マルチモーダル対話データベースに基づく音声と身振りの認識系統合モデル

湯浅夏樹† 三谷純司† 外川文雄†

本稿では，音声と頭の動きを統合して扱い，ユーザの意図（肯定/否定）の識別をマルチモーダル対話データベース中のデータに基づいて行う「認識系統合モデル」を提案する．そしてこのモデルの意図識別率とマルチモーダルインタフェースの快適度の2点について評価した結果について述べる．その結果，実際の対話システムを開発して評価したときの識別率は，音声だけを用いた場合が86.3%，頭の動きだけを用いた場合が55.6%なのに対し，音声と頭の動きを両方とも用いた場合が91.9%となり，本認識系統合モデルが有効に機能することを確認できた．また，今回試作したマルチモーダルインタフェースを快適に感じるかどうかには個人差があることが分かった．

Integrated Recognition Models of Keywords in Speech and Head Movements Using a Multimodal Interaction Database

NATSUKI YUASA,† JUNJI MITANI† and FUMIO TOGAWA†

This paper describes an integrated recognition model of keywords in speech and head movements using a multimodal interaction database, and describes this system's accuracy and comfortable degree. In the results, in use of speech recognition only, a success rate of 86.3% was achieved, and in use of head movements recognition only, a success rate of 55.6% was achieved, but in use of both recognition, a success rate of 91.9% was achieved. And we confirmed that there are differences among individuals of feeling the system's comfortable degree.

1. はじめに

従来の音声認識のみによる対話システムでは，人が発する情報の一部分しか利用できなかったために，自然な対話を実現するのは困難であった．たとえば，人間どうしの対話においては，声を出さなくても頭の振りや顔の向き，視線，身振り手振り等で意図を伝えている場合がかなりある．したがって，より自然なヒューマンインタフェースを実現するためには，音声認識だけでなく，頭の動きやジェスチャ等も認識して，これらの情報を統合して扱う必要があると考えられる．

近年，このような複数の情報を扱うマルチモーダル対話システムに関して様々な研究が行われている[1]．たとえば，音声だけでなくジェスチャ等も使用することで入力の自由度を向上させた対話システムの研究[2),3)]，音声入力とタッチパネルを用いた対話システムの検討[4)]，音声とマウスを用いた対話システムの研究[5)]，音声・マウス・キーボードによる統合入力環境についての検討[6)]，音声認識と唇画像の認識を組み合わせて音声の認識率向上を図る試み[7),8)]，また，対話における身振り手振り，うなずきや視線などの役割を示した研究[9)~11)]等がある．しかし，人間どうしの対話過程の詳細な解析結果に基づいて対話システムを開発した例はほとんどなかった．

我々はこれまで対話における動作や発話のデータを収集し，マルチモーダル対話データベースを構築して，データに基づいた対話の解析を行ってきた[12)~16)]．今回認識系統合モデルとしてユーザの意図の識別をこのマルチモーダル対話データベース中のデータに基づいて行うモデルを提案する．また，この認識系統合モデルの適用例として「肯定/否定」の識別を音声認識と頭の振りの認識を統合することで行う対話システムを開発して意図識別率を評価した．そして，音声や頭の振りや顔の向きを用いてシステムと対話ができる「商品紹介システム」を試作し，マルチモーダルインタ

† シャープ株式会社映像メディア研究所内 RWCP 新機能シャープ研究室
Real World Computing Partnership Novel Functions Sharp Laboratory in Image and Media Laboratories, Sharp Corporation

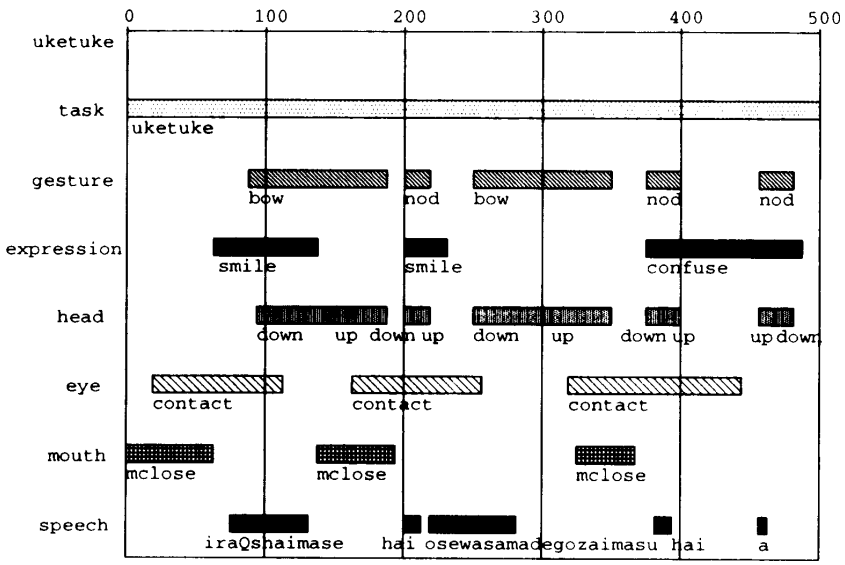

図1 マルチモーダル対話データベース
Fig. 1 Multimodal interaction database.

フェースの快適度の主観評価を行った.

本稿では,この認識系統合モデルと対話システムについて説明し,「肯定/否定」の識別率を評価した結果および「商品紹介システム」を用いたマルチモーダルインタフェースの主観評価の結果について報告する.

2. マルチモーダル対話データベース

我々は,人と人との対話過程を忠実に記録・解析するために,被験者が100インチのスクリーンに映し出された人(コンピュータシステムが表示する疑似人間の場合もある)と対話をしている様子をマイクおよびカメラによってとらえ,ランダムアクセス可能な光磁気ディスクに記録できる実験設備を構築した.記録されたデータは1フレーム(1/30秒)単位で再生でき,様々なモード(音声,頭の動き,目の動き等)について視察でラベリングが行えるようになっている.このようにして記録した音声や動作を含むデータのことを,本論文中ではマルチモーダル対話データベースと呼ぶ.

図1に示す例は,受付担当者と訪問者役の被験者との対話の例で,受付担当者側のデータである.これは会社の受付での対話という設定で,訪問者役の被験者には「加藤さんを訪ねてR社の受付にやってきた」という課題で受付担当者と自由に対話してもらった.

図1の横軸の単位はフレーム(1/30秒)である.縦軸には以下のラベリング情報が並んでいる.

