Vol.2015-MUS-107 No.5
2015/5/23

[招待講演] 音声と音楽による人間・機械間メタコミュニケーション

伊藤 彰則[1,a)]

概要：音声・音楽を中心とした人間と機械のコミュニケーションについて考察する。音声認識技術の発展により、音声認識・合成を用いたユーザインタフェース（音声対話システム）が実現可能になってきているが、現在の音声対話システムに欠けているのはメタコミュニケーション（コミュニケーションチャネルに関するコミュニケーション）であると考えている。同様のことは、音楽エンタテインメントであるカラオケでの歌声評価などにも言うことができる。本発表では、音声・音楽（特に歌声）の機械による評価におけるメタコミュニケーションの意義と可能性について論じる。

1. はじめに

音声言語による人間と機械の対話の研究は古くから研究されており [1]，多くのシステムが作られている [2]．このような音声対話システムの多くは，ある特定のタスクを遂行するために設計されている．これらのタスクとは，例えば施設案内 [3]，観光案内 [4]，カーナビゲーション [5]，ロボット操作 [6] などである．

このような研究の多くは，ユーザとシステムの関係，あるいはユーザの「利用態度」について，いくつか暗黙の仮定を置いていると考えられる．これらは例えば次のようなものである．

- ユーザは自分が行うべきタスクを知っている．
- ユーザは，システムに情報を入力することで，そのタスクを遂行したいと考えている．
- ユーザはタスクに関する情報を速やかに入力できる．
- ユーザは，システムとの対話中には，その対話に集中していて，他のことをしたりしていない．
- ユーザはタスクと無関係な発話をしない．

これらは，人間と機械のコミュニケーションチャネルが理想的であるという仮定である．人間同士の通常の対話では，これらの仮定が常に成り立っているわけではない（もちろん，人間と機械の対話でも，実際は成り立っていないと思われる）．そのため，人間同士の対話では，対話のためのコミュニケーションチャネルに関する様々な調整が常に話者間で行われている．

コミュニケーションに関するコミュニケーションは「メタコミュニケーション」と呼ばれる [7]．その中には言語的なものと非言語的なものが混在している [8]．これらのメタコミュニケーションは人間間の対話では自然に行われているが，機械との対話においてどのようなメタコミュニケーションが必要かは明らかではない．一方，直接の対話だけでなく，音楽演奏のようなパフォーマンスとその鑑賞も，対人関係の中で行われる限りにおいてコミュニケーションである．例えばカラオケ等で，すでにその場にいる全員が知っている曲を歌う，という行為は，その行為自体に新たな情報がないという点で，メタコミュニケーションだけからなるコミュニケーションなのかもしれない．

2. メタコミュニケーション

「メタコミュニケーション」という用語を使い始めたのは Bateson であると言われ [7]，そこでは人間同士のインタラクションだけでなく，動物間の闘争とじゃれ合いの違いなどにもメタコミュニケーションの概念を適用している [7]．一方，「コミュニケーションに関するコミュニケーション」のうち，言語現象のみをメタコミュニケーションと呼んだり [9], [10]，言語的なものをメタコミュニケーション，非言語的なものをパラコミュニケーションと呼んで区別している研究もある [8]．音声対話システムの文脈では，その多くが言語的な現象（聞き返し，確認など）をメタコミュニケーションと呼んでいるようである [11], [12]．本稿では，言語的なものと非言語的なものをまとめてメタコミュニケーションと呼ぶ．

[1] 東北大学 大学院工学研究科
Grad. Sch. Eng., Tohoku University, Sendai 980–8579, Japan
[a)] aito@spcom.ecei.tohoku.ac.jp

メタコミュニケーションにも様々なものが考えられる．これらを分類すると，例えば次のようになる．

1. 物理的な位置関係 [13]
2. 相手の外見 [14], [15]
3. コミュニケーションチャネルの開始と終了
   (a) 話しかけやすさ [16], [17]
   (b) 対話の開始を促す発話
4. コミュニケーションチャネルの維持
   (a) 相手の話を聞いていることの確認（バックチャネル）[18]
   (b) ターンテイキング [19]
   (c) 雑音への対応 [20], [21]
   (d) 発話相手の判別
5. コミュニケーションの内容と処理
   (a) 相手の話を理解しているかどうかの表出と理解 [22], [23], [24], [25]
   (b) 聞き返し，確認
6. コミュニケーションの評価，態度
   (a) 内容への興味 [26]
   (b) 話の盛り上がり [27], [28]
   (c) 真摯さ，自信，信頼性 [29], [30]

