音声対話システムにおけるユーザの許容範囲を考慮した多様な同調応答の検討

Study on Various Adaptive Responses in Spoken Dialogue Systems Considering Users' Acceptable Range

菊池浩史 [1*]　楊潔 [1]　菊池英明 [1]
KIKUCHI Hirofumi[1]　YANG Jie[1] and KIKUCHI Hideaki[1]

[1] 早稲田大学
[1] WASEDA University

Abstract: ユーザが許容できないパラ言語情報での応答を音声対話システムがすることによって、ユーザの対話継続欲求が下がる問題が存在する。本研究では、このような破綻の問題の解決を目指す。我々はこれまでにユーザ発話へのシステム応答に対する許容範囲の存在を、一名の話者によるユーザ発話音声を用いて確認した。本稿では、9 名の話者によるユーザ発話音声を収録したうえで、聴取評価実験を行い、多様なユーザ発話での許容範囲を確認した。

1　はじめに

情報技術の発展により音声対話システムの普及が進んでいる。近年、我が国では少子高齢化などの社会情勢により対話の機会が減っている。そうしたなか、音声対話システムは孤独感の解消や介護・子守の現場への活用が期待されている。しかしながら、音声対話システムが同じ応答を繰り返すことによってユーザが飽きてしまう問題がある。対話を続けたい・また対話したいと思う欲求、対話継続欲求が低くなることが原因の一つとして挙げられる。以上の問題を解決するために、音声対話システムの同じ応答の繰り返しを避け、多様なシステム応答を生成する必要がある。

鈴木らはシステムがユーザ発話のパラ言語情報を反響的に模倣しビープ音で応答することでユーザの志向的な姿勢が誘発されることを示唆 [1] し、発話の音調の多様性をイントネーションで表現できる可能性を示した。一方、宮澤らは人と音声対話システムの対話において、ユーザの対話継続欲求を高めるにはシステムがユーザに対して「話を聞いてもらえるという実感を与えること」が有効であると示唆している [2]。このことから、ユーザ発話に対するシステム応答のイントネーションを調節し多様なシステム応答を表現することで、ユーザの対話継続欲求を高めることが期待できると推測する。さらに鈴木らは、ユーザがシステムに対し人間の成人と同等な自律的振る舞いを期待することが考えられるため、模倣では必ずしもポジティブな印象を持つとは限らない [1] とも述べている。つまり、音声対話システムでは模倣のみではない多様な同調応答の実装が必要であると推測できる。一方、竹内らは、人間と人格化したエージェントとのインタラクションが人間同士のインタラクションと同様に社会的であるとの示唆を得たうえで、「一般的な社会性から逸脱したエージェントの振る舞いは，人間とエージェントによる社会的関係を形成するうえで障害となる.」[3] と述べている。

多様な応答を実現する際、ユーザが許容できないパラ言語情報での応答を音声対話システムがすることによって対話が破綻する恐れがある。これまでに、筆者らは、ユーザ発話とシステム応答のパラ言語情報として表出された発話者の快不快状態に着目し、ユーザ発話へのシステム応答に対するユーザの許容範囲について、一名のユーザ発話音声を用いた音声聴取評価実験によって調査した。一名のユーザ発話音声を用いた音声聴取評価実験では、被験者による共通する許容評価が広がり、分布していることが確認され、ユーザ発話へのシステム応答に対するユーザの許容範囲の存在を示唆した [4]。また、許容範囲の特性として、

1. ユーザ発話とシステム応答の快不快状態が同じ快または不快のときに許容される傾向がある
2. ユーザ発話とシステム応答の快不快状態が同じ快・不快であっても、ユーザ発話の快不快状態に

*連絡先：早稲田大学
〒 359-1192 埼玉県所沢市三ヶ島 2-579-15
E-mail: hirofumi.kikuchi@toki.waseda.jp

対してシステム応答の快不快状態が過剰に強い快または不快を表出するときに許容されない傾向がある

3. メッセージ性の強度によって、どのユーザ発話に対しても許容されないシステム応答音声がある

を確認した。ただし、[4] ではユーザ発話として話者 1 名の音声のみを扱っていた。ユーザの許容範囲を音声対話システムに実装するためには、ユーザ発話について広く調査する必要がある。そこで、本稿ではユーザ発話者 9 名によるユーザ発話を収録し、音声聴取評価実験を行なった。

