いった、音声メディアの本来の利点をスポイルしてしまうことになる。

そこで、本研究では、画像処理による視線検出技術と、エージェント CG 技術[5]を用い、(1) 画面上の擬人化エージェントに対するユーザの注視を検知し、(2) エージェントの表情によってユーザへフィードバックを返すことで、(3) ユーザとシステムとのアイコンタクトを実現し、このアイコンタクトによって (4) 音声入力の受付可否を制御する “GazeToTalk” システムを開発した。図 1は、本システムの内部構成を示しており、また図 2は、本システムの使用の様子と画面例を表している。

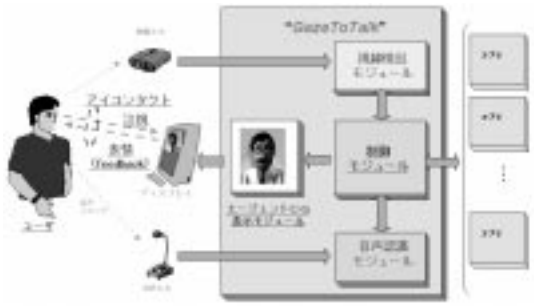
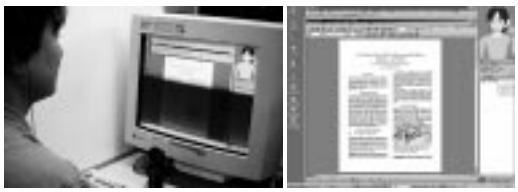


図 1: “GazeToTalk” システムの内部構成



(a) (b)

図 2: システム使用の様子 (a) と画面例 (b)

図 2(a) に於いて、ディスプレイ前下部に設置されているのがカメラであり、ここから得られる動画像情報が視線検出モジュールに渡される。またユーザの装着しているマイクからの音声信号は、音声認識モジュールへと渡され認識処理される。制御モジュールは、各モジュールから得られる情報に基づいて各モジュールを制御し、エージェント CG 表示モジュールは、エージェントの表情および身振りを動画像として生成し、ディスプレイを通じてユーザに提示する。

続いて、各モジュールの概要を説明する。

3.1 視線検出モジュール

本視線検出モジュールは、リアルタイムでユーザ注視位置の検出を行なう。ここでは、カメラから逐次得られる動画像に対する画像処理 [福井 97] を拡張した処理が行なわれ、(1) 画像からの顔領域の抽出および位置トラッキングによるユーザ検出と、(2) 目鼻等の部品領域候補の同定および目周辺のパターン情報の照合による注視位置の判定を行なっっている。本モジュールでの認識処理はソフトウェアのみによって実現されており、ディスプレイ面 9 分割の分解能での高速 (Indy R5000 180MHz 使用で 5 回 /sec) な視線検出を実現している。

さらに本モジュールは、画面特定領域への注視の他、ユーザの到来 / 離脱の検出や、ユーザの視線が画面上のエージェント以外の不特定領域内を推移している状態である ” 非注視状態 ” の検出なども行なう。 この非注視状態は、ユーザが本システムを使用中ではあるが、例えば画面上の他のアプリケーションを操作中であって音声入力を意図していない状態であることを検出するために利用している。

3.2 音声認識モジュール

今回のシステムでは、音声入力の受け付けの可否の制御に注目しているため、音声認識モジュールには、 PC 上で動作するオーソドックスな不特定話者対応の音声認識 [金沢 96] を用いた。認識対象語彙は、あらかじめ用意した語彙セットの中から、その時点でアクティブなアプリケーションに対応して随時自動的に設定され、音声コマンドとして認識できるようにしている。

[5] HI の窓口として擬人化されたキャラクターを CG で生成し利用する技術を、こう呼ぶこととする。