次節以降，これらのメタコミュニケーションに関連して，筆者のグループが行った研究について紹介する．

3. 対話におけるメタコミュニケーション

3.1 うるさい場所では丁寧に話してください

雑音環境下での音声対話は難しい課題である．その原因は雑音環境下での音声認識が難しいことであり，これに対しては現在でも多くの研究がなされている [31]．現在の雑音下での自動音声認識性能は人間よりもだいぶ低いが，人間であっても雑音下での音声の認識は難しい．このような時に，人間同士の会話では「近づいて話す」「大きな声で話す」「ゆっくり話す」などの現象が起きる [32]．人間と機械が対話をする時に，機械の方が人間に対して「ゆっくり話してください」と頼む，というシステムは十分あり得るのではないだろうか．

そこで筆者らは，雑音環境に合わせて話し方を変えるようユーザに頼む対話システムを開発した [20]．このシステムは，SN 比などの観測値からニューラルネットで推定した単語正解精度に基づき，精度が高い場合はユーザに自由発話を許し，精度が低い場合には限られた内容だけを答えるようユーザに依頼する（図 1）．ユーザが限られた内容しか話さないことが事前に分かっているのであれば，小さい辞書と文法を用意することによって，雑音環境下でも認識精度を高く保つことができる．ただし，システムがユーザに回答方法を依頼しても，ユーザがその通りに答えるとは限らない点がポイントである．ユーザへの依頼方法と「ユーザがその通り話すかどうか」について調べた結果，「単語で答えて下さい」という依頼をしても，ユーザは必ずしもその通り答えてくれないことがわかった．この研究では，ユーザが答えやすい依頼の仕方を何通りか選び，それを切り替えることで認識精度を高く保つことができた．

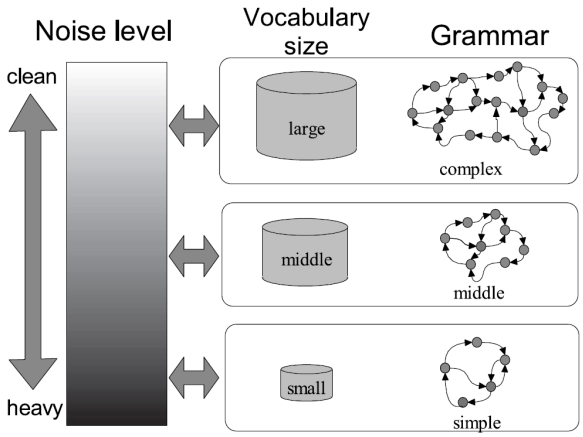

図 1 雑音環境に適応した対話戦略 [20]
Fig. 1 Adaptation of dialog strategy according to noise environment [20]

3.2 話し相手はいた方がいいのか

対話システムでは，仮想的な話し相手として話者のアニメーションを表示することが多い [14].「話し相手が見える」という状況は非常に強力なメタコミュニケーションである．表示される仮想的話者（エージェント）は，デフォルメされたゆるキャラ的な外観 [3] からリアル〜劇画調の外観 [33] まで様々であるが，「エージェントがいた方がいいのか」「エージェントがいた場合といない場合で何かが変わるのか」についての研究は案外少ない印象である．Bickmore らの研究 [14] では，「単純な会話のときには画像があった方がよいが，雑談のような会話では音声だけの方がよい」という結果が示されているが，これもエージェントの外観や動作などに依存するようである．

