。登場人物の顔が、すべて来場者の顔に置き換わっており、照明条件も考慮され、顔のレンダリング結果は、シーンにマッチしていることが見て取れる。

５．むすび

世界初のエンタテインメントシステムである、フューチャーキャストシステムの概要について述べた。これはアトラクションのストーリーに同期して登場人物の動きを定めたキャストシナリオデータを実時間で処理し、来場者のＣＧキャラクタを実時間処理して、動作させる点に特徴がある。よって来場者のキャラクタをアトラクション映像に登場させ演技させることが可能となり、この結果、来場者へのストーリーへの没入感を増大させ、臨場感を飛躍的に増大させることができる。

この技術は、エンタテインメントの１つの実現形態であると同時に、コンピュータビジョン、コンピュータグラフィックスなどの要素技術の集大成によって、初めて実現できるものである。個々の要素技術は、まだ進化を遂げている最中であるが、今後その進化の過程で、よりリアリティが高く、大きな感動を生むエンタテインメントの創造につながることは、理解に難くない。

是非とも、愛・地球博の三井・東芝館を訪れて頂き、多くの方々にこのフューチャーキャストシステムを体験して頂くことを希望してやまない。

図４　原画像のスキャンによる個人CGモデル生成
（左：原画像，右：個人CGモデル）

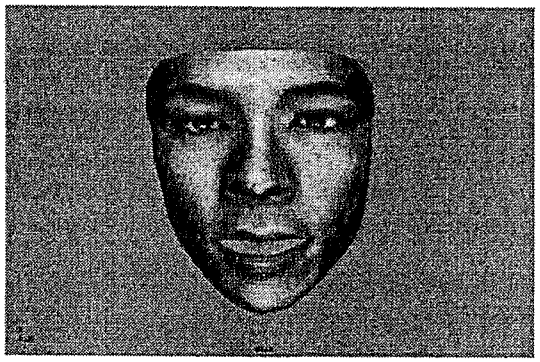


図５．表情合成結果

図６　ストーリーのひとこま

EURASIP Journal on Applied Signal Processing 2004:11, 1637–1647

Multimodal Translation System Using Texture-Mapped Lip-Sync Images for Video Mail and Automatic Dubbing Applications

Shigeo Morishima

School of Science and Engineering, Waseda University, Tokyo 169-8555, Japan
Email: shigeo@waseda.jp

ATR Spoken Language Translation Research Laboratories, Kyoto 619-0288, Japan

Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories, Kyoto 619-0288, Japan
Email: satoshi.nakamura@atr.jp

Received 25 November 2002; Revised 16 January 2004

We introduce a multimodal English-to-Japanese and Japanese-to-English translation system that also translates the speaker's speech motion by synchronizing it to the translated speech. This system also introduces both a face synthesis technique that can generate any viseme lip shape and a face tracking technique that can estimate the original position and rotation of a speaker's face in an image sequence. To retain the speaker's facial expression, we substitute only the speech organ's image with the synthesized one, which is made by a 3D wire-frame model that is adaptable to any speaker. Our approach provides translated image synthesis with an extremely small database. The tracking motion of the face from a video image is performed by template matching. In this system, the translation and rotation of the face are detected by using a 3D personal face model whose texture is captured from a video frame. We also propose a method to customize the personal face model by using our GUI tool. By combining these techniques and the translated voice synthesis technique, an automatic multimodal translation can be achieved that is suitable for video mail or automatic dubbing systems into other languages.

Keywords and phrases: audio-visual speech translation, lip-sync talking head, face tracking with 3D template, video mail and automatic dubbing, texture-mapped facial animation, personal face model.

1. INTRODUCTION

The facial expression is thought to send most of the nonverbal information in ordinary conversation. From this viewpoint, many researches have been carried on face-to-face communication using a 3D personal face model, sometimes called an "Avatar" in cyberspace [1].

For spoken language translation, ATR-MATRIX (ATR's multiligual automatic translation system for information exchange) [2] has been developed for the limited domain of hotel reservations between Japanese and English. A speech translation system has been developed for verbal information, although it does not take into account articulation and intonation. Verbal information is the central element in human communications, but the facial expression also plays an important role in transmitting information in face-to-face communication. For example, dubbed speech in movies has the problem that it does not match the lip movements of the facial image. In the case of making the entire facial image by computer graphics, it is difficult to send messages of original nonverbal information. If we could develop a technology that is able to translate facial speaking motion synchronized to translated speech where facial expressions and impressions are stored as effectively as the original, a natural multi-lingual tool could be realized.

There has been some research [3] on facial image generation to transform lip shapes based on concatenating variable units from a huge database. However, since images generally contain much larger information than that of sounds, it is difficult to prepare large image databases. Thus conventional systems need to limit speakers.

Therefore, we propose a method that uses a 3D wire-frame model to approximate a speaker's mouth region and captured images from the other regions of the face. This approach permits image synthesis and translation while storing the speaker's facial expressions in a small database.

If we replace only the mouth part of an original image sequence, the translation and rotation of head have to

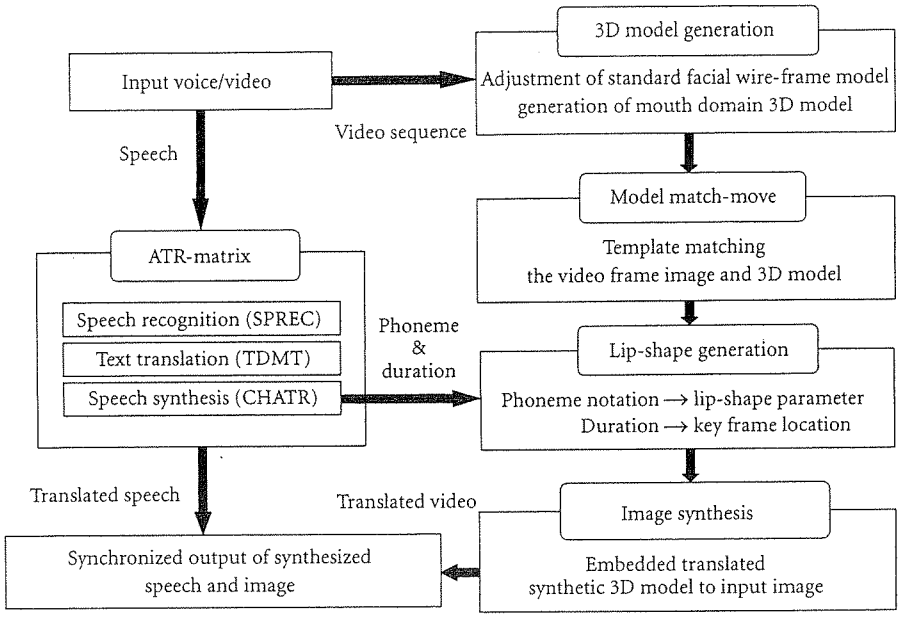


FIGURE 1: Overview of the system.

be estimated accurately while keeping smooth motion between frames. We propose a method to generate a 3D face model with a real personal face shape and to track face motion such as translation and rotation automatically for audio-visual speech translation. The method enables the proposed system to detect movement and rotation of the head from the 3D shape of the face by template matching using a 3D personal face wire-frame model.

We describe a speech translation system, the method to generate a 3D personal face model, an automatic face-tracking algorithm, and experiments to evaluate tracking accuracy. Finally, we show generated mouth motions that were never spoken by a speaker and introduce a method to evaluate lip synchronization.

2. OVERVIEW OF MULTIMODAL TRANSLATION

Figure 1 shows an overview of the system developed in this research. The system is divided broadly into two parts: the *speech translation part* and the *image translation part*.

The speech translation part is composed of ATR-MATRIX [2], which was developed in ATR-ITL. ATR-MATRIX is composed of ATR-SPREC to execute speech recognition, transfer-driven machine translation (TDMT) to handle text-to-text translation, and CHATR [4] to generate synthesized speech. The two parameters of phoneme notation and duration information, which are outputs from CHATR, are applied to facial image translation.

The first step of the image translation part is to make a 3D model of the mouth region for each speaker by fitting a standard facial wire-frame model to an input image. Because of the differences in facial bone structures, it is necessary to prepare a personal model for each speaker, but this process is required only once for each speaker.

The second step of the image translation part is to generate lip movements for the corresponding utterance. The 3D model is transformed by controlling the acquired lip-shape parameters so that they correspond to the phoneme notations from the database used at the speech synthesis stage. Duration information is also applied and interpolated linearly for smooth lip movement. Here, the lip-shape parameters are defined by a momentum vector derived from the natural face at lattice points on a wire frame for each phoneme. Therefore, this database does not need speaker adaptation.

In the final step of the image translation part, the translated synthetic mouth region 3D model is embedded into input images. In this step, the 3D model's color and scale are adjusted to the input images. Even if an input movie (image sequence) is moving during an utterance, we can acquire natural synthetic images because the 3D model has geometry information.

Consequently, the system outputs a lip-synchronized face movie for the translated synthetic speech and image sequence at 30 frames/second.

3. SPEECH TRANSLATION SYSTEM

The system is based on the speech-to-speech translation system developed at ATR [2]. This system is called ATR-MATRIX. The system consists of speech recognition, language translation, and speech synthesis modules. The speech recognition module is able to recognize naturally spoken utterances in the source language. The language translation module is able to translate the recognized utterances to sentences in the target language. Finally, the translated sentences are synthesized by the text-to-speech synthesis module. In the following section, each of the modules is described.

3.1. Speech recognition system

The long research history and continuous efforts of data collection at ATR have made a statistical model-based speech recognition module possible. The module is speaker-independent and able to recognize naturally spoken utterances. In particular, the system drives multiple acoustic models in parallel in order to handle differences in gender and speaking styles.

Speech recognition is achieved by the maximum a posterior (MAP) decoder, which maximizes the following probability:

$$\begin{aligned} \hat{W} &= \arg\max P(W|O) \\ &= \arg\max P(O|W)P(W). \end{aligned} \tag{1}$$

Here, $P(O|W)$ and $P(W)$ are called "acoustic model probability" and "language model probability, respectively."

Parameters of the acoustic model and the language model are estimated by speech data and text data. For the acoustic model, a hidden Markov model (HMM) is widely used. However, the conventional HMM has problems in generating optimal state structures. We devised a method called HMnet (hidden Markov network), which is a data-driven automatic state network generation algorithm. This algorithm iteratively increases the state network by splitting one state into two states by considering the phonetic contexts so as to increase likelihood [5]. Speech data is sampled at 16 kHz and 16 bits. Short-time Fourier analysis with a 20-millisecond-long window every 10 milliseconds is adopted. Then, after a Mel-frequency bandpass filter and log compression, the twelfth-order Mel-frequency cepstrum coefficients and their first- and second-order time derivatives are extracted. For acoustic model training, we used 167 male and 240 female speech samples of travel dialogue and phonetically balanced sentences. Total length of the training data is 28 hours. Using this data, the structure and parameters of the HMnet acoustic model are determined. The estimated model is composed of 1400 states with 5 Gaussian mixtures for speech and 3 states with 10 Gaussian mixtures for the silence model.

For the language model, the statistical approach called the "N-gram language model" is also widely used. This model characterizes a probability of the word occurrence by the conditional probability-based previous word history. A trigram language model defined by the previous two words is widely used. The length of "N" should be determined by considering the trade-off between the number of parameters and the amount of training data. Once we get a text corpus, the N-gram language model can be easily estimated. For the word triplets that occur infrequently, probability smoothing is applied. In our system, a word-class-based N-gram model is used. This method reduces the training data problems and the unseen triplets problem by using a word class as part-of-speech. For language model training, we used 7,000 transcribed texts from real natural dialogues in travel domain. The total number of words is 27,000. Using this text corpus, the class-based N-gram language model is estimated for 700 classes.

Finally, the speech recognition system searches the optimal word sequence using the acoustic models and the language models. The search is a time-synchronous two-pass search after converting the word vocabulary into a tree lexicon. The multiple acoustic models can be used in the search but get pruned by considering likelihoods scores.

The performances of speaker-independent recognition in the travel arrangement domain were evaluated. The word error rates for face-to-face dialogue speech, bilingual speech, and the machine-friendly speech are 13.4%, 10.1%, and 5.2%, respectively.

3.2. Speech synthesis system

The speech synthesis system generates natural speech from the translated texts. The speech synthesis system developed at ATR is called CHATR [6]. The CHATR synthesis relies on the fact that a speech segment can be uniquely described by the joint specification of its phonemic and prosodic environmental characteristics. The synthesizer performs a retrieval function, first predicting the information that is needed to complete a specification from an arbitrary level of input and then indicating the database segments that best match the predicted target specifications. The basic requirement for input is a sequence of phone labels, with associated fundamental frequency, amplitudes, and durations for each. If only words are specified in the input, then their component phones will be generated from a lexicon or by rule; if no prosodic specification is given, then a default intonation will be predicted from the information available.

The CHATR preprocessing of a new source database has two stages. First, an analysis stage takes as its input an arbitrary speech corpus with an orthographical transcription and then produces a feature vector describing the prosodic and acoustic attributes of each phone in that corpus. Second, a weight-training stage takes as its input the feature vector and a waveform representation and then produces a set of weight vectors that describe the contribution of each feature toward predicting the best match to a given target specification.

At synthesis time, the selection stage takes as its input the feature vectors, the weight vectors, and a specification of the target utterance to produce an index into the speech corpus for random-access replay to produce the target utterance.

3.3. Language translation system

The translation subsystem uses an example-based approach to handle spoken language [7]. Spoken language translation faces problems different from those of written language translation. The main requirements are (1) techniques for handling ungrammatical expressions, (2) a means for processing contextual expressions, (3) robust methods for speech recognition errors, and (4) real-time speed for smooth communication.

The backbone of ATR's approach is the translation model called TDMT [8], which was developed within an example-based paradigm. TDMT's constituent boundary parsing [9] provides efficiency and robustness. We have also explored the processing of contextual phenomena and a method for

TABLE 1: Quality and time.

Language conversion	Japanese-to-English	Japanese-to-German	Japanese-to-Korean	English-to-Japanese
A (%)	43.4	45.8	71.0	52.1
A + B (%)	74.0	65.9	92.7	88.1
A + B + C (%)	85.0	86.4	98.0	95.3
Time (seconds)	0.09	0.13	0.05	0.05

dealing with recognition errors and have made much progress in these explorations.

In TDMT, translation is mainly performed by a transfer process that applies pieces of transfer knowledge of the language pair to an input utterance. The transfer process is the same for each language pair, that is, Japanese-English, Japanese-Korean, Japanese-German, and Japanese-Chinese, whereas morphological analysis and generation processes are provided for each language, that is, Japanese, English, Korean, German, and Chinese.

The transfer process involves the derivation of possible source structures by a constituent boundary parser (CB-parser) [9] and a mapping to target structures. When a structural ambiguity occurs, the best structure is determined according to the total semantic distances of all possible structures. Currently, the TDMT system addresses dialogues in the *travel domain*, such as travel scheduling, hotel reservations, and trouble-shooting. We have applied TDMT to four language pairs: Japanese-English, Japanese-Korean [10], Japanese-German [11], and Japanese-Chinese [12]. Training and test utterances were randomly selected for each dialogue from our speech and language data collection, which includes about 40 000 utterances in the travel domain. The coverage of our training data differs among the language pairs and varies between about 3.5% and about 9%. A system dealing with spoken dialogues is required to realize a quick and informative response that supports smooth communication. Even if the response is somewhat broken, there is no chance for manual pre-/postediting of input/output utterances. In other words, both speed and informativity are vital to a spoken-language translation system. Thus, we evaluated TDMT's translation results for both *time* and *quality*.

Three native speakers of each target language manually graded translations for 23 dialogues (330 Japanese utterances and 344 English utterances, each about 10 words). During the evaluation, the native speakers were given information not only about the utterance itself but also about the previous context. The use of context in an evaluation, which is different from typical translation evaluations, is adopted because the users of the spoken-dialogue system consider a situation naturally in real conversation.

Each utterance was assigned one of four ranks for translation quality:

(A) perfect: no problem in either information or grammar;
(B) fair: easy to understand with some unimportant information missing or flawed grammar;
(C) acceptable: broken but understandable with effort;
(D) nonsense: important information has been translated incorrectly.

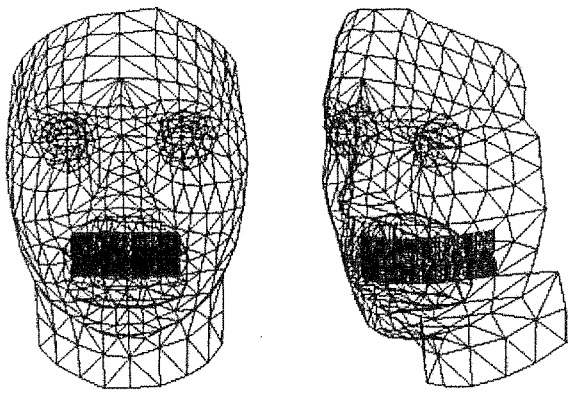


FIGURE 2: 3D head model.

Table 1 shows the latest evaluation results for TDMT, where the "acceptability ratio" is the sum of the (A), (B), and (C) ranks. The JE and JG translations achieved about 85% acceptability, and the JK and EJ translations achieved about 95% acceptability. JK's superiority is due to the linguistic similarity between the two languages; EJ's superiority is due to the relatively loose grammatical restrictions of Japanese.

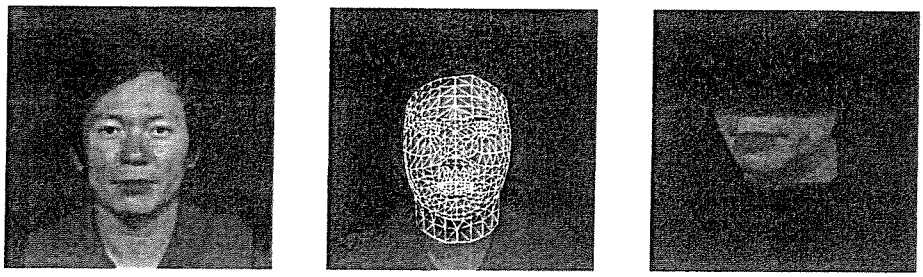
The translation speed was measured on a PC/AT Pentium II/450 MHz computer with 1 GB of memory. The translation time did not include the time needed for a morphological analysis, which is much faster than a translation. Although the speed depends on the amount of knowledge and the utterance length, the average translation times were around 0.1 seconds. Thus, TDMT can be considered efficient.

4. GENERATING PERSONAL FACE MODEL

It is necessary to make an accurate 3D model that has the target person's features for the face recreation by computer graphics. In addition, there is demand for a 3D model that does not need heavy calculation load for synthesis because this model is used for both generating a face image and tracking face location, size, and angle.

In our research, we used the 3D head model [13, 14] shown in Figure 2 and tried to make a 3D model of the mouth region. This 3D head model is composed of about 1,500 triangular patches and has about 800 lattice points.

The face fitting tool developed by IPA (Facial image processing system for Human-like "kansei" Agent, http://www.tokyo.image-lab.or.jb/aa/ipa/) is often used to generate a 3D face model using one photograph. However, the manual fitting algorithm of this tool is very difficult and requires a lot of time for users to generate a 3D model with a real personal



(a) (b) (c)

FIGURE 3: 3D model generation process. (a) Input image. (b) Fitting result. (c) Mouth model.

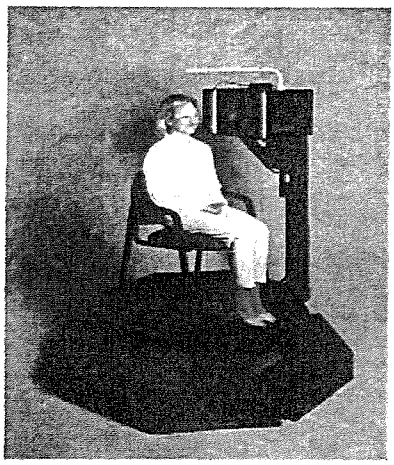


FIGURE 4: Head and face 3D color range scanner.

face, although it is able to generate a model with a nearly real personal shape along with many photographs. Figure 3 shows a personal face model. Figure 3a is an original face image, Figure 3b shows the fitting result of a generic face model, and Figure 3c is the mouth model constructed by a personal model used for the mouth synthesis in lip synchronization.

In order to raise the accuracy of face tracking by using the 3D personal face model, we used a range scanner like *Cyberware* [14], shown in Figure 4. This is a head-and-face 3D color scanner that can capture both range data and texture as shown in Figure 5.

We can generate a 3D model with a real personal shape by using a standard face wire-frame model. First, to fit the standard face model to the *Cyberware* data, both the generic face model and the *Cyberware* data are mapped to a 2D cylindrical plane. Then, we manually fit a standard model's face parts to the corresponding *Cyberware* face parts by using texture data. This process is shown in Figure 6. Finally, we replace the coordinate values of the standard model to *Cyberware* range data coordinates values and obtain an accurate 3D personal face model shown in Figure 7.

The face fitting tool provides a GUI that helps the user to fit a generic face wire-frame model onto texture face data accurately and consistently with coarse-to-fine feature points selection.

FIGURE 5: Acquired shape and texture.

(a) (b)

FIGURE 6: Face parts fitting on 2D plane. (a) Before fitting. (b) After fitting.

5. AUTOMATIC FACE TRACKING

Many tracking algorithms have been studied for a long time, and many of them have been applied to tracking a mouth edge, an eye edge, and so on. However, because of such problems as blurring of the feature points between frames or occlusion of the feature points by rotation of a head, these algorithms have not been able to provide accurate tracking.

In this chapter, we describe an automatic face-tracking algorithm using a 3D face model. The tracking process using template matching can be divided into three steps.

First, texture mapping of one of the video frame images is carried out using the 3D individual face shape model created in Section 3. Here, a frontal face image is chosen from the video frames for the texture mapping.

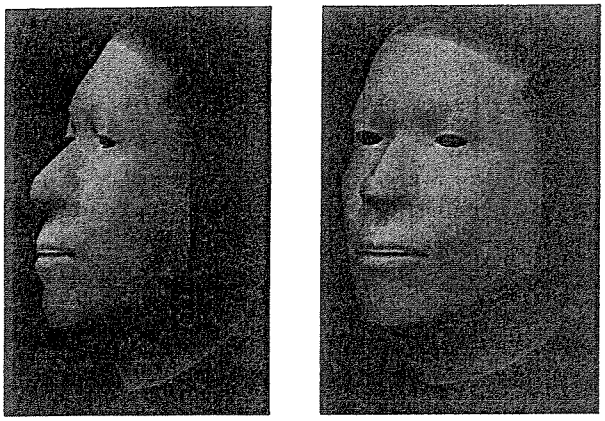


FIGURE 7: Generated 3D personal model.

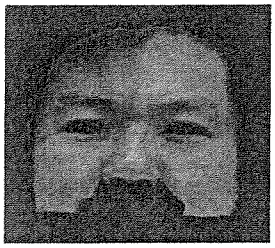


FIGURE 8: Template face image.

FIGURE 9: Template matching mechanism.

Next, we make 2D template images for every translation and rotation by using the 3D model shown in Figure 8. Here, in order to reduce matching errors, the mouth region is excluded from a template image. Consequently, even while the person in a video image is speaking something, tracking can be carried out more stably.

Expression change also can be handled by modifying a 3D template with the face synthesis module. The face synthesis module in the IPA face tool (http://www.tokyo.image-lab.or.jb/aa/ipa) can generate a stereotype face expression and also introduce personal character.

Currently, the test video image sequence includes slight and ordinary expression change but does not include an emotional expression. Therefore, modifying the face template is not currently considered, and the face template is treated as a rigid body.

Finally, we carry out template matching between the template images and an input video frame image and estimate translation and rotation values so that matching errors become minimum. This process is illustrated in Figure 9.

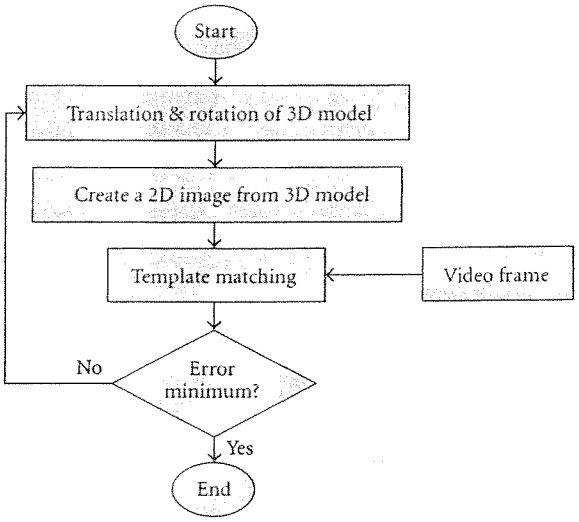


FIGURE 10: Flow of face tracking.

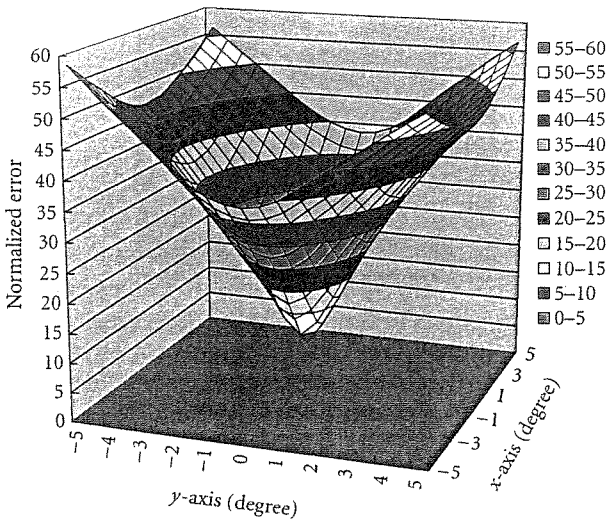


FIGURE 11: Error graph for rotation.

We show a flow chart for the search process of a face position and a rotation angle in one frame in Figure 10. The template matching for the tracking is carried out by using a euclid error function in the RGB value of all pixels normalized by the pixel number within a template.

Since template matching is performed only in the face region except for the blue back of template images and thus the number of pixels is different for each template image, we apply normalization in the error function based on the number of pixels.

By searching for a certain area, we obtain an error graph as shown in Figure 11. An approximation shows that there is only one global minimum. Therefore, we set initial values of the position and angle to those in the previous frame and search for desired movement and rotation from a $3^n - 1$ hypothesis near the starting point. We show a conceptual figure of the minimum error search in Figure 12.

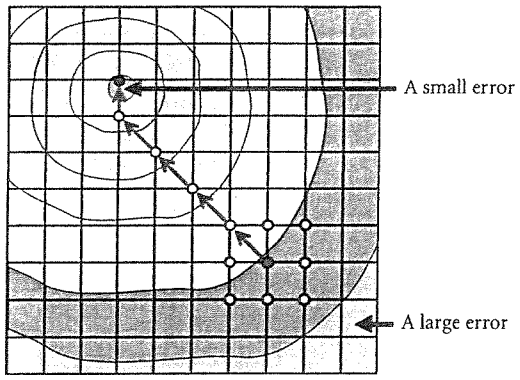


FIGURE 12: $3^n - 1$ gradient error search.

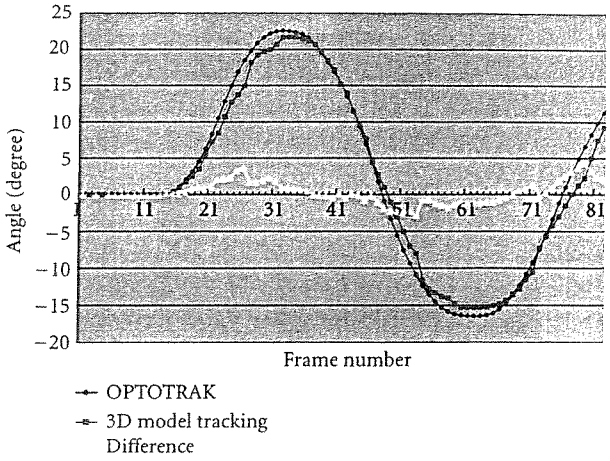


FIGURE 13: Evaluation of rotation angle with y-axis.

6. EVALUATION OF FACE TRACKING

We carried out tracking experiments to evaluate the effectiveness of the proposed algorithm.

6.1. Measurement by OPTOTRAK

To evaluate the accuracy of our tracking algorithm, we measured the face movement in a video sequence using *OPTOTRAK* (see [15]), the motion measurement system. We measured the following head movements:

(1) rotation of x-axis,
(2) rotation of y-axis,
(3) rotation of z-axis,
(4) movement of x direction.

In the following, we treat the data obtained by OPTOTRAK as the correct answer value for tracking.

6.2. Evaluation of the tracking

As an example of a tracking result, a graph computing the rotation angle to y-axis is shown in Figure 13. The average of the angle error between the angle obtained by our algorithm and that by OPTOTRAK is about 0.477 (degree).

This graph shows that the error increases as the rotation angle becomes large. This is because the front image is mapped on the 3D model.

An example of a model matching movement in a video frame is shown in Figure 14. The top row is the original video frame chosen from the sequence randomly. The second row is a synthetic face according to the position and rotation angle estimated by our algorithm. The third row is the image generated by replacing the original face with a synthetic one. From a subjective test, the quality of the synthesized image sequence looks so natural that it is impossible to distinguish the replacement face from the original one.

6.3. Processing speed

The system configuration is as follows:

(i) CPU: Xeon 2 GHz;
(ii) Memory: 1 GB;
(iii) OS: Microsoft Windows 2000 Professional;
(iv) VGA: 3D labs Wild Cat 5110.

In the first frame of the video sequence, it takes about 30 seconds because a full screen search is needed. In the succeeding frames with little head motion, the searching region is limited locally to the previous position so this becomes 3 seconds. When the head motion becomes bigger, the searching path becomes deeper and convergence takes a longer time of up to 10 seconds.

Currently, this is too slow to realize a real-time application, but the delay time is only one video frame theoretically, so a higher-speed CPU and video card can overcome this problem in the future.

7. LIP SHAPE IN UTTERANCE

When a person says something, the lips and jaw move simultaneously. In particular, the movements of the lips are closely related to the phonological process, so the 3D model must be controlled accurately.

As with our research, Kuratate et al. [16] tried to measure the kinematical data by using markers on the test subject's face. This approach has the advantage of accurate measurement and flexible control. However, it depends on the speaker and requires heavy computation. Here, we propose a method by unit concatenation based on the 3D model, since the lip-shape database is adaptable to any speaker.

7.1. Standard lip shape

For accurate control of the mouth region's 3D model, Ito et al. [14] defined seven control points on the model. These are shown in Figure 15. Those points could be controlled by geometric movement rules based on the bone and muscle structure.

In this research, we prepared reference lip-shape images from the front and side. Then, we transformed the wire-frame model to approximate the reference images. In this process, we acquired momentum vectors of lattice points on the wire-frame model. Then, we stored these momentum vectors in the lip-shape database. This database is normalized by the mouth region's size, so we do not need speaker adaptation. Thus, this system has achieved talking face generation with a small database.

FIGURE 14: Examples of model match-move.

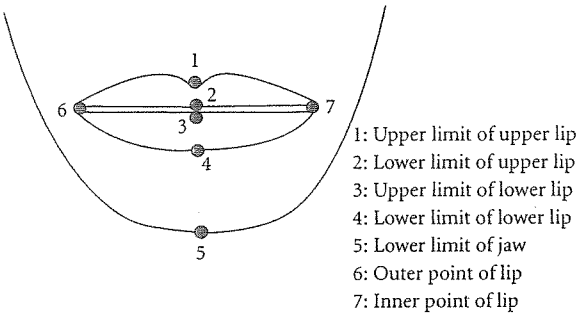


FIGURE 15: Positions of control points.

7.2. Lip-shape classification by viseme

Viseme is a word created from "phoneme," which is the smallest linguistic sound unit. Visemes are generally also defined for lip movement information like [au] and [ei] of the phonetic alphabet, but in this research we decomposed those visemes further into shorter and more static units.

We classified English phonemes into 22 kinds of parts based on visemes. In addition to English, we classified Japanese vowel phonemes into 5 kinds of parts. We also prepared a silent interval viseme. Table 2 shows the correspondences of these 28 visemes to the phonemic notation outputs from CHATR.

The system has as many standard lip-shapes in its database as the number of visemes. Japanese consonant lip-shape data come from 60% of the standard English consonant lip-shape data.

In English phonemes, some kinds of visemes are composed of multiple visemes. For example, these include [au], [ei], and [ou] of the phonetic alphabet. As stated previously, those visemes are decomposed into standard lip-shapes. We call these multiplicate visemes.

Each parameter of phonemic notations from CHATR has duration information. Furthermore, the decomposed visemes need to be apportioned by duration information. We experimentally apportioned 30% of the duration information to the front part of multiplicate visemes and the residual duration information to the back part of them.

7.3. Utterance animation

The lip-shape database of this system is defined by only the momentum vector of lattice points on a wire frame. However, there are no transient data among the standard lip shapes. In this section, we describe an interpolation method for lip movement by using duration information from CHATR.

The system must have momentum vectors of the lattice point data on the wire-frame model while phonemes are being uttered. Therefore, we defined that the 3D model configures a standard lip shape when a phoneme is uttered at any point in time. This point is normally the starting point of a phoneme utterance, and we defined the keyframe at the starting point of each phoneme segment.

Thereafter, we assign a 100% weight of the momentum vector to the starting time and a 0% weight to the ending time and interpolate these times by a sinusoidal curve between them.

For the next phoneme, the weight of the momentum vector is transformed from 0% to 100% as well as the current phoneme. By a value of the vector sum of these two weights, the system configures a lip shape that has a vector unlike any in the database. Although this method is not directly connected with kinesiology, we believe that it provides a realistic lip-shape image. The sinusoidal interpolation is expressed as follows. When a keyframe lip-shape vector is defined as V_n located at $t = t_n$, and a previous keyframe vector is defined as V_{n-1} at $t = t_{n-1}$, an interpolation between these keyframes is realized by the weight W_n, and the lip-shape vector is described as $M(t)$:

$$M(t) = W_n(V_{n-1} - V_n) + 0.5(V_{n-1} + V_n), \quad t = [t_{n-1}, t_n],$$
$$\text{where } W_n = 0.5\cos\left(\frac{t - t_{n-1}}{t_n - t_{n-1}}\right)\pi. \tag{2}$$

TABLE 2: Classification of visemes.

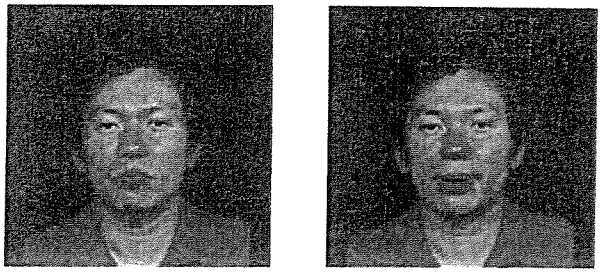
Viseme number	Phoneme notation from CHATR	
1	/ae/	
2	/ah/, /ax/	
3	/A/	
4	/aa/	
5	/er/, /ah r/	
6	/iy/, /ih/	
7	/uh/	
8	/uw/	
9	/eh/	
10	/oh/, /ao/	
11	/ax r/	English
12	/l/	
13	/r/	
14	/b/, /p/, /m/	
15	/t/	
16	/d/, /n/	
17	/k/, /g/, /hh/, /ng/	
18	/f/, /v/	
19	/s/, /z/, /sh/, /zh/, /ts/, /dz/, /ch/, /jh/	
20	/th/, /dh/	
21	/y/	
22	/w/	
23	/a/, /A/	
24	/i/, /I/	
25	/u/, /U/	Japanese
26	/e/, /E/	
27	/o/, /O/	
28	/#/	Silence interval

8. EVALUATION EXPERIMENTS

We carried out subjective experiments to evaluate effectiveness of the proposed image synthesis algorithm. Figures 16 and 17 show examples of the translated speaking face image. In order to clarify the effectiveness of the proposed system, we carried out subjective digit discrimination perception tests. The test audio-visual samples are composed of connected 4 to 7 digits in Japanese.

We tested using original speech and speaking face movies in speech. The original speech is used under the audio conditions of SNR = −6, −12, −18 dB using white Gaussian noise. Figure 18 shows the results. Subjects are 12 Japanese students (10 males and 2 females) in the same laboratory. Discrimination rate means the rate users can recognize each digit accurately by listening with headphones.

In every case, according to the low audio SNR, the subjective discrimination rates degrade. "Voice only" is only playback of speech without video. Even in the case of



(a) (b)

FIGURE 16: Translated synthetic image from Japanese to English. (a) Original image. (b) Synthetic image.

(a) (b)

FIGURE 17: Translated synthetic image from English to Japanese. (a) Original image. (b) Synthetic image.

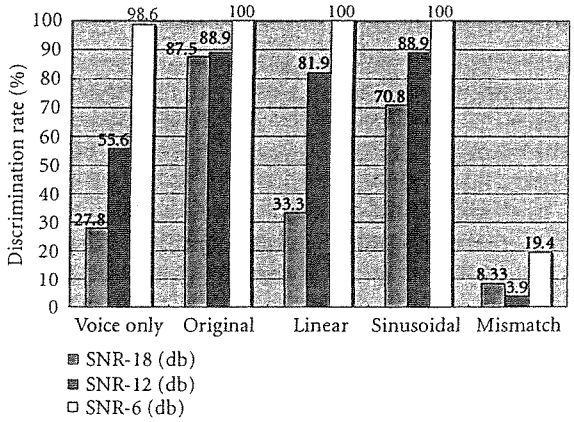


FIGURE 18: Subjective digit discrimination rate: evaluation test result.

SNR = −6 dB, the discrimination rate is not 100%. However, by adding a matched face movie, the rate becomes 100% in all cases. "Original" is a combination of the original voice and the video-captured natural face image. In this case, even at −18 dB, a high discrimination rate can be achieved. "Linear" indicates linear interpolation of keyframe parameters of the basic mouth shape. Lip-shape vector $M(t)$ is expressed as

follows:

$$M(t) = V_{n-1} + \alpha t \quad t = [t_{n-1}, t_n], \tag{3}$$

where $\alpha = (V_n - V_{n-1})/(t_n - t_{n-1})$. "Sinusoidal" is nonlinear interpolation using a sinusoidal curve between keyframes as described in Subsection 7.3.

"Mismatch" is using a digit voice and an unsynchronized video-captured face saying another digit number. The discrimination rate drastically degrades in the case of "Mismatch" between voice and image, even at −6 dB.

As a result nonlinear interpolation using a sinusoidal curve while considering coarticulation is able to reach a high score, and the proposed system significantly enhances perception rates. This method provides a good standard for evaluation of lip synchronization. A better interpolation method for lip synchronization will be pursued in order to more closely match the original image sequence.

9. CONCLUSIONS

As a result of this research, we propose a multimodal translation system that is effective for video-mail or applications for automatic dubbing into other languages. For a video phone application, a few seconds delay is inevitable depending on speech recognition and translation algorithm, and this will never be overcome theoretically, therefore real-time telecommunication cannot be realized. Video tracking requires high cost now, but this will be overcome by increases of CPU power.

Currently, speech recognition and machine translation strongly depend on context. However, the size of context will grow bigger and bigger, and context-independent systems will be realized in the future by changing databases.

Our proposed system can create any lip shape with an extremely small database, and it is also speaker-independent. It retains the speaker's original facial expression by using input images besides those of the mouth region. Furthermore, this facial-image translation system, which is capable of multimodal English-to-Japanese and Japanese-to-English translation, has been realized by applying the parameters from CHATR. This is a hybrid structure of the image-based and CG-based approaches to replace only the part related to verbal information.

In addition, because of the different durations between original speech and translated speech, a method that controls duration information from the image synthesis part to the speech synthesis part needs to be developed.

A method to evaluate lip synchronization was proposed and it will provide a standard method for lip-sync performance evaluation.

REFERENCES

[1] T. Yotsukura, E. Fujii, T. Kobayashi, and S. Morishima, "Generation of a life-like agent in cyberspace using media conversion," IEICE Tech. Rep. MVE97-103, Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, pp. 75–82, 1998.

[2] T. Takezawa, T. Morimoto, Y. Sagisaka, et al., "A Japanese-to-English speech translation system: ATR-MATRIX," in *Proc. 5th International Conference on Spoken Language Processing (ICSLP '98)*, pp. 2779–2782, Sydney, Australia, November–December 1998.

[3] H. P. Graf, E. Cosatto, and T. Ezzat, "Face analysis for the synthesis of photo-realistic talking heads," in *Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG '00)*, pp. 189–194, Grenoble, France, March 2000.

[4] N. Campbell and A. W. Black, "CHATR: a multi-lingual speech re-sequencing synthesis system," IEICE Tech. Rep. SP96-7, Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, pp. 45–52, 1995.

[5] M. Ostendorf and H. Singer, "HMM topology design using maximum likelihood successive state splitting," *Computer Speech and Language*, vol. 11, no. 1, pp. 17–41, 1997.

[6] W. N. Campbell, "CHATR: A high definition speech re-sequencing system," in *Proc. 3rd ASA/ASJ Joint Meeting*, pp. 1223–1228, Honolulu, Hawaii, USA, December 1996.

[7] E. Sumita, S. Yamada, K. Yamamoto, et al., "Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach," in *Proc. 7th Machine Translation Summit VII*, pp. 229–235, Singapore, Singapore, September 1999.

[8] O. Furuse, J. Kawai, H. Iida, S. Akamine, and D. Kim, "Multilingual spoken-language translation utilizing translation examples," in *Proc. 3rd Natural Language Processing Pacific Rim Symposium (NLPRS '95)*, pp. 544–549, Seoul, Korea, December 1995.

[9] O. Furuse and H. Iida, "Incremental translation utilizing constituent boundary patterns," in *Proc. 16th International Conference on Computational Linguistics (COLING '96)*, vol. 1, pp. 412–417, Copenhagen, Denmark, August 1996.

[10] E. Sumita and H. Iida, "Experiments and prospects of example-based machine translation," in *Proc. 29th Annual Meeting of the Association for Computational Linguistics (ACL '91)*, pp. 185–192, Berkeley, Calif, USA, June 1991.

[11] M. Paul, E. Sumita, and H. Iida, "Field structure and generation in transfer-driven machine translation," in *Proc. 4th Annual Meeting of the NLP*, pp. 504–507, Fukuoka, Japan, 1998.

[12] K. Yamamoto and E. Sumita, "Feasibility study for ellipsis resolution in dialogues by machine learning techniques," in *Proc. 17th International Conference on Computational Linguistics-Annual Meeting of the Association for Computational Linguistics (COLING-ACL '98)*, pp. 1428–1435, Montreal, Quebec, Canada, August 1998.

[13] K. Ito, T. Misawa, J. Muto, and S. Morishima, "3D head model generation using multi-angle images and facial expression generation," IEICE Tech. Rep. 582, Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, pp. 7–12, 2000.

[14] K. Ito, T. Misawa, J. Muto, and S. Morishima, "3D lip expression generation by using new lip parameters," IEICE Tech. Rep. A-16-24, Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, pp. 328, 2000.

[15] T. Misawa, K. Murai, S. Nakamura, and S. Morishima, "Automatic face tracking and model match-move automatic face tracking and model match-move in video sequence using 3D face model in video sequence using 3D face model," in *Proc. IEEE International Conference on Multimedia and Expo (ICME '01)*, pp. 234–236, Tokyo, Japan, August 2001.

[16] T. Kuratate, H. Yehia, and E. Vatikiotis-Bateson, "Kinematics-based synthesis of realistic talking faces," in *Proc. International Conference On Auditory-Visual Speech Processing (AVSP '98)*, pp. 185–190, Terrigal, New South Wales, Australia, December 1998.

Shigeo Morishima was born in Japan on August 20, 1959. He received the B.S., M.S., and Ph.D. degrees, all in electrical engineering, from the University of Tokyo, Japan, in 1982, 1984, and 1987, respectively. From 1987 to 2001, he was an Associate Professor and from 2001 to 2004, a Professor at Seikei University, Tokyo. Currently, he is a Professor at School of Science and Engineering, Waseda University. His research interests include computer graphics, computer vision, multimodal signal processing, and human computer interaction. Dr. Morishima is a Member of the IEEE, ACM SIGGRAPH, and the Institute of Electronics, Information and Communication Engineers, Japan (IEICE-J). He is a Trustee of Japanese Academy of Facial Studies. He received the IEICE-J Achievement Award in May, 1992 and the Interaction 2001 Best Paper Award from the Information Processing Society of Japan in February 2001. He was having a sabbatical staying at University of Toronto from 1994 to 1995 as a Visiting Professor. He is now a Temporary Lecturer at Meiji University and Seikei University, Japan. Also, he is a Visiting Researcher at ATR Spoken Language Translation Research Laboratories since 2001 and ATR Media Information Science Laboratory since 1999.

