Vol.14 No.3, 2012

原著論文

ボイスチャット中の音声情報に基づく会話活性度推定方法の検討

守屋 悠里英[*1] 田中 貴紘[*1] 宮島 俊光[*1] 藤田 欣也[*1]

Study of Activation Level Estimation using Phonetic Information during Voice Chat

Yurie Moriya[*1], Takahiro Tanaka[*1], Toshimitu Miyajima[*1]and Kinya Fujita[*1]

Abstract – **In this paper, we discuss the feasibility of conversational activation level estimation by using phonetic and turn-taking features. At first, we experimentally collected six groups of 3-persons conversational voice at three different activation levels. Then, we calculated the phonetic and turn-taking features, and analyzed the correlation between the features and the activity level. The analysis revealed that response latency, overlap rate and speech rate correlate with activation level and are less sensitive to individual deviation. Then, we formulate multiple regression equation, and examined the estimation accuracy using another six 3-persons groups. The results demonstrated the feasibility to estimate activation level at about 24% root-mean-square-error.**

Keywords : **activity level of conversation, phonetic feature, estimate, speech information processing, avatar**

1. はじめに

高速広帯域ネットワークの普及に伴い，様々な遠隔コミュニケーションシステムが開発され，普及している．中でも，電子メールや電子掲示板，テキストチャットなどのテキストを主たる媒体とするシステムは，マイクロブログのように，さらに形態を変化させながら社会に浸透している．その一方で，テキストコミュニケーションは，感情伝達が難しい点なども指摘されている [1]．他方，ボイスチャットやビデオチャットなどの音声や映像を媒体とするシステムも，無料ソフトウェアの配布や機器の低価格化を背景に，利用者層が広がっている．ビデオチャットは会話相手の状態を視認可能であるという利点を有するが，背景画像等を介した意図しない情報流出の可能性が懸念される．ボイスチャットは導入が容易であるものの，多人数会話においては，発話者識別の困難さや，発言の少ないユーザの存在感の希薄化といった課題がある．そこで，仮想空間とユーザの分身となるアバタを利用することで，意図しない情報流出を抑制しつつ，他のユーザの存在を感じながら会話することを可能とする，様々なシステムが開発されている．

他方，一般的な対面会話場面においては，相手の言葉だけでなく，表情やジェスチャといった非言語（ノンバーバル，以下 NV) 媒体を介したコミュニケーションを行っていることが知られている．マジョリーは，メッセージの伝わり方の 65%は NV コミュニケーションによるとし [2]，マレービアンは，言語と非言語で相反する情報を同時に相手に送った場合，メッセージ伝達要素は顔の表情が 55%，音声が 38 %，言葉が 7%を占めると報告した [3]．これらの研究は，ヒトがコミュニケーションにおいて，多くの NV 情報を認識していることを示すものであり，アバタを介したコミュニケーションにおいても，NV 情報の欠落は，会話感を損ない，コミュニケーションを阻害すると考えられる．したがって，適切な NV 情報をアバタに表出させることが，自然なコミュニケーションを実現する上で必要になってくる．

アバタに視線や表情，ジェスチャといった NV 情報の表出機能を実装した場合，手動での制御は煩雑であり，ユーザの負担が増加する．そこで，NV 情報の自動表出制御に関する研究が行われてきた．渡辺らの InterActor は，話者の音声情報を基に自動でアバタの頷きやジェスチャ制御することで，会話が促進されることを示した [4]．筆者らも，音声情報に基づくアバタの表情や視線制御，発話強調動作の生成手法を提案するとともに，実験的に有効性を検討してきた [5]．これらの研究は，個々の NV 情報の制御の有効性を示したが，NV 情報間の整合性に関しては，特に検討されていなかった．しかし，ヒトの動作は自身や場の状態に応じて変化することから，アバタの動作の大きさや速さ，動作の発生頻度などを統一的に制御 [6] することで，より自然な印象が得られ，ひいては会話の促進にもつながると期待される．そこで，筆者らは，会話の活性度に応じて身体各部の動作の大きさや発生頻度などを統一的に変化させる，アバタの動作制御アルゴリズムを検討してきた [7]．しかし，次章に詳述するように，会話活性度に関連する先行研究の大半が，各種音

*1: 東京農工大学大学院
*1: Tokyo University of Agriculture and Technology

声指標と活性度の相関の報告に留まっており，指標を用いて活性度を推定した際の誤差や，個人差等の観点からの推定指標としての適否に関しては，ほとんど検討されていなかった．

そこで，本論文では会話の盛り上がりの程度の主観的印象を会話活性度と定義し，自然会話における活性度の自動推定の可能性を検討する．これまでに，活性度の推定には，手の動きなどのジェスチャから推定する手法も提案されている [8]．しかし，動作に基づく活性度推定は，カメラなどの機材を必要とし利用環境が制限される．そのため，本研究では，音声情報のみに基づく活性度推定を検討対象とした．会話実験により収集した 3 人会話音声から算出した各種音声指標と，主観的活性度の相関を分析したところ，オーバーラップ率と平均発話潜時，発話率の 3 者に高い相関が認められたため，これらを用いて活性度の推定を試行したところ，キャリブレーションが不要で個人差や環境の影響を受けにくい，会話活性度推定の可能性を示唆する結果が得られたため，報告する．

2. 関連研究

2.1 会話活性度と音声指標

感情等の影響により会話中の音声が変化することは，多くの先行研究によって指摘されている．感情の変化は韻律情報の変化を伴い，例えば怒りではピッチと強度は高く [9]，話速は速くなり [10]，喜びでも同様に，ピッチと強度が高く，話速が速くなる [9]．他方，嫌悪のときにはピッチと強度は低く [10]，話速は速くなる [11]．悲しみのときには，ピッチと強度は低くなり，話速は遅くなる [10] とされている．また，弛緩感情ではピッチが低くなり抑揚もなくなることが指摘されている [12]．先行研究で報告されている，感情による韻律指標の変化を表 1 にまとめる．一度生じた感情は，特に強い外的要因が作用しない限り，ある程度の時間継続すると考えられるため，会話中に生じた感情は，十分に検出可能な変化を会話音声にもたらすものと考えられる．

人間の感情について，ラッセルらは感情を快-不快と活性-非活性の 2 軸で表す円環モデルを提唱している [13]．なんらかの方法を用いて，アバタの感情表出を自動制御した場合，誤推定による意図しない不快感情の表出は，コミュニケーションを阻害すると考えられる．しかし，自然会話音声からの感情種の推定は，現段階では容易とは言い難い [14]．また，不快感情は理性によって抑制することが，社会通念上望ましいとされることから，実際の会話場面での表出頻度は必ずしも高くない．そこで，本研究では，知人同士によるカジュアルな会話場面での利用を想定し，不快感情の表出は，真に必要な場面において手動で制御するものとして，活性度のみの 1 軸の自動制御について考える．ここで，韻律情報と感情の関連性に関する先行研究の知見を円環モデルにあてはめてみると，怒りや喜びは活性度が高く，嫌悪や悲しみ，弛緩は活性度が低い状態に相当する．したがって，韻律情報と会話活性度の間には相関関係があるものと推測される．すなわち，韻律指標に基づく会話活性度推定の可能性が期待される．

表 1　韻律指標と感情の関係 (先行研究より)
Table 1　Relationship between emotions and phonetic features (summary of preceding studies).

	感情				
	怒り	喜び	嫌悪	悲しみ	弛緩
ピッチ	高い	高い	低い	低い	低い
強度	強い	強い	弱い	弱い	-
話速	非常に速い	速い	速い	遅い	-
抑揚	-	-	-	-	ない

他方，話者の内部状態は，個人の音声の韻律だけではなく，会話の構造にも影響することが指摘されている．例えば，活性度が高くなると，話者間の発話が重なるオーバーラップの頻度が高くなることや [15], [17]，肯定的発話で話者交替時の発話潜時が短くなることが報告されている [16]．発話全体においても，活性度が高くなると無音時間が減少し，発話率が上昇することが指摘されている [17]．また，会話を続けることで話者間の行動が類似する同調傾向は，特に女性グループで観察され [15]，会話活性度との関連が期待される．活性度により変化すると予想される，会話構造指標を表 2 に示す．

2.2 個人と場の活性度とアバタ制御への応用

会話の盛り上がりの程度に連動して，うなずきや上肢による強調動作が速く大きくなるなど，アバタ動作の大きさや速度等が適切に変化すれば，音声から受ける印象とアバタの挙動が一致し，より自然なコミュニケーションにつながるものと期待される．瀬島らは，オーバーラップ頻度の重み付き移動平均によって会話

表 2　会話構造指標と活性度の関係 (先行研究より)
Table 2　Relationship between activation level and conversational features (summary of preceding studies).

	活性度	
	高い	低い
オーバーラップ頻度	高い	低い
発話潜時	短い	長い
発話率	高い	低い
同調傾向	強い	弱い

の活性化を検出し，仮想聴衆の動作に前傾などの強調を加えると，主観的な楽しさや好ましさが向上したと報告している [18]．そこで筆者らも，会話の盛り上がりの程度を活性度と定義し，活性度に応じたアバタ動作の自動制御法を検討してきた [6]．なお，先に述べたように，本研究においても先行研究と同様に，主観的な会話の盛り上がりの程度を会話活性度と定義する．

会話音声からの活性度推定が可能になれば，アバタ動作の自動制御が可能になると期待される．複数人会話における心理的な盛り上がりの程度には，参加者ごとのばらつきが伴うため，活性度は個人ごとに異なる値を取ると考えられる．一方で，会話の盛り上がり感は必ずしも話者のみが感じるものではなく，受話が主体の参加者も，会話の場の総体的な変化に同調して漠然とした盛り上がりを感じる可能性が考えられる．このとき，場の総体的な活性度は，会話参加者によってもたらされるものであるが，立場や発話率の相違などを考えると，個人の活性度の単純平均等で代表されるとは考えにくい．そこで本研究では，発話行動に反映される個人ごとの活性度とは別に，受話主体の参加者に影響を及ぼす会話の場全体の活性度を想定し，“ 個人の活性度 ”と“ 場の活性度 ”の 2 つの会話活性度の推定に向けた実験的検討を行う．

また，実際にアバタの動作を制御する場面では，発話の多いユーザと発話が少ないユーザ，両方のアバタの活性度を制御する問題と考えることもできる．発話の多寡にかかわらず，適切に活性度を制御するためには，本人の発話に基づく個人の活性度と，場全体の発話に基づく場の活性度を推定するとともに，加重平均などの手段によって，両者を個々のアバタに反映することが望ましいと考えられる．

3. 実験

活性度の各音声指標への影響を調べるため，以下の実験手順で，活性度の異なる会話音声を取得した．実験には，自然な音声会話が可能であり，かつ，遅延や画角などが，NV 情報を自動制御したアバタ音声会話に近くなるように，ビデオチャットシステムを用いた．

3.1 実験方法

実験は，被験者を 3 人 1 組とし，異なる 3 つの条件でビデオチャットによる会話を行わせ，音声を記録した．被験者は 20 代の男女各 9 人の合計 18 人である．会話グループは，男性 3 人によるグループを 3 組，女性 3 人によるグループを 3 組の計 6 組とし，全組で被験者全員が顔見知りとなるよう組み合わせた．実験に用いたシステムを図 1 に示す．被験者は，それぞれ防音室に入り，P2P 方式で，リアルタイムに各被験者の縦 46mm 横 61mm のバストアップ画像を相互通信し，VoIP によって音声通信を行った．

また，実験中の話題に，それぞれ異なる制約を課すことで，会話活性度の低・中・高 3 段階の制御を行った．話題以外には特に教示を与えず，自然に近い状態で会話を行わせた．実験は，各条件 15 分間の計 45 分間とし，順序効果が相殺されるように統制した．さらに，各条件での会話終了後，被験者に会話活性度を 0〜100 %で主観評価させた．評価は会話全体の活性度に加え，会話の前半・中盤・後半の各 5 分に対して，自身の活性度と場の活性度の 2 種類の活性度について主観評価させた．このとき，場はグループ全体の活性度，個人は被験者自身の活性度であり，両者を一致させる必要はないと教示した．

3.2 実験条件

会話活性度の 3 段階制御を目的に，実験で課した話題に関する条件を以下に示す．実験前の聞き取り調査により，験者が予め用意したニュースや社会問題に関する被験者の知識や興味を調査した上で，話題を設定した．

- **条件 L，想定する活性度“ 低 ”:** 被験者 3 人の知識と興味が最も低い深刻なニュースについて真剣に議論させ，他の話題についての会話は禁止した．
- **条件 M，想定する活性度“ 中 ”:** 被験者 3 人とも知識を有してはいるが，興味が低い社会問題について議論させた．他の話題についての会話は禁止した．
- **条件 H，想定する活性度“ 高 ”:** 話題の制限は行わなかった．また，話題選定の助けとなるように，各被験者の趣味をアンケートで事前調査し，結果を全員に教示した．さらに，できるだけ盛り上が

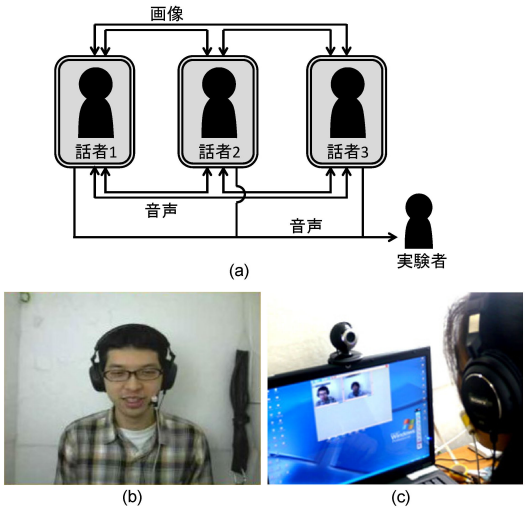

図 1 実験に使用したビデオチャットシステム
(a) 情報の流れ，(b,c) 利用風景

Fig. 1 Used video chat system (a) flow of information, (b,c) scene of conversation

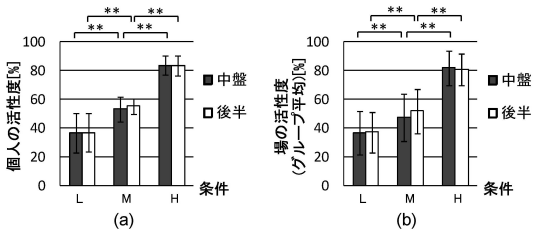

図 2　被験者による活性度の主観評価値
(a) 個人の活性度，(b) 場の活性度
Fig. 2　Evaluation of subjectively activation level. (a) personal activation level, (b) group activation level

るよう教示を与えた．

条件 L においては，知識も興味もない話題であるため，話題を広げられず，活性度は特に低くなると予想される．一方，条件 H では，話題制限がなく，話者の好きな内容について話して良いため，活性度は特に高くなると予想される．条件 M は，知識はあるため話を広げることは可能であるが，特に興味のない話題であるため，活性度は条件 L と H の間の中程度になると予想される．

3.3　実験結果概要

被験者自身が会話後に回答した，活性度の主観評価結果を図 2 に示す．なお，会話開始直後の数分間は，どの条件においても，何を話すかという話題探しの会話が多く発生し，被験者の活性度評価値にも差が見られなかったため，15 分間の会話データの内，前半の 5 分間を除外し，中盤以降の 10 分間を分析対象とした．

図 2 の結果を，実験条件を要因とした Kruskal-Wallis 検定を行ったところ，有意差が確認された（$p<0.01$）．そこで， Steel-Dwass 法による多重比較を行った結果，各条件間に有意差が認められ，会話条件の設定によって活性度が 3 段階に制御されていたことが確認される．なお，個人と場それぞれについて主観評価を行わせたが，個人と場で全く同一の回答をした被験者はいなかったことから，個人と場の活性度を区別して評価していたことがうかがえた．

4.　音声指標と活性度の関係の分析

本章では，先行研究に基づき，韻律を反映する正規化音圧・正規化ピッチ・音圧の変動係数・ピッチの変動係数・話速・発話長・同調傾向，そして会話構造を反映するオーバーラップ率・平均発話潜時・発話率の 10 の指標について，活性度の反映の有無や，個人差の大小などについて分析を行う．そうして，活性度自動推定に適した指標を検討する．

4.1　指標の算出方法

会話実験によって得られた音声データの例を，図 3 に示す．各指標を算出するため，まず，発話区間の切り出しを行った．音声の有無を判定する音圧の閾値は，ホワイトノイズの約 5 倍，平均音圧の約 1/10 に実験的に決定した．さらに有音区間と有音区間の間の無音区間が 0.3 秒以内であれば，発話中の小休止と見なし有音区間と見なした．最後に有音区間が 0.1 秒以上であれば発話区間とし，その他はノイズと見なした．各指標の算出方法を以下に示す．

- **正規化音圧：**30ms 毎に算出した音圧を，各話者の音圧の最大値で正規化し，0〜1 の範囲の値とした．
- **正規化ピッチ：**30ms 毎にピッチを算出し，各話者のピッチの最大値で正規化した．
- **音圧の変動係数：**抑揚が強くなると音圧の変動も大きくなると考え，発話区間中の音圧の標準偏差を平均値で除算した．
- **ピッチの変動係数：**抑揚が強くなるとピッチの変動幅も大きくなると考え，発話区間中のピッチの標準偏差を平均値で除算した．
- **話速：**話速算出には，1 秒あたりのモーラ数の計測が必要である．そこで音圧の谷と谷の間を音節と近似する手法 [19] を用い，音圧の極小値から極小値までを 1 音節と見なした．さらに，各音節の時間の平均を算出することで，近似的に話速を求めた．
- **発話長：**話速が速くなれば，同一文の発話時間は短くなる．そのため，一発話あたりの発話時間も，話速の変化を反映して変化する可能性が考えられる．そこで，発話区間の終了時刻から開始時刻を引いた値を発話長として算出した．
- **同調傾向：**先行研究と同様に，音圧・ピッチ・話速・発話長に関して，被験者 3 人の間に正の相関が確

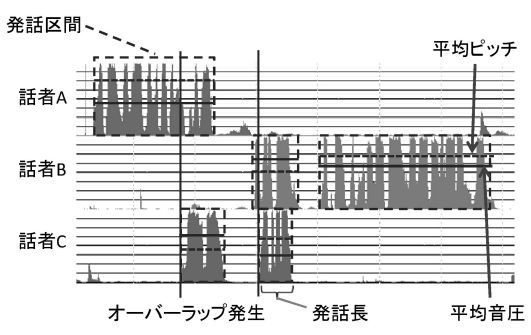

図 3　会話実験中の音声データと各指標検出例
Fig. 3　Example of actual phonetic information and calculated phonetic features.