我々は，拡張現実感を使って，エージェントを通じた「モノとの対話」を試みている [15]．このシステムでは，VR 技術によって操作対象の近くにエージェントを出現させ，そのエージェントとの対話を通じて機器を操作する（図 2）．主観評価の結果，エージェントがいることによって「システムへの話しかけやすさ」が向上することが確かめられた．また，ゆるキャラ的な外観のため，話しかける口調が親しい口調となり，エージェントなしの場合に比べてユーザの発話単語数が減少した．

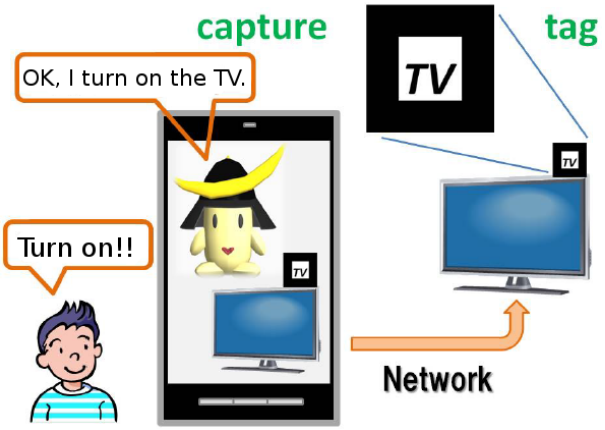

図 2 仮想エージェントを使ったモノとの対話 [15]
Fig. 2 Dialog with an object through a virtual agent [15]

3.3 対話ロボットの高さはどれくらいが適切か

話者同士の物理的な上下関係は，発話における話者間の関係に影響する．実際，背の高い人の方が収入や社会的地位が高い傾向にあることが知られており [34]，テレビ会議などでは話者と相手を表示するモニタの位置関係によって会話の支配率が変化する [35]．したがって，ロボットのように物理的実体があるものとの対話では，相手の大きさが対話に影響すると考えられる．

我々は，高さが変更できるロボットを使い，どのくらいの高さのロボットが最も対話しやすいかを調べた [13]．使用したロボットを図 3 に示す．実験の結果，最も話しやすいロボットの高さは，話者の目の高さよりおよそ 30cm 低い位置であった．また，最も好まれる高さよりも ± 20cm 高さをずらして対話をすると，話しやすさが悪化することがわかった．

高さだけでなく，相手が移動ロボットの場合は「どこまで近づいてよいか」も問題である [36]．これはパーソナルスペース [37] と関連しており，ロボットと対話をする時に「どこまで近づいてもよいか」「どこまで離れてもよいか」は重要な研究課題に思われる．

3.4 聞いてない相手に叫ぶ

音声で命令できるはずのロボットに話しかけたとき，ロボットが動かなかったらどうするだろうか．我々は，ロボットがコマンド音声の通りに動かなかった場合にユーザの音声がどのように変化するのかを調査した [38]．図 4 は，実験の概略である．小型のロボットを音声コマンドで止めるというタスクにおいて，故意にコマンドを無視する試行を設け，その時にオーバーランしたロボットに対する音声のパラ言語的特徴を分析した．当初は（図にあるように）コマンド音声は大きく，速くなると予想したが，予想に反して声の大きさには有意差はなく，発話速度はかえって遅くなった．また，F0 はわずかに大きくなった．発話が遅

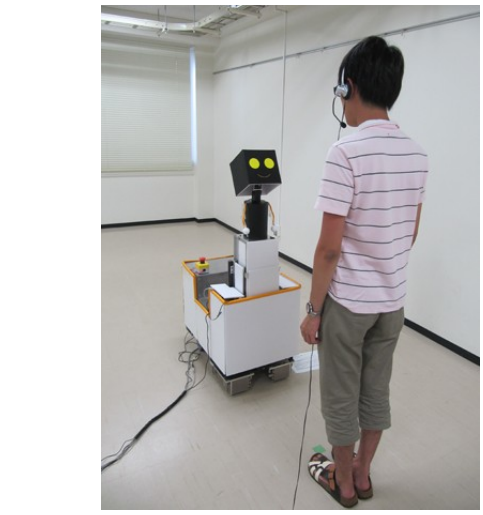

図 3 高さ可変なロボットとの対話実験 [13]
Fig. 3 Dialog with a variable-height robot [13]