Satoshi Nakamura was born in Japan on August 4, 1958. He received the Ph.D. degree in information science from Kyoto University in 1992. He worked with ATR Interpreting Telephony Research Laboratories from 1986 to 1989. From 1994 to 2000, he was an Associate Professor at the Graduate School of Information Science, Nara Institute of Science and Technology, Japan. In 1996, he was a Visiting Research Professor of the CAIP Center of Rutgers University of New Jersey, USA. He is currently the Head of Acoustics and Speech Research Department at ATR Spoken Language Translation Laboratories, Japan. He also serves as an Honorary Professor at University Karlsruhe, Germany, since 2004. His current research interests include speech recognition, speech translation, spoken dialogue systems, stochastic modeling of speech, and microphone arrays. He received the Awaya award from the Acoustical Society of Japan in 1992, and the Interaction 2001 Best Paper Award from the Information Processing Society of Japan in 2001. He served as an Editor for the Journal of the IEICE Information from 2000 to 2002. He is currently a Member of the Speech Technical Committee of the IEEE Signal Processing Society.

Future Cast System

Shigeo Morishima* Waseda University
Akinobu Maejima** Waseda University
Shuhei Wemler*** Siliconstudio Corp.
Tamotsu Machida**** VINO AZUL Inc.
Masao Takebayashi***** DENTSU TEC Inc.

1 Introduction

Most likely, everyone has gone through dreaming of acting a hero in the movie. The system called "Future Case System"(FCS) is a new entertainment system that is easily able to accomplish that dream. FCS has been implemented in MITSUI TOSHIBA pavilion at Expo 2005 Aichi Japan first in the world. This attraction is to allow 240 theater attendances to have a role of the movie they are watching. Comparing FCS with Previous Attraction Systems (PASs), PASs provide the visual information to visitors one-sidedly. On the other hand, since attendances' faces come out on the screen, they can feel as if they are acting in the movie. This is the main idea of FCS and makes attendances' feeling of immersion increase.

2 Overview of Future Cast System

Fig 1 shows the overview of FCS. First, FCS requires the attendances to be scanned by a range-scanner to measure their 3D face information. This 3D information has attendance's personal characteristic such as nasal height and tumor of cheek. Then FCS automatically constructs a CG character face based on this information. Attendance's facial texture is captured by front camera of the range scanner at the same time. Based on this information, personal information that includes skin color, iris color, gender, and age are estimated from the facial texture in the same process. The cast is decided by estimated information. CG character's face is synthesized in real-time by pre-making scenario data that correspond to each scene of the movie. Finally, the movie is screened with blending the synthesized CG character's face and pre-rendered movie.

3 Scanning 3D information of personal face

The range-scanner consists of seven digital cameras that are placed in a half circle. Using this scanner, the 3D information of attendance's face is measured by image processing. Due to constraint of processing time, FCS only deals with front face data. However, this algorithm can handle 3D information of all around personal head data if this system uses more cameras. Additionally, the facial texture of attendance's front face corresponds to 3D data on each pixel. The facial texture and 3D data transfer to the synthesis process.

4 Generating 3D personal face model

In order to generate realistic CG character from attendance's personal 3D face information, the 3D standard wire frame model is fitted to the facial texture automatically. In this fitting process, facial feature points such as eye contour, eyebrows and chin are detected by image processing automatically. Mesh vertex of the wire frame are placed based on our interpolation rule. Through this process, personal face model is generated. In addition, this process carries out estimation of gender, age, and iris color.

*email: shigeo@waseda.jp
**email: akinobu@toki.waseda.jp
***email: wemler@siliconstudio.co.jp
****email: tmach@tkk.att.ne.jp
*****email: take-m@bb.e-mansion.com

5 Real time facial animation and Rendering

In order to express facial emotions such as happiness, or mouth shapes with utterance, we prepare basic facial emotional patterns and mouth shapes called "Target Shape" and FCS synthesizes facial expression by blending target shapes. Based on the pre-planned story called "Grand Odyssey" we prepare Environment Scenario Data (ESD) and Cast Scenario Data (CSD) that has time-code data for each scene of the story. ESD defines lighting environment. CSD defines the blend ratio of each target shape for each CG character. In synchronization with Facial expression, rendering is performed based on ESD and CSD for each frame. The cast is selected by age and gender information after sorting. One of the pre-recorded (pre-scoring) voice track is selected by gender recognition result. That is, the cast is changed by age composition and the ratio of male and female of all attendances. There is no relation with the order of scanning. This is the big advantage of FCS to make a game feeling and a repeater. Even though attendances come back to FCS, they may be able to enjoy another role and watch unique movie each time.

6 Results

Fig 2 shows one scene of "Grand Odyssey". All of the faces are replaced to attendances' faces. The rendering result based on ESD seems to match to the movie seamlessly. In this paper, we have described the world's first entertainment system called Future Cast System. FCS makes attendances appear on a pre-planned movie as a CG character. As a result, attendances can be absorbed in the movie and feel high realistic sensation and immersion feeling in the story. In addition, if there are pre-planned movie and scenario data, FCS can be applied to any attractions. Improving the FCS to personalize not only attendance's face but also body and voice is a future work.

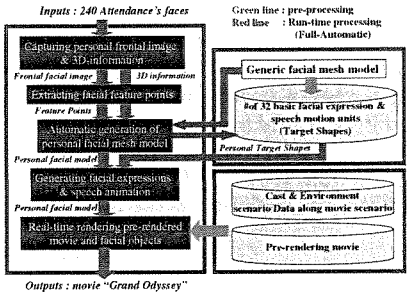


Figure 1: Overview of the Future Cast System

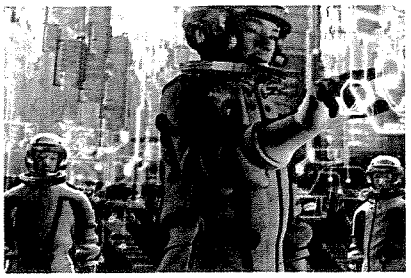


Figure 2: One scene of "Grand Odyssey"

TVRSJ Vol.7 No.4 pp.533-542, 2002

総説論文

エンタテインメントのための表情分析・合成技術

森島　繁生

Face Analysis and Synthesis for Entertainment

Shigeo MORISHIMA

Abstract – Copying human action accurately is main scheme in entertainment VR area to control and generate a virtual human or cartoon character in a screen. Motion capture system is generally used to generate a crone human in the scene of motion picture or interactive game, however, huge manual operation at post processing is inevitable to generate high quality image. In this paper, especially copying facial action is focused on and high quality and accurate method to generate and copy natural impression of face with semi-interactive process using face image analysis and face image synthesis scheme. Finally, some application systems in entertainment area are introduced.

Keywords : facial expression, lip synchronization, multimodal communication, face tracking, multimodal translation

1．まえがき

人物の動作を忠実に分析し、そのデータからスクリーンの中の仮想人物やキャラクターをコントロールして、コピーの動作を作り出す技術は、エンタテインメントVRに必要不可欠の技術となっている。映画やゲーム製作におけるモーションキャプチャ技術は、コピー人物作成のための実用化された技術であるが、現実には手作業による多大なポスト処理を必要とし、インタラクティブにかつ自然にコントロールする技術の確立が課題となっている。

一方、マルチモーダルを駆使して人間と機械がフェーストゥーフェースで対話する環境を実現しようとする試みは、ポストGUIに向けて特に注目を集める分野となっている。擬人化されたエージェントのリアリティに関する議論もしばしば行われるが、カートゥーンキャラクターのような本来の動きというものがもともと存在せず架空に仕立て上げられたものに多くの人々が愛着を覚えるのに比して、人間そのもののクローンを実現する技術は、実物が存在するがゆえに要求される条件も厳しくなり、僅かな粗も無視することはできない。したがって、現時点の技術では、人物描写において生身の人間と同等のリアリティを追及し、まさにビリーバブルな姿や動きを実現するには、まだまだ年月を要するであろう。

いずれにしても、高忠実な個人モデルの構築とその物理的な制約モデルの実現が、結果としてモデルベースビジョンの精度の向上に寄与し、Analysis by Synthesisのアプローチによって、より自然でインタラクティブな人物動作再現に寄与することは明らかである。

本稿では、エンタテインメントのための人物描写技術の中で特に表情の表現方法に的を絞り、オフラインで一般的に行われている映像特殊効果のような現実の映像処理を、よりインタラクティブな処理に近づけようとする試みについて述べる。また完全にオンラインで行うインタラクションの部分とオフライン処理のハイブリッドにより、高いクオリティ

成蹊大学工学部

Faculty of Engineering, Seikei University

の実現を目指す。したがって、コンピュータビジョン等の技術によって全自動のオンライン処理を前提にするのではなく、1フレームづつ行われていたマニュアル処理の負担を大幅に軽減して自動で行えるところはできる限り自動で行い、かつ最終的に実現されるクオリティは妥協しないという思想に基づく。

イメージベースレンダリングやビデオリライトは、イメージのみに拘ることで、クオリティの向上に寄与する技術であるが、本稿では完全にイメージのみに拘るのではなく、幾何構造も考慮したハイブリッド構造をとり、イメージベースのメリットと3次元処理のリアリティの双方を考慮して、よりアプリケーションに柔軟性を持たせている点が特徴である。

以下第2章では、表情モデリングの現状について述べる。また3章ではオリジナルの映像シーンから顔情報を抽出する方法について述べ、4章ではコミュニケーションギャップ克服の一例として言語障壁をとりあげ、翻訳音声とのリップシンクの実現について述べる。5章では別の応用としてのインタラクティブムービーシステムを紹介する。6章はリップシンクの評価方法を紹介し、7章では、より高いリアリティ実現に向けて、表情の動的なモデリング手法としての顔面筋肉モデルと、頭髪のダイナミクスモデルについて述べる。8章では現在進行中の音声対話擬人化エージェントプロジェクトについて紹介する。

2. 表情のモデリング

まずビデオシーンの顔部分を別人物と入れ替えるようなエンタテインメントアプリケーションを想定し、このための表情合成技術について述べる。シーン中の顔の方向が変化したり、表情が変化することへも対応するため、正確な3次元幾何形状とその変形が必要となる。

2.1 個人顔モデルの作成

人間の顔は、マクロに見れば同様の幾何構造と部品配置をしているが、ミクロに見れば凹凸の具合など個人差を有することは自明である。よってマクロな表現の顔モデルを用意して表情変形規則をあらかじめ幾何構造変形ルールとして定義しておく。このルール化については、例えばFACS[1]を定量化したもの、さらにその発展形としてMPEG4で定義されるFAPs (Facial Action Parameters) [2]に基づくもの、表情筋をバネモデルでシミュレーションしたもの[3]等がある。いずれもステレオタイプを定義したにすぎず、誰の顔も同じ印象の表情となる。

個人への適用化は、レンジファインダ等を使用したり、さらに精度を上げる場合にはデスマスクを作成した後に接触型の3次元デジタイザを使用して実測する方法もあるが、できるだけコストと手間を省くことを目指し、筆者らは正面顔画像もしくは複数方向から撮影された静止画像に対して標準モデル（図1）を整合する簡易な方法を提案した[4]。

また最近では安価で高精度のレンジファインダ（たとえばNEC Danae-R）も登場しているため、比較的容易に忠実な個人顔モデルが作成できる（図2）。表情変形ルールに関しても、このレンジファインダを利用して3次元の構造変形ルールについて実測が可能となった[5]。ただし、レンジデータは規則正しい点の集合に過ぎないので、標準顔モデルの各特徴点の影響力を考慮して整合することにより、格子点に意味づけを行うことができる（図3）。

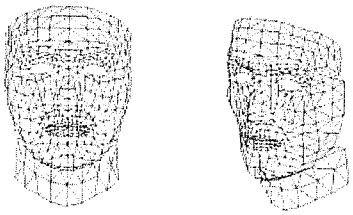


図1　標準ワイヤフレーム
Fig.1　Generic wireframe model

図2　個人モデル作成手順
Fig.2 Making of personal model

図３　表情変形ルールの定義

Fig.3 Control rule of facial expression

図４　標準口形の定義

Fig.4 User defined mouth shape

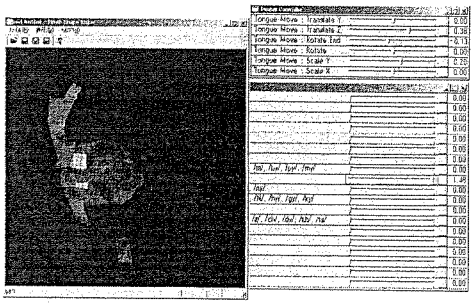


図５　舌および口内部エディター

Fig.5 Inside mouth editor

また表情変形のダイナミクスを定義するためには、表情筋モデルが有効であるが、実際に幾何変形の様子をリアルタイムで実測する手段が存在しないために評価が難しいこと、また表情筋の個人適応の方策がないことなどから、無表情と特定の表情までの変化の始点と終点の幾何形状もしくは表情筋パラメータを決定してそのパラメータ変化を線形的に補間して変形を実現しているに過ぎない現状にある。

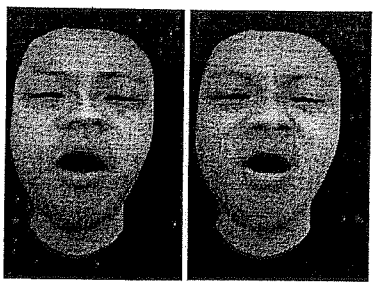
2.2 発話モデル

発話の際の口形状についても、やはり 3 次元での忠実な変化を表現しなくてはならない。筆者らは口形状を17個のパラメータで表現する方式を提案し、そのパラメータの制御で口形状を編集するツールを開発した[6]。さらに舌のモデルと口内部のモデル、歯モデルを定義して、任意の口内部の状況を表現できるツールを実現した（図５）[7]。

図４に口形状の一例を示す。すでにすべての英語および日本語の音韻に対して視覚素に相当する Viseme の定義を完了しており、発話の際の口形状アニメーションを作成可能である。時間方向の表現については、後に述べる主観評価実験から、標準口形を配置したキーフレーム間を正弦波でパラメータ補間する方法が現在もっとも良好としたが、さらに最適な補間法検討のための実測とコーパス作りを進めている。

2.3 テクスチャブレンディング

皺を忠実に表現するため、個々の基本表情単位毎にレンジデータの取得とともにテクスチャの取得も行い、特定の表情は基本表情単位の混合によって実現される。この際、その混合率によってテクスチャのブレンディングを行う。図６は、テクスチャブレンディングを行った場合と行わない場合の比較である。左は 1 枚の正面無表情テクスチャを、変形された幾何モデル上にマッピングしたもの、右は基本となる表情のテクスチャをブレンドして同様にマッピング合成したものである。同一の幾何形状に基づくにも拘らず印象が全く異なることが分かる。またこのブレンド率を入力データとして階層型ニューラルネットワークを恒等写像学習することにより中間層に感情空間を構築する手法を提案し、任意の表



a)ブレンドなし　b)ブレンドあり

図６　テクスチャブレンディング

a)Without blending　b)With blending

Fig.6　Result of texture blending

情変化がこの感情空間の軌跡として表現できることを示した[8]。

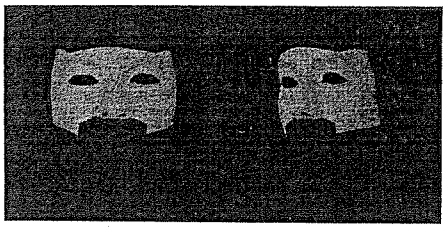


図7　3次元テンプレート
Fig.7　3D face template

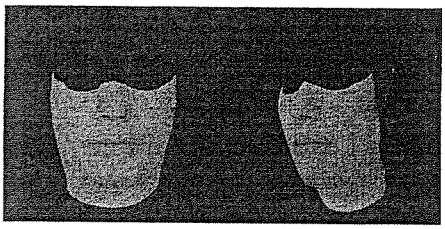


図8　合成用口形状モデル
Fig.8　Mouth model for synthesis

3．顔のトラッキング

ムービーシーンの人物画像の置換のため、オリジナルの映像中からターゲットとなる人物の顔の位置およびその向いている方向、大きさを推定することを目的とする。これは表情を強調したり、発話シーンを合成するために不可欠の前処理である。顔のトラッキング技術および特徴点抽出技術として、Eyematic社の製品が注目されているが[9]、特徴点ベースで顔の入れ替えを行った場合、特徴点点数は十分でなく僅かな推定誤差によってジッターを生じてしまうので、合成画像を原画像に重ねた場合、僅かなジッターにも人間の目は敏感に反応して、それに違和感を感じてしまう。そこで、シーンの照明環境は一連のシーケンスの中で大きく変化しないと仮定して、動画中の 1 フレームから取り出したテクスチャを個人の顔モデルにマッピングして 3 次元テンプレートを構成する。これを回転および移動と拡大縮小を行い、2 次元画像平面への投影像を作成し、ビデオフレーム中においてマッチング誤差最小となる点を繰り返し探索する。また 1 フレーム前の探索結果に基づいて探索範囲を限定することで、フレーム間の大きな変動が生じないように考慮され、動きがスムースになり、違和感が減少する。

現在は、無表情の発話シーンを想定してモデルは剛体として扱っているが、個人の表情変形ルールに基づいて表情変化をも考慮し、テンプレート自体の幾何形状も変形してトラッキングを行うモデルベース方式の検討を行っている。これが実現できれば、ターゲットがどのような表情を表出しているかも推定できることになる。いずれにしてもモデル合成の精度が認識の性能を左右することになる。

図7はトラッキングに用いた 3 次元テンプレートである。剛体として近似するために動きに敏感な部分は取り除いている。また図8は、唇部分を入れ替えるためのモデルである。図7のモデルで計算された回転角度と位置情報、大きさ情報に基づいて、図8のモデルの位置と角度と大きさを決定して発話合成を行い、オリジナル画像と置換することで、別の言葉を発声する唇部分を入れ替えた画像が合成できる。すなわち、言語に関わる唇の動き情報は置き換えられるが、目の動きや頭部の動きなどはオリジナルのままであり、ノンバーバル情報は保存される。

OPTOTRAK(LEDマーカによる高精度なモーションキャプチャシステム)を使った評価では、高い精度で 3 軸周りの回転角度と位置が再現されていることが分かった[10]。また撮影中のズームの変化にもある程度耐えることが確認できた。

もちろん、シーンチェンジや急激な照明環境変化には追従できないため、その都度モデルのテクスチャを手動で更新する必要が生じる点が現在の唯一の欠点である。

4．マルチモーダル翻訳システム

使用言語の異なる人間同士の対話を想定し、音声翻訳部分は既存の技術（たとえば ATR MATRIX システム[11]）を利用する。音声合成装置である CHATR[12]からは、各音素の継続長と音素表記が得られるので、この音素表記に相当する Viseme の標準口形状を選択し

てキーフレームとし、時間軸補間をすることで、滑らかな発話アニメーション（図9）が実現できる[13]。これによって、映画の吹き替えも自動的に行え、また演者の口の動きは翻訳された音声に同期するように合成して置換される。

このシステムの特徴は、音声翻訳に関係する唇周辺の情報のみが置換され、頭部の動きや眼球運動などオリジナルのままに保存される点である。したがって、言語障壁というコミュニケーションギャップに関連する部分のみが置換されて、他のノンバーバルに関わる部分については変換が加えられない。さらにATR の音声翻訳システムは、本人の声の特徴が保存されたまま多言語間で音声翻訳されるので、言語情報のみが同様に置換された形となり、個人情報は保存される。

同様の研究として、ビデオ素片を結合することによるもの[14]、コーパスに基づくもの[15]が存在するが、表情の変化や顔の向きに自由度がない点で性能的に問題があり、本システムは優位な位置づけである。

このシステムの応用として映画の自動吹き替えシステムが考えられる。出演者自体の他言語データベースを完備することにより、本人の声で別の言語の音声に吹き替えが可能であり、さらにリップシンクも自動的に実現できる。しかし実際には、音声認識や機械翻訳はタスクを狭い範囲に限定する必要があるため、任意の台詞をフルオートで音声翻訳することは現実的ではない。また音声に含まれるのは言語情報だけではないので、ノンバーバルな情報も含めて音声合成するのは現在の技術では困難な課題である。この解決策として、実際には声優に台詞を発声してもらい、この声優の声を自動的に分析して、台詞のテキストも併用することによって、音素継続長のみを声優の発声した音声から獲得し、リップシンクを自動的に実現する方法がもっとも身近な実現方法と考えられる。このため、テキストを併用したセグメンテーション手法[16]、あるいは HMM に基づく自然音声からの発話口形の自動再現の研究が行われている[17]。

5．インタラクティブムービーシステム

インタラクティブシステムの実現例として、翻訳システムで行われていた口部分の入れ替えだけにとどまらず、顔全体を別の人物と入れ替えることによって、あたかも映画の主人公に自分がなったような錯覚にとらわれ、ストーリーの中に埋没していくことができるシステムを実現する[18]。フィルム全体に対するオフラインの顔のトラッキングおよびモデル整合の自動化（図 10）が課題であるが、このプロセスが完了すれば、オンラインで違和感のない顔全体の置換が実現できる（図 11）。インタラクティブモードでは、声優がマイクに向かって喋った声にシンクロしてリアルタイムで口形状を合成できるアルゴリズム[19]を採用し、映像中の顔画像がインタラクティブに反応するシステムを構築できる[20]。

図 9　マルチモーダル翻訳
Fig.9 Multimodal translation image

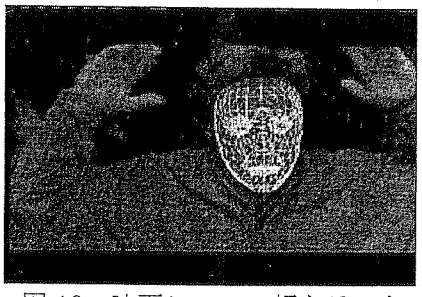


図 10　映画シーンの顔トラッキング
Fig.10 Face tracing in movie scene

図 11　置換された主人公の顔
Fig.11　User's face replaced with hero's

したがって、実際に映画を見ながらマイクに向かって語りかけると、その台詞に同期した人物映像がリアルタイムで再現される。オリジナルのストーリーにオーバーラップされて、あたかも自分がストーリーの中で役を演じているような気分になる。顔以外は主人公のアクションそのものであり雰囲気はオリジナルのままに保存されるため、本来のエンタテインメント応用のみならず、実際に表現の乏しい人物にとって、相手に豊かに自分の感情を伝えることができるシステムとして利用可能である。

6．リップシンク評価方法

どれだけ合成された口形状が音声信号とシンクロしているかを評価する方法として、白色雑音を音声に付加して明瞭性を劣化させ、それに合成された顔を併せて提示することによって被験者の聞き取りの精度がどれだけ向上するかで判断する方法を提案する[21]。

図 12 はこの評価結果の一例を示す。被験者には 6 つの数字を連続的にランダムに発声した音声が提示され、SNR を-6,-12,-18dB と変化させたものをランダムに聞かされる。同時に顔画像が提示され、聞き取った数字を書きとめる。被験者は 12 名である。

一番左のグラフから、音声だけ提示した場合には、聞き取り性能が SNR の劣化につれて大胆に低下していくことが分かる。この場合は、自然音声をサンプルとして使用しているため、その言葉を喋っているオリジナルの顔動画像が存在するが、左から 2 番目のグラフでは、これを同時に被験者に提示すると、認識率が劇的に向上することが分かる。ノンバーバル情報の重要性が裏づけされた形となる。

ちょうど中央のグラフは、合成顔画像で置き換えた場合であるが、音韻に対応するキーフレーム間を線形補間した場合である。右から 2 番目は正弦波補間した場合であり、滑らかに変化する正弦波補間の方が線形補間よりも優れていることが分かる。よって、いかにオリジナルの顔画像の場合の認識率に近づけられるが、顔画像合成手法の性能を反映していることになる。ちなみに音声と画像がミスマッチした場合には、右端のグラフのように認識率は著しく低下する。よって、中途半端な顔の表示はむしろ逆効果ということもありうる。この評価方法はリップシンク性能を判断する上で有効と考えられる。

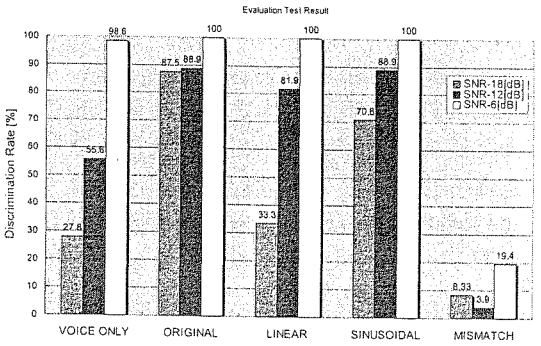


図 12　リップシンクの主観評価
Fig.12　Evaluation of lip-sync

7．より高いリアリティを求めて

人間同士がコミュニケーションを行っている場合、微妙な表情の変化によって受ける印象が大きく異なる場合がある。また表情とは関係がなく、普段はあまり意識することがない頭髪の表現も、風に靡くことがなければ不自然な印象が焼き付いてしまう。

7.1　表情の動的なモデリング

ハイスピードビデオカメラ(250frames/sec)を用いて、作為的に表出した表情と、映像を見て自然に生じた感情に基づいて表出された

表情との違いを観察した。その結果、顔のパーツごとに時間的な変位に差が生ずることがわかった[22]。また顔面筋の実際の解剖に立ち会ってモデル作成を行ったところ、顔面筋の発達の度合いや配置等において個人差が大きいことが分かった。したがって、喚起された感情と直接的に関係づけられた顔面筋肉モデルの構築がよりリアルな顔アニメーションにとって必要となろう。

すでに提案されている顔面筋モデル[3]は、筋肉繊維を線形のバネで近似しているに過ぎない。著者のグループでは顔面筋に実際にボリュームを持たせて、筋肉の収縮を忠実に再現できる新しい顔面筋モデルの構築を行っている（図 13)。また、その物理特性を筋繊維毎に変化させたり、表情筋の配置を個人の顔形状に応じてカスタマイズできる機能を市販の Alias/Wavefront 社のモデリングツールである Maya のプラグインとして実現した[23]。これによって、キャラクターアニメーションのデザイナーのこれまでの労力が大幅に軽減されるばかりでなく、よりリアルな表情の動的なアニメーションが実現できると思われる。

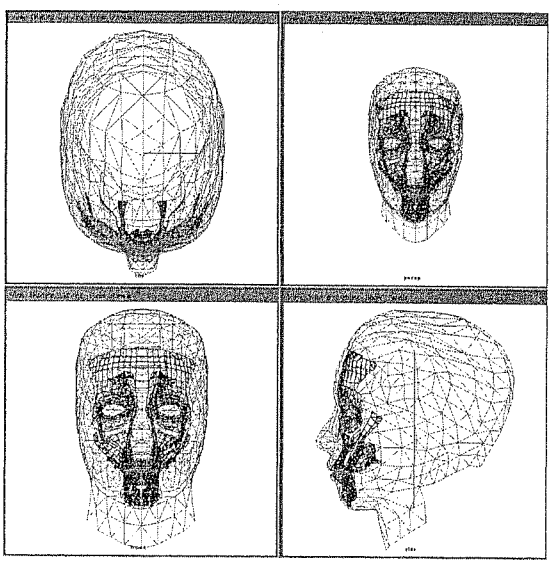


図 13　ボリュームをもった顔面筋モデル
Fig.13　Volumetric facial muscle model

7.2　頭髪の運動モデリング

頭髪のモデル化の研究は 80 年代後半に一つのピークを迎えたが、最近、再び活発化している[24][25]。ハードウェアの高性能化に伴う研究環境の劇的な進歩によって 10 年前には実現し得なかった頭髪運動のデモ映像が、安価なＰＣでも簡単に作れるようになったからである。筆者らも、剛体セグメントモデルにより運動方程式を解いて制御点の運動を求め、それらを自由曲線で結ぶという手法を 10 年来進めてきた[26]。問題はいかにしてヘアデザインを簡単に実現するか。またそのスタイルを維持したまま運動させるかに重点が置かれる[27]。最近ではカートゥーンの髪型制御にも着手している。

図 14 は風に靡くパーマヘアの一例である。

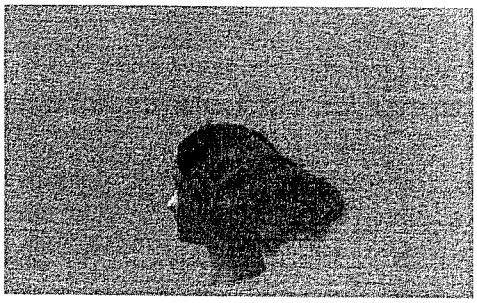


図 14　風に靡くウェーブ
Fig.14　Wavy hair in the wind

8．音声対話擬人化エージェントツール

音声認識・合成・対話の研究者と顔画像合成の研究者が一丸となり、擬人化エージェントによるマルチモーダル対話を実現するフリーソフトウェアの開発を進めている[28]。
このシステムの構成を図 15 に示す。

対話環境を構築する VoiceXML ベースの Rapid Prototyping Tools、音声認識、音声合成、顔画像合成の各モジュールを独立したコンポーネントとして統合的に制御する Agent Manager、対話中の突然の割り込みを考慮した対話タスクをも制御する Task manager を含み、誰でも容易に対話システムの構成が可能であり、合成音声とのリップシンクも自動

的に実現される。さらにすべてのソースリストは公開される予定である[29]。よって関連の研究者に及ぼす影響は大きいものと推測される。このツールによってさらなるマルチモーダル研究が発展することが期待される。

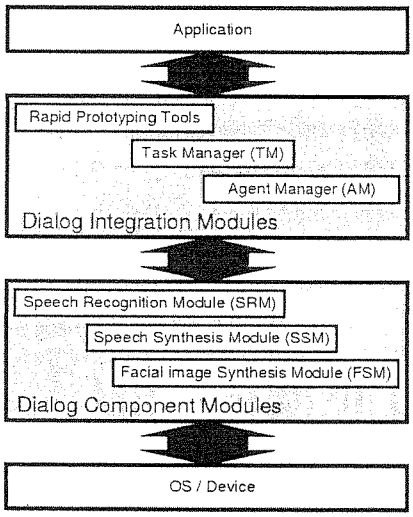


図 15　音声対話擬人化エージェントの構成
Fig.15　Configuration of
Anthropomorphic Dialog Agent

9．むすび

本稿では、エンタテインメントVRへの応用を考慮した表情の分析・合成手法と、関連技術について紹介した。スタティックな表情のリアリティのみならず、ダイナミックな特性が必要とされている。またリアリティが向上すればするほど、欠点が浮き彫りにされ、さらなる問題点が浮上してくる。現在は、感情の表現のみならず、個人情報をいかに表現すべきかという観点からも研究を進めている。これは、現在までステレオタイプを追及してきたアプローチとは異なり、より特定の個人の印象に近づける試みである。

最後に紹介した音声擬人化対話エージェントツールは、オープンソースということで、新規に研究を開始する研究者にとって極めて有効なツールとなり得る。したがってマルチモーダル研究がエンタテインメント分野をはじめ、ヒューマンコミュニケーション分野や様々な領域において活発化することを願ってやまない。

ともあれ、映像のクオリティは今後ますますリアリティを増し、実在する人物とも区別が難しいビリーバブルな擬人化エージェントが実現されるのも時間の問題であろう。

参考文献

[1] P. Ekman and W.V. Friesen, "Facial Action Coding System", Palo Alto, CA: Consulting Psychologists, 1978.

[2] Information technology-generic coding of audio-visual objects part 2:Visual, ISO/IEC 14996-2, Final draft of international standard, ISO/IEC, JTCI/SC29/WG11 N2501.

[3] Victor Lee, Demeteri Terzopoulos and Keith Waters, "Realistic Modeling for Facial Animation", Proc. SIGGAPH95, pp.55-62, 1995.

[4] 伊藤圭、三澤貴文、武藤淳一、森島繁生, "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.HIP99-65, No.582, pp.7-12, 2000.

[5] 大橋俊介、杉崎英嗣、伊藤圭、森島繁生,"レンジファインダを用いた表情変形ルールと表情編集ツールの構築", 信学技報, Vol.101, No.610, pp.1-7, 2002.

[6] 伊藤圭、三澤貴文、武藤淳一、森島繁生, "仮想空間上におけるリアルな三次元口形状の作成", 信学全大, A-16-24, p.328, 2000.

[7] 鳥飼友美、伊藤　圭、緒方　信、森島繁生, "仮想人物の舌モデル構成と発話アニメーション作成", 信学総合全大, A-16-34, 2002.

[8] 柳澤尋輝、高橋光紀、森島繁生、"テクスチャブレンディングによる皺の表現と基本顔モデルによる感情空間の構築", 信学技報, Vol.101, No.693, HCS2001-49, pp.17-24, 2002.

[9] http://www.eyematic.co.jp/

[10] 長田誉弘、大室学、緒方信、森島繁生, "ズーム変化を含む動画中の顔自動トラッキングとマッチムーブによる表情合成", 信学技報、Vol.101, No.693, HCS2001-50, pp.25-32, 2002.

[11] 菅谷、竹澤、横尾、山本、"日英双方向音声翻訳システム(ATR-MATRIX)の対話実験", 音響春季大会, pp.107-108, 1999.

[12] Nick Campbell, Alan W. Black, "Chatr: a multi-lingual speech re-sequencing synthesis system", 信学技報, Vol.SP96-7, pp.45-51, 1995.
[13] 緒方信、中村哲、森島繁生, "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-", 情報処理学会インタラクション論文集, Vol.2001, No.5, pp.203-210, 2001.
[14] Fu Jie Huang, Eric Cosatto, Hans Peter Graf, "Triphone Based Unit Selection for Concatenative Visual Speech Synthesis", Proc. of ICASSP2002, Vol.2, pp.2037-2040, 2002.
[15] Tony Ezzat, Gadi Geiger, Tomaso Poggio, "Trainable Videorealistic Speech Animation", ACM Transaction on Graphics, pp.388-398, Vol.21, No.3, 2002.
[16] 吉住悟、緒方信、森島繁生, "テキスト情報を利用した発話音声の自動セグメンテーションと感情音声分析への応用",信学会総合全大, A-14-1,2001.
[17] 垣原清次, 中村哲, 鹿野清宏, "HMM を用いた自然な発話動画像合成", 電子情報通信学会論文誌, Vol. J83-D-II, No.11, pp.2498-2506, 2000.
[18] Shoichiro Iwasawa, Tatsuo Yotsukura, Shigeo Morishima, "Face Analysis and Synthesis for Interactive Entertainment", International Workshop on Entertainment Computing (IWEC2002), pp. 143-150, 2002.
[19] Shigeo Morishima and Hiroshi Harashima, "A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface", IEEE JSAC, Vol. 9, No. 4, pp. 594-600, 1991.
[20] Shigeo Morishima, Shoichiro Iwasawa, Tatsumi Sakaguchi, Fumio Kawakami, Makoto Ando, "Better Face Communication", Visual Proceedings of ACM SIGGRAPH'95, Interactive Communities, p.117, 1995.
[21] Shigeo Morishima and Satoshi Nakamura, "Multi-modal Translation and Evaluation of Lip-synchronization using Noise Added Voice", The First International Joint Conference on Autonomous Agents & Multi-Agent Systems, Proc. of Workshop 14: Embodied conversational agents – let's specify and evaluate them!, 2002.
[22] 四倉達夫、内田英子、山田寛、赤松茂、鉄谷信二、森島繁生, "高速度カメラによる動的な顔面表情の分析および合成", 信学技報, HCS2002-33, Vol.101, No.610, pp7-12, 2002.
[23] Tatuo Yotsukura, Mitsunori Takahashi, Shigeo Morishima, Kazunori Nakamura, Hirokazu kudoh, "Magical face: Integrated Tool for Muscle Based Facial Animation", Conference Abstracts and Applications, ACM SIGGRAPH, p.212, 2002.
[24] Johnny T.Chang, Jingyi Jin, Yizhou Yu, "A Practical Model for Hair Mutual Interactions", 2002 ACM SIGGRAPH Symposium on Computer Animation, pp.73-80, 2002.
[25] Tae-Yong Kim, Ulrich Neumann, "Interactive Multiresolution Hair Modeling and Editing", ACM Transaction on Graphics, pp.620-629, Vol.21, No.3, 2002.
[26] 岸 啓補、森島繁生、"頭髪のスタイリングとアニメーション表現", 電子情報通信学会論文誌, Vol.J83-D-II, No.12, pp.2716-2724, 2000.
[27] 杉森大輔、杉崎英嗣、森島繁生,"コンピュータグラフィックスによる髪型を保存する復元力を用いた頭髪の自然な運動表現", グラフィックスとCAD シンポジウム 2002, pp.83-86, 2002.
[28] 川本真一、下平博、新田恒雄、西本卓也、中村哲、伊藤、克亘、森島繁生、四倉達夫、甲斐充彦、李晃伸、山下洋一、小林隆夫、徳田恵一、広瀬啓吉、峯末信明、山田篤、伝康晴、宇津呂武仁、嵯峨山茂樹, "カスタマイズ性を考慮した擬人化音声対話ソフトウェアツールキットの設計", 情報処理学会論文誌, vol.43, no.7, pp.2249-2263, Jul 2002.
[29] http://iipl.jaist.ac.jp/IPA/

(2002 年 9 月 10 日受付)

[著者紹介]

森島 繁生

1987 年東京大学大学院工学系研究科修了. 同年成蹊大学工学部専任講師. 88 年同助教授. 01 年同教授, 現在に至る. 人物像の合成・認識の研究に従事 (工学博士).

HyperMask – projecting a talking head onto a real object

T. Yotsukura[1], S. Morishima[1], F. Nielsen[2], K. Binsted[3], C. Pinhanez[4]

[1] Faculty of Engineering, Seikei University, 3-3-1 Kichijoji-Kitamachi, Musashino-shi, Toyko 180-8633, Japan
E-mail: {yotsu,shigeo}@ee.seikei.ac.jp
[2] Sony Computer Science Laboratories Inc., 3-14-13 Higashi-Gotanda, Shinagawa-ku, Tokyo 141-0022, Japan
E-mail: nielsen@csl.sony.co.jp
[3] I-chara Inc., 2-34-1 Uehara, Shibuya-ku, Tokyo 151-0064, Japan
E-mail: kimb@i-chara.com
[4] IBM Research, Watson, Route 134, P.O. Box 218, Yorktown Heights, N.Y. 10598, USA
E-mail: pinhanez@us.ibm.com

Published online: 15 March 2002

HyperMask is a system which projects an animated face onto a physical mask worn by an actor. As the mask moves within a prescribed area, its position and orientation are detected by a camera and the projected image changes with respect to the viewpoint of the audience. The lips of the projected face are automatically synthesized in real time with the voice of the actor, who also controls the facial expressions. As a theatrical tool, HyperMask enables a new style of storytelling. As a prototype system, we put a self-contained HyperMask system in a trolley (disguised as a linen cart), so that it projects onto the mask worn by the actor pushing the trolley.

Key words: Talking heads – Homography – Neural networks – Computerized theatrical performances – Lip-synch

Correspondence to: T. Yotsukura

1 Introduction

HyperMask is a demonstration technology for a theatrical tool. It enables a new style of storytelling, in which a human actor's performance is enhanced by the system in an entertaining manner. However, the same technology could also be useful for other applications in which active projection is necessary. For example, in the so-called "Office of the Future" (Raskar et al. 1998) or an interactive playground, we would like to be able to project dynamically images and information onto moving, irregularly shaped objects.

Also, HyperMask is an interesting demonstration system for its integrated component technologies. Basically, HyperMask consists of camera that observes the stage, and a retro-projector that projects image information (e.g. onto the masks of the actors). Note that the retro-projector can be considered as a camera whose direction of propagation of light is inverted. Our first technical step was to implicitly calibrate the geometry implied by the camera and projector without explicitly calculating all intrinsic and extrinsic parameters, which is time-consuming and error-prone.

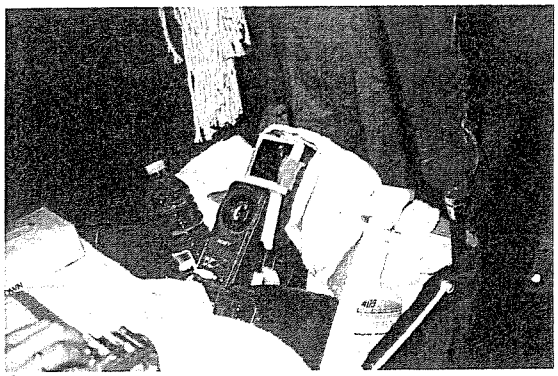
Another technology is real-time lip synchronization using user's own texture mapping. This system allows the user to quickly fit a face texture to a 3D polygonal model. Then, a neural network is trained for predicting lip movements based on vowels. The system can then synchronize the lip movements of the face model with the voice of the user in real time. The expression of the projected face can also be altered by the user.

The HyperMask project improves technologically some seminal work explored in different ways by contemporary artists. Projecting human faces and bodies onto mostly static objects and puppets has appeared in many art pieces, notably in the works of Tony Oursler and Laurie Anderson in the 1990s. Before that, some artists have also experimented with video-based heads for performers. In particular, Otavio Donasci has explored "video-creatures" since the beginning of the 1980s, in performances where actors used video monitors as their heads. The actors' faces were hidden from the public and substituted for the faces of off-stage actors captured live by a video camera.

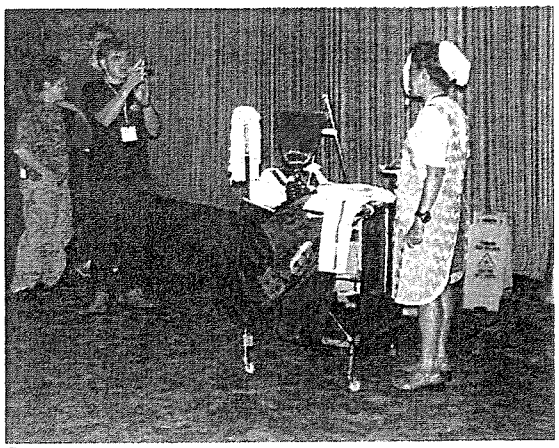
2 HyperMask prototype

The first prototype of the HyperMask was used in a performance at the Sony Computer Science Labo-

The Visual Computer (2002) 18:111–120
Digital Object Identifier (DOI) 10.1007/s003710100140



1



2

Fig. 1. Installed camera and projector
Fig. 2. HyperMask prototype

ratory in the summer of 1998. After that, we incorporated the lip-synching and facial expression control software described in this paper. Based on these early experiments, we created a performance piece for the SIGGRAPH'99 Emerging Technologies exhibition that used a portable version of the HyperMask system. The equipment (camera, projector and computer) is loaded into a trolley, and the actor wheels the trolley around the performance area and chats with the audience. The faces projected onto the mask reflect the tone and content of the various stories and interactions.

Figures 1 and 2 show this HyperMask prototype. The camera on the trolley is always tracking the actor's mask, and the LCD projector is always projecting a synthesized facial expression onto the mask. The actor's speech, picked up through a microphone in the mask, is converted into a lip shape in real-time, and the lip shape image is generated. Then a face image, with a facial expression chosen by the user via a small keypad, is synthesized using a 3D face model and texture mapping. The actor can also change the face model and texture using the keypad.

3 Camera and projector calibration

A major goal of the HyperMask project was to allow the actor using the mask to move her face, so she can use facial gestures, look to the audience, nod, etc. To accomplish this we made the projected face considerably smaller than the projectable area, so if the actor moves around the projector's cone of light, it is possible to keep the computer-graphics (CG) face projected on her face by simply moving the CG face around the projectable area.

In the HyperMask system a camera is employed to track the position of the mask by finding on the camera imagery the position of 4 infrared markers on the mask. To project the CG face exactly on the mask worn by the actor, it is necessary to calibrate the camera to the projector, i.e., to determine for any point in the camera image its corresponding point in the projector's image.

The relationship between points observed on a planar surface from two different cameras is known to be a homography (Faugeras 1993). A homography (also called collineation, since it preserves lines) is a 3×3 matrix defining a linear application in the projective space that, for a given planar surface of the real world, maps all projected points in one camera's image into the other camera's image.

The fundamental observation is that from a geometrical point of view, "ideal" pinhole projectors and cameras are identical (see Fig. 3). Let H denote the homography that relates the image of the projector image frame to the camera image frame. This means that a 2D point homogeneous coordinates on the camera image,

$$c = (x_c/z_c, y_c/z_c) ,$$

matches a 2D point,

$$p = (x_p/z_p, y_p/z_p) ,$$

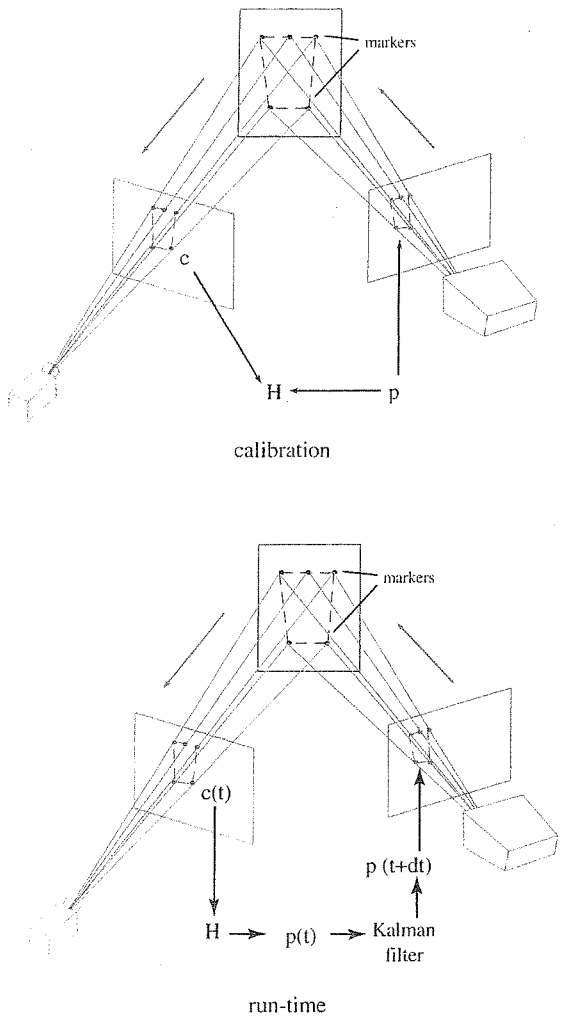


Fig. 3. Calibration process and run-time system

on the projector image as follows:

$$\boldsymbol{p} = \begin{pmatrix} x_p \\ y_p \\ z_p \end{pmatrix} = \boldsymbol{H}\boldsymbol{c} = \boldsymbol{H} \begin{pmatrix} x_c \\ y_c \\ z_c \end{pmatrix} .$$

A homography is completely defined if the projection of four 3D points of the world on both image planes is known. To determine the homography between a camera and a projector, we need simply to obtain the four needed points while manually aligning a projection of the surface with the real surface (see Fig. 3).