認された場合を話者が協応していると見なし，同調傾向を示す値として，相関が確認された発話の割合を算出した[15]．非発話区間は，先行研究と同様に指標を線形補間し，50 発話ごとの話者間の指標値の相関係数の正値の割合を算出した．

- **オーバーラップ率：**会話が盛り上がると，笑いなどの発話の重畳が多く発生すると予想される．そこで，話者が発話を開始した時に，既に他の話者が発話していた場合に，オーバーラップ発生と見なした．また，発話数の多少によって，オーバーラップの発生数に差が生じるため，発話数に対するオーバーラップの発生割合を算出した．
- **発話潜時：**話者交替の発生を検出し，後発発話の開始時刻から先発発話の終了時刻を減算した値を求め，後発発話者の発話潜時とした．発話潜時は正値と負値の両方を取り，発話の重なりが生じた話者交替では負値をとる．会話の活性度が高くなると発話権授受のリズムが速くなり，潜時が小さくなると予想される．

  ただし，相槌や笑い声などの他話者の発話に対して完全に重なる発話では，発話権の授受は発生しない．そこで，発話権を持つ発話に遅れて開始した発話が後から終了した場合のみ，先行話者からの発話権の移譲，すなわち話者交代があったと見なして，潜時を算出した．
- **発話率：**個人の発話率は，当該被験者の発話が検出されている時を発話状態とし，発話状態の時間割合を求めた．

4.2 分析結果

３つの会話条件における，前半を除く中盤および後半 10 分間の，被験者およびグループごとの各指標の平均値を図 4,5 に示す．場の指標のうち，発話率は，話者全員が発話していない状態を沈黙状態として，沈黙状態ではない時間の割合とした．発話率以外は，話者 3 人の各指標値の平均とした．また，同調傾向については，話速と発話長については傾向が見られず，音圧とピッチで同様の傾向が見られたため，音圧の結果のみを記載する．グループ G1 は話者 a から c で構成され，同様に，話者 d から r がグループ G2 から G6 に相当する．なお，話者 a から i は男性，話者 j から話者 r は女性である．

分析の結果，先行研究と同様に，平均正規化音圧，平均正規化ピッチ，オーバーラップ率，発話率は，高活性度において増加し，平均発話潜時は高活性度において減少する傾向が確認された．一方で，音圧とピッチの変動係数，発話長及び話速は活性度との顕著な相関は見られず，同調傾向は先行研究[15]と同様に，女性グループのみで確認された．また，全指標について実験条件を要因とした Kruskal-Wallis 検定を行ったところ，平均正規化音圧，平均正規化ピッチ，オーバーラップ率，平均発話潜時，発話率の 5 指標で有意差が確認された（$p<0.05$）．

平均正規化音圧および平均正規化ピッチ，オーバーラップ率，発話率，平均発話潜時は，ほとんどの話者において活性度に応じた増減傾向が確認された．ただし，音圧とピッチが正規化後も大きな個人差やグループ差を示した一方で，オーバーラップ率と発話率は，グループ間の差が小さい傾向が見られる．平均発話潜時は，ほとんどの被験者で高活性度時に 0 秒程度まで減少しており，グループ差に加えて個人差も小さい傾向が認められた．

なお，G1 のみで高活性度条件での会話構造関連 3 指標の変化が小さい傾向が見られるが，他の 5 グループでは条件 LH 間の活性度差が約 50%だったのに対し，G1 は約 20%であった．すなわち，G1 は条件 H において他グループほど盛り上がっておらず，話題統制による活性度制御効果が小さかったことが一因と考えられる．

4.3 推定に用いる指標の選定

分析した指標のうち，平均正規化音圧・平均正規化ピッチ，オーバーラップ率，発話率，平均発話潜時はほとんどの話者において活性度に応じた増減傾向が見られ，活性度とよく相関していた．しかし，図 4 に見られるように音圧とピッチは正規化を行っても個人差が大きく，マイク感度など環境の影響も受けやすいことから，アバタ制御への応用には，キャリブレーション方法等の検討が必要である．一方，オーバーラップ率と平均発話潜時，発話率の会話構造に関連する 3 指標は，条件 H での指標値の個人差やグループ差が小さく，有意差も確認されたことから，他の話者でも同様の変化を示すことが予想される．さらに，これら 3 指標は発話の時間関係に基づくため，マイク感度などの環境差の影響を受けないことからも，キャリブレーション無しでの活性度推定につながる可能性が期待される．

よって，次章以降では，オーバーラップ率と平均発話潜時，発話率の 3 指標に基づく，場と個人の会話活性度推定の可能性を検討する．

5. 音声指標に基づく推定法の検討

5.1 活性度推定式の検討

5.1.1 会話音声指標と活性度の関係

オーバーラップ率と平均発話潜時，発話率による活性度推定可能性の検討のため，3 指標と主観的活性度の相関関係を分析した．話者およびグループごとの，3 指標の中盤・後半の平均値と，各個人あるいはグルー