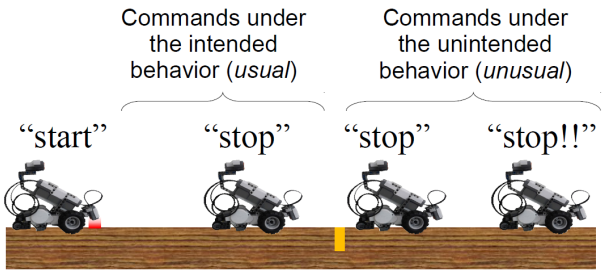

図 4 コマンドを無視するロボットに対する発話 [38]
Fig. 4 Uttrances toward a robot that does not listen to the command [38]

く，F0 が高くなるという特徴は雑音下発声と類似している [32] という点が興味深い，ディスコミュニケーションに起因する音声の変化は，雑音の影響に限らず一般的にみられる特徴なのかもしれない．

3.5 しゃべらない相手にどうしたらいい？

誰かに何かを質問されたとき，すぐに答えられない場合がある．これは，相手の言うことが理解できずに戸惑っているのかもしれないし，言っていることはわかるが答えが思い出せないのかもしれない．この様子を図 5 に示す．音声対話システムでユーザから一定時間返答がない場合に，システムがプロンプト発話を繰り返すという手法 [39] が広く使われているが，これはユーザが必死に答えを思い出そうとしている場合は大きなお世話である．そこで，ユーザのこの 2 つの状態を識別するという研究を行った [24], [25], [40]．最初は特徴量として最初の無音区間の長さ，フィラーの長さ，顔の向きなどを利用して識別を行った [24]．その後，Bag-of-Features に類似の方法を利用し，特徴量にも視線情報などを加えて実験を行っている [25]．最終的に，ユーザのターンから 10 秒以内に 70%程度の精度で前記の 2 状態が識別できた．特徴量の組み合わせ結果

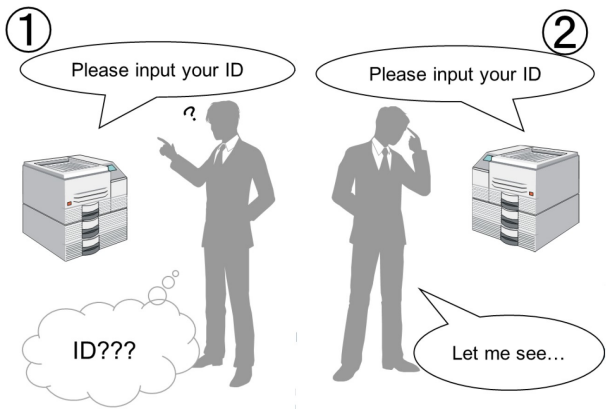

図 5 応答発話ができないユーザの 2 種類の状態 [40]
Fig. 5 Two states of the user who has difficulty making an answer utterance [40]

から，最も強力な手掛かりは視線であるという結果が得られている．このようなメタコミュニケーションに関するマルチモーダル情報の重要性が定量的に得られるのは珍しいのではないだろうか．

このように「相手が自分の問いに答えようとしているか」を推測する能力は FOAK (Feeling Of Another's Knowing) と呼ばれている [22]．このように，自分の考えている状態を顔や声に（無意識的に）出してしまう性質と，それを読み取る能力が対になってメタコミュニケーションを形成している．

3.6 「早く答えろよ」というプレッシャー

対話をしているときには，片方の話者が話してからもう片方の話者が話すまでに少し間があくことが多い．この「話者が後退するときの間」(交替潜時) は対話において観測される重要な現象の一つであり，多くの研究がなされている [41]．自然な交替潜時の値は多くの言語で 300〜600ms あたりに分布しているが [41], [42]，第二言語，特に十分習得していない第二言語では素早い話者交替ができず，交替潜時が大きくなる．特に，CALL システムのように対話相手が機械である場合には，システムが用意するエージェントの制約により，話者交替のためのキューを十分表出できないことなどから，交替潜時が大きくなりがちである．CG キャラクタに自然なインタラクションをさせるというアプローチもありうるが [43], [44]，韻律だけでなく CG キャラクタの視線やジェスチャなどをすべて制御することは簡単ではない．そこで，CG キャラクタを利用した英会話練習をターゲットに，「自然ではないが明確な方法」で話者交替のための手がかりを出す方法（タイムプレッシャー）を開発した [45]．

