対話システムの応答決定のためのパラ言語情報を用いた発話態度認識

Speech Attitude Recognition using Paralanguage Information for Determining the Response of the Spoken Dialogue System

宮澤 幸希 * 佐藤 可直

Kouki Miyazawa, Yoshinao Sato

フェアリーデバイセズ株式会社
Fairy Devices Inc.

Abstract: In human-to-human communications, paralinguistic information plays an essential role in conveying the speakers' attitude. Similarly, it is beneficial to give a spoken dialogue system the ability to interpret paralinguistic information for reducing the user's cognitive load and making human-to-machine interactions more smooth. In this study, we investigate a machine learning model that recognizes paralinguistic attitudes of speech. To be specific, we assume four attitude classes essential for determining the system response: agreement, disagreement, question, and stalling. We collected a speech corpus for paralinguistic speech attitude recognition (PAR) and evaluated the model on it. Furthermore, we discuss a plan for evaluating a spoken dialog system to which the PAR model is incorporated.

1 はじめに

音声対話システムに対するユーザの命令やフィードバックを正確に解釈し，タスクを成功させるためには自然言語理解（Natural Language Understanding; NLU）の技術が欠かせない．また近年では，非タスク指向対話におけるユーザの多様な談話行為 [1] の識別（Dialog Act Classification; DAC）も研究されている．しかし従来の NLU や DAC は主にテキスト情報を扱っており，ユーザ発話のパラ言語情報（韻律や声質など）やノンバーバル情報（表情や視線）は補助的な特徴量として用いられてきた [2, 3].

人対人コミュニケーションにおいてテキストとパラ言語が伝える話者の態度は必ずしも常に相補的な関係とは限らず，独立した別個のチャネルである．例えば，語彙的には無意味な短い感動詞が，発話の長さや声質・韻律の違いによって話者の肯定・不満・驚きなどの多様な態度を伝える [4]．怒りの声質は話者の否定的態度の表れと解釈される [5]．養育者は言葉を理解できない乳児に対して，韻律の強調によって受容的態度や危険などを伝える [6]．パラ言語とテキストが矛盾したメッセージを伝える場合，これは皮肉と解釈される．このように，パラ言語情報は単独でも話者の発話態度を伝達する役割を担う．しかし音声対話システムにおいて，パラ言語情報独自のチャネルを活用しようという試みは少数の研究のみにより検討されてきた [7][1].

そこで我々は，音声対話システムへの応用を前提に，パラ言語による発話態度認識（Paralinguistic speech Attitude Recognition; PAR）技術を開発している．音声対話システムの NLU/DAC と PAR を分離することで，ユーザは語彙的な制約に囚われずに，システムをより自然かつ円滑に制御できるようになると期待される．図 1 は PAR によってシステムが理解可能になるユーザ発話の例である．例えば，ユーザの「ええ？」という語尾上げの発話から疑問の意図を察してメッセージを再度提示するシステムや，「この部屋暑い。」という平叙文からユーザの気分を察して空調を作動させるシステムなどが実現できる．さらに，ユーザがシステムに対して完全文（テキスト情報のみから正確に話者の意図が伝わる文）を考えて発話する手間を省けるため，認知負荷を軽減できると考える．

* 連絡先：フェアリーデバイセズ株式会社
〒113-0034 東京都文京区
E-mail: miyazawa@fairydevices.jp

[1] いっぽう感情認識においては，パラ言語で表現される emotion とテキストで表現される sentiment に切り分けることが多い．

システム：「予約をキャンセルしますか？」
ユーザ：
「ええ。」　（了解。キャンセルして）
「ええ？」　（聞こえない。もう一度言って）
「ええ!?」　（勝手にキャンセルしないで）
「ええ―…」　（考えるからちょっと待って）

