音声言語情報処理 2－3
（1994．7．7）

リアルタイム発話視覚化システムの試作

木山 次郎　伊藤 慶明　岡 隆一

新情報処理開発機構つくば研究センタ

画像によるリアルタイムフィードバックは、人間と計算機の間の新しい種類の interaction を実現するための有望なアプローチである。今回我々は、ユーザの発話内容を漸次的に画像に反映するシステムを試作した。本システムにおいて表示される画像系列は、音声を介してユーザの思考を反映したものであり、一種の「思考の視覚化」に相当すると考えられる。視覚化の対象と提示方法を適切に設定することにより、ユーザに計算機との一体感、臨場感を与えることができると予想され、音声と画像を結び付ける、計算機の新しい適用分野となる可能性を持つ。

A Real-time Visualization System of Thinking

Jiro Kiyama　Yoshiaki Itoh　Ryuichi Oka

Tsukuba Research Center
Real World Computing Partnership

Real-time visual feedback seems to be a promising approach to realize a new interactive system between a human and a machine. We implemented a system which enabled to convert a sequence of utterances into a sequence of images in real-time. The converting process is a partial reflection of a sequence of thinking events through user's utterance. If we obtain a model which has objects suitable to be displayed in images, a user feels as if he is in a machine. Only speech provides this kind of feeling so that it reveals a new application field.

1 本研究の背景

1.1 画像によるフィードバックの意義

近年、人間と計算機との間の円滑な対話を実現するための研究が各所で進められている[1-4]。その中において、画像を計算機からのフィードバックとして積極的に活用しようという研究はそれの持つ可能性に比べると少ないのが現状である。「百聞は一見に如ず」の言葉があるように、画像というメディアには「一覧性」「即解性」という性質があり、多くの有用な情報を一度に解りやすくユーザに伝えることができるという点で、音声、文字のような他のメディアに勝る。また、ユーザをよりタスクに集中させるという効果も期待できる。もちろんすべての情報が画像で表現できるわけではないが、可能な範囲で画像表示を導入することで人間と計算機との間のよりスムーズなコミュニケーションが可能になると思われる。

対話システムにおいて画像によるフィードバックを用いることで実現できるであろう機能をいくつか検討する。ユーザの発話に対するシステムの理解内容の視覚化は、自分の話したことがシステムに正しく理解されているかユーザが一目で確認できるという機能を提供することができる。また、ユーザの発話内容に対するシステムの状態の視覚化は、自分の発話によるシステムの反応を直観的に知ることができるという機能を提供する。Nagao ら[2] は、計算機上の仮想的な人間の表情を用いてユーザの発話に対するシステムの反応の表現、例えば認識理解に失敗した場合、自信のない表情をさせる、を試みその有用性を示している。

1.2 漸次的フィードバックの重要性

発話中にも何らかのフィードバックを適宜ユーザに与えることで、対話の進行を円滑にすることが可能である。例えば、畑崎ら[7] は、ユーザ発話と同時進行的な認識理解結果の出力は、システムの理解が正しいことを確認しながら入力できる、発話内容を考えつつ発話ができる等の利点があるとし、実験によってその効果を確認している。

対話システムにおける、画像による漸次的フィードバックはほとんど試みられてないが、上述した画像の特性を考慮に入れると画像化可能な領域において非常に有用であることは想像に難くない。例えば、ユーザ発話に対するシステムの認識理解の程度を人間の表情を用いて漸次的に表わせば、人間同士の face-to-face の対話と同様の、相手の顔色、表情を見ながら内容や話し方を変えるといった自然な interaction が生まれるであろう。また、ユーザ発話の理解内容の漸次的な画像化は、文字で認識理解結果を出力するのに比べ、対話履歴を含めたシステムの理解内容を常時一目で確認することができるという利点を持つ。

1.3 リアルタイム発話視覚化システム

以上のように、画像による漸次的なフィードバックは、より豊かなコミュニケーションの手助けとなる可能性があるが、我々はそこで行なわれる音声と画像による interaction 自体が、計算機を応用することで実現できる新しい機能となり得るのではないかと考えた。

