適応型効果音提示システムの評価

圓尾俊樹 [1,a)]　原野響 [2,b)]　大澤博隆 [1,c)]

概要：YouTube や TikTok のような動画コンテンツが，世界中で普及している．また，効果音には，会話にメリハリを生み出すことで，飽きを和らげる効果があると考えられる．以上のことから，これらのコンテンツで利用されている効果音を，実際の会話に応用することで，会話の質を向上させることができるのではないかと考えた．本研究では，会話の質を向上させることを目的として，話者の感情を推定することによって，適切な効果音を提示するシステムを提案する．評価実験の結果，本システムは会話を妨げることなく，会話の深まりに貢献することが，明らかになった．また，会話と効果音の提示にラグが発生してしまうことによって，本システム本来の効果を最大限に活用することができないという問題点も明らかになった．

1. はじめに

近年では, YouTube や TikTok といった動画コンテンツが, 世界的に普及している．これらの動画コンテンツを魅力的にしている要因の 1 つとして，内容や状況に応じて適切に流れる効果音が挙げられる．強調すべきであると投稿者が考えた場所に効果音が付されていることによって，動画にメリハリが生まれて，動画が長かったとしても，飽きずに楽しむことができる. そこで, 効果音を付すことによって，飽きずに楽しむことができるという効果は，YouTube や TikTok のような編集されたコンテンツに留まらず，実際の会話にも応用できるのではないかと考えた．

発話に対し，効果音を自動的に追加する研究としては，例えばラジオの音声に対し，効果音を自動的に鳴らす方法の提案がある［1］．しかしながら，これはあらかじめ放送形式が決まったラジオ上の音声であり，リアルタイムの処理ではない．多くの人が，日常生活を行ううえで頻繁に行う，議論や雑談などを行う場面における効果音をリアルタイムで鳴らし，飽きを減らすという効用を得る内容の研究はあまりない．

本研究においては,「議論の内容に応じて，議論の雰囲気に適切に対応しているような効果音を鳴らすシステム」を提案し，その影響を評価する．近年，生成 AI の進歩により，リアルタイムで人の音声を認識して，その内容に基づいた感情を推定することが可能になっている．生成 AI による感情推定の結果に基づいて, 効果音を提示することで，会話のメリハリを生み出して，会話の質を向上させることを考える．具体的には，ユーザに議論を実施してもらい，会話内容や文脈に合わせて適切な効果音を生成 AI で判断し，リアルタイムで鳴らすシステムを構築する．

論文の構成を示す．2 章は本研究と関連する研究の紹介であり, 3 章では, 提案した手法について詳細に説明する．4 章では，本システムの評価手法を示したうえで, 5 章では評価結果, 6 章で評価結果に基づいた考察を示す. 最後に，7 章では, 研究のまとめと今後の展望について, 記述する．

2. 関連研究

本研究では，話者の感情推定と感情表出の両方が必要である．基礎となる感情推定技術として，BERT のような小規模な言語モデルや大規模言語モデルを使用して感情推定を実施した研究及び Plutchik の感情の輪という感情分類手法について記述する．現時点で，言語モデルによる感情推定は実施されているが，感情推定をどのようにして，人の生活を豊かにするために利用するのかについての議論は，あまり行われていない．

2.1 言語モデルによる感情推定

近年, 言語モデルを用いた感情分析の研究がある［2］［3］．以前まで，感情分析や感情推定における研究は，感情辞書を用いるのがほとんどであった．しかし，Nakazawa らは，Pre-training と Fine-tuning を行った BERT を用いた感情極性の推定方法を提案し，従来は自然言語処理の分野において代表的な文章解析の手法であった，形態素解析と構文解析という 2 つの手法と比較することによって，言語モデルを用いた感情推定法が, 従来の手法より優れることを示した．