利用した CG キャラクタとタイムプレッシャー表現を図 6 に示す．キャラクタは円筒と立方体からなる必要最小限のロボットの形をしている．キャラクタが発話した後，

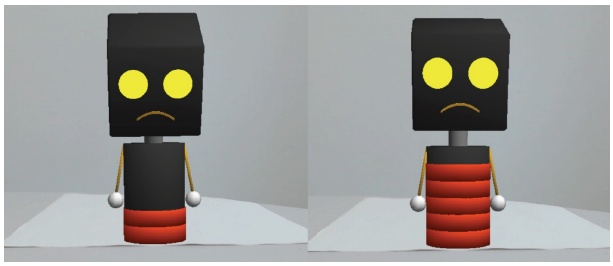

図 6 CG キャラクタとタイムプレッシャー表現 [45]
Fig. 6 CG character with time pressure expression [45]

キャラクタの色が下から赤くなっていくことで，人間の話者に早く発話するよう促す．タイムプレッシャー表現の導入により，交替潜時の値は 200ms ほど短くなり，人間同士の交替潜時の値に近づく．

4. 音楽におけるメタコミュニケーション

4.1 熱唱度

音楽は自然言語による対話のように直接意味内容を伝えるためのメディアではないが，音楽演奏や歌唱によって，音楽が表現しているリズムやメロディに加えて「感情」や「ムード」が伝達すると考えられている [46], [47]．

一方我々は，「歌唱者がどれだけ一生懸命歌っている」(ように聞こえる) かを表す「熱唱度」(singing enthusiasm) という概念を提案した [48]．熱唱度には，「歌唱者がどれだけ熱唱のつもりで歌ったか」(本人熱唱度) と，「聴取者がその歌声をどの程度熱唱に感じたか」(知覚熱唱度) の 2 つがある．熱唱度は感情と似ているところもあり，感情の次元でいう覚醒の軸と近い．しかし，「熱唱」をしている歌唱者が興奮しているとは限らず，また熱唱の歌を聞いた聴取者が興奮するとも限らないので，熱唱度は感情とは別な評価軸だといえる．歌唱者の実際の興奮とは独立に「興奮したかのような」歌唱をするという点で，「演じられた感情」(acted emotion) と類似しているといえる．

4.2 知覚熱唱度の推定

聴取者が知覚する熱唱度の推定には，基本的に声の大きさが関連している．しかし，歌唱音声のパワーを一定に正規化したとしても，複数の評価者が比較的安定して熱唱度を知覚することがわかっている．また，カラオケの採点等に「熱唱」を導入することを考えると，声のパワーの絶対値は口とマイクの距離やマイクアンプのゲインによって変化するため，パワーをそのまま利用するのは好ましくない．そこで，声のパワーを正規化したうえで，知覚熱唱度を自動推定するための手法について検討した [48]．これを図 7 に示す．

ターゲットとなる「知覚熱唱度」は，コーパスに含まれる歌唱に対して評価者が 3 段階で点数付けし，それを平均した値である．これを推定するための特徴量として，「A 特

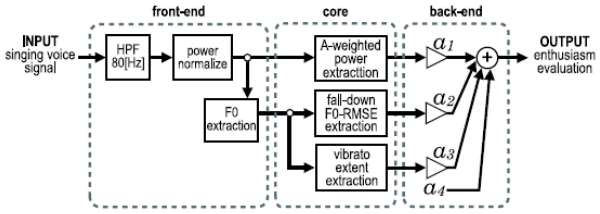

図 7 知覚熱唱度推定手法の概略 [48]
Fig. 7 Overview of estimating the singing enthusiasm[48]

性パワー」「ビブラート」「ずり下げ特徴量」の 3 つを利用した．これらを重回帰によって組み合わせることにより，推定値と知覚熱唱度の間で 0.66 程度の相関を得ることができた．