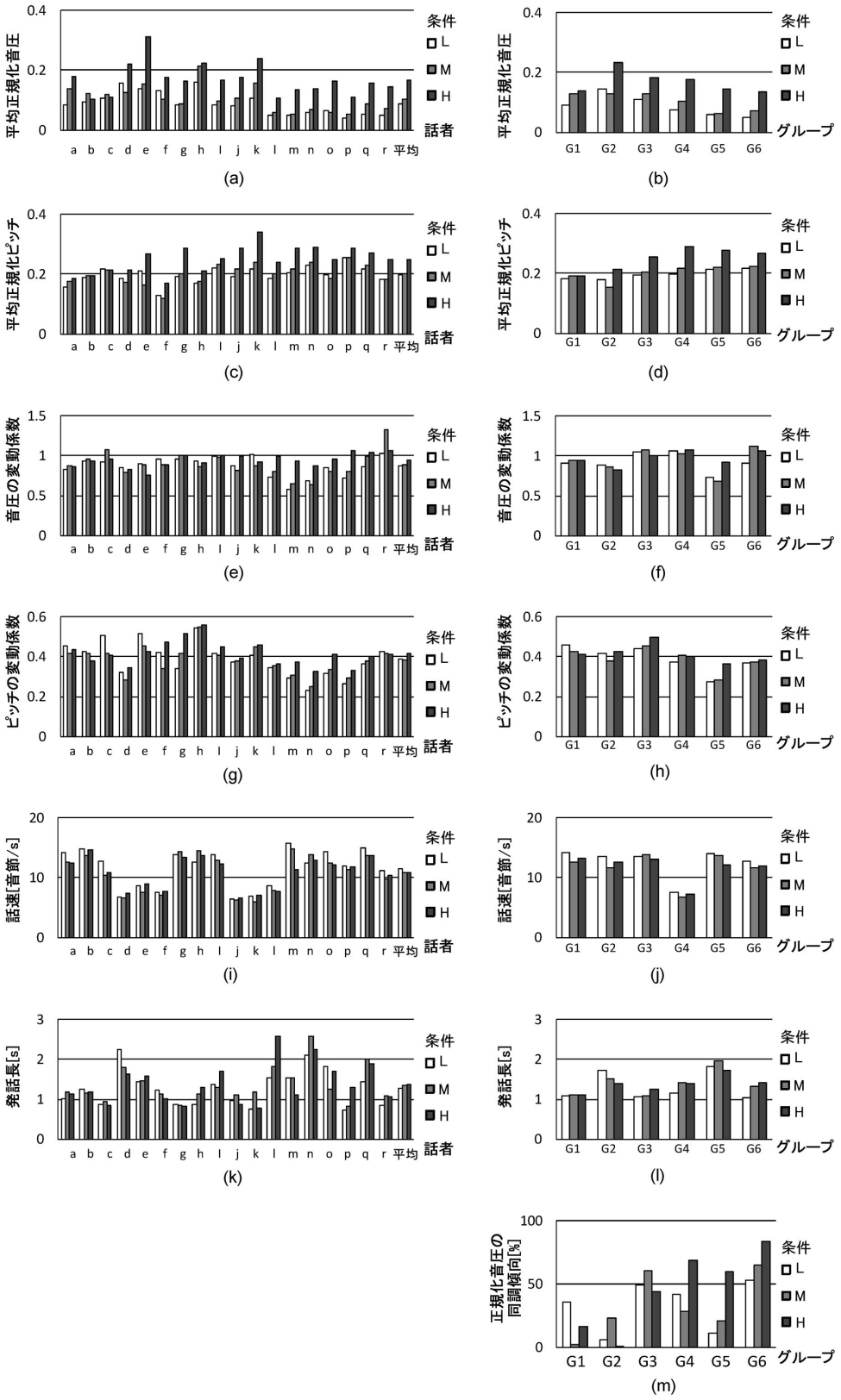

図 4　指標の分析結果 1 (a,b) 平均正規化音圧，(c,d) 平均正規化ピッチ，(e,f) 音圧の変動係数，(g,h) ピッチの変動係数，(i,j) 話速，(k,l) 発話長，(m) 正規化音圧の同調傾向

Fig. 4　Results of phonetic feature analysis. (a,b) average of normalized sound pressure, (c,d) average of normalized pitch, (e,f) variation coefficient of sound pressure, (g,h) variation coefficient of pitch, (i,j) utterance rate, (k,l) utterance length, (m) average of normalized sound pressure synchrony

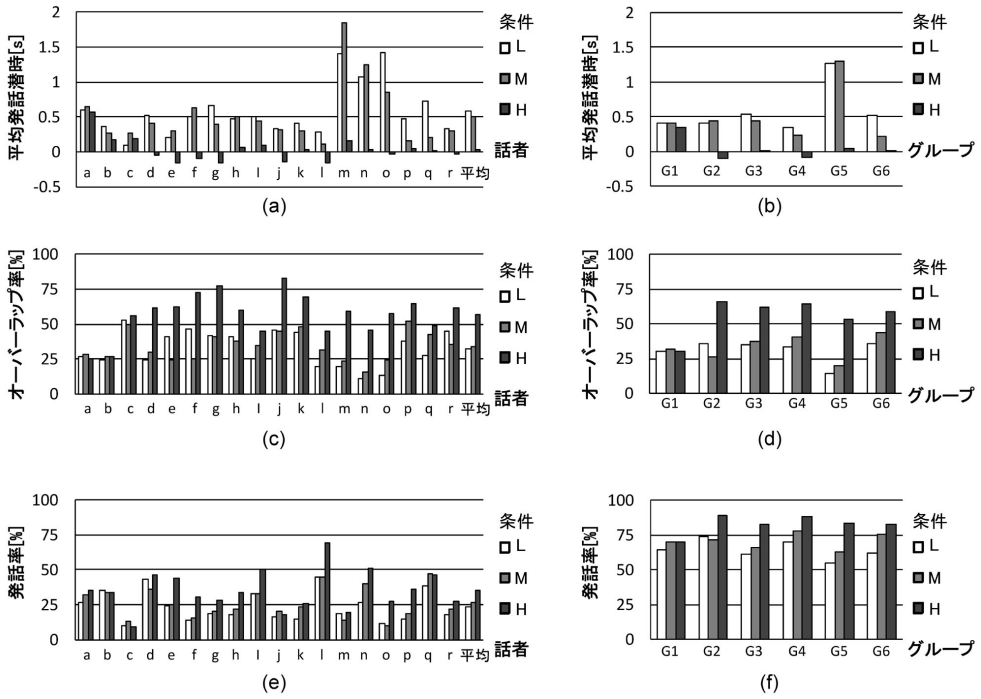

図 5　指標の分析結果 2 (a,b) 平均発話潜時，(c,d) オーバーラップ率，(e,f) 発話率
Fig. 5　Result of phonetic features 2. (a,b) average of response latency, (c,d) overlap rate, (e,f) speech rate

