Vol.2013-HCI-151 No.21
Vol.2013-SLP-95 No.21
2013/2/2

対話中のユーザ状態逐次推定のための多段階識別手法に関する検討

千葉 祐弥[1,a)]　伊藤 仁[2]　伊藤 彰則[1]

概要：従来の音声対話システムは，ユーザが入力した発話の音声認識結果を基準として処理を行うため，ユーザの入力を待機している間にユーザの状態を推定することはできなかった．しかしながら，実環境下においては，ユーザがシステムのプロンプトに戸惑ってしまうなどで，入力を行うことができないという状況が度々起こる．こういったユーザに対して適切な応答を行うためには，従来の音声対話システムでは無視されていた「発話を行う前のユーザ状態」を考慮する必要がある．我々は，発話前のユーザ状態を 2 種類定義し，その推定手法について研究を行ってきた．ここまでの分析結果から，マルチモーダル情報を用いることで対象とするユーザの状態がある程度推定できることを結論づけた．この結果を踏まえ，本報告では動画像と音声から得られる情報を統合し，逐次的にユーザの状態を推定する手法について検討を行う．

1. はじめに

音声対話システムが柔軟な応対を行うためには，ユーザ状態の推定が必要である [1]．これまで，多くの研究が感情 [2], [3] や嗜好 [4]，システムへの習熟度 [5], [6] といった様々なユーザの内部状態に着目してきた．これらの研究は暗黙のうちに，対話システムがプロンプトを提示すれば，ユーザは常に入力を行うということを想定している．しかしながら，実環境下のシステムにおいては常にユーザの発話が入力されるとは限らない．例えば，システムのプロンプトの意図がわからなければ，ユーザは発話入力できず対話を放棄してしまうかもしれないし，入力が即答できるものでなければ入力内容を考える時間が必要である．

このような想定から，我々は少なくとも 2 つのユーザ状態が新たに考慮されるべきだと考えた．一つは，ユーザがどんな入力を行うべきかわからない状態であり，もう一つはユーザがシステムのプロンプトへの入力を考えている状態である．ここでは，前者を State A，後者を State B とする．これらのユーザ状態は，従来の対話システムでは区別せずに単に入力に時間がかかっているユーザとして扱われていた．本研究の目的は，上記の 2 状態に「円滑に対話できている状態」に相当する State C を加えた 3 つの状態を識別することである．これらのユーザ状態はユーザの入力の有無によらず現れるので，「対話ターン中の」ユーザ状態と定義した．人間同士の対話においては，多かれ少なかれ，本研究が目指すような対話相手の状態推定が行われている．例えば，Feeling of Another's Knowing (FOAK) [7] と呼ばれる，対話相手が自分の質問に答えられそうかどうかを推量する能力などがこれに当てはまる．この推量は，視覚的情報，音声情報の連動によって行われている [8]．このような人間同士の対話のやり方を模倣することで音声対話システムの性能は向上できると考えられる．

本研究では，上述したユーザ状態の推定を行うために視覚的特徴量と音声特徴量を同期させて用いることを考えた．近年，バイモーダル特徴量の統合は自動感情認識の分野で盛んに検討されている [9], [10]．対象とする対話ターン中のユーザ状態はユーザの感情と似た性質を持つため，推定の特徴量はこれらの研究で用いられている特徴量とほぼ同質のものを採用した．すなわち，音声に含まれる韻律的情報，声質的情報，顔の特徴点である．本研究が従来の取り組みと決定的に異なるのは，目的とする対話システムの性質上，ユーザの発話終端を処理のトリガーとして扱えないため，逐次的な推定が必要となる点である．

以前の報告 [11] では，収集したマルチモーダル対話データを人間によって評価し，対象とするユーザ状態が視覚的情報，音声情報のみが与えられた場合も全ての情報を与えられた場合と同様の傾向で評価されることを確かめた．この結果をもとに，本報告では対話ターン中のユーザ状態を自動推定する手法を検討する．本稿で提案する手法では，

[1] 東北大学
Tohoku University
aoba 6-6-5, Aramaki, Aoba-ku, Sendai, Miyagi
[2] 東北工業大学
Tohoku Institute of Technology
Kasumicho 35-1, Yagiyama, Taihaku-ku, Sendai, Miyagi
[a)] yuya@spcom.ecei.tohoku.ac.jp