The homogeneous coordinates of four points to be projected,

$$p_i = (x_p^i, y_p^i, 1) \qquad i = 1, 2, 3, 4 ,$$

are determined arbitrarily, making sure that the points are visible and there is a way to move the real surface so it aligns with the projection. Then, we consider the homogeneous coordinates of the four points on the camera image as sensed by the tracking system,

$$c_i = (x_c^i, y_c^i, 1) \qquad i = 1, 2, 3, 4 .$$

Let the homography matrix, $\boldsymbol{H}$, be defined as follows:

$$\boldsymbol{H} = \begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{pmatrix} .$$

Matrix $\boldsymbol{H}$ is defined up to a scalar coefficient. Assuming $h_9 \neq 0$, we set $h_9 = 1$.
A point with homogeneous coordinates $(x\ y\ w)$ is transformed to $(x''y''w'')$ as below:

$$x'' = h_1 x + h_2 y + h_3 w ,$$
$$y'' = h_4 x + h_5 y + h_6 w ,$$
$$w'' = h_7 x + h_8 y + h_9 w .$$

Setting $w = 1$ yields

$$x' = \frac{x''}{w''} = \frac{h_1 x + h_2 y + h_3}{h_7 x + h_8 y + 1} ,$$
$$y' = \frac{y''}{w''} = \frac{h_4 x + h_5 y + h_6}{h_7 x + h_8 y + 1} .$$

This can be written as

$$x' = h_1 x + h_2 y + h_3 - h_7 x x' - h_8 y x' ,$$
$$y' = h_4 x + h_5 y + h_6 - h_7 x y' - h_8 y y' .$$

We obtain the following linear system to solve (we need to invert the 8×8 matrix):

$$\begin{pmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1'x_1 & -x_1'y_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -y_1'x_1 & -y_1'y_1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -x_2'x_2 & -x_2'y_2 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -y_2'x_2 & -y_2'y_2 \\ x_3 & y_3 & 1 & 0 & 0 & 0 & -x_3'x_3 & -x_3'y_3 \\ 0 & 0 & 0 & x_3 & y_3 & 1 & -y_3'x_3 & -y_3'y_3 \\ x_4 & y_4 & 1 & 0 & 0 & 0 & -x_4'x_4 & -x_4'y_4 \\ 0 & 0 & 0 & x_4 & y_4 & 1 & -y_4'x_4 & -y_4'y_4 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \end{pmatrix} = \begin{pmatrix} x_1' \\ y_1' \\ x_2' \\ y_2' \\ x_3' \\ y_3' \\ x_4' \\ y_4' \end{pmatrix} ,$$

where $x'_i = x''_p$, $y'_i = y''_p$ and $x'_i = x''_c$, $y'_i = y''_c$ for $i \in \{1, 2, 3, 4\}$.

Computing homographies is quite an unstable numerical process. Indeed we need to invert a 8×8 matrix. Therefore, we may use singular value decompositions or pseudo-inverse if $\boldsymbol{H}$ is ill-conditioned. Another alternative is not to use points but corners, as suggested in Zoghlami et al. (1997). A corner is defined by two half-lines joining in an intersection junction. If more fiducials are available, we can compute more reliably the homography by using least median square methods or even better the statistical approach of Kanatani (1998).

However, if the plane is far enough from both the projector and camera (in relation to their baseline distance), we can relax the homography by an affine transform (or even similitude) as described below. In fact, our final system used this simplified approach: Taking the matrices corresponding to these two sets of four points,

$$\boldsymbol{P} = \left(p_1^{\mathrm{T}}, p_2^{\mathrm{T}}, p_3^{\mathrm{T}}, p_4^{\mathrm{T}}\right)$$

and

$$\boldsymbol{C} = \left(c_1^{\mathrm{T}}, c_2^{\mathrm{T}}, c_3^{\mathrm{T}}, c_4^{\mathrm{T}}\right) ,$$

we want $\boldsymbol{P} = \boldsymbol{HC}$, whose solution is

$$\boldsymbol{H} = \boldsymbol{P}\boldsymbol{C}^{\mathrm{T}} \left(\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}\right)^{-1} .$$

During run-time, we simply take a point in camera image $c = (x_c, y_c, 1)$, project it through the homography $\boldsymbol{H}$, obtain $p = \boldsymbol{H}c$ and compute the position on the projector's image plane,

$$p = (x_p/z_p, y_p/z_p) .$$

Surprisingly, this calibration step is numerically stable even with only four points, and can be done, in practice, in a few seconds. We believe that the stability is also related to the fact that in our experiments the projection centers of the camera and the projector are close to being aligned. Note that there is no need to determine the camera's intrinsic parameters or those of the projector.

4 Tracking the projection surface

In our experiment, we used plain markers on the projection surface. In particular, we employed infrared LEDs that can be easily tracked by a camera with an infrared filter. However, if we move the mask too quickly, we observe that the projected image "falls behind" the moving surface. That is, there is a "shifting" effect, where the observations at discrete time t on the camera image, $c(t)$, are displayed by the projector at time $t + \mathrm{d}t$ using the estimate at time t, $p(t) = \boldsymbol{HC}(t)$. To reduce the "shifting" problem we employ a predictive Kalman filter (Gelb 1974) that estimates the most likely position of every point at time $t + \mathrm{d}t$, using equations of dynamics as the underlying model of the Kalman filter, as shown in Fig. 3. The parameter $\mathrm{d}t$, corresponding to the average delay between sensing and displaying, is determined experimentally. The Kalman filtering approach proved to be very effective in our experiments.

5 Handling a 3D mask

The method described above works quite well for a planar surface. However, when transferring from a 2D mask to a 3D mask, we have to handle the projected pattern more carefully. Given our projection setup, two kinds of problems occur. First, when the mask is panned to the left or to the right, hidden (or occluded) parts of the virtual projected mask do not appear on the physical mask. One ideal solution is to have a set of cameras and projectors covering the whole stage. Each projector would have to project an image on the parts of the mask it can effectively hit through a ray emanating from its optical center. However, in a performance situation it is possible to constrain the interaction with the audience so almost always the actor is looking forward, so projection occlusion is minimized.

The second problem arising from projection onto a 3D mask is related to the fact that the projection has to be corrected for the differences in depth in the projected surface. For example, suppose we have a mask in the shape of a human face, onto which a "clown" mask with a red nose is projected. Now suppose we are projecting a 2D CG rendition of a face. Also assume that the tracking system is able to correctly detect the borders of the mask and to deform the CG face to match the borders of the mask. As we rotate the mask from center to left, the projection of the corrected 2D face will, in general, put the clown's red nose in the incorrect place. This is because the nose, when viewed in a profile, moves more to the right

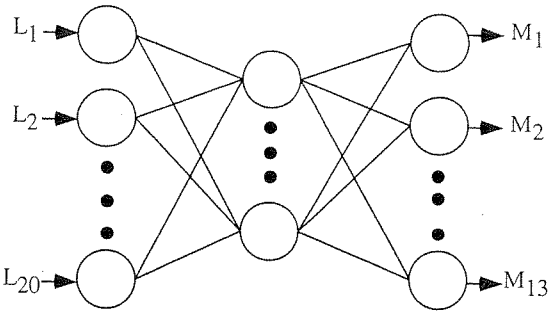
than the rest of the face, simply because its 3D position is in front of the other parts of the face.
The simple way to correct for this is to project a 3D CG model of a face so it replicates in the virtual world the movements of the 3D mask in the real world. Therefore, we have to recover the attitude (Lowe 1991), i.e. the 3D coordinates and 3D orientations in the frame world, of our 3D mask, in order to put its model in a virtual 3D scene so that we can perform corresponding CG occlusion and project the observed 3D scene (a 2D image) onto the mask in the 3D world. Our system uses simple mechanisms based on the relative lengths of the observed position of the LED to estimate the 3D attitude of the mask. Based on those measurements, the system is able to project a 3D face onto the mask that correctly occludes areas of the face as the user rotates her head.

6 Real-time talking head

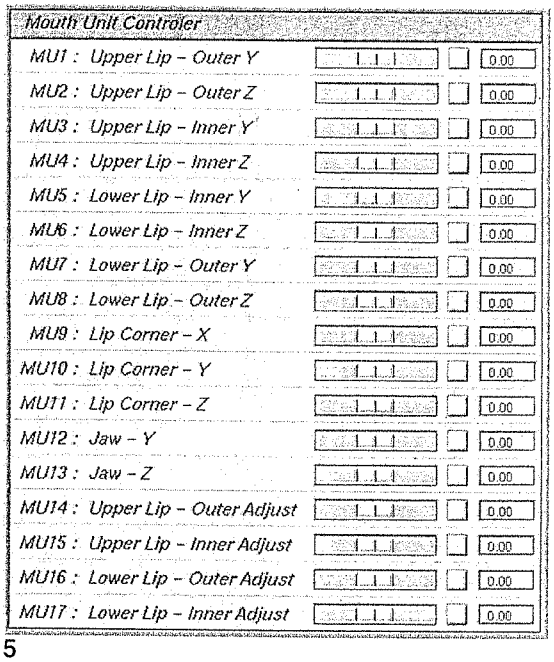
To realize real-time lip synchronization, the user's voice (captured by a microphone) is phonetically analyzed and converted to a mouth shape and expression parameters on a frame-by-frame basis. LPC Cepstrum parameters are converted into mouth shape parameters by a neural network trained on vowel features. Figure 4 shows the neural network structure for parameter conversion. The 20-dimensional Cepstrum parameters are calculated every 32 ms with a 32 ms frame length. The mouth shape is then synthesized according to these mouth-shape parameters. The facial expression is chosen by the user, from Anger, Happiness, Disgust, Surprise, Fear and Sadness. Each basic emotion is associated with specific facial expression parameters described by FACS (Ekman and Friesen 1978).

7 Designing the mouth shape

The set of mouth shapes can be easily edited by our mouth-shape editor (see Fig. 5). We can change each mouth parameter to determine a specific mouth shape, which can be seen in the preview window. Typical vowel mouth shapes are shown in Fig. 6. Our special mouth model has polygons for the teeth and the inside of the mouth. A tongue model is now under construction. When converting from the LPC Cepstrum parameters to the mouth shape, only the mouth shapes for 5 vowels and nasals are defined in the



4



5

Fig. 4. Network for parameter conversion from voice to mouth shape
Fig. 5. Mouth shape editor

training set. We have defined all of the mouth shapes for Japanese phonemes and English phonemes using this mouth-shape editor.

8 Customizing the face model

To generate a realistic avatar's face, a generic face model is manually adjusted to the user's face im-

age. To produce a personal 3D face model, both the user's frontal face image and profile image are necessary. The generic face model represents all of the control rules for facial expressions (defined by FACS parameters) as a 3D movement of grid points, which modify the geometry of the model.

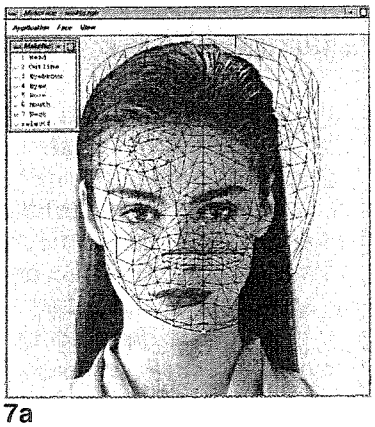
Figure 7 shows a personal model both before and after the fitting process for a front-view image, using our original GUI-based face-fitting tool. The front-view image and the profile image are loaded into the system, and then the corresponding control points are manually moved to an approximately correct position, using the mouse. The synthesized face results from mapping a blended texture (generated from the user's frontal image and profile image) onto the modified personal face model.

However, sometimes self-occlusion happens, and we cannot capture the whole texture using only the front and profile face images. To construct the 3D model more accurately, we introduce a multi-view, face-image-fitting tool. Figure 8 shows the fitted result with face images from any oblique angle. The rotation angle of the face model can be controlled in the GUI preview window to achieve the best fit for face images captured from any arbitrary angle. Figure 9 shows examples of reconstructed faces. Figure 9a uses 9 view images, and Fig. 9b uses only frontal and profile views. As you can see, much better image quality is achieved by the multi-view fitting process.

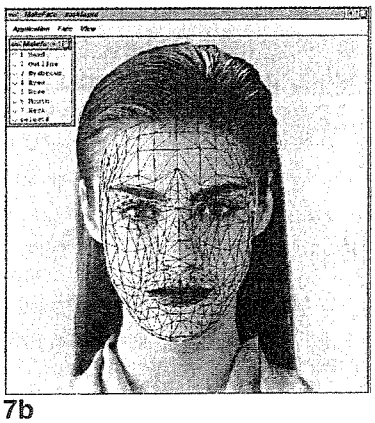
9 User adaptation of voice

When a new user comes in, the voice model, as well as the face model, has to be registered before operation. Ideally, the neural network has to be retrained in each case. However, it takes a very long time to get convergence using back-propagation. So, 75 subjects' voice data, including 5 vowels, were pre-captured, and a database of weights of the neural network and the voice parameters were constructed. So, speaker adaptation is performed by choosing the optimum weights from the database. When a new non-registered speaker comes in, s/he has to speak 5 vowels into a microphone. The LPC Cepstrum is calculated for the 5 vowels, and this is fed into the neural network. The mouth shape is then calculated by selected the weight, and the error be-

6



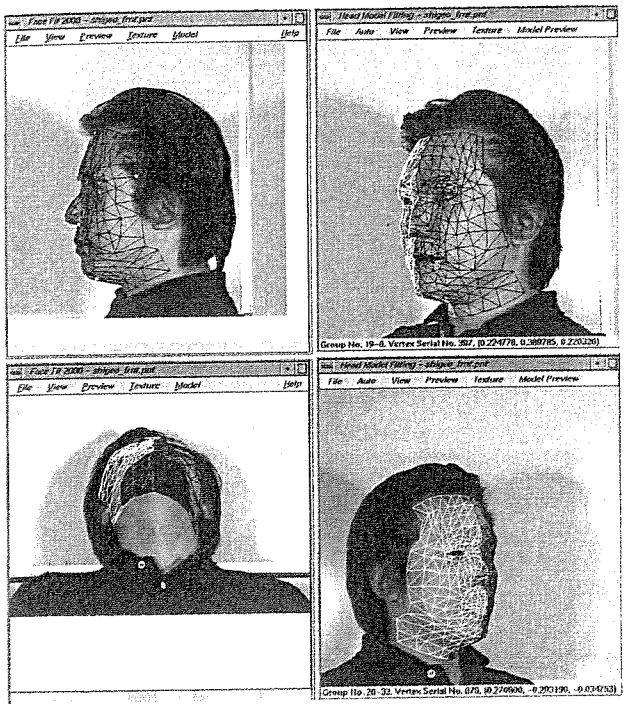
7a



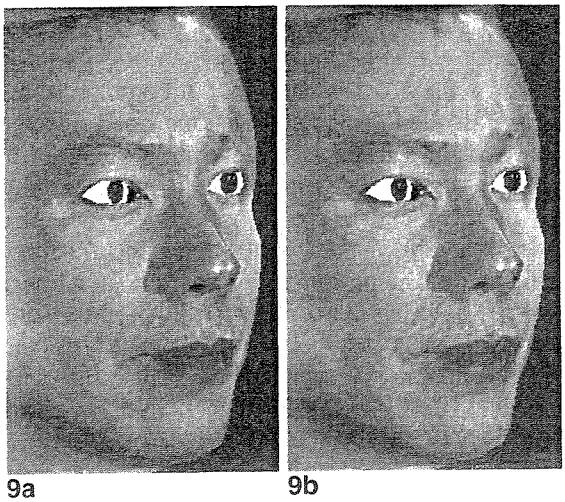
7b

Fig. 6. Typical mouth shapes

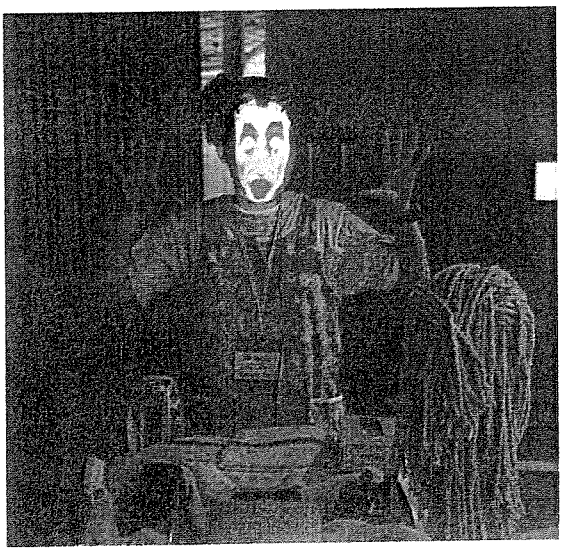
Fig. 7a,b. Frontal model fitting using the GUI tool. **a** Initial model; **b** fitted model



8



9a 9b



10

Fig. 8. Multi-view fitting to oblique angle
Fig. 9a,b. Reconstructed face. **a** Multi-image view; **b** two-image view
Fig. 10. Projected face on mask

tween the true mouth shape and the generated mouth shape is calculated. This process is applied to all of the database entries one by one, and the optimum weight is selected when the minimum error is detected.

10 Interactive experience

In our proposed performance, the user is an actor portraying a storytelling character (see Fig. 10). During the stories, the attendees are the audience

at a live computer-assisted performance. Between stories, however, they can chat with the character. The actor can improvise because the combination of real-time lip synchronization, active projection, and user-controlled facial expressions does away with the need for a fixed script. Surprisingly, coordinating the story-telling process with the manual control of the facial expressions took the actors only about 1 hour to master. After that period, it became quite natural to produce any desired facial expression by clicking the corresponding button on the keypad.

The HyperMask system uses an SGI Indigo2 workstation (MIPS 10000, 128 MB, IRIX6.5), a camera (Sony EVI-G20), an LCD projector (Sony), and a LED-marked mask. The chambermaid costume, wig, shopping cart, and linen are optional. A scene of live demo is shown in Fig. 10. This demonstration was made in the SIGGRAPH'99 Emerging Technology exhibition area. Hundreds of people watched the stories and interact with the two performers behind the mask (Kim Binsted and Claudio Pinhanez). Normally, 5 to 10 people at a time gathered around the performance.

11 Future vision of HyperMask

The HyperMask system is a combination of different technologies, and each will have different social, cultural and technical implications. Active projection could be useful in a number of different applications. For example, in the so-called "Office of the Future", we would like to be able to project dynamically images and information onto moving, irregularly shaped objects. We plan to extend the system to use several cameras and projectors, so that objects can be covered with projected images, which can then be viewed from any direction. More interestingly, we are also considering the use of a system with one fixed projector whose image is deflected by a rotating mirror, similar to the "everywhere displays projector" proposed by Pinhanez (2001). We also hope to be able to make the object markers more subtle, or even remove the need for them completely.

Talking heads with real-time lip synchronization also have a number of potential applications, most obviously as avatars for virtual communities and gaming. We also like to imagine people being able to put themselves into famous movies, by substituting their face for Harrison Ford's (Morishima 1996) (see Fig. 11). In addition, we have proposed "Danger Hamster 2000" (Binsted et al. 2000), which is an entertainment system that uses the HyperMask's technologies (see Fig. 12).

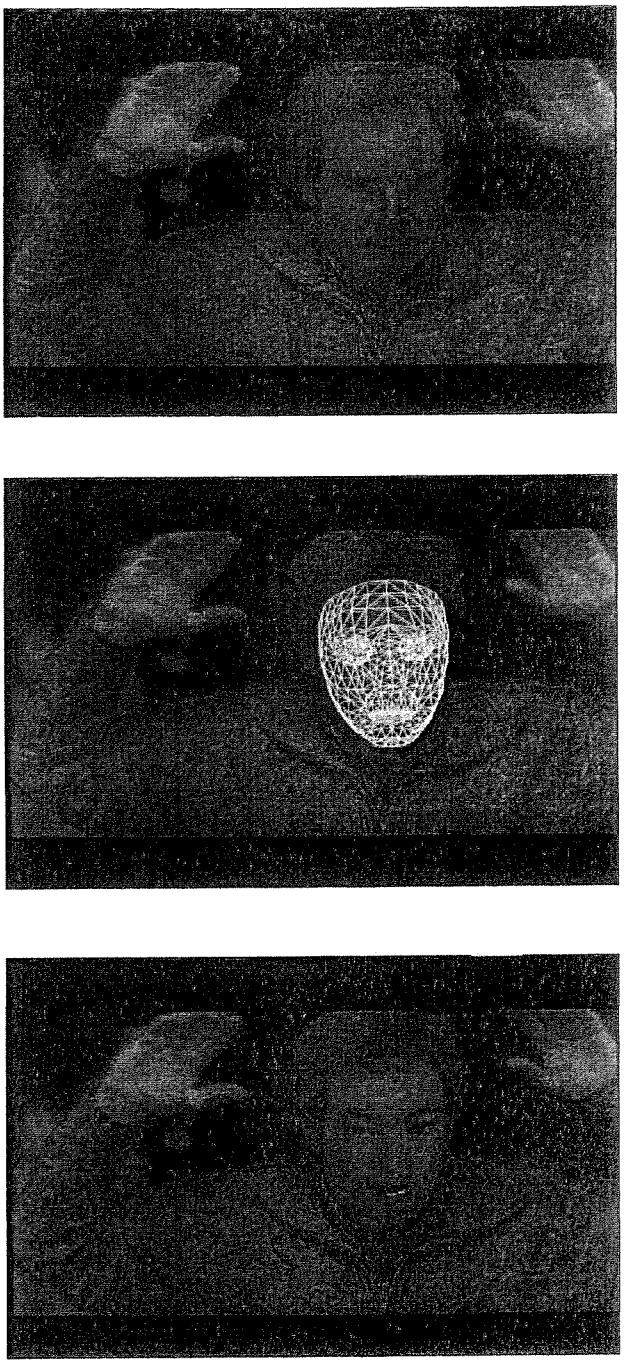


Fig. 11. Interactive movie

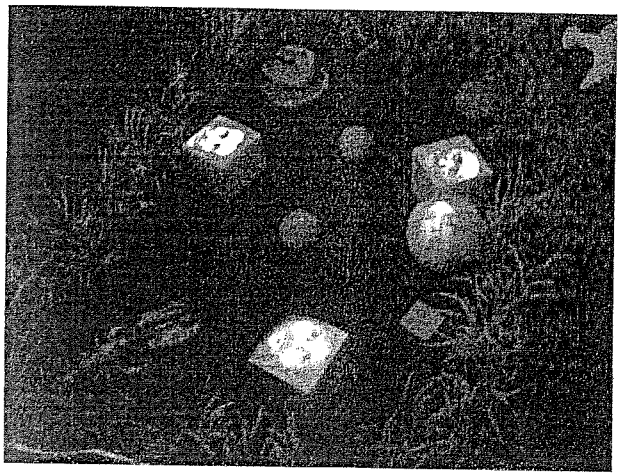


Fig. 12. Danger Hamster 2000 system

Computer-enhanced live performance in general shows a lot of promise. In order to support human performers in their task of entertaining and interacting with a live audience, the technology needs to be flexible, fast, and provide new creative opportunities. We believe that HyperMask is a first step in this direction.

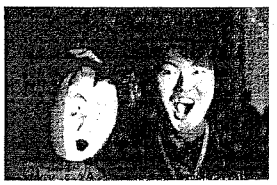
12 Conclusion

We have described HyperMask, a system for projecting images onto an actor's mask as that mask moves around in a performance area. The projected image is an animated face with real-time lip synchronization with the actor's voice. The face's expression is controlled by the actor to fit with the tone and content of the story being told. We also described the HyperMask prototype system, which was put into a linen cart pushed around by a chambermaid, a character who tells amusing stories and chats with the audience.

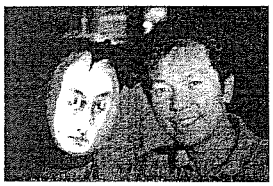
References

1. Rasker R, Welch G, Cutts M, Lake A, Stesin L, Fuchs H (1998) The office of the future: a unified approach to image-based modeling and spatially immersive displays. In: SIGGRAPH Annual Conference Proceedings. ACM, New York, pp 179–188
2. Faugeras O (1993) Three-dimensional computer vision: a geometric viewpoint. The MIT Press, Cambridge, Mass.
3. Gelb A (1974) Applied optimal estimation. The MIT Press, Cambridge, Mass.
4. Lowe DG (1991) Fitting parameterized three-dimensional models to images. IEEE Trans PAMI 13(5):441–450
5. Ekman P, Friesen WV (1978) Facial action coding system. Consult Psychol 13(5):441–450
6. Morishima S (1996) Modeling of facial expression and emotion for human communication system. Displays 17:15–25
7. Kanatani K (1998) Optimal homography computation with a reliability measure. In: Ikeuchi (ed) Proceedings of MVA'98, IAPR Workshop on Machine Vision Applications, Tokyo, IAPR, pp 426–429
8. Zoghlami I, Faugeras O, Deriche R (1997) Using geometric corners to build a 2D mosaic from a set of images. In: Proc IEEE Computer Vision and Pattern Recognition. IEEE, Piscataway, NJ, pp 420–425
9. Pinhanez C (2001) The everywhere displays projector: a device to create ubiquitous graphical interfaces. In: Proc of Ubiquitous Computing (Ubicomp '01). Atlanta, GA
10. Binsted K, Nielsen F, Morishima S, Misawa T (2000) Danger Hamster 2000. In: ACM SIGGRAPH Conference Abstracts and Applications. ACM, New York, pp 81

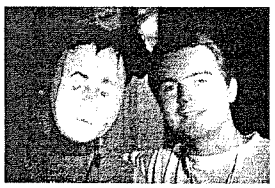
Photographs of the authors and their biographies are given on the next page.



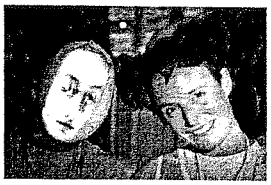
TATSUO YOTSUKURA received his B.S. and M.S. degrees, both in the Faculty of Engineering, from Seikei University, Tokyo, in 1998 and 2000, respectively. Currently, he is a Ph.D. student at Seikei University and an intern researcher at the ATR Media Integration & Communication Laboratories. His research interests include facial animation and realtime face-to-face communication systems. He is a member of IEICE-J. He received the NICOGRAPH/MULTIMEDIA best paper award in 2000 and the IEICE-J young engineer award.



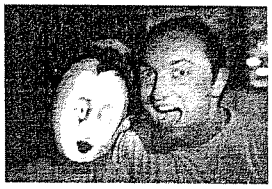
SHIGEO MORISHIMA received his B.S., M.S. and Ph.D. degrees all in electrical engineering from the University of Tokyo in 1982, 1984, and 1987, respectively. Currently, he is a professor at Seikei University, Tokyo. His research interests include physics-based modeling of face and body, facial-expression recognition and synthesis, human–computer interaction, and future interactive entertainment using speech and image processing. He was a visiting researcher at the University of Toronto from 1994 to 1995. He has been engaged in the Multimedia Ambiance Communication TAO research project as a sub-leader since 1997. He has been a temporary lecturer at Meiji University, Japan, since 2000 and a visiting researcher at the ATR Spoken Language Translation Research Laboratories since 2001. He is an editor of Transactions of the Institute of Electronics, Information and Communication Engineers, Japan (IEICE-J). He received the IEICE-J achievement award in May 1992.



FRANK NIELSEN received his B.S. and M.S. degrees from École Normale Supérieure (ENS) at Lyon in 1992 and 1994, respectively. He defended his Ph.D. thesis on "Adaptive Computational Geometry" prepared at INRIA Sophia-Antipolis under the supervision of Pr. Boissonnat in 1996. As a civil servant of the University of Nice (France), he gave lectures at the engineering schools ESSI and ISIA (École des Mines). In 1997, he served in the army as a scientific member of the computer-science laboratory of École Polytechnique (LIX). In 1998, he joined Sony Computer Science Laboratories, Tokyo, as an associate researcher. His current research interests include computational geometry, algorithmic vision, combinatorial optimization for geometric scenes and compression.



KIM BINSTED is CEO of I-Chara Inc., a Tokyo-based mobile agent company (www.i-chara.com). Formerly, she was a researcher at the Sony Computer Science Laboratories, working on Human Computer Interaction and Artificial Intelligence (AI). She received her Ph.D. in AI at the University of Edinburgh and her B.Sc. in physics at McGill University, Montreal.



CLAUDIO PINHANEZ received his B.S. degree in mathematics and his M.S. degree in computer science from the University of Sao Paulo in 1985 and 1989, respectively. He received his Ph.D.in 1999 from the Media Arts & Sciences Graduate Program at the Media Laboratory, Massachusetts Institute of Technology. Currently, he is research scientist at the IBM T.J. Watson Research in New York. His interests are interactive spaces; user models for human–computer interaction; machine intelligence; computer vision; computerized entertainment; interactive stories; and computer theater.

論　文

HyperMask：3次元顔モデルを用いた仮面の構築

四倉　達夫[†,††]　Kim Binsted[†††]　Frank Nielsen[††††]
Claudio Pinhanez[†††††]　鉄谷　信二[††]　中津　良平[††]
森島　繁生[†]

HyperMask: Reactive Talking Head for Storytelling

Tatsuo YOTSUKURA[†,††], Kim BINSTED[†††], Frank NIELSEN[††††],
Claudio PINHANEZ[†††††], Nobuji TETSUTANI[††], Ryohei NAKATSU[††],
and Shigeo MORISHIMA[†]

あらまし　HyperMaskとは従来単一の顔表情や人物を表現する仮面の概念を進化させ，一つの仮面からあらゆる表情や人物を自由に生成及び表現可能なシステムである．本システムを用いることで，その仮面を装着した役者の表現の幅や新しい演出方法が生み出されていくと考えられる．顔の表出手法として，仮面に装着された五つのLEDを，カメラにより追跡することで仮面の位置及び方向を求め，プロジェクタによって算出されたパラメータをもとに顔画像の投影を行う．また投影されている顔画像は演技者の音声を分析することによりリアルタイムで音声同期して口形状のアニメーションを行い，顔表情や人物の切換はユーザが任意に選択可能である．本論文ではHyperMaskシステムを用いた演出支援装置を紹介し，新たな仮面の表現技法の確立を目指す．

キーワード　仮想空間，顔合成，ニューラルネットワーク，ホモグラフィ

1. まえがき

HyperMaskは仮面を装着した演劇への支援手法であり，観客に対する演技者の表現力を大きく広げ，新たな演出が構築可能なシステムである．俳優や演劇者などに，白色の仮面を装着させ，プロジェクタによって表情・口形状変形及び人物の切換が可能な3次元顔モデルを投影しストーリーや場面，状況の変化に応じて容易に操作ができるように構成されている．また，演技者の動きに応じて仮面に正しく顔モデルを投影でき，自由度の高いシステムとなっている．

一般的に仮面と呼ばれているものは文化・宗教・地域によって様々なものが存在している．また仮面制作において写実的な表出を施してあるものもあれば逆に抽象的表現を施したものもあり多種多様である．それらは一様に単一の表情・人物が描かれており，演技するキャラクタのストーリー上での素性や立場に応じて使い分けている．また日本の古典芸能である能面のような演技者の動きや観客の視線方向，照明条件，ストーリーや音楽等様々な舞台環境から観客の心理状態，創造により仮面に内的な表情を付加させ演出を行う手法も存在する[1]．HyperMaskの演出手法は従来の手法と異なり，外的な表情の変化，そして役柄自体の切換が可能で従来の仮面の概念を超えた自由度の高い演出が表現できると考えられる．また単純な表情から内的な感情を付加させるような複雑な表情を容易に表出可能にするため，仮面に投影する顔画像はリアルな顔モデルを用いるよう工夫した．それにより従来の仮面を使った独特な演出を熟知していない演技者でも簡単に表情の変化が操作可能で，観客もまた直感的に演技

† 成蹊大学工学部，武蔵野市
Faculty of Engineering, Seikei University, 3-3-1 Kichijoji-kitamachi, Musashino-shi, 180-8633 Japan
†† ATR 知能映像通信研究所，京都府
ATR Media Integration & Communication Laboratories, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto-fu, 619-0288 Japan
††† I-Chara，東京都
I-Chara K.K., 2-34-1 Uehara, Shibuya-ku, Tokyo, 151-0064 Japan
†††† ソニーコンピュータサイエンス研究所，東京都
Sony Computer Science Laboratories, Inc., 3-14-13 Higashigotanda, Shinagawa-ku, Tokyo, 141-0022 Japan
††††† IBM T.J. ワトソンリサーチセンタ，米国
IBM T.J. Watoson Research, 30 Saw Mill River Rd. (Route 9A) – Hawthorne, NY 10532, U.S.A.

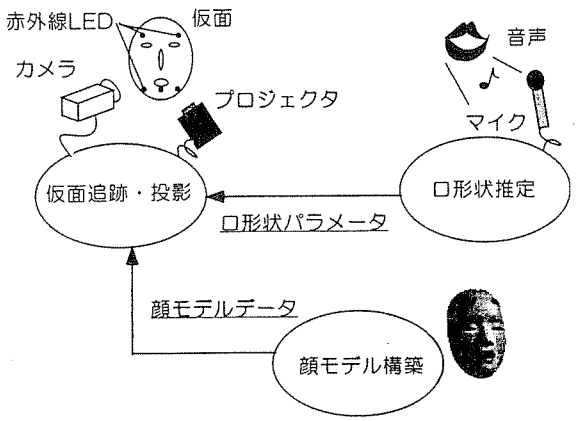


図 1　システム構成
Fig. 1　System feature of HyperMask.

者の表情を読み取ることができ，没入度の高い演劇空間が広がると考えられる．

本システムの構築にあたり，(1) 演技者が装着している仮面の動きを正しく追跡し，仮面上に顔画像を正しく投影する手法，(2) 仮面に投影する顔モデルの生成及び口形状・表情の制御法，(3) 音声から口形状への推定，計三つの基盤技術を用いている（図 1）．(1) において，本論文では簡単かつ短時間でキャリブレーションを行え，また精度の高い追跡・投影が可能な方法を提案し，(2) ではフレシキブルな顔モデルの構築を目指し，3 次元ワイヤフレーム上に顔のテクスチャ画像を容易にフィッティング可能なツールを開発した．また口形状・表情変化に必要なワイヤフレームの特徴点制御ルールを紹介する．(3) では仮面上の顔画像をリアルタイムにて制御を行うため，演技者の声をニューラルネットワークによって分析を行い，リアルタイムに音声と同期させ合成を行った．表情変形及び投影を行う顔モデルの変更は演技者がマニュアルで操作可能である．

紹介した三つの基盤技術は HyperMask を構築するために大変重要な要素となっているが，本システムのみ利用可能ではなく，他の分野へ容易に応用・転用可能できると期待される．例として技術 (1) を用いて “The Office of the Future” [2], [3] と呼ばれる実世界と仮想空間との共有インタラクティブスペースの構築に利用できると考えられる．また技術 (2), (3) では顔画像生成技術は制御パラメータのみで表情・口形状変形が可能であることから MPEG-4 [6] プロトコルに類似したテレビ電話やサイバスペース上での低ビットレートコミュニケーションシステムなど幅広い応用が期待できる．

以下，これらの基盤技術を **2.** (1) 仮面の追跡・投影，**3.** (2) 顔モデル構築，**4.** (3) 口形状推定，と順に追って紹介し，**5.** で実際に HyperMask システムを用いたプロトタイプの演出支援システムの構築，そしてプロトタイプを用いてデモンストレーションを行い，そのシステム評価結果を述べていく．

2.　仮面の追跡・投影

本章では 1 台のカメラとプロジェクタを用いた基盤技術の一つである (1) 仮面の追跡・投影について述べる．演技者は舞台上で静止していることはほとんどなく，演出に応じた動きを行う．そのために投影する顔モデルと仮面とが常に正しく投影され，かつ演技者の動きの制約を可能な限りなくしたシステム設計が求められる．またカメラとプロジェクタのキャリブレーションに関しても利用者の専門知識の必要なく短時間で各種パラメータが設定不要の単純な操作が望まれる．そこで本手法ではそれらの問題を解決すべく，次に述べる手法でキャリブレーションを行った．

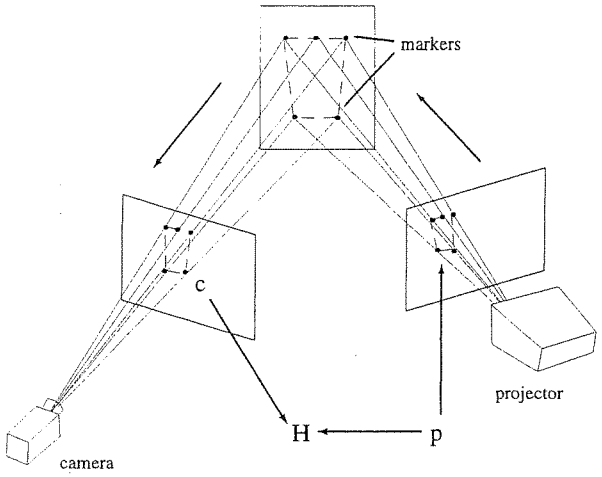
2.1　キャリブレーション [4]

カメラ対と平面上に存在する観測点との関係を調べる際，一般的にホモグラフィ [7]（共線変換とも呼ばれる）が用いられている．ホモグラフィは実空間中の同一平面上に乗る複数点を 2 台のカメラで撮影したときの画像間での対応を表現し，ホモグラフィで記述される対応は同一平面上のみ有効で，2 台のカメラの位置や対象となる平面に依存する 3 行 3 列の行列で定義される．

投影モデルを考える際，基礎的な概念として理想状態のピンホールカメラモデルがよく知られているが，プロジェクタもまた理想状態でのピンホールモデルとして考えても差し支えない（図 2）．H をホモグラフィとするとプロジェクタ画像フレーム $\bar{p} = (x_p/w_p, y_p/w_p)$ とカメラ画像フレーム $\bar{c} = (x_c/w_c, y_c/w_c)$ との関係は次式のように示される．ただし，座標系は同次座標 $p = (x_p, y_p, w_p)$，$c = (x_c, y_c, w_c)$ を用いる．

$$p = \begin{pmatrix} x_p \\ y_p \\ w_p \end{pmatrix} = Hc = H \begin{pmatrix} x_c \\ y_c \\ w_c \end{pmatrix} \tag{1}$$

もし両画像平面上の 4 点の基準点となる座標がわかれば，ホモグラフィは完全に定義できる．本手法では仮面上に設定した 4 点とプロジェクタから仮面表面の 4



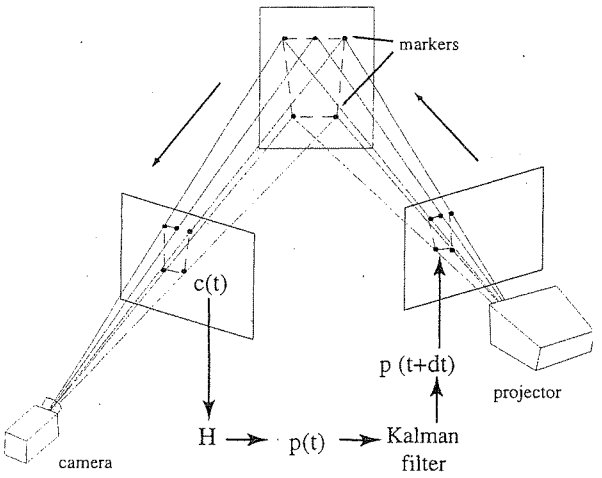


図 2　キャリブレーションと実行時のプロセス
Fig. 2　Calibration process and run-time process.

点の位置関係が対応した 4 点の画像をマニュアルで合わせることでカメラとプロジェクタ間のホモグラフィを求めた（図 2）.

プロジェクタの 4 点の同次座標（$w_p = 1$ とする）を

$$p_i = (x_p^i, y_p^i, 1) \qquad i = 1,2,3,4 \tag{2}$$

とし，次にカメラの 4 点の同次座標（$w_c = 1$ とする）を

$$c_i = (x_c^i, y_c^i, 1) \qquad i = 1,2,3,4 \tag{3}$$

とした.

ホモグラフィ行列 H を

$$H = \begin{pmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{pmatrix} \tag{4}$$

とし，同次座標 (x, y, w) が (x'', y'', w'') に変換されると定義する．同次座標系は定数倍の自由度があり，ホモグラフィ行列も定数倍の自由度があるため $h_9 = 1$ としておくことができ，8 パラメータの変換として考えることが可能である.

$$x' = \frac{x''}{w''} = \frac{h_1 x + h_2 y + h_3}{h_7 x + h_8 y + 1} \tag{5}$$

$$y' = \frac{y''}{w''} = \frac{h_4 x + h_5 y + h_6}{h_7 x + h_8 y + 1} \tag{6}$$

上式は以下のようにも示される.

$$x' = h_1 x + h_2 y + h_3 - h_7 x x' - h_8 y x' \tag{7}$$

$$y' = h_4 x + h_5 y + h_6 - h_7 x y' - h_8 y y' \tag{8}$$

よって次式のような線形系で示され，H を 8 行 8 列 S の逆行列で求めることができる.

$$S = \begin{pmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x_1' x_1 & -x_1' y_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -y_1' x_1 & -y_1' x_1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -x_2' x_2 & -x_2' y_2 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -y_2' x_2 & -y_2' x_2 \\ x_3 & y_3 & 1 & 0 & 0 & 0 & -x_3' x_3 & -x_3' y_3 \\ 0 & 0 & 0 & x_3 & y_3 & 1 & -y_3' x_3 & -y_3' x_3 \\ x_4 & y_4 & 1 & 0 & 0 & 0 & -x_4' x_4 & -x_4' y_4 \\ 0 & 0 & 0 & x_4 & y_4 & 1 & -y_4' x_4 & -y_4' x_4 \end{pmatrix}$$

$$H'^T = (h_1 \quad h_2 \quad h_3 \quad h_4 \quad h_5 \quad h_6 \quad h_7 \quad h_8)$$

$$Z^T = (x_1' \quad y_1' \quad x_2' \quad y_2' \quad x_3' \quad y_3' \quad x_4' \quad y_4')$$

$$SH' = Z \tag{9}$$

また

$$x_i' = x_p'^i,\ y_i' = y_p'^i,\ x_i = x_c^i,\ y_i = y_c^i,\ i \in \{1,2,3,4\}$$

と置き換えることができる.

先ほども述べたとおり，H' を求めるとき S の逆行列が必要であるが，特異値分解や擬似逆行列を用いて逆行列を求める必要性もあり，安定した H' を求めることが困難な場合がある．他の手法として四つ以上の点を配置して最小 2 乗法や Kanatani [5] らの手法があるがこれ以上点を増やすことは投影する顔画像の印象に影響が出るため好ましくない．そこで本手法ではカメラ，プロジェクタ両方とマスクの距離が十分に遠いと仮定することにより，アフィン変換によって H を導出する.

4 点のカメラ，プロジェクタ

$$P = (p_1^T, p_2^T, p_3^T, p_4^T) \tag{10}$$

$$C = (c_1^T, c_2^T, c_3^T, c_4^T) \tag{11}$$

とし，$P = HC$ を用いて H を求めると

$$H = PC^T(CC^T)^{-1} \tag{12}$$

のようになる．実行時はカメラ画像 $c = (x_c, y_c, 1)$ と，キャリブレーションの際に求まった H を用いて p を求めることができる．本手法で用いたキャリブレーションは予備実験の結果4点のみを用いて非常に安定した処理を実現しており，キャリブレーションに要する時間も短時間で済む．またカメラやプロジェクタ固有のパラメータを必要としない．

2.2 投影面の追跡

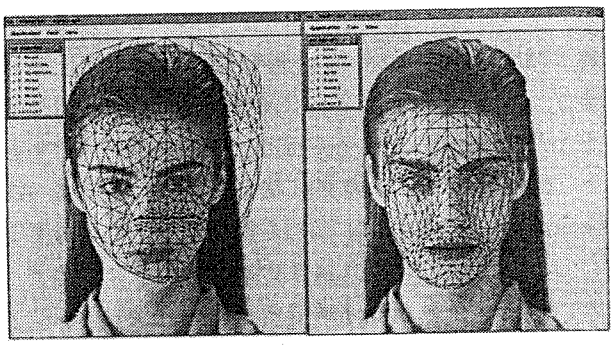
投影面の追跡を行う際，本論文では赤外線 LED を用い，カメラ部には赤外線フィルタを装着させている．また，4点 LED のラベリングを行うためにもう1点ラベリング用の LED を用意し，計五つの LED を仮面に付けた．投影面の追跡及びトラッキングは非常に安定していると前節で述べたが，マスクの移動が非常に早い場合，投影画像に「遅延」が生じてしまう問題がある．カメラ画像の取込み時間を t とし，プロジェクタへの投影 $t + dt$ とすると dt の遅延が生じてしまう．この問題を解決するため本論文ではカルマンフィルタ [8] を用いた（図 2）．パラメータ dt はカメラ画像の取込みからプロジェクタの投影までの平均時間とし，経験的に求めた．カルマンフィルタを用いた場合と用いない場合を実験で比べてみたところ，フィルタによって遅延現状が軽減されており満足のいく結果となった．

3. 顔モデル構築

HyperMask では実際の人物の顔と同レベルに近いクオリティでの表出を目的にしているため，微妙な表情が表出可能で観客に対し違和感を与えない顔モデル，また口形状や表情の制御ルールが必要となる．本章では基盤技術（2）にあたる表示用顔モデルの作成法の紹介，また作成した顔モデルの表情や口形状の制御方法を紹介する．

3.1 3次元顔モデル

リアルな顔モデルの製作のため，演技に使用する対象人物の正面画像を三角形ポリゴンで構成させる顔の標準ワイヤフレームモデルをマニュアル整合し，個人モデルを作成する．このモデルは約 850 ポリゴンの三角形パッチにより構成されていて，格子点数は約 480 点から形成される．ポリゴン数は形状の変化の際，演算量及びレンダリングの処理時間に直接関係する．ここでは実時間でのアニメーション実現のため，動きの変化の激しい部分（唇，眉，眼周辺部）にのみ細かいポリゴンを割り当て，全体的な演算量の軽減を行っている．そしてこのモデルにテクスチャマッピングを施すことによって顔合成画像を作成する．また唇を開けたときを考慮し歯及び口内部のモデルを追加した．歯のモデルは白色系のグローシェーディングを施しており，口内部は袋状のモデルにペイントツールで作成したテクスチャイメージを付けた．



(a) Initial Model (b) Fitted Model

図 3 整合ツールウィンドウ
Fig. 3 Fitting tool's window.

顔モデルを対象人物に整合した様子を図 3 に示す．顔モデルの整合を容易に行うため，GUI ツールを開発した [9]．まず演出の際に必要な顔画像を読み込む．顔モデルのワイヤフレームモデルの格子点を動かし画像と顔モデルの整合を行う．点の移動ははじめマクロに制御して，次第に細かく位置合せでできるように考慮されている．また実際に表情変形してみて，不自然な部分はインタラクティブに位置修正できるように配慮されている．特に目と唇の部分は表情変形に重要であるため綿密な整合が必要である．図 3 (a) は整合前の編集画面であり，(b) は整合された後の画面を示している．このツールを用いて顔モデルを完成させる所要時間は全くの初心者でも約5分程度で完成できる．

3.2 表情及び口形状のパラメータ化

仮面に投影された顔モデルの表情や口形状変化を表現する顔画像を構築するために，3次元顔モデルの幾何学的変形の基準となる特徴点の設定と，その移動量の記述，そして特徴点の周囲の格子点の移動規則などを定める必要がある．ここではモデル変形の基礎とな

る表情と口形状の制御パラメータについて述べる．

3.2.1 表情パラメータ

表情パラメータとして心理学の分野で提案されている FACS（Facial Action Coding System）[10] と呼ばれる動きの方向を解剖学的に考慮して顔の表情を AU（Action Unit）と呼ばれる 44 個の基本動作に分類している．あらゆる表情は AU の組合せで表現できるとされ，FACS は表情記述単位として顔画像の分析，合成分野で広く用いられている．各 AU は顔面上の特徴点の 3 次元移動ベクトルとして定義されている．表情変化は 3 次元モデルの特徴点の AU の強さによって移動させ，特徴点以外の格子点は，特徴点の移動に基づく補間によって制御される．特徴点は 48 点設定して，各特徴点の位置は唇・眉・目輪郭部と頬周辺部，額上部と耳に配置している．感情の種類としてこの AU の組合せによって表現された，怒り，喜び，悲しみ，嫌悪，驚き，おそれの 6 基本感情を標準として用意した．もちろん，この AU のポリゴン数は形状の変化の際，演算量及びレンダリングの処理時間に直接関係する．ここでは実時間でのアニメーション実現のため動きの変化の激しい部分にのみ細かいポリゴンを割り当て，全体的な演算量の軽減を行っている．このモデルにテクスチャマッピングを施すことによって顔合成画像を作成する．また，歯及び口内部のモデルを追加した．編集によってユーザ自身で感情をカスタマイズするための AU エディタも用意されている．図 4 に基本 6 感情の合成画像の一例を示す．これはあくまで標準として用意するもので，ユーザによるカスタマイズは容易に実行可能である．

3.2.2 口形パラメータ

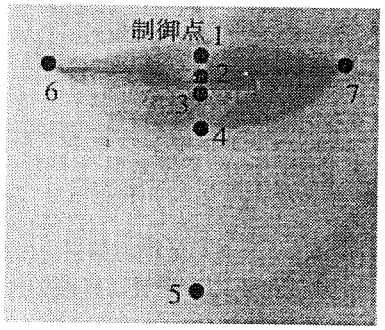
発話時の口形状を表現するために，先に述べた AU とは異なる，口領域の変形に限定したパラメータを用いる．パラメータは五つの母音（/a/, /i/, /u/, /e/, /o/）と閉口の口形状を基準とし，すべての口形状はこれらの補間によって再現できると仮定している．

口領域の動きを少数のパラメータで表現するために，口領域の制御点のパラメータとして図 5 のような 13 個を定めた．3 次元計測結果に基づいて，この制御点自体の移動量の算出，更に制御点以外の格子点の移動量算出ルールを定めた．この 13 個の座標値によって，唇の形状を一意に決定することができる．図 6 にこの口形パラメータによって表現された口形/u/の合成画像を示す．

(a) Anger (b) Sadness

(c) Happiness (d) Surprise

図 4 顔モデルによる各表情合成画像
Fig. 4 Example of face synthesis images.



制御点	口形パラメータ	制御
1	1	上唇上側の縦方向の動き
	2	上唇上側の奥行方向の動き
2	3	上唇下側の縦方向の動き
	4	上唇下側の奥行方向の動き
3	5	下唇上側の縦方向の動き
	6	下唇上側の奥行方向の動き
4	7	下唇下側の縦方向の動き
	8	下唇下側の奥行方向の動き
5	9	あごの縦方向の動き
6	10	口角の縦方向の動き
	11	口角の横方向の動き
	12	口角の奥行方向の動き
7	13	唇の横方向の開き具合

図 5 口形パラメータ
Fig. 5 Location of mouth parameters.

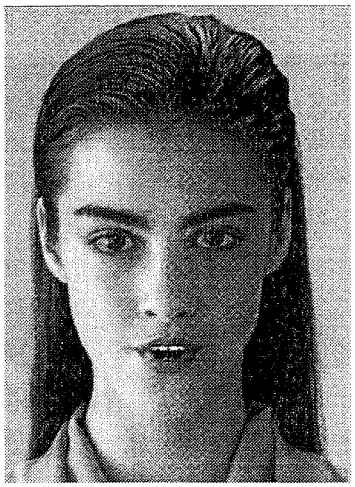
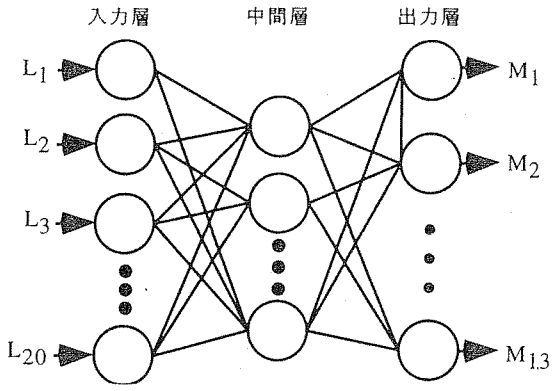


図 6　口形/u/の合成画像
Fig. 6　Face synthesis image of vowel 'u.'



L_n: LPC cepstrum, M_n: Mouth parameter

図 7　口形パラメータへの変換に用いるニューラルネットワーク
Fig. 7　Neural Network for conversion to mouth shape parameters.

4.　音声情報から口形状の推定 [11]

三つ目の基盤技術となる口形状の推定手法は顔モデルの口形状をリアルタイムに決定するため，ユーザから入力された音声をフレームごとに分析することによって，毎フレーム口形パラメータを推定する．特徴パラメータとして計算時間が比較的少なく，また発話者の声動特性と放射特性の特徴を表現していると考えられる LPC ケプストラム係数とした．入力音声は 16 [kHz]，16 [bit] とし，分析フレーム長及び周期は 32 [ms] で切り出す．

LPC ケプストラムから口形パラメータへの変換は図 7 のような 3 層フィードフォワード型ニューラルネットワークを用いている．入力層は LPC ケプストラム次数と同じ 20 ユニット，出力層は 13 個の口形パラメータに相当する．更に中間層は経験的に 20 ユニットとした．学習パターンは 5 母音の LPC ケプストラムとそれぞれの発話時の口形パラメータ，及び無発音時の周囲の環境雑音から求めた LPC ケプストラム係数と閉口口形とした．収束までに 100 万回の学習を行った．このニューラルネットの重み係数は基本的に話者依存性が強く，話者ごとに事前に学習を行う必要がある．この問題を解決するために後述する話者適応処理によってこの学習を省略することもできる．

4.1　話者適応

本システムは不特定多数の方が利用すると考えられるが，ユーザが更新されるたびにニューラルネットによって学習を行うことは非能率的である．そこであらかじめ収録した 100 人分の学習データで重み係数のデータベースを構築した．この中からユーザに最適な重み係数を自動的に選択する．新しいユーザには，実験開始直前に 5 母音を発生してもらい，データすべての中から一つずつ選択された重み係数によって順次口形状推定を行い，基準の 5 母音の口形状に最も近いものを生成できる重み係数をその人物の最適な係数と判断して，話者適応を実施した．

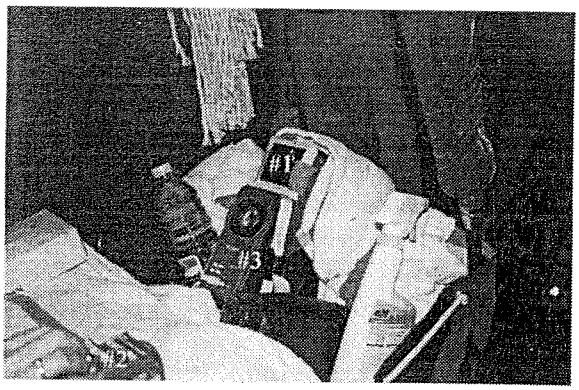
4.2　口形状推定評価

1995 年 8 月にロサンゼルスで行われた ACM の SIGGRAPH '95 において，インタラクティブデモ展示を行った [12]．このデモでは，会場に訪れた人物の顔正面画像と 5 母音の音声をその場で取り込み，モデル整合と話者適応処理の後に，リアルタイムでマイクから入力された音声を分析して，口形を合成する処理を行い，合成された顔画像を通じて 2 者間で対話を行うというものであった．このデモにおいて，来場者 160 人の整合処理と話者適応を実施し，すべての人物において自然な口形状と表情の合成が可能であることが明らかとなった．なお，この際の表情合成速度は毎秒 10 フレームであり，すべての外国人を対象として対話は英語で行われた．整合処理は経験のある人物によって実施されたが，平均 1 分程度の所要時間であった．

5.　プロトタイプ

前章まで述べた手法を用いて図 8 のようなプロトタイプを構築した．プロトタイプを構築するにあたり，演技者が舞台内を自由に動け，演技に支障をきたさないシステムを構築するために図 9 のようなカート内にカメラやプロジェクタ，コンピュータ等を収納し，装

図 8 プロトタイプ
Fig. 8 Prototype of HyperMask.



#1: Camera (covering Infrared Filter)
#2 Key-pad #3 Projector.

図 9 カート内部
Fig. 9 Trolley inside.

置自体が自由に移動可能なポータブルシステムを構築した．

5.1 システム構成

処理用のワークステーションとして，SGI 社製 Indigo2（MIPS 10000, 123 MByte, IRIX6.5），を使用した．このワークステーション 1 台で（1）仮面の追跡・投影，(3) 口形状推定のプロセスを並列処理している．仮面追跡用カメラ（Sony EVI-G20），顔画像投影用プロジェクタ（Sony），そして赤外線 LED が埋め込まれた白色の仮面を用いた．仮面の目にあたる部分は演技者の視界を確保するため数箇所小さな穴を空けた．これらの穴は直径 1.5 mm 程度であるため投影画像に影響を受けず，プロジェクタの輝度による演技者への影響は今回使用したプロジェクタの輝度が 50ANSI ルーメンで演技者の目に与える負担は非常に少なく，また予備実験から視界に影響を与えないことがわかった．そして演技者が台詞を発するとき声がこもり，観客に聞こえないおそれがあるため，仮面内側には小型マイクを付け，カート内にあるスピーカと接続させ声量の確保を行っている．

プロトタイプを運用する前にいくつかの準備を行う必要がある．まず基盤技術（2）を用いて演出に用いる顔モデルの制作を行う．次に口形状の推定を行うために必要なニューラルネットの重みデータの学習を行う．声の収録の際，演技者には仮面を装着してもらいデモ環境と同じ状態で録音した．今回演技者が 2 名であり，最適な推定を行うため先述した話者適応の手法は使用しなかった．最後にカメラのキャリブレーションを行って準備が完了する．

演技者はカートを押しながら舞台を移動しパフォーマンスを行い，また観客とのインタラクションを行う．仮面に投影された顔は様々なストーリー展開や口調，観客とのインタラクションによって変化する．カート上のカメラは常時演技者の仮面を追跡し，プロジェクタもまたリアルタイムに顔表情を合成させたモデルを投影する．演技者が発話したセリフすべて最適な口形状へとリアルタイムに処理され，顔モデルの口形が生成される．顔表情や投影する人物の顔画像の変更はカート上に装備してあるテンキーによって演技者が任意に変更が可能である．

5.2 評価実験

本プロトタイプを用いて，1999 年 8 月 SIGGRAPH '99 のエマージングテクノロジーにて実際に一つのオリジナルストーリーを作成し，観客とのインタラクティブなコミュニケーションも取り入れたデモンストレーションを行った [13]．顔画像は図 10 のように投影され，ストーリーやアドリブによって口形状をリアルタイムに推定，合成を行い，表情表出及び演出する役の切換はカートに備え付けているキーパッド（図 9）によって任意に変更可能となっている．図 11 にデモンストレーションの様子を示す．

1 回のパフォーマンスで約 10 人から 30 人の観客が訪れ，開催期間中約 1000 人参加した．デモに参加した観客の大多数が本システムの演出法に対し大変興味をもち，斬新かつ応用性をもつシステムであると好評を得た．また投影した顔表情の合成フレームレートは毎秒 8〜10 フレーム前後ではあったが仮面に投影された顔モデルの口形状に対して同期や表情表出が自然に

図 10　顔画像投影例
Fig. 10　Projected face on mask.

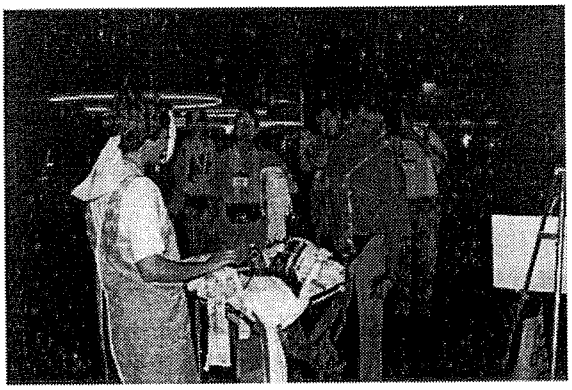


図 11　デモンストレーション風景
Fig. 11　Snapshot of demonstration.

表現されているとの意見を頂いた．実際に音声と口形状との遅延時間を計測したところ約 30 ms ほど音声の方が早く出力していたが同期には影響しなかった．

6. むすび

本論文ではプロジェクタによって口形状や表情が変化し，投影する人物の顔が選択可能な仮面を用いた演技支援システム “HyperMask” を提案した．本システムは（1）仮面追跡，(2) 顔モデル構築，(3) 口形状推定，以上三つのプロセスで構成され，各々のプロセスで用いた手法の考察は以下のとおりである．

（1）　仮面追跡・投影

赤外線 LED を装着した白色の仮面のトラッキング及びプロジェクタによる投影に関し，ホモグラフィを用いることで 4 点（LED のラベル付けを含めると 5 点）の LED のみで容易かつ短時間にキャリブレーションでき，仮面の追跡の精度も良い結果がでた．仮面 LED のキャプチャから投影までのプロセス処理時間で発生する「遅延」の問題はカルマンフィルタを用いることで，ほぼ仮面の動きと顔画像とが遅延なく投影されていることが確認できた．このアルゴリズムでの仮面の回転運動の許容範囲は実験結果から，ロール ϕ・ピッチ θ・ヨー ψ で示すと $0° \leq \phi < 360°$, $-20° \leq \theta \leq 20°$, $-30° \leq \psi \leq 30°$ である．すべての LED がカメラによってとらえなくてはならないためにこのような制限があるが，演技者の動作に支障が生じることはなかった．

（2）　顔モデル構築

HyperMask で使用する仮面の顔画像はリアルな顔を再現することをが目的となっている．そこで標準ワイヤフレームモデルを用意し演技用の顔画像とのフィッティングを行い，ワイヤフレームモデルにあらかじめ設定した表情・口形状制御ルールで表出を行った．これらのルールは顔の一つひとつの基本的な動きをパラメータとして定義したので基本 6 感情のみならず複雑な表情やすべて音素発音時の口形状に対応できる．ワイヤフレームモデルと顔画像とのフィッティングでは GUI ベースの整合ツールを用意し簡単に処理可能である．

（3）　口形状推定

演技者の音声情報からニューラルネットワークを用いることで母音推定が可能となった．母音のみの推定ではあるが先に述べたデモンストレーションで英語発生時の口形状評価を行ったところ，自然な表出が出ているとの回答を頂いた．また推定に要する処理時間も大変少ないため音声と合成画像との同期も問題がなかった．今後母音のみならず唇の形状と密接な関係をもつ他の音素（例えば破裂音の/Ba/, /Pa/）の推定を行いより精度の高い推定システムの構築を考えている．

そしてこれらを用いて実際に HyperMask システムのプロトタイプを製作し，演技者が舞台内を移動できるようカート状のポータブルなシステムを提案した．観客とのインタラクションやオリジナルストーリーのデモンストレーションを行うことで HyperMask システムの実現性が確認でき，従来までの仮面の概念を超えた全く新しい演出技法の有効性が明らかとなった．今後の展開として，システム全体の軽量・小型化，使いやすさの向上を図っていく．現状ではワークステーションやプロジェクタ等をカートに入れているが，近年の急速な電子機器の発展により実現は困難ではない．

また表情の操作，人物の入換にはキーパッドを使っているが今後，操作デバイスの改良や音声による感情推定システムによる自動化を考慮していく．

本システムはカメラによる追跡・投影，顔合成，音声分析技術を融合している．そのためこれらの技術はHYPERMASKシステムのみならず今後，様々な分野への応用化が可能であると考えられる[14]．現状では演劇に特化したシステム構成となっているが，カメラ追跡・投影技術（追跡プロセス）を用いることで先に述べた“The Office of the Future”への転用も可能である．また顔合成・音声分析システム（音声分析・投影プロセス）を用いてフェース・トゥ・フェースでの多人数コミュニケーションシステム，電子会議システム等への応用化も検討中である．

文　献

[1] M.J. Lyons, A. Plante, M. Kamachi, S. Akamatsu, R. Campbell, and M. Coleman, “Viewpoint dependent facial expression recognition: Japanese noh masks and the human face,” Proc. 22nd Annual Conference of the Cognitive Science Society, pp.322–327, 2000.

[2] R. Rasker, G. Welch, M. Cutts, A. Lake, L. Stesin, and H, Fuchs, “The office of the future: A unified approach to image-based modeling and spatially immersive displays,” SIGGRAPH Annual Conference Proceedings, pp.179–188, 1998.

[3] C. Cruz-Neira, S.J. Daniel, and T.A. DeFanti, “Surround-screen projection-based virtual reality: The design and implementation of the CAVE,” Computer Graphics, SIGGRAPH Annual Conference Proceedings, pp.135–142, 1993.

[4] C. Pinhanez, F. Nielsen, and K. Binsted, “Projecting computer graphics on moving surfaces: A simple calibration and tracking method,” SIGGRAPH Sketches and Application, p.266, 1993.

[5] K. Kanatani, “Optimal homography computation with a reliability measure,” Proc. MVA ’98, IAPR Workshop on Machine Vision Applications, pp.426–429, 1998.

[6] E. Patajan, “Approaches to visual speech processing base on the MPEG-4 face animation standard,” Proc. ICME2000, 2000.

[7] O. Faugeras, Three-Dimensional Computer Vision: A Geometric Viewpoint, The MIT Press, Cambridge, Massachusetts, 1993.

[8] A. Gelb, Applied Optimal Estimation, The MIT Press, Cambridge, Massachusetts, 1974.

[9] 森島繁生，八木康史，金子正秀，原島　博，谷内田正彦，原　文雄，橋本周司，“顔の認識・合成のための標準ソフトウェアの開発，” 信学技報，PRMU97-282, 1998.

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

[11] S. Morishima, “Modeling of facial expression and emotion for human communication system,” Displays, vol.17, pp.15–25, 1996.

[12] S. Morishima, “Better face communication,” SIGGRAPH Sketches and Application, p.117, 1995.

[13] K. Binsted, “Virtual reactive face for storytelling,” SIGGRAPH Sketches and Application, p.186, 1999.

[14] K. Binsted, T. Misawa, S. Morishima, and F. Nielsen, “Denger hamster 2000,” SIGGRAPH Sketches and Application, p.81, 2000.

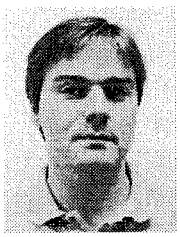
（平成13年2月5日受付，6月12日再受付）

四倉　達夫　（学生員）

平10成蹊大・工卒．平12同大大学院修士課程了．現在同大学院博士課程在学中，及び（株）ATR知能映像通信研究所研修研究員．超高精細顔モデルの構築・仮想空間上でのコミュニケーションシステムに関する研究に従事．平12本会学術奨励賞受賞．

Kim Binsted

is CEO of I-Chara Inc., a Tokyo-based mobile agent company (www.i-chara.com). Formerly, she was a researcher at the Sony Computer Science Laboratories, working on Human Computer Interaction and Artificial Intelligence (AI). She received her PhD in AI at the University of Edinburgh, and her BSc in Physics at McGill University, Montreal.



Frank Nielsen

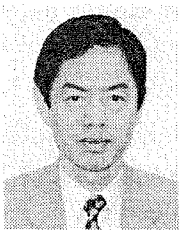
received the B.S. and M.S. degrees from École Normale Supérieure (ENS) of Lyon in 1992 and 1994, respectively. He defended his Ph. D. thesis on “Adaptive Computational Geometry” prepared at INRIA Sophia-Antipolis under the supervision of Pr. Boissonnat in 1996. As a civil servant of the University of Nice (France), he gave lectures at the engineering schools ESSI and ISIA (École des Mines). In 1997, he served army as a scientific member in the computer science laboratory of École Polytechnique (LIX). In 1998, he joined Sony Computer Science Laboratories, Tokyo (Japan) as an associate researcher. His current research interests include computational geometry, algorithmic vision, combinatorial optimization for geometric scenes and compression.

Claudio Pinhanez

is a computer scientist and a media artist. He has been a researcher at IBM TJ Watson Research Center since 1999, and currently is part of the Pervasive Computing Group, working in the design and development of interactive spaces and on physical interfaces to information. Claudio got his PhD. from the MIT Media Laboratory in 1999 with Prof. Aaron Bobick, working on the design and construction of physically interactive environments. In particular, he investigated new paradigms for computational representation of human action and the problem of scripting stories in interactive environments. During his PhD. he created and produced innovative theatrical experiences involving computers interacting with human actors on stage, including the computer theater plays "SingSong" and "It/I." Claudio was a visiting researcher at ATR-MIC laboratory (Kyoto, Japan) in 1996 and at Sony Computer Science Laboratory (Tokyo) in 1998.

鉄谷　信二　（正員）

昭 55 北大工学部大学院修士課程了．同年電電公社（現 NTT）入社以来，ファクシミリにおける画像信号処理，電子写真記録，立体表示技術等の研究実用化に従事．平 3 ATR 通信システム研究所に出向，臨場感表示技術に従事．平 6 NTT に復帰，高速ネットワーク用アプリケーション開発に従事．平 12 ATR 知能映像通信研究所に出向，コミュニケーション環境生成に関する研究に従事．現在，同研究所第 1 研究室長．工博．



中津　良平　（正員）

昭 44 京大・工・電子卒，昭 46 同大大学院修士課程了．同年日本電信電話公社（現 NTT）武蔵野電気通信研究所入所．昭 55 横須賀電気通信研究所．主として音声認識の基礎研究，応用研究に従事．平 2 NTT 基礎研究所研究企画部長，平 3 NTT 基礎研究所情報科学研究部長．平 6 より ATR に移り，現在（株）ATR 知能映像通信研究所代表取締役社長．マルチメディア要素技術の研究及びマルチメディア技術を応用した通信方式の研究などに従事．工博（京大）．昭 53 年度本会学術奨励賞，平 8 IEEE Multimedia Systems and Computing '96 最優秀論文賞，平 9 ロレアル賞，平 11 映像情報メディア学会論文賞，平 11・12 テレコムシステム技術賞，平 11・12 日本バーチャルリアリティ学会論文賞，平 12 人工知能学会論文賞，平 13 文部科学大臣賞各受賞．平 13 IEEE フェロー．日本音響学会，情報処理学会，人工知能学会，画像電子学会，日本バーチャルリアリティ学会，映像情報メディア学会，日本芸術科学会，日本情報考古学会各会員．

森島　繁生　（正員）

昭 57 東大・工卒．昭 59 同大大学院修士課程，昭 63 同大学院博士課程了．工博．平 13 成蹊大学工学部教授，現在に至る．平 6 から 1 年間，トロント大学客員研究員．平 8 から通信放送機構 3 次元空間共有プロジェクトサブリーダー．明治大学非常勤講師．(株) ATR 音声言語通信研究所客員研究員を併任．本会論文誌編集委員．グラフィックス，ビジョン，マルチモーダルインタフェース等の研究に従事．平 4 本会業績賞受賞．

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 －

緒方　信†‡　　中村　哲‡　　森島　繁生†
†成蹊大学工学研究科　　‡ ATR 音声言語通信研究所

あらまし　本稿では、従来より研究されてきた音声翻訳技術に加え画像をも翻訳する、日英双方向翻訳システムを紹介する。本手法は顔画像翻訳において、話者の表情を保つ為に口やその周囲の情報以外は原言語発話時の動画像をそのまま用い、口領域については任意の話者に適合可能な3次元ワイヤフレームモデルを用意し双方を合成することを試みた。この手法により、小規模なデータベースより顔画像の合成、翻訳が可能となった。

Multi-modal Translation System
- Model Based Lip Synchronization with Automatically Translated Synthetic Voice -

Shin OGATA†‡　　Satoshi NAKAMURA‡　　Shigeo MORISHIMA†
† Faculity of Engineering, SEIKEI Univ.　　‡ ATR Spoken Language Translation Research Lab.

Abstract　In this paper, we introduce the multi-modal English to Japanese and Japanese to English translation system, which translates the speaking motion synchronized to the translated speech. To retain the speaker's facial expression, we substitute only speech organ's image with the synthesized one, which is made by a three-dimensional wire frame model that is adaptable to any speaker. Our approach enables the image synthesis and translation with extremely small database.

1．はじめに

音声翻訳の研究は、あらゆる言語間で、またさまざまな目的に応じて盛んに行われており、その発展は目覚しいものがある。

1993年に発足したATR音声翻訳通信研究所における研究の結果、話題の対象が限定されるなどの一定の条件下で、異なる言語での対話を翻訳するシステム(ATR-MATRIX[1])として利用可能であるという段階に達している。そしてこの分野の研究は、2000年に発足したATR音声言語通信研究所においても引き継がれ、より広いドメインにおける日常の自然な話し言葉に拡張されることが期待されている。

音声翻訳技術は韻律情報等を除けば、主に言語情報を扱う研究分野として発展してきた。しかし言語情報は、意思疎通の為のひとつの手段に過ぎず、Face-to-Faceのコミュニケーションにおいて、顔は言葉と共にさまざまなメッセージを伝えている。例えば映画などにおける吹き替えでは、音声のみを翻訳している為、口の動きと発話内容が一致しないという課題がある。また顔画像全体をコンピュータグラフィックスにより合成した場合、ノンバーバルな情報を再現して伝えることが困難となる。これらの課題を克服し音声翻訳と共に画像の翻訳、つまり話者の表情を保ちつつ口形状の翻訳が可能となれば、より親しみのあるコミュニケーションを実現できるであろう。

口形状を変形した顔画像を生成する研究[2]は、過去にもアプローチはいくつかあった。しかし、画像は音声に比べ情報量が多い為、大規模データベースを用意するのは困難であり、話者が限定される等の汎用性が乏しいという制約があった。

そこで本研究では、話者の表情を保つ為に、口やその周囲の情報以外は原言語発話時の顔動画像をそのまま用い、口領域については任意の話者に適合できる3次元モデルを用意し、双方を合成することを試みた。3次元モデルを用いれば、音声合成に用いた音素の表記と継続長情報を基に口形を生成することができ、顔の位置や向きにも対応する合成画像を得ることが可能である。また3次元モデルには画像データベースは必要がなく、用意するのは口形状を表現する為のワイヤフレーム格子点のベクトル移動量のみである。これにより、比較的小規模のデータベースによって音声のみならず顔画像をも翻訳できる、ビデオ翻訳システムが可能となる。

本稿ではまず、このビデオ翻訳システムの全体像について触れ、顔画像合成における3次元口形モデルの生成について記述する。次に音声合成部より

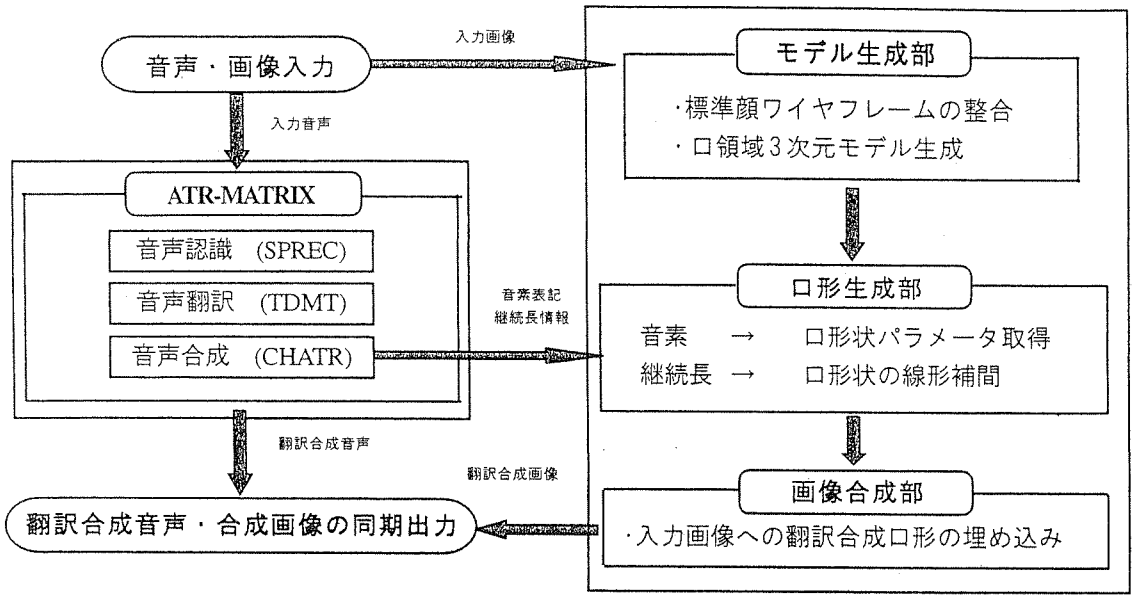


図1 システム全体像

得られる音素表記と継続長情報の2つのパラメータから、発話に対応する口形をモデル上に生成する手法について説明する。その後モデルと入力画像の合成手法について触れ、そして最後に、このシステムにおける研究課題について考察する。

2．システム全体像

図1に、本研究におけるシステムの全体像を示す。システムは大別すると音声翻訳部と画像翻訳部に分かれている。音声翻訳部はATR音声翻訳通信研究所で開発されたATR-MATRIX[1]により行われる。ATR-MATRIXは、音声認識を行うSPREC、テキストに変換された言語情報を翻訳するTDMT、翻訳されたテキストから合成音声を生成するCHATR[3]より構成されている。このうちの翻訳合成音声を生成するCHATRより出力される音素表記と音素継続長の情報は、顔画像翻訳に利用される。

画像翻訳部における第一段階は、入力画像から標準顔ワイヤフレームを整合することにより、話者別の口領域の3次元モデルを生成することである。話者により顔面の骨格が異なる為に、個人ごとにモデルを生成しておかなければならないが、この工程は話者1人につき1度踏まえればよい。

画像翻訳部第二段階は、発話に対応する口形生成部である。音声合成に用いた音素表記から、各音素に対応する口形状パラメータをデータベースより取得し、口領域モデルを変形させる。また音素継続長情報は口形状の線形補間に利用する。このときに使用する口形状パラメータは音素ごとに定めた口形状のワイヤフレームのベクトル移動量としている為、話者に依存することはない。

画像翻訳部の最終段階は、入力画像に3次元口形モデルを埋め込む画像合成部である。この工程でモデルと入力画像の色、スケールを一致させる。入力した顔画像が発話時に運動していても、モデルは3次元情報を所持している為、自然な画像合成を行うことが可能である。

システムの最終工程では、翻訳された合成音声と合成画像を30[frame/sec]で同期させて出力する。

3．口領域の3次元モデルの生成[5][6]

3-1. 3次元頭部モデル

人間の顔は基本的な形状や構造は同じといってよいが、目、鼻、口等の要素を構成する形状や位置

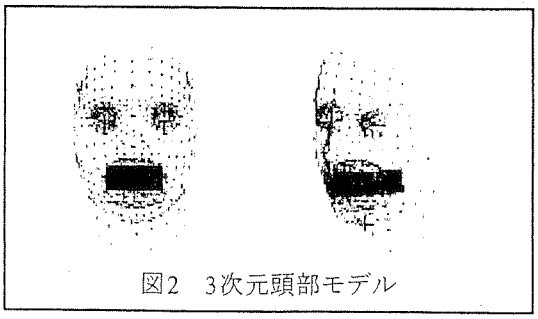
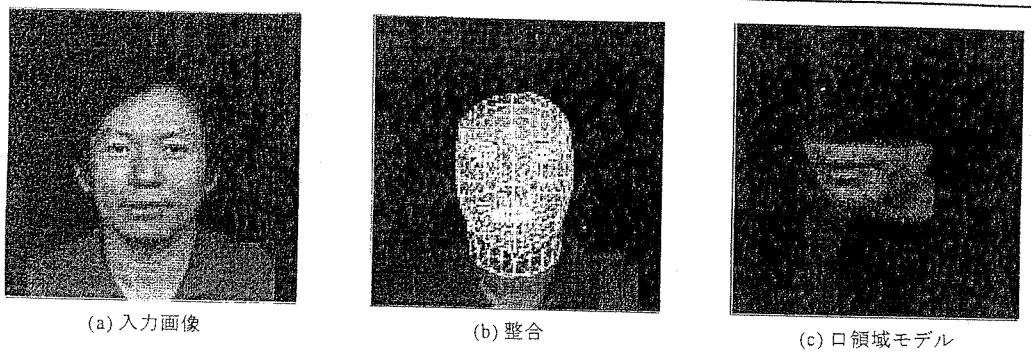


図2 3次元頭部モデル



(a) 入力画像　(b) 整合　(c) 口領域モデル

図3　生成された口領域3次元モデル

は、個人によって微妙に異なる。CG(Computer Graphics)によって自然な表情を合成するには、対象人物の顔により忠実でかつ演算量の少ない3次元モデルを構成することが必要となる。

本研究では、成蹊大学情報通信研究室において研究・開発されている、3次元頭部モデル[図2]を用いて、口領域の3次元モデルを生成することを試みた。本研究では顔領域全体の整合を行った。

このモデルは約1500ポリゴンの三角形パッチより構成されていて、格子点数は約800からなる。

3-2. 口領域モデルの作成過程

3-1節で導入した3次元頭部モデル上に、人物画像の口領域を正確にテクスチャマッピングする為には、ワイヤフレームモデルと対象人物の入力画像の整合を行わなければならない。整合は、任意方向から撮影した複数画像と、専用のGUIツールを用いて行い、この工程を経てモデルに3次元情報を付加することが可能である。現状では、整合に多くの画像を用いるほど精密なモデルを生成することができるが、人手による作業量も増すことになる。

本研究においては、話者1人につき、正面・側面に加えさらに2つの斜め方向より、計4方向から撮影した画像を用いて、口領域モデルの整合を行った。図3(c)が、入力画像(a)の人物の生成された3次元口領域モデルを表している。

4. 発話口形の生成[4][6]

人間が会話をする際、動作の大きい部分として、唇、顎などが挙げられる。特に唇の動きは音韻と密接な関係がある為、正確な制御が要求される。

本研究と同様、発話顔動画像の生成をモデルベースからアプローチしている、文献[4]の報告では、被験者の口領域にマーカを置き、3次元的移動量を計測することで運動学的データを採取し、アニメーションに再現する方法を試みている。しかし今回は、話者への依存性を削減するという点から、次節に示す手法を用いた。

4-1. 標準口形状データの設定[6]

口領域の動きを定量的に表現する為に、文献[6]では口領域の制御点として図4のように7点を定めている。各々の制御点はワイヤフレームモデルの格子点と対応しており、骨格と筋肉の運動に基づく、3次元の移動法則が定められている。

本研究では、表現する音韻を表している正面と側面の2方向から撮影された、口形状の参照画像を用意し、上記の制御点を移動することでワイヤフレームモデルをその参照画像に近づけるよう変形させたとき得られる、各格子点のベクトル移動量を基本口形のデータベースとしている。このデータは口領域の大きさで正規化したベクトル移動量としている為、一度基本口形を用意すれば、すべての話者に適用することが可能である。このように、本研究では話者に依存しない小規模なデータベースしか必要としないシステムを実現している。

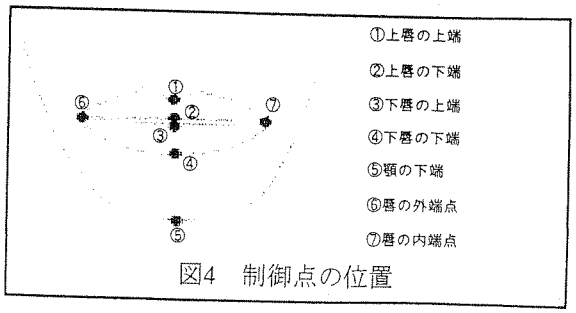


図4　制御点の位置

4-2.VISEMEによる音韻分類

4-2-1.VISEMEの定義

VISEMEとは、音素である “phoneme” から作られた造語である。音声学的に異なった音であっても同一言語の中で同一音とみなされる最小の音単位の意である。例として、英語における “me” と “knee” という単語の発音は、騒音下などでは音のみよって区別するのは非常に困難である為、同一音とする場合があるが、同一音であっても視覚的要素によってどちらの単語であるか区別するのは容易である。もし話者の口が閉じていればそれは “me” であればあり、そうでなければ “knee” である。またそれとは逆に、“bat” と “pat” のような単語にみられる、視覚では区別できず聴覚によって区別が可能な /b/, /p/ 等の音韻を「視覚素」、すなわちVISEMEと呼ぶ。

本研究では、音声合成部CHATRより出力される音素表記をVISEMEに基づき、英語については22種類に、日本語については5母音をそれとは別に分類し、さらに無音区間を加えた計28種類の基本口形をデータベースとして用いた。

表1　VISEMEの分類

VISEME No.	CHATRより返還される音素表記
1	/ae/　　英語
2	/ah/, /ax/
3	/A/
4	/aa/
5	/er/, /ah r/
6	/iy/, /ih/
7	/uh/
8	/uw/
9	/eh/
10	/oh/, /ao/
11	/ax r/
12	/l/
13	/r/
14	/b/, /p/, /m/
15	/t/
16	/d/, /n/
17	/k/, /g/, /hh/, /ng/
18	/f/, /v/
19	/s/, /z/, /sh/, /zh/, /ts/, /dz/, /ch/, /jh/
20	/th/, /dh/
21	/y/
22	/w/
23	/a/, /A/　　日本語
24	/i/, /I/
25	/u/
26	/e/, /E/
27	/o/, /O/
28	/#/　　無音

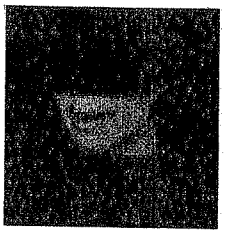


図5　音素表記/ah/に対応する基本口形

本来、VISEMEは発音記号[au],[ei]等に現れる口唇運動の情報まで定義されるのであるが、本研究においては、そのようなVISEMEは複合VISEMEとしてさらに分解し、運動情報は所持しない形状の情報のみで分類した。複合VISEMEについては後述する。

表1に本研究で分類したVISEMEとCHATRの音素表記の対応を、図5に3次元モデルで表現される音素表記/ah/の基本口形を、例として示す。

CHATRでは英語合成音声に、英国英語（British English）と米国英語（American English）を用意している。音声合成には、それぞれ別々の音素辞書を用いるが、本研究ではBritish Englishの音素辞書をVISEMEに対応付けた。また日本語の音素辞書には、外来語（カタカナ語）に多く用いられる「ヴェ」、「デュ」等の音素が存在するが、これらについては対応付けるには至らなかった。その他、共通にみられる「笑い声」や「いびき」等についてもCHATR側には表記として存在したが、本研究においては対応付けはしていない。

4-2-2. 英語における複合VISEME

英語には、音素表記1つに対して時間的に複数の基本口形から構成されるVISEMEが存在する。発音記号[au]や[ei]、[ou]等の音素がそれに当たる。本研究ではこのようなVISEMEに対して基本口形単位で分解し新たに分類し、複合VISEMEと呼ぶことにする。表2はCHATRより出力される、これらの複合VISEMEで表される音素表記を示す。

音素表記は個別に音素継続長の情報を持っている。しかしこのように音素表記が複合VISEMEと対

表2　英語における複合VISEME

音素表記	VISEME No.
/aa r/	4+2
/ia/	6+5
/ia r/	6+5
/ua r/	8+11
/ea r/	9+11
/aw/	4+8
/ey/	9+6
/oy/	5+6
/ow/	5+8
/ao r/	5+2
/ay/	4+6

応する場合はその継続長情報についても基本口形の個数によって分解する必要がある。

本研究では、音素表記が2つの基本口形から構成される場合において、前半に現れる基本口形に30%の音素継続時間を、後半に残りの継続時間を経験的に割り当てた。これはあくまで経験的に定めた値であるが、音素別に特徴を定量化できた場合、データベースを容易に変更することが可能であり、この点は本手法の特徴であるといえる。

4-2-3. 日本語の子音分類

日本語の子音は英語に現れるものより少ない為、本研究では英語のデータベースより引用した。しかし一般的に日本語の子音口形は英語に比べ運動の変化が少ないことが知られている。そこで今回、データベースより引用する日本語の子音基本口形は、英語の基本口形の移動量の60%におけるものと定めた。

また日本語では、文末の母音が無声化することが多い。CHATRでは、母音「う」の音素表記に有声音 /u/ と無声音 /U/ がある。そこで本システムでは無声化した場合の唇の移動量を考慮にいれて、子音のときと同様に、/u/ の基本口形60%を /U/ の基本口形とした。

さらに日本語子音の特殊な例として、「は行」がある。発音記号[h]に表される子音は、主に口内で生成される音である為、唇の動作等の視覚要素に反映されることは少ない。これを考慮に入れ、システムにおける音素表記 /h/ に対しては、後に続く母音基本口形を割り当てた。

4-3. 口形状の補間

システムの口形状データベースには28種類の基本口形があることは前節までに触れた。しかしある基本口形から次の基本口形に移行するまでのデータは存在しない。

本節では、音声合成部より出力されるもう1つのパラメータである音素継続長情報より、基本口形間の線形補間を行う手法を述べる。

4-3-1. 口唇運動の軌跡

人間が言葉を話すとき、唇は絶えず運動をする。しかし同じ発話内容であっても、1音ずつ音節を区切りながら発音するときと、文章として発話するときとでは口唇運動の軌跡が異なる。これは人間が滑らかに発話する場合、口唇の運動軌跡は効率良く推移していく傾向がある為である。

本システムに使用する口形状データベースは、基本口形に変形させたワイヤフレームの格子点のベクトル移動量で定義されている。そこで、ある基本口形の移動量と次に現れる基本口形のベクトル移動量を加減算することにより連続的な口唇運動を実現する手法について次節で説明する。

4-3-2. 口形状の線形補間

発声された音素が継続している間は基本口形の要素を持ったベクトル移動量の情報がモデルワイヤフレーム上に存在しなければならない。本研究において、音素が発声される開始時は、基本口形状を構成しているものと定義した。従って図6に示すように、音素継続時間の始点における、基本口形状を構

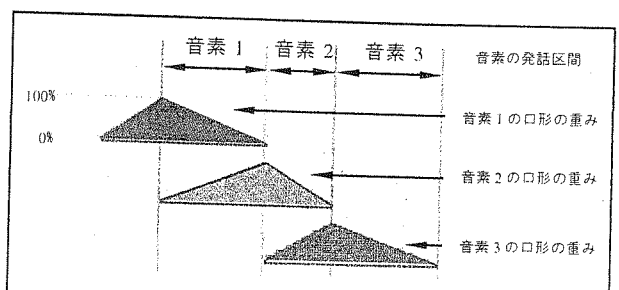


図6　口形状の線形補完に用いた重み付け

成する格子点のベクトル移動量を100%とするとき、音素継続時間の終点では0%になるように線形補間を行った。同様に、現時点で扱っている音素の次に現れる音素についても、現音素の継続時間長を基に、格子点のベクトル移動量を0%から100%に線形補間する。こうして得られる時系列上の2つのベクトル移動量を加算した値が基本口形間におけるワイヤフレームを変形する為のベクトル移動量となる。すなわちデータベースに存在しない口形状も算出することが可能となる。本手法は人間の発話時における、骨格や筋肉の運動に直接的には結びついてはいないが、口唇運動を近似的に再現できるものと考える。

4-3-3. 合成音声との同期

システムの最終工程では合成音声と合成画像を同期させなければならない為、口領域モデルもそれに合わせて生成する必要がある。

音声合成部CHATRで扱う最小時間長は、1[msec]である。そこで本研究では、前節で説明した線形補間の手法を用いて、事前にベクトル移動量の百分率を1[msec]単位で算出した。最終的に30[frame/sec]で動画を生成する事を考慮し、33[msec]毎にそのベクトル移動量を改めてサンプリングする。演算量を少なくする為に、口領域モデルの生成に必要なデータのみを操作するのである。

実際には、100[msec]毎に1[msec]の割合でこのサンプリングされたベクトル移動量の補正を行うことにより、時間的誤差の減少を図っている。

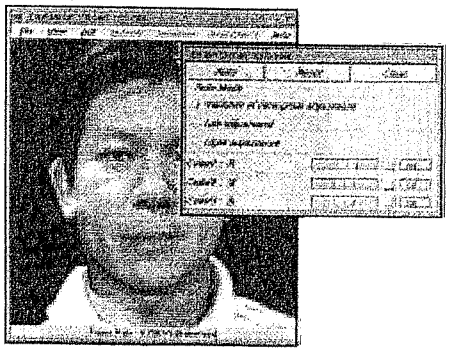


図7　画像合成部のGUIツール

5. 出力合成画像の生成

本システムの画像合成部には入力画像に対して口領域モデルのスケールを一致させる工程と入力画像とモデルの色調補正を行う工程、そして入力動画像にモデルを追跡し自然な合成画像を得る工程があるが、これらの工程は現時点では全て人の手による補正が必要となり、完全な自動化には至っていない。

そこで本研究では、この画像合成の工程に必要な一連の機能を備えたGUIを開発し[図7]、このツール上で合成画像の生成を行った。ツールにはCHATRより出力されるパラメータファイルを読み込む機能が実装されており、口形状は自動で生成することが可能である。

(a) 補正前

(b) 補正後

図8　色補正を行った例

(a) 入力画像

(b) 翻訳合成画像

図9　入力画像と翻訳合成画像

5-1. 口領域モデルの色補正

翻訳された合成口形を入力画像に埋め込む際、話者別に生成したモデルは、入力画像の撮影条件によっては色補正を必要とする。より自然な画像に見せる為、色補正の後、口領域モデルと入力画像の境界は、色調の透過率を徐々に変化させた[図8]。

色補正の手法に、モデルと入力画像の色ヒストグラムの平均値の調整と、L*a*b*(注1)を用いた光源情報の補正を用いたが、今回は完全な自動化には至っていない。今後は補正手法についても検討の必要があるものと考える。

5-2. 入力画像の間引き・繰り返し

入力画像は入力音声の継続時間長分の情報を所持している。しかし、言語を翻訳することにより、音声の時間長は変化する。

本研究では、入力動画像を連続する静止画像の時系列、画像シーケンスとみなし、翻訳された合成音全体の継続時間長から、合成時に使用する入力画像のシーケンス数を操作する手法をとった。合成音声が入力音声に対して短くなる場合には、一定の割合で画像シーケンスを間引き、反対に合成音声が入力音声に比べて長くなる場合においては、時系列に沿って一定の割合で画像を繰り返し用いることで、合成音の継続長と合成画像の継続長を調整し、同期出力させた。

5-3. 翻訳合成口形の埋め込み

前節までの工程を経た口形状モデルは、3次元形状、発話口形、発話時間長、スケール、色の情報を所持している為、入力顔画像に対して自然な合成を得ることが可能である。図9に合成画像の1例を示す。

モデルは対象人物の鼻の下から喉仏までの情報を所持している。入力画像の口形はモデルによって覆い隠される為、これより入力画像の原国語の発話口形に依らず翻訳画像を生成することができる。

また、モデルに覆い隠されない顔の部位に、ノンバーバル情報が現れている場合、その情報を維持することが可能であるのも、本手法の特徴である。

4章で述べたように、発話時間に応じて口形の補間が行われ、その情報に基づき画像シーケンスが生成される。画像シーケンスを30[frame/sec]でCHATRの生成する合成音声と同期させ出力することにより、翻訳合成動画像を得る。

以上の処理を用いて、日本語から英語の翻訳、反対に英語から日本語の双方向の翻訳が可能なシステムを構築した。

6. まとめ

本研究の成果として挙げられるのは、小規模な口形データベースから話者に依存することなくさまざまな発話口形画像を生成できることと、口領域以

(注1) CIEL*a*b* 色空間を使用。

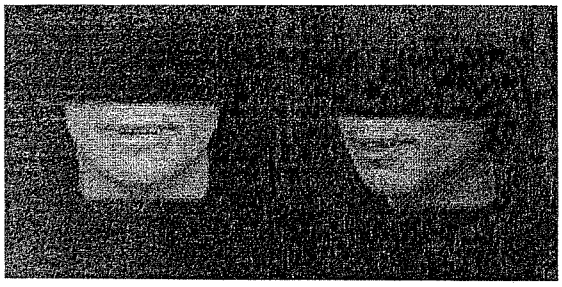


図10　現状における/th/の口形

図11　入力画像に対してモデルが適合しない例

外は入力動画像を用いることで、ノンバーバルな情報を維持できたこと、それによって翻訳合成音声を生成する際に出力されるパラメータから、日英双方向ビデオ翻訳システムを実現できたことである。

さらに、今回の研究において確認できた今後の研究課題について述べる。

まず、口形状モデルについては現在、舌のモデル化が行われていない為、英語の[th]の発音等は不完全である[図10]。舌モデルも唇と同様にパラメータ制御が可能なモデルを作成することが必要と考える。また歯のモデルについては、スケール変換はできるが、未だ人工的な印象を与える。これについては、ライティングの設定或いはテクスチャマッピングを施すことで改善できるものと考える。

今回、入力画像に対するモデルのトラッキングは全て手作業で行った。自動トラッキングの分野の研究は盛んである為、システムに取り入れる余地はあると考える。

そしてまた、今回のシステムは全てオフラインで行ったに過ぎない。このシステムのリアルタイム化の実現にはまず第1に、画像処理の高速化が必要である。その1つの方法として提案するのが、入力動画像のキャプチャと口領域モデルの完全分離処理である。また、オンラインでシステムを稼動させる為には、今回のように入力画像の継続時間を調節するのではなく、音声合成側の継続時間を操作する手法を確立することが必要であると考える。

また発声方法や発話様式、話者の感情も考慮したシステムを構築する必要がある。図11に示すのは、入力画像において話者が笑顔であるとき、合成画像に笑顔を再現できていない状況である。これについては、感情によってワイヤフレームの移動量を新に定義することを考えている。

本手法に用いた基本口形を構成するベクトル移動量を定義した値、線形補間法についてもまた、更なる改善の余地があるものと考える。

7. 参考文献

[1] 菅谷，竹澤，横尾，山本
「日英双方向音声翻訳システム(ATR-MATRIX)の対話実験」
日本音響学会1999年春季研究発表会講演論文集，pp 107-108, 1999

[2] Hans Peter Graf, Eric Cosatto, Tony Ezzat
"Face Analysis for the Synthesis of Photo-Realistic Talking Heads"
PROCEEDINGS FOURTH INTERNATIONAL CONFERENCE ON AUTOMATIC FACE AND GESTURE RECOGNITION 28-30 MARCH,2000.,GRENOBLE,FRANCE pp189-194

[3] Nick Campbell, Alan W. Black
"Chatr : a multi-lingual speech re-sequencing synthesis system"
電子情報通信学会技術研究報告，sp96-7, pp.45, 1995

[4] T. Kuratate, H. Yehia, E. Vatikiotis-Bateson
"KINEMATICS-BASED SYNTHESIS OF REALISTIC TRACKING FACE"
International Conference on Auditory-Visual Speech Processing - AVSP'98, pp.185-190, 1998

[5] 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」
電子情報通信学会技術研究報告，Vol99, No582, pp7-12, 2000

[6] 伊藤，三澤，武藤，森島
「仮想空間上におけるリアルな三次元口形状の作成」
電子情報通信学会総合大会，A-16-24, pp328, 2000

2．4　表情合成アルゴリズムの提案

表情筋モデル研究の知見に基づき、新たな表情合成アルゴリズムに関して、検討を行った。

Face and Gesture Cloning for Life-like Agent

Shigeo Morishima

Waseda University
3-4-1 Okubo Shinjuku-ku, Tokyo 169-8555, Japan
shigeo@waseda.jp

Abstract

Face and gesture cloning is essential to make a life-like agent more believable and to give it a personality and a character of target person. To realize cloning, an accurate face capture and motion capture are inevitable to get corpus data about face expressions, speaking scenes and gestures. In this paper, our recent approach to capture the personal feature of face and gesture is presented.

For the face capturing, a face location and angles are estimated from video sequence with personal 3D face model and then a synthetic face model data is imposed into frames to realize automatic stand-in system or multimodal translation system..

A stand-in is a common technique for movies and TV programs in foreign languages. The current stand-in that only substitutes the voice channel results awkward matching to the mouth motion. Videophone with automatic voice translation are expected to be widely used in the near future, which may face the same problem without lip-synchronized speaking face image translation. In this paper, we introduce a method to track motion of the face from the video image and then replace the face part or only mouth part with synthesized one which is synchronized with synthetic voice or spoken voice. This is one of the key technologies not only for speaking image translation and communication system, but also for an interactive entertainment system. Also, an interactive movie system is introduced as an application of entertainment system.

Capturing and copying a facial expression based on a physics base facial muscle constraint has been already presented[6]. So in this paper, this part is not described.

For a gesture capturing, commercially available motion capture products give us fairly precise movements of human body segments but do not measure enough information to define skeletal posture in its entirety. This paper describes how to obtain the complete posture of skeletal structure with the help of marker locations relative to bones that are derived from MRI data sets.

1 Introduction

Recently, CG technology is used in various media, such as TV programs, movies and amusements. The CG quality rapidly progresses in these several years, and has reached even at the same level as a photograph taken on the spot. Furthermore, the researches which generate and control human's facial images and body motion by CG also performed briskly, and their applications in various fields are expected.

In a human communication scene, it seems that verbal information is the most important and additionally, non-verbal information also plays an important role in natural communication. The facial expression is thought to send the most of the non-verbal information. From this viewpoints many researches have been studied such as face-to-face communication using 3D face model, called an avatar on cyberspace as one of the communication forms, and we proposed HYPERMASK as an interaction experiment with theatrical tool using a real object, on which various synthesized facial expressions are projected.

As for speech communication, a spoken language translation has been studied at ATR and a prototype system, ATR-MATRIX[2] is developed for the limited domain of hotel reservation between Japanese and English. On the other hand, the stand-in of a foreign language for overseas movies and TV programs, has been performed conventionally. But its speech sounds and mouth motions are always not synchronized.

A gesture also includes non-verbal information. Sometimes body motion includes personality, skill, emotion and etc.. However, these information cannot be captured easily. A motion capture system is widely used to get body motion for synthetic movie actor or game character. Current commercial motion capture system can only capture the markers' position put on a actor's body and then approximate bone data is estimated to generate a character motion. So huge handwork in post-process is inevitable to make a target motion.

The first part of this paper proposes a method to generate 3D personal face model with real personal face shape, and to track the face motion like movement and rotation automatically and accurately for audio-visual speech translation and interactive movie system. The method enables to detect translation and rotation of the head by template matching using a 3D personal face model.
The second part of this paper proposes a method to capture an accurate motion of bones and joints of human body with combination of motion capture system and MRI.

2 3D Personal Face Model

Face fitting tool developed by our group[3] is a tool to generate a 3D face model using one's photograph. But the manual fitting algorithm requires a lot of time for users to generate a good 3D model with real personal face, although it is able to generate an approximate model every time with real personal shape by fitting many photographs without any lack of facial vertices.

In order to raise accuracy of face tracking using the 3D face model, we used 3D color range scanner to generate a 3D model with a real personal shape. We show data acquired by Cyberware range scanner in Fig.1. This data is only a uniform vertices, so we cannot control facial expression because there is no meaning in every point. So at first, to fit a generic face model to the range data, both a generic model and range data are mapped to 2D plane. Then, we manually fit a generic model's face parts to corresponding parts in captured texture (Fig.2 left). Finally, we replace the depth coordinates values of the standard model with range data values and obtain a real 3D personal face model (Fig.2 right).

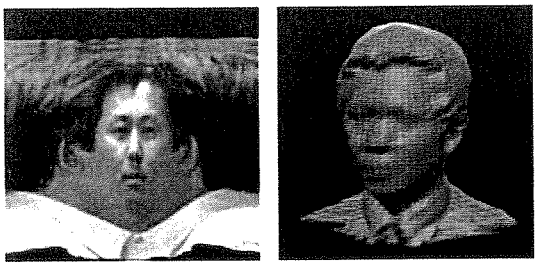


Figure 1. Captured Range Data and Texture

Figure 2. Fitting Result and Personal 3D Model

3 Tracking Face in Video Sequence

Many tracking algorithms have been studied by many researchers for a long time, and a lot of algorithms are applied to track a mouth contour, an eyes contour, and so on. However, because of blurring feature points between frames, or occlusion of the feature points by rotation of a head etc., these algorithms were not able to do accurate tracking enough to regenerate natural expression. In this chapter, we describe an automatic face tracking algorithm using a 3D face model. This is not a point tracking but a surface tracking by treating a part of face surface as a rigid object.

Tracking processing using template matching can be divided into the three steps. First, texture mapping of one of the video frame images is carried out to the 3D individual face shape model created in Chapter 2. Here, the frontal face image is chosen out of video frame images for the texture mapping. The reason is that the video images used in this experiment are mainly frontal in movement and the lighting condition does not change so drastically in the same video sequence. Anyway, we have to fit a generic face model to one frame shot from video sequence by manual operation.

Next, we make a 2D template images for every translation and rotation using a 3D model (Fig.3). Here, in order to reduce a matching error, a mouth region is excluded in a template image. Thereby, even while the person in a video image is speaking something, tracking can be carried out more stably.
Finally, we carry out template matching between the template images and an input video frame image and estimate translation and rotation values so that a matching error becomes minimum using mean square error of RGB values. Fig.4 shows an example of original video sequence. After face tracking, face part is extracted from original sequence in Fig.5. By replacing with synthesized mouth shape, new image sequence is coming out in Fig.6.

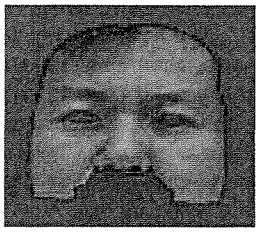


Figure 3
3D Face Template

Figure 4 Original Video Sequence

Figure 5 Extracted Face Location

Figure 6 Synthesized Image after Replacing Mouth Part

4 Video Translation System

In this chapter, a video translation system is introduced. The system is divided broadly into two parts: one is a speech-translation part and a image-translation part. A speech-translation part is composed of ATR-MATRIX[2], which was developed in ATR-ITL. ATR-MATRIX is composed of ATR-SPREC to execute speech recognition, TDMT to handle text-to-text translation, and CHATR to generate synthesized speech. In the process of speech translation, the two parameters of phoneme notation and duration information, which are outputs from CHATR, are applied to facial image translation.

The first step of a image-translation part is to make a 3-D model of the mouth region for each speaker by fitting a standard facial wire-frame model to an input image. Because of the differences in facial bone structures, it is necessary to prepare a personal model for each speaker, but this process is required only once for each speaker. This process is already shown in Chapter 2.

The second step is to generate lip movements for the corresponding utterance. The 3-D model is transformed by controlling the acquired lip-shape parameters so that they correspond to the phoneme notations from the database used at the speech synthesis stage. Duration information is also applied and interpolated by some interpolation for smooth lip movement. Here, the lip-shape parameters for each Viseme are defined by a momentum vector derived from a neutral face at lattice points on a wire-frame. Therefore, this database does not need speaker adaptation.

In the final step of a image-translation part, the translated synthetic mouth region's 3-D model is embedded into input images. In this step, the 3-D model's color and scale are adjusted to the input images. Even if an input image sequence is moving during an utterance, we can acquire natural synthetic images because tracking result is accurate and a 3-D face model has personal geometry information. Consequently, the system outputs a lip-synchronized face image to the translated synthetic speech at 30 frames/sec. Fig.7 shows an example of translated image frame shot.

Figure 7 Original and Translated Image

5 Interactive Movie System

When people watch movies, they sometimes overlap their own figure with the actor's image. An interactive movie system we constructed is an image creating system in which the user can control facial expression and lip motion of his or her face image inserted into a movie scene. The user submits a voice sample by microphone and pushes keys that determine expression and special effect. His or her own video program can be generated in real-time.

At first, once a frontal face image of a visitor is captured by camera, a 3-D generic wireframe model is fitted onto the user's face image to generate a personal 3-D surface model. A facial expression is synthesized by controlling the

grid point of the face model and texture mapping. For speaker adaptation, the visitor has to speak five vowels to choose an optimum weight from the database.

In the interactive process, a famous movie scene is going on and the facial region of an actor or actress is replaced with visitor's face. Facial expression and lip shape are also controlled synchronously by the captured voice. Figure 11 shows the result of face model fitting into original movie scene. Figure 12 shows an user's face inserted into actor's face after color correction. Any expression can be appended and any scenario can be given by the user independent of the original story in this interactive movie system.

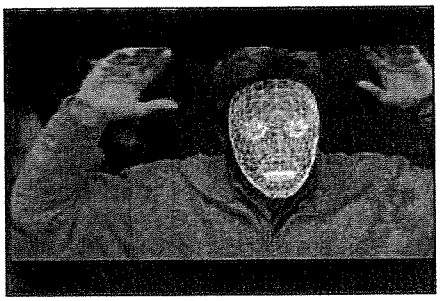


Figure 9 Fitted Model

Figure 10 Replaced Face

6 Skeleton Capturing System

Today, ordinary motion capture (mocap) products do not have the ability to measure the actual posture of an underlying skeleton. They only measure moment-to-moment displacements from the movements of markers attached to the skin surface or clothing. Each joint angle is calculated from these displacements based on simplified bone structure. In a typical post-processing of commercial mocap products, a user has to determine the correspondence between skeleton structure and marker position at an arbitrary initial posture.

A more anatomically correct way to determine skeletal posture is based on certain types of knowledge such as the relative locations of surface landmarks and the hip joint's center location. In the field of biomechanics, various estimation methods [4] were introduced based on the geometric relation of the bone and a landmark on the anatomical skin's surface. However, this landmark can only be observed by skilled professionals. The method is based on statistical analysis and depends on various factors such as race, age and gender. Furthermore, each individual is not necessarily different from the statistical results if the estimation method is employed and individual characteristics are completely ignored. In addition, an estimation method for Europeans is obviously not suitable for Asians.

As described above, the absolute posture of a skeleton cannot be determined through an ordinary mocap process. Such stringent determination is not so important in the entertainment industries because acceptable motions need not be scientifically realistic. On the other hand, capturing traditional motion like classical ballet requires highly detailed and precise posture for educational and archiving purposes.

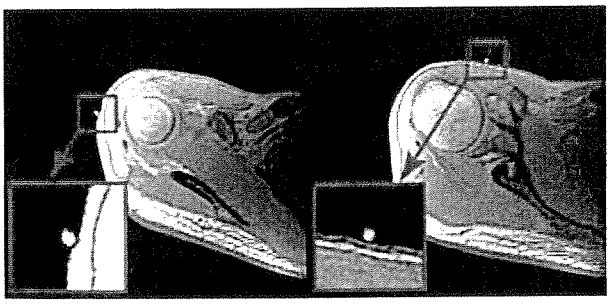


Figure 11
MR images of Markers on the shoulder

Human motion of ballet dance is our primary target. Ballet dance includes extreme postures like keeping a foot tip higher than eye height [5]. That is why we chose ballet dance as a primary target motion. Highly reflective ball-

like markers are attached to the dancer's skin surface. These markers are filled with oil for the reason described below.

We have to observe the interior of a human body and markers on the surface at the same time. Computer tomography (CT) seems an appropriate solution for such a requirement. It has the capability to acquire images of the whole body all at once. However, a CT scanner works with X-ray, which can cause serious health problems. Therefore, magnetic resonance imaging (MRI) might be considered as an alternative. To use MRI, we focus on small volumes such as the shoulder or hip joint so that the scan area is strictly limited compared to CT. Mocap markers are specially designed for this purpose. Oil inside a marker can be easily identified in a MRI image as shown in Figure 11.

Once MRI data sets of the target dancer are acquired, bone segments are then semi-automatically extracted. At specific bone segments with a few markers, the centroid of each marker is located by hand and represented relative to local coordinates on the attached bone segment. Describing a bone-to-marker transformation is most essential. We then perform a mocap operation on ballet dancing. After that, each marker is labeled, and its position is recovered in three-dimensional space by the conventional method. The 3D positions of markers and the bone-to-marker mapping description help to locate bones. Figure 12 illustrates recovered skeletal posture from a capture session of ballet.

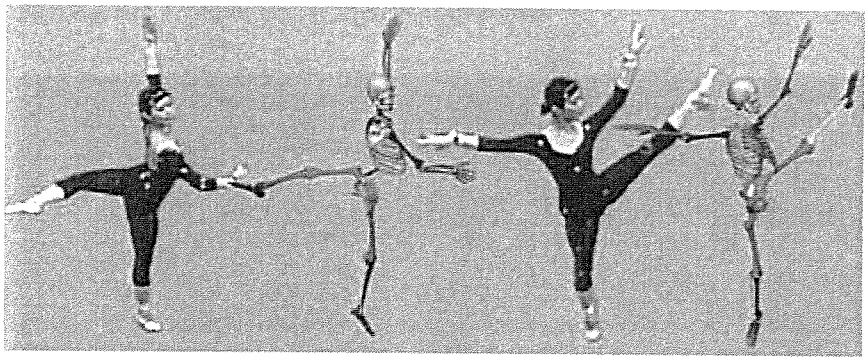


Figure 12
Recovered Skeleton

7 Estimation of Forearm Motion

Our goal is to describe or model the correspondences between actual skeletal postures and mocap markers attached to the skin surface. A mocap system can be used to generate an accurate posture for a specific joint of a human body.

Commercial mocap systems use the coordinates of mocap markers to clearly indicate the skeletal posture of a subject. In fact, the relative coordinates between mocap markers and skeletal structures are not perfectly fixed, so they are slightly articulated in general.

Ordinary mocap systems usually map marker data to a simple skeletal structure, which is far from accurate anatomically and individually. The ordinary mocap process does not consider skin artifacts, such as the skidding effect between the skin (a marker) and a bone. You will never get anatomically and individually correct skeletal postures, since mocap systems can only look at the surface of the subject.

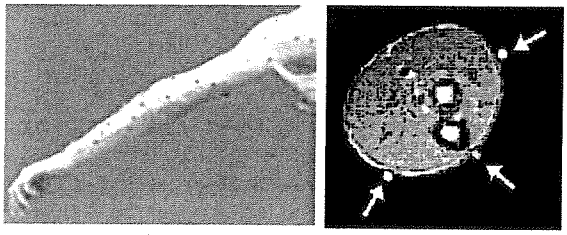


Figure 13
Markers on the arm(left) and MR image(right)

First, we need to extract the bones by observing the internal structure of a subject's body. A forearm has two stick-like bones, called long bones, and a slice MR image of them shows different features in the middle of the bone, a diaphysis, and on each end, epiphysises. An epiphysis holds much more water than other bony parts. This means

that a simple thresholding technique is sufficient for extraction. A diaphysis includes soft coarse tissues (bone marrow) which fill the central core, and a hard dense tissue (compact bone) which wraps the core. A core of diaphysis appears as a clear circle region on a single slice of an MR image. Otherwise, a part of the compact bone does not show sharply and is hard to distinguish(Fig.14). The center of the diaphysis core is located, then the edge that has almost the same distance from the center is detected. Then, the edges of the slices are combined and transformed into a 3D geometric representation(Fig.15).

Once the 3D bone geometries of the forearm are ready, you can visualize the geometric relation between the bones and the mocap markers, which are also extracted from MR images. We chose the poses under several conditions, such as bending or stretching the elbow joint with supination and pronation (outer and inner rotation of the elbow joint). The subject had to stay still for each specific pose during the MRI scan. Figure 16 reconfirms that there is some kind of tight constraint between the mocap markers and the bones, but the constraint does not seem to be a non-rigid type. Based on the observation results, we were able to make an approximate mapping model that describes the relative position between the mocap markers and bones in local space coordinates that are fixed to the bones. Only the mocap data (marker positions) of arbitrary joint movements on an elbow are needed to estimate an internal bone state if the mapping model has already been derived from the same subject.

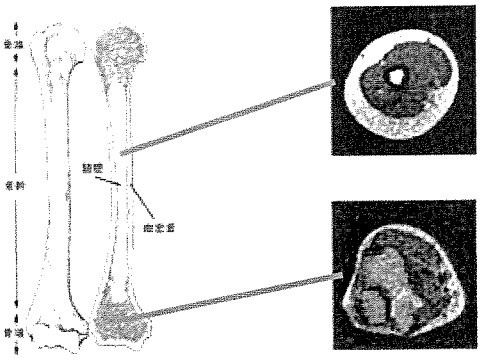


Figure 14 Extracting Bone Shape

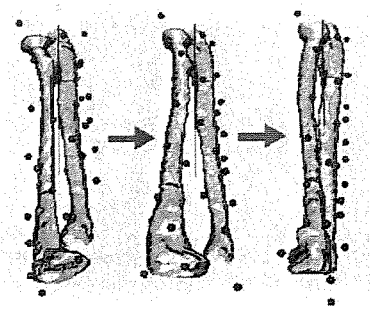


Figure 16 Mocap Markers and Forearm Bones

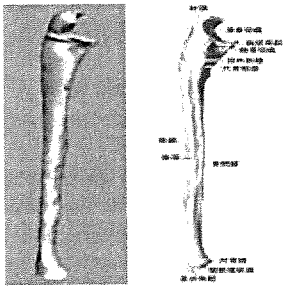


Figure 15 Extracted Skeleton and Reference

8 Conclusion

In a face capturing part, a location and angle are accurately estimated in video sequence. And then video translation system and interactive movie system can be realized as applications.

In a gesture capturing part, the proposed mocap process powered by medical imaging technology seems to be helpful for examining the actual internal state of a skeletal structure. A sort of skidding skin artifact over the skeletal structure can be observed that was not derived from a ordinary mocap process. We now want to study applying the proposed method to other parts of a skeleton.

References

[1] Shigeo Morishima, **"Face Analysis and Synthesis"**,
IEEE Signal Processing Magazine, Vol.18, No.3, pp.26-34, May 2001.
[2] T. Takezawa, et.al., "**Japanese-to-English speech translation system:ATR-MATRIX**", Proc. of ICSLP, pp. 957-960, 1998.
[3] **Facial Image Processing System for Human-like "Kansei" Agent** web site :
http://www.tokyo.image-lab.or.jp/aa/ipa/
[4] P. De. Leva, "**Joint Center Longitudinal Positions Computed from a Selected Subset of Chandler's Data**". Journal of Biomechanics, 29, 1231-1233,1996.
[5] C. Sparger, "**Anatomy and Ballet**"; 5th Edition. Theatre Arts Books, 1972.
[6] T. Ishikawa, S. Morishima and D.Terzopoulos, "**3D Face Expression Estimation and Generation from 2D Image Based on a Physically Constraint Model**", IEICE Transactions on Information and Systems, Vol.E83-D, No.2, pp.251-258, February 2000.
[7] S. Iwasawa, K. Kojima, K. Mase, S. Morishima, "**How to Capture Absolute Human Skeletal Posture**", Sketches and Applications, ACM SIGGRAPH 2003.

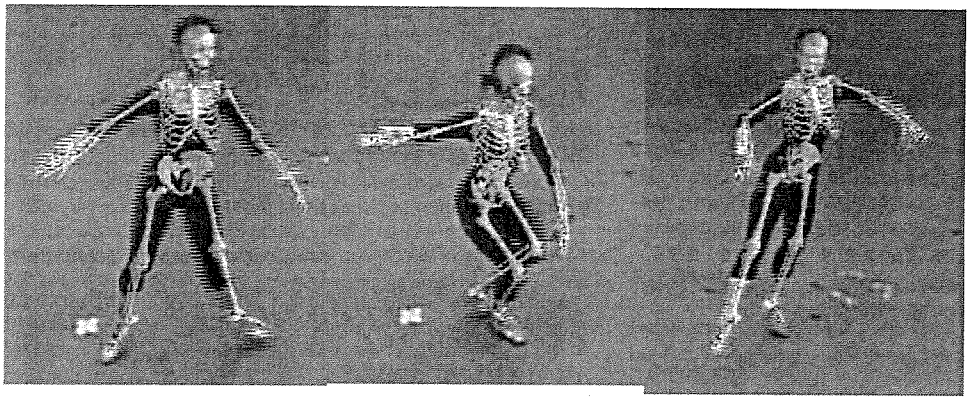


Figure 17 Estimated Skeleton of Ballet Dancer

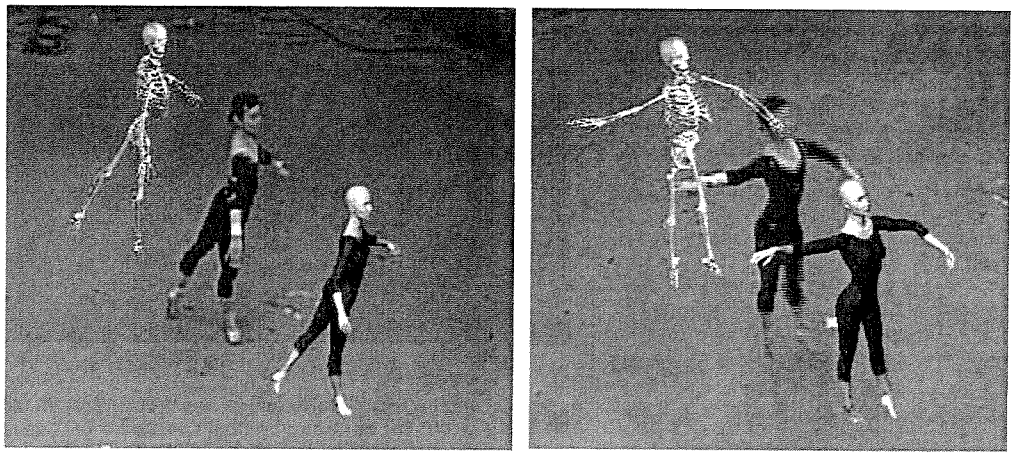


Figure 18 Estimated Skeleton and Generated/Rendered Body

フェイスキャプチャによる顔表情合成及び顔表情の定量表現

Expression Synthesis and Quantitative Representation of Face Expression Using Motion Capture System

○柳澤博昭[1]　前島謙宣[1]　四倉達夫[2]　森島繁生[1]

（1　早稲田大学理工学研究科・2　ATR 音声言語コミュニケーション研究所）

E-mail: h-yanagisawa@moegi.waseda.jp

1　はじめに

本研究では、モーションキャプチャシステムを用いて計測された表情表出時における顔表面の遷移データを用いて表情合成する手法、及び遷移データに対して主成分分析を行うことで得られる直交基底を用いて表情合成を行う手法を提案する。

表情表出には、Facial Action Coding System[1](以後、FACS)を使用した。FACS は、顔面筋の位置および動きの方向を解剖学的に考慮し、顔の表情を 44 種類の Action Unit(以後、AU と呼ぶ。)と呼ばれる基本動作に分類され、AU の組み合わせにより表情を記述できる。

本研究では、まず最初に、AU による表情をモーションションキャプチャシステムにより撮影することで、顔表面の動きの定量化を行った。次に、撮影された顔表面の遷移データを用いて表情合成、及びその全データに対して、段階的に Principal Component Analysis (PCA)を行い、時系列も考慮に入れた表情変化パラメータ(以後、固有 AU と呼ぶ)を得た。最終的に、固有 AU の線形結合により、表情合成を行った。

2　フェイスキャプチャ

本研究では、時系列に対する表情変化の遷移を詳細に計測するため、高い時間解像度を持つ Vicon 社製の光学式モーションキャプチャシステムを使用した。被験者は、AU 表出可能な男性 1 名である。被験者の顔表面に、3mm マーカを 146 点配置した。特に表情表出時に動きが見られる、鼻、唇、頬、顎の部位に重点的に配置した。被験者は、無表情の状態から AU を表出し、また無表情に戻るという動作を 3 回繰り返す。これを、AU 単体 11 種類、AU 複合を 52 種類、計 63 種類についてその様子を撮影した。図 1 に実際にマーカを配置した様子を示す。

3　表情合成

表情合成は、モーションキャプチャマーカのフレーム毎の動きを、3 次元顔ワイヤフレームモデルの頂点に与えることで行う。このワイヤフレームモデルは、無表情の顔画像に対して、独自の顔ワイヤフレームモデルを整合することで得られる。モーションキャプチャから得られる点群データは、計 146 点の 3 次元座標で与えられるが、実際に表情合成を行う顔ワイヤフレームモデルの頂点は 759 点存在するため、マーカの移動量を直接 3 次元顔ワイヤフレームの全頂点に割り当てることは不可能である。このため、ワイヤフレーム頂点の移動に関しては、近接するマーカから移動量を推定し、移動量を反映する必要がある。本研究では、モーションキャプチャマーカの移動量を 3 次元ワイヤフレームモデルの全ての頂点に適用するための補間手法として、RBFT(Radial Basis Function Transform)を用いた。

4　固有 AU

時系列の動きを含んだ表情変化パラメータを得るために、撮影された全ての顔表面の遷移データに対して PCA を行う。結果として、遷移データに対する圧縮と直交化がなされ、AU 間の相関が排除された固有ベクトル(以後、Face ベクトル と呼ぶ。)が得られる。次に、各フレームの Face ベクトル係数を 1 つのベクトルとみなし再度 PCA を行う。結果、Face ベクトル係数に対応された固有ベクトル(以後、Motion ベクトル と呼ぶ。)が得られる。Face ベクトルと Motion ベクトルを表情変化のパラメータとし、固有 AU と呼ぶ。最終的に、固有 AU の線形合成により表情合成を行う。図 2 に固有 AU の例を示す。

5　結果

本研究では、モーションキャプチャシステムを用いて、AU 表出時の顔表面の遷移について測定し、AU を定量化した。また、計測された顔表情の遷移データを用いて表情合成する手法を提案した。さらに、固有 AU を用いて表情合成する手法を提案した。

本実験において撮影した AU は、AU 表出に熟練した男性 1 名のみであったため、表情変化パラメータに個人依存性があると考えらる。今後は、さらに撮影人数を増やし、平均化することでパラメータを汎用的にする予定である。また、固有 AU の一般化と固有 AU を用いた表情合成の検証を行う予定である。

図 1:　マーカ配置　　図 2:　固有 AU1

謝辞

本研究は、科学技術振興機構の CREST プロジェクト助成の支援による。記して謝意を表します。また、撮影に協力して頂いた日本大学山田寛教授に深く謝意を表します。

参考文献

[1] P.Ekman and W.Friesen. "*Facial Action Coding System.*" Consulting Psychologists Press.1977

Speech to Talking Heads System Based on Hidden Markov Models

Tatsuo Yotsukura*
ATR Spoken Language
Communication Research Laboratories

Shigeo Morishima**
Waseda University
ATR Spoken Language
Communication Research Laboratories

Satoshi Nakamura***
ATR Spoken Language
Communication Research Laboratories

1 Introduction

This paper describes a technique to create human-like talking head speech animation with levels of naturalness and realism by mapping from speech information to facial movement sequences. Speech animation techniques for human-like natural talking head systems have traditionally included both key-framing methods [Cohen and Massaro 1993]and physics-based methods[Waters 1987]. Recent machine learning methods provide a new technique for speech animation systems [Yamamoto 1998], allowing them to be trained from recorded data and then used to synthesize new motion. Hidden markov models (HMMs) are frequently used for these methods and have demonstrated their effectiveness for speech animation. However, those methods lack producing high-quality, natural facial speech animation. In particular key-framing and physics-based methods are difficult to animate in precise synchronization with input audio signals (LipSync). Machine learning methods have an advantage over LipSync because the training data have explicit phonetic information and speech animation can be synthesized considering the surrounding phoneme context. However, the training data for facial movements is not high quality nor does it have precise values. It follows that the creation of natural talking head animation has difficulties in estimating the mapping from speech information to facial movements using inaccurate training data.

This paper presents a novel speech animation method with triphone based HMMs using a high-precision audio-visual corpus as the HMMs' training data. The audio-visual corpus is built by a 3D optical motion capture system (VICON inc.) that captures over 100 3D measurement points on the face. The output is a series of facial motion vectors, suitable for driving many different kinds of speech animation applications with 3D human-like characters.

2 System Overview

Figure 1 schematically outlines the main steps of the talking head generation system from input speech signals. The system synthesizes facial speech animation that is synchronized with the input audio speech signals, and outputs precise LipSync animation. Step (1) is a collection of motion capture data (facial motion vectors) and an audio signal synchronous corpus. Step (2) parameterizes input audio speech signals to phonemes with their durations automatically using Viterbi alignments. Step (3) maps the input audio parameters to the facial motion vectors based on the audio HMMs. Finally, step (4) creates 3D character speech animation from the facial motion vectors.

3 Corpus Collection

The facial motion vectors and speech are recorded by the Vicon MX System. The number of motion vectors is 111 3D positions (333 vectors) that are recorded at 120 Hz along with the simultaneous recording of speech at 16 kHz. Since the influence of head motion on the face surface points is large, we propose a method that applies an affine transformation to compute a singular-value decomposition from eight corrective points and, only for facial movements, apply an affine matrix to non-corrective points. Speech data is segmented to phonemes with their duration using Viterbi alignments. In addition, their phonemes (43 phonemes) are converted into visemes (9 visemes) to reduce the number of phoneme classes.

4 HMM-Based Speech Animation

The HMM-based method used in this paper is based on the mapping of input audio parameters to facial motion vectors through audio triphone-based HMM states. The triphone model has advantages compared to a monophone model because the face configurations of those phonemes depend on the preceding and succeeding phonemes. All training data for this method can be analyzed into synchronous audio and facial motion parameter sequences. As the audio parameter sequences can be converted to HMM state sequences, the synchronous facial parameter sequences can be segmented per HMM state. Motion vectors estimated by HMMs are mapped to the measurement positions of motion capture data. We have proposed a speech animation method of mapping from motion vectors to 3D characters' face. This method be applied to optimum motion vectors to facial object nodes.

5 Summary

Facial animation using the proposed talking heads was created from input speech signals, as shown in the video demonstration. We have confirmed facial animations of various facial objects.

Acknowledgements This research was supported by the Japan Science and Technology Agency (JST), as part of the CREST Project.

References

COHEN, M. M., MASSARO, D. W. 1993. *Modeling coarticulation in synthetic visual speech.* Models and Techniques in Computer Animation, pp.139-156.

YAMAMOTO, E., NAKAMURA, S., SHIKANO, K. 1998. *Lip movement synthesis from speech based on hidden markov models.* Speech Communication, pp.105-115

WATERS, K. 1987. *A muscle model for animating three-dimensional facial expressions.* ACM SIGGRAPH 87, pp.17-24.

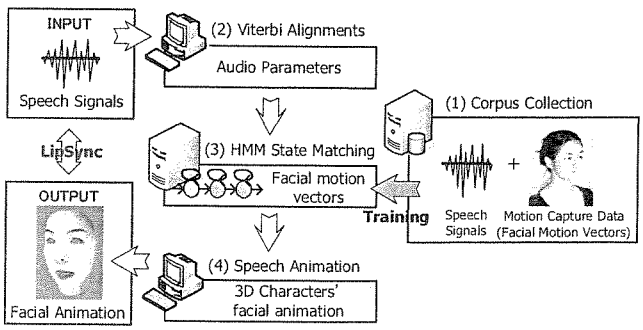


Figure 1: System Overview

* email: tatsuo.yotsukura@atr.jp, **email: shigeo@waseda.jp
***email: satoshi.nakamura@atr.jp

Automatic Head-Movement Control for Emotional Speech

Shin-ichi Kawamoto*
ATR SLC Labs.

Tatsuo Yotsukura†
ATR SLC Labs.

Shigeo Morishima‡
Waseda University
ATR SLC Labs.

Satoshi Nakamura§
ATR SLC Labs.¶

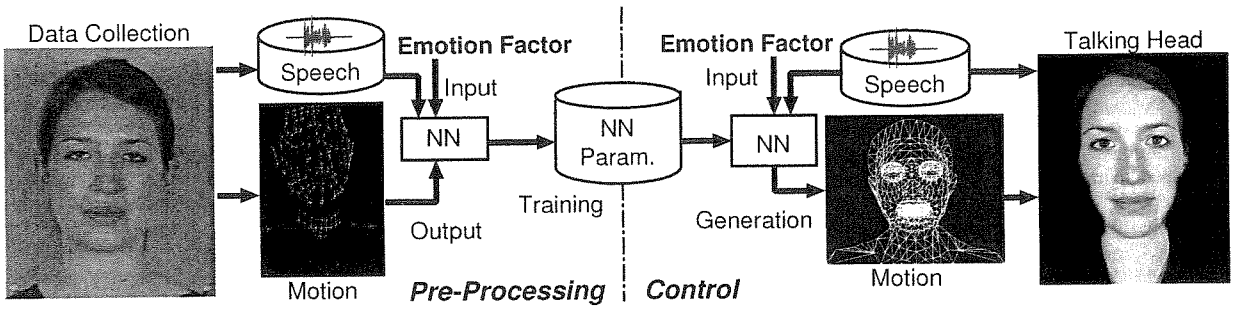


Figure 1: Processing Flow

1 Introduction

Creation of Speaking animation for CG characters needs to synchronously control various factors, such as lips, facial expression, head-movement, gestures, and so on. Lip movement helps a listener to understand speech. In addition, facial expression, head movement, and gestures help to make emotions understandable. There have been many studies that focused on automatic lip animation [Ezzat et al. 2002; Brand 1999]. However, in these studies, facial expression and gesture were realized heuristically or using directly motion-captured data. There have been no studies that tried to automatically control head movement for emotional speech. In this paper, an automatic head-movement control method for emotional speech is proposed.

2 Automatic Head-Movement Control

Automatic head-movement control for speech needs a function to convert from speech to head-movement. We tried to derive this conversion function using actual motion-captured data with emotional speech. This paper presents a Speech-To-Head conversion method using a neural network (NN). The NN we used with a 3-layer feed-forward NN based on a least square error minimization between speech features and head movement parameters.

In addition, the emotional factor should be controllable on user demand. Thus, a user defined emotional factor was added to the input features of the NN to customize emotional expression.

For the input features of the NN, we used fundamental frequency (F_0), power, normalized duration of utterance (ND), and emotional factor. ND is timing information that is normalized by an utterance. ND = 0 means the beginning of an utterance or non-speech region, and ND = 1 means the end of an utterance. The F_0, power, and ND used include from the current frame to the past 10 frames. Namely, the dimension of the input features is 34. The output features of the NN used 3 angles of head rotation that are collected by a motion capture system.

*e-mail: shinichi.kawamoto@atr.jp
†e-mail: tatsuo.yotsukura@atr.jp
‡e-mail: shigeo@waseda.jp
§e-mail: satoshi.nakamura@atr.jp
¶ATR Spoken Language Communication Research Laboratories

3 Data Collection

We collect emotional motion and speech data to achieve automatic head movement control. In this paper, we focus on anger as our target emotion. We collected 3 emotion data: natural, anger, and rage. Each emotion data was heuristically assigned an emotional factor of 0.0, 0.5, and 1.0 respectively.

Motion data was collected using a VICON motion capture system with 181 markers. Then six head movement parameters were calculated: three rotation and three translation. Speech data was collected using a DV camera synchronous to the motion capture system.

4 Implementation

We implemented the talking head with head movement in a Galatea [Kawamoto et al. 2002]. The talking head achieved lip synchronization with recorded English speech. Head movement parameters were generated by speech features using trained NN automatically. In addition, a heuristic low-pass filter was applied to the NN outputs for a smoother head movements. We used an angry face for facial expression, whose intensity corresponds to an emotional factor.

The processing flow of the head movement control for the talking head is shown in Figure 1.

5 Summary

Head movements could be automatically generated from speech data. The expression of head movement could be also controlled by user-defined emotional factors, as shown in the video demonstration.

Acknowledgement This work was supported by Japan Science and Technology Agency, as part of the CREST Project.

References

BRAND, M. 1999. Voice puppetry. In *SIGGRAPH*, 21–28.

EZZAT, T., GEIGER, G., AND POGGIO, T. 2002. Trainable videorealistic speech animation. In *SIGGRAPH*, 388–398.

KAWAMOTO, S., SHIMODAIRA, H., NITTA, T., NISHIMOTO, T., NAKAMURA, S., ITOU, K., MORISHIMA, S., YOTSUKURA, T., KAI, A., LEE, A., YAMASHITA, Y., KOBAYASHI, T., TOKUDA, K., HIROSE, K., MINEMATSU, N., YAMADA, A., DEN, Y., UTSURO, T., AND SAGAYAMA, S. 2002. Open-source software for developing anthropomorphic spoken dialog agent. In *PRICAI-02, Int'l Workshop on Lifelike Animated Agents*, 64–69.

Quantitative Representation of Face Expression Using Motion Capture System

Hiroaki Yanagisawa*
Waseda University.

Akinobu Maejima**
Waseda University

Tatsuo Yotsukura***
ATR SLT

Shigeo Morishima****
Waseda University

1 Introduction

In this paper, we propose a new synthesis method for facial expression movie. The facial expression synthesis method has been proposed in many researchers. In previous research, [Ekman and Friesen, 1977] have proposed Facial Action Coding System (FACS) and anatomically categorized the facial expressions as 44 basic action units (AUs). Facial expression is generated by linear combination of AUs. However, AUs are psychological approach and are not quantified engineered. Also, [Blanz and Vetter, 1999] have proposed reconstruction and facial animation method from still image to 3-D face model using Principal Component Analysis (PCA).

In this research, we firstly capture movement during facial expression, and quantify captured movement engineered . Secondly, we implement PCA on all the measured facial expressions. As a result, we obtained Eigen Face Vectors (EFVs) as an expression change parameter. Moreover, we align the PCA coefficient corresponding EFVs of each frame as a vector, and arrange 63 facial expressions as a matrix. For this matrix, we perform PCA again (we called second stage PCA). As a result of second stage PCA, we estimate parameters that correspond to EFV's coefficients (called Eigen Motion Vectors(EMVs)). Finally, facial expressions are created by linear combination of EFVs and EMVs. The advantage of proposed method is able to generate various facial expression animations using engineered quantified action units with motion.

2 Facial Motion Capture

In order to capture transition of skin surface, we used Motion Capture System (Vicon Co, Ltd) with high quality temporal resolution. Subject is male and an expert in AU representation. We put 146 motion capture markers on subject's face (marker size = 3 mm), especially appeared wrinkle during facial expression. In capturing process, we capture 11 single AUs and 52 combinations of AUs using 8 motion capture cameras (120fps) . Figure 1 shows the placement of motion capture markers on subject's face.

3 Expression synthesis

The facial expression synthesis is generated by allocating the movement of each motion capture marker to vertex of 3-D generic wire frame model. The wire frame model is fitted to subject's frontal image. Number of wire frame vertices (759) is not equal to number of motion capture markers (146). Therefore, it is difficult to allocate the movement of the marker to the vertices. In order to solve this problem, we use Radial Basis Functions Transformation (RBFT) as an interpolation method for not correspondence of the marker to the vertex.

*email: h-yanagisawa@moegi.waseda.jp
**email: akinobu@toki.waseda.jp
***email: tatsuo.yotsukura@atr.jp
****email: shigeo@waseda.jp

4 Eigen Action Units

In order to obtain eigenvectors that include motion element (we called Eigen Action Units(EAUs)), we implement PCA for temporal transition of skin surface. Firstly PCA is implemented on all data in the facial expression database. In the result, the movements between each AU is orthogonalized and compressed dimension of parameter. so, we defined Eigen Face Vector s(EFVs).Secondly, we align PCA coefficients corresponding each frame in the facial expression as a vector, and implement PCA again for arranged the matrix. As a result of second stage PCA, we decide the coefficients of EFVs by using linear combination of EMVs. We can generate any quantified AU expression movie by combination of EAUs. Figure 2 shows EAU1.

5 Results

In this research, we measured temporal transition on skin surface during AUs using the motion capture system, and engineered quantify AUs. Moreover, we proposed synthesis method for generating facial expression animations by combination of Eigen Action Units. In future works, we have to consider to an individual dependency in the EAUs. It's necessary to consider generalization of EAU and controlling speed of facial expression change.

Acknowledgment

This research is supported by Japan Science and Technology Agency,CREST project. and Prof. Hiroshi Yamada, Nihon-University.

References

P.Ekman and W.Friesen. 1977. *Facial Action Coding System.* Consulting Psychologists Press.
V.Blanz and T.Vetter. 1999. A morphable model for the synthesis of 3D faces. In *SIGGRAPH Conference Proceedings*, 187-194

Figure 1: Left Placement of Motion Capture marker
Figure 2: Right Example of EAU1

Face Expression Synthesis
Based on a Facial Motion Distribution Chart

Tatsuo Yotsukura*
ATR Spoken Language Translation
Research Laboratories

Shigeo Morishima**
Waseda University
ATR Spoken Language Translation
Research Laboratories

Satoshi Nakamura***
ATR Spoken Language Translation
Research Laboratories

1 Introduction

In recent years, 3D computer graphic techniques are used for virtual human and cartoon characters in the entertainment industry. Their facial expressions and mouth movements are natural and smooth. However, these successful results require a tremendous amount of time and effort on the part of accomplished CG creators. In fact, the technique of producing facial expressions for a face mesh object typically calls for preparing all of the transformed objects after changing expressions. Then, using blend shape (another way of saying "morphing") deformers, we can change the neutral face object into the transformed objects.

To solve these problems, we propose a technique to create deformed models of the expressions from any user-created face object in a short period of time (Figure 1). First, we create "Facial Motion Distribution Chart" (FMD Chart) which describes the 3D displacement difference of mesh nodes of the face object between an neutral object and a deformed objects. In order to deform a object from the user object using the FMD Chart, we modified the chart smoothly using Radial Basis Function Translation (RBFT). The user who created the face object can thereby create deformed objects with expressions automatically. In other words, the user face object can attain the necessary deformed expressions by preparing various charts of expressions that include realistic, cartoon like and personalized facial expressions. This technique is similar to expression cloning [Noh and Neumann 2001] using RBFT. However we improve accuracy and usability of cloning the face object. For example, we employ FMD-chart of the image-space field (similar to u-v texture-space), which can exactly match the target mesh to separate the upper and lower lips.

This paper describes the definition of FMD Chart, the method for creating the chart, and fitting the user face object and the deformed object. We also demonstrate the results of a re-synthesized user object with facial expression.

2 Facial Motion Distribution Chart

This chart consists of grid structures (pixels), and each pixel stores a 3D displacement. The method for generating a chart from previously defined user face objects with expressions is as follows: First, we fit the frontal face mesh object and an FMD chart that doesn't have any displacements. Next, we determine which pixels of the chart are inside the meshes of the face. When there is a pixel inside a mesh, the 3D displacement is calculated by interpolating each node of mesh that has a corresponding displacement and store it in the pixel. The figure 2 shows an image that is colored in response to the displacements.

3 Fitting FMD Chart and User Face Object

The number of meshes and the structure of the shape are different between the source models used fore the FMD-chart and the user face model. Therefore we have to translate the chart to fit the user face object using RBFT. The anchor points are put on the previously defined frontal user-created face objects and the frontal user face object around the eyes, eyebrows, nose, mouth and outline of the face manually. The total number of the points is about 100, but it takes less than 10 minutes using an easy rule-based operation tool. The pair of anchor points is used to determine the warping of the FMD Chart using RBFT. Figure 2 shows the dedicated FMD chart for the user face object.

* email: tatsuo.yotsukura@atr.jp, **email: shigeo@waseda.jp
***email: satoshi.nakamura@atr.jp

4 Results

The user-created face object with expressions is created to apply the dedicated FMD Chart, as shown in video demonstraion. These objects have a different number of polygons and shape topologies. But the user face object with expression copies the FMD Chart's displacement correctly.

Acknowledgements

This research was supported in part by the National Institute of Information and Communications Technology of Japan.

References

NOH, J., AND NEUMANN, U. 2001, Expression Cloning. In Proceedings of ACM SIGGRAPH 2001, ACM Press / ACM SIGGRAPH, New York. E. Fiume, Ed., Computer Graphics Proceedings, Annual Conference Series, ACM, 277-288.

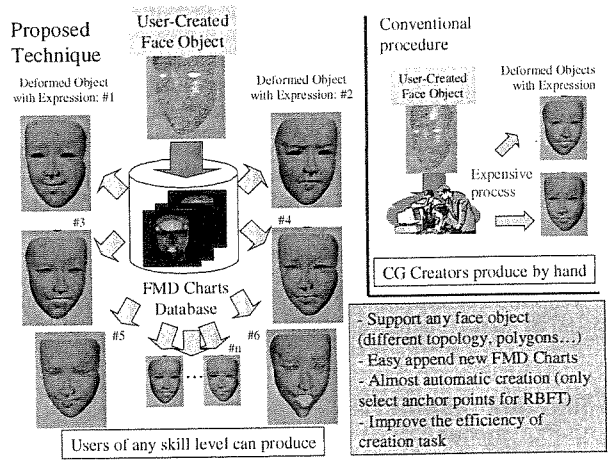
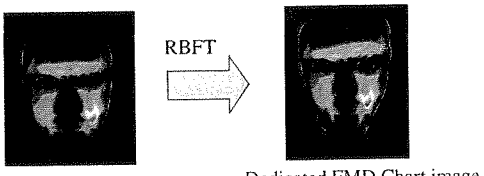


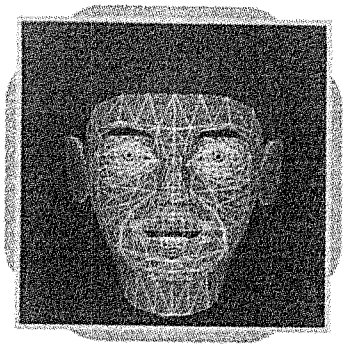
Figure 1: Overview of Proposed Technique



Original FMD Chart image

Dedicated FMD Chart image for user object

Figure 2: Convert FMD-Chart using RBFT



Face Analysis and Synthesis

For Duplication Expression and Impression

Shigeo Morishima

Recently, research in creating friendly human interfaces has flourished. Such interfaces provide smooth communication between a computer and a human. One style is to have a virtual human, or avatar [3], [4] appear on the computer terminal. Such an entity should be able to understand and express not only linguistic information but also nonverbal information. This is similar to human-to-human communication with a face-to-face style [5].

A very important factor in making an avatar look realistic or alive depends on how precisely an avatar can duplicate a real human facial expression and impression on a face precisely. Especially in the case of communication applications using avatars, real-time processing with low delay is essential.

Our final goal is to generate a virtual space close to the real communication environment between network users or between humans and machines. There should be an avatar in cyberspace that projects the features of each user with a realistic texture-mapped face to generate facial expression and action controlled by a multimodal input signal. Users can also get a view in cyberspace through the avatar's eyes, so they can communicate with each other by gaze crossing.

In the first section, the face fitting tool from multiview camera images is introduced to make a realistic three-dimensional (3-D) face model with texture and geometry very close to the original. This fitting tool is a GUI-based system using easy mouse operation to pick up each feature point on a face contour and the face parts, which can enable easy construction of a 3-D personal face model.

When an avatar is speaking, the voice signal is essential in determining the mouth shape feature. Therefore, a real-time mouth shape control mechanism is proposed by using a neural network to convert speech parameters to lip shape parameters. This neural network can realize an interpolation between specific mouth shapes given as learning data [1], [2]. The emotional factor can sometimes be captured by speech parameters. This media conversion mechanism is described in the second section.

For dynamic modeling of facial expression, a muscle structure constraint is introduced for making a facial expression naturally with few parameters. We also tried to obtain muscle parameters automatically from a local motion vector on the face calculated by the optical flow in a video sequence. Accordingly, 3-D facial expression transition is duplicated from actual people by analysis of video images captured by a two-dimensional (2-D) camera

without markers on the face. These efforts are described in the third section. Furthermore, we tried to control this artificial muscle directly by EMG signal processing.

To get greater reality for the head, a modeling method of hair is introduced, and the dynamics of hair in a wind stream can be achieved at low calculation cost. This is explained in the last section. By using these various kinds of multimodal signal sources, a very natural face image and impression can be duplicated on an avatar's face.

Face Modeling

To generate a realistic avatar's face, a generic face model is manually adjusted to the user's face image. To produce a personal 3-D face model, both the user's frontal face image and profile image are necessary at the least. The generic face model has all of the control rules for facial expressions defined by the facial action coding system (FACS) parameter as a 3-D movement of grid points to modify geometry.

Figure 1 shows a personal model fitted to a front image and profile image using our original GUI-based face fitting tool. In this system, corresponding control points are manually moved to a reasonable position by mouse operation.

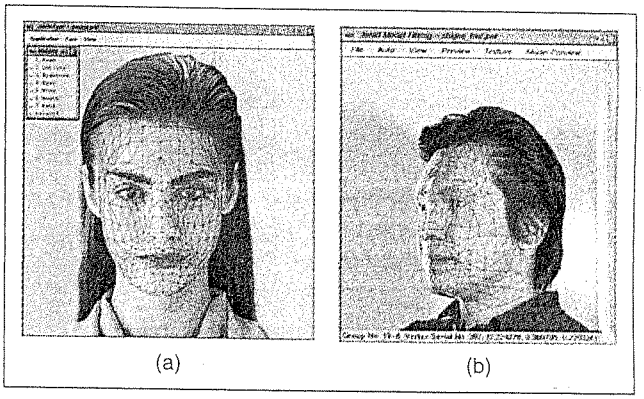
The synthesized face emerges by mapping of the blended texture generated by the user's frontal image and profile image onto the modified personal face model. However, sometimes self-occlusion occurs and then we cannot capture texture from only the front and profile face images in the occluded part of the face model. To construct a 3-D model more accurately, we also introduce a multiview face image fitting tool.

Figure 2(b) shows the fitting result to the captured image from the bottom angle to compensate for the texture behind the jaw. The rotation angle of the face model can be controlled in the GUI preview window to achieve the best fitting to the face image captured from any arbitrary angle. Figure 3 shows the full face texture projected onto a cylindrical coordinate. This texture is projected onto a 3-D personal model adjusted to fit to multiview images to synthesize a face image. Figure 4 shows examples of reconstructed faces. Figure 4(a) uses nine view images and (b) uses only frontal and profile views. Much better image quality can be achieved by the multiview fitting process.

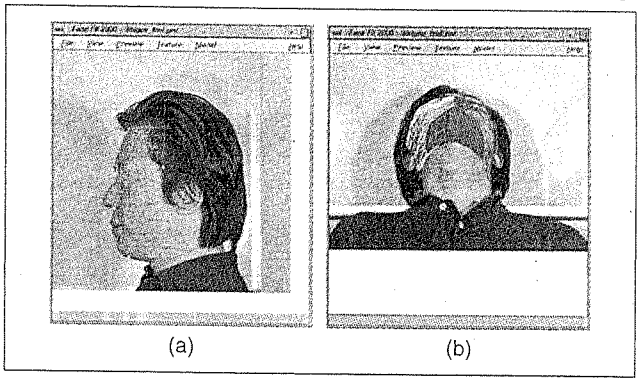
Voice Driven Talking Avatar

The relation between speech and mouth shape is studied widely [1], [6], [7], [9]. To realize lip synchronization, the spoken voice is analyzed and converted into mouth shape parameters by a neural network on a frame-by-frame basis.

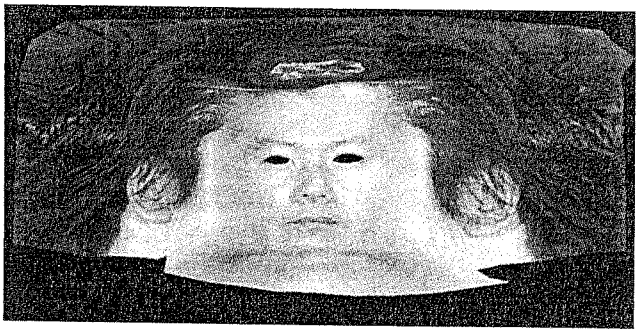
Multiuser communication systems in cyberspace are constructed based on a server-client system. In our system, only a few parameters and the voice signal are transmitted through the network. The avatar's face is synthesized by these parameters locally in the client system.



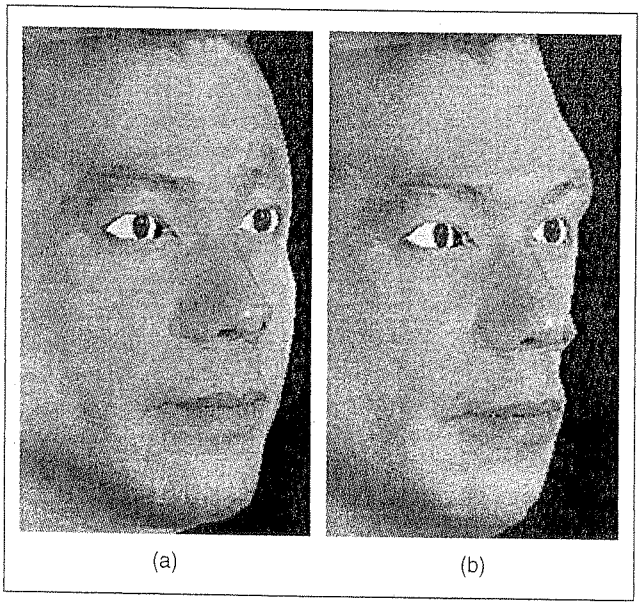
▲ *1. Fitting front and profile model. (a) Front fitting, (b) profile fitting.*



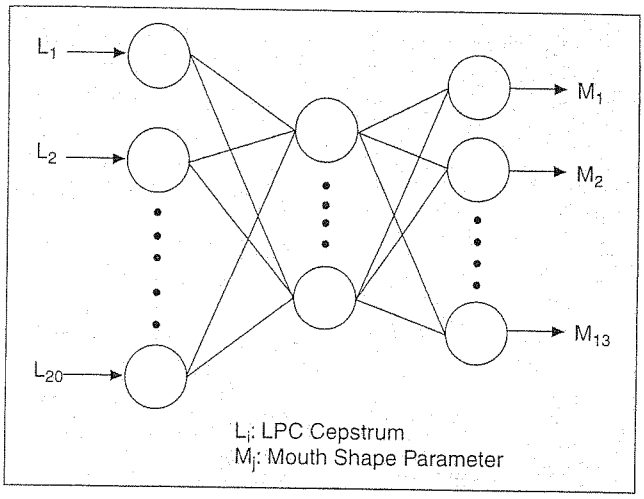
▲ *2. Fitting front and profile model. (a) From bottom, (b) from diagonal.*



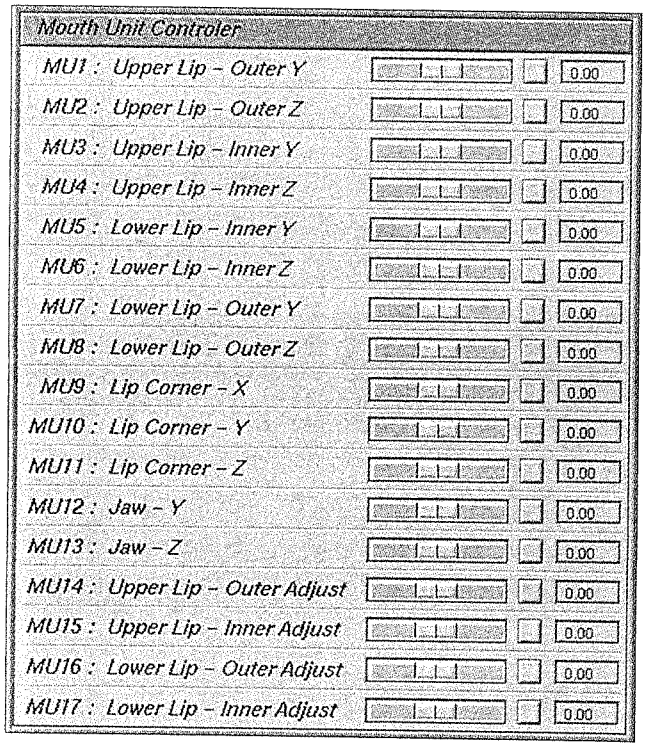
▲ *3. Cylindrical face texture.*



▲ *4. Reconsructed face. (a) Multiview, (b) two view.*



▲ 5. Network for parameter conversion.



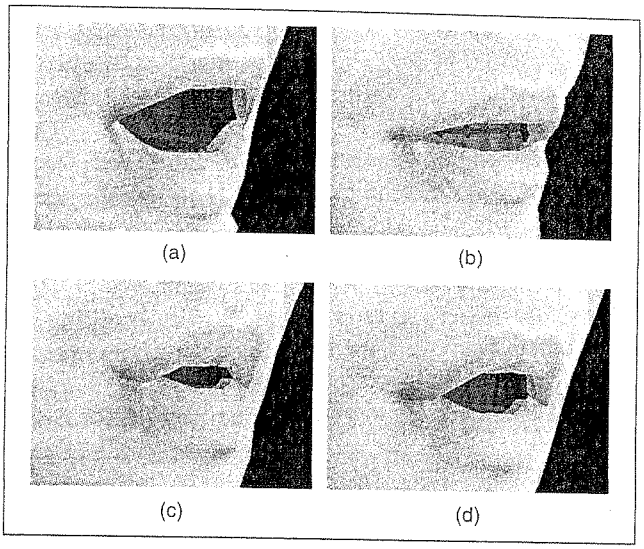
▲ 6. Control panel for mouth shape editor.

Parameter Conversion

In the server system, the voice from each client is phonetically analyzed and converted to mouth shape and expression parameters.

LPC Cepstrum parameters are converted into mouth shape parameters by a neural network trained by vowel features. Figure 5 shows the neural network structure used for parameter conversion. Twenty dimensional Cepstrum parameters are calculated every 32 ms with 32 ms frame length and given into input units of neural network. The output layer has 17 units corresponding to the mouth shape parameters.

In the client system, the on-line captured voice of each user is digitized to 16 kHz and 16 bits and transmitted to the server system frame-by-frame through a network in the communication system. Then the mouth shape of each avatar in cyberspace is synthesized by this mouth shape parameter received at each client.



▲ 7. Typical mouth shapes. (a) Shape for /a/, (b) shape for /i/, (c) shape for /u/, (d) shape for /o/.

The training data for back-propagation are composed of data pairs with voice and mouth shape for vowels, nasals, and transient phoneme captured from video sequence.

Hidden Markov model–based mouth shape estimation from voice is proposed [9]. However, their system cannot realize no-delay real-time processing.

The emotion condition is also determined by LPC Cepstrum, power, pitch frequency, and utterance speed by using discriminant analysis into anger, happiness, or sadness. Each basic emotion has a specific facial expression parameter described by FACS [8].

Mouth Shape Editor

Mouth shape can be easily edited by our mouth shape editor (see Fig. 6). We can change each mouth parameter to set a specific mouth shape on the preview window. Typical vowel mouth shapes are shown in Fig. 7. Our special mouth model has polygons for inside the mouth and the teeth. A tongue model is now under construction. For parameter conversion from LPC Cepstrum to mouth shape, only the mouth shapes for five vowels and nasals are defined as the training set. We have defined all of the mouth shapes for Japanese phonemes and English phonemes by using this mouth shape editor. Figure 8 shows a synthesized avatar face speaking phoneme /a/.

Multiple User Communication System

Each user can walk through and fly through cyberspace by mouse control, and the current locations of all users are always monitored by the server system. The avatar image is generated in a local client machine by the location information from the server system. The emotion condition can always be determined by voice, but sometimes a user

gives his or her avatar a specific emotion condition by pushing a function key. This process works as first priority. For example, push anger and the angry face of your avatar emerges.

The location information of each avatar, mouth shape parameters, and emotion parameters are transmitted every 1/30 seconds to the client systems. The distance between any two users is calculated by the avatar location information, and the voice from every user except himself or herself is mixed and amplified with gain according to the distance. Therefore, the voice from the nearest avatar is very loud and that from far away is silent.

Based on facial expression parameters and mouth shape parameters, an avatar's face is synthesized frame-by-frame. Also, the avatar's body is located in cyberspace according to the location information. There are two modes for display: the view from avatar's own eyes for eye contact and the view from the sky to search for other users in cyberspace. These views can be chosen by a menu in the window. Figure 9 shows a communication system in cyberspace with an avatar.

User Adaptation

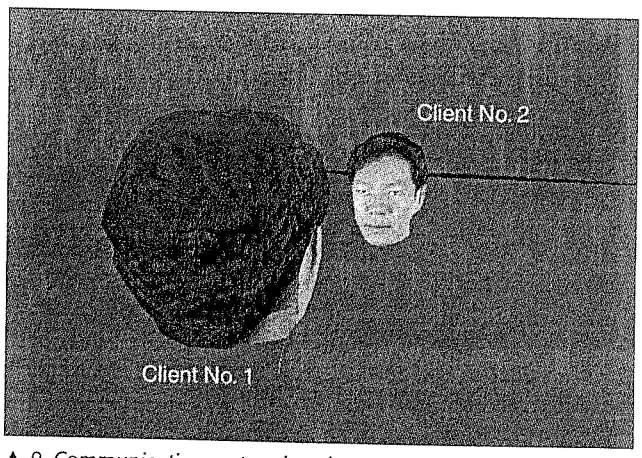
When a new user enters the system, his or her face model and voice model have to be registered before operation. In the case of voice, ideally new learning for the neural network should be performed. However, it takes a very long time to get convergence of back propagation. To simplify the face model construction and voice learning, a GUI tool for speaker adaptation has been prepared. To register the face of a new user, a generic 3-D face model is modified to fit on the input face image. Expression control rules are defined in the generic model, so every user's face can be equally modified to generate basic expressions using the FACS-based expression control mechanism.

For voice adaptation, 75 persons voice data including five vowels are precaptured, and a database for weights of the neural network and voice parameters is constructed. Accordingly, speaker adaptation is performed by choosing the optimum weight from the database. Training of the neural network for the data of each of the every 75 persons is already finished before operation. When a new nonregistered speaker comes in, he or she has to speak five vowels into a microphone. LPC Cepstrum is calculated for each of the five vowels, and this is entered into the neural network. Then the mouth shape is calculated by the selected weight, and the error between the true mouth shape and the generated mouth shape is evaluated. This process is applied to the entire database one-by-one and the optimum weight is selected when the minimum error is detected.

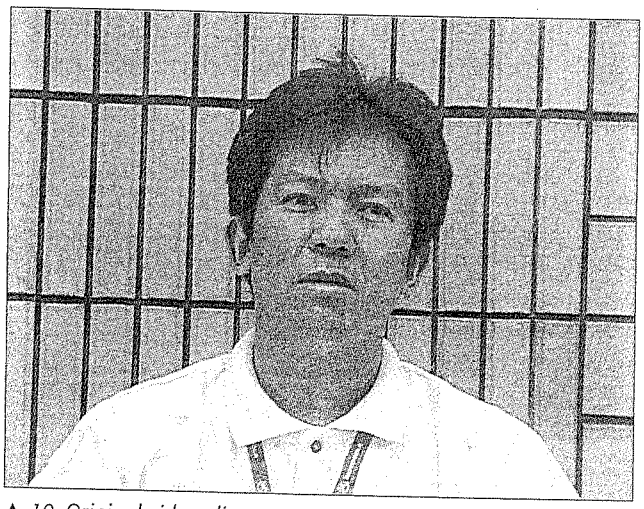
System Features

A natural communication environment between multiple users in cyberspace by transmission of natural voice and real-time synthesis of an avatar's facial expression is realized. The current system works with three users in an intranetwork environment at the speed of 16 frames/s on an SGI Max IMPACT. An emotion model [12] is intro-

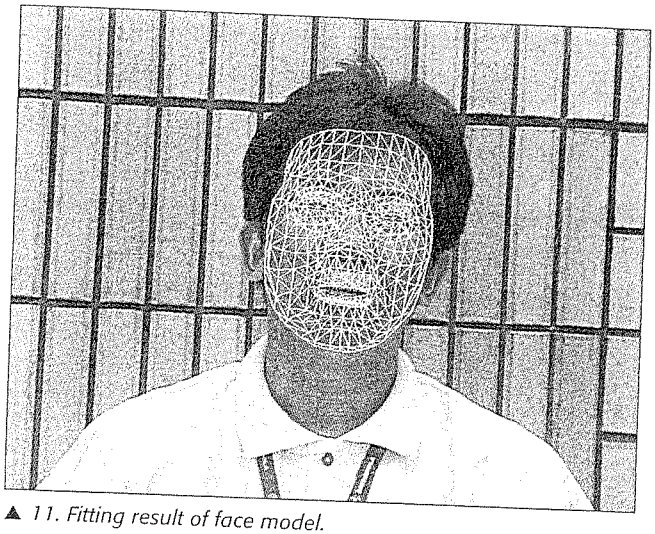
▲ 8. Synthesized face speaking /a/.



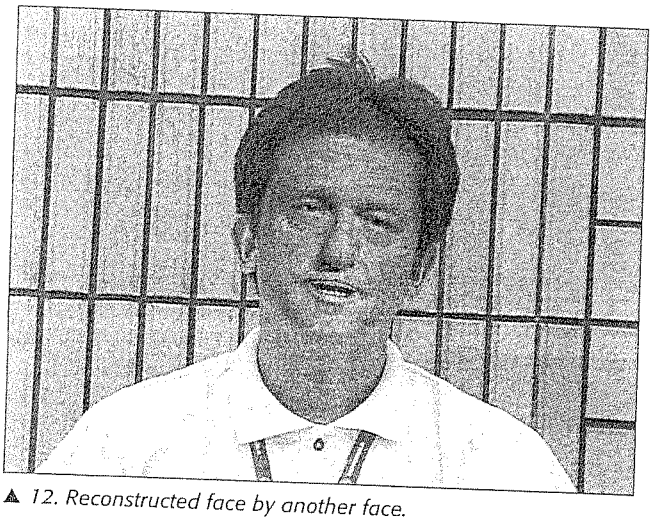
▲ 9. Communication system in cyberspace.



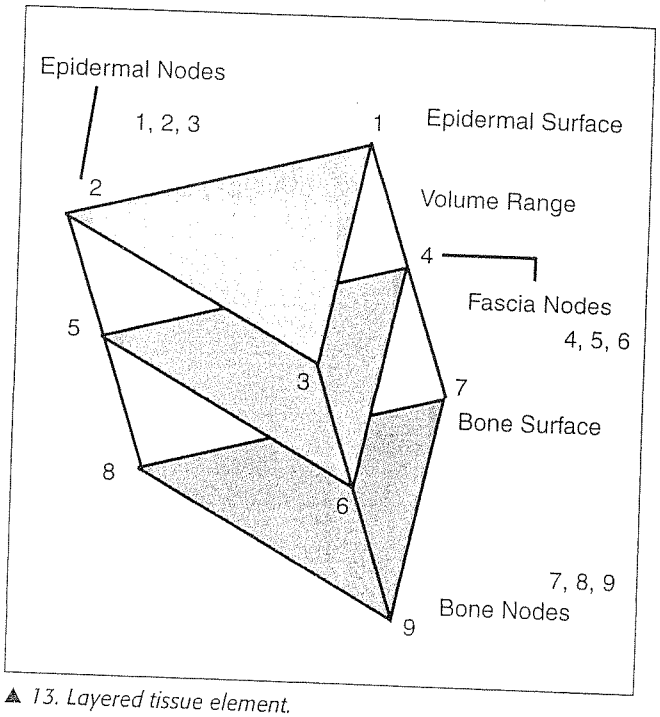
▲ 10. Original video clip.



▲ 11. Fitting result of face model.



▲ 12. Reconstructed face by another face.



▲ 13. Layered tissue element.

duced to improve the communication environment. Gaze tracking and mouth closing point detection can also be realized by a pan-tilt-zoom controlled camera.

Entertainment Application

When people watch movies, they sometimes overlap their own figure with the actor's image. An interactive movie system we constructed is an image creating system in which the user can control the facial expression and lip motion of his or her face image inserted into a movie scene. The user submits a voice sample by microphone and pushes keys that determine expression and special effect. His or her own video program can be generated in real time.

At first, once a frontal face image of a visitor is captured by camera, a 3-D generic wireframe model is fitted onto the user's face image to generate a personal 3-D surface model. A facial expression is synthesized by controlling the grid point of the face model and texture mapping. For speaker adaptation, the visitor has to speak five vowels to choose an optimum weight from the database.

In the interactive process, a famous movie scene is going on and the facial region of an actor or actress is replaced with the visitor's face. Facial expression and lip shape are also controlled synchronously by the captured voice. An active camera tracks the visitor's face, and the facial expression is controlled by a CV-based face image analysis. Figure 10 shows the original video clip, and Fig. 11 shows the result of fitting a face model into this scene. Figure 12 shows a user's face inserted into actor's face after color correction. Any expression can be appended, and any scenario can be given by the user independent of the original story in this interactive movie system.

Muscle Constraint Face Model

Muscle-based face image synthesis is one of the most realistic approaches to the realization of a lifelike agent in computers. A facial muscle model is composed of facial tissue elements and simulated muscles. In this model, forces are calculated to effect a facial tissue element by contraction of each muscle string, so the combination of each muscle's contracting force decides a specific facial expression. This muscle parameter is determined on a trial-and-error basis by comparing a sample photograph and a generated image using our muscle editor to generate a specific face image. In this section, we propose the strategy of automatic estimation of facial muscle parameters from 2-D optical flow by using a neural network. This corresponds to the 3-D facial motion capturing from 2-D camera images under a physics-based condition without markers.

We introduce a multilayered back-propagation network for the estimation of the muscle parameter. A neural network is used to classify 3-D objects from a 2-D view [15], recognize expressions [17], and model facial emo-

tion condition [12], and a self-organizing mechanism is well studied [16].

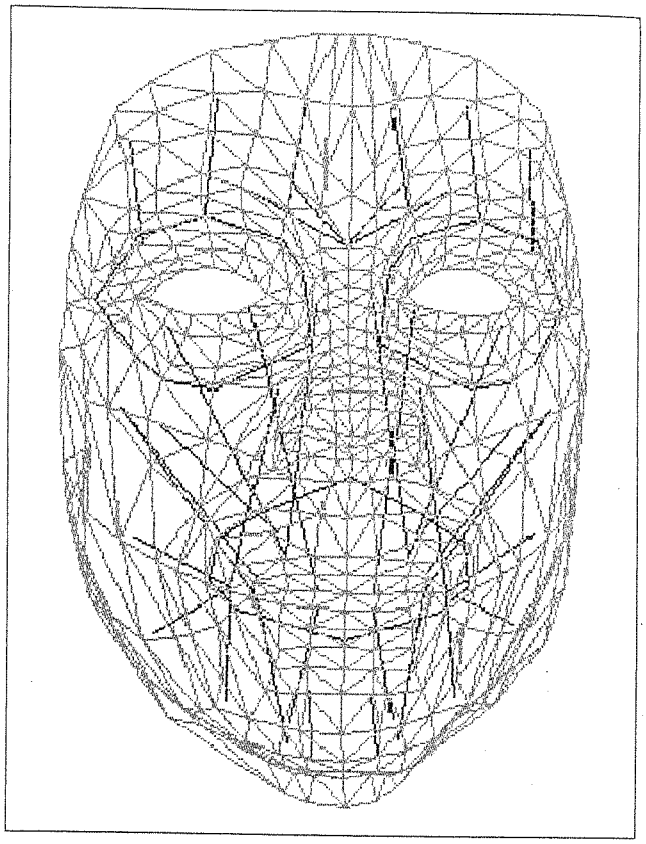
The facial expression is then resynthesized from the estimated muscle parameter to evaluate how well the impression of an original expression can be recovered. We also tried to generate animation by using the captured data from an image sequence. As a result, we can obtain and synthesize an image sequence that gives an impression very close to the original video.

Layered Dynamic Tissue Model

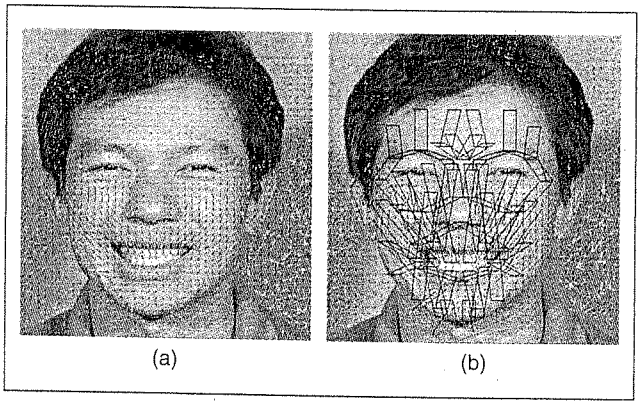
The human skull is covered by a deformable tissue that has five distinct layers. In accordance with the structure of real skin, we employ a synthetic tissue model constructed from the elements illustrated in Fig. 13, consisting of nodes interconnected by deformable springs (the lines in the figure). The facial tissue model is implemented as a collection of node and spring data structures. The node data structure includes variables to represent the nodal mass, position, velocity, acceleration, and net force. Newton's laws of motion govern the response of the tissue model to force [13].

Facial Muscle Model

Figure 14 shows our simulated muscles. The black line indicates the location of each facial muscle in a layered tissue face model. Normally, muscles are located between a bone node and a fascia node. But the orbicularis oculi has an irregular style, whereby it is attached between fascia nodes in a ring configuration; it has eight linear muscles that approximate a ring muscle. Contraction of the ring muscle makes the eyes thin. The muscles around the mouth are very important for the production of speaking scenes. Most Japanese speaking scenes are composed of vowels, so we mainly focused on the production of vowel mouth shapes as a first step and relocated the muscles around the mouth [14]. As a result, the final facial muscle model has 14 muscle springs in the forehead area and 27 muscle springs in the mouth area.



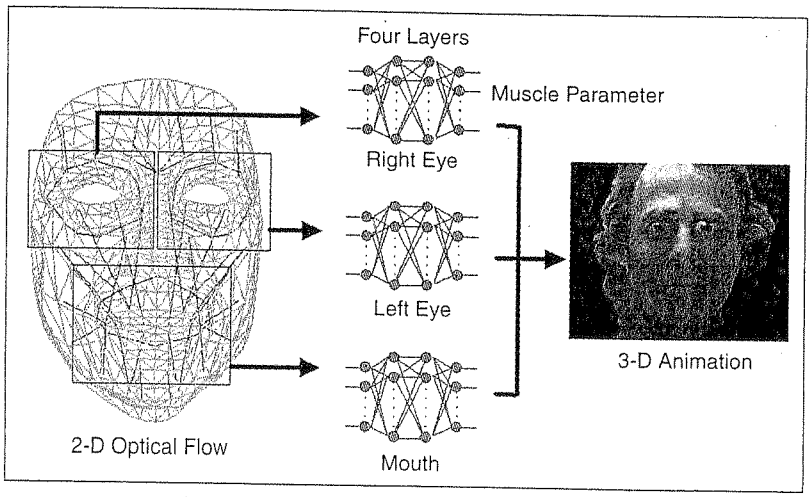
▲ *14. Facial muscle model.*



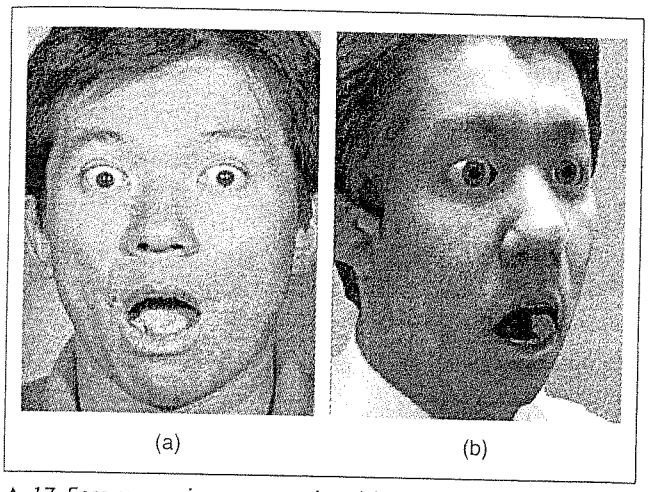
▲ *15. Motion feature vector on face. (a) Optical flow, (b) feature window.*

Motion Capture on Face

Face tracking [10] and face expression recognition [11] based on optical flow have already been proposed. Their purpose is to find a similar cartoon face by some matching processes; however optical flow is very sensitive to noise. Our goal is to estimate precisely the muscle parameters and resynthesize a texture-mapped face expression using muscle model with the original impression. A personal face model is constructed by fitting the generic control model to personal range data. An optical flow vector is calculated in a video sequence, and we accumulate the motion in the sequence from neutral to each expression. Then motion

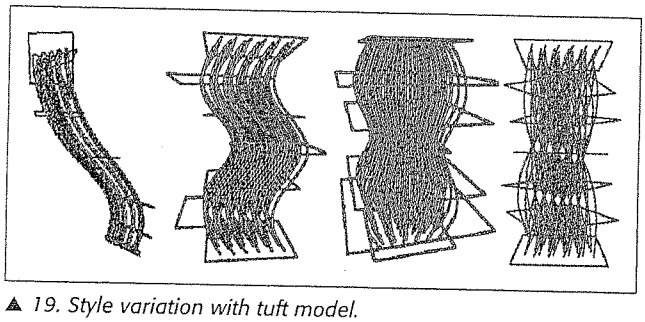


▲ *16. Conversion from motion to animation.*

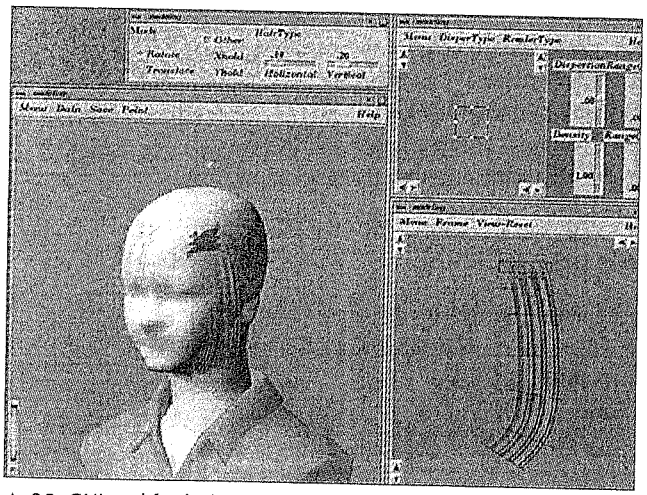


▲ 17. Face expression regeneration. (a) Original, (b) regenerated.

▲ 18. Mouth shape control by EMG.



▲ 19. Style variation with tuft model.



▲ 20. GUI tool for hair designing.

vectors are averaged in each window as shown in Fig. 15(b). This window location is determined automatically in each video frame. Our system is not limited to recognize only six basic expressions [17], [18].

A layered neural network finds a mapping from the motion vector to the muscle parameter. A four-layer structure is chosen to model effectively the nonlinear performance. The first layer is the input layer, which corresponds to a 2-D motion vector. The second and third layers are hidden layers. Units of the second layer have a linear function, and those of the third layer have a sigmoid function. The fourth layer is the output layer, corresponding to the muscle parameter, and it has linear units. Linear functions in the input and output layers are introduced to maintain the range of input and output values.

A simpler neural network structure can help to speed up the convergence process in learning and reduce the calculation cost in parameter mapping, so the face area is divided into three sub-areas as shown in Fig. 16. They are the mouth area, left-eye area, and right-eye area, each giving independent skin motion. Three independent neural networks are prepared for these three areas.

Learning patterns are composed of basic expressions and the combination of their muscle contraction forces.

A motion vector is given to a neural network from each video frame, and then a facial animation is generated from the output muscle parameter sequence. This is a test of the effect of interpolation on our parameter conversion method based on the generalization of the neural network.

Figure 17 shows the result of expression regeneration for surprise from an original video sequence.