プの活性度の主観評価値の散布図を図 6 に示す．ここで，被験者による場の活性度の主観評価値は，評価者の尺度の差によってばらつくと考えられることから，場の活性度は 3 人の評価値の平均値を用いた．また，発話潜時のうち，話題が尽きた場合などに生じる例外的な長い発話潜時は，会話グループの状態を反映した結果ではあるものの，定常的な会話リズムとはやや性質が異なると考えられることから，個人でも場でも，実験的に閾値を 5 秒とした外れ値処理を行い，平均値を算出した．各活性度と各指標の相関係数を表 3 に示す．

個人の活性度は，散布図に見られるように，平均発話潜時と強い相関を示し，オーバーラップ率と発話率ではばらつきが大きくなるものの，中程度の相関が認められた．場の活性度は，3 指標すべてにおいて個人の場合よりもばらつきが小さく，強い相関が確認された．すなわち，精度の差はあるが，これらの指標に基づく個人や場の活性度推定の可能性が示唆された．

表 3　各活性度と各指標の相関係数
Table 3　Correlation coefficients between activation levels and values of phonetic features.

	オーバーラップ率	平均発話潜時	発話率
個人	0.48	-0.60	0.37
場	0.72	-0.79	0.82

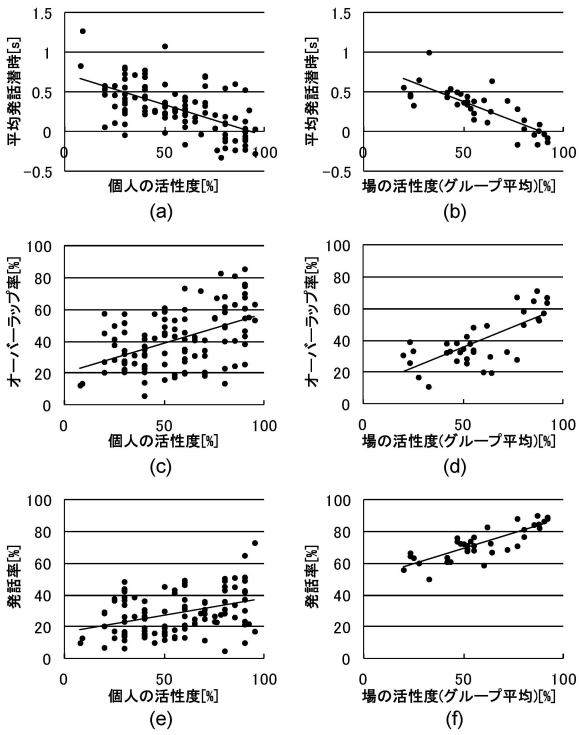

図 6　個人と場の活性度と指標 (a,b) 平均発話潜時，(c,d) オーバーラップ率，(e,f) 発話率
Fig. 6　Scatter plot, activation levels vs. phonetic features. (a,b) average of response latency, (c,d) overlap rate, (e,f) speech rate

5.1.2 活性度推定式の検討

3 指標すべてに会話活性度との相関が確認されたこと，また，より多くの指標を用いた方が多様な会話パターンに対してロバストな推定が可能と考え，3 指標と活性度の重回帰式を求め，活性度推定式とした．個人の活性度推定式を (1) 式，場の活性度推定式を (2) 式に示す．

$$\begin{aligned} \text{個人の活性度} &= 0.276 * \text{オーバーラップ率 [\%]} \\ &\quad -27.2 * \text{平均発話潜時 [s]} \\ &\quad +0.423 * \text{発話率 [\%]} + 41.1 \quad (1) \\ \text{場の活性度} &= 0.363 * \text{オーバーラップ率 [\%]} \\ &\quad -33.8 * \text{平均発話潜時 [s]} \\ &\quad +1.44 * \text{発話率 [\%]} - 22.0 \quad (2) \end{aligned}$$

5.2 推定式の妥当性の確認

上記の会話活性度推定式の妥当性を検討するため，分析データである 6 組の会話音声に対して，活性度推定を試行した．推定には，中盤・後半の 5 分間ずつの各指標値を用いた．なお，推定値が 0 未満となった場合は 0，100 を超えた場合は 100 とした．そうして，推定値と主観評価値から，RMSE（二乗平均平方根誤差）と，相関係数を算出した．結果を表 4 に示す．

分析データに対する推定結果は，個人の活性度の RMSE が 18%で，場の活性度が 12%であった．場の活性度に比べて個人の活性度の推定誤差が大きい結果となったが，いずれも高い相関を示すことが確認された．すなわち，個人差やグループ差を考慮せず，キャリブレーションを行わずに算出可能な指標からであっても，ある程度の精度で活性度が推定できる可能性が示唆された．

6. 推定実験

分析で得られた推定式による会話活性度推定の可能性を検証するために，新たに会話音声を収集し，推定実験を行った．

6.1 実験方法

分析データ取得時と同様の実験条件下で，新たに 20 代の男女各 9 人の合計 18 人，6 グループの会話音声を収集した．その後，収集した会話音声から，中盤 5 分と後半 5 分のオーバーラップ率，外れ値処理を行った平均発話潜時，および発話率を算出した．そうして，算出された各指標値を，(1,2) 式にそれぞれ代入し，活性度の推定値を算出した．

6.2 実験結果

推定された活性度と，主観評価による活性度評価値の散布図を図 7 に，RMSE および，主観評価値と推定値の相関係数を表 5 に示す．

散布図に見られるように，個人の活性度も場の活性度も，主観評価値との正の相関傾向を示し，前章で求めた推定式による活性度推定の可能性が確認される．場の活性度は，個人の活性度と比較してばらつきが小さく，回帰直線も，ほぼ主観評価値と一致する結果となった．個人の活性度推定値は，被験者自身による主観評価値よりやや高くなる傾向を示したが，傾きは 1 に近く，平均 RMSE は 24%であった．また，分析実験で重回帰式を求めた際に，標準偏回帰係数が最大となった，発話潜時（個人）と発話率（場）に関して，分析実験データから各指標の 1 時近似式を求め，推定実験データに対する単一指標による活性度推定も行った．単一指標の結果と比較すると，平均 RMSE は同程度であったが，相関係数が重回帰式を用いた推定では個人が 0.02，場が 0.01 向上していた．以上のように，会話構造に関連する指標を用いることで，個人差やグループ差を考慮せずキャリブレーションを行わずに，RMSE20%程度で，おおよその活性度が推定できる可能性が示された．

7. 考察

7.1 推定精度

今回の実験では，分析データと推定データの両方で，場の活性度が個人の活性度よりも推定精度が高い結果となった．場の活性度は個人ごとの主観尺度のばらつ

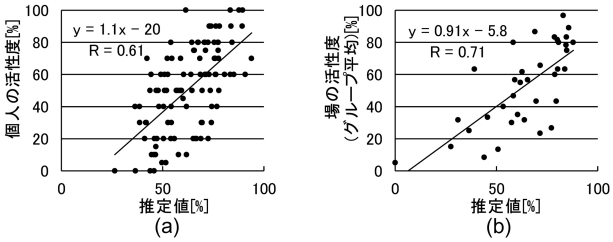

図 7 活性度推定値と主観評価値 (a) 個人の活性度，(b) 場の活性度

Fig. 7 Scatter plot, subjective activation levels vs. estimated features. (a) group, (b)personal

表 4 分析データの推定誤差と相関係数

Table 4 RMSE and correlation coefficients for analyzed data set.

	RMSE	相関係数
個人	18%	0.64
場	12%	0.83

表 5 分析対象と異なる群の推定誤差と相関係数

Table 5 RMSE and correlation coefficients for another data set.

	RMSE	相関係数
個人	24%	0.61
場	21%	0.71

きが平均されることに加えて，個人の活性度推定では，発言が少ない話者は母数が小さくなってオーバーラップ率や平均発話潜時が大きく変動しやすいことや，集団は，沈黙が続かないように発話するなどの個人のばらつきを補完する挙動を示すことなどが理由と考えられる．そのため，異なる話者の組み合わせでも，今回と同様の傾向を示すと予想される．すなわち，場の活性度の方が，個人の活性度と比較してロバストな推定結果が得られるものと考えられる．

重回帰式による推定は， RMSE 誤差は単一指標による推定と同程度であったが，わずかに相関が向上していた．また，平均発話潜時のみを用いて個人の活性度推定を行った場合，発話数の少ない話者では母数が小さくなるため，個々の発話のばらつきの影響が原理的に大きくなる．場の活性度に関しても，発話率は話者交代リズムの変化や発話の重なりを反映しないため，発話量の個人差の影響を受けやすくなると考えられる．したがって，個人差や環境の変化に対する推定のロバスト性の観点からは，複数指標を併用する方が大きな誤差を発生する可能性が低減され，望ましいと考えられる．

推定実験では，分析対象と異なる 6 グループに対して，平均 RMSE24%で個人の活性度の推定が可能であったが，被験者自身の主観評価よりも高めに推定される傾向が見られた．推定誤差の発生要因の一つとして，活性度評価が主観評価であったため，被験者ごとの尺度の差が影響した可能性が考えられる．音声情報に基づく活性度推定が，外部から観察可能な指標に基づく推定であることから，今後は，被験者自身ではなく，会話相手や第三者が評価した個人の活性度と，推定値の関係を調べることも必要と考えられる．

また，推定精度の向上のためには，個人差の影響を軽減するために，事前にキャリブレーションを行う，あるいは機械学習によってパラメータを個人適合するなどの方法が考えられる．しかし，話題やその場の雰囲気のように，動的に変化する個人差以外の要因も影響する可能性があることから，個人適合の有効な指標とそうでない指標の分別が必要になると予想される．また，個人差が大きく環境の影響を受けやすいため，本研究では採用しなかった音圧やピッチなどの韻律情報を，相対的変化を求めるなど，個人差の影響を軽減するなんらかの処理をして併用することも考えられる．さらに，場の活性度推定がロバストであることを利用して，個人の活性度推定値を改善する方法も考えられる．例えば，活性度が低く推定されるユーザに対しては，場の活性度を基準に，個人の活性度の相対的変化のみを反映するなどの方法が考えられる．

7.2 アバタ制御への活性度の反映

本研究では，個人の活性度を被験者自身の主観評価によって決定した．しかし，自身の内観と他者の客観的評価が，必ずしも一致するとは限らない．特に，活性度が上がっても見た目の動作や会話音声に変化が表れにくい話者の場合，推定式によって低めに推定された活性度は，むしろ，他者が韻律や会話構造から判断した活性度と一致する可能性が高いと考えられる．そのため，アバタ制御における活性度推定式の妥当性に関しては，推定値を実際の動作に反映して挙動の自然さなどを評価する必要がある．

今回の実験では，5 分間の会話音声からの会話活性度の推定可能性を検討した．しかし，アバタの動作制御への適用を考えた場合，過去数分間の音声から活性度を推定すると，アバタ活性度も同じ時定数を持って変化する．そのため提案手法をそのまま適用した場合には，穏やかな会話ムードの変化は適切に反映すると期待される一方で，冗談などによる突発的な会話活性度の変化が生じた時には，アバタは適切な挙動を行うことができない．しかし，単純に推定時間を短縮すると推定精度の低下が予想されることに加え，冗談などによる一時的な活性は，必ずしもそのまま持続するとは限らない．そこで，アバタ笑顔の自動制御 [5] と同様に，ゆっくりと連続的に変化する成分と，瞬時的な変化を反映する成分の 2 成分で活性度をモデル化し，前者を発話潜時やオーバーラップ率，発話率を用いて推定し，瞬時的な成分は，笑顔の自動制御で有効だった発話初期音圧を利用するなどの方法が考えられ，今後の検討が必要である．

8. まとめ

本研究では，会話活性度によるアバタの動作自動制御に向けて，3 人会話音声を収集し，活性度と音声指標の関係を分析した．その結果，オーバーラップ率と平均発話潜時および発話率に，活性度との相関に加えて，ピッチ等と比較して個人差が小さい傾向が認められた．そこで，3 指標を用いた会話活性度推定式を重回帰分析によって求め，異なるユーザに対して推定実験を行った結果，キャリブレーションを行わなくとも，自然会話中の個人や場の活性度を定量的に推定できる可能性が示された．今後の課題は，アバタ動作の活性度制御に向けて，推定式を補完するような制御と適切な推定時間の検討を行い，アバタの活性度制御を実装したいと考えている．

謝辞

本研究の一部は，文部科学省特別教育研究費ならびに科学研究費補助金によるものである．ここに記して

感謝する．

参考文献

[1] Kato, Y., Kato, S., Akahori, K.: Effects of emotional cues transmitted in e-mail communication on the emotions experienced by senders and receivers; Computers in Human Behavior, Vol. 23, No. 4, pp. 1894-1905 (2007).