ⓒ 2013 Information Processing Society of Japan

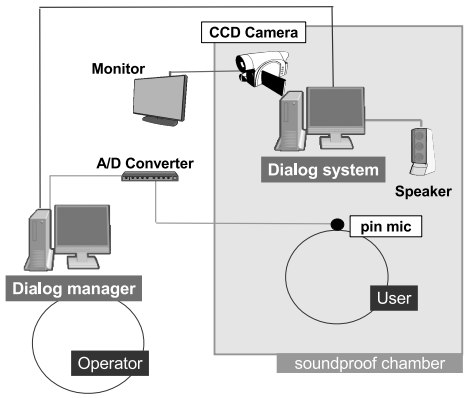

図 1 実験環境

ニューラルネットワークを多段に用いて複数のモダリティを統合し，ユーザ状態の推定結果を逐次的に出力する．

2. 実験用データ

実験用のデータは Wizard of Oz（WOZ）法に則って収集を行った．WOZ 法は被験者に，実験者が操作する対話システムと対話してもらうことで，より自然な対話データを収録する方法である．図 1 に実験環境を示す．

収録は防音室内で行われた．被験者はモニターと対面して座り，モニターには被験者の視線をできるだけ正面に集める目的で，簡単な表情変化ができるエージェントが表示された．発話を収録するため被験者にはピンマイクを装着してもらい，同時にモニターの上方に設置したビデオカメラによって被験者を正面から撮影した．実験者は防音室の外でユーザの発話とビデオカメラの映像を観察しながらエージェントを操作した．音声信号は 16 kHz，16 bit の PCM 形式で保存し，動画像は 29.97 fps，色深度 24 bit のカラー画像として保存した．対話のタスクとして，システムがユーザに質問を行い，ユーザはそれに対する入力を行うという「一問一答」型のタスクを用意した．これは，できるだけユーザに対象とする状態を表出させる目的で導入されたものである．ユーザは例えば「今年の最高気温は何度でしたか」といったような，常識的な知識だがすぐに答えるのは難しい質問や，「会員番号を入力して下さい」といった事前に記憶しておいた内容を答えさせる質問などの 44 個の質問に答えた．ユーザが入力内容を考えているなどで入力が行われないときは，15 秒経過する度に同じ内容の質問が繰り返された．また，どうしても答えられない質問には「わかりません」と入力することを許可したが，最低限の利用に留めるよう求めた．対話の収集には男性 14 名，女性 2 名の計 16 名が参加した．データ収集後，対話データは一組のシステム発話とそれに対するユーザの応答を含むセッションとして分割した．この時，ユーザの入力が行われなかったセッションは質問の繰り返しの直前までで分割を行った．ここで，我々はシステムのプロンプト終了直後からユーザの入力が始まるまでの区間を *latency* と定義する．以前の検討により，*latency* が短いものは State C と評価されがちであることがわかっているので，*latency* が 5 秒よりも大きいものを対象として 5 人（男性 4 名，女性 1 名）の評価者によって評価を行なってもらった．尚，評価者 5 名は対話の収集には参加していない．

表 1 評価結果

State A	State B	State C	Total
59	195	538	792

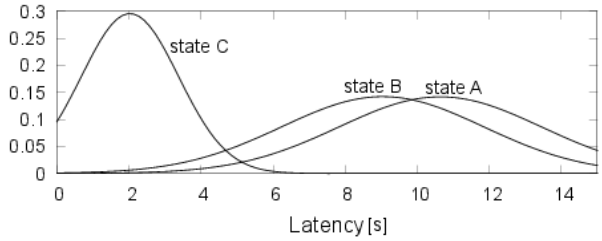

図 2 latency の分布

表 1 に評価結果を示す．最終的なラベルは 5 人の評価結果の多数決で決定された．*latency* が 5 秒よりも短いセッションは全て State C に分類した．また，1 つのセッションは発話がシステムのプロンプトと重なり，音響的特徴量が得られなかったため除外した．

2.1 *latency* の分布

図 2 は *latency* の分布を示している．State C とその他のセッションは *latency* の大きさによって明らかに分割できる．実際，事後確率による 5 交差検定では，State C とその他のクラスを 98.99% の精度で識別することができた．一方で，State A と State B のセッションは分布が重なっており，*latency* による識別だけでは不十分であることがわかる．