ユーザ：
「この部屋暑い。」（クーラーをつけてほしい）
「この部屋暑い？」（今の室温が知りたい）

図 1: PAR によって可能になるユーザ発話例.
（下線部はパラ言語情報で表現される音声，丸括弧内は完全文で記述したユーザの意図を示す）

パラ言語認識器の有用性を検証するために，本稿では以下の 2 つのリサーチクエスチョンを扱う．

1. 一般的なユーザはパラ言語情報を用いて複数の発話態度を表出することが可能か？
2. それを機械学習モデルで識別可能か？

ユーザがシステムを制御するためには，少なくとも 4 つの基本的な態度，すなわち「肯定・否定・疑問・考え中」が必要であると仮定する（表 1）．

本稿では，ユーザの発話態度を表 1 の 4 クラスに分類する問題を考える．各態度で読み上げた音声を収集して人手で評価し，機械学習モデルを開発することでリサーチクエスチョンの検証を行う．また，次のステップとして PAR を組み込んだシステムが実際に特定の利用状況において対話の自然性・円滑性を向上させ，かつユーザの認知負荷を下げるかを検証するための実験計画について論じる．

2 パラ言語による態度認識器の開発

2.1 音声コーパス収録

2.1.1 テスト収録コーパス

初めにテスト収録として，プロの音響監督のもとで 6 名の演技経験者に表 1 の各態度で文章を読み上げてもらった．文章は日本語話し言葉コーパス[8]から語彙的な態度が中立で，かつパラ言語的な特徴（発話長，アクセント句の音節数・アクセント核の位置・母音の種類・子音の有声性）のバランスを考慮した 315 文（1 セット 63 文を 5 セット）を選んだ. いずれかのセットの 63 文を 4 通りの態度で読んでもらい，合計 1,512 文を収録した．

著者らはテスト収録に立ち会い，監督及び作業者らと議論しながら，4 つの態度をパラ言語のみで表現するために注意すべき点を文書化した（図 2）．

表 1: システム応答決定のための態度 4 クラス.

クラス	ユーザの態度	システム応答例
肯定	賛成である，満足である	現在のアクションを続行する，次に進む
否定	反対である，不満がある	アクションをキャンセルする，指示を求める
疑問	質問がある，理解できない	質問に回答する，前の発話を言い直す
考え中	懸念がある，熟慮している	指示を待つ，追加情報を提示する

肯定（文章の最後が「。」）
　相手に賛成するとき、話の内容に異論がないとき、相手の話にあいづちを打つとき、自分の考えを述べるときなどの、**下がり調子**のイントネーションでお読みください。

否定（文章の最後が「!?」）
　話の内容にどちらかといえば反対のとき、相手の話に納得がいかないとき、話の流れを変えたいときなどの、**やや短い上がり調子**のイントネーションでお読みください。

疑問（文章の最後が「？」）
　質問をするとき、相手の返事が欲しいとき、事実確認をしたいとき、わからなくて聞き返したいときなどの、**やや長い上がり調子**のイントネーションでお読みください。

考え中（文章の最後が「…」）
　難しい話をしているときや、悩んだり戸惑ったりしているときなどの、**語尾を引き伸ばした**イントネーションでお読みください。

図 2: 話し方マニュアル（本収録の話者に提示）.

2.1.2 本収録コーパス

続いて本収録では，演技経験を問わずクラウドソーシングで募集した 138 人の作業者に図 2 のマニュアルを教示した上で，テスト録音と同じ文を読み上げてもらった．別の作業者によって低品質と評価された録音は除外して，32,148 件を収録した．

2.2 コーパスの分析

本収録とは別の 20 人のクラウドソーシング作業者に，本収録コーパスの各音声の発話態度が何に聞こえるかを回答してもらった. 各音声には 2 名または 3 名の作業者を割り付けて複数人で評価した．

表 2 は評価結果の集計である．「肯定」の精度は 0.956 で，ほとんどの肯定の態度が正しく認識できている．「疑問」「考え中」の精度はやや下がるが比較的正確に認識できている．したがって，これらの態度クラスをパラ言語のみで表現することは，演技経験を問わず一般ユーザでも可能であることが示された．いっぽう「否定」の精度は 0.344 であり芳し