- uketuke (タスクの名称)
- task (タスクが行われている期間)
- gesture (「おじぎ」や「うなずき」)
- expression (表情)
- head (頭の動き)
- eye (相手に視線を向けているかどうか)
- mouth (口が閉じているかどうか)
- speech (発話内容)

被験者の視線および頭の振りは,被験者の顔の正面に据えられたカメラを通して光磁気ディスクに記録されたデータを,ラベリング担当者(1名)が視察でラベリングを行った.

eye contactは,被験者が話相手の顔に視線を向けているか否かを視察で観測した.頭の振りは,被験者の頭の上下左右方向の動きを視察で観測した.

2.1 マルチモーダル対話データベースの解析

本論文では,認識系統合モデルの適用例として,「肯定/否定」の識別を行う対話システムを開発し,評価した.この「肯定/否定」の識別率の評価に用いたマルチモーダル対話データベースはQ&A(イエス/ノー質問に答える)タスクのデータである.被験者(回答者)は7名.スクリーンに映し出された人(質問者)の質問に「イエス/ノー」で答えてもらうというタスクで対話してもらった.具体的な質問内容は,「出身地は千葉県ですか?」「今朝コーヒーは飲みましたか?」「富士山の高さは3500 mより高いですか?」等である.

このようなデータ(回答が「肯定/否定」になっている)に注目し,その一塊のやりとりの中で回答者に発生している「発話」「頭の振り」「視線の動き」等を調べてみたところ,次のようなことが分かった.

なお,ここでは「キーワード」とは,"質問-答"という各ペアにおいて,質問者側が回答者に答えてもら

うことを期待していると推察される（回答者の回答を最も誘引していると推察される）単語である．

- 質問者の質問文中の一番重要な単語（キーワード）が発話された後に，回答者の意図を示す反応（「発話」や「頭の振り」）が現れる．
- 難しい質問ですぐに回答することができない場合でも，質問者の発話終了後0.5秒以内に，回答者は何らかの反応（「頭のかしげ」や「目の泳ぎ」を含む）を示している．
- 回答者の意図が「肯定」の場合は「『はい』,『そうです』，質問の述部を繰り返す等の発話」や「頭の縦振り」や「質問文中のキーワードの発話」が起こる（例：「今日は暑いですか？」に対して「はい，暑いです」と回答）．
- 回答者の意図が「否定」の場合は「『いいえ』，質問の述部の否定形等の発話」や「頭の横振り」や「質問文中のキーワードの反意語の発話」「質問文中のキーワードに関連した単語（正解）の発話」が起こる（例：「今日は月曜日ですか？」に対して「いいえ，火曜日です」と回答）．
- 「頭のかしげ」や「目の泳ぎ」が現れると，頭がまっすぐになったり目が正面を向くまで回答は行われない．
- 回答者の意図が「否定」の場合でも，発話の後半で「頭の縦振り」が何回か見られることがあるが，これは回答者自身の発話自体に対するうなずきであり，質問文に対する肯定を意図してはいない「頭の縦振り」である．

3. 音声と身振りの認識系統合モデル

マルチモーダル対話データベースの解析結果に基づいて，次のような「音声と身振りの認識系統合モデル」を構築した．この概念図を**図2**に示す．

これは対話システム（コンピュータ側）のキーワード発話時刻や発話終了時刻に基づいてレスポンスウィンドウという時間枠を設定して，このレスポンスウィンドウ内での各モード（発話単語，頭の振り等）の生起に基づいて回答者の意図をベイズ識別するものである．ベイズ識別関数のための学習データはマルチモーダル対話データベース中のデータを使用し，発話単語をカテゴリ分けするために大規模テキストデータベースから学習した概念ベクトルというものを使用している．

3.1 レスポンスウィンドウ

レスポンスウィンドウは次の3つの規則に従って設定される．

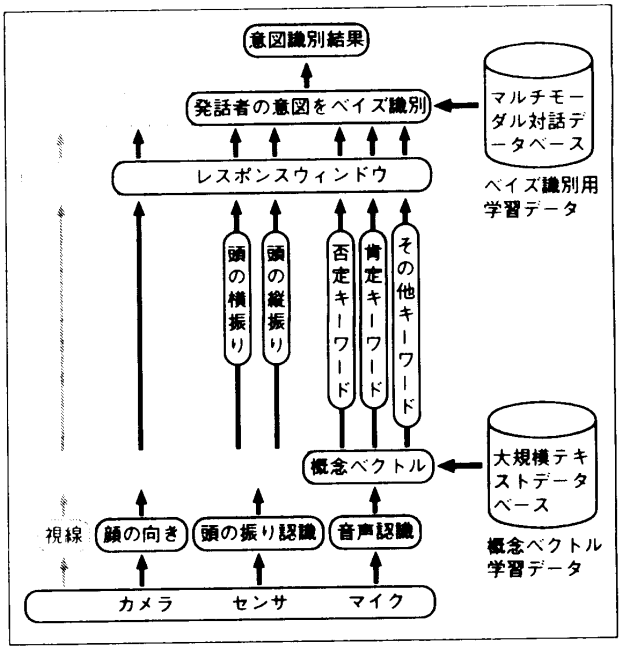

図2 音声と身振りの認識系統合モデル
Fig. 2 Integrated recognition models of keywords in speech and head movements.

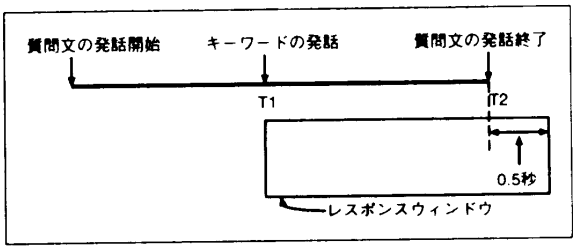

図3 レスポンスウィンドウ
Fig. 3 The response window.

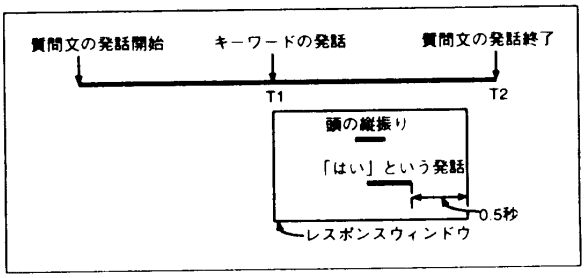

図4 レスポンスウィンドウの短縮
Fig. 4 Shortening the response window.