2 研究手法

2.1 発話内容

本研究では、ユーザとシステムの 1 発話ずつの対話に着目する。ユーザ発話は、特定の感情の影響を比較的受けにくく、直前に接続する文脈によって強い当事者意識で発話できる「連絡を待っています」である。システム応答は、汎用性が高く発話時間がある程度ある相槌である「そうですか」を用いた。なお、本研究では快不快次元を扱うため、基本 6 感情をステレオタイプで発話した音声ではなく、表現豊かな多様な音声が必要となった。そこでシステム応答には、女性声優 5 名による、人物像 10 種類、シチュエーション 28 種類、平静音声 1 種類、合計 1405 種類の「あーそうですか」が収録されている「表現豊かな音声コーパス」[5] から抜粋して使用した。本実験では「あーそうですか」の「そうですか」部分のみをシステム応答として 45 種類用意した。45 種類の音声試料は不快から快までほぼ均等に分布（識別値：-0.069〜0.807）している。

2.2 快不快識別器

本研究において、快不快状態を数値で扱う上では、Fairy Devices 株式会社が提供する音声感情識別器の実装を用いた。この識別器は LLD (Low Level Descriptor) 音響特徴量の BoAW (Bag of Audio Words) 表現 [6] を用いた SVR (Support Vector Regressor) である。快不快の推定のために、UUDB[7]（宇都宮大学パラ言語情報研究向け音声対話データベース）release 1 の全データ（4840 発話、1 時間 53 分）を使用して感情識別器の訓練を行った。訓練の際、UUDB のパラ言語情報ラベルのうち ”快-不快” の全評価者の平均評価値を用いた。以下、この訓練済みの識別器を「快不快識別器」と呼ぶ。また快不快識別器が産出する値を「識別値」と呼ぶ。識別値は -1〜1 の範囲を取り、小さいほど不快、大きいほど快を表す。

3 音声収録

3.1 概要

ユーザ発話者の多様性を高めるため、9 名の音声収録を行った。被験者は 20〜60 代の 9 名（女性 5 名、男性 4 名、被験者名を A〜I と呼称）で、5 段階の快不快状態（強い不快・弱い不快・平静・弱い快・強い快）を表出してもらった。

収録にあたって、当事者意識を高く持つことと、5 段階の快不快状態が強い不快から強い快の表出がおおよそ均等に分布することを教示した。環境からの騒音をできるだけ抑え、発話のみを収録するために、単一指向性ヘッドセットマイク（DPA CORE 4288-DC-F-F00-MH）を、口角から横に 3cm の位置で使用した。マイクを IC レコーダ（TASCAM DR-100MKⅢ）に接続し録音することで音声データとして保存する。収録風景を図 1 に示す。収録終了の条件を以下に示す。

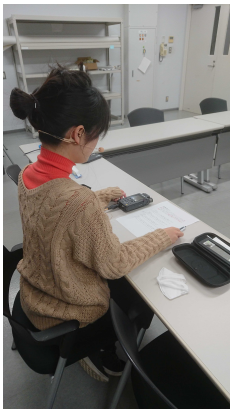

図 1: 収録風景

1. 全ての発話において、当事者意識が高いものであったと、実験者の判断が下された
2. 全ての快不快状態について収録が終わった
3. 収録された音声の識別値が偏りなく分布していることが確認された

3.2 結果

収録した音声の識別値の分布を図 2 に示す。収録終了の条件に記したように、被験者それぞれ 5 段階の快不快状態において、識別値はおおよそ均等に分布していた。一方、発話者によって識別値のとりうる範囲が異なることが明らかになった。