3. ユーザ状態の二段階推定

以前の検討 [12] では，システムのプロンプト終了直後からユーザの入力発話の開始直前までで観測された全ての特徴を用いて，SVM による固定次元の識別を行なった．しかしながら，本研究の最終的な目標はシステムのプロンプトに対して応答が困難なユーザを補助できるシステムを構築することであり，ユーザ状態の逐次的な推定が不可欠である．本研究では，対話の各ターンで観測されるユーザの仕草や音声をビデオカメラとマイクロフォンにより常時観測し，観測された特徴系列を利用することでユーザの状態推定を行う．識別器としてはニューラルネットワークを用い，フレームごとのユーザ状態を推定する．各時刻のユーザ状態は図 3 に示すように，多段に用いて推定する．前段

ⓒ 2013 Information Processing Society of Japan

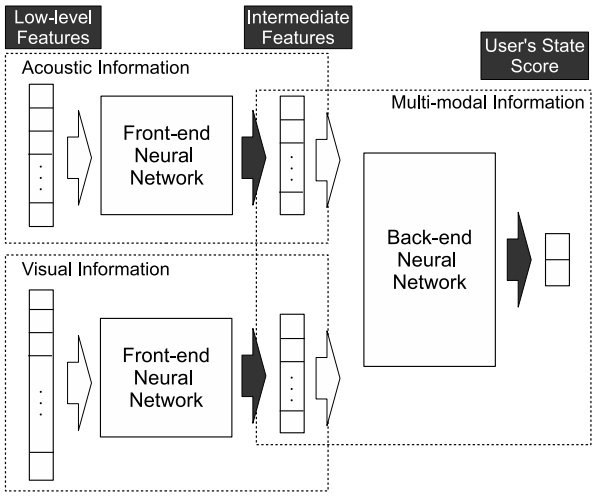

図 3 ユーザ状態の二段階推定

表 2 音響特徴量抽出の条件

	MFCC	ΔPitch	Zero cross ratio
Frame width	25.0 ms	7.5 ms	10.0 ms
Frame shift	10.0 ms	10.0 ms	10.0 ms

のニューラルネットワークは低次の特徴量を入力とし，声質のカテゴリや顔表情といった，記号的なクラスの尤もらしさをスコアとして出力する．出力されたスコアはそのまま後段のニューラルネットワークの入力として用いられ，最終的にユーザ状態のスコアを結果として出力する．以降では後段の特徴量（すなわち前段の出力値）を中間特徴量と呼ぶ．多段階の識別には，中間特徴量に一般的なクラスを用いることで，1) 結果の解釈と分析が容易になる，2) 中間特徴量の学習に汎用的なデータが使える，3) 感情推定や相槌の生成など，より一般的なマルチモーダル対話の問題への応用が可能になる，といった利点がある．

4. 時系列マルチモーダル特徴量の抽出

4.1 低次の音声特徴量

対話ターン中のユーザ状態は言語学におけるパラ言語情報に関係がある．パラ言語情報は音声の声質やイントネーションによって伝達されることが知られている [13]．したがって，本研究においても音声スペクトルの周波数的な特徴や基本周波数（$F0$）を表現する特徴量を利用することを考えた．本稿では 3 種類の音声特徴量を導入する．以下でそれぞれについて概説する．

スペクトルの周波数的特徴を表現するため，低次の音声特徴量として MFCC を用いた．MFCC は音声認識に一般的に用いられる特徴量であり，間接的にスペクトルの包絡構造を表現している．ここでは，フィルタバンクのチャネル数を 24 とし，下から 12 番目までの係数にパワーを含めた 13 次元を基本の係数として用いる．それぞれの係数に対しては Δ 成分と ΔΔ 成分を計算し，全体で 39 次元の特徴量となる．差分特徴量計算の窓幅はいずれも 5 とした．

音声のイントネーションは $F0$ 軌跡によって表現される．$F0$ の変動は，ユーザの思考状態の表出である有声休止 [14] の判定にも有用である．本研究は $F0$ を正規化相互相関 [15] によって取り出し，対数尺度に変換することで抽出する．音声の基本周波数は個人差，男女差が大きいため，一次差分を低次の音声特徴量とする．差分特徴量計算には MFCC と同じく前後 2 フレームの値を用いた．

上述した特徴量に加えて，零交叉率を低次の音声特徴量として用いた．これは，有声区間と無声区間を識別するためである．零交叉率は振幅とともに音声区間検出（VAD）の特徴量として一般的に用いられている．振幅については MFCC のパワーで代替した．

