ノンバーバルな話者感情表出を考慮した統計的対話システムとその拡張

Statistical Dialogue System Considering Nonverbal Speaker Emotion Expression and Its Extension

海原　颯馬　　溝口　和輝
KAIHARA Soma　　MIZOGUTI Kazuki
広島市立大学大学院　言語音声メディア工学研究室
Language and Sound Media Laboratory, Hiroshima City University

概要　本論文では非言語的な表出感情を絵文字として発話文字列に付与した中間表現を作成し，その中間表現を統計的応答手法への入力として応答を生成する非タスク指向型音声対話システムを提案し，印象評価を行う．さらに感情推定部分の精度を表情を用いて向上させる手法と，エージェント自身に感情を持たせる手法についての考察を行う．

1　はじめに

現在，自然言語対話システムは様々な分野で普及しつつあり，雑談のような非タスク指向型の対話についても研究が進められている．しかし従来の音声対話システムはノンバーバル情報を考慮できないため，言葉に表れない話者感情の違いに対応することができない．例えば図 1 のように疲れた様子で「おはよう」と発言しても，元気そうな様子で「おはよう」と発言しても従来の方法だとすべて同じ返答になってしまう．

本研究では，ユーザの非言語的な感情表出を考慮して応答できる非タスク指向型音声対話システムを実現する．しかし，ルールベース方式による応答発話生成手法では 1 パターンの入力発話文字列に対してユーザ感情の種類の分だけ応答ルールのバリエーションを追加しなくてはならないため，想定されていなかった入力発話に対しても頑健に応答を行うことができる統計的応答手法を用いてユーザの非言語的な感情表出を考慮した応答を行う手法を提案する．しかし統計的応答手法は文字列を対象とした手法であることから，非言語的な表出感情をそのまま扱うことができない．そこで，非言語的な表出感情を絵文字として発話文字列に付与した中間表現を発話音声から作成し，その中間表現を統計的応答手法への入力とする．その際，表出感情を絵文字で表現することにより，統計的応答手法に必要となる学習用応答ペアを Twitter やチャットログなどから容易かつ大量に収集することができる．

2　話者感情を考慮した統計的対話システム

本研究のシステムの流れを図 2 に示す．本システムは大きく“入力音声を文字列と絵文字に変換する”入力音声変換部，“入力音声の情報から絵文字付きテキストの応答を返答する”応答発話生成部，“生成された応答テキストを感情がこもった音声合成へ変換する”音声発話生成部からなる．

入力音声変換部では，入力音声の音響的特徴から推定した感情を絵文字に変換し発話文字列とともに応答生成部へ入力する．応答発話生成部では入力音声変換部で生成された中間表現を絵文字付きの応答ペアを学習した統計的応答システムに入力することで、応答発話テキストを生成する．音声発話生成部では生成した応答発話テキストを絵文字で表された感情を考慮した合成音声に変換する．各部の詳細については次節以降で説明する．