(1) 質問者の質問文中の最重要単語（キーワード）が発話された時刻を開始時刻とし，質問の発話終了後0.5秒の時刻を終了時刻とする（**図3**）．
(2) 回答者が何らかの反応を示した後に，0.5秒以上何の反応も示さない場合に短縮される（**図4**）．
(3) 回答者の「頭のかしげ」「目の泳ぎ」「不要語の発話」によって伸長される（**図5**）．

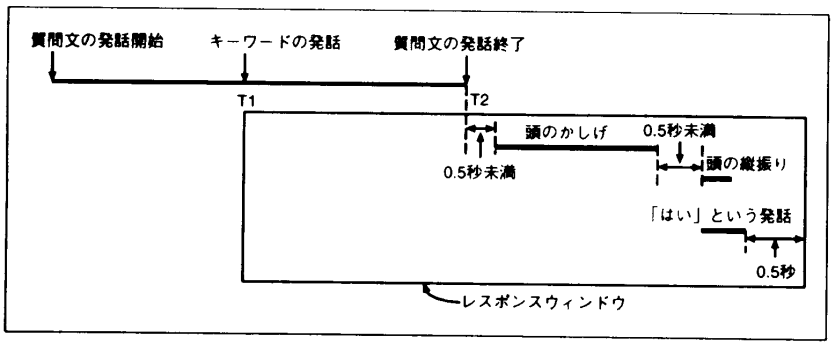

図5 レスポンスウィンドウの伸長
Fig. 5 Extending the response window.

3.2 認識系統合モデルによる「肯定/否定」の意図識別

今回は認識系統合モデルの最も簡単な適用例として，回答者の意図が「肯定/否定」の場合の対話システムを開発し，評価を行った．意図を「肯定/否定」にしぼったことにともない，回答者の反応のうち認識するものは「音声」「頭の縦振り」「頭の横振り」とし，「音声」に関しては次の3種類にカテゴリ分けをした．

(1) 肯定キーワード：「はい」「そうです」等の肯定に用いる常套句や，質問文中のキーワード
(2) 否定キーワード：「いいえ」等の否定に用いる常套句や，質問文中のキーワードの反意語あるいは関連している単語
(3) その他キーワード：肯定キーワードでも否定キーワードでもない単語

なお，否定キーワードの「質問文中のキーワードの反意語あるいは関連している単語」の設定には，後述する概念ベクトルを用いている．

複数のモードの入力データを統合する方法としては，(1) ベイズ識別，(2) ニューラルネットワーク，(3) HMM，… 等が考えられる．現状ではマルチモーダル対話データベースに蓄積されているデータ数はまだそれほど多くないが，ベイズ識別は学習データ数が比較的少なくても，その学習データの範囲内で最適な識別ができる．また，ニューラルネットワークやHMMでは学習結果をブラックボックス的に扱うのに対し，ベイズ識別の学習結果は人間に分かりやすい形態になっているため，後々の分析等もしやすいと考えられる．以上の理由から，本認識系統合モデルでは，複数のモードの入力データを統合する方法としてベイズ識別[17]を採用した．

表1 中にある $x_1,\cdots,x_5$ が今回の統合モデルのベイズ識別に用いられる入力データである．y_1 については「肯定/否定」のベイズ識別には用いていないが，後述する「商品紹介システム」において，商品の選択を顔の向きによっても行えるようにするために用いている．

表1 認識系統合モデルのベイズ識別に用いられる入力データ
Table 1 Input data used in Bayesian decision.

入力装置	モード	生起データ
マイク	音声	x_1：肯定キーワードの発話 x_2：否定キーワードの発話 x_3：その他キーワードの発話
センサ	頭の振り	x_4：頭の縦振り x_5：頭の横振り
カメラ	顔の方向	y_1：顔の向き

3.2.1 ベイズ識別関数の学習

マルチモーダル対話データベース中の対話の中で，発話者の意図が「肯定/否定」になっているものを集め，それぞれについてレスポンスウィンドウを設定し，そのレスポンスウィンドウ内での $x_1,\cdots,x_5$ の生起状況とその時の発話者の意図（肯定/否定）とのペアを，ベイズ識別関数の学習データとする．たとえば，**図6** のように，発話者の意図が「肯定」の場合に，「肯定キーワードの発話」と「頭の縦振り」が生じているデータがあったとすると，

肯定，$x_1=1,\ x_2=0,\ x_3=0,\ x_4=1,\ x_5=0$

という学習データが得られる．また，**図7** のように，発話者の意図が「否定」の場合に，「否定キーワードの発話」と「頭の横振り」が生じているデータがあったとすると，

否定，$x_1=0,\ x_2=1,\ x_3=0,\ x_4=0,\ x_5=1$

という学習データが得られる．

このような学習データをたくさん用意しておけば，認識系統合においては，設定されたレスポンスウィンドウ内での $x_1,\cdots,x_5$ のパターンから次の関数 f で発話者の意図が識別できる．

$$f(x_1,x_2,x_3,x_4,x_5)=\begin{cases}\text{肯定} & (p_y > p_n)\\ \text{否定} & (p_y \le p_n)\end{cases} \quad (1)$$

ただし p_y，p_n はマルチモーダル対話データベース中

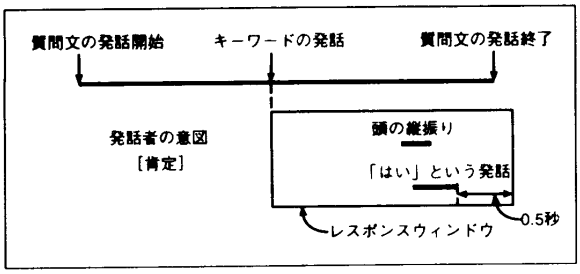

図6 「肯定」の学習データの例
Fig. 6 Example data of affirmation.

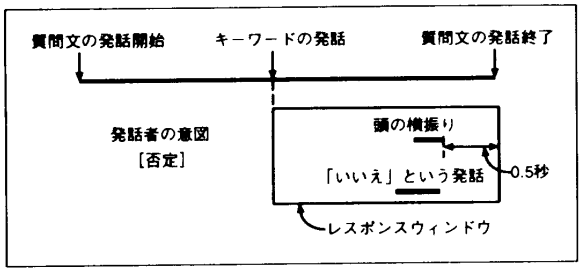

図7 「否定」の学習データの例
Fig. 7 Example data of negation.

の学習データから求めた以下の確率である.

$p_y = p(x_1, x_2, x_3, x_4, x_5|C_{yes})$
= 発話者の意図が肯定のときの $x_1, \cdots, x_5$ の出現確率

$p_n = p(x_1, x_2, x_3, x_4, x_5|C_{no})$
= 発話者の意図が否定のときの $x_1, \cdots, x_5$ の出現確率

今回開発したシステムでは，ベイズ識別関数の学習データとして，マルチモーダル対話データベース中の被験者 A のデータで，回答が「肯定/否定」になるもの 31 対話分を使用した.

なお，今回のシステムでは回答者の意図のベイズ識別は「肯定/否定」のみの識別に使用したが，意図の種類をどのように増やしていけるかについては今後の課題としたい.