基本的な低次音声特徴量の抽出条件を表 2 に示す．音声特徴量は 10 ms を同期の基準としており，フレームシフトを統一した．

4.2 低次の画像特徴量

動画像から得られる情報のうち，最も重要なものはユーザの顔の運動である．そのため，本稿では Constraint Local Model (CLM) [17] によって顔の特徴点を抽出した．この方法では，まず画像中から顔検出によって顔の領域を決定した後，特徴点モデル（図 4）の当てはめが行われ，それ以降のフレームでは検出された特徴点が追随される．図 5 は顔特徴点の検出結果の例を示している．

顔特徴点の検出は頑健であるが，主に顔検出の誤りとオクルージョンによって誤検出が起こる．顔の検出誤りは実環境下での利用においても比較的容易に対策がとれると考えられるが，オクルージョンによる誤検出は基本的に対話中に生じるため訂正が困難である．したがって，最初のフレームでの顔の誤検出のみ人手で修正を行った．この操作により，顔特徴点の誤検出が含まれるのは全体の約 5% に限られる．顔の特徴点は検出された顔領域の大きさによって正規化され，各特徴点はモデルの中心からの相対座標で表現した．特徴点は全部で 66 点あり，これらが二次元空間上の座標で表されるため，低次の画像特徴量の次元数は全体で 132 である．

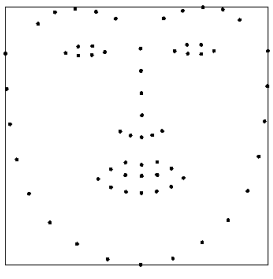

図 4 顔の特徴点

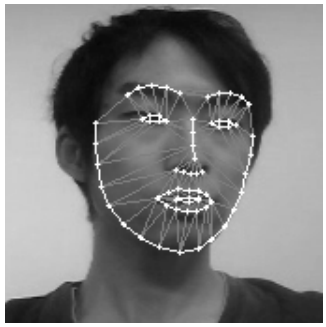

図 5 特徴点検出の例

ⓒ 2013 Information Processing Society of Japan

表 3 音声クラスのラベル

System (AS)	システムのプロンプト
Input (AI)	ユーザの入力発話
Filler (AF)	フィラー
Breath (AB)	気息音
Self speech (ASE)	ユーザの独り言
Whisper (AW)	ささやき声
Silence (ASI)	無音区間

表 4 画像クラスのラベル

顔向き	表情
システムの方を向いている（DON）	平静（EN）
システム以外の方を向いている（DOF）	笑顔（ES）
	驚き（EO）
	しかめ（EW）

4.3 中間特徴量

後段の識別器は，定義された音声のクラスと画像のクラスそれぞれに対する推定スコアを特徴量として用いる．音声及び画像クラスは実験者が定め，各クラスの観測に基づいて人手でラベル付けを行った．このラベルは前段の識別器の学習の教師信号として用いられる．表 3，4 にそれぞれ音声クラス，画像クラスのラベルの種類を示した．前段の識別器の出力はそのまま後段のニューラルネットワークに入力される．ここで，対話ターン中のユーザ状態の推定には現在のフレームで観測された特徴量だけでなく，各クラスの継続時間や出現回数などの文脈的な情報が重要であると考えられる．そこで，観測系列の時間的な性質を取り込むため，スコアの一時差分を中間特徴量に加える．これは，

$$\Delta p_{tc} = \frac{\sum_{\theta=1}^{\Theta} \theta(p_{(t+\theta)c} - p_{(t-\theta)c})}{2\sum_{\theta=1}^{\Theta} \theta^2} \qquad (1)$$

によって計算される．ここで，p_{tc} は各中間特徴量の値，すなわち時刻 t におけるクラス c のスコアである．また，Θ は差分計算の窓幅であり，実験によって $\Theta = 4$ と定めた．一時差分を含めた中間特徴量は 26 次元の特徴量である．

5. 識別実験

5.1 各モダリティの同期処理

音声特徴量はフレーム処理の条件から 10 ms おきに特徴量が抽出されるが，画像特特徴量はフレームレートの値に依存して取得されるため，本稿の条件では約 33 ms おきにしか抽出できない．各モダリティの同期に関しては様々なレベルでの統合が検討されているが，ここでは，中間特徴量を後段のニューラルネットワークに入力する時点で同期処理を行うこととした．本稿では，直前に得られた各画像クラスのスコアを 10 ms おきに複製することによって同期を行う．