表 2: 本収録コーパスに対する人手の評価結果.
（表の右下は全体の精度を示している）

		評価者が回答した態度				Recall
		肯定	否定	疑問	考え中	
話者が発話した態度	肯定	15586	72	236	404	0.956
	否定	2156	8403	13579	231	0.344
	疑問	1297	3356	19126	380	0.791
	考え中	4848	96	481	10221	0.653
Precision		0.652	0.704	0.572	0.909	0.662

表 3: 本収録コーパスに対するモデルの認識結果.
（表の右下は全体の精度を示している）

		モデルが認識した態度				Recall
		肯定	否定	疑問	考え中	
話者が発話した態度	肯定	7277	32	513	327	0.893
	否定	1269	1246	5439	169	0.153
	疑問	1080	405	6346	222	0.788
	考え中	2035	26	462	5300	0.677
Precision		0.624	0.729	0.497	0.880	0.627

くない. パラ言語のみを使って否定の態度を表出し，またそれを正しく聞き取ることは難易度が高いタスクであることが示唆される.

2.3 認識器

以下の 17 次元の特徴量を使用した：基本周波数（F0），第 1・第 2 フォルマント（F1，F2），振幅の二乗平均平方根（RMS）とゼロ交差率（ZCR），12 次元のメル周波数ケプストラム係数（MFCC12）. 16kHz へのダウンサンプリングと振幅の正規化の後，MFCC 以外はフレーム長 10ms とフレームシフト 10ms で計算され，MFCC はフレーム長 25ms とフレームシフト 10ms で計算された.

本研究で使用したニューラルネットワークモデルは tanh 活性化関数を備えた LSTM 2 層（サイズはそれぞれ 300 と 100 ），ドロップアウト層（ドロップアウト率 0.5）および softmax 活性化関数を備えた全結合層で構成した. 出力は 4 つの態度クラスの確率分布である．訓練には RMSProp オプティマイザを使用し，バッチサイズは 32，学習率は 0.0001 とした.

テスト収録コーパスと本収録コーパスを混合したデータを 6 分割し交差検定を行った．各 fold では訓練に 4/6，最適なエポックの選択に 1/6，評価に 1/6 を使用した.

訓練の際は，本収録コーパスの中で態度表出が不明瞭と評価された音声（[9] において提案された手法を用いて求めた信頼度のスコアが 0.8 以下の音声）を除外した．評価の際は本収録コーパスのみを使用し，データの除外は行わなかった.

表 3 はモデルの認識結果の集計，表 4 は人手およびモデルによる結果の F スコアである．「肯定」「考え中」の態度について，機械学習モデルのパフォーマンスは人手による評価結果とほぼ同等であった．「疑問」はやや F スコアが低く，「否定」は人手評価よりも顕著に F スコアが悪化した．テスト収録の際に，否定発話は他のクラスよりも精緻な発話制御が必要であり，特に疑問発話と区別することが難しいと報告されているため，本モデルの表現力が不足していることが考えられる．モデルの誤認識のパターンは人手評価と類似しており，考え中を肯定に，否定を疑問に誤認識する傾向があった.

表 4: 人手評価およびモデル認識結果の F スコア.

	態度				平均
	肯定	否定	疑問	考え中	
人手	0.775	0.463	0.664	0.760	0.697
モデル	0.734	0.253	0.609	0.765	0.654

3 システム評価実験計画

3.1 パラ言語による態度認識が有用な状況

ここまでの検討によって，少なくとも肯定・疑問・考え中の 3 つの態度については，機械学習モデルによって実用的な精度で認識できることが示された．そこで，パラ言語による態度認識器を組み込んだ音声対話システム（以下，「提案システム」）のユーザビリティ評価が次の課題である.

提案システムが有用なのは，曖昧性のない完全文の発話を考えてシステムに話しかけるための認知的コストが高い状況であると考えられる．例えば以下の状況では，提案システムはユーザにとって使いやすく，負担の少ないシステムとなりうる.

1. ユーザの認知的資源が限定される状況．簡単なゲームや作業を実行している最中や，視聴覚コンテンツの視聴中など.
2. ユーザに認知的負担をかけたくない状況．早朝や夜間，カジュアルなシステムとのくだけた対話，緊急性のないタスクの実行中など.

3.2 評価実験の概要

3.1 で述べた 1. または 2. の利用状況を想定して，提案システムと NLU/DAC のみを行う従来システムの比較実験を行う予定である．

両システムは Wizard of Oz 法でオペレータが操作し，ユーザには任意の話し方でシステムと対話しながら特定のタスクを解決するように要求する．その上で，タスク達成までにかかった発話数及び時間を測定する．また，システムに対する印象（自然性：どちらのほうが自然に会話できたか，円滑性：どちらが使いやすかったか，継続性：また使いたいのはどちらか，等）のアンケート調査を実施する．

4 まとめ

パラ言語による発話態度認識 PAR を提案した．本稿では，一般的なユーザが発話長や韻律などのパラ言語情報のみによって機械との対話に必要な態度を表出することは可能か，またそれを機械学習モデルで認識可能かについて検討した．

その結果,「否定」の態度の表出は演技経験者には可能であるものの，一般ユーザにはやや困難であることが示唆された．いっぽう「肯定・疑問・考え中」の 3 種類の態度に関しては主にイントネーションの上げ方・下げ方・伸ばし方によって表出可能であることが示された．この話し方は簡単なインストラクションでほとんどのユーザが習得できた．さらに，開発した態度認識器はこれらの態度を人と同程度の精度で推定できた．以上より，我々が提案する PAR を組み込んだ音声対話システムが実用的であることが示された．

パラ言語を認識可能なシステムによって，ユーザは発話文の内容を熟慮することなくシステムを制御できるようになると期待される．今後，提案したシステムによって，ユーザが実際に自然に，少ない認知的負担で音声対話システムを使用できるかどうかの評価実験を計画している．

謝辞

本稿執筆にあたり，社内で音声収録環境の管理を担当いただいた石原道佳氏に感謝します．

参考文献

[1] Pareti S., and Lando T.: Dialog intent structure: A hierarchical schema of linked dialog acts, Proc. of LREC, (2018)

[2] Si Y., Wang L., Dang J., Wu M., and Li A.: A hierarchical model for dialog act recognition considering acoustic and lexical context information, Proc. of ICASSP, (2020)

[3] Bothe C., Weber C., Magg S., and Wermter S.: A context-based approach for dialogue act recognition using simple recurrent neural networks, Proc. of LREC, (2018)

[4] Ishi C. T., Ishiguro H., and Hagita N.: Automatic extraction of paralinguistic information using prosodic features related to f0, duration and voice quality, Speech Communication, Vol.50, No.6, pp.531–543, (2008)

[5] Saha T., Patra A., Saha S., and Bhattacharyya P.: Towards emotion-aided multi-modal dialogue act classification, Proc. of ACL-58, (2020)

[6] Fernald A.: Intonation and communicative intent in mothers' speech to infants: Is the melody the message?, Child Development, Vol.60, No.6, pp.1497–1510, (1989)

[7] Takatsu H., Yokoyama K., Matsuyama Y., Honda H., Fujie S., and Kobayashi T.: Recognition of intentions of users' short responses for conversational news delivery system, Proc. of INTERSPEECH, (2019)

[8] Maekawa K.: Corpus of spontaneous japanese : its design and evaluation, Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.7–12, (2003)

[9] Sato Y., and Miyazawa K.: Quality estimation for partially subjective classification tasks via crowdsourcing, Proc. of LREC, (2020)