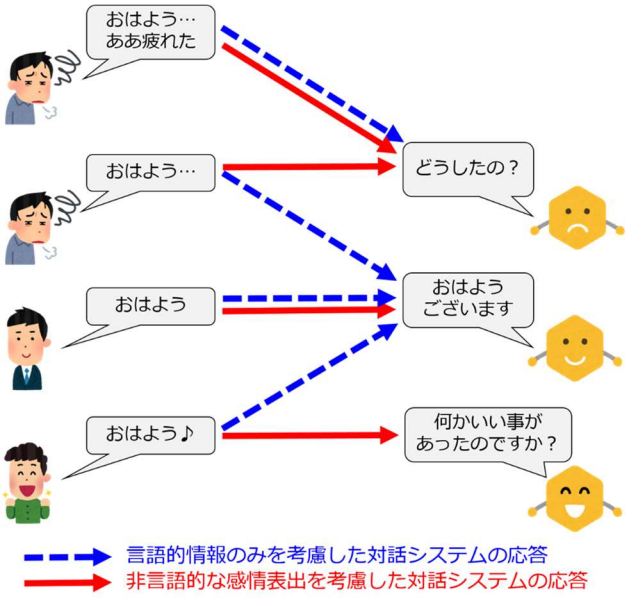

図1　非言語的な感情表出を考慮した応答の変化

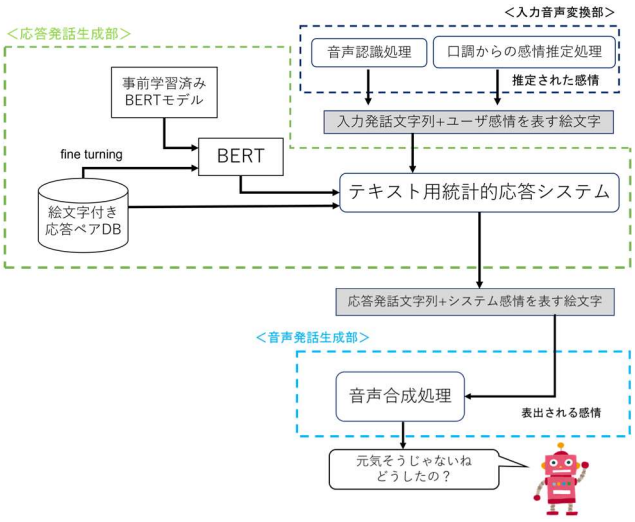

図2　システムの流れ

表 1　推定されたユーザ感情と絵文字の対応

口調から推定されたユーザ感情	絵文字
怒り	💢
嫌い，軽蔑	😑
不安・緊張，恐れ	😱
狂喜・楽しい，余裕・嬉しい，リラックス・気楽	😍
憂鬱・悲しい，眠い・疲れた	😢
冷静	無し

2.1　入力音声変換部

入力音声変換部では音声認識処理によって入力された発話を文字列として取得すると同時にユーザ感情を推定し応答発話生成部へ入力される．推定する手法として機械学習を使った音声からの感情推定手法を用いる [2]．その際推定した感情を絵文字で表し，入力発話の文字列の末尾へ付与する．

ユーザ感情の推定感情クラスは 6 種類（怒り，嫌悪，恐れ，幸福，悲しみ，驚き）である．これは Ekman の基本 6 感情モデルを参考にしている．推定されたユーザ感情と絵文字の対応を表 1 に示す．

2.2　応答発話生成部

応答発話生成部では，非言語的な感情表出を絵文字として表現した応答ペアデータを学習させた統計的応答手法によって応答発話を生成する．

絵文字付きテキスト用の応答発話生成システムは Python でつくる対話システム[3]を参考にして用例ベース方式で構築する．用例ベース対話システムは，全文検索エンジン Elasticsearch と応答選択用の自然言語処理モデル BERT（Bidirectional Encoder Representations from Transformers）[4]を組み合わせることで実装する．

実行手順を以下に示す．まず Elasticsearch の検索機能を使って，絵文字付き応答ペアデータベースから「入力発話テキストと一致度が高いクエリを持つ応答ペア」を上位から 100 件取り出す．次に入力発話テキストと上記で取得した応答ペアのレスポンスの受け答えの適切さについて，fine-tuning 済みの BERT を用いて適切な受け答えである確率を計算する．Fune-turning には Google が公開している事前学習済み BERT モデル（bert-base-multilingual-cased[5]）を用いる．これを応答ペアデータに含まれる発話をランダムに組み合わせて負例となるペアデータを作成し，二値分類を行う．上記から得た結果から適切な受け答えである確率が最も高いレスポンスをシステムの応答発話テキストとして出力する．

統計的応答手法で絵文字を考慮した応答を行うためには絵文字付きの発話応答ペアを大量に収集する必要があるため tweet-reply ペアを対話システムに用いる．集めたデータから公序良俗に反する tweet を削除し絵文字付き応答ペア 1,929,792 件，絵文字無し応答ペア 1,376,644 件，合計で 3,306,436 件となった．