4.3 本人熱唱度の推定

知覚熱唱度に比べて，本人熱唱度の推定は難しい．これは，熱唱度推定に用いる特徴量の大小が個人に大きく依存しており，個人差の方が本人熱唱度の大小による変動よりも大きいためである．知覚熱唱度の推定には短い時間での特徴量を利用していたが，本人熱唱度を推定するためには，より長い（曲全体の）区間での特徴量の分布を手掛かりにする必要がある．

そこで，曲全体での入力音声のパワーの変動を使って本人熱唱度を識別する手法を検討した [49]．入力音声を数秒程度のセグメントに区切って，セグメント内のパワーの平均と分散を求めたところ，本人が熱唱している場合には，通常の部分とサビの部分でのパワーの平均・分散に差があることがわかった．そこで，すべてのセグメントにおけるパワーの「平均の平均」「平均の分散」「分散の平均」「分散の分散」を特徴量とすることで，「熱唱」「非熱唱」の 2 クラス識別において 72%の識別率を得ることができた．

5. むすび，あるいは我々はどこへ行くのか

筆者の行ってきた研究を中心に，コミュニケーションの対象であるコンテンツの周辺にあるコミュニケーション，「メタコミュニケーション」について述べた．メタコミュニケーションというと何やら怪しい響きがあるが，筆者は「音声や音楽のメタコミュニケーション」としてまとめられる研究をこれまで行ってきたし，こういう切り口で見れば他の多くの研究者もこの分野での研究を行ってきていることがわかる．

音声認識の研究は，「声を使った入力インタフェース」つまりしゃべった内容がそのままコンピュータに入力されることが理想であって，「どういう状況で入力するのか」「なぜ声で入力するのか」「なんのために入力するのか」を必ずしも問わない．それは，「音声認識」という技術がちょうどよいまとまりをもった技術だからだと思う．では，「音声対話」はどうだろうか．入力の内容，入力の目的，入力装置とユーザの関係と切り離された「音声対話」というものがありうるだろうか．メタコミュニケーションの機械実装を目指す研究は，それを具備した機械が人間にとって何なのか，という議論と切り離して論じることはできないと思う．

人間同士で行われているようなメタコミュニケーションをコンピュータが精密に複製することは難しいであろうが，本気でやろうとすれば原理的には不可能ではないのだと思う．それが完全にできた時，コンピュータは人間にとってどういうものになっているだろうか．「人間に対して情緒的にふるまっているかのように見えるロボットは開発すべきでない」という意見がある [50]．このような議論は主にロボット分野で進んでいるが [51]，音声や音楽の分野でこういう議論は必要ないだろうか．いまの技術段階では少々背伸びした議論なのかもしれないが，いずれ必要になるように思われる．

謝辞

ここで紹介した研究内容は，多くの方々との共同研究の成果である．ここに謝意を示したい．

参考文献

[1] S. R. Young. Use of dialogue, pragmatics and sematics to enhance speech recognition. *Speech Communication*, 9(5–6):551–564, 1989.
[2] C. Lee, S. Jung, K. Kim, D. Lee, and G. G. Lee. Recent approaches to dialog management for spoken dialog systems. *Journal of Computing Science and Engineering*, 4(1):1–22, 2010.
[3] T. Cincarek, H. Kawakami, R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano. Development, long-term operation and portability of a real-environment speech-oriented guidance system. *IEICE Trans. Inf. & Syst.*, E91-D(3):576–587, 2008.
[4] H. Kashioka, T. Misu, E. Mizukami, Y. Shiga, K. Kayama, C. Hori, and H. Kawai. Multimodal dialog system for kyoto sightseeing guide. In *Proc. APSIPA ASC*, 2011.
[5] T. Misu, A. Raux, I. Lane, J. Devassy, and R. Gupta. Situated multi-modal dialog system in vehicles. In *Proc. of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction*, pages 25–28, 2013.
[6] X. Chen, J. Ji, J. Jiang, G. Jin, F. Wang, and J. Xie. Developing high-level cognitive functions for service robots. In *Proc. 9th Int. Conf. on Autonomous Agents and Multiagent Systems*, pages 989–996, 2010.
[7] R. W. Mitchell. Bateson's concept of "metacommunication" in play. *New Ideas in Psychology*, 9(1):73–87, 1991.
[8] B. D. Beitman and G. I. Viamontes. Unconscious role-induction: Implications for psychotherapy. *Psychiatric Annals*, 37(4):259–265, 2007.
[9] J. A. Levin and J. A. Moore. Dialogue-Games: Metacommunication structures for natural language interaction. *Cognitive Science*, 1(4):395–420, 1977.
[10] D. L. Sanford and J. W. Roach. Representing and using metacommunication to control speakers' relationships