表 5 各層のユニット数

	前段			後段
	音声	顔向き	表情	ユーザ状態
Input	42	133	133	27
Hidden	80	15	20	15
Output	7	2	4	2

5.2 実験条件

今回構築したニューラルネットワークは全て入力層，隠れ層，出力層を持つ三層のネットワークであり，各層はバイアスユニットを持つ．隠れ層の活性化関数にはシグモイド関数を用い，出力層の活性化関数にはそれぞれのクラスの推定値をスコアとして出力するためソフトマックス関数を用いた．表 5 に各層のユニット数を示す．ニューラルネットワークの入力層のユニット数は入力特徴量の次元数に等しく，表中ではバイアスユニットの数を含んだものを示した．隠れ層のユニット数に関しては実験により定めている．以降の実験は全て 5 交差検定による識別の結果である．

5.3 実験結果と考察

はじめに，中間特徴量の精度を表 6，表 7 に示す．ただし，本稿の枠組みでは前段の識別器から得られた各クラスのスコアを全て後段の識別器の入力とするため，必ずしも前段の識別精度の向上が最終的な識別精度の向上につながるとは限らない．前段の識別器の性能はフレームごとに評価を行った．すなわち，

$$\hat{c}_t = \arg\max_c p_{tc} \qquad (2)$$

である．ここで，$\hat{c}_t$ は時刻 t における識別結果である．表より，前段の識別器は，音声に関しては AS，AI，ASI といったクラスで高い性能を示した．これは，これらのクラスがその他のクラスと比較して音響的に明らかに異なっているためである．一方で，ASE や AW といったクラスは AI や AB といった他のクラスと似た特徴を持つため，混同される傾向にあった．この結果は，本稿で設定したクラスラベルの種類が冗長である可能性を示しており，ラベルの選択には再考の余地があると言える．画像クラスについては，EN や顔向きクラスについて高い精度が得られた．しかしながら，いくつかの表情は頭部運動が雑音になって上手く識別できない傾向があった．したがって顔の特徴点については得られた特徴点の座標をそのまま用いるよりも，ある程度冗長性を取り除いた上で識別に利用する必要があると考えられる．

後段の識別に関しては，State A と State B の 2 クラス識別を行った．これは，前述した通り，State C とその他のクラスが *latency* によって明らかに判別できるためである．対話ターン中のユーザ状態の推定に関しては，逐次的