3.3 概念ベクトル

2.1 節で述べたように，マルチモーダル対話データベース中のデータのうち，回答が「肯定/否定」になっているデータに注目すると，回答者の意図が「否定」の場合には「『いいえ』，質問の述部の否定形等の発話」や「頭の横振り」や「質問文中のキーワードの反意語の発話」「質問文中のキーワードに関連した単語（正解）の発話」が起こることが分かった.

したがって，質問文中に含まれるキーワードの反意語あるいは関連している単語を否定キーワードとして定義すると，単語を扱いやすくなることが分かる.

この「質問文中に含まれるキーワードの反意語あるいは関連している単語」を設定するために，本システムでは大量のテキストデータから作成した単語の特徴ベクトル（概念ベクトル）を用いている.

3.3.1 概念ベクトルの学習

この特徴ベクトルの作成方法について簡単に説明する[18].

同一の記事中や，対話の一塊のやりとりの中などで同時に出現（共起）している単語は互いに関連している単語であると考えられる．そこで，これらの共起している単語どうしが類似した値を持つような特徴ベクトルを構成すれば，特徴ベクトルが類似している単語を検索することで，関連する単語が得られることになる．今回は大規模テキストデータベースとして，朝日新聞の記事 1 年分を使用し，その各記事の単語出現頻度分布をその記事中に共起している各単語の特徴ベクトルに加算していくようにした．このようにすれば，同じ記事中で共起している単語どうしの特徴ベクトルは類似するようになる．本稿では，この単語の特徴ベクトルのことを概念ベクトルと呼ぶことにする.

単語の概念ベクトルの作成方法をより詳しく説明する.

単語出現頻度分布を調査する単語を $word_1, word_2, \cdots, word_n$ の n 個とし，記事は m 個あるとする．記事 i に含まれる単語の出現頻度ベクトル $\mathbf{V}_i$ を

$$\mathbf{V}_i = (v_{i1}, v_{i2}, \cdots, v_{in}) \quad (2)$$

v_{ij} ： 記事 i 中に出現する $word_j$ の個数

で表すと，単語の概念ベクトル $\mathbf{W}_1, \mathbf{W}_2, \cdots, \mathbf{W}_n$ は，

$$\mathbf{W}_i = \sum_{j=1}^{m} v_{ji} \cdot \frac{\mathbf{V}_j}{|\mathbf{V}_j|} \quad (3)$$

となる.

概念ベクトルを学習するための大規模テキストデータベースの代表例としては新聞記事データがあげられるが，新聞記事と対話テキストとはかなり性質の異なる文章であるから，新聞記事データのようなものだけから学習すると期待どおりの概念ベクトルが得られない可能性がある．しかし，我々はマルチモーダル対話データベースを構築しており，このデータベースから音声テキストデータ部分を抜き出すことで，概念ベクトルの学習データとして用いることができるようになっている.

また，概念ベクトルの計算対象とする単語の種類は，音声認識に用いる単語のうち，特にキーワードとなりうるものを最低限設定する．したがって，今回の実験では「好き」「嫌い」「高い」「低い」「日曜日」「月曜

日」…「土曜日」「一日」「二日」…「三十一日」が最低限設定すべき単語となり，実際には新聞記事中から出現頻度順に選び出した単語（名詞，形容詞，形容動詞，動詞）もこれらに加え，1024単語として概念ベクトルを計算した．

今回開発したシステムでは，概念ベクトルの学習データとして，1990年の朝日新聞1年分[19]のみを使用したが，「好き/嫌い」「高い/低い」などが類似した概念ベクトルとなり，否定キーワードの設定に利用できることを確認した．

3.3.2 概念ベクトルによる否定キーワード設定法

単語の概念ベクトルのノルムを1に正規化してから内積を取ることで，単語間の類似度を計算することができる．否定キーワードの「質問文中に含まれるキーワードの反意語あるいは関連している単語」は，質問文中のキーワードの概念ベクトルとの類似度が閾値以上の概念ベクトルを持つ単語（あるいは，類似度の高い順に適当な候補数だけとってきたもの）とすればよい．

今回の実験の音声認識対象語彙の中での類似度を調べてみると，「高い」と「低い」や，「好き」と「嫌い」の類似度が高くなった．したがって，質問文中に「高い」が含まれている場合は「低い」が否定キーワードとなり，質問文中に「好き」が含まれている場合は「嫌い」が否定キーワードとなる．

なお，「関連している語」の中には同義語も含まれてしまう．しかし，これまでのマルチモーダル対話データベースの解析結果からは，肯定する意志がある回答者が「質問文中のキーワード」ではなく，わざわざその同義語を発話するということは確認されていない．「質問文中のキーワード」を発話せず，表現の違う言葉を発話したということは，回答者の「肯定したくない」という意志が働いていると考えられる．したがって，本方式においては「関連している語」には同義語も含まれたまま処理して問題ないということにしている．

4. 各種認識部品

認識系統合モデルを検証するための対話システムを開発するためには，統合する情報を認識するための認識部品が必要である．それらについて簡単に説明しておく．

4.1 音　　声

ユーザの発話する音声から連続DP法を用いて単語を検出する．これは，予め登録した各単語の標準パタンと入力音声を始端から1フレーム（20 msec）ずつずらしながら，入力音声の部分区間とDPマッチングを行う手法である．

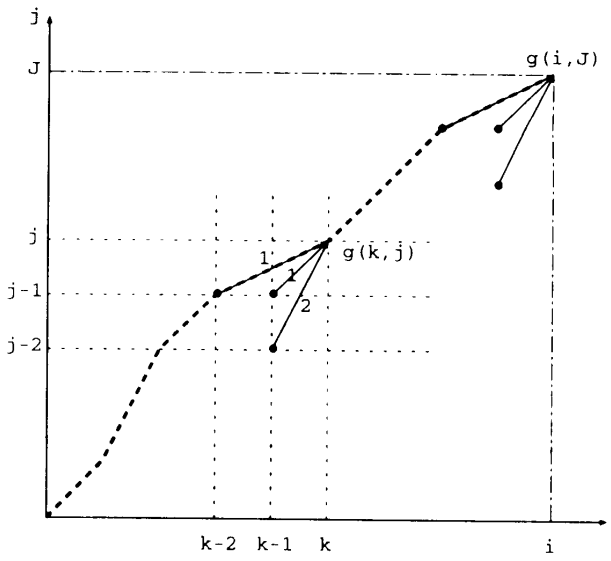

図8 連続DP法の時間伸縮関数
Fig. 8 Warping function of continuous DP method.