Xue らの研究は, 大規模言語モデル（以下, LLM とする）を用いて 48 種類の緻密な感情推定の手法を提案している．本研究は, Plutchik の感情の輪に基づく感情空間を用いて，感情推定を指示するプロンプト文（指示文）と分析対象文を LLM に入力することにより, 感情推定結果を得ている．ここで, 本研究では, WRIME データセットを用いている．WRIME データセットとは, 計 43200 件の各投稿に対して，投稿主の感情を値として付与しているデータセットである．LLM の感情推定結果と WRIME データセットを比較して．LLM の性能比較を実施して，その有効性を検証している．

1. 慶應義塾大学理工学部管理工学科
2. 慶應義塾大学大学院理工学研究科 開放環境科学専攻
a. t-maruo2713@keio.jp
b. hibiki23@keio.jp
c. hailabsec@gmail.com

2.2 Plutchik の感情の輪

Plutchik の感情の輪とは，R. Plutchik によって提案された感情モデルである. Plutchik の感情の輪は, 8 個の基本感情，各基本感情に対応する 16 個の強弱派生感情，及び 8 個の応用感情から構成されている. 各基本感情は, joy と sadness のように対極の感情が配置されている. 各強弱派生感情は，弱い感情は基本感情の外側，強い感情は基本感情の内側に配置されている．また，隣り合う二つの基本感情によって成立している応用感情は，基本感情の間に配置されている［4］. Plutchik の感情の輪と感情語類似度ネットワークの構造を比較することによって，感情の輪の構造的妥当性を検討した研究もある［5］. 両者は，一部的に構造の違いが見られたものの，全体的に構造的特徴は近いと考えられるという結果が得られており, 妥当であると考えられている. 本研究に用いる感情は，Plutchik の感情の輪で定義される感情のなかから，「喜び」,「驚き」,「怒り」及び「悲しみ」という 4 つの感情に限定した．感情を限定した理由は，1) Plutchik の感情の輪にある全ての基本感情と効果音を対応させることが困難であったこと, 2) 本システムの有用性を検証するにあたり,「喜び」のような明るい感情と「悲しみ」のような暗い感情に分けた結果,「驚き」と「怒り」以外は，同じ効果音にしてしまっても問題ないと考えたためである.

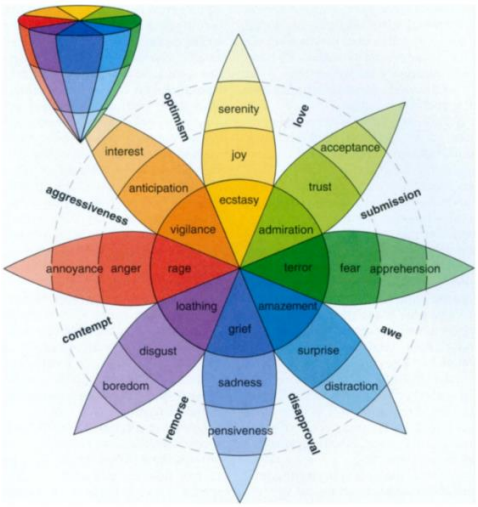

図 1　Plutchik の感情の輪

3. 提案手法

3.1 システムの設計

図 2 に示すように, 20 秒ごとに取得される音声データを文字に起こした．次に，文字に起こしたテキストデータを LLM に送信して，感情を推定した．そして，LLM からの返答内容に応じて，効果音を実際に鳴らした.

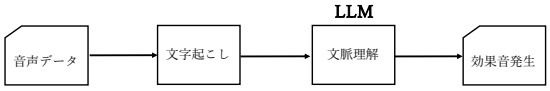

図 2　システム構成図

3.2 システムの実装

本実験においては, 20 秒ごとに取得される発話データを OpenAI 社の Whisper を使用して文字に起こした. 本研究において，OpenAI 社の Whisper を用いた理由は，文字起こしの精度と高速性を勘案した結果，本モデルが最適であると考えたためである. その後に, ローカル PC で動作する LLM, Gemma-2-2B-JPN を用いて，文脈を理解させた．本研究において, Gemma-2-2B-JPN を用いた理由は, 安全性や費用，及び効率を勘案した結果，本モデルが最適であると考えたためである．会話の文脈を理解するための尺度としては，「感情」を用いた．具体的には，LLM に対して，「Use the word “joy” if the conversation is positive, “surprise” if the conversation is unusual, “sadness” if the conversation is negative, and “anger” if the conversation is angry. Each emotion category should be determined according to the context of the conversation. Be sure to output only words.」と指示した．前述した指示に応じて LLM が出力した単語に基づいて，事前に用意した各感情に対応すると考えられる効果音を，Wizard of Oz 法（WOZ 法）によって, 効果音を MP3 ファイルで再生した.

3.3 効果音の選定

効果音は，実験参加者がどの感情を表現しているのかを判断することが可能なように選定する必要がある．また，効果音が利用されている場面を考えると, YouTube がある．具体的には, YouTube の企画で, 人が会話しているなかで，会話の内容を強調させる目的で効果音が利用されている．そのため，YouTube において，前述した 4 つの感情を表現する際に, しばしば用いられる効果音を選定し, 使用した．効果音は,「効果音ラボ」というサイトから選定した［6］.

4. システムの評価

4.1 参加者

参加者は、次の条件に基づいて募集した．: 1) 日本語をネイティブレベルで理解できること，2) 身体・精神が共に健常であること, 3) Google Forms に回答できることである．その結果，20 名が研究に興味を示して，実験に参加した．実験参加者の年齢は，21 歳から 24 歳（平均年齢＝23 歳，標準偏差＝0.82）で，男性が 12 名，女性が 8 名であった.

4.2 実験内容

実験参加者には, 二人一組で, 会話を実施してもらった．本研究においては，システムの有効性を検証するために，

2 水準の条件下で，実験を実施した（4.3 節で詳述する）．各条件で 5 分間ずつの会話を実施してもらい，20 秒おきに，効果音を，スピーカーから鳴らした．本研究においては，題目を「今年の夏休みの思い出」に指定した．これにより，実験グループ間の題目による影響を排除することができる．また，感情に基づいて効果音を鳴らすため，様々な感情が会話に出てくる題目が望ましいと考え，本題目に設定した．実験時における機材と実験者の配置を，以下の図 3 に示す．

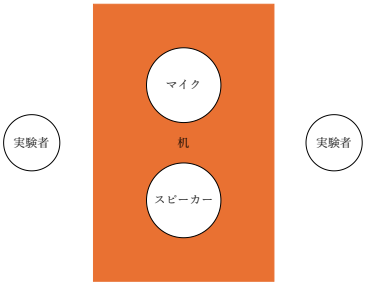

図 3　実験実施時の配置

4.3 評価方法

本システムを評価するために，アンケートを実施した．前節において設定した題目について，1) 効果音がランダムで変化，2) 効果音が感情に沿って変化，の二条件で実験を実施して，システムに関する評価をアンケートで集めた．アンケートは，各実験終了時にそれぞれ回答してもらった．実験者に応じて，(1)と(2)の実施順序をランダム化したため，各システム実験の実施順序による影響はないものとする．具体的なアンケート内容を，表 1 に示す．本アンケートは，実験者における，システムの受容性及びユーザーエクスペリエンス（UX）の評価を行うために設定した［7］［8］［9］．

表 1　本実験で実施したアンケート内容

Q1	効果音によって雑談や議論がより楽しくなりましたか
Q2	このシステムでもっと話したいと思いましたか
Q3	効果音はあなたの感情を動かしましたか
Q4	効果音は会話の深まりに貢献したと思いますか
Q5	将来このシステムを使用したいと思いますか
Q6	このシステムを他の会話や状況でも使いたいと思いますか
Q7	このシステムによって会話が妨げられていると感じましたか
Q8	効果音は聴覚的に魅力的でしたか
Q9	効果音は会話内容に適切に対応していると感じましたか

研究の仮説は，「会話の文脈に沿った効果音を提示するシステムに従って，効果音を鳴らすことによって，議論や雑談の質を高めることが可能である」というものである．本仮説に基づくと，Q1-Q6，Q8 及び Q9 に関しては，提案システム条件の方が，ランダム条件よりも点数が高くなると考えられるが，Q7 に関しては，提案システム条件の方が，ランダム条件よりも点数が低くなると考えられる．