ⓒ 2013 Information Processing Society of Japan

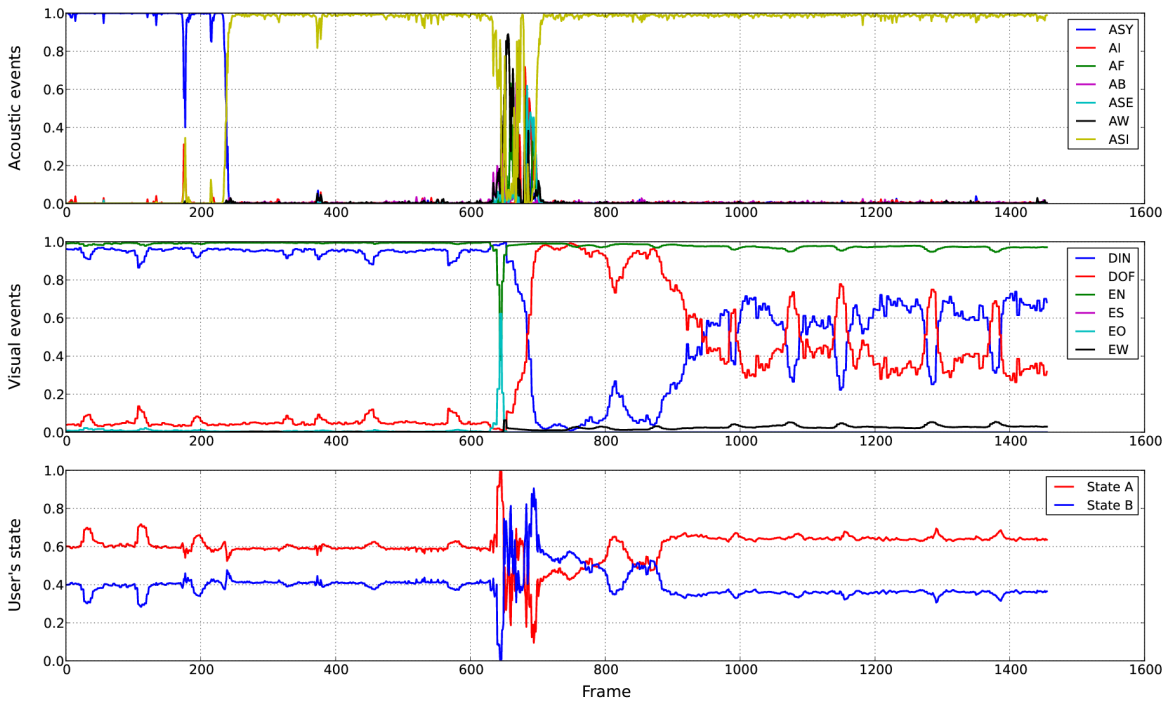

図 6 識別結果の例

表 6 音声クラスの識別精度（%）

AS	97.26
AI	80.11
AF	49.03
AB	36.06
ASE	17.02
AW	13.84
ASI	97.92

表 7 画像クラスの識別精度（%）

顔向き		表情	
VDI	94.74	EN	96.37
VDO	59.72	ES	36.49
		EO	0.56
		EW	14.56

表 8 ユーザ状態の識別精度（%）

State A	State B	Total	Harm.
52.58	65.13	62.22	58.18

に推定結果を出力しなければならないという性質上，その精度の評価方法が問題になる．本稿では最終的な識別結果 $\hat{c}$ を

$$\hat{c} = \arg\max_{c}\left(\max_{1\le t\le T}(p_{tc})\right) \qquad (3)$$

によって選択することとした．ここで，T は推定結果を決定する時刻に相当するパラメータである．今回の検討では，初期検討として T をそれぞれのセッションの継続長と等しく設定し，セッション終了までの全ての区間を考慮する．また，全体の識別結果はデータの量が State B に偏っているため（表 1 参照），識別結果が State B に偏った方が高くなる傾向がある．本稿では，識別結果はなるべくバランスしていた方が望ましいと考え，5 交差検定によって得られた平均の識別精度の調和平均（Harm.）を識別器の性能の指標として用いることとした．表 8 に最も高い調和平均値が得られたときの識別結果を示す．

ここで，識別結果の例を図 6 として示した．例は「ゲートウェイの IP アドレスはいくつですか？」という質問が行われたセッションである．ユーザはこの質問に対して 600 フレーム付近で「ゲート？」という独り言を行った後，900 フレーム付近で首をかしげる，といったような応答を行った．これは，ゲートウェイという言葉を知らなかったために起こった応答である．この例では，独り言付近で AW（ささやき声）のスコアが ASE（独り言）などのその他のクラスのスコアと共に上昇している．また，900 フレーム以降では顔向きの変動が大きくなっていることがわかる．音声に関しては，識別誤りが含まれるものの，概ね期待した結果が得られていると言えるが，画像に関しては推定誤りである．これは，傾げに関するラベルを用意していなかったことが原因であり，この結果からも特に画像クラスに関するラベルや特徴量の選択には改善の余地があることがわかる．また，最終的なユーザ状態の推定では，フィラーやささやき声など特殊な音声クラスが生じる区間で State B（応答を考えているユーザの状態）のスコアが高くなり，顔向きや表情が変化する区間で State A（質問の意味がわから

ⓒ 2013 Information Processing Society of Japan

表 9 単階層の識別器の精度（%）

State A	State B	Total	Harm.
38.79	87.18	75.96	53.69

ないユーザの状態）のスコアが高くなる傾向があった．これは，ある程度予め想定した結果に近いものである．しかしながら，最終的な精度は実際の対話システムに導入するにはまだ不十分であり，識別精度の向上を行う必要がある．

5.4 単階層の識別手法との比較

本稿では，中間特徴量を用いることで二段階の識別を行う手法を提案した．一方で，同様の枠組みでは低次の特徴量から直接ユーザの状態を推定するといった構造が考えられる．すなわち，173(41 + 132) 次元の特徴から 2 つのユーザ状態のスコアを出力する単一のニューラルネットワークを学習する方法である．本稿の最後にこれらの識別結果の比較を行う．