理解内容の漸次的視覚化を継続的に行なっていくことで、図1のような画像→思考→発話→画像→ ...のループができることになる。ここで提示される漸次的な視覚的フィードバックは、発話を介してユーザの思考が反映されたものであり、一種の「思考の視覚化」ともとらえることができる。ユーザにとってみると自分の考えていることが自発話を通じて時々刻々と視覚化されているように見えるが、ここで適切な視覚化の対象や提示方法をとることで、提示される画像が自分の思考の一部を構成しているような感覚や自分の思考が拡張されたような感覚をユーザに与えることができるのではないかと予想している。本研究における最終的な目標の１つは、そういった臨場感、計算機との一体感をユーザに与えることのできる視覚化モデルを探ることにある。

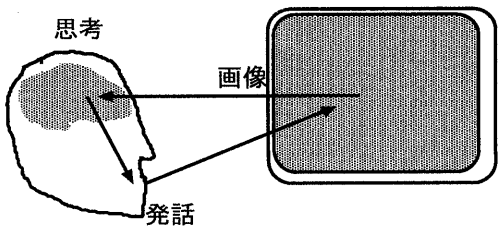

図 1: 視覚化の概念

そのためには、画像、思考、発話の各要素間のギャップをできるだけ小さくすることが重要であろう。例えば、思考→発話のパスにおいては、考えていることがそのまま発話として表れるように、発話に制限を加えないことが必要である。すなわち、任意のタイミングの、しかも不要語や言い直し、非文等を含む発話を適切に処理できなければならない。また、発話→画像のパスにおいては、ユーザを混乱させないよう発話に対し適切なタイミング、手段で画像を提示することが求められる。

今回、小規模なドメインにおいて発話のリアルタイム視覚化システムの試作を行なったので、以下それについて報告を行なう。

2 システムの構成

ドメインには「家の設計」を採用した。「設計」といっても専門家が行なう「設計」ではなく、非専門家が自分の家の大雑把な仕様を決定していくというものである。対象として、視覚的フィードバックによる効果が大きく、またユーザによって身近で興味を惹かれるものが望ましいが、「家の設計」はそれに適合するものだと考えられる。ユーザには図2に示すような画像が提示され、希望するイメージをユーザが述べるとそれが即座に画像に反映されていくことになる。

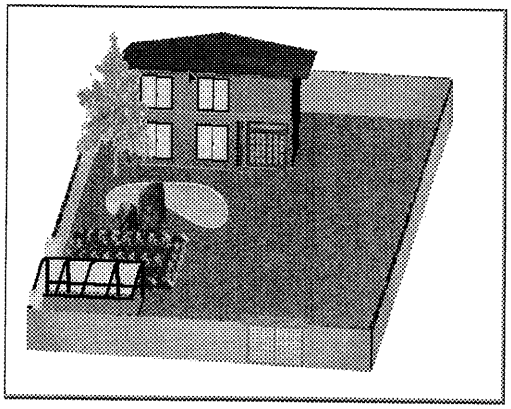

図 2: 提示画像例

システムの構成を図3に示す。大きく意図抽出部と状態管理部の 2 つに分かれる。ここで「意図」とは、システムにアクションを起こさせることのできる、意味の最小単位と定義し、本ドメインの場合 (家の構成要素, 属性値) のペアで表すことにする。これは、ある構成要素をある属性値に変更するという意味を持ち、例えば (家の階数, 二階建て) という意図であれば、家の階数を二階建てに変更することに相当する。システムの受理可能な意図は、用いているドメインからトップダウンに定めることが可能である。

意図抽出部は入力される音声中から意図の抽出を行ない、それを状態管理部に送る。状態管理部では、送られてきた意図情報と直前の理解状況を基に最新の理解状況を更新し、対応する画像の生成を行なう。システム全体は、音声処理に用いている 8msec のフレーム周期と同期しているため、意図の発話とほぼ同期した画像の更新が可能である。