5. 結果

5.1 アンケート結果

アンケートに基づいて算出した平均点を，図 4 に示す．図 4 において，* p 値＜0.05，** p 値＜0.01 を表している。

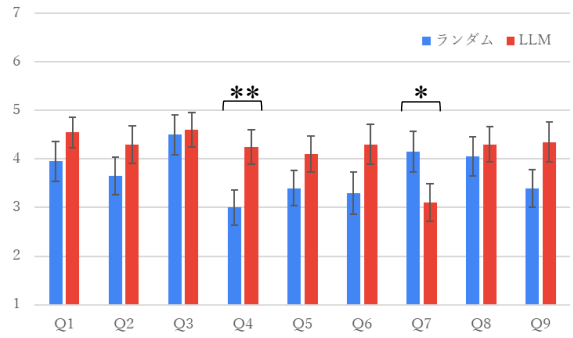

図 4　アンケート結果

5.2 検定結果

アンケート結果に基づいて，対応のある t 検定を行った．ここで，本検定を実施するにあたり，正規性を仮定した．本実験においては，2 水準の条件下で実施した実験間での有意差が認められるかを検証するために，本手法を用いた．

まず，以下のアンケート項目に，有意差が認められた．

1) Q4：p 値＝0.0097＜0.01，t＝−2.88
2) Q7：p 値＝0.0367＜0.05，t＝2.25

また，以下のアンケート項目に，有意傾向が認められた．

1) Q2：p 値＝0.091＜0.1，t＝ −1.78
2) Q5：p 値＝0.069＜0.1，t＝ −1.93
3) Q9：p 値＝0.067＜0.1，t＝ −1.94

6. 考察

5 章において，Q4 及び Q7 に関する有意差が認められた．

提案システム条件の方が，会話の深まりに貢献した理由としては，効果音により，話者の感情が肯定されることで，会話に対する積極性が向上したことにあると考えられる．実際，「嬉しい話題のときに、明るい効果音が鳴ると嬉しい」という意見や「喜びの効果音のときは、会話がさらに盛り上がる気がして良かったです」という意見があった．一方

で，ランダム条件下においては，自分の感情と異なる，または対照的な効果音が鳴る場合がある．実際,「話の内容に沿ってないと，効果音は邪魔に感じた」という意見があった．効果音が会話実施時に邪魔に感じてしまうと，会話に対して話者が消極的になり，会話の深まりが得られなかったのではないかと考えられる．

次に，提案システム条件の方が，会話が妨げられているように感じられなかった理由としては，効果音がより会話に馴染んだことが考えられる.実際,「ランダム条件の方が，効果音が鳴る頻度が多く感じた」という意見が多数あった．これは，効果音が会話に馴染まなかったために，会話から浮いていたことが原因であると考えられる．提案システム条件が優れているのではなく，ランダム条件時の効果音が妨害していただけの可能性があるが，日常生活において，人間は，自分とは関係ない騒音の下でも，自分と関係ある話題だけを選択することが可能である．これを,「カクテルパーティー効果」という．議論や雑談を行う場面においては，雑音があることが想定されるため，それよりも会話を妨げることがないのであれば，本システムは有用であると考えられる．

また，Q9 に関しては，有意傾向が認められた．提案システムにおいては，会話内容から感情を推定しているため，「適切に対応していると感じましたか」に有意差が認められないことは問題のように感じられる．有意差が認められなかった原因としては，効果音の種類が少なかったことが考えられる.効果音を 4 種類しか用意していなかったため，25%の確率で適切な効果音が提示されてしまう．また，実験を行うにあたって，最適なタイミングで 1 つの効果音が鳴ると実験参加者の記憶に残るという傾向が見られた．ランダム条件の下において最適なタイミングで，感情に応じた効果音が鳴ったグループが複数あったため，それが記憶に残ってしまったことが，有意差が認められなかった一因として考えられる．

