音声対話エージェントの社会的存在感向上のための感情表現の音楽的強調

Emotional Expression Emphasized by a Melody Enhances Social Presence of a Spoken Dialogue Agent

髙橋ともみ[1]　田中一晶[1]　小林賢一郎[2]　岡夏樹[1]

Tomomi Takahashi[1], Kazuaki Tanaka[1], Kenichiro Kobayashi[2], and Oka Natsuki[2]

[1] 京都工芸繊維大学
[1] Kyoto Institute of Technology
[2] TIS 株式会社
[2] TIS Inc.

Abstract: A spoken dialogue agent such as a virtual assistant has few emotional expressions. Therefore, we propose a method to emphasize emotional expression by adding a melody which expresses the emotion of the agent. The experiment investigating effectiveness of the method showed that emphasis of melodic emotional expression can not only convey emotions clearly, but also enhance the humanity and ease of speaking to the agent. For practical use, we are also working on automatic generation of the melodies according to emotions.

1. はじめに

日常生活において AI スピーカや AI アシスタント等の音声対話エージェントがみられるようになってきた．しかしながら，2017 年に行われた調査では 7 割を超える日本人が音声操作に抵抗感を感じており[1]，この抵抗感がエージェントの普及を妨げる要因となっている可能性が考えられる．実際に，2018 年に行われた調査においても AI スピーカの認知度 76%に対して所有者は 6%と少なく，所有していない理由の中には依然として音声操作への抵抗感が挙げられていた[2]．そこで本研究では，この音声操作への抵抗感を低減することを目指す．

人がエージェントへ話しかける際に抵抗感が生じるのは，エージェントの社会的存在感が低く人から社会的に接してもらいにくい[3][4]ことが起因していると考えた．社会的存在感とは人がエージェントから人間のような性質を感じ取る度合いであり，ヒューマンエージェントインタラクションの分野において社会的存在感を強化する手法はこれまでにも研究されてきた[5][6][7][8][9][10]．その基本的な手法はエージェントの外見や振る舞いを人間らしくすることである．しかしながら，音声対話エージェントはユーザが運転中や料理中等の作業中であっても支援できることが利点であり，そのようなエージェントに視覚的な工夫を適用することは必ずしも適切とは言えない．

そこで，我々は音声対話エージェントの人間らしさを向上させる手法として，聴覚情報による感情表現に着目した．人間は表情や話し方等によって自分の感情を表出したり相手の感情を認知しつつ対話を行うが，既存の多くの音声アシスタントの発話は人の音声と比べて平坦な合成音が用いられており，感情表現が乏しい状態である．そのため，エージェントの感情や意図が伝わりにくく，そのことが人間らしさや AI スピーカへ話しかけやすさを低下させる一因となっていることが考えられる．合成音声の韻律を調整して感情を表現する技術（以降，感情的な合成音声）も存在するが，我々は Mehrabian の実験結果[11]から感情的な合成音声のみではエージェントの感情表現としては不十分であると考える．Mehrabian によると，他者の感情を判断する上での影響は，視覚情報である表情は 55%であるのに対し，聴覚情報である声色は 38%，言語情報である発話内容は 7%であった．この知見から，感情の判断に最も影響する視覚情報がない場合には，平坦な合成音声はもちろん，声色を人間と同レベルの精度で再現できたとしても音声対話エージェントの感情表現として不十分である可能性がある．したがって，聴覚情報のみを用いて，視覚的な感情表現の欠落と聴覚的

な感情表現の不十分さを補う新たな感情表現の手法が必要である．

本研究では，音声対話エージェントの感情意図と適切に伝達するため，感情に対応する音楽等（以降，感情音）を合成音声に付与することで感情を強調して提示する手法を提案する．本研究ではこの手法を強調的感情表現と呼ぶ．強調的感情表現は，アニメやドラマ等に登場する架空の人物の感情表現から着想を得たものである．アニメやドラマの登場人物は実世界の人間が行っているより大袈裟に，すなわち強調的に感情表現を行っており，併せて Background Music（BGM）や，Sound Effect（SE）によっても感情が強調されているが，視聴者はそのような人間は行わないような強調された感情表現をむしろ人間らしく自然であると感じている．このことから，エージェントのように人間でないものが感情を表現する場合には，人間と同じ表現をするのでは不十分であり，感情を強調提示することが必要であると考えた．本研究では，強調的感情表現によって平坦な合成音声であっても適切に感情を伝達し，人間らしさを向上させることができるか調査するための実験を行った．