4 音声聴取評価実験

4.1 目的

ユーザ発話者 1 名による許容範囲の存在の示唆より、ユーザ発話者の多様性を高めた場合のユーザ発話へのシステム応答に対する許容範囲を明らかにする。

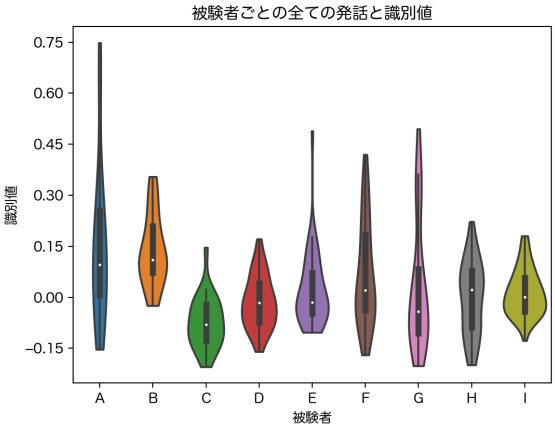

図 2: ユーザ発話の識別値の分布

4.2 仮説

本研究の目的を実現するために下記の仮説を立てた。

1. ユーザ発話とシステム応答の識別値は許容評価に影響を与える
2. ユーザ発話とシステム応答の識別値と許容評価の関係は線形ではない

4.3 概要

ユーザ発話「連絡を待っています」とシステム応答「そうですか」を結合した音声試料を被験者が聴取し、ユーザ発話が自分自身の発話であると仮定した時にシステム応答がどのくらい許容できるかを 7 件法（4. をどちらでもないとし、1. に近づくほど許容できない、7. に近づくほど許容できるとした）で回答してもらった。以後、得られた値を許容評価値と呼ぶ。また、教示は以下の通りにした。

1. ユーザ発話は自分自身の発話であると仮定
2. システムのキャラクター性は考慮しない
3. 深く考え込まず直感で答える

本実験は対話システムの応答が自分に向けてであることを前提としている。そのため、音声試料のユーザ発話が被験者自身の発話であるという仮定が必要不可欠である。また、システムの声そのものが許容できないといった、システムのキャラクター性は考慮しないこととした。最後に、対話は背景によって多様な解釈ができてしまうため、対話の背景を限定せず、被験者に素早く直感で答えてもらうことを優先とした。本実験はクラウドソーシングサイトであるクラウドワークス上で実施した。被験者は 20 代から 60 代の男女合計延べ 500 名であった。

4.4 音声試料

ユーザ発話は第 3 章音声収録で収録した発話者 9 名による音声を用いる。発話者ごとに、5 段階の快不快状態それぞれ 1 音声を以下の条件で抜粋した。

1. 5 段階の快不快状態の識別値がおおよそ等間隔になる
2. できるだけ識別値の幅が広くなる

ユーザ発話者 9 名、5 段階の快不快状態それぞれ 1 音声、合計 45 音声をユーザ発話として用いた。システム応答は 2.1 発話内容の 45 音声を用いた。ユーザ発話 45 音声とシステム応答 45 音声を総当たりで接続し、先行発話がユーザ発話、後続発話がシステム応答となる、2025 音声を用意した。2025 音声からランダムで 81 音声・25 組み用意し、実験 1 回あたり 81 音声を用い 25 回実施した。1 回の実験は 20 名ずつ行い、1 音声あたり 20 名の評価を得た。なお、ユーザ発話とシステム音声の間に 0.78 秒の無音を挿入する。無音時間は、オフライン状態の Amazon Echo Show 5 において、ウェイクワードから応答までの平均応答時間を用いた。

4.5 手順

実験サイトは Google ドライブ、Google サイト、Google フォームを用いて作成した（図 3）。図 3 左側の埋め込みプレーヤで音声を再生し、右側のフォームにて評価を行う。クラウドワークスのユーザ名、年齢、性別、全ての評価を入力することでフォームが送信できる。フォームを送信することで実験が終了する。

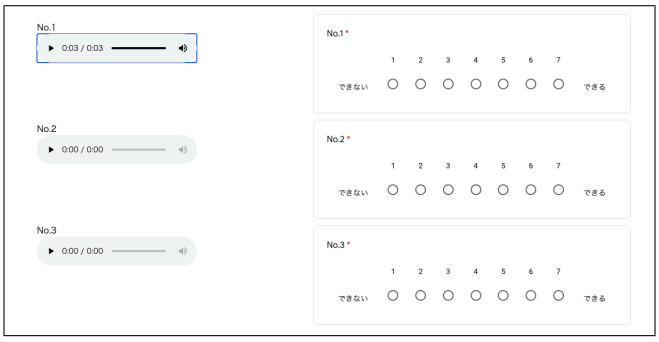

図 3: 実験サイト（一部抜粋）

4.6　結果

実験結果を散布図 4 に示す。横軸はユーザ発話の識別値、縦軸がシステム応答の識別値である。マーカの色はそれぞれユーザ発話とシステム応答の識別値が対応する音声の許容評価値の平均の大きさを色分けしている。許容評価値の平均が、1 以上 2 未満が紫、2 以上 3 未満が青、3 以上 4 未満が緑、4 以上 5 未満が黄色、5 以上 6 未満がオレンジ、6 以上が赤となっている。