以下に，計算方法について述べる．入力音声の k フレームと標準パタンの j フレームとの距離を $d(k,j)$ とすると累積距離 $g(k,j)$ は，次式で計算される（図8）．

$$g(k,j) = \min\left(\begin{array}{l} g(k-2,j-1) + d(k,j), \\ g(k-1,j-1) + d(k,j), \\ g(k-1,j-2) + 2d(k,j) \end{array} \right) \quad (4)$$

また，DPパスの長さ $c(k,j)$ は次式で計算される．ただし，上式において，

$g(k,j) = g(k-2,j-1) + d(k,j)$ のときを case a,
$g(k,j) = g(k-1,j-1) + d(k,j)$ のときを case b,
$g(k,j) = g(k-1,j-2) + 2d(k,j)$ のときを case c
とする．

$$c(k,j) = \begin{cases} c(k-2,j-1) + 2 & \text{case a} \\ c(k-1,j-1) + 1 & \text{case b} \\ c(k-1,j-2) + 1 & \text{case c} \end{cases} \quad (5)$$

そして，標準パタン長を J としたとき，i フレームにおける累積距離を最適DPパスで正規化した距離 $D(i)$ を求める．

$$D(i) = g(i,J)/c(i,J) \quad (6)$$

ある単語の標準パタンとの距離 $D(i)$ が設定した閾値以下でかつ最小値になった時点で，その単語が存在したと判定し出力する．

一般に，マッチング区間が短いほうが累積距離が小さくなる傾向があるため，式(4)や式(5)の重み付けをすることで，標準パターン長に比べ極端に短い発声に対してマッチングすることを防いでいる．

次の章で記述する意図識別率の評価実験では，ユー

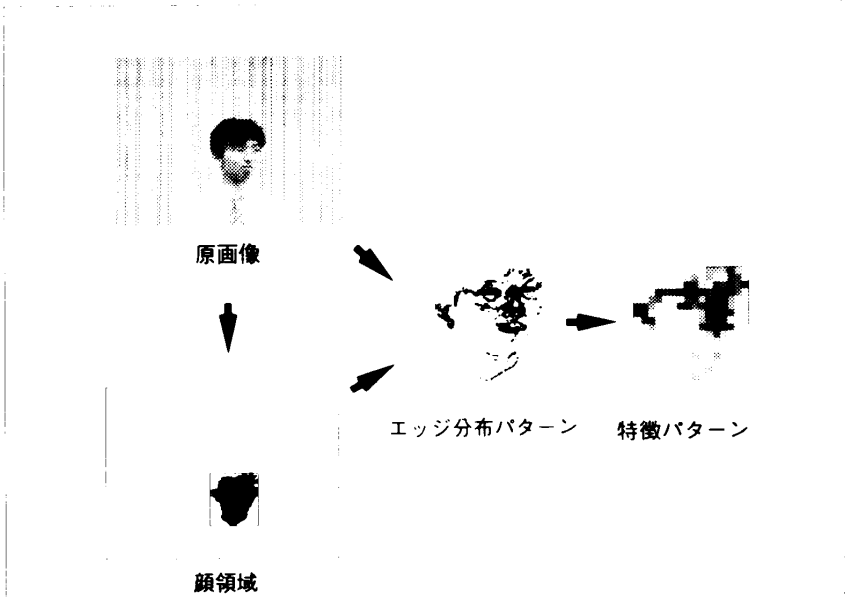

図9 顔の向きの判定
Fig. 9 Discrimination of facial direction.

ザの返答内容を想定し，語彙として「はい」「そうです」「いいえ」「違います」「あります」「ありません」「好きです」「嫌いです」「高いです」「低いです」「水曜日です」「八日です」の12語を用い，商品紹介システムでは,「はい」「いいえ」「ビューカム」「ザウルス」「右」「左」「それ」の7語を用いた．

質問文の内容によって音声認識語彙を変化させるようなことは行っていないので，たとえば質問文が「～は好きですか？」の場合，肯定キーワードは「はい」「そうです」「好きです」，否定キーワードは「いいえ」「違います」「嫌いです」，その他キーワードは「あります」「ありません」「高いです」「低いです」「水曜日です」「八日です」となる．

4.2 頭の動き

頭の動きとして，顔の向きと頭の振りを認識する認識部品を開発した．入力手段としては，非接触な手段が望ましい．しかし，実時間での使用を考慮して，頭の振りについては頭部に付けた磁気センサによって検出している．顔の向きについては，テレビカメラから入力した画像データから画像認識により非接触な入力を行っている．

4.2.1 顔の向きの判定

色情報に基づいて顔領域の抽出を行い，抽出された顔領域を囲む矩形内におけるエッジ分布パターンを特徴として用いる．顔の各向きに対してあらかじめ複数の顔画像データから特徴パターンを求めて顔の向きごとに平均して標準パターンを作成しておく．顔の向きの判定は，特徴パターンと標準パターンとの類似度によって行う．

顔画像から顔の向きを判定する処理は以下のように行う[20]．

4.2.1.1 顔領域の抽出

色情報に基づいて顔領域の抽出を行う．まず，テレビカメラより入力したカラー画像から肌色領域を抽出し，肌色領域の中で面積最大となる連結領域を選ぶ．この面積最大の領域に対して，穴となる部分を埋める処理をして得られる領域を顔領域とする．

4.2.1.2 特徴抽出

抽出された顔領域を囲む矩形内におけるエッジの分布パターンを特徴として用いる．この矩形内においてエッジの強さを求め，エッジの強さが閾値以上なら1に，エッジの強さが閾値未満なら0とするパターンを作成する．マッチングのために，モザイク処理により，矩形を一定の大きさになるように正規化し，さらにノルムが1になるように正規化して，特徴パターンとする（図9）．

4.2.1.3 判　　定

顔の各向きに対して，あらかじめ複数の顔画像データから特徴パターンを求めて，顔の向き d_i ごとに平均して標準パターン R_{d_i} を作成する．判定は，パターンをベクトルと見なして，特徴パターン F と標準パターン R_{d_i} との内積 $F \cdot R_{d_i}$ により求められる類似度 S_{d_i}

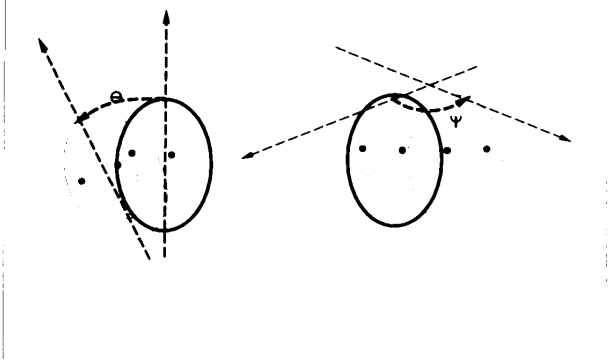

図10 頭の振りの検出
Fig. 10 Detection of head movements.