単階層の識別器は，入力層，隠れ層，出力層の三層のネットワークとし，隠れ層のノード数は実験により 10 と定めた．結果を表 9 に示す．結果より，全体としては高い識別率が得られるものの，識別結果は偏っていることがわかる．また，スコアも一方の状態に傾きがちであり，推定結果の判定が難しいという問題があった．二段階の識別では図 6 に示した通り，このような極端な変化は起こりにくい．以上のことから，本研究では多階層の推定結果が望ましいものと考え，前述した提案法のメリットも勘案し，検討を深めていく予定である．

6. まとめと今後の課題

音声信号の MFCC, $\Delta F0$, 零交叉率と顔画像系列から得られる顔の特徴点を識別の特徴量として，対話ターン中のユーザ状態をフレーム毎に推定する手法の検討を行った．提案した手法では，最終的な識別結果に関して 62% の識別精度が得られた．

今後は識別精度の向上のために，まずは特徴量の改善を検討する予定である．特に，画像特徴量である顔特徴点は今回の実験により非常に冗長性が大きいことがわかった．個人差による影響も大きいため，ある程度の前処理が必要であると考えられる．そのため，特徴点抽出の段階で PCA などを用いた次元削減を導入する予定である．また，人間による対話データの評価実験では，被験者の視線の動きが判定の基準となるケースが多いことがわかっている．本稿で用いた顔特徴点モデルは視線の動きを追随することはできないため，眼球運動を表現する特徴量を改めて抽出することが必要である．加えて，中間特徴量のラベルの選択についても再考の余地がある．現在のラベルの種類は冗長性が大きく，タスクへの依存性も大きいため，より汎用性の高いラベル選択を検討する予定である．

謝辞 本論文は、総務省の「大規模災害時における移動通信ネットワーク動的制御技術の研究開発」（平成 23 年度一般会計補正予算（第 3 号））による委託を受けて実施した研究開発による成果である．

参考文献

[1] A. Kobsa. User modeling in dialog systems: Potentials and hazards. *AI&Society*, 4:214–231, 1990.
[2] K. Forbes-Riley and D. Litman. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53:1115–1136, 2011.
[3] A. Metallinou, A. Katsamanis, and S. Narayanan. A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs. In *Proc. ICASSP*, pages 2401–2404, 2012.
[4] A. N. Pargellis, H. K. J. Kuo, and C. H. Lee. An automatic dialogue generation platform for personalized dialogue applications. *Speech Communication*, 42:329–351, 2004.
[5] K. Jokinen and K. Kanto. User expertise modelling and adaptivity in a speech-based e-mail system. In *Proc. COLING*, 2004.
[6] F. Rosis, N. Novielli, V. Carofiglio, A. Cavalluzzi, and B. D. Carolis. User modeling and adaptation in health promotion dialogs with an animated character. *J. Biomedical Informatics*, 39:514–531, 2006.
[7] S. E. Brennan and M. Williams. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *J. Memory and Language*, 34(3):383–398, 1995.
[8] M. Swerts and E. Krahmer. Audiovisual prosody and feeling of knowing. *J. Memory and Language*, 53(1):81–94, 2005.
[9] J. C. Lin, C. H. Wu, and W. L. Wei. Error weighted semi-coupled hidden markov model for audio-visual emotion recognition. *IEEE Trans. Multimedia*, 14:142–156, 2012.
[10] Y. Wang, L. Guan, and A. N. Venetsanopoulos. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Trans. Multimedia*, 14:597–607, 2012.
[11] Y. Chiba, M. Ito, and A. Ito. Effect of linguistic contents on human estimation of internal state of dialog system users. In *Proc. IWFBD*, 2012.
[12] Y. Chiba, M. Ito, and A. Ito. Estimation of user fs internal state before the user fs first utterance using acoustic features and face orientation. In *Proc. HSI*, 2012.
[13] C.T. Ishi, H. Ishiguro, and N. Hagita. Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality. *Speech Communication*, 50:531–543, 2008.
[14] M. Goto, K. Itou, and S. Hayamizu. A real-time filled pause detection system: Toward spontaneous speech dialogue. In *Proc. Eurospeech*, pages 227–230, 1999.
[15] B.S. Atal. Automatic speaker recognition based on pitch contours. *J. Acoustical Society of America*, 52:1687–1697, 1972.
[16] L. R. Rabiner and M. R. Sambur. An algorithm for determining the endpoints of isolated utterances. *The Bell System Technical Journal*, 54:297–315, 1975.
[17] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *Int. J. Computer Vision*, 91:200–215, 2011.

ⓒ 2013 Information Processing Society of Japan