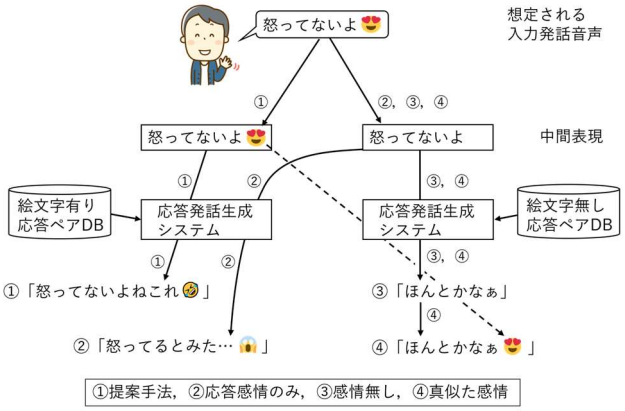

図 3　提案手法と比較手法による応答発話生成手順

2.3　音声発話生成部

音声発話生成部では，応答発話生成部で取得した応答発話テキストを音声合成処理により音声へ変換し出力する．音声合成処理には日本語音声合成エンジン OpenJTalk[4]の中の感情モデル 5 種類中 4 種類を用いる（normal，angry，happy，sad）．

なお，収集した tweet や reply には異なる感情を表す絵文字が同時に含まれていることがあるため，先に現れた絵文字を優先して音声合成時の感情モデルを選択している．

3　評価実験

本節では，提案手法である「クエリとレスポンスの両方に絵文字を含む応答ペアを用いた統計的応答手法」と「ユーザ感情と同じ感情をまねて表出する手法」との比較実験について述べる．

3.1　実験の構成

評価実験用に準備した入力発話文字列に喜び，怒り，悲しみ，嫌悪の絵文字を付与した入力発話テキストを作成し，それらに対して提案手法と比較手法が出力した応答発話テキストについて，「応答発話において文言に合った感情表出ができているか」および「好感が持てる応答をしているか」の観点から Scheffé の一対比較法（中屋の変法）[10, 11]を用いて主観評価実験を行う．

応答発話生成手順を図 3 に示す．本稿では紙面の都合上，①提案手法と④真似た手法の比較実験についてのみ紹介する．①の提案手法による応答は，2 節で構築した音声対話システムの応答発話生成部を用いて作成する．④の「ユーザ情報と同じ感情を真似て表出する応答生成手法」による応答は③の感情無し応答生成システムの応答テキストの後にユーザ感情を表す絵文字と同じ絵文字を付与することで生成する．

実験用入力発話文字列は，「食べ過ぎた」，「怒ってないよ」，「気にしてないよ」，「ありがとうございます」，「そうなんですか」の 5 つの発話文字列に喜び，怒り，悲しみ，嫌悪の 4 感情を表す絵文字を付与したものを用いた．

実験参加者は 19 歳から 23 歳の大学生および大学院生 11 名（男性 9 名，女性 2 名）である．実験参加者は，入力発

話テキストおよび各手法による応答発話テキストを提示され，前述の 2 観点について 5 段階のリッカート尺度で回答してもらった．回答は，「そう思わない（-2）」，「あまりそう思わない（-1）」，「普通（0）」，「ややそう思う（+1）」，「そう思う（+2）」の 5 つから選択させた．

さらに，提案手法による応答発話テキストと比較手法による応答発話テキストをまとめて提示し，各手法に対する印象を「応答に単調さを感じるか」と「このような応答をする対話システムと話してみたいと思うか」という 2 観点から 5 段階リッカート尺度による主観評価実験を行った．

3.2.2 実験結果

まず，提案手法と比較手法における印象評価結果に有意差があるか Wilcoxon の符号付順位検定を用いて検査を行った．「応答発話において文言にあった感情表出ができているか」についての印象評価結果を図 4 に示す．ユーザ感情種別ごとに提案手法と比較手法間の有意差検定の結果は，怒り，悲しみおよび応答発話全体において有意水準 5%を下回っていた．結果から提案手法のほうが比較手法より文言に合った感情表出ができているといえる．一方，「好感が持てる応答をしているか」（図 5）については怒りへの応答以外で有意差は確認できなかった．

次に対話システム全体の印象評価の比較結果を図 6 に示す．「単調だ」，「話したい」は「応答に単調さを感じるか」と「このような応答をする対話システムと話してみたいと思うか」のことである．結果として提案手法の評価結果が有意に高いことが確認できた．このことから応答発話単体では比較手法も好感が持てる応答を行っているものの，非言語的情報も考慮した多様な応答ができる提案手法のほうが単調さを感じさせないため，より話してみたいと思わせていることが確認できた．

4 今後の改良について

現在の対話システムでは音声からユーザ感情を推定しているが，音声のみでは推定の精度に限界があるため，音声だけでなく表情情報も利用した感情推定を行う．表情はユーザ感情が顕著に出やすいため今回の音声対話システムの精度の向上を行いやすい．また，2 つの感情推定手法を別々に適用することでユーザの偽りの感情（表情は笑っているけれど音声では怒りと推定されているなど）を見抜き，「本心」を特定することができると考える．

そのため現在は提案手法の対話システムをもとに表情を感情推定に追加し，性能の向上を図る実験を行っている．表情推定に用いるものとして EmotionNet2[8]と Py-feat[9]を用いた．EmotionNet2 では 8 種類の感情ラベル（平常，幸福，怒り，嫌悪，恐怖，悲しみ，驚き，軽蔑）がある．EmotionNet2 を利用する理由としては作り笑顔の検出ができると推測できたためである．これは「本心」の特定と日本人が欧米人に比べて感情を表情に表しにくく，作り笑顔を多用する傾向があるためである．例としてユーザの感情が「怒り」のとき表情は笑顔だとシステムは「幸福」と誤認識してしまい本当のユーザ感情を読み取ることができず適切な応答ができない．ただ，実際に利用してみなければ作り笑顔を判断できるのかわからないため簡易な性能実験をおこなった．