によって評価する．顔の向きごとに標準パターンとの類似度 S_{d_i} を計算し，最大類似度 $S_{d_m} = \max_{d_i} S_{d_i}$ となる方向 d_m を顔の向きとする．

4.2.2 頭の振りの検出

頭部につけた磁気センサから 60 msec ごとに頭の縦方向の角度 Θ，横方向の角度 Ψ を求める（**図10**）．角度の変化 $\Delta\Theta$，$\Delta\Psi$ を閾値処理することで頭の縦振りと横振りを検出している．

4.3 対話制御

認識系統合モデルを用いて人間と対話を行うためには対話制御を行う必要がある．今回の対話システムでは状態遷移による単純な対話制御を行っている．

これは，現在の状態から，次にシステムが発話すべき内容が決定され，その発話に対する回答が肯定なのか否定なのかによって，次の状態に遷移していくものである．

この処理を行っている対話制御部は，システムが発話する時刻やシステムが発話する内容の中のキーワードが何であるかを把握しているので，この情報と各認識部品から送られてくる情報とからレスポンスウィンドウや肯定キーワードや否定キーワードを設定することができ，対話している人の意図の識別（肯定/否定）ができるようになる．

5. 認識系統合モデルの評価

認識系統合モデルの意図識別率の評価は，マルチモーダル対話データベース中のデータを用いた場合と，実際に対話システムを開発し人間とそのシステムとの対話結果を用いた場合との2種類で行った．また，マルチモーダルインタフェースの快適度についても主観評価を行った．

表2 認識系統合モデルの意図識別率（データベースを用いた評価）
Table 2 Understanding rate of user's intentions (on the experiments using the database).

マルチモード	音声と頭の振りの両方の場合	98%
シングルモード	音声だけの場合	91%
	頭の振りだけの場合	72%

5.1 マルチモーダル対話データベースを用いた評価

マルチモーダル対話データベース中で，回答が「肯定/否定」になるものだけを64対話集め（ただし認識系統合モデルのベイズ識別関数を学習するときには用いなかったデータ），それらを本認識系統合モデルに通した場合の意図（肯定/否定）の識別率を調べた．なお，発話や頭の振りがなかったり学習データにない等の理由で肯定/否定の判断ができない場合は「否定」と識別するようにした．

なお，ベイズ識別関数学習用データの31対話も評価用データの64対話も，同じQ&Aタスクの対話であり，学習用データは被験者Aのデータを使用し，評価用データは被験者A以外の6名のデータを使用した．

意図識別率を調べた結果は**表2**のようになった．

この結果を見ると，音声と頭の振りを両方とも用いることで，意図識別率が向上することが確認できる．

表2において，音声だけの場合に識別できなかった9%分は意図が「肯定」であるにもかかわらず，頭の振りしか見られなかった場合と発話内容に「その他キーワード」しか見られなかった場合である．頭の振りだけの場合に識別できなかった28%分は，意図が「肯定」であるにもかかわらず，頭の振りが見られなかった場合である．音声と頭の振りの両方の場合に識別できなかった2%分は，意図が「肯定」であるにもかかわらず，発話内容に「その他キーワード」しか見られなかった場合である．

5.2 対話システムを用いた評価

本認識系統合モデルと前章で説明した各認識部品を用いて，回答者が「肯定/否定」で答えるような質問をする対話システムを開発し，そのシステムによる回答者の意図（肯定/否定）の識別率を調べた．ただし，音声認識の誤認識にともなう「その他キーワード」の認識は「肯定/否定」識別に悪影響を与えることが分かったため，「その他キーワード」の認識を無視するようにした．また，発話や頭の振りがなかったり学習データにない等の理由で肯定/否定の判断ができない場合は「否定」と識別するようにした．

10人の被験者に，各16問の質問対話を行ってもらった．その結果は**表3**のようになった．なお，この表の結果は同じ対話データについて，「音声と頭の振りの両

表3 認識系統合モデルの意図識別率（対話システムを用いた評価）
Table 3 Understanding rate of user's intentions (on the experiments using the system).

マルチモード	音声と頭の振りの両方の場合	91.9%
シングルモード	音声だけの場合	86.3%
	頭の振りだけの場合	55.6%

方のデータを用いた場合」「音声だけの場合」「頭の振りだけの場合」の3種類のフィルタをかけてから意図識別率を求めたものである．つまり16問の質問対話自体は各被験者が1回ずつ行っただけで，これらの3つの意図識別率を求めている．

この結果を見ると，実際の対話システムを使った場合でも，音声と頭の振りを両方とも用いることで，意図識別率が向上することが分かる．しかし，音声だけの場合の86.3%と比較すると，音声と頭の振りの両方を使っても91.3%と，マルチモーダル対話データベースを用いた評価ほどは識別率は向上していない．これは頭の振りだけの場合の識別率が55.6%とかなり低い（「肯定/否定」の識別なので，ランダムに回答しても50%の識別率が得られる）ことが原因だと考えられる．

なお，音声認識において「その他キーワード」の認識を無視しないようにすると，音声と頭の振りを両方とも用いた場合の識別率は78.1%にまで低下してしまった．

ベイズ識別関数の学習データは人手でラベリングを行ったマルチモーダル対話データベース中のデータであり，各認識部品の認識率が100%でのデータと考えられるから，現実のシステムのように認識率が100%でないシステムに適用した場合には，誤認識の影響で識別率が下がってしまう．今回のシステムでは「その他キーワード」についての誤認識が，意図識別に悪影響を与えることが分かったため，最終的にはこの「その他キーワード」の認識結果を無視するようにしたところ，音声と頭の振りを両方とも用いた場合の識別率が，どちらか片方だけ用いた場合の識別率より高くなった．そこで，発話内容が「肯定キーワード」だった場合の音声認識結果と，発話内容が「否定キーワード」だった場合の音声認識結果を調査した．その結果を**表4**に示す（1つの単語を発話しても，音声認識部は複数の単語を認識結果として出力する場合があるので，この表の認識率を横方向に加算すると100%を越える）．

表4より，発話内容が肯定キーワードの場合でも否定キーワードの場合でも，「その他キーワード」と誤認識されてしまった確率が高いことが分かる．

したがって，本認識系統合モデルが有効に機能するためには，ベイズ識別に用いる入力データに「その他キーワード」のような，誤認識率が高いものは入れないほうがよいことが分かる．

表4 音声認識結果
Table 4 Result of speech recognition.

