Vol.2011-ICS-162 No.5
2011/3/8

音声対話型観光案内システムにおける誤応答リカバリー効果の評価

香山 健太郎[†1] 小林 亮博[†1] 水上 悦雄[†1]
翠 輝久[†1] 柏岡 秀紀[†1] 河井 恒[†1]

本稿では，対話システムの誤応答に対するリカバリー発話について，そのユーザに与える影響についての予備実験を行った．その結果，誤応答の場合に謝罪するとシステムに対する評価が上がること，謝罪タイミングは通常発話終了後一呼吸置いてからあるいは気付き次第通常発話を中断して行うことが良いことがわかった．また，ユーザの反応を映した画像情報のみからシステムの誤応答を推定することは難しいことがわかった．

Evaluation of the Effect of the Recovery from Inadequate Response on Tourist Guide Spoken Dialog System

Kentaro Kayama,[†1] Akihiro Kobayashi,[†1]
Etsuo Mizukami ,[†1] Teruhisa Misu,[†1]
Hideki Kashioka[†1] and Hisashi Kawai

We performed pilot experiments on the effect of the recovery from inadequate response on spoken dialog system. They show that apology for inadequate response from the system increases subjective evaluation of the system from users, and that adequate timing of apology is three seconds after completing response or to do as soon as possible by terminating response. It is also shown to be difficult to estimate inadequate response of the system from image information of the responses of user only.

1. はじめに

音声対話システムは，これまで，飛行機予約・バス案内・サポートセンターなど，電話対話を中心に実用化がなされ，特に，欧米において普及してきた．

一方，日本では，一部の予約システム等に音声対話システムは利用されていても，普及はさほど進んでいない．しかし，技術の発展によって，今後は日本でも予約・案内業務に音声対話システムが普及していくことが期待される．

この音声対話システムの将来の発展領域として，我々は観光案内業務に着目し，それを題材とした対話システムを構築している．本システムはまた，自然な感じでのコミュニケーションの実現のため，非言語情報をも利用した対話システムを目指している．現在我々が作成している大型ディスプレイを用いた音声対話型京都観光案内システム”HANNA” を**図 1** に示す．

しかし，このような音声対話の普及に対する問題点の一つとして，音声認識の誤りなどによって，システムがユーザの望まない情報を提示したまま，特に訂正・謝罪等もすることなく対話が進行してしまい，ユーザが不満を抱いてしまうことがあげられる．

この問題に対し，我々はユーザの顔向き・動きなどを画像情報から取得し，それを利用してシステムの誤応答を判別して，適切なタイミングでユーザに謝罪したり言い直しを示唆したりするようなシステムを構築しようとしている．この誤応答リカバリーが実現すれば，ユーザの対話満足度が上がることが期待できる．

そこで，まず誤応答リカバリーを行うことの有効性を検証するために予備実験を行った．本稿ではそれについて述べる．以下，2 節にて関連研究について述べた後，3 節で誤応答リカバリーについてどのようなことを解明すべきかを述べる．そして，4 節で現在我々が構築しているシステムについて詳細を述べた後，それを用いた誤応答リカバリーに関する評価実験の内容とその結果について 5 節で述べる．

2. 関連研究

人間と機械システムとの対話において，システムの誤応答を検出するためには，そのときのユーザの状態を観察することが有効であると考えられる．

†1 情報通信研究機構
National Institute of Communications and Technology (NICT)

Vol.2011-ICS-162 No.5
2011/3/8

図 1 大型ディスプレイを用いた音声対話型京都観光案内システム”HANNA”
Fig. 1 ”HANNA”, Kyoto tourist guide system by spoken dialog with large display

藤江らは，音韻情報からロボットと対話中のユーザの機嫌を判定している[1)]．また，中島らはユーザの頷き・首振りを画像から判定[2)]している．Kaliouby らも，表情から愉快な映像を見ているか不快な映像を見ているか判定するような研究を行っている[3)]．

しかし，これらの研究においてはユーザは椅子に座るなど頭部の位置を限定したものであり，用いている情報も表情および音韻のみであって，立っているユーザの総合的な動きから応答の正誤を判別しようとするものではない．これに対し，本システムでは，より自然な対話シーンとして，据え置きの大画面ディスプレイを持つシステムに対し，ユーザが近づいてきて立ったまま対話を行うという状態を設定して実験を行っている．

3. 誤応答リカバリー問題

音声対話システムにおいて，システムの誤応答を自動的に検出し，それに基づいてリカバリー発話を行うことは有効であると期待されるが，その明確な効果は明らかではない．また，システムの誤応答の検出については，人間が通常の対話で行っているのと同様に，ユーザの反応を手掛かりにすることが有効と考えられるが，どのような特徴量・手法を用いればそれが精度良く可能になるかについても明らかになっていない．

これらが明らかになれば，誤応答を検出する部分のソフトウェアを構築することが容易になる上，誤応答検出結果を対話制御戦略においてどのように用いれば良いかが明らかになり，音声対話システムの対話能力向上に貢献できると考えられる．

そこで，本稿ではこれらの誤応答リカバリーに関する性質を明らかにするための予備実験を行う．本稿で検証する問題は次の 4 つである．

(1) システムが誤応答リカバリー発話を行うことによって，ユーザのシステムに対する主観的評価はどのように変動するか
(2) 誤応答の検出失敗によって発生する 2 種類の齟齬，すなわち，正応答の場合に誤応答リカバリー発話を行うこと，および，誤応答なのにリカバリー発話を行わないことについて，それぞれがユーザのシステムに対する主観的評価にどの程度影響するか
(3) 誤応答リカバリー発話を行う適切なタイミングはどのようなものか
(4) 誤応答の検出を画像情報のみから行うことはどの程度可能か

この問題を検証するにあたって，音声対話システムとしては，我々が開発している大型ディスプレイを用いた音声対話型京都観光案内システム”HANNA” を用いた．本システムは，国際学術会議や一般向け展示会でデモを行う程度の完成度を持っており，システムそのものに対する不満がユーザの主観的評価に与える影響を減らすことができると期待できる．このシステムの概要について次節で述べる．

また，上記の問題に対応する実験として，(1) および (2) については 5.2 節の実験 1 で，(3) については 5.3 節の実験 2 で，(4) については 5.4 節の実験 3 で述べる．

4. 大型ディスプレイを用いた音声対話型京都観光案内システム

4.1 プロアクティブな対話システム

我々が開発している大型ディスプレイを用いた音声対話型京都観光案内システム”HANNA” は，様々な音声対話技術のみならず，対話中のユーザの非言語情報を検出する画像認識技術をも統合して，状況に応じて push 型にも pull 型にもなれるような，より自然な音声対話システムの実現を目指して研究開発を行っているものである[4)]．

これは，我々が提案している，新しいインタラクティブ情報ディスプレイシステムの一環でもある．プロアクティブな対話システムとは，システム側からも積極的に気の利いた情報を気の利いたタイミングで提示するものである[5)]．

Vol.2011-ICS-162 No.5
2011/3/8

そして，HANNA では，

- 画像認識によるユーザ検出技術
- 顔向きや視線の検出技術・大きな動作の検出技術などの非言語情報検出技術
- 音声認識・合成技術
- 画像・音声の認識結果を統合する柔軟な対話制御技術

を集約した統合システムを構築している．

使われ方としては，観光案内所等，屋内の公的スペースにおいて不特定多数の利用者に情報を提示するために据え置きで設置されることを想定している．そして，観光案内に限らず，役所や商業施設の入り口での簡単な案内業務を肩代わりするというようなアプリケーションを見据えている．

本節ではそのシステムの概要について述べる．

4.2 ハードウェア構成

本システムは，50 インチプラズマディスプレイ・姿勢制御可能な単眼カメラ 3 台・ステレオカメラ 1 台・USB カメラ 1 台・マイク＋超音波センサ・スピーカー・処理用 PC9 台からなる（図 1）．PC を除くシステム全体のサイズは幅 135cm，奥行き 100cm，高さ 205cm であり，重量は約 150kg である．

本稿における実験は，対話の状態遷移を手動で行う Wizard of Oz 方式にて行ったため，カメラ・マイク等の入力機器は基本的には状況の記録に用いているのみである．

このうち，ステレオカメラの片方の目からの入力画像を 5.4 節の実験 3 で用いている．ステレオカメラとしては PointGrey 社の Bumblebee2 を用いている．これは水平方向に約 60 度の画角を持っている．

4.3 ソフトウェア構成

HANNA は合計 20 種類強のモジュールからなっている．モジュールは，その機能から画像処理部・音声認識および構文解析部・対話制御部・音声合成部・画面制御部の 5 つに大別できる．

本稿で述べる実験では，前項で述べたように画像および音声の認識部分は Wizard of Oz 方式で行うため，判断は人間が行う．

また，対話制御においては，様々な質問が受け入れられるようになっており，実験においても被験者にそのように教示は行うが，実際に発話してもらう文，およびそれに対するシステムの状態遷移・発話内容は固定である．

音声合成部では，声優の音声データに基づいた HMM 音声合成を行うことにより，滑らかで自然な感じの音声合成を実現している．

(i) 初期画面　(ii) 概要説明　(iii) 正応答　(iv) 誤応答

図 2 実験 1 における画面表示
Fig. 2 Screenshot in experiment 1

画面制御部では，**図 2** に示すような表示を行う．画面には基本的にキャラクターエージェントが表示されている．このキャラクターは，ユーザの仮想的な対話相手として様々な動作を行う．ユーザがマイクに顔を近づけたときは耳を傾ける，音声入力から画面遷移までの間は考え中のような動作をする，誤応答リカバリー発話の際には申し訳なさそうな顔をして頭を下げるなど，ジェスチャや表情でシステムの状態をわかりやすく表出する．これは，キャラクターエージェントがユーザに同調的な動作をすることによって，システムからの提案への受容度があがるという角らの研究成果[6] にも基づいている．

5. 実験と結果

5.1 誤応答リカバリー評価実験

前節で述べたシステムを用いて，誤応答リカバリーの評価に関する実験を行った．

まず，20 代・30 代の男女各 24 名ずつ，計 48 名の被験者を集めて次の実験を行った．

実験 1 誤応答リカバリーのユーザ評価に与える効果の測定
実験 2 誤応答リカバリーのタイミングに関する評価の測定

そのときの様子を**図 4** に示す．

さらに，実験 1 のときの様子を 3 名の女性に見せることで次の実験を行った．

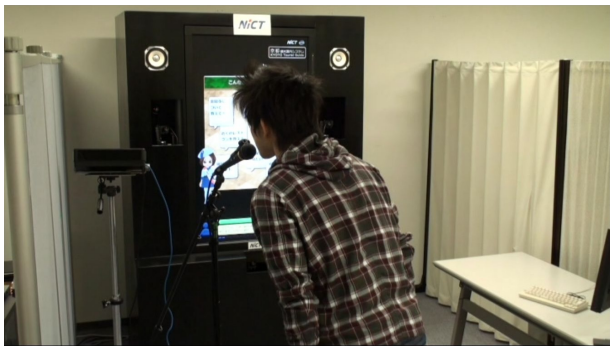

図 3 誤応答リカバリー評価実験の様子
Fig. 3 Experiments for the evaluation of the effect of the recovery from inadequate response

実験 3 誤応答の画像からの判定に関する予備実験

本節では，これらの実験の概要と結果について述べる．

5.2 実験 1:誤応答リカバリーのユーザ評価に与える効果

まず，システムによる誤応答リカバリーが対話の自然さおよびユーザのシステムに対する好感度に与える影響についての実験を行った．

本実験では，ユーザに，システムに対して次のような対話を 4 回行うよう指示した．

(1) 「金閣寺について教えて」と質問する
(2) システムからの応答に続き「近くのレストランを教えて」と質問する
(3) システムからの応答が間違っていた場合は，再度「近くのレストランを教えて」と質問する

これに対し，システムは (1) の段階では必ず金閣寺についての概要を表示・説明する（正応答，図 2(ii)）．(2) 以降については，正しく近くのレストランを表示する場合（正応答，図 2(iii)）と金閣寺までの行き方を表示する場合（誤応答，図 2(iv)）とを用意し，また，表示から数秒後にシステムが「何か間違えましたでしょうか？申し訳ありませんがもう一度お願いします.」と発話する場合（謝罪あり）とそのまま何もしない場合（謝罪なし）とを用意した．すなわち，次の 4 種類の応答をするように設定した．

(*) 正応答・謝罪なし（ユーザがシステムに慣れるためのテストとして，必ず 1 回目に行う）
(a) 正応答・謝罪あり
(b) 誤応答・謝罪なし
(c) 誤応答・謝罪あり

(3) の後は必ず正応答をするようにした．これらの応答は，ユーザの発話に応じてオペレータがシステムの状態を遷移させるという方法で行った．

被験者は 3 グループに分け，(a)〜(c) の順番を変えて 1 人ごとに各 1 回ずつ，(*) も含めて 4 回の対話をさせた．さらに，(a)〜(c) の直後にそれぞれ対話の自然さおよびシステムに対する好感度を 5 段階で評価させた．その結果を**表 1** および図 4 に示す．

なお，(a)〜(c) の実施順序パターン 3 群の間で，条件間の交互作用は見られなかったため，混合して条件間比較を実施している．また，教示どおりに対話を行わない被験者が 1 名いたため，その人を除いて分析を行った．したがって，本実験の総被験者数 $N = 47$ である．

自然さに関しては，条件間に有意傾向差があり,「誤応答・謝罪あり」の場合が,「正応答・謝罪あり」および「誤応答・謝罪なし」より良い結果となっている ($F(92, 2) = 3.029, p = 0.0523$)．また,「正応答・謝罪あり」と「誤応答・謝罪なし」では有意差はなかった．

表 1 誤応答リカバリーに対する主観的評価
Table 1 Subjective evaluation of the recovery from inadequate response

	自然さ		好感度	
	平均値	標準偏差	平均値	標準偏差
(a) 正応答・謝罪あり	3.681	1.187	3.702	0.987
(b) 誤応答・謝罪なし	3.702	1.128	3.617	0.980
(c) 誤応答・謝罪あり	4.021	0.887	4.064	0.998

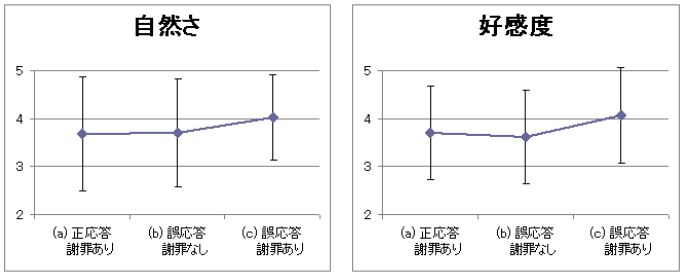

図 4 誤応答リカバリーに対する主観的評価
Fig. 4 Subjective evaluation of the recovery from inadequate response

Vol.2011-ICS-162 No.5
2011/3/8

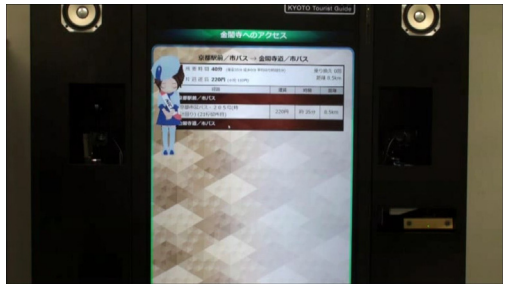

図 5 誤応答リカバリータイミング評価のための動画例
Fig. 5 An example of video files for evaluating adequate timing of the recovery from inadequate response

図 6 ステレオカメラからの入力画像例
Fig. 6 Example of the inputted image from stereo vision camera

システムに対する好感度については，条件間に有意差があり，自然さと同様に,「誤応答・謝罪あり」の場合が,「正応答・謝罪あり」および「誤応答・謝罪なし」より良い結果となっている ($F(92,2) = 7.012, p < 0.01$)．また,「正応答・謝罪あり」と「誤応答・謝罪なし」では有意差はなかった．

5.3 実験 2:誤応答リカバリーのタイミング

次に，誤応答リカバリーを行うタイミングに関する実験を行った．

この実験では，ユーザがシステムと対話している動画を，誤応答リカバリーを行うタイミング別に５種類用意し，それを被験者に見せて，それぞれについての自然さ・好感度，およびこの中でどれがもっとも最適なタイミングかを判定させた．なお，動画は 5 種類とも自由に何度も見返すことができるようにした．

動画内でのユーザとシステムの対話は

(1) ユーザ「近くのレストランを教えて」
(2) システム「京都駅から金閣寺への行き方を表示します」
(3) システム「何か間違えましたでしょうか？申し訳ありませんがもう一度お願いします」

というもので，(2) の通常発話と (3) のリカバリー発話との間を-2,0,+1,+3,+5[秒] の 5 段階に変えたものを用意した．(2) と (3) の間が負になっているものは，(2) の発話終了前に中断して (3) の発話を始めたことを示す．また，ユーザの反応による影響を排除するため，被験者に見せる動画ではシステムの画面のみが映され，ユーザについては声のみが聞こえるようになっている（**図 5**）．

その結果を**表 2** および**図 7** に示す．自然さ，好感度ともシステム通常発話終了後 3 秒経ってからリカバリー発話を始めるものがピークとなっている．しかし，最適タイミングについては，ピークが 2 つ存在することから，一呼吸おいてからの謝罪を自然とするグループと，間違いに気づき次第発話を中断して謝罪することを自然とするグループとが存在し，またピークの大きさから前者の方がやや多いと考えられる．

表 2 誤応答リカバリーのタイミングの主観的評価
Table 2 Subjective evaluation of the timing of the recovery from inadequate response

	自然さ		好感度		最適
タイミング	平均値	標準偏差	平均値	標準偏差	[人数]
-2[秒]	2.58	1.43	3.06	1.16	7
0	2.88	1.35	3.63	1.10	1
+1	3.44	1.27	3.94	0.93	13
+3	3.77	1.13	4.06	0.93	22
+5	2.83	1.36	3.33	1.23	5

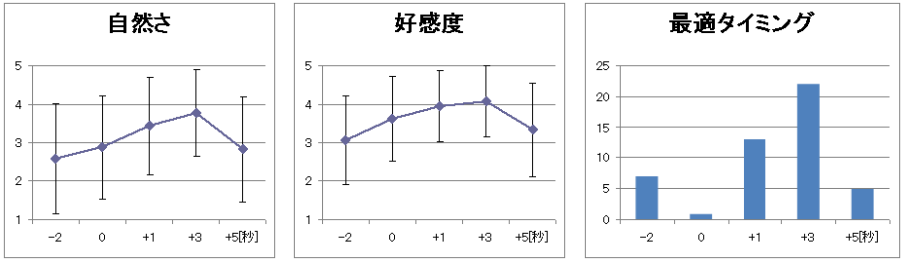

図 7 誤応答リカバリーのタイミングの主観的評価
Fig. 7 Subjective evaluation for the timing of the recovery from inadequate response

5.4 実験 3:誤応答の画像からの判定に関する予備実験

発話直後からシステムの応答を視聴している間の約 10 秒間について，システムのステレオカメラから撮影したユーザの画像（**図 6**）を動画として 3 人の評価者に見せ，ユーザに対してフォローを入れたいかどうかを 4 段階で評価させた．画像は解像度が 320 × 240[pixel] で 3〜4fps で撮影されており，それが実際と同じタイムラインで再生されるよう動画を作成した．

5.2 節で述べた 4 種類の応答分類に加え，最初の質問に対する金閣寺の概要表示・説明を (0) としたときの結果を**表 3** に示す．この結果から，フォローを入れたいかどうかの評価はかなり個人差があることがわかる．しかし，誤応答のときに評価値が大きくなっている傾向

Vol.2011-ICS-162 No.5
2011/3/8

表 3 フォローを入れたいかどうかの主観的評価
Table 3 Subjective evaluation whether the user should be cared

応答分類	評価者 A		評価者 B		評価者 C		総合	
	平均値	標準偏差	平均値	標準偏差	平均値	標準偏差	平均値	標準偏差
(0) 概要表示	0.676	0.897	1.199	0.454	0.272	0.599	0.716	0.775
(*) 正応答・謝罪なし	0.307	0.687	1.133	0.523	0.421	0.777	0.620	0.764
(a) 正応答・謝罪あり	0.188	0.583	1.031	0.337	0.198	0.533	0.472	0.634
(b) 誤応答・謝罪なし	0.354	0.763	0.990	0.810	1.292	1.145	0.878	1.001
(c) 誤応答・謝罪あり	0.240	0.625	1.083	0.607	1.115	1.189	0.813	0.943

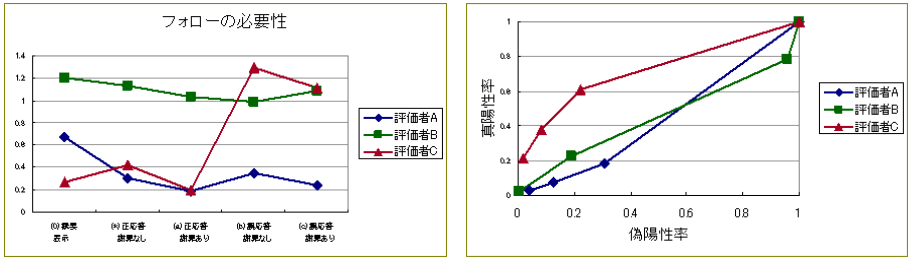

図 8 フォローを入れたいかどうかの評価の誤応答判別能力
Fig. 8 Discriminant ability for inadequate response of the system by using subjective evaluation whether the user should be cared

は存在するようである．

また，この評価を閾値として正応答・誤応答を判別した際の ROC 曲線を**図 8** に示す．この評価が正応答・誤応答の分類にある程度合致するのは評価者 C のみであり，それでも精度はそれほど高いとは言えない．

今回の設問は，直接誤応答と思うかどうかを評価させたものではなく，フォローを入れたいかどうかを評価させたものであるが，このように，画像のみからの判定は人間でも難しいことがわかる．

6. おわりに

本稿では，音声対話システムにおいて，システムの誤応答をユーザの反応から判別すること，およびそれを対話制御戦略に利用することを目指し，そのための予備実験を行った．

まず，システムが誤応答リカバリー発話を行うことによって，ユーザのシステムに対する主観的評価はどのように変動するか，および正応答の場合に誤応答リカバリー発話を行う・誤応答なのにリカバリー発話を行わないという誤った行動をシステムがとることによってユーザのシステムに対する主観的評価にどの程度影響するかを調べた．その結果,「誤応答・謝罪あり」の場合が,「正応答・謝罪あり」および「誤応答・謝罪なし」より良い結果となっていること,「正応答・謝罪あり」と「誤応答・謝罪なし」では有意差がないことがわかった．

次に，システムが通常発話を行った後，それに対する誤応答リカバリー発話を行うタイミングはどのようなものが妥当かを調べた．その結果，通常発話終了後 3 秒後に誤応答リカバリー発話を行うものが一番評価が高かったが，通常発話を中断して誤応答リカバリー発話を行うことが最適であるとするユーザ群も一定数存在することがわかった．

さらに，システムの誤応答の検出をユーザを映した画像情報のみから行うことはどの程度可能かを調べた．その結果，画像情報のみでは，人間でも判別が難しいことがわかった．

今後は，これらの結果がより複雑な対話シーンでも一致するかどうかを調べるとともに，この結果を対話制御戦略に組み込んだ対話を実現し，その評価を行う予定である．

また，画像情報に加え音韻情報や音声認識の信頼度をも手がかりとして，誤応答をより精度よく判別できるようなアルゴリズムを考案していく予定である．

参 考 文 献

1) 藤江真也, 江尻康, 菊池英明, 小林哲則: 肯定的/否定的発話態度の認識とその音声対話システムへの応用, 電子情報通信学会論文誌 D(II), Vol. J88-D-II, No. 3, pp. 489–498 (2006).
2) 中島慶, 江尻康, 藤江真也, 小川哲司, 松坂要佐, 小林哲則: 対話ロボットの動作に頑健な頭部ジェスチャ認識, 電子情報通信学会論文誌 D, Vol. J89-D, No. 7, pp. 1514–1522 (2006).
3) Rana El Kaliouby and Peter Robinson: Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures, *Proc. of the 2004 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops* (CVPRW'04) (2004).
4) 香山健太郎: 映像情報を用いた音声対話, 情報処理, Vol. 52, No. 1, pp. 79–86 (2011).
5) 河原達也, 川嶋宏彰, 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, Vol. 49, No. 8, pp. 912–918 (2008).
6) 角薫, 長田瑞恵: エージェントの表情と言葉が応諾行動に与える影響, 電子情報通信学会技術研究報告 ヒューマンコミュニケーション基礎, Vol. 108, No. 238, HCS2008-42, pp. 7–13 (2008).