2. 実験

実験には，20 代以上の男女 120 名（主に 20～60 代で，120 名のうち男性 90 名，女性 30 名）が参加した．

2.1 タスク

実験は，AI スピーカが実験参加者に対してメールを受信したことを通知する場面を想定して行った．実験参加者は合成音声のみの通知音声と強調的感情表現を行う通知音声の 2 種類を聴取し，その後それぞれについて印象評価を行った．なお，実験参加者には AI スピーカに搭載された AI がメールの内容に応じて通知を行っていると教示した．また，通知音声の再生順序を実験参加者ごとに交互に入れ替えることでカウンターバランスを取った．

実験は図 1 に示す環境で行った．本来であればスマートスピーカーから直接音声が流れることが望ましいが，イベント会場の一角で行ったため雑音への対策として，実験参加者はヘッドフォンをつけて通知音声を聴取した．

2.2 使用した感情音

本実験では，感情音として音楽（BGM／SE）と擬態語を使用した．古くから音楽は感情の言語であると述べられており [12]，音楽と感情が密接な関係にあることは直観的にも理解されるため使用した．擬態語については，SNS でメッセージの発信者が感情を表す際に顔文字やスタンプと併せて用いられているほか，漫画でキャラクタの感情を表す際に付与されていることから使用した．

図 1 : 実験環境

各感情音の付与方法の詳細について述べる．BGM を用いる場合には合成音声の開始直前に付与し，以降合成音声と重複させて再生し続け，合成音声の再生終了後にフェードアウトさせた．また，SE を用いる場合には合成音声の開始直前に付与し，合成音声と重複しないようにした．さらに，擬態語を用いる場合には合成音声の末尾に 150 ミリ秒の間ののち，擬態語を付与した．

実験では，ポジティブ／ネガティブの両方の感情表現について調査した．BGM や SE についてはフリー音楽素材サイトよりダウンロードしたものを使用し，それらが意図した感情（ポジティブ／ネガティブ）を表現できているかを事前に複数人に確認した．擬態語については，ポジティブな感情の表現には「にこにこ」を，ネガティブな感情の表現には「しょぼん」を使用した．

2.3 実験条件

BGM／SE／擬態語のそれぞれにおける実験要因は次の 2 要因であり，実験条件は 4 条件となる．

強調要因（参加者内計画）：強調的感情表現の有無による違いを検証するための要因．なし（統制条件）／あり（提案手法）の 2 水準が存在する．

合成音声要因（参加者間計画）：合成音声の種類による要因．合成音声が平坦な場合と感情的な場合の両方で有効であるかを調べるため設定した．平坦／感情の 2 水準が存在する．

たとえば，一人の実験参加者が聴取する 2 種類の通知音声は，「1 回目：平坦な合成音声のみ，2 回目：平坦な合成音声＋ポジティブな BGM を用いた強調的感情表現」のようになる．

2.4 印象評価アンケート

2 種類の通知音声の聴取後に行った印象評価アンケートは下記の通りである．いずれも 7 段階の SD 法による評価とし，1 回目，2 回目のそれぞれの音声について回答してもらった．

1. どのようなメールが届いたと感じましたか？
ネガティブなメール ― ポジティブなメール
2. 通知をした AI についてどのように感じましたか？
2-1. 機械的 ― 人間らしい
2-2. 話しかけにくい ― 話しかけやすい

3. 結果

印象評価アンケートは, 2.4 節に記載した形容詞対のうち左側を 1，右側を 7 として点数化した．このデータに対し，感情音ごとに 2 要因分散分析を行った．

3.1 ポジティブな感情の表現

まず，図 2 は感情音として BGM を用いた場合の結果である．メールの内容（$F(1, 18) = 100.660$, $p < .001$），人間らしさ（$F(1, 18) = 39.710$, $p < .001$），話しかけやすさ（$F(1, 18) = 25.462$, $p < .001$）の全項目において，強調要因の主効果が見られた．メールの内容については，合成音声要因の主効果も有意傾向であった（$F(1, 18) = 3.176$, $p = .092$）．したがって，感情音として BGM を用いれば，合成音声の種類を問わず，感情の伝わりやすさ，人間らしさ，話しかけやすさを向上させられることがわかった．

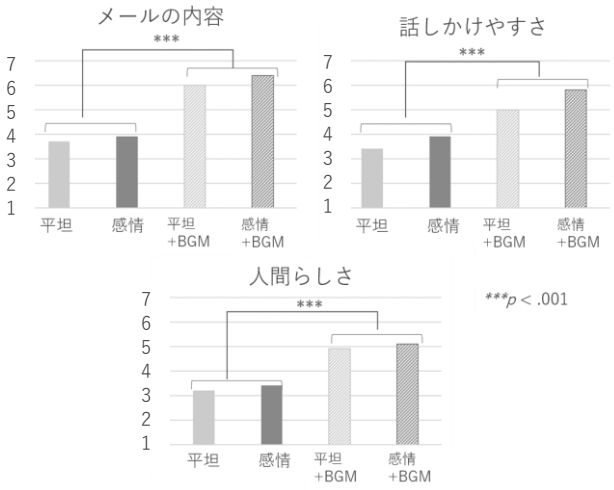

図 2：BGM を使用した場合の印象評価結果

次に, 感情音として SE を用いた場合の結果を図 3 に示す．メールの内容（$F(1, 18) = 59.648$, $p < .001$），話しかけやすさ（$F(1, 18) = 5.345$, $p < .01$）の項目において，強調要因の主効果が見られた．人間らしさについては交互作用が有意傾向であったため（$F(1, 18) = 3.927$, $p = .063$），TukeyHSD による下位検定を行ったところ，合成音声が平坦である場合のみ SE により有意に人間らしさが向上していた（$p < .05$）．したがって，感情音として SE を用いれば，合成音声の種類を問わず，感情の伝わりやすさ，話しかけやすさを向上させられることがわかった．また，合成音声が平坦な場合には SE を用いることで人間らしさを向上させることが可能であった．

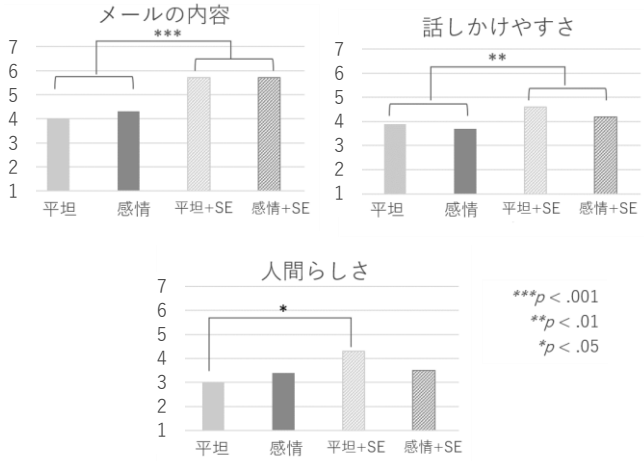

図 3：SE を使用した場合の印象評価結果

さらに，感情音として擬態語を用いた場合の結果を図 4 に示す．メールの内容においては強調要因の主効果が見られたが（$F(1, 18) = 10.055$, $p < .01$），そのほかには有意な結果は見られなかった．したがって，感情音として擬態語を用いると，合成音声の種類を問わず，感情の伝わりやすさを向上させられることがわかった．

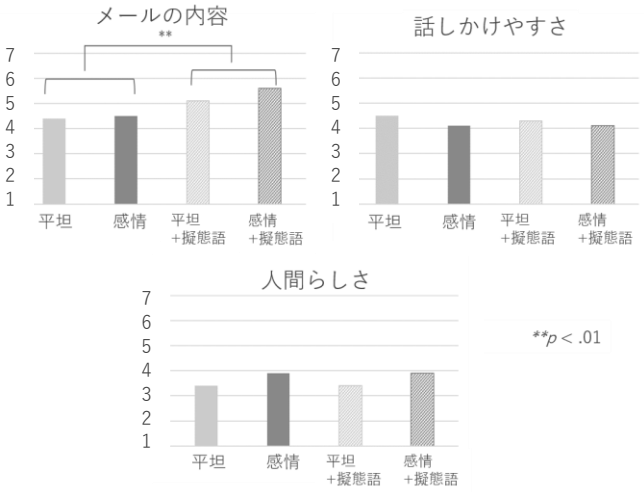

図 4：擬態語を使用した場合の印象評価結果

3.2 ネガティブな感情の表現

まず，感情音として BGM を用いた場合の結果について述べる．メールの内容においてのみ，強調要因の主効果と（$F(1, 18) = 13.926, p < .01$），合成音声要因の主効果が見られた（$F(1, 18) = 8.901, p < .01$）．したがって，感情音として BGM を用いれば，合成音声の種類を問わず感情の伝わりやすさを向上させられることがわかった．また，BGM の有無を問わず合成音声を感情的にした場合にも感情の伝わりやすさを向上させられることがわかった．

次に，感情音として SE を用いた場合の結果について述べる．メールの内容においてのみ交互作用が見られたため（$F(1, 18) = 8.727, p < .01$），TukeyHSD による下位検定を行ったところ，合成音声が平坦である場合と感情的である場合のいずれにおいても SE を付与することにより有意にネガティブな印象を伝達できていた（それぞれ $p < .001$，$p < .05$）．さらに，SE を付与しない場合には合成音声を感情的にした場合にも有意にネガティブな印象を伝達できていた（$p < .001$）．したがって，感情音として SE を用いれば，合成音声の種類を問わず感情の伝わりやすさを向上させられることがわかった．また，SE を付与しない場合には合成音声を感情的にすることでも感情の伝わりやすさが向上することがわかった．

さらに，感情音として擬態語を用いた場合の結果について述べる．メールの内容において強調要因の主効果が有意であり（$F(1, 18) = 81.000, p < .001$），合成音声要因の主効果が有意傾向であった（$F(1, 18) = 3.491, p = .078$）．また，話しかけやすさにおいても強調要因の主効果が見られた（$F(1, 18) = 4.455, p < .05$）．したがって，感情音として擬態語を用いれば，合成音声の種類を問わず，感情の伝わりやすさ，話しかけやすさを向上させられることがわかった．また，擬態語の有無を問わず，合成音声を感情的にした場合にも内容の伝わりやすさの向上には効果があることがわかった．

4. 考察

前章で得られた結果より，ポジティブ／ネガティブのいずれの感情を表現する場合であっても，感情音の種類を問わず強調的感情表現を用いることで，平坦な合成音声であっても明確に感情を伝達することが可能であることが明らかになった．さらに，感情的な合成音声と共に用いた場合にも，合成音声のみの場合と比較して明確に感情を伝達できることがわかった．

また，ポジティブな感情を表現する場合には，感情の伝わりやすさ，人間らしさ，話しかけやすさのすべての項目を向上させられた BGM を用いることが，社会的存在感の観点からも有効であるといえる．

一方，ネガティブな感情を表現した場合は，人間らしさや話しかけやすさを向上させる効果はほとんど見られなかった．アンケート後に行った簡単なインタビューでは，ネガティブな内容のメールの場合は通知までネガティブにしてほしくないという意見が多く聞かれたことから，メールの内容に応じた通知を行うというタスク設定がネガティブな感情表現と相性が悪く，話しかけやすさに悪影響を及ぼしたことが考えられる．唯一話しかけやすさを向上させることができた擬態語は，茶化しているような感じがしたという意見が得られており，ネガティブな内容であることを伝えつつも深刻さを感じさせないことで話しかけやすい印象を与えることができたと考えられる．したがって，ネガティブな感情を表現する際には用途に応じて適切な感情音を使い分けることが望ましく，タスク設定を変更することで新たな知見が得られると考えられる．

5. 今後の展望

5.1 感情音としての BGM の自動生成

今回の実験では BGM 等の感情音は実験者が人手であらかじめ選定したが，実用化に向けては表現したい感情に応じて適切な感情音を自動で生成するシステムを作成することが必要である．現在は Google が公開している Magenta[12]を利用して，指定した感情に対応する BGM を生成する取り組みに着手している．Magenta は生成したい楽曲の初めの数音等の楽曲の元となるシード情報を入力することで，それに対応した楽曲を出力するシステムである．出力される楽曲が意図した感情を表現できるようにするためには適切なシード情報を生成する必要があり，現段階ではシード情報の生成モデルを試験的に構築している．

