音声に含まれる感情を考慮した自然言語対話システム

Natural Language Dialogue System considering Emotion Guessed from Acoustic Features

Tang Ba Nhat[1]　目良和也[1]　黒澤義明[1]　竹澤寿幸[1]

Tang Ba Nhat[1], Kazuya Mera[1], Yoshiaki Kurosawa[1], