ヒューマンコンピュータインタラクションのための音声から画像へのリアルタイムメディア変換

A Real-time Media Conversion from Voice to Image for Human-Computer Interaction

宮下 直也 Naoya Miyashita
坂口 竜己 Tatsumi Sakaguchi
森島 繁生 Shigeo Morishima
{shigeo, tatsu, miyasita}@ee.seikei.ac.jp

成蹊大学工学部 電気電子工学科 情報通信研究室
〒180 東京都武蔵野市吉祥寺北町3-3-1
Tel. 0422-37-3726, Fax. 0422-37-3871

1. はじめに

人間同様の顔を持つエージェント(擬人化エージェント)がヒューマンインタフェース分野のホットなトピックとなっている。これは、あたかも人と人とが直接、接しているような高度な現実感を持った環境を実現することが要求される。その第一の条件としてエージェントが作り物であることをユーザに意識させない、自然な顔表情合成と実時間での音声との同期表示が挙げられる。

このような環境の実現に向けて、マイクから入力された、あるいは記録された自然音声の分析に基づいて会話時の口唇の動き、および表情をリアルタイムに合成するリアルタイムメディア変換システムを提案する。本稿では、ユーザとエージェントとのインタラクション[1]を実現するプロトタイプシステムについて報告する。

2. 音声から顔画像へのメディア変換[2]

音声から顔画像へのメディア変換とは、テキストなどの段階を経ることなく、音声から直接発話時の口の形状および感情を推定することで顔画像を合成するものである。これは、音声から特徴パラメータを抽出し、このパラメータを何らかのマッピングルールに従って、その音声の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)および表情変形パラメータ(以下、表情パラメータ)へと変換することと換言できる。本稿で提案するシステムでは、口形パラメータへの変換ルールには、ニューラルネットワークを用いている。図1に音声から顔画像へのメディア変換のブロック図および変換に用いたニューラルネットワークを示す。

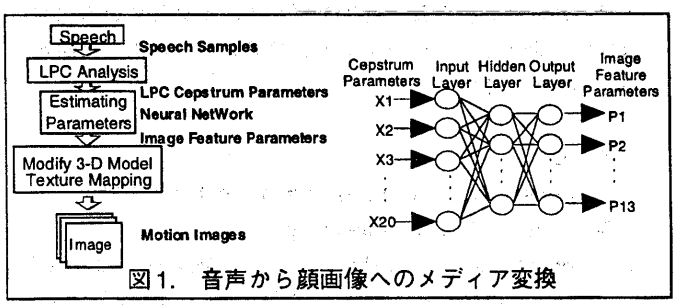

図1. 音声から顔画像へのメディア変換

3. 3次元ワイヤフレームモデル

リアルな顔画像を合成するためには対象人物の顔に忠実な3次元モデルを作成する必要がある。ワイヤフレームモデルの頂点数が多ければ、忠実な形状モデルの再現が可能となるが、その反面、計算コストが大きくなる。そこで、本稿では、人物に忠実で、変形が容易で、変形処理にコストのかからない標準モデルを構築した。そのモデルを図2に示す。このモデルには、歯のモデルが付加されており、歯のモデルは3次元レーザスキャナより歯の型をスキャンすることで構築した。このモデルにテクスチャマッピングを施すことで合成画像を作成する。

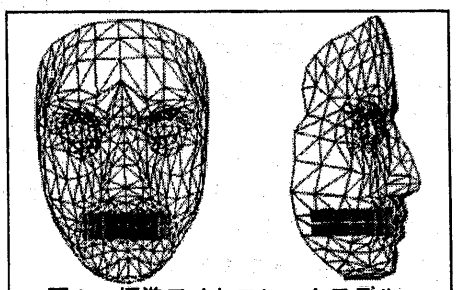

図2. 標準ワイヤフレームモデル

4. システム構成

図3にユーザと仮想エージェントとのインタラクションシステムの構成を示すが、大きく分類して3つのプロセスから構成されている。

(1) 音声分析、パラメータ変換プロセス

このプロセスは、ユーザと仮想エージェントの音声を分析して、3次元モデルの変形パラメータへと変換する部分である。まず、エージェント、ユーザの自然音声を音声入力プロセスにおいて16KHz、16bitでA/D変換し、リングバッファに蓄積する。これを、音声分析プロセスで線形予測分析(LPC分析)を用い、ユーザ、エージェントそれぞれのLPCケプストラムを算出する。ユーザ側の分析では、感情推定のためにピッチ分析等も併せて行う。次にパラメータ変換プロセスでLPCケプストラムからニューラルネットワークを用い、口形パラメータに変換し、ピッチ情報は、6基本感情である喜び、怒り、嫌悪、恐れ、悲しみ、驚きの表情パラメータへと変換される。次にそれぞれのパラメータはネットワーク制御プロセスにおいて、Ethernetを通じてエージェントの画像合成プロセスに送出される。音声分析、パラメータ変換プロセスとネットワーク制御プロセスは完全に独立したプロセスとして、Silicon Graphics社のIndy上で実行される。

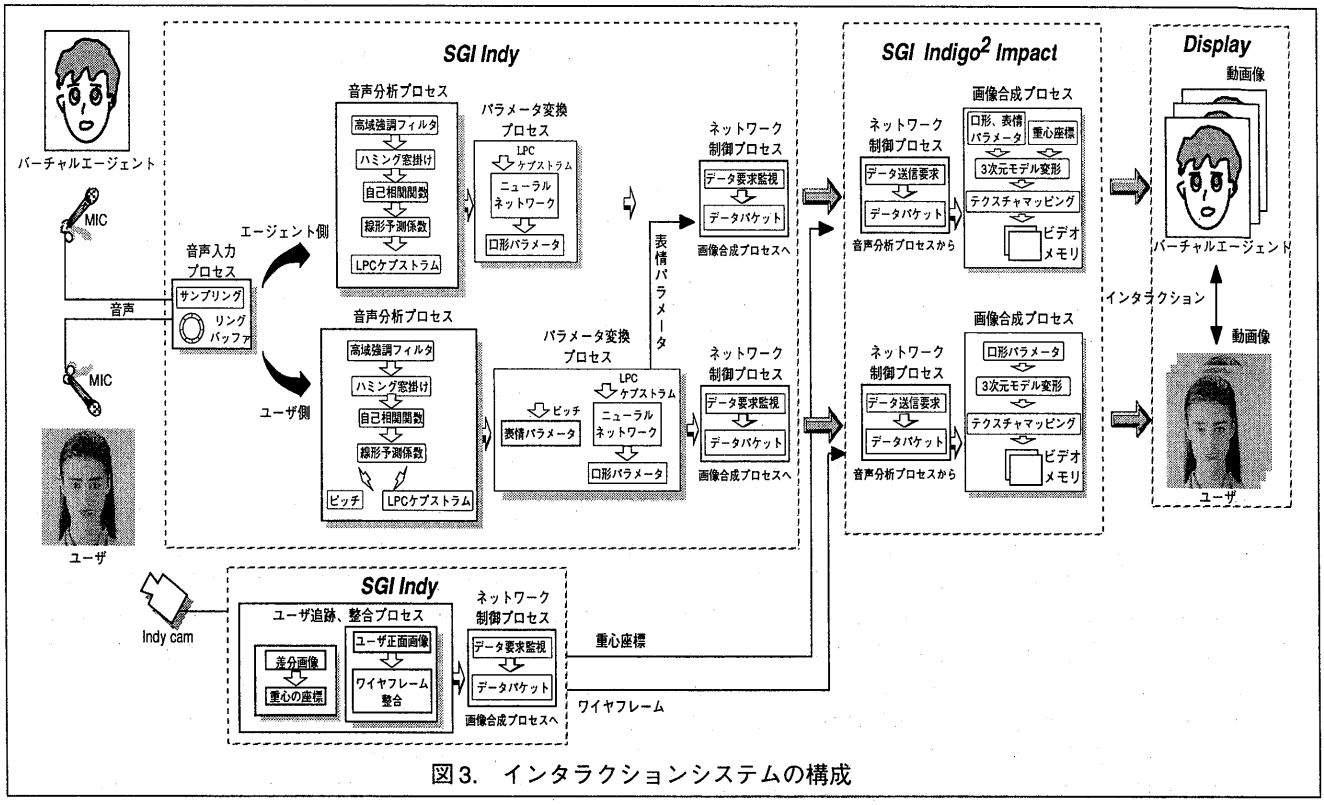

図3. インタラクションシステムの構成

(2) ユーザ追跡と3次元モデル整合プロセス

エージェントがユーザの動きを追跡するために、ディスプレイ上部に設置されたカメラにより、ユーザの正面画像をつねにキャプチャする。その連続画像よりフレーム間差分を計算し、重心座標を求める。この重心座標は、ほぼユーザの両目の中間に位置するため、エージェントの視線をユーザの重心座標に一致させることでエージェントが常にユーザに注意を向けているような印象を与えることができる。ここでは、計算コストを考え、このような簡単なアルゴリズムを用いている。この重心座標はネットワーク制御プロセスによって、エージェントの画像合成プロセスに逐次Ethernetを通じて送出される。また、ユーザ自身の合成画像も常時モニタできるようにするため、開始時にユーザの正面画像1枚から3次元ワイヤフレームモデルをユーザに整合させて個人のモデルを作成する。この整合プロセスは、現在のところ手動で行っている。このモデルは、ユーザ側の画像合成部において変形用のモデルとなる。これらのプロセスは、音声分析、パラメータ変換プロセスで用いたIndyとは別のIndy上で実行される。

(3) 画像合成プロセス

エージェントの顔画像を合成するために、画像合成プロセスでは、表情パラメータ、口形パラメータ、ユーザの重心座標より3次元モデルを変形、移動し、それにテクスチャマッピングを行うことで動画像を生成する。現在は推定されたユーザの感情をエージェントの表情としてプレイバックしているにすぎないが、ノンバーバルな対話ルールによってユーザの感情状態に対するエージェントの表情を決定できる。この対話ルールについては、現在検討段階にある。画像合成プロセスにおいては、テクスチャマッピングに多大な演算量を必要とするため、ネットワーク制御プロセスと、画像合成プロセスとを別プロセスとして実行し、プロセス間通信によって高速化を図っている。これらのプロセスは、Silicon Graphics 社の Indigo² Maximum Impact 上で実行される。

5. システム評価

以上のシステムを構築した結果、音声分析・パラメータ変換レートは平均30.9frame/sec、プロセス間通信に要する時間は2.0msecであった。画像合成レートは、平均23.1frame/secであった。音声が入力されてから画像が合成されるまでの遅延時間は、約87.3msecであるが、エージェントの反応等、インタラクションシステムのプロトタイプとしては、十分自然な対話環境が構築できた。

6. おわりに

本稿では、ユーザと仮想エージェントとのインタラクションシステムについて述べた。現時点では、エージェントは、ユーザの表情及び音声をプレイバックしているにすぎない。今後は、エージェントとのノンバーバルな対話環境の実現を目指す。

参考文献

[1] S. Morishima, Better Face Communication, ACM SIGGRAPH'95, Visual Proceeding p117, 1995

[2] S. Morishima, H. Harashima, "Human Machine Interface Using Media Conversion and Model-Based Coding Schemes", Visual Computing, CG International Series, Spring-Verlag, pp.95-105, 1992