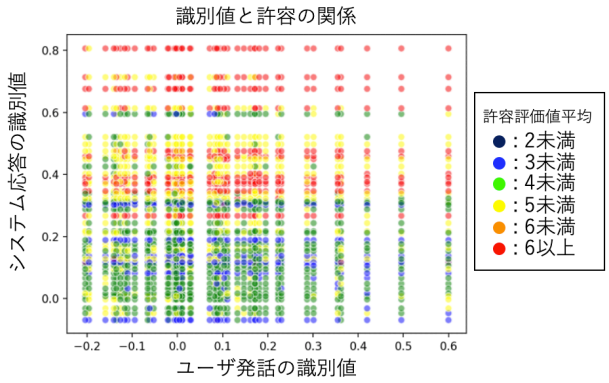

図 4: ユーザ発話とシステム応答の識別値と許容評価値平均

図 4 より、システム応答の識別値が小さいと全体的に許容評価が低くなる傾向がわかった。一方、システム応答の識別値が大きいと全体的に許容評価が高くなる傾向が確認された。ユーザ発話の識別値が大きくなるにつれて、システム応答の識別値が 0.4 付近の音声について、より許容できるようになることがわかった。このことから仮説 1 を支持している。一方、どのユーザ発話に対しても許容評価があまり変わらないシステム応答が多く見られることから、仮説 2 を支持している。

5　おわりに

本稿は、ユーザ発話へのシステム応答に対するユーザの許容範囲の調査を行なった。システム応答の識別値、大きいと許容評価が常に高くなる、小さいと許容評価が常に低くなることが確認された。このことは、ユーザが不快であってもシステム応答が快であるとき、評価が高くなる傾向があったとする [4] の考察を裏付けていると推測できる。一方、用意した全ての音声試料に対する結果では、識別値の変化による許容の分布の変化がみてとれたものの、ユーザ発話の識別値の変化によって大きな許容の変化は見られなかった。本研究では、表現豊かな音声をシステム音声に用いている。つまり、システム応答自体にシチュエーションに応じたメッセージが含まれている。そのため、システム応答全ての音声を一つのモノサシで観察することで、音声に含まれるメッセージを考慮しないことになってしまう。そのため、システム応答音声 1 つ 1 つとユーザ発話の識別値との許容評価の関係を見ていく必要があると考えられる。

謝辞

本研究は Fairy Devices 株式会社との共同研究により実施された。音声感情識別器をご提供いただいた同社に感謝する。

参考文献

[1] 鈴木紀子, 竹内勇剛, 石井和夫, 岡田美智男: 非分節音による反響的な模倣とその心理的影響, 情報処理学会論文誌, Vol.41, No.5, pp.1328-1338 (2000)

[2] 宮澤幸希, 小川義人, 松尾智信, 中山真太郎, 常世徹, 桝井祐介, 菊池英明: 音声対話システムにおける継続性向上の要因, 研究報告ヒューマンコンピュータインタラクション *(HCI)*, Vol.2011-HCI-142, No.1, pp.1-8 (2011)

[3] 竹内勇剛, 片桐恭弘: ユーザの社会性に基づくエージェントに対する同調反応の誘発, 情報処理学会論文誌, Vol.41, No.5, pp.1257-1266 (2000)

[4] 菊池浩史，楊潔，菊池英明: 音声対話システムの応答に対するユーザの許容範囲の調査–パラ言語に着目して–，*HIA* シンポジウム, *2020*，P-43 (2020)

[5] 宮島崇浩, 菊池英明, 白井克彦, 大川茂樹: 演技指示の工夫が与える音声表現への影響: 表現豊かな演技音声表現の獲得を目指して, 音声研究, Vol.17, No.3, pp.10-23 (2013)

[6] Schmitt, M., Ringeval, F., Schuller, B.: At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions, *Speech. Proc. Interspeech*, pp.495-499 (2016)

[7] 「宇都宮大学パラ言語情報研究向け音声対話データベース」, NII 音声資源コンソーシアム, URL: http://research.nii.ac.jp/src/UUDB.html[閲　覧日:2020 年 2 月 13 日]