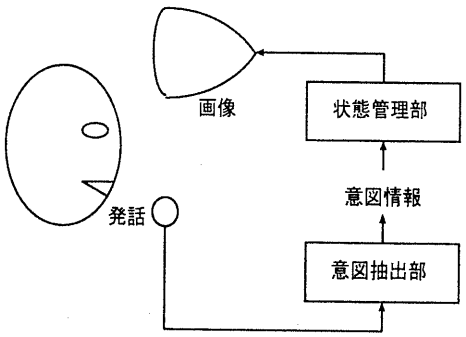

図 3: システムの概略

3 意図抽出部

既に述べたように、このシステムにおいてはユーザの発話に際して、タイミングや語彙、文法等の拘束をできるだけ少なくしなければならない。また、意図の発せられた直後に結果を出力できることも必要である。我々はこれらの条件と認識精度を考慮して次のようなアプローチをとった。

本システムでは、意図はドメインからトップダウンに定められる (構成要素, 属性値) の組で表現されると既に述べた。その枠組を利用し、「構成要素」と「属性値」に対応するキーワードがこの順序で入力音声中に検出された場合にのみ、その意図が発せられたと判断することにする。この制限を導入することで、意図の抽出はコンテキスト独立のキーワード列のスポッティング問題として扱うことができ

る。例えば (家の階数, 二階建て) という意図は、「家」「二階建て」という 2 つのキーワードで表すことにし、図4のように意図の抽出を行なう。本アプローチは、不要語、言い直し等に対する頑強性とある程度の認識精度を両立でき、文法の記述が容易であるという利点を持つ。しかし、上記の制限により省略や倒置に対応できず、話者の発話を制約するという問題を抱える。この点については今後の課題とする。

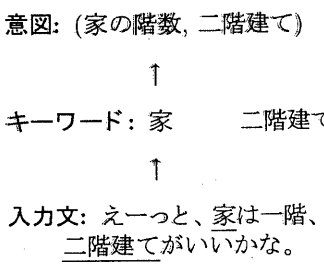

図 4: 意図抽出の例

キーワード列のスポッティングには、連続構造化法 [5] を用いた。連続構造化法は、文法的制約を満たしたキーワード列をフレーム同期に直接出力することができる手法であり、今回のようなタスクに適している。本手法は、キーワードスポッティングをベースとするがラティスを経由しないため、トップダウン的制約をスポッティングに活かせるという利点を持つ。

意図抽出部は、図5に示すように特徴抽出部とキーワードマッチング部とキーワード列スポッティング部で構成されている。以下、各処理部について説明を行なっていく。

3.1 特徴抽出部

サンプリング周波数 15kHz で A/D 変換を行ない、フレーム周期 8msec、フレーム長 17msec で hamming 窓をかけ、36 次元のボカシスペクトルベクトル場 [6] に変換する。

3.2 キーワードマッチング部

連続 DP による各キーワードとの非線形マッチングを行ない、各キーワードに対する現時点を終端としたときのマッチング距離と入力音声における始端時刻をフレーム毎に出力する。各キーワードの標準パターンは、発声変

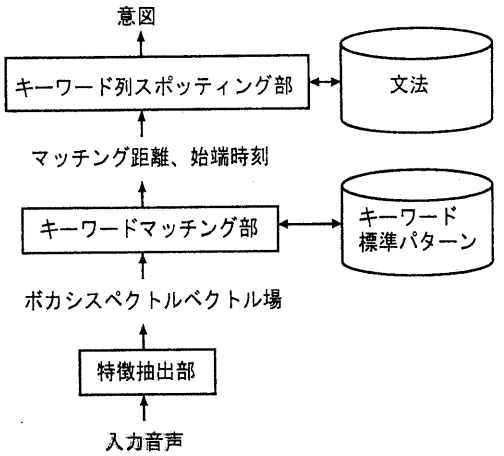

図 5: 意図抽出部

家 ベランダ 門 塀 駐車場
一階建て 二階建て
洋風 和風
生け垣 コンクリート レンガ
カーポート 車庫
いらない いりません 不要です

図 6: キーワードの一覧

動に対応するために複数コンテキスト内で発声された同一キーワードのパターンを平均化したものとした。用いたキーワード数は 17 で一覧を図 6に示す。

現在のところユーザはキーワード音声を登録した人に限られるが、複数話者のキーワードのパターンを平均化することで、不特定話者にも対応することが可能であると考えている。