Muscle Control by EMG

EMG data is the voltage wave captured by a needle inserted directly into each muscle so that the feature of wave expresses the contraction situation of a muscle. In particular, seven di-ball wires were inserted into the muscles around the mouth to make a model of mouth shape control.

The conversion from EMG data to muscle parameters is as follows. First, the average power of an EMG wave is calculated every 1/30 seconds. Sampling frequency of EMG is 2.5 kHz. The maximum power is equal to the maximum muscle contraction strength by normalization. Then the converted contraction strength of each muscle is given every 1/30 seconds and the facial animation scene is generated. However, the jaw is controlled directly by the marker position located on the subject's jaw independent of EMG data.

EMG data is captured from a subject speaking five vowels. Seven kinds of waves can be captured and seven contractions of muscles are determined. Then the face model is modified and the mouth shape for each vowel is synthesized. The result is shown in Fig. 18. In Fig. 18, the camera captured image and synthesized image can be compared, and the impression is very close to each other. Next, EMG time sequence is converted into muscle pa-

rameters every 1/30 seconds, and facial animation is generated. Each sample is composed of a three-second sentence. Good animation quality for speaking the sentence is achieved by picking up impulses from the EMG signal and activating each appropriate muscle. This data offer the precise transition feature of each muscle contraction.

Dynamic Hair Modeling

The naturalness of hair is a very important factor in a visual impression, but it is treated with a very simple model or as a texture. Because real hair has huge pieces and complex features, it's one of the hardest targets to model by computer graphics [19], [20]. In this section, a new hair modeling system is presented [21]. This system helps the designer to create any arbitrary hair style by computer graphics with easy operation by using an editor of tufts. Each piece of hair has an independent model, and dynamic motion can be simulated easily by solving a motion equation. However, each piece has only a few segments and the feature is modeled by a 3-D B-spline curve, so the calculation cost is not so huge.

Modeling Hair

It is necessary to create a model of each piece of hair to generate a natural dynamic motion when wind is flowing. However, the hair feature is very complex, and a real head has more than 100,000 pieces. Therefore, a polygon model for hair needs a huge amount of memory for the entire head to be stored. In this article, each piece has only a few segments to control and is expressed with a 3-D B-spline curve.

Tuft Model

In our designing system, a head is composed of about 3,400 polygons, and a 3-D B-spline curve comes out from each polygon to create hair style. The hair generation area is composed of about 1,300 polygons, and one specific hair feature is generated and copied for each polygon. To manipulate hair to get a specific hairstyle, some of the hair features are simultaneously treated as one tuft. Each tuft is composed of more than seven squares in which hair segments pass through. This square is the gathering control point of the hair segments, so manipulation of this square can realize any hairstyle by rotation, shift, and modification. This tuft model is illustrated in Fig. 19.

Hair Style Designing System

There are three processes to generate hair style by GUI, and each process is easily operated by mouse control. The first process is to decide the initial region on the head from which the tuft comes out. The second process is to manipulate a hair tuft by modifying the squares by rotation, shift, and expansion. The third process is matching a tuft onto the surface of the head. After these processes, each polygon gets one hair feature, and in total 1,300 features are available. Finally, 20,000 to 30,000 pieces of hair for static images, or 100,000 to 200,000 pieces for animation, are generated by copying the feature in each polygon. The GUI tool for hair designing is shown in Fig. 20.

▲ *21. Example of hair style.*

▲ *22. Copying real hair style.*

Rendering

Each hairpiece is modeled with a 3-D B-Spline curve and also modeled as a very thin circular pipe. Therefore, the surface normal can be determined in each pixel to introduce the Lambert model and Phong model. The typical hairstyle Dole Bob is shown in Fig. 21. Figure 22 shows an example modeling of a real hairstyle. The impression of the generated hair is very natural.

This hair piece is modeled by stick segment and can be controlled dynamically by solving motion equation. Collision detection between hair and head is also considered to generate natural dynamic hair animation.

Conclusion

To generate a realistic avatar's face for face-to-face communication, a multimodal signal is introduced to duplicate original facial expressions. Voice is used to realize lip synchronization and expression control. Video captured images are used to regenerate an original facial expression under a facial muscle constraint. EMG data is used to control directly an artificial muscle. Finally, a hair modeling

method is proposed to make an avatar appear more natural and believable.

Shigeo Morishima received the B.S., M.S., and Ph.D. degrees, all in electrical engineering, from the University of Tokyo, Tokyo, Japan, in 1982, 1984, and 1987, respectively. Currently, he is a Professor at Seikei University, Tokyo, Japan. His research interests include physics-based modeling of face and body, facial expression recognition and synthesis, human computer interaction, and future interactive entertainment using speech and image processing. He was a Visiting Researcher at the University of Toronto from 1994 to 1995. He has been engaged in the multimedia ambiance communication TAO research project as a Subleader since 1997. He has been also a temporary lecturer at Meiji University, Japan, since 2000 and a visiting researcher at the ATR Spoken Language Translation Research Laboratories since 2001. He is an editor of the *Transaction of the Institute of Electronics, Information and Communication Engineers* (IEICE), Japan. He received the 1992 Achievement Award from IEICE.

References

[1] S. Morishima and H. Harashima, "A media conversion from speech to facial image for intelligent man–machine interface," *IEEE J. Select. Areas Commun.*, vol. 9, no. 4, pp. 594-600, 1991.

[2] S. Morishima, "Virtual face-to-face communication driven by voice through network," in *Workshop on Perceptual User Interfaces*, 1997, pp. 85-86.

[3] N.I. Baddler, C.B. Phillip, and B.L. Webber, *Simulating Humans, Computer Graphics Animation and Control*. Oxford, U.K.: Oxford Univ. Press, 1993.

[4] N.M. Thalmann and P. Kalra, "The simulation of a virtual TV presenter," in *Computer Graphics and Applications*. Singapore: World Scientific, 1995, pp. 9-21.

[5] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents," in *Proc. SIGGRAPH'94*, pp. 413-420.

[6] E. Vatikitotis-Bateson, K.G. Munhall, Y. Kasahara, F. Garcia, and H. Yehia, "Characterizing audiovisual information during speech," in *Proc. ICSLP-96*, pp. 1485-1488.

[7] K. Waters and J. Frisbie, "A coordinate muscle model for speech animation," in *Proc. Graphics Interface'95*, pp. 163-170.

[8] P. Ekman and W.V. Friesen, *Facial Action Coding System*. Palo Alto, CA: Consulting Psychologists, 1978.

[9] S. Nakamura, "HMM-based transmodal mapping from audio speech to talking faces," in *IEEE Int. Workshop on Neural Networks for Signal Processing*, 2000, pp. 33-42.

[10] I. Essa, T. Darrell, and A. Pentland, "Tracking facial motion," in *Proc. Workshop on Motion and Nonrigid and Articulated Objects*, 1994, pp. 36-42.

[11] K. Mase, "Recognition of facial expression from optical flow," *IEICE Trans.*, vol. E-74, no. 10, pp. 3474-3483, 1991.

[12] S. Morishima, *Modeling of Facial Expression and Emotion for Human Communication System Displays*. Amsterdam, The Netherlands: Elsevier, 1996, pp. 15-25.

[13] Y. Lee, D. Terzopoulos, and K. Waters, "Realistic modeling for facial animation," in *Proc. SIGGRAPH'95*, pp. 55-62.

[14] H. Sera, S. Morishima, and D. Terzopoulos, "Physics-based muscle model for mouth shape control," in *Proc. Robot and Human Communication*, 1996, pp. 207-212.

[15] S. Suzuki and H. Ando, "Unsupervised classification of 3D objects from 2D view," in *Advances in Neural Information Processing System 7*. Cambridge, MA: MIT Press, 1995, pp. 949-956.

[16] Y. Katayama and K. Ohyama, "Some characteristics of self-organizing back propagation neural network," in *Annu. Convention Rec. IEICE*, 1989, pp. SD-1-14.

[17] H. Kobayashi and F. Hara, "Analysis of the neural network recognition characteristics of 6 basic expressions," in *Proc. IEEE ROMAN'94*, pp. 222-227.

[18] M. Rosenblum, Y. Yacoob, and L. Davis, "Human emotion recognition from motion using a radial basis function network architecture," in *Proc. Workshop on Motion and Non-Rigid and Articulated Objects*, 1994, pp. 43-49.

[19] N. Thalman, T. Kurihara, and D. Thalman, "An integrated system for modeling, animating and rendering hair," in *Eurographics'93*, pp. 211-221.

[20] L.-H. Chen, S. Saeyor, H. Dohi, and M. Ishizuka, "A system of 3D hair style synthesis based on wisp model," *Visual Computer*, vol. 15, no. 4, pp. 159-170, 1998.

[21] K. Kishi and S. Morishima, "Modeling and animating human hair," *Trans. IEICE*, vol. J83-D-2, no. 12, pp. 2716-2724, 2000.

2．5　表情合成の評価

表情合成結果の評価方法について、提案を行った。ハイスピードカメラにより、より繊細な表情合成規則の設定が必要であることを確認した。また音声との併用によって、リップシンクアニメーションの評価方法を確立した。

論　文

雑音環境下での音声の聞き取り実験による合成発話顔アニメーションの評価

前島　謙宣[†a)]　四倉　達夫[††]　森島　繁生[†,††]　中村　哲[††]

Subjective Evaluation of Synthetic Talking Face in Acoustically Noisy Environments

Akinobu MAEJIMA[†a)], Tatsuo YOTSUKURA[††], Shigeo MORISHIMA[†,††], and Satoshi NAKAMURA[††]

あらまし　人間のような見た目をもつ擬人化エージェントの実現は，コンピュータを介して人間同士のコミュニケーションの幅を広げるための重要な研究課題である．筆者らは，このようなコミュニケーションを可能にするための，自然な発話顔アニメーションの合成手法を提案している．しかし，発話顔アニメーションに対する性能の評価方法は課題として残されていた．発話顔アニメーションの性能は，（1）読唇をできる程度に再現されているか，（2）視覚的に自然であるか，（3）音声と正確に同期しているかの3点により決定される．本論文では，まず雑音環境下において発話顔アニメーションと音声とを被験者に提示し，発話内容の聞き取り実験を行うことにより（1）を検証する．次に（2）について，発話顔アニメーションの視覚的な自然さ及び発話口形の滑らかさを5段階評価する．最後に（3）について，ある一定間隔で音声と発話顔アニメーションとの同期をずらしたものを被験者に提示し，同期のずれの主観値を調査するとともに，違和感の程度を5段階評価により評価する．加えて音声と発話顔アニメーションとの同期のずれが音声の知覚に及ぼす影響についても評価する．これらの評価実験を通じて，筆者らが提案する合成発話顔アニメーションの品質を評価するとともに，合成発話顔アニメーションと音声との自然な同期について検証した．

キーワード　合成発話顔アニメーション，雑音環境下，再現性，自然性，音声との同期

1. まえがき

近年，人間と機械との自然で豊かなインタラクションの実現を目的とした，ヒューマンコンピュータインタフェースの研究が盛んに行われている．人間のような外見をもつ擬人化エージェントの実現は重要な課題の一つといえる[1]．エージェントには，その性質上ユーザである人間とFace-to-Faceの自然なコミュニケーションを行えることが強く望まれている．

ところで，Face-to-Faceの対話においては，音声のような聴覚的・言語的な情報だけでなく，口の動きや表情といった視覚的・非言語情報も重要な役割を果たしている．一般に人間は，発話内容を知覚する場合，聴覚的な情報だけでなく，視覚的な情報，会話の話題，文章の前後関係，過去の経験といった情報を統合し，発話内容を認識しているとされ，特に音声や口の動きには互いに強い相関があると考えられている．このような二つのモダリティ間の同期がとられていない場合，人は不自然な感覚を覚えたり，誤った知覚をする[2]．したがって，エージェントには，自然かつ音声との同期がとれた正確な口の動きが要求される．自然なエージェント実現のための，合成発話顔アニメーションに関する研究は，既にいくつか報告がなされている[3]～[6]，[8]．

Ezzatら[8]は，顔画像コーパスに基づいたイメージベースの自然な発話顔画像の合成手法を実現している．あらかじめ，発話時の各顔器官画像をコーパスと

† 早稲田大学大学院理工学研究科，東京都
Graduate School of Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
†† （株）国際電気通信基礎技術研究所，音声言語コミュニケーション研究所，京都府
ATR Spoken Language Translation Research Laboratory, 2-2-2 Hikaridai, Seika-cho, Souraku-gun, Kyoto-fu, 619-0288 Japan
a) E-mail: akinobu@toki.waseda.ac.jp

して用意し，合成する際に各器官の画像を必要に応じてコーパスから取得し，基板となる顔上に二次元的に再配置することで発話顔アニメーションを生成するものである．しかしながら，このような手法の場合，顔の回転や，移動に対する制約が必要であり，顔器官の画像を蓄えておくデータベースも巨大になるという問題がある．

垣原らは，HMM に基づく発話顔画像の合成手法を提案している．これは，人物の発話時の三次元座標を計測し，計測された座標に対して主成分分析を行い，音響特徴と顔の特徴である主成分を，HMM の状態単位で対応づけ，連結したものを HMM で学習している [4]．合成の際には，入力音声に対してビタビアラインメントを行うことで，音声と同期した発話顔アニメーションを得ることができる．

それに対して筆者らは，標準的な三次元顔モデルを個人の顔に適用するモデルベースの発話顔画像の合成手法を提案しており，これを応用したビデオ翻訳システムを構築している [9]．提案手法の利点は，顔が三次元モデルで表現されているため，移動・回転に制約がなく，発話口形を生成するためのルールを構築することが容易でかつ，特定の個人に限らず適用できることが挙げられる．

発話口形状は，三次元顔モデルの頂点のベクトル移動法則を定めた口形パラメータ [10] と，無音時及び発話時の顔の三次元座標のベクトル移動量を用いて生成される．このベクトル移動量は個人性をできるだけ除去するため，数人分のベクトル移動量を用いることで平均化されている．これによりイメージベースの手法と比較して，個人に依存しない，比較的少数のデータベースから発話口形状の生成を可能にしている．

ところで，このような合成発話顔アニメーションの品質の評価には，一般的に客観評価法と主観評価法が用いられている [7]．しかしながら提案手法の場合，発話口形の品質を客観評価法により確かめることは困難である．このため筆者らは，主観評価法により発話口形の品質を検証している [12]．しかし発話顔アニメーションの再現性や，視覚的な顔画像の自然さ，合成発話顔アニメーションと音声との同期に関する厳密な評価は行われていなかったといえる．

合成発話顔アニメーションの品質は，（1）発話顔アニメーションが読唇をできる程度に再現されているか（以後，発話アニメーションの再現性と呼ぶ）．（2）発話顔アニメーションが視覚的に自然であるか（以後，発話アニメーションの自然性と呼ぶ）．（3）発話顔アニメーションが音声と正確に同期して表現されているか（以後，発話顔アニメーションの同期性と呼ぶ）の 3 点により決定される．

本研究では，まず雑音環境下において発話顔アニメーションと音声とを被験者に提示し，発話内容の聞き取り実験を行うことにより（1）を検証する．次に（2）について，発話顔アニメーションの視覚的な自然さ及び，発話口形の滑らかさに対して 5 段階評価を行う．最後に（3）について，ある一定間隔で音声と発話顔アニメーションとの同期をずらしたものを被験者に提示し，同期のずれの主観値を調査するとともに，違和感の程度を 5 段階評価により評価する．加えて音声と発話顔アニメーションとの同期のずれが音声の知覚に及ぼす影響についても評価する．

以上の評価実験を通じて，筆者らが提案する合成発話顔アニメーションの品質を評価するとともに，合成発話顔アニメーションを作成する上で重要な音声との自然な同期について検証した．

2. 合成発話顔アニメーションの生成

本研究では，[9] に基づく手法により合成発話顔アニメーションを作成した．手順を図 1 に示す．まず，データベース中の各画像シーケンスに対して，その初期フレームを用いて，個人の三次元顔モデルを作成する．次に映像中の顔の動きを推定するために，顔モデルより作成される三次元顔テンプレートモデルを用いて自動顔トラッキングを行う．

更に，発話内容を既知として音声を音素セグメンテーション（自動＋手動補正）することで音素継続長を得る．ここで，発話内容の音素表記に対応する発話口形を基本口形データベースより取得する．更に，音素継続長情報を用いて基本口形間の口形状の補間を行う．最終的に得られた発話している口領域モデルを，顔トラッキングにより推定された位置・角度へと埋め込むことで，合成発話顔アニメーションが作成される．

2.1 三次元顔モデルの生成

人間の顔は，基本的な形状や構造は同じといってよいが，目，鼻，口等の各部位の形状や位置は個人に依存する．このため CG により自然な顔を合成するには，対象となる人物の顔により忠実かつ演算量の少ない三次元顔モデルを構築する必要がある．そこで本研究では，標準ワイヤフレームモデルを個人の顔へと整合することにより，個人の三次元顔モデルを生成した．

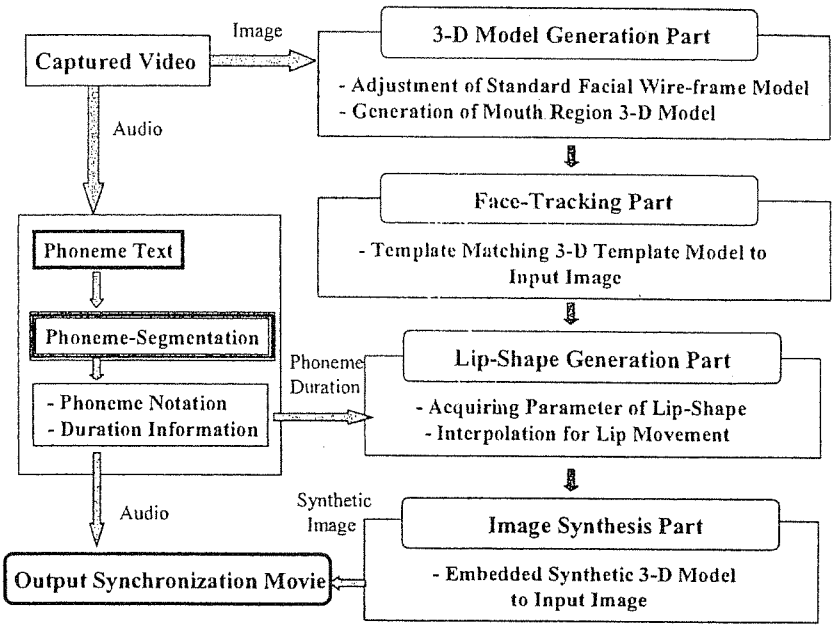


図 1　合成発話顔アニメーションの生成手順
Fig. 1　Creation procedure for synthetic talking face.

図 2　原　画　像
Fig. 2　Original image.

図 3　ビデオフレームへの整合
Fig. 3　Fitting to video frame.

図 2 は画像シーケンスの任意の 1 フレームから取得した原画像を示している．図 3 は，原画像に対して標準顔ワイヤフレームの整合を行った結果を示している．この整合には専用の GUI を用いて約 1 分ほどの作業を必要とする [10]．そして最後に，整合により得られた画像を三次元顔モデルにテクスチャマッピングを施すことにより，個人の三次元顔モデルが得られる．

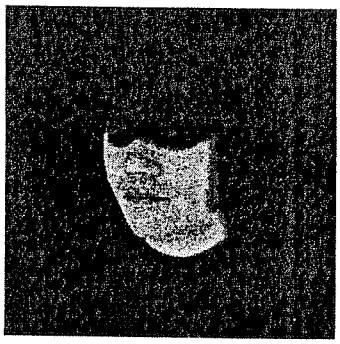


図 4　取得した口領域モデル
Fig. 4　Mouth model.

合成発話顔は，生成された三次元顔モデルを画像フレーム中に埋め込み，口領域のワイヤフレームの頂点を移動させることにより生成される．しかしながら，顔領域全体のモデルを用いたのでは，口の動き以外の瞬きや表情変化等のプロソディが失われてしまう．そこで，このような情報をできる限り保持するために，三次元顔モデルの口唇領域以外を削除した図 4 に示すような口領域のみのモデルを用いた．また口内に関しては，あらかじめ用意したモデルを適用した．

2.2　自動顔トラッキング

合成発話顔アニメーションは，2.1 で述べた三次元

口領域モデルを画像上に埋め込むことで生成される．しかしながら三次元口領域モデルを，画像上のどの位置にどのくらい回転させて埋め込めばよいのかは明らかになっていない．このため，あらかじめ映像中の人物の顔の位置・角度を知る必要がある．本研究では，三次元顔テンプレートモデルを用いた自動顔トラッキングにより，映像中の人物の顔の位置・角度を推定した [9]．

2.3 音素セグメンテーション

本研究では，音声の発話内容は既知であるとして，音素セグメンテーションに HTK [13] を用いた．音響特徴は，16 kHz サンプリング周波数，フレーム長 25 ms，フレーム周期 10 ms で抽出された十二次元 MFCC，十二次元 ΔMFCC，Δ 対数パワーを使用した．音響モデルには，話者非依存の IPA 準拠のモデルを用いた．また，音素セグメンテーションに誤りが生じている場合は，音声と口の動きの同期がとれなくなるため，手動で音素境界を補正した．

2.4 発話顔アニメーションの生成

2.4.1 発話口形の生成

人間が会話する際の動作の大きな部位として，唇，顎などが挙げられる．とりわけ，唇の動きは音韻と密接な関係をもつため，正確な制御が必要とされる．倉立らは，被験者の顔にマーカを置き，運動情報を計測している [5]．このようなアプローチは，運動を正確に計測でき，柔軟な制御を行えるという利点がある．

本研究では，ビデオ翻訳システム [9] で使用されている発音記号を VISEME に基づいて日本語について 12 種類に分類し，更に無音状態を加えた 13 種類の基本口形を，基本口形データベースとして使用した（図 5）．基本口形は，ワイヤフレームモデルの口領域のベクトル移動量を定義した口形パラメータと，三次元計測器を用いて計測することで得られる無音及び発話時における口形状の三次元位置情報のベクトル移動量を用いて作成される．ここで，後者のベクトル移動量については話者数人分について計測し，その平均値を用いることにより個人性を除去している．また，本来 VISEME は二重母音 [au][ei] 等に現れる口唇運動の情報まで定義されるのであるが，そのような VISEME は，二つの VISEME から構成されるものとし，二つの基本口形を割り当てることにより表現した．前半に現れる口形については音素継続長の 30%を，後半の口形については残りの音素継続長を経験的に割り当てた．これにより，話者に依存しない小規模なデータベースを構築している．例として図 6 に母音 a を発話している際の口形を示す．

VISEME No.	Phoneme Notation
1	/a/
2	/i/, /y/
3	/u/
4	/e/
5	/o/
6	/r/, /ry/
7	/b/, /p/, /m/, /by/, /py/, /my/
8	/t/
9	/d/, /n/, /ny/
10	/g/, /k/, /N/, /hy/, /gy/, /ky/
11	/f/
12	/j/, /s/, /z/, /ch/, /dy/, /sh/, /ts/
0	/#/ silence

図 5 音素と VISEME の対応
Fig. 5 Correspondance phoneme to VISEME.

図 6 発話口形（左：無音 右：母音 a 発話時の口形）
Fig. 6 Synthetic image of talking face. (left: Silence, right: Vowel a)

2.4.2 発話口形状の補間

基本口形データベースより取得された基本口形をキーフレームとして，キーフレームアニメーションを行う．ビデオレートに依存したキーフレーム以外のビデオフレームにおける発話口形状は，基本口形データベースには存在しておらず，このままでは音声と発話口形状との完全な同期を得ることができない．そこで，音素継続長情報を利用して基本口形間の発話口形状の補間を行う．

音素の発声開始時は必ず基本口形状を構成しているとして，これをキーフレームとした．そして音素継続時間の始点では，基本口形状を構成する格子点のベクトル移動量の重みを 100%とし終点では 0%になるように，また後続する音素についてもベクトル移動量を始点から終点へ 0%から 100%となるようにし，二つのベクトル移動量の加算を行うことにより発話口形状の補間を行った．

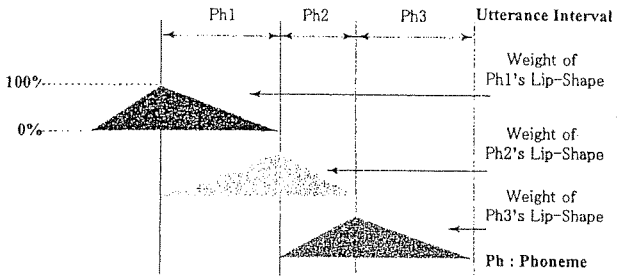


図 7　線形補間の概念図
Fig. 7　Linear interpolation.

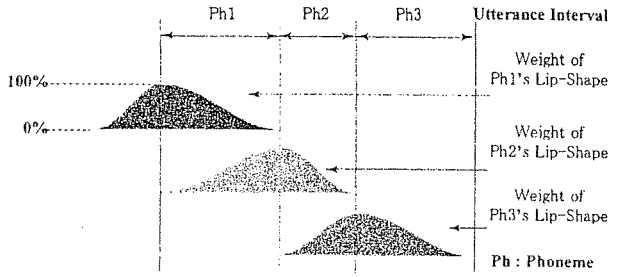


図 8　正弦波補間の概念図
Fig. 8　Sinusoidal interpolation.

本研究では，口形状の補間に線形補間法，正弦波補間法を用いた．両者の補間法の概念図を図 7，図 8 に示す．一般的にこのような補間を行う際には，トライフォンを考慮することで再現性が増すと考えられている [4]．しかしながらここに示すような補間法により，三つの基本口形のベクトル移動量を加算することでトライフォンに対する口形を表現するのは困難であり，演算量も増大するという問題もある．そこで本研究では，後続する音韻のみを補間の対象とした．

2.4.3　合成発話顔アニメーションの生成

2.4.2 で得られた三次元口領域モデルをビデオフレーム画像に重ね合わせることで合成発話顔アニメーションが生成される．しかしながら，単に口領域モデルを画像に重ね合わせただけでは，モデルと画像との間に境界が発生してしまう．これを避けるために，モデルの境界に対して α ブレンディング [9] を施すことにより境界を除去する．こうして得られた口領域モデルをビデオフレーム画像へ重ね合わせることで境界のない自然なアニメーションを作成することができる．

3.　評価対象の作成方法

本章では，実験に用いる評価対象の作成方法について述べる．評価実験に用いる発話顔アニメーションの発話内容には，10 進 4 けたのランダムな数字列を使用した．また実験では，雑音環境下における発話内容の聞き取りを行うため，音声に SNR が 0，−6，−20，−30 dB になるようにホワイトノイズを付加している．具体的な評価実験の方法については，次章以降で述べることとする．

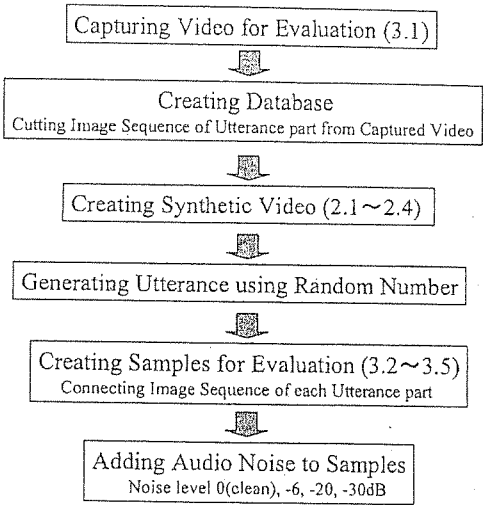


図 9　評価対象動画像の作成手順
(() 内は章，節番号を示す)
Fig. 9　Evaluation sample creation procedure of synthetic talking face.

3.1　評価対象動画像の撮影及び編集

評価対象動画像の作成手順は以下のとおりである (図 9)．

1. 発話の開始を知らせる合図（以後，発話トリガと呼ぶ）となる「番号は」と，0 から 9 までの発話を DV カメラで撮影する．

2. 撮影された動画像から各単語（数字）が発話されている区間の画像シーケンスの切出しを行い，これをデータベースとする．

3. 各単語の画像シーケンスに対して合成発話の動画像を作成し，データベースに保存しておく．

4. 評価対象の発話内容となる 4 けたの数字列を，乱数を用いて生成する．

5. データベースから発話トリガと，発話内容に該当する数字の動画像と音声を取得し，それらを接続する．

6. 5. で生成された動画像に対して，音声に SNR が −6，−20，−30 dB のホワイトノイズを加える．

以上により評価対象となる発話顔アニメーションが作成される．ここで，自然発話顔アニメーションに関しては，自然発話の画像シーケンスを，音声のみの場合は，背景をすべて黒にした画像を用いている．音声のみの場合，背景が黒であることを除いて，自然発話顔アニメーションと同様の手法により作成される．

3.2 自然動画像の撮影

まず，自然動画像として，正面から被験者が発話している様子を DVCAM (SONY DSR-PD-150) を用いて撮影した．このときフレームレート 29.97 [fps]，サンプリングレート 48 [kHz] である．発話内容は，「番号は」と 0 から 9 までの数字の離散発話である．また，カメラ～被写体間の距離は 3 m とし，照明は被写体の顔に影ができないよう一定とした．このときの撮影環境を図 10 に示す．ここで，撮影された動画像に対して，あらかじめ単語（数字）単位に切出しを行っておく．これをデータベースとして，以降で説明する評価対象の作成を行った（図 11）．

3.3 自然発話顔アニメーションの作成

評価対象として用いる自然発話顔アニメーションは，発話内容となる 4 けたの数字列を，乱数により生成後，データベース中の画像シーケンスと音声とを，数字列に従って各単語（数字）単位で接続することにより作成した（図 11）．例えば，図 11 において，数字列が「0941」となった場合，まずはじめに，発話内容の開始を知らせる発話トリガとなる「番号は」，次に，「0」「9」「4」「1」と順に画像シーケンスと音声が接続されていく．ここで，接続された映像フレーム及び音声の境界付近における不連続性が問題となる．そこで，映像フレーム間の境界における映像の不連続性を軽減するため，撮影時に被験者に極力顔以外の部分を動かさないよう，そして単語発話後は必ず口を閉じるように依頼した．更に，境界間の二つの映像フレームに対して α ブレンディングにより映像フレームを混合し，生成された画像を新たなフレーム画像とした．本研究では，画像境界の前後 2 フレームを α ブレンディングの対象とした．

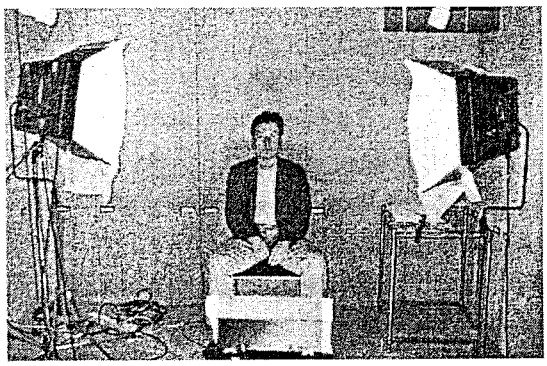


図 10 撮影風景
Fig. 10 Environment of capturing original video.

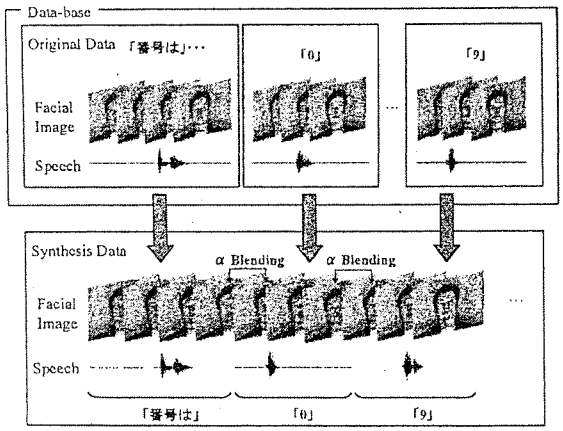


図 11 評価対象の作成
Fig. 11 Making sample of evaluation.

また，音声の境界については聞き取り実験の際に雑音が付加されるため，特別な処理は行わず，数字列に従い単に音声波形を接続した．

3.4 合成発話顔アニメーションの作成

評価実験の際に用いる合成発話顔アニメーションは，データベース中の各画像シーケンスに対して **2.** の手法により作成され，データベース中に保存される．そして乱数により数字列が生成された際に，**3.2** と同様の手法で画像シーケンスと音声が接続されることで合成発話顔アニメーションが作成される．また実験の際に，被験者に合成であることに気づきにくくさせるため，発話トリガとなる「番号は」に関しては，自然発話の画像シーケンスを用いた．

3.5 不一致発話顔アニメーションの作成

音声と発話顔アニメーションの同期の重要性を検証するために，音声と発話口形との同期が，一定時間ずらされた発話顔アニメーションの作成を行った．本論文では，今後この発話顔アニメーションを不一致発話顔アニメーションと呼ぶことにする（本論文では，時間的な同期の不一致にのみ言及することとする）．不一致発話顔アニメーションは，自然・合成発話顔アニメーションにおける音声と発話口形との同期を，Adobe Premiere を用いて ±33 ms 間隔で ±500 ms までずらすことにより作成した．

4. 実験 1：合成発話顔アニメーションの再現性の評価

合成発話顔アニメーションの発話口形が実際の口形に対して正確に再現できているかどうかについて評価実験を行った．被験者は，大学生 20 代男性 13 名である．

4.1 実験方法

自然音声 (0 dB) 及び，−6 dB，−20 dB，−30 dB のホワイトノイズを付加した音声と，自然発話，及び 2 種類の口形状補間法を用いて作成した合成発話顔アニメーションと，全体が黒の画像（音声のみの状況を意図的に作り出すため）を被験者に提示し，雑音に埋もれた音声の発話内容の聞き取り実験を行った．

4.2 考　　察

合成発話顔アニメーションの発話口形が実際の発話口形と比較して正確に再現できているか考察する．結果を図 12 に示す．ここで，被験者が，発話内容である 4 けたの数字列のうちいくつ聞き取ることができたかを表す度合として，数字識別率を定義する．図中，横軸は提示した発話顔アニメーションを，縦軸は数字識別率を表している．なおエラーバーは，数字識別率 ± 標準偏差を表す．

図 12 について，有意水準 5%で検定を行った結果，SNR が 0 dB，−6 dB のとき，音声のみの場合と自然・合成発話顔アニメーションを提示した場合との間に有意差はなく，SNR が −20 dB，−30 dB のとき有意差が存在した．このことから，音声とともに発話顔アニメーションを提示することで，音声の明りょう度が改善されていることがうかがえる．特に，SNR が −30 dB の場合，被験者にはほとんど雑音しか聞こえていないが，発話顔アニメーションが存在することにより，自然発話の場合において 70%程度の改善がなされている．これは，発話内容が 10 進 4 けたの数字に限定されていて，かつ提示されている発話顔アニメーションから読唇により発話内容を推定し，認識した結果であるといえる．つまり，提示されている発話顔アニメーションの再現性が高ければ，音声の明りょう度も改善され，結果数字識別率が高くなるということである．したがって，合成発話アニメーションの数字識別率が自然発話に近いならば，合成発話顔アニメーションの口形の再現性も高いといえるであろう．

ここで，正弦波補間法，線形補間法 [9] により作成された 2 種類の発話顔アニメーションについて検定を行った結果，SNR が 0 dB，−30 dB については有意差はなく，−6 dB，−20 dB については有意差が存在した．全体的には正弦波補間法が数字識別率が高い結果となったが，自然発話の場合と比較すると，いまだに 50%近い差があり，再現性の性能の改善が必要であると考えられる．

5. 実験 2：合成発話口形の自然性の評価

4. の実験では，発話顔アニメーションの再現性のみを評価した．ここでは，合成発話顔アニメーションにおける発話口形の自然性について評価する．

5.1 実験方法

4. での実験の際に，自然・合成発話顔アニメーションに対して，視覚的な自然さ，口の動きの滑らかさについて「非常に良い，良い，普通，悪い，非常に悪い」の 5 段階評価を行った．評価実験の被験者は，大学生 20 代男性 13 名である．

5.2 考　　察

発話顔アニメーションの自然性について考察する．

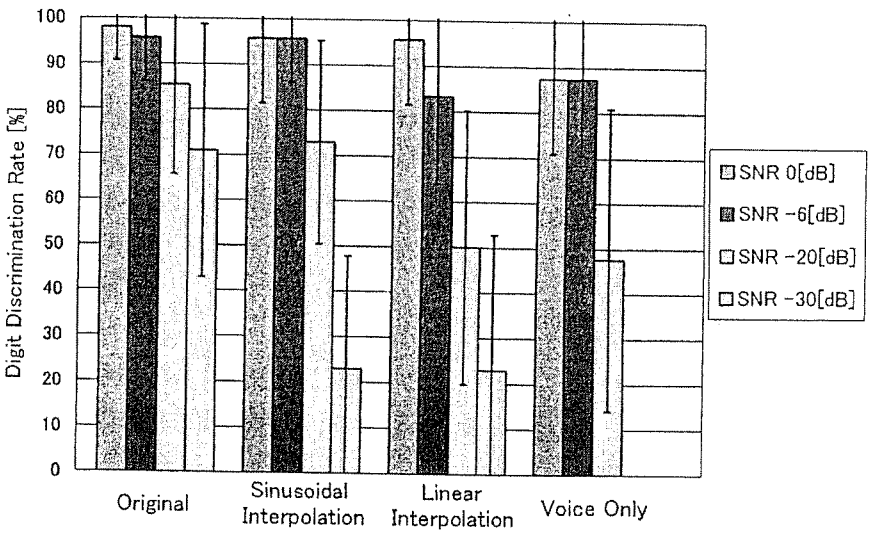


図 12　合成発話顔アニメーションの再現性評価
Fig. 12　Evaluation of representation synthetic talking face.

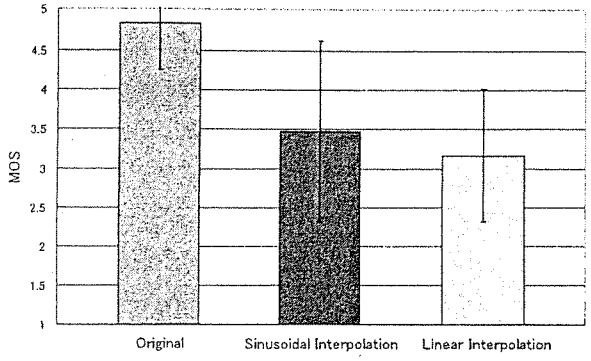


図 13 発話顔アニメーションの視覚的な自然さ
Fig. 13 Naturalness evaluation of synthetic talking face.

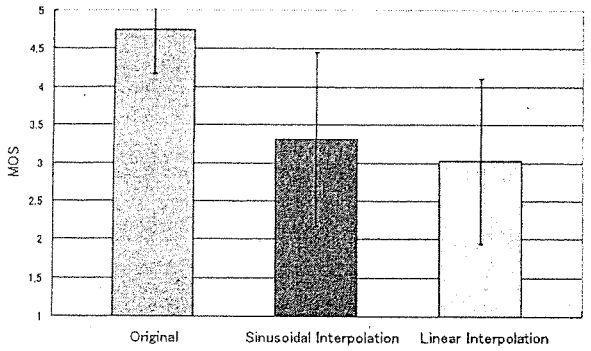


図 14 発話口形の滑らかさ
Fig. 14 Smoothness evaluation of mouth movement.

図 13 に発話顔アニメーションの視覚的な自然さ，図 14 に発話口形の滑らかさに対する評価結果を示す．図中縦軸は MOS 値を，横軸は提示している発話顔アニメーションを表す．またエラーバーは MOS 値 ± 標準偏差を表す．

図 13，図 14 に対して有意水準 5%で有意差検定を行ったところ，自然発話顔及び 2 種類の合成発話顔アニメーションとの間には，各々有意差が存在した．よって，合成発話顔アニメーションは，視覚的な自然さ及び発話口形の滑らかさの点において自然発話よりも劣る結果となった．原因には，口の動きが自然発話の場合と比較して不自然であること，口内モデルや舌モデル，歯モデルの視覚的な不自然さなどが挙げられる．特に後者は，顔モデルに張られているテクスチャが実際の人物のものであるため，顔モデルと口内モデルとの視覚的な不整合が不自然さを強調しているものと考えられる．

また，2 種類のキーフレーム補間法である正弦波補間法と線形補間法とを比較したところ，正弦波補間法がやや自然であるとの回答が得られた．これは線形補間法が実際の人間の口の動きと比較するとやや機械的であるのに対し，正弦波補間法が人間の口の動きに近い表現が行えているためであると考えられる．

6. 実験 3：発話顔アニメーションの同期性の評価

音声と発話口形との同期が正確かつ自然にとられているのか検証する．同期には，音韻的な同期と時間的な同期とが存在するが，実験で用いる発話内容の音韻情報が既知であることを考慮して，ここでは時間的な同期についてのみに言及することにする．

6.1 音声と発話顔アニメーションの主観的な同期のずれ

6.1.1 実験方法

自然音声と，±30 ms 間隔で ±500 ms まで同期がずらされた自然・合成の不一致発話顔アニメーションを被験者に提示し，被験者にどの程度同期がずれていたか主観値で回答させた．ここでは，音声よりも口の動きの早い場合を＋，遅い場合を－とした．実験の被験者は，大学生 20 代男性 8 名及び社会人 20 代男性 1 名である．

主観値は，自然な同期状態を 0，音声よりも口の動きが早く，同期のずれが +500 ms の場合を +100，逆に音声よりも口の動きが遅く，同期のずれが −500 ms の場合を −100 とし，被験者に自然な同期からのずれを主観値により回答させた．

6.1.2 考察

図 15 に，音声と発話顔アニメーションの主観的な同期のずれに対する結果を示す．各点は，発話顔アニメーションにおける被験者の回答の分布を，曲線は，分布に対する二次の最小二乗曲線を表している．

図 15 に注目すると，音声よりも口の動きの方が早い場合においては，被験者にはさほどの違和感がなく，実際の同期のずれよりも小さいと感じている．これに対して，音声が口の動きよりも早い場合は，実際の同期のずれとほぼ同じ値となっている．これは人間が発話するときには，調音器官が開口等の動作をした後に音声が発せられるという事実に起因するものと考えられる．実際，+66 ms 付近に同期タイミングがあると被験者は感じており，これを裏づけている．

6.2 音声と発話顔アニメーションの非同期による違和感

音声と発話顔アニメーションとの同期がずれた場合に，人間の知覚に与える影響について検証した．実験の被験者は，大学生 20 代男性 17 名及び社会人 20 代

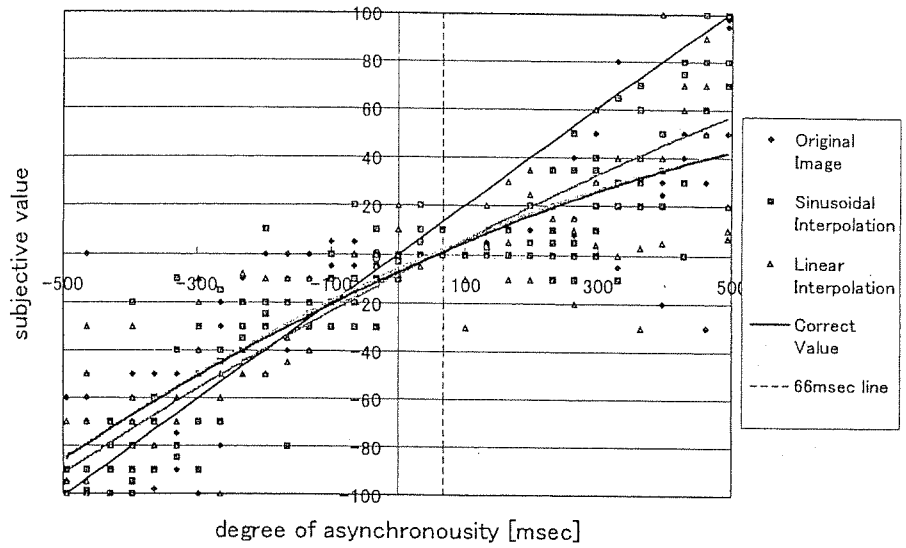


図 15　音声と発話顔アニメーションの主観的な同期のずれ
Fig. 15　Subjective asynchronousity between audio and visual-speech.

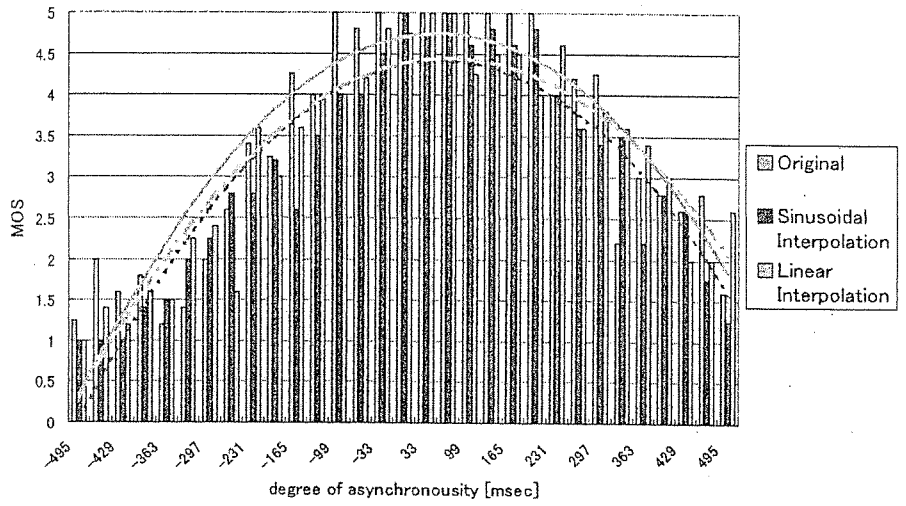


図 16　音声と発話顔アニメーションの同期のずれによる違和感
Fig. 16　Sense of incongruty by asynchronous between audio and visual-speech.