in natural-language dialogue. *International Journal of Man-Machine Studies*, 26(3):301–319, 1987.

[11] J. Bateman, E. Hagen, and A. Stein. Dialogue modeling for speech generation in multimodal information systems. In *Proc. ESCA Workshop on Spoken Dialogue Systems*, pages 225–228, 1995.

[12] L. Dybkjær and N. O. Bernsen. Usability evaluation in spoken language dialogue systems. In *Proc. of the workshop on Evaluation for Language and Dialogue Systems*, volume 9, 2001.

[13] Y. Hiroi, T. Nakayama, H. Kuroda, S. Miyake, and A. Ito. Effect of robot height on comfortableness of spoken dialog. In *Proc. 5th Int. Conf. on Human System Interaction*, 2012.

[14] T. Bickmore and J. Cassell. *Social Dialongue with Embodied Conversational Agents*, volume 30 of *Text, Speech and Language Technology*, chapter 1, pages 23–54. Springer, 2005.

[15] S. Miyake and A. Ito. A spoken dialogue system using virtual conversational agent with augmented reality. In *Proc. APSIPA ASC*, 2012.

[16] S. Hudson, J. Fogarty, C. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. Lee, and J. Yang. Predicting human interruptibility with sensors: a wizard of oz feasibility study. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, pages 257–264, 2003.

[17] S. Satake, T. Kanda, D.F. Glas, M. Imai, H. Ishiguro, and N. Hagita. How to approach humans? — strategies for social robots to initiate interaction. In *Proc. 4th ACM/IEEE Inter- national Conference on Human-Robot Interaction*, 2009.

[18] R. Poppe, K. P. Truong, and D. Heylen. Perceptual evaluation of backchannel strategies for artificial listeners. *Autonomous Agents and Multi-Agent Systems*, 27(2):235–253, 2013.

[19] A. Raux and M. Eskenazi. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Trans. Speech Lang. Process.*, 9(1):1:1–1:23, May 2012.

[20] A. Ito, T. Oba, T. Konashi, M. Suzuki, and S. Makino. Selection of optimum vocabulary and dialog strategy for noise-robust spoken dialog systems. *IEICE Trans. on Inf. & Syst.*, E91-D:538–548, 2008.

[21] K. Kogure, M. Yoshinaga, H. Suzuki, and T. Kitahara. A spoken dialogue system for noisy environment. In *HCI International 2014 - Posters' Extended Abstracts*, volume 435 of *Communications in Computer and Information Science*, pages 577–582, 2014.

[22] 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. *Journal of Memory and Language*, 34(3):383–398, 1995.

[23] M. Swerts and E. Krahmer. Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, 53(1):81–94, 2005.

[24] Y. Chiba and A. Ito. Estimating a user's internal state before the first input utterance. *Advances in Human Computer Interaction*, Article ID 865362:10 pages, 2012.

[25] Y. Chiba, T. Nose, A. Ito, and M. Ito. User modeling by using bag-of-behaviors for building a dialog system sensitive to the interlocutor's internal state. In *Proc. SIGdial*, 2014.

[26] W. Y. Wang and J. Hirschberg. Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning. In *In Proc. SIGDial*, pages 151–161, 2011.

[27] R. Tokuhisa and R. Terashima. Relationship between utterances and "enthusiasm" in non-task-oriented conversational dialogue. In *Proc. SIGdial Workshop on Discourse and Dialogue*, pages 161–167, 2006.