3.3 キーワード列スポッティング部

各キーワードのマッチング距離、始端時刻を入力として、連続構造化法により語彙中の各キーワードを終端とするその時点における準最適なキーワード列候補が得られる。図 7に示す文法によりキーワード列の終端となり得るキーワードが定義されるので、そのキーワードを終端とするキーワード列候補のみに注目し、最小の評価値 (距離) とその時の候補を保持しておく。そのようにして得られる最小値の時系列から局所最小値かつ閾値未満の点を

検出し、その時点におけるキーワード列を出力する。

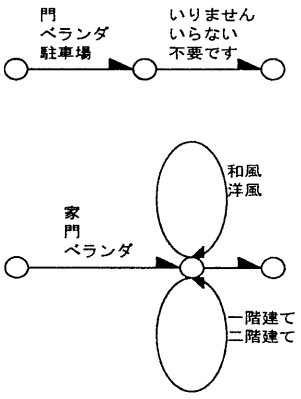

図 7: 用いた文法の例

本システムでは局所最小値の検出に長さ 50 フレームの窓を用いているため、実際のキーワード発話終了から結果出力までに 400msec の遅延があり、完全なリアルタイムにはなっていない。しかし、実際にはキーワード単独で発声されることは少なく、付属語を後続させて文節の形で発声されることが多いと考えられる。その場合、結果的にその文節の発声直後付近において結果が出力されることになり、ユーザとしては文節を発声直後に結果が返ってきているように解釈できる。そのためこの遅延がユーザに違和感を与えるケースは少ないと予想される。

連続構造化法のアルゴリズム自体の説明は [5] に任せ、ここでは本システムで用いたキーワード列評価関数について述べる。

$$D = \frac{\sum_{i=1}^{L} d(i)}{\sum_{i=1}^{L} l(i)} + C_i \cdot \frac{\sum_{i=1}^{L-1} (t_e(i) - t_b(i+1))^2}{L-1} + C_o \cdot (t - t_e(n)) + \frac{\sum_{i=1}^{L-1} C_s(s(i), s(i+1))}{L-1}$$

ここで L はそのキーワード列に含まれるキーワードの数に対応する。$d(i)$ はキーワード列中の i 番目のキーワードのマッチング距離、$l(i)$ はそのキーワードの標準パターンの長さ、$t_b(i)$、$t_e(i)$ はそのキーワード区間候補の入力音声中で対応する開始フレーム、終了フレーム、$s(i)$ はそのキーワードの文法的属性をそれぞれ表す。

右辺の第 1 項は、キーワード列中の各キーワードの距離の和をそれぞれの標準パターン長の和で正規化したものを表す。第 2 項はキーワード間の時間的整合度に対するコストであり、時間差の二乗に定数をかけたものをペナルティとしている。第 3 項は現時刻とそのキーワード列の終端時刻との時間差へのペナルティである。現時刻からの時間差に応じたペナルティをかけることで、最近の事象を優先している。

第 4 項は文法的接続コストである。構文的制約には、図 7に例示する正規文法を用いた。接続可能なキーワード間には 0 のペナルティを、不可能なキーワード間には無限大のペナルティをかける。なお、C_o、C_i等の各項の重みは経験的に決定している。

4 状態管理部

状態管理部は図 9の構成をとる。家を構成する各要素について、取り得る属性値を予め図 8のように定めておき、理解状況は属性値の決定した各要素の組合せによって表現した。

家の階数 (一階建て, 二階建て)

家の様式 (洋風, 和風)

塀の様式 (なし, 生け垣, コンクリート, レンガ)

駐車場 (なし, カーポート, 車庫)

ベランダの様式 (なし, 洋風, 和風)

門の様式 (なし, 洋風, 和風)