男性 2 名である．

6.2.1　実 験 方 法

実験には前節と同様に，自然音声と自然・合成の不一致発話顔アニメーションを用い，提示された評価対象が自然であるかどうか「非常に良い，良い，普通，悪い，非常に悪い」の 5 段階評価を行った．

6.2.2　考　　察

結果を図 16 に示す．図中横軸は音声と発話顔アニメーションとの同期のずれを，縦軸は MOS 値を表す．また，曲線は MOS 値に対する最小二乗曲線を表す．図 16 に着目すると，同期のずれに対して感じる違和感の程度にはばらつきがあり，図 15 のように，発話顔アニメーションと音声との同期が一番自然にとられていると感じる明確な同期位置を決定することが困難な結果となった．しかしながら，図中の曲線のように音声よりも口の動きが早い場合に，より違和感を感じにくいという傾向が得られた．今後より多くの被験者に対して実験を行うことで，発話顔アニメーションと音声との同期が最もよくとられている位置を模索していく予定である．

次節では，図 16 において被験者が「非常に良い，普通，非常に悪い」と回答している五つの同期タイミングについて着目し，同期のずれが音声認識に及ぼす影響について検証する．

6.3　同期のずれが音声認識に及ぼす影響

音声と発話顔アニメーションとの同期がずれたとき

に，人間の音声認識に及ぼす影響について検証した．

6.3.1 実験方法

前節の結果から，被験者が「非常に良い，普通，非常に悪い」と回答している五つの同期タイミング (−462, −297, +66, +396, +495 ms) について，4. と同様の実験を実施した．ただし，雑音レベルは読唇による効果が生かされ，かつ音声を聞き取れる雑音レベルとして，SNR が −20 dB のときに限定して実験を行った．

6.3.2 考察

図 17 に，図 16 において被験者が，「非常によい，普通，悪い」と回答している五つの点 (−462, −297, +66, +396, +495 ms) における数字識別率を示す．

図 17 に着目すると，発話顔アニメーションよりも音声が早い方が，特に非同期の影響を受けやすく，自然発話の場合においてそれが顕著となった．これは合成発話の場合，それが合成されたものであると分かると無意識に音声の方に重点をおいて聞き取りを行ったためであると考えられる．逆に音声よりも口の動きの方が早い場合に着目すると，非同期の影響を受けにくいことが判明した．

以上の結果から，音声よりも口の動きを数十 ms 程度早くすることにより，自然な発話スタイルに近い発話顔アニメーションが合成できると考えられる．また，発話顔アニメーションを合成する場合に限らず，音声と顔画像を用いて音声認識を行う場合においても音声と発話顔アニメーションの間の非同期性を考慮に入れることにより，認識率が上がる可能性があるということがいえる．

7. 合成発話顔アニメーションの品質評価

以上の考察から，合成発話顔アニメーションの品質を評価する．そして，提案手法である，2 種類の口形状補間法のうち線形補間法と正弦波補間法のどちらが適切であるかを検討する．

線形補間法は，キーフレームのみ 100%の口形を表すため，キーフレームがフレームレートの間隔と一致しない場合，基本口形を 100%表すフレームが存在しないという問題点がある．このため，正弦波補間法の場合と比較して口の動きが読み取りづらいと感じたと考えられる．図 12 に着目すると線形補間法が正弦波補間法よりも，5〜10%程度聞き取り率が下回っている．このため線形補間法よりも正弦波補間法の方が口形状の再現性が高くなっている．

一方で正弦波補間法は，基本口形を 100%表すフレームを線形補間法よりも長く持続させることができるが，後続する口形への変化が線形補間法と比較して急しゅんになる．しかしながら，人間の口の動きは，非線形な動きをするため，多くの被験者には正弦波補間法の方が自然であると感じたと考えられる．図 13，図 14 に着目すると顔画像の自然性，口の動きの滑らかさにおいて正弦波補間法が線形補間法よりも 0.3〜0.5 点程度上回る結果となり，現状では正弦波補間法により作成した合成発話顔アニメーションが品質が高く，キーフレーム補間法として妥当であるといえる．

しかしながら，これらの評価における両者の差はわずかであり，自然発話顔アニメーションと同程度の発

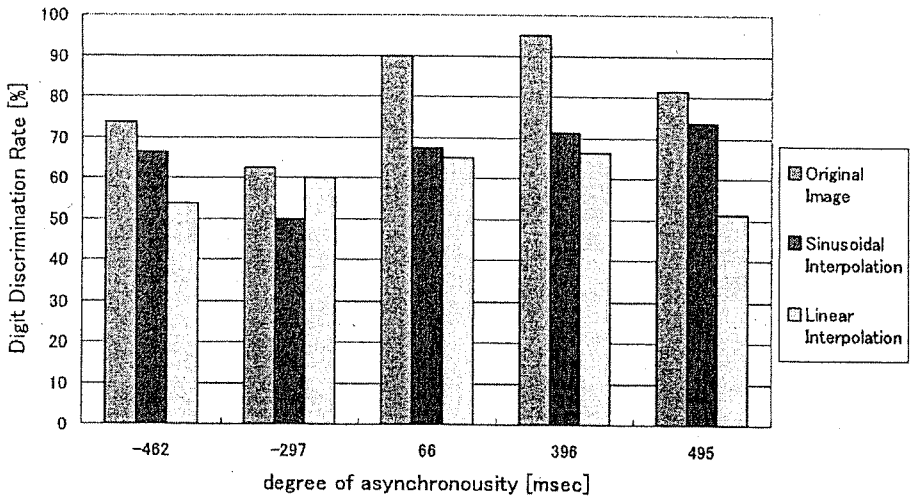


図 17 同期のずれが音声認識に与える影響

Fig. 17 Effect which asynchronous between audio and visual-speech on speech recognition.

話顔が合成できていないというのが現状であるといえる．今後，より自然な発話顔アニメーションを合成する手法を検討する必要がある．

8. むすび

本論文では，雑音環境下において音声とともに発話顔アニメーションを被験者に提示し，音声の明りょう度や，発話顔アニメーションの視覚的な自然さ，音声と発話顔アニメーションの同期に対して検証することで，筆者らの提案する合成発話顔アニメーションの評価を行った．

評価結果より合成された発話口形の再現性は，いまだに自然発話のものよりも低く（約50%），自然性においてもやはり自然発話口形との差はあるといえる．今後，再現性の高い発話口形を生成する手法を検討する必要がある．また，音声と発話顔アニメーションとの同期について検証した結果，発話顔アニメーションを合成する場合には，音素の開始と同時に音素に対応する口の動きを始めるのではなく，数十 ms 程度口の動きを早くすることで，合成発話顔アニメーションの自然さを改善できると考えられる．これは音声と発話顔画像を用いて音声認識を行う上でも同じことがいえる．

本評価実験では，数字列が聴取者にとって馴染みのある単語であり，次の数字への出現確率が一定で予測が不可能であることから，無意味単語での評価と同類であるとして，発話内容に数字を用いている．今後は，発話内容として一般文章を用いた大規模実験を検討している．また，**6.2** で示した音声と発話アニメーションとの非同期による影響について，被験者を増やして最も自然な同期位置を検討し，加えて，本論文では考慮の対象としなかった音声と発話顔アニメーションとの音韻的な同期についても今後検証していく予定である．

ここで用いた評価方法は，主に筆者らの提案手法のような客観評価の困難な個人性の除去された特徴から合成された発話顔アニメーションに対して有効であると考える．更に，合成発話顔アニメーションの評価だけでなく，例えば顔画像の表情と音声のモダリティ間の同期が失われた場合，被験者の受ける印象はどのように変化するのか，また合成音声とともに顔画像を提示することにより，単に SNR では表すことのできない合成音声の印象や明りょう性の評価等にも応用が可能であると考える．

謝辞 本研究は，通信放送機構の研究委託により実施したものである．また，本評価実験に参加して下さった皆様に心より感謝致します．

文　献

[1] 川本真一，下平　博，新田恒雄，西本卓也，中村　哲，伊藤克亘，森島繁生，四倉達夫，甲斐充彦，李　晃伸，山下洋一，小林隆生，徳田恵一，広瀬啓吉，峯末信明，山田篤，伝　康晴，宇津呂武仁，嵯峨山茂樹，“擬人化音声対話エージェントツールキットの基本設計，”情処学音声言語情報処理研報，2002-SLP-40-11, pp.61–66, Feb. 2002.

[2] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” Nature, vol.264, pp.746–748, 1976.

[3] 酒向慎司，徳田恵一，益子貴史，小林隆夫，北村　正，“HMM に基づいた視聴覚テキスト音声合成—画像ベースアプローチ，”情処学論，vol.43, no.7, pp.2169–2176, 2002.

[4] 垣原清次，中村　哲，鹿野清宏，“HMM を用いた自然な発話動画像合成，”信学論（D-II），vol.J83-D-II, no.11, pp.2498–2506, Nov. 2000.

[5] T. Kuratate, H. Yehia, and E.V. Bateson, “Kinematics-based synthesis of realistic tracking face,” Proc. International Conference on Auditory-Visual Speech, Processing, AVSP ’98, pp.185–190, 1998.

[6] H.P. Graf, E. Coatto, and T. Ezzat, “Face analysis for the synthesis of photo-realistic talking heads,” Proc. 4th International Conference on Automatic Face and Gesture Recognition, pp.189–194, 2000.

[7] E. Yamamoto, S. Nakamura, and K. Shikano, “Subjective evaluation for HMM-based speech-to-lip movement synthesis,” AVSP ’98, pp.225–230, 1998.

[8] T. Ezzat, G. Geiger, and T. Poggio, “Trainable videorealistic speech animation,” ACM SIGGRAPH, pp.388–398, 2002.

[9] 緒方　信，森島繁生，中村　哲，“ビデオ翻訳システム—自動翻訳合成音声とモデルベースリップシンクシステムの実現，”情報処理学会シンポジウム，インタラクション2001 論文集，vol.2001, no.5, pp.203–210, 2001.

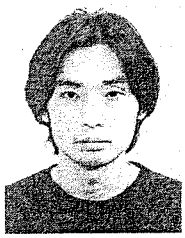
[10] 伊藤　圭，三澤貴文，武藤淳一，森島繁生，“仮想空間上におけるリアルな三次元口形状の作成，”2000 信学総大，A-16-24, p.328, 2000.

[11] T. Misawa, K. Murai, S. Nakamura, and S. Morishima, “Automatic face tracking and model-match-move in video-sequence using 3D-face model,” IEEE International Conference on Multimedia and Expo., pp.349–352, Aug. 2001.

[12] S. Morishima and S. Nakamura, “Multi-modal translation system and its evaluation,” Proc. ICMI ’02, pp.241–246, 2002.

[13] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, The HTK Book, Microsoft Corporation, 2000.

（平成 16 年 2 月 25 日受付，7 月 6 日再受付，
9 月 2 日最終原稿受付）



前島　謙宣

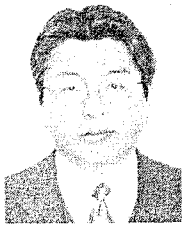
平 14 成蹊大・工卒．平 16 同大大学院修士課程了，平 16 早大大学院博士課程入学，現在に至る．ヴァーチャルヒューマンの構築に関する研究に従事．

四倉　達夫　（正員）

平 10 成蹊大・工卒．平 12 同大大学院修士課程了，平 15 同大学院博士課程了，同年（株）ATR 音声言語コミュニケーション研究所に入所，現在に至る．工博．ヴァーチャルヒューマンの構築，マルチモーダル音声認識に関する研究に従事．平 12 年度本会学術奨励賞受賞．

森島　繁生　（正員）

昭 63 東大・工・大学院博士課程了，工博．平 13 成蹊大工学部教授．平 6 から 1 年間，トロント大学客員研究員．平 8 から 5 年間，通信放送機構プロジェクトサブリーダ．明治大学非常勤講師．成蹊大学非常勤講師．（株）ATR 音声言語コミュニケーション研究所，同メディア情報科学研究所客員研究員．平 4 本会業績賞受賞．平 16 早稲田大学理工学部応用物理学科教授．



中村　　哲　（正員）

昭 56 京工繊大・工芸・電子卒．昭 56～平 6 シャープ（株）中央研究所及び情報技術研究所に勤務．昭 61～平元 ATR 自動翻訳電話研究所に出向．平 4 京都大学博士（工学）．平 6～12 奈良先端科学技術大学院大学情報科学研究科助教授．平 8 年 3 月～8 月 Rutgers University CAIP Center Visiting Research Professor．平 12 年 4 月より ATR 音声言語通信研究所及び音声言語コミュニケーション研究所第一研究室長．平 14 年 4 月より豊橋技術科学大学，及び立命館大学客員教授．平 16 年 1 月よりドイツ・カールスルーエ大学客員教授．音声認識，多次元信号処理などの音声・音響情報処理，マルチモーダル情報処理，ヒューマンインタフェースの研究に従事．平 4 日本音響学会粟屋学術奨励賞，平 13 インタラクション 2001 ベストペーパ賞受賞．IEEE，情報処理学会，日本音響学会，人工知能学会各会員．平 13.4～15.3 本会英文論文誌 (ED) 編集幹事，平 13 より IEEE Speech Technical Committee 委員，情報処理学会音声言語情報処理研究会主査．

Dynamic Micro Aspects of Facial Movements in Elicited and Posed Expressions Using High-Speed Camera

Shigeo Morishima[1], Tatsuo Yotsukura[1,2], Hiroshi Yamada[3],
Hideko Uchida[4], Nobuji Tetsutani[2], and Shigeru Akamatsu[5,6]

1 Seikei University, 2 ATR MI&C, 3 Nihon University,
4 San Francisco University, 5 Hosei University, 6 ATR International

3-3-1 Kichijoji-Kitamachi Musashino-shi, Tokyo, 180-8633 Japan
E-mail: shigeo@ee.seikei.ac.jp

Abstract

The present study investigated the dynamic aspects of facial movements in spontaneously elicited and posed facial expressions of emotion. We recorded participants' facial movements when they were shown a set of emotional eliciting films, and when they posed typical facial expressions. Those facial movements were recorded by a high-speed camera of 250 frames per second. We measured facial movements frame by frame in terms of displacements of facial feature points. Such micro-temporal analysis showed that, although it was very subtle, there exits the characteristic onset asynchrony of each part's movement. Furthermore, it was found the commonality of each part's movement in temporal change although the speed and the amount of each movement varied along with expressional conditions and emotions.

1 Introduction

Movements of the face and body are important components of nonverbal communication. Most of the psychological literature on facial expression has relied on observers' inferential judgments about what emotion is shown in facial behavior, and many of those studies were focused on still photographs of posed facial expressions. It is much more time-consuming to measure facial behavior itself with the Facial Action Coding System (FACS) or other technique than to obtain inferences about emotion from observers. Although it is extremely labor intensive to measure facial behavior itself, it is much more interesting and meaningful to study facial expressions that occur unintentionally instead of still images of posed facial expressions.

A handful of recent studies have asked participants to make judgments of facial expressions of morphed photographs that feature linear facial movements [1]-[4]. These studies raised a number of interesting points, and yet those stimuli were created under assumptions of linear movements. Since actual human facial movements are nonlinear, future studies need to utilize stimuli of linear movements.

The purpose of this study was to investigate dynamic aspects of facial movements between "posed" (intended) facial expressions and "elicited" (unintended) emotional responses. Participants' facial movements were recorded via a high-speed video camera (NAC, HSV-500c3, 250 frames per sec.), which allowed us to analyze facial movements very closely in image sequences. These movements cannot be seen with a regular video camera (30 frames per sec.). We also simulated facial synthesis by using results from an analysis. The animations that we produced confirmed differences in the intensity of the facial expressions.

2 Method

2.1 Participants

Twenty-four subjects (12 Japanese female, 12 Japanese male) were selected from a pool of eighty-four candidates. Participants ranged in age from 20 to 30 years.

The study's inclusion criteria was as follows: (a) be born in Japan; (b) have both parents be Japanese; (c) be between age 20 and 30 years; (d) have eyesight better than at least 1.0 without glasses; and (e) have experience in theater or modeling.

2.2 Apparatus

Figure.1 illustrates the apparatus used in the study. To provide a video recording of facial expressions, a high-speed video camera (NAC, HSV-500c3) was hidden behind a 21-inch prompter. Images of the high-speed video camera were recorded onto an S-VHS videocassette. The recording speed was 250 frames per second, and a shutter speed was 250/1 second. The camera faced the participants to capture a frontal view of the face. These images were shown on a 21-inch color television monitor, which was located in the back of the room. All of the equipment was placed behind a partition so that the

experimenter was hidden from view except when giving instructions. Thus, the subject could not see the images that the experimenters were videotaping.

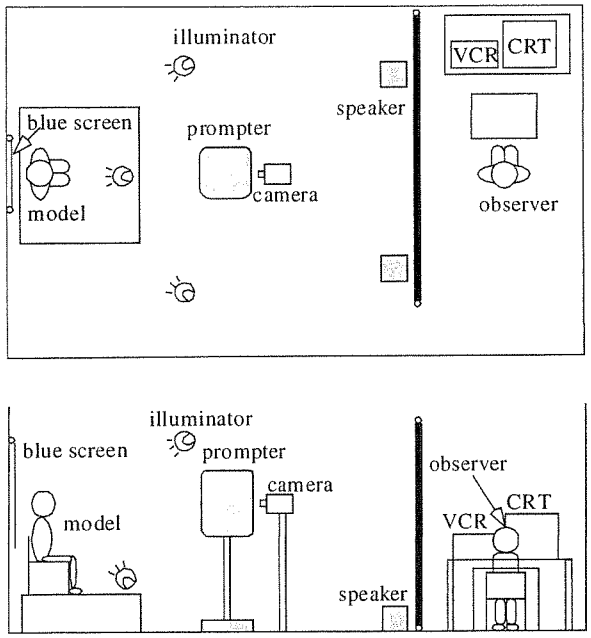


Figure 1. Apparatus

The film stimuli used in this study were adapted from Gross & Levenson's set of standardized emotional film stimuli [5]. Nine out of twelve film clips were excerpted from commercial films, and the rest were developed from non-commercial sources, which were obtained from Gross & Levenson. Each film clip previously had been validated as an elicitor of a target emotional state [5][6]. These film stimuli were originally produced in English; however, Japanese subtitles were provided for participants in this study. The order of presentation of these films was systematically varied and edited so that the films' embedded neutral film stimuli appeared between each emotion film clip (Table.1).

Table 1. Film clip

Film item	Target emotion	Clip length
1. Waves	Contentment	**00:59**
2. Beach	Contentment	**00:23**
3. Sea of Love	Surprise	**00:27**
4. The Champ	Sadness	**02:59**
5. The Shining	Fear	**01:24**
6. Pink Flamingos	Disgust	**00:38**
7. Silence of the Lambs	Fear	**03:35**
8. Cry Freedom	Anger	**02:37**
9. Amputation	Disgust	**01:04**
10. Capricorn One	Surprise	**00:57**
11. When Harry Met Sally	Amusement	**02:53**
12. Bambi	Sadness	**03:22**

Ekman & Friesen [7][8] developed the facial action coding system (FACS), which is an objective method for quantifying facial movements. FACS is an anatomically based coding scheme that codes the facial muscular movements in terms of 44 action units (AUs) or action unit combinations (AU combinations).

Participants were instructed to perform the six basic facial expressions of emotion (happiness, surprise, fear, anger, sadness, and disgust) based on combinations of AUs. Table.2 shows a list of Action Unit for creating a posed expression for happiness.

Table.2 Combination of expression for "happiness"

AU No.	AU name
AU12	Lip corner puller
AU16	Lower lip depressor
AU25	Lips part

2.3 Procedure

Experiments were conducted individually in an experiment room at ATR Human Information Research Laboratories. The experiment consisted of three sessions: (1) a Training Session; (2) a Film Session; and (3) a Posed Expressions Session. The order of the sessions was not counterbalanced in order to eliminate the possibility of an initial Posed Expression Session biasing participants' responses in the Film Session.

1) Training Session

The training session was conducted to accustom participants to the experimental setting, to ease their tension, and to relax their facial muscles prior to the film session. The Training Session lasted between 45 and 60 minutes. The single AUs (e.g., AU 12, lip corner puller) listed on the guide sheet were described and demonstrated, and the image cues were reviewed and indicated on the example images excerpted from Ekman & Friesen [8][14]. The participants' task was to produce the facial muscular movements indicated in the image on the sheet. During this session, participants were able to view their faces on the 21-inch television monitor (inside a prompter) and a hand mirror. The experimenters checked participants' facial movements for accuracy, and provided feedback when necessary.

2) Film Session

For the second session, the participants viewed Gross & Levenson's set of film stimuli, which elicits the emotional states, on the 21-inch color television monitor through the prompter at a distance of 1 m [5]. The participants viewed this film alone in the experiment room. The Film Session lasted for approximately 50 minutes. The experimenters recorded the participants' facial expressions to the film stimuli. The participants were given a lunch break at the end of the session.

3) Posed Expressions Session

For the third session, the procedures were the same as those of the Training Session. First, the experimenters reviewed the single AU listed on the guide sheet. The participants' task was to produce the facial muscular movements indicated in the images on the sheet. The participants were allowed to take as much time as they needed to perform the task. After practicing single AUs (e.g., AU 26, jaw drop), the experimenters instructed participants to perform combinations of AUs (e.g., AU 1 + 2, brow raise) by emotion [8]. To eliminate possible participant' bias to certain emotion terms, the experimenters did not use emotion terms such as happy, disgust, and surprise while giving instructions. The participants were instructed to make a neutral face, then an emotional expression involving a high-magnitude muscle contraction, then to relax.

The participants were instructed to repeat this task until the experimenters told them to stop. The participants were given a short break when needed. This session lasted for an average of 4 hours.

3 Results

For the present study, only images of disgust, happiness, and surprise expressions via the high-speed camera included in this study were analyzed.

Feature Point Tracking. The authors developed a Graphical User Interface (GUI) tool, a feature point tracking tool, to analyze the facial movements of emotion in this study. The images were downloaded onto a computer. Twenty-eight dots were manually plotted on the face (4 dots for an outline of the eyebrows, 4 dots for an outline of both eyes, 5 dots for an outline of the nose, 6 dots for an outline of the mouth, and 1 dot at the corner of the jaw) in the initial image (Figure.2). Once the 28 dots were plotted, the dots remained on the next frame. The experimenter then manually moved the dots as facial movements occurred. The authors selected these twenty-eight dots because they were necessary to recreate the movements of basic facial expressions and to design computer graphics (CG) programs that aim to simulate natural facial movements.

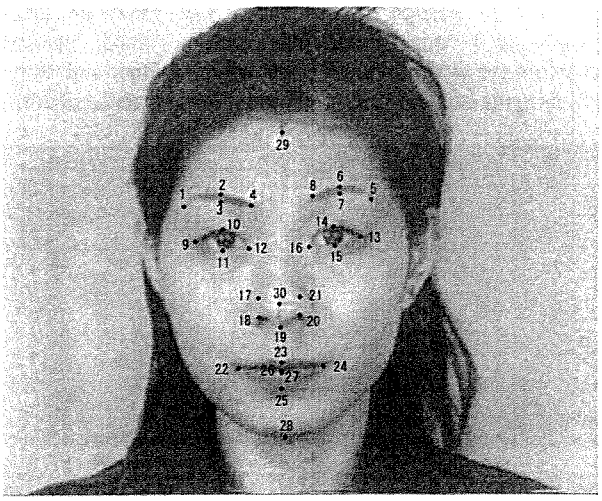


Figure 2. Feature points

We examined patterns of facial movements by measuring the movements of the feature points. We analyzed the feature points by facial region (i.e., eyebrows, eyes, and mouth). We computed the mean ratings of each region per emotion for all subjects (Table 3). Overall, our results indicated that the mean ratings of the eye region were higher than those of the eyebrows and mouth regardless of the emotion category or experimental conditions.

Table 3. Mean ratings of each region per emotion

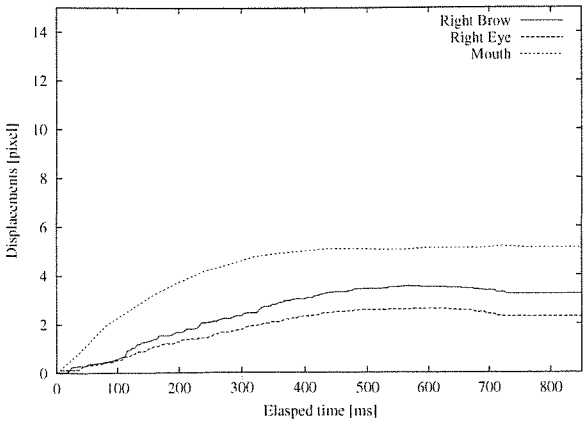
Expression	Condition	1st	Time difference	2nd	Time difference	3rd
Happiness	Elicited	eye	23.3ms	brow	26.1ms	mouth
	Posed	eye	9.2ms	brow	18.5ms	mouth
Surprise	Elicited	eye	5.6ms	mouth	6ms	brow
	Posed	eye	12.8ms	mouth	15.9ms	brow
Disgust	Elicited	brow	0.2ms	mouth	6.4ms	eye
	Posed	eye	11.6ms	mouth	4.2ms	brow

Next, we compared the total duration (the time from neutral, the start of the expression, and to reach the peak of intensity) between posed and elicited emotion conditions. Our results demonstrate that the total duration was shorter for the posed expressions for all emotions (Table 4). We also found differences in the magnitude of the facial movements between the posed and elicited emotion conditions. The data showed that the magnitude of the posed expressions was greater than that of the elicited emotion across emotions. In addition, our data revealed that the facial movements were nonlinear as a function of time.

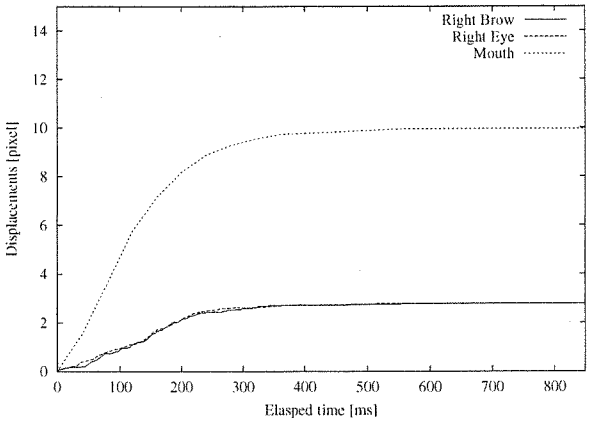
Table 4. Total duration between "Elicited" and "Posed" expression

Expression	Elicited	Posed
Happiness	724ms	556ms
Surprise	228ms	412ms
Disgust	836ms	388ms

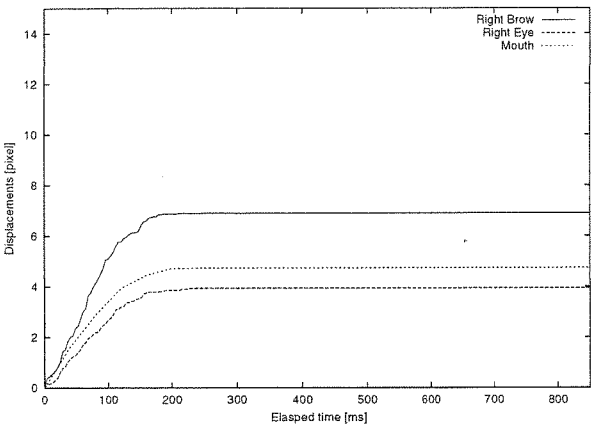
Furthermore, We analyzed the characteristic displacement magnitude of the feature points. Figure.3 a) b) show the result of an expression for "Happiness". The movement of an elicited expression is generally faster then that of a posed expression. One characteristic is that the mouth involves a large and quick movement. With the "Posed" expression, these are hardly any different feature points in terms of magnitude of the eyes and eyebrows. On the other hand with the "Elicited" expression, the motion of the eyebrows is larger than that of the eyes is characterizing.



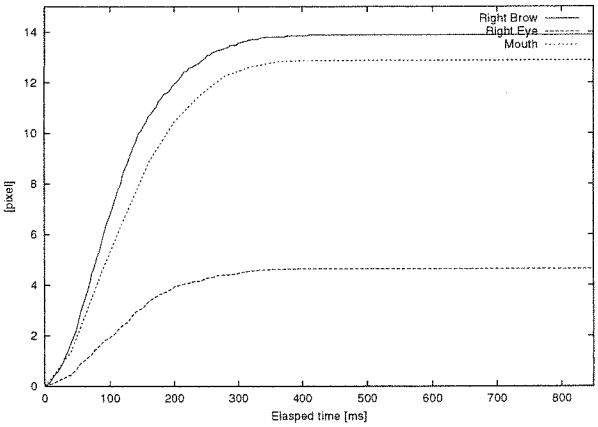
a) Elicited expression of happiness



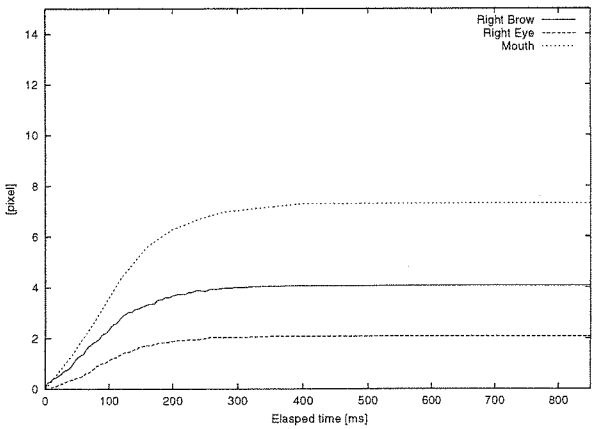
b) Posed expression of happiness



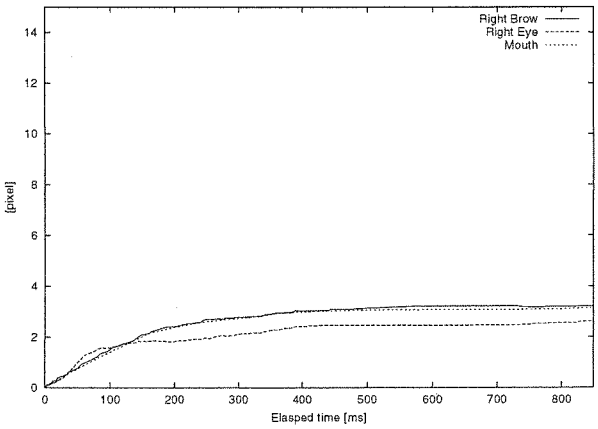
c) Elicited expression of surprise



d) Posed expression of surprise



e) Elicited expression of disgust



f) Posed expression of disgust

Figure 3. Average number of movements by facial part

Figure.3 c) d) graphs are about expression for "surprise". It had already been proven that the motion of the "Elicited" expression is faster than that of the "Posed" expression in this case.

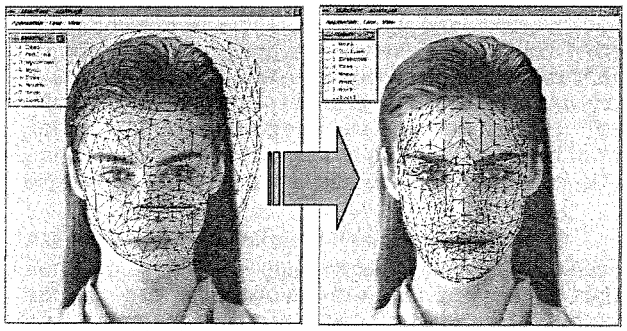
Figure.3 e) f) graphs are about an expression of "disgust". The motion of the "Posed" expression is smaller than that of the "Elicited" expression, and there is no difference between facial parts.

All of these graphs prove that face morphing is not linear. As a general tendency, the amount of change is small when the face moves. Afterwards, the movement becomes. And then, the amount of change decreases before facial movement is finish.

4 Simulation of facial movement

From these results, the next step was to simulate a facial movement. To generate a real synthetic face, a generic face model was manually adjusted to a person's face image (Figure.4). These images show personal models each before and after the fitting process for a front-view image by using our original GUI based face-fitting tool [9].

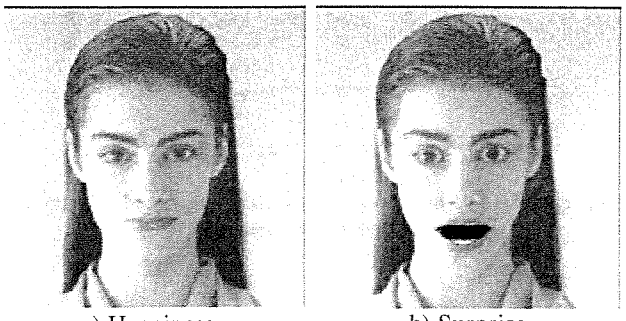
The front view was put into the system and then the corresponding control points were manually moved to reasonable positions by mouse operations. The fitting process was finished by an expert in about five minutes.



a) Initial image b) Fitted image

Figure 4. Face-fitting tool

Expression control rules are defined in the generic model, so every user's face can be equally modified to generate basic expressions using the FACS-based expression control mechanism. Figure.5 show examples of synthesized emotional faces for happiness and surprise.



a) Happiness b) Surprise

Figure 5. Synthesized emotional face

Using the method to generate to real synthesized faces and graphs of the average numbers of movements by facial part, we created three facial animations for the expression of "happiness" (Figure.6). Figure.6 a) is a linear animation and b) is a non-linear animation of "Elicited" expression. The animation used by our technique is more natural than conventional generic animation.

0ms 5ms 10ms 15ms 20ms

a) Linear animation [30 fps]

0ms 5ms 10ms 15ms 20ms

b) Non-linear animation [30 fps]

Figure.6 Reconstructed face: expression of "happiness"

5 Conclusion

Our study was designed to investigate dynamic aspects of the facial movements of posed expressions and elicited emotional expressions. To the best of our knowledge, this study was the first that attempt to examine facial movements using a high-speed camera, which allowed us to measure very subtle movements of facial expressions of emotion. Although the coding of the facial movements was incredibly labor intensive, the findings from this study provide some evidence for dynamic aspects of the movements of facial expressions.

Nishio, Koyama, and Nakamura [10] examined temporal differences in the facial movements of smiles, and they classified smiles into three categories. The findings of happiness are partially consistent with Nishio et al.'s notion that the eyes move prior to the mouth in a smile of unpleasantness.

Hager & Ekman [11] suggest that smiles are more symmetrical when children respond to a joke than when they smile deliberately. The results of happiness or smile showed that facial movements are asymmetrical particularly in the eye region.

The findings partially support the notion that posed expressions are longer than spontaneous expressions. However, not all results were consistent with the previous literature [12][13].

One limitation of our findings was that some of the film clips might not be universal. Since our participants were native Japanese, it is probable that some of the film clips did not elicit the emotions that we intended to measure. Future research needs to test the validity and reliability of the film stimuli of universality. Furthermore, our study is limited in terms of the generalizability of our

findings. Further studies are necessary to continue to examine facial movements during interactions and context influences.

Certainly, further data needs to be collected to test the validity of our findings using a larger sample and participants from cultures other than Japan.

References

[1] Kamachi, M., Yoshikawa, S., Gyoba, J., & Akamatsu, S. The dynamics of facial expression judgments. Paper presented at the meeting of Twenty-second European Conference on Visual Perception, Trieste, Italy. 1999

[2] Kato, T., Saeki, M., Takuma, M., Kamei, M., Mukaida, S., & Akamatsu, S. Sensitivity to subtle visual differences in faces seen in similarity judgment and impression formation. Technical Report of the Institute of Electronics, Information and Communication Engineers, 51, pp25-30. 1999

[3] Yamaguchi, M. & Oda, M. (1999, Cognition of gaze and facial expression: Interaction of detecting gaze and facial expression. Poster session presented at the annual meeting of ATR Symposium on Face and Object Recognition, Kyoto, Japan. July, 1999

[4] Matsumoto, D., Consolacion, T., Yamada, H., Suzuki, R., Franklin, B., Paul, S., Ray, B., & Uchida, H. American-Japanese Cultural Differences in Judgments of Emotional Expressions of Different Intensities. Manuscript submitted for publication. 2000

[5] Gross, J. J. & Levenson, R. W. Emotion Elicitation Using Films. Cognition and Emotion, 9, pp89-108. 1995

[6] Ekman, P., Friesen, W. V., & O'Sullivan, M. Smiles when lying. Journal of Personality and Social Psychology, 54, pp414-420. 1988

[7] Ekman, P. & Friesen, W. V. Facial Action Coding System (FACS): A technique for the measurement of facial movement. Palo Alto, CA: Consulting Psychological Press. 1978

[8] Ekman, P. & Friesen, W. V. Manual for the Facial Action Coding System and Action Unit Photographs. Palo Alto, CA: Consulting Psychological Press. 1978

[9] Morishima, S. Modeling of Facial Expression and Emotion for Human Communication System Display 17. pp15-25, Elservier, 1996

[10] Nishio, S., Koyama, K., & Nakamura, T. Temporal differences in eye and mouth movements classifying facial expressions of smiles. Proceedings of the IEEE Third International Conference on Automatic Face and Gesture Recognition. 1998

[11] Hager, J. C. & Ekman, P. The asymmetry of facial actions is inconsistent with models of hemispheric specialization. In P. Ekman & E. Rosenberg (Ed.), What the face reveals New York, NY: Oxford University Press, Inc. pp. 40-57. 1997

[12] Weiss, F., Blum, G. S. & Gleberman, L. Anatomically based measurement of facial expressions in simulated versus hypnotically induced affect. Motivation and Emotion, 11, pp67-81. 1987

[13] Hess, U. & Kleck, R. E. Differentiating emotion elicited and deliberate emotional facial expressions. European Journal of Social Psychology, 20, pp369-385. 1990

[14] Ekman, P. & Friesen, W. V. Unmasking the face. (Kudoh, T., Trans.). Englewood Cliffs, NJ. 1987.

Analysis and Simulation of Facial Movements in Elicited and Posed Expressions Using High-Speed Camera

Tatsuo Yotsukura
ATR Media Integration & Communications Research Laboratories
/ Seikei University
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 6190288 JAPAN
yotsu@mic.atr.co.jp

Introduction

The purpose of this research was to examine dynamic aspects of facial movements involving "posed" (intended) facial expressions versus "elicited" (unintended) emotional facial expressions. The participants were shown Gross & Levenson's set of standardized emotional film stimuli[1], and their facial expressions to the film stimuli were recorded by a high-speed video camera (250 frames/ sec), which allowed us to analyze facial movements very closely in image sequences. These movements cannot be seen with a regular video camera (30 frames/sec). In addition, the participants were asked to produce the facial expressions of happiness, surprise, and disgust based on the Facial Action Coding System (FACS). The findings suggested that the patterns of facial movements between posed facial expressions and elicited emotional facial expressions are not significantly different, but that there were differences in the intensity of the facial expressiveness.

Method

The subjects ware 24 participants (12 Japanese female, 12 Japanese male). To obtain the video recordings of facial expressions, a high-speed video camera was hidden behind a 21- inch prompter. The film stimuli used in this study were adapted from Gross & Levenson's set of standardized emotional film stimuli[1]. The protocol for Posed Expressions was utilized. Ekman & Friesen's facial action coding system (FACS), which is an objective method for quantifying facial movements was also utilized. FACS is an anatomically based coding scheme that codes facial muscular movements in terms of 44 action units (AUs) or action unit combinations (AU combinations). The participants were instructed to perform six basic facial expressions of emotions based on combinations of AUs.

Analysis and Simulation of Facial Movements

For the present study, only images of disgust, happiness, and surprise expressions were analyzed via the high-speed camera. Feature Point Tracking. The authors developed a GUI tool, a feature point tracking tool, to analyze the facial movements of emotions in this study. The images were downloaded onto a computer. 28 dots were manually plotted on the face (4 dots for an outline of the eyebrows, 4 dots for an outline of both eyes, 5 dots for an outline of the nose, 6 dots for an outline of the mouth, and 1 dot at a corner of the jaw) in the initial image. The authors selected these twenty-eight dots because they were found to be necessary to recreate the movements of basic facial expressions and to design computer graphics programs aiming to simulate natural facial movements.

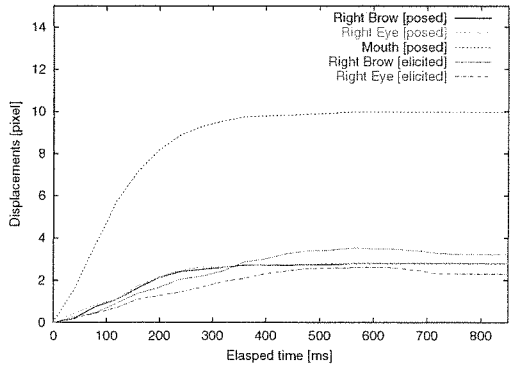


Figure.1 The amount of average movements by each facial part (expression of "Happiness")

We examined patterns of facial movements by measuring the movements of the feature points. We analyzed the feature points by the facial regions (i.e., eyebrows, eyes, and mouth). We computed the mean ratings of each region per emotion for all of the subjects.
Overall, our results indicated that the mean ratings of the eye region were higher than those of the eyebrows and mouth regardless of the emotion category or experimental conditions.Furthermore, we compared the total duration (the time from neutral, the start of the expression, and the attainment of the peak intensity) between posed and elicited emotion conditions. Our results demonstrated that the total duration was shorter for the posed expressions for all emotions. Figure. 1 shows an example of the expression of "Happiness". We also found differences in the magnitudes of facial movements between the posed and elicited emotion conditions. The data showed that the magnitudes of posed expressions were greater than those of elicited emotions across the emotions. In addition, our data revealed that the facial movements were nonlinear as a function of time.

We also simulated facial synthesis using Morishima's system[2]. Figure. 2 shows a reconstracted facial movement of all frames. The generic facial animation was created through linear animation by morphing. The animation used by our technique is more natural than conventional generic animation.

Contributors:Hideko Uchida (San Francisco State Universtiy), Hiroshi Yamada (Nihon Univ.), Nobuji Tetsutani (ATR), Shigeru, Akamatsu (ATR), Shigeo Morishima (Seikei Univ.)

References

1 Gross, J. J. & Levenson, R. W. (1995). Emotion Elicitation Using Films. Cognition and Emotion, 9, 89-108.
2 Morishima, S. (1996). Modeling of Facial Expression and Emotion for Human Communication System. Displays 17, pp. 15-25, Elsevier

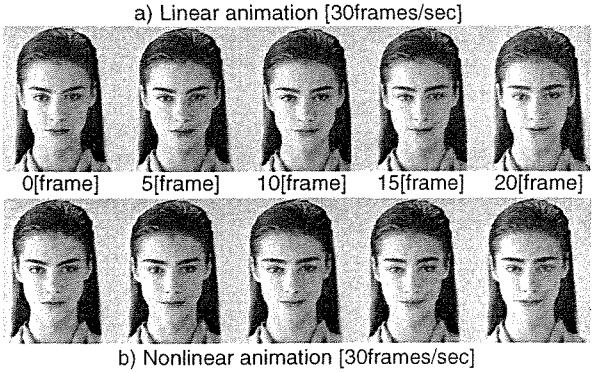


Figure 2. Reconstructed face (expression of "Happiness")

第3章　まとめ・今後の課題

顔表情合成や表情アニメーションのニーズは、エンタテインメント分野や コンテンツ分野などさまざまな分野でますます増加している。しかし、高いリアリティを実現できる一般的な方法論は存在せず、未だ手作業に依存する部分が多く、コンテンツ製作に多大なる手間とコストを発生させていると考えられる。今回、研究対象となった表情筋モデルによる顔表情合成は、まだまだユニバーサルな解をもたらすレベルには至らない。しかし、表情筋の編集機能を付加することで、個人性の表現が可能となり、従来の物理シミュレーションに基づくアプローチにカスタマイズ機能を付加した点で、この分野に与える影響は大きく、貢献度も大きいと判断できる。

また、今回の成果を応用し、 愛・地球博のFuture Cast Systemという現実のエンタテインメントシステムを構築するに至り、博覧会の中でもベスト3のアトラクションに選ばれるほどの人気を博す結果となった。100万人以上にもおよぶ多くの来場者に感動を与えることができた点は、何よりも現在の研究レベルが高い評価を得た証に他ならないと考えられる。

今後は、これらの研究成果をさらに発展させ、コンテンツやエンタテインメントにおいて人を感動させる要因、あるいはCGモデルのリアリティを向上させる手法について、さらに深い検討を進めていく予定である。また、顔表情のみならず、人物全体を研究対象として、感性を豊かに反映させるモーションキャプチャリングの方法や表現方法についても、検討を進めていく予定である。