[28] Y. Moriya, T. Tanaka, T. Miyajima, and K. Fujita. Estimation of conversational activation level during video chat using turn-taking information. In *Proc. 10th Asia Pacific Conf. on Computer Human Interaction*, pages 289–298, 2012.

[29] M. Kurisu, K. Mera, R. Wada, Y. Kurosawa, and T. Takezawa. A method using acoustic features to detect inadequate utterances in medical communication. In *IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC)*, pages 116–119, 2012.

[30] C. Jost, B. Le Pévédic, and D. Duhaut. Study of the importance of adequacy to robot verbal and non verbal communication in human-robot interaction. In *Proc. Int. Symp. on Robotics*, 2012.

[31] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach. An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 22:745–777, 2014.

[32] J.-C. Junqua. The Lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93(1):510–524, 1993.

[33] T. Kuratate, K. Ayers, J. Kim, and D. Burnham. Exploring the uncanny valley effect with talking heads. In *Proc. Interspeech*, 2008.

[34] T. Judge and M. D. Cable. The effect of physical height on workplace success and income: Preliminary test of a theoretical model. *Journal of Applied Psychology*, 89(3):428–441, 2004.

[35] W. Huang, S. J Olson, and M. G. Olson. Camera angle affects dominance in video-mediated communication. In *Proc. Conf. on Human Factors in Computing Systems (CHI02)*, pages 716–717, 2002.

[36] Y. Hiroi and A. Ito. Effect of the size factor on psychological threat of a mobile robot moving toward human. *KANSEI Engineering International*, 8(8):51–58, 2009.

[37] K. B. Little. Personal space. *Journal of Experimental Social Psychology*, 1:237–247, 1965.

[38] N. Totsuka, Y. Chiba, T. Nose, and A. Ito. Robot: Have I done something wrong? —Analysis of prosodic features of speech commands under the robot's unintended behavior—. In *Proc. Int. Conf. on Audio, Language and Image Processing*, 2014.

[39] N. Yankelovich. How do users know what to say? *Interactions*, 3(6):32–43, 1996.

[40] Y. Chiba, M. Ito, and A. Ito. Modeling user's state during dialog turn using hmm for multi-modal spoken dialog system. In *Proc. Int. Conf. on Advances in Computer-Human Interactions*, pages 343–346, 2014.

[41] M. Heldner and J. Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

[42] D. Reidsma, I. de Kok, D. Neiberg, S. C. Pammi, B. van Straalen, K. Truong, and H. van Welbergen. Continuous interaction with a virtual human. *J Multimodal User Interfaces*, 4:97–118, 2011.

[43] D. Traum and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proc. the 1st Int. Joint Conf. on Autonomous agents and multiagent systems: part 2*, pages 766–773, 2002.

[44] P. Wik and A. Hjalmarsson. Embodied conversational

agents in computer assisted language learning. *Speech Communication*, 51(10):1024–1037, 2009.

[45] N. Suzuki, T. Nose, Y. Hiroi, and A. Ito. Controlling switching pause using an ar agent for interactive call system. In *HCI International 2014 - Posters' Extended Abstracts*, volume 435 of *Communications in Computer and Information Science*, pages 588–593. Springer, 2014.

[46] A. Gabrielsson and P. N. Juslin. Emotional expression in music performance: Between the performer's intention and the listener's experience. *Psychology of Music*, 24(1):68–91, 1996.

[47] P. N. Juslin and P. Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5):770–814, 2003.

[48] R. Daido, M. Ito, S. Makino, and A. Ito. Automatic evaluation of singing enthusiasm for karaoke. *Computer Speech & Language*, 28:501–517, 2012.

[49] A. Ito. Assessing the intended enthusiasm of singing voice using energy variance. In *Proc. Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, pages 558–561, 2014.

[50] S. Bringsjord and M.H. Clark. Red-pill robots only, please. *IEEE Trans. Affective Computing*, 3(4):394–397, 2011.

[51] G. Veruggio. Roboethics. *IEEE Robotics & Automation Magazine*, 17(2):105–109, 2010.