5.2 応用的効果の検証

エージェントの感情表現に関する研究で，ロボットの表情による感情表現を目元と口元に分けてモデル化し，口元は社会的に微笑み（社会的表現），目元は感情に基づいて不快感示す（情動的表現）によって Non-DuchenneSmile と呼ばれる苦笑いのような表情を表出するロボットは，一方の表現しか行わないロボットと比較して，より人に一緒に暮らしたいと感じさせられることが報告されている．本研究で提

案した手法においてもこれと同様に，社会的な言葉を発する一方でネガティブな BGM を併せて提示することで Non-DuchenneSmile のような高次の感情表現が可能かもしれない．このように本研究の知見を応用した場合の効果を検証する実験を行うことも今後の課題である．

6. むすび

本研究では，音声対話エージェントに対しても適用できる社会的存在感の向上手法の検討を行った．具体的には，現在の音声対話エージェントの感情表現の希薄さに着目し，合成音声に対して音楽等の聴覚情報を付与して感情を強調的に表現する手法を提案した．

エージェントの感情表現を強調する聴覚情報として BGM，SE，擬態語を用い，それぞれを合成音声に付与した際にポジティブ／ネガティブの両感情が伝わりやすくなるかを調べた結果，どの聴覚情報を用いてもポジティブ／ネガティブの両感情ともに意図通りに伝達することができた．また，ポジティブな感情の表現に BGM や SE といった音楽を用いた場合，エージェントの人間らしさを向上させることができ，さらに，エージェントに対して話しかけやすい印象を与えることができた．これらの効果は特に BGM を用いた場合において強く表れた．以上より，BGM を用いて感情表現を音楽的に強調することで，音声対話エージェントの社会的存在感を強化することができたといえる．

謝辞

JSPS 科研費 JP18H05076，JP19K12081 からの支援を受けた．

参考文献

[1] “日本人の音声操作に対する意識調査 2017”，http://news.kddi.com/kddi/corporate/newsrelease/2017/10/05/2726.html, (最終検索日：2020 年 1 月 28 日)

[2] “スマートスピーカーの日本における利用実態についてインターネット調査”，dentsudigital.co.jp/release/2019/0218-000164/, (最終検索日：2020 年 1 月 28 日)

[3] M.K. Lee, U. States, S. Kiesler, J. Forlizzi, and P. Rybski.: Ripple Effects of an Embedded Social Agent: A Field Study of a Social Robot in the Workplace, Proc.CHI2012, pp.695–704, (2012).

[4] K. Tanaka, N. Yamada, H. Nakanishi, and H. Ishiguro.: Teleoperated or Autonomous?: How to Produce a Robot Operator’s Pseudo Presence in HRI, Proc. HRI2016, pp.133-140, (2016).

[5] N. Yee, J.N. Bailenson, and K. Rickertsen.: A Meta-analysis of the Impact of the Inclusion and Realism of Human-like Faces on User Experiences in Interfaces, Proc.CHI2007, pp.1–10, (2007).

[6] A. Zaraki, D. Mazzei, M. Giuliani, and D.D. Rossi.: Designing and Evaluating a Social Gaze-control System for a Humanoid Robot, IEEE Transactions on Human-Machine Systems, vol.44, no.2, pp.157–168, (2014).

[7] N. Koyama, K. Tanaka, K. Ogawa, and H. Ishiguro.: Emotional or Social?: How to Enhance Human-robot Social Bonding, Proc. HAI2017, pp.203–211, (2017).

[8] C. Breazeal.: Emotion and Sociable Humanoid Robots, International Journal of Human-Computer Studies, vol.59, no.1-2, pp.119–155, (2003).

[9] D. Cameron，S. Fernando，E. Collins，A. Millings，R. Moore，A. Sharkey，V. Evers, and T. Prescott.: Presence of Life-like Robot Expressions Influences Children’s Enjoyment of Human-robot Interactions in the Field, Proc. AISB2015，pp.36‑41, (2015).

[10] A. Pereira, R. Prada, and A. Paiva.: Improving Social Presence in Human-agent Interaction, Proc. CHI2014, pp.1449–1458, (2014).

[11] A. Mehrabian.: Nonverbal Communication, Aldine Transaction, (1972).

[12] D. Cooke.: The language of music, Oxford University Press, (1959).

[13] “Magenta”，https://github.com/tensorflow/magenta, (最終検索日：2020 年 1 月 28 日)