また,提案システムの問題点も浮かび上がった.それは，音声認識から効果音の提示までにラグがあることである．本実験においては，20 秒に 1 回の頻度で効果音を鳴らすため，会話と効果音にラグが生じてしまっていた．これは，Q9 に関して，有意差が認められなかった要因の 1 つとも考えられる．実際，会話の話題が頻繁に変化したグループでは，2 人ともランダム条件下の方が，適切に対応していると回答していた．その他の多くのグループにおいても，ラグを感じたとの意見は，多く見られた．これは，議論や雑談の質を向上させるためには，改善すべきであると考えられる．具体的には，録音の秒数を 20 秒よりも短く設定することが挙げられる．これによって，本実験よりも直近の会話に適切に対応した効果音を提示することができると考えられる．

また，効果音を鳴らす方法にも問題があったと考えられる．実際,「効果音自体が，なんの感情に対応しているのかが，あまり分からなかった」という意見があった．本研究では，スピーカーから効果音を鳴らしたが，音質がわるくなってしまい,「怒り」と「悲しみ」の効果音が雑音のように感じられる場合があった.機材を変更することによって，本問題点は改善することができると考えられる．

本研究を実施したことによって，多くの問題点が明らかになったが，会話内容から，話者の感情を推定して，感情に合わせた効果音を提示するシステムに従って，効果音を鳴らすことによって，会話を妨げることなく，会話の深まりに貢献することが可能であると考えられる.少なくとも，Q2 及び Q5 において，有意傾向が認められているため，実験参加者には，提案システムが好印象だったと言える．

7. まとめと今後の展望

本研究では，会話の質を向上させることを目的として，話者の感情を推定して，適切な効果音を提示するシステムを提案した．評価実験を実施した結果，会話を妨げることなく,会話の深まりに貢献することが分かった.その一方，会話から効果音の提示までにラグが生じてしまった結果，会話と対応していない効果音を提示してしまう問題点や，スピーカーの性能により，効果音が雑音に聞こえてしまうという問題点が明らかになった．より良いシステムを作成するために,以上の点を改善すべきであることが分かった．

本研究では，二人での会話の場面に注目した．今後は，会話の場面に留まらず，演説のような場面でも活用できるかどうかの検証を実施し,本システムの可能性を模索する．

謝辞 本研究の実験に参加していただいた 20 名の皆様に心よりお礼申し上げます．本研究は，JST ムーンショット型研究開発事業「身体的共創を生み出すサイバネティック・アバター技術と社会基盤の開発」(Grant number JPMJM2013)および JSPS 科研費「ポストヒューマン社会のための想像学」の助成を受けたものです．

参考文献

［1］ Songwei Ge, Curtis Xuan, Ruihua Song, Chao zou, Wei Liu, and Jin Zhou. From Text to Sound: A Preliminary Study on Retrieving Sound Effects to Radio Stories. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.865-868, 2019.
［2］ 中澤政孝，亀井且有，前田陽一郎，クーパーエリック．BERT を用いた単文の感情極性推定手法の提案とその有効性．FSS2020, TA2-3, 2020.
［3］ 薛沛欽，小林亜樹．大規模言語モデルによる細かな感情推定手法. DEIM 2024, T1-B-8-03, 2024.
［4］ Robert Plutchik. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist, Vol. 89, No.4, pp.340-350, 2001.
［5］ 岩城史享，横須賀天臣，高橋 達二．感情語類似度ネットワ

ークの遍在的コミュニティ抽出による解析. The 37th Annual Conference of the Japanese Society for Artificial Intelligence, 3Xin4-11, 2023
［6］ “効果音ラボ”. https://soundeffect-lab.info/, (参照 2024-11-08).
［7］ Venkatesh, Viswanath, James YL Thong, and Xin Xu. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS quarterly, pp.157-178, 2012.
［8］ Fred D. Davis and Viswanath Venkatesh. Toward preprototype user acceptance testing of new information systems: implications for software project management. IEEE Transactions on Engineering management, Vol. 51, No. 1, pp.31-46, 2004.
［9］ Effie L-C. Law, Virpi Roto, Marc Hassenzahl, Arnold P.O.S. Vermeeren, Joke Kort. Understanding, Scoping and Defining User eXperience: A Survey Approach. CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp.719-728, 2009