[2] マジョリー・F・ヴァーカス著, 石丸 正訳: 非言語 (ノンバーバル) コミュニケーション; 新潮社 (1987).

[3] A. マレービアン著, 西田 司, 津田幸男, 岡村輝人, 山口常夫訳: 非言語コミュニケーション; 聖文社 (1986).

[4] 渡辺富夫, 大久保雅史, 中茂睦裕, 檀原龍正: InterActor を用いた発話音声に基づく身体的インタラクションシステム; ヒューマンインタフェース学会論文誌, Vol. 2, No. 2, pp.21-29 (2000).

[5] 宮島俊光, 藤田欣也: 音声チャットシステムにおける基本周波数と音圧を利用したアバタ表情制御法; ヒューマンインタフェース学会論文誌, Vol. 9, No.4, pp.85-94 (2007).

[6] Hartmann, B., Mancini, M., Pelachaud, C.: Implementing Expressive Gesture Synthesis for Embodied Conversational Agents; Gesture Workshop 2005: pp.188-199 (2005).

[7] 佐伯悠子, 田中貴紘, 藤田欣也: 会話活性度に基づくアバタ動作の統一的制御方法の検討; ヒューマンインタフェース学会研究報告集, Vol. 11, No.2, pp.71-74 (2009).

[8] 前田貴司, 高嶋和毅, 梶村康祐, 山口徳郎, 北村喜文, 岸野文郎, 前田菜穂, 大坊郁夫, 林 良彦: 3 人会話における非言語情報と「場の活性度」に関する検討; 電子情報通信学会技術研究報告, Vol. 109, No.27, pp.73-78 (2010).

[9] Bezooyen. R, V.: Characteristics and Recognizability of Vocal Expressions of Emotion; Foris Pubns USA (1984).

[10] Murray, I, R., Arnott, J, L.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion; J. Acoust. Soc. Am, Vol. 93, No.2, pp.1097-1108 (1993).

[11] Coleman, R, F., Williams, R.: Identification of emotional states using perceptual and acoustic analyses; Transcripts of the 8th symposium on care of the professional voice, The Voice Foundation Vol. 1, pp. 77-83 (1979).

[12] 森山 剛, 斎藤英雄, 小沢慎治: 音声における感情表現語と感情表現パラメータの対応付け; 電子情報通信学会論文誌, Vol. J82-DII, No.4, pp703-711 (1999).

[13] Barrett, L, F., Russell, J, A.: Independence and bipolarity in the structure of current affect; Journal of Personality and Social Psychology, VOL. 74, pp.967-984 (1998).

[14] Nwe, T, L., Foo, S, W.: Speech emotion recognition using hidden Markov models; Speech Communication, VOL. 41, pp.603-623 (2003).

[15] 西村良太, 北岡教英, 中川聖一: 音声対話における韻律変化をもたらす要因分析; 音声研究, Vol. 13, No.3, pp.66-84 (2009).

[16] 河原達也, 川島宏彰, 平山高嗣, 松山隆司: 対話を通じてユーザの意図・興味を探り情報検索・提示する情報コンシェルジェ; 情報処理学会誌, Vol. 49, No.8, pp.912-918 (2008).

[17] 豊田薫, 宮越喜浩, 山西良典, 加藤昇平: 発話時間長に着目した Tree-Augmented Naive Bayes による対話雰囲気推定; 人工知能学会全国大会論文集, Vol. 25, pp.1I2-5 (2011).

[18] Sejima, Y., Ishii, Y., Watanabe, T.: A Virtual Audience System for Enhancing Embodied Interaction Based on Conversational Activity; Lecture Notes in Computer Science, Vol. 6772, pp. 180-189 (2011).

[19] Maeran, O., Piuri, V., Gajani, G, S..: Speech recognition through phoneme segmentation and neural classification; Instrumentation and Measurement Technology Conference, 1997. IMTC/97. Proceedings. 'Sensing, Processing, Networking'., IEEE, Vol. 2, pp.1215-1220 (1997).

（2011 年 12 月 21 日受付，2012 年 4 月 17 日再受付）

著者紹介

守屋　悠里英　（学生会員）

2011 年東京農工大学工学部情報工学科卒業．現在，同大学院修士前期課程在学中．会話音声を利用したアバタ自動制御の研究に従事．人の特性に合わせたコミュニケーション支援に興味を持つ．

田中　貴紘　（正会員）

2006 年東京工業大学大学院総合理工学研究科博士後期課程修了．博士（工学）．2007 年東京農工大学大学院共生科学技術研究院助教，現在に至る．知的エージェント，ヒューマンエージェントインタラクションなど，人と人，人とエージェント間のコミュニケーション支援に興味を持つ．

宮島　俊光

1997 年早稲田大学教育学部理学科数学専修卒業．現在，東京農工大学大学院工学部情報工学専攻技術専門職員として，音声情報を利用したコミュニケーション支援の研究に従事．

藤田　欣也　（正会員）

1988 年慶應義塾大学大学院理工学研究科修了．相模工業大学，東北大学医学部，岩手大学を経て，現在東京農工大学大学院教授．共有仮想空間コミュニケーション，ＶＲシステムや感覚の遠隔共有など，人と共生する情報システムのためのヒューマンインタフェースの研究に従事．工学博士．

(C) NPO法人ヒューマンインタフェース学会