認識結果 / 発話内容	肯定キーワード	否定キーワード	その他キーワード
肯定キーワード	100%	9.71%	25.2%
否定キーワード	15.2%	93.5%	39.1%

5.3 商品紹介システムを用いた評価

音声や頭の振りや顔の向きを用いてシステムと対話ができる「商品紹介システム」を開発して，そのシステムを使用した場合の快適度の主観評価と，対話にかかった時間の計測を行った．

タスクは，商品紹介システムから2つの商品についての解説を聞き出すというものである．商品紹介システムとの対話は，音声と頭の振りと顔の向きを使って行うことができるが，このすべてを用いることができる場合と，頭の動きだけ，あるいは音声だけしか用いることができない場合とで，快適度や対話にかかった時間がどのように変化するかを調査した．対話は「音声と頭の動きを両方認識できる場合」「頭の動きしか認識できない場合」「音声しか認識できない場合」の3種類をこの順番で行ってもらったが，被験者には事前に，何が認識できるのかを説明してから対話をしてもらった．

この「商品紹介システム」が識別するユーザの意図としては，「肯定/否定」以外に，商品名や，「それ」と発話したときの顔の向きがある．商品名については単純に音声認識だけを用いることにした．また，顔の向きは常時（約1秒間隔で）認識させており，「それ」という単語を音声認識したら，その単語が発話された時刻における顔の向きを調べ，これをユーザの意図と見なした．ただし，「頭の動きしか認識しない場合」における顔の向きは，約2秒間同じ方向を見続けた場合に，その方向をユーザの意図と見なした．

快適度の主観評価の尺度は以下のとおりである．
1：とても不快，2：不快，3：普通，4：快適，5：とても快適

10人の被験者に，この商品紹介システムとの対話を行ってもらい，快適度の主観評価と，対話にかかった時間を計測した．

快適度の主観評価の平均値と，対話にかかった時間の平均値とを**表5**に示す．

表5より，快適度に関しては，音声と頭の振りを両方とも用いた場合が一番快適になることが分かったが，頭の動きだけの場合や音声だけの場合に比べてそれほ

表5 快適度の主観評価と対話時間の平均値
Table 5 The comfortable degree and the conversation time.

用いたモード	快適度の主観評価の平均値	対話時間の平均値
音声と頭の動き	3.2	27.4 秒
頭の動きだけ	2.6	27.9 秒
音声だけ	2.9	25.2 秒

表6 3種類の群の快適度の主観評価と対話時間（秒）の平均値
Table 6 The comfortable degree and the conversation time (second) of the three groups.

感じ方	快適		変わらない		負担	
人数	4人		3人		3人	
	快適度	時間	快適度	時間	快適度	時間
声と頭	4.0	28.5	3.2	24.7	2.2	28.7
頭だけ	1.8	27.3	3.1	25.0	2.5	31.7
声だけ	2.6	25.0	2.9	24.7	3.2	26.0

ど大きな違いはないことが分かった．

対話に要する時間は「音声だけ」の場合が最も短くなったが，これは対話を3回続けて行ったことによる慣れによる影響と，「頭の動きだけ」の場合は顔の向きの判定に約2秒の時間を要してしまうことが理由と考えられる．

また，主観評価結果をもう少し調べてみると，被験者は次の3種類の群に分かれることが分かった．

(1) マルチモーダルを快適に感じる人（4人）
(2) マルチモーダルもシングルモーダルも変わらないと感じる人（3人）
(3) マルチモーダルを負担に感じる人（3人）

音声と頭の振りを両方とも用いた場合と，頭の動きだけの場合や音声だけの場合とで，主観評価にそれほど大きな違いが生じなかった原因としては，マルチモーダルを負担に感じたり，マルチモーダルもシングルモーダルも変わらないと感じる人が過半数存在していることによるのではないかと考えられる．

そこで，被験者を各群に分けて，快適度の主観評価の平均値と，対話時間の平均値を調査したところ**表6**のようになった．

それぞれの群に属する被験者に，その主観評価を行った理由を聞いてみたところ，以下のような回答が得られた．

(1) マルチモーダルを快適に感じる人
- 1種類のみの入力しかできないのに比べると，複数の入力が可能なほうが直観的に使える．
- 複数の入力が可能な方が不自然ではない．
- 1種類のみの入力しかできないとそれ（音声なら音声，頭の動きなら頭の動き）をしっかりと入力する義務を感じてしまうが，複数の入力が可能ならどの入力も中途半端に行っても大丈夫なのではないかと思える．

(2) どちらも変わらないと感じる人
- 自分の意図が正しく伝わるのなら，入力の種類は何でもかまわない．
- 入力のコツがつかめれば，どんな入力方法でもかまわない．

(3) マルチモーダルを負担に感じる人
- 複数の種類の入力が可能だとどれを使ったらよいか迷ってしまう．
- 複数の種類の入力が可能だとそのことが気になってしまう．
- 1種類の入力のみ可能なら，それに専念できる．

表7 アンケート内容と点数
Table 7 The questionnaires.

	設問	点数 YES	-	NO
1	コンピュータが好きである	2	0	−2
2	コンピュータが恐い	−2	0	2
3	キーボードをブラインドタッチできる	1	0	0
4	パソコンを持っていて，使っている	2	0	0
5	ワープロか電子手帳を持っていて，使っている	1	0	0
6	コンピュータを使うのは楽しい	2	0	−2
7	あまりコンピュータとはかかわりたくない	−2	0	2
8	ビデオの録画予約を難しいと感じる	−2	0	2
9	コンピュータとは対話したくない	−2	0	2
10	コンピュータと対話できたら楽しいだろうと思う	2	0	−2

そこで，**表7**のようなアンケートをとってみた．この表7の右側の点数は，コンピュータに好意的な人は大きくなり，コンピュータが嫌いな人は小さくなるように設定してある．たとえば，1問目の設問を見ると，コンピュータが好きな人は +2 点，好きではない人は −2 点，どちらでもない人は 0 点となる．このようにして，10問の設問に答えてもらい点数の合計を調査した．各設問の点数は基本的には ±2 点に設定し，経済的な影響等コンピュータに関する感情以外の要因がからむものについては 1 点や 0 点に設定してある．

各群の人々のアンケートの点数の合計の平均点数は**表8**のようになった．

つまり，コンピュータが嫌いな人（点数が負）は，

表8 アンケート結果
Table 8 Result of the questionnaires.

主観評価	人数	平均点数	各人の点数
快適に感じる	4人	6.5	4, 6, 7, 9
変わらない	3人	13.0	6, 15, 18
負担に感じる	3人	−0.7	−7, 0, 5

マルチモーダルインタフェースを負担に思い，非常に好意的な人（点数が10以上）はどんなインタフェースでもかまわないと思い，両者の中間層がマルチモーダルインタフェースを快適に感じる傾向があるのではないかと考えられる．

なぜ，今回の「商品紹介システム」のインタフェースを負担に感じる人がいたかを検討すると，以下の点があげられる．

- 音声認識できる単語が限られていた．
- 音声や頭の振りの認識率がそれほど高くなかった．
- 磁気センサを身に付けなければいけなかった．
- 手の動きを認識してくれなかった．

また，今回のタスクは，音声と頭の動きとを両方使うことで得られるメリットがあまりないものを選んだが，音声と頭の動きとを両方とも使わざるをえないタスク（たとえば，同じ色・形をした呼び方の分からない物体が画面上にいくつか置かれている場合に，その中のどれかを指定するタスク等，顔を物体に向けながら「それ」と発話して指示せざるをえないタスク）を用いれば，マルチモーダルインタフェースを快適に感じる人の割合が増えることが期待できる．

6. まとめ

本稿では，ユーザの意図の識別をマルチモーダル対話データベースに基づいて行う認識系統合モデルを提案し，このモデルの適用例として「肯定/否定」の識別を行う対話システムを開発した．

開発した対話システムを用いて意図識別率を評価した結果は，音声だけを用いたシングルモードの場合が86.3%，頭の動きだけを用いたシングルモードの場合が55.6%なのに対し，音声と頭の動きを両方とも用いたマルチモードの場合は91.9%となった．このように，音声と頭の動きを統合して扱うことで意図識別率が高くなり，本認識系統合モデルが有効に機能することを確認できた．

また，今回試作した「商品紹介システム」におけるマルチモーダルインタフェースを快適に感じるかどうかについては個人差があり，どんな人も快適に感じてくれるとは限らないことが分かった．

今後，さらにマルチモーダル対話データベースの拡充につとめ，より複雑な対話過程をモデル化し，また各認識部品の性能を向上させることで，より良い対話システムを開発していきたいと考えている．

謝辞 本研究は，新情報処理開発機構（RWC）の研究テーマ「多元情報を用いたヒューマンインタフェース」によってなされたものである．関係各位に深謝いたします．

また，本研究を行うに際して，日ごろご指導いただくマルチメディア事業化推進本部中島隆之技監ならびにご討議いただく新機能シャープ研究室の皆様に感謝いたします．特にマルチモーダル対話データベースに関しては新機能シャープ研究室の綿貫啓子氏に，音声認識に関しては新機能シャープ研究室の坂本憲治氏に多大な協力をいただきました．この場を借りてお礼申し上げます．

参考文献

1) Bolt, R.A.: Put-That-There: Voice and Gesture at the Graphics Interface, *Computer Graphics*, Vol.14 No.3, pp.262–270 (1980).
2) 末永康仁，間瀬健二，福本雅朗，渡部保日児：Human Reader: 人物像と音声による知的インタフェース，電子情報通信学会論文誌，Vol.J75-D-2, No.2, pp.190–202 (1992).
3) 山之内毅，大橋 健，松永 敦，江島俊朗：指示棒と音声を用いるコミュニケーション環境CoSMoSの設計（音声と統合処理），信学技報，PRU94-81, pp.71–78 (1994).
4) 新田恒雄ほか：自由発話音声入力と直指（直接指示）を利用したマルチモーダル対話システムの検討，信学技報，SP92-120, pp.37–42 (1993).
5) 吉岡 理，南 泰浩，鹿野清宏：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と音声入力の評価，信学技報，SP93-128, pp.1–8 (1994).
6) 小林哲則，竹内陽児，白井克彦：音声・マウス・キーボードによる統合的入力環境，信学技報，HC92-68, pp.19–24 (1993).
7) 村松光浩，大川茂樹，小林哲則，白井克彦：読唇の併用による音韻認識，信学技報，SP93-124 pp.47–54 (1994).
8) Vo, M.T. and Waibel, A.: Multimodal Human-Computer Interaction, *Proc. the International Symposium on Spoken Dialogue (ISSD-93)*, pp.95–101 (1993).
9) Bolt, R.A.: The Integrated Multi-Modal Interface, 電子情報通信学会論文誌，Vol.J-70-D, No.11, pp.2017–2025 (1987).
10) Ekman, P. and Friesen, W.V.: The Repertore of Nonverbal Vehavior, *Semiotica 1*, pp.49–98 (1969).

11) 黒川：人と機械のノンバーバル・コミュニケーション，第46回ヒューマン・インタフェース研究会資料 (1994).
12) 坂本，綿貫，外川：マルチモーダル対話解析，人工知能学会，SIG-SLUD-9401, pp.39–46 (1994).
13) Watanuki, Sakamoto, Togawa: Analysis of Multimodal Interaction Data in Human Communication, *ICSLP94*, 17.8, pp.899–902 (1994).
14) 外川，坂本：マルチモーダルデータベースに基づく対話の解析，1994年電子情報通信学会春季大会，A-342 (1994).
15) 綿貫，坂本，外川：マルチモーダル対話データの解析，日本音響学会平成6年度春季研究発表会講演論文集，1-7-20, pp.39–40 (1994).
16) 綿貫，坂本，外川：マルチモーダル対話データベースにもとづく対話解析，情報処理学会音声言語情報処理研究会研究報告，95-SLP-5, pp.17–22 (1995).
17) 上坂吉則，尾関和彦：パターン認識と学習のアルゴリズム，文一総合出版 (1990).
18) 湯浅夏樹，上田 徹，外川文雄：大量の文書データから自動抽出した名詞間共起関係による文書の自動分類，情報処理学会自然言語処理研究会報告，NL98-11，pp.81–88 (1993).
19) CD-HIASK，朝日新聞全文記事情報1990年版，紀伊国屋書店 日外アソシエーツ.
20) 三谷純司，外川文雄：複数カメラによる顔の向きの識別，第49回情報処理学会全国大会，7F-2 (1994).

(平成7年10月20日受付)
(平成8年 4 月12日採録)

湯浅 夏樹（正会員）

1967年生．1990年東京工業大学工学部情報工学科卒業．1992年同大学院理工学研究科修士課程（情報工学専攻）修了．同年シャープ（株）入社．現在，同社映像メディア研究所に勤務．自然言語処理，ヒューマンインタフェース等の研究に従事．

三谷 純司（正会員）

1965年生．1989年京都大学工学部情報工学科卒業．1991年同大学院工学研究科修士課程（情報工学専攻）修了．同年シャープ（株）入社．現在，同社映像メディア研究所に勤務．画像認識，ヒューマンインタフェース等の研究に従事．

外川 文雄

1953年生．1976年金沢大学工学部電子工学科卒業．同年シャープ（株）入社．1989～1991年米国オレゴン州でニューロコンピュータ技術共同開発．現在，同社映像メディア研究所に勤務．音声認識，文字認識，ニューラルネット，ヒューマンインタフェース等の研究に従事．