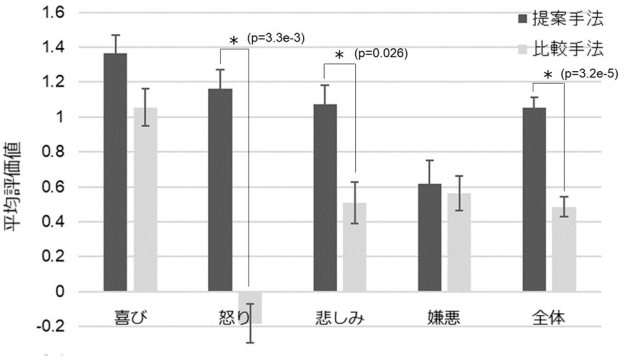

図 4　文言にあった感情表出ができているかについての評価結果

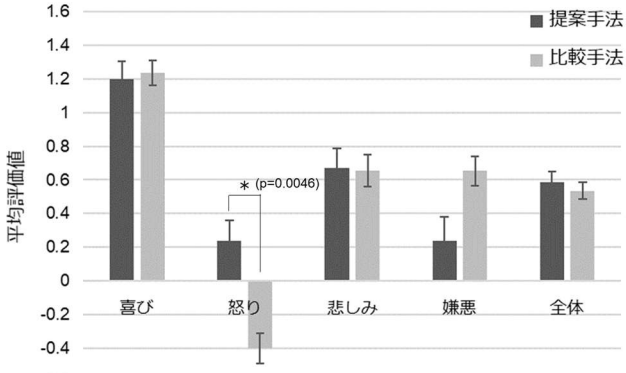

図 5　好感が持てる応答をしているかについての評価結果

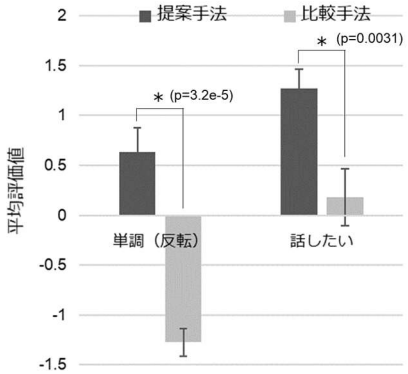

図 6　対話システム全体の印象評価結果

表 2　作り笑顔の検出度実験結果

	平常	幸福	怒り	嫌悪	恐怖	悲しみ	驚き	軽蔑	計
幸福笑顔	0	7	2	0	9	2	0	0	25
作り笑顔	0	2	7	2	8	16	0	0	30

実験方法としては笑顔（幸福）なもの 20 枚とそうでない（作り笑顔）の写真 25 の計 55 枚（自分で判断したもの）を用意し表情推定を行った．結果を表 2 に示す．として笑顔（幸福）20 枚中 7 枚を幸福と推定し，そうでない（作り笑顔）では 25 枚中 2 枚を幸福と推定された．この結果より作り笑顔を検出できると判断した．

今後としては表情を音声のユーザ感情と文章を記憶し，急な感情の変化に対しての適切な応答（怒り→喜びの場合感情の変化が大きすぎるため平常を挟むことや文字列で「急にどうした？」などを投げかける）やそのユーザの質問に対して答えることができる（今日のご飯がおいしかった．[幸福]→お寿司を食べたのですよね？どのネタが美味しかったの？）ようにしていく予定である．

5　おわりに

本研究は，ユーザの非言語的な感情表出を絵文字によって表現することで，テキストベースの統計的応答生成システムによる応答を行う音声対話手法を提案した．喜び，怒り，悲しみ，嫌悪の 4 感情を対象とした提案手法と比較手法における印象評価結果に有意差があるか Wilcoxon の符号付き順位検定を行った結果「応答発話において文言にあった感情表出ができているか」について怒りと悲しみにおいて有意差が確認された．

参考文献

[1] 東中竜一郎, 船越孝太郎, 高橋哲朗, 稲葉通将, 角森唯子, 赤間怜奈, 宇佐美まゆみ, 川端良子, 水上雅博, 小室允人, ドルサ・テヨルス: “対話システムライブコンペティション 3,” 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol.B5, No.02, pp.96-103, 2020.

[2] 目良和也, 谷有希, 村田唯, 黒澤義明, 竹澤寿幸: “演技感情と推定感情のタグを付与した感情音声コーパスの構築,” 日本音響学会 2017 年春季研究発表会講演論文集, pp.1471-1474, 2017.

[3] 東中竜一郎, 稲葉通将, 水上雅博: “Python でつくる対話システム,” オーム社, 2020.

[4] J. Devlin, M. Chang, K. Lee, and K. Toutanova: “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol.1, pp.4171-4186, 2019.

[5] BERT: https://github.com/google-research/bert/blob/master/README.md [accessed Jan. 28, 2022]

[6] 大浦圭一郎, 橋本佳, 南角吉彦, 徳田恵一: “隠れマルコフモデルに基づく日本語音声合成ソフトウェア入門,” システム／制御／情報（システム制御情報学会誌）, Vol.62, No.2, pp.57-62, 2018.

[7] 目良和也, 市村匠, 黒澤義明, 竹澤寿幸: “情緒計算手法と心的状態遷移ネットワークを用いた音声対話エージェントの気分変化手法,” 知能と情報（日本知能情報ファジィ学会誌）, Vol.22, No.1, pp.10-24, 2010.

[8] GitHub, EmotionNet2, https://github.com/co60ca/EmotionNet2 [2022/7/17 アクセス]

[9] Py-Feat, https://py-feat.org/pages/intro.html [2022/7/17 アクセス]

[10] 高木英行: “使える！統計検定・機械学習―Ⅲ―主観評価実験のための有意差検定,” Institute of Systems, Control and Information Engineers, システム／制御／情報, Vol.58, No.12, pp.514−520, 2014.

[11] 中屋澄子: “Scheffé の一対比較法の一変法,” 第 11 回官能検査大会報文集, pp.1-12, 1970.