図 8: 各要素のとり得る属性値の一覧

入力された意図は、理解状況更新部においてシステムの最新の理解状況と矛盾しないかどうか調べられ、矛盾があった場合その意図は無視される。例えば現在の理解状況において、家が一階建てである場合、(家の階数, 一階建て) という意図が状態管理部に送られたなら、その意図は現在の状況に対し何ら変更を加えないので理解状況の更新は起こらない。もし矛盾がなければ、直前の理解状況に最新の意図を反映したものを最新の理解状況とする。

理解状況の更新が起こった場合、その理解状

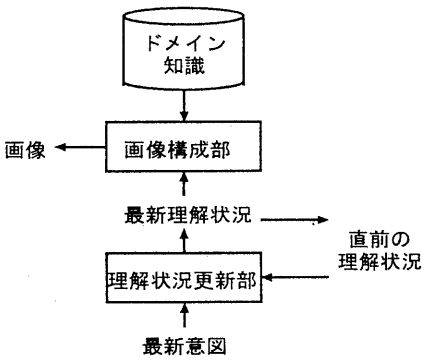

図 9: 状態管理部の構成

況を表す各要素の属性値に則した画像の描画が画像構成部により行なわれる。本システムにおいては、画像構成部は図8に示す以上の変更可能な表示要素を持つが、現状では意図抽出部のリアルタイム化のために語彙数を制限する必要があり、その表示能力を活かしていない。

表示画面は 4 分割され、現時刻と 3 イベント前までの過去の理解画像の履歴を常に表示している。履歴表示は思考過程の視覚化に相当し、以降のユーザの思考を助けるものとして有効であると考えられる。

5 インプリメント

本システムのリアルタイム化を複数の計算機の組み合わせで実現した。図5中の特徴抽出部はIRIS Indigo、キーワードマッチング部はHP9000/755、キーワード列スポッティング部はIRIS Crimson 上にそれぞれインプリメントした。図9中の理解状況更新部はIRIS Indigo、画像構成部はMacintosh Quadra800 上に実装した。画像構成部と理解状況更新部の間のデータのやりとりはsocket 機能で実現し、その他のデータのやりとりはUNIX のパイプ機能を利用した。

6 まとめ

ユーザの発話に対する、画像による漸次的フィードバックの有用性について述べた。また、そのようなフィードバックを積極的にユーザに提示することにより臨場感、一体感を与えることのできる可能性について考察し、その検証のためのシステムを試作した。

今後は本システムに関して評価実験を行ない、その際に得られる被験者の反応等の知見を基にシステムの完成度を上げていきたいと考えている。また、臨場感を与えるのに適した画像の提示方法 (表示タイミング、表示デバイス等) についての検討も重要課題の 1 つである。

謝辞 本研究に日頃支援頂く、新情報処理開発機構 島田潤一所長に深謝いたします。Macintosh 上の表示ソフトウエアを開発頂いた (株) メディアドライブ研究所の中越、松村両氏に深謝いたします。

参考文献

[1] S. Hayamizu, K. Itou, M. Tamono and K. Tanaka "A Spoken Language Dialog System for Automatic Collection of Spontaneous Speech", Proc. ICSLP-92, (1992)

[2] K. Nagao and A. Takeuchi, "A New Modality for Natural Human-Computer Interaction: Integration of Speech Dialogue and Facial Animation", Proc. ISSD-93, (1993.11)

[3] S. Seto, Y. Nagata, H. Kanazawa, H. Hashimoto, H. Shinchi and Y. Takebayashi, "Spontaneous Speech Dialogue System TOSBURG II and its Evaluation", Proc. ISSD-93, (1993.11)

[4] 吉岡, 南, 山田, 鹿野,"電話番号案内を対象としたマルチモーダル対話システムの作成", 音講論,1-8-19 (1993.10).

[5] 木山, 伊藤, 岡,"連続構造化法の性能評価", 音講論,2-7-10 (1994.3).

[6] R. Oka and H. Matumura, "Speaker Independent Word Speech Recognition using the Blurred Orientation Pattern Obtained from the Vector Field of Spectrum", Proc. IJCPR (1988.11)

[7] 畑崎, 野口,F. Ehsani, 渡辺,"発話同時理解による音声対話インタフェースの検討", 音講論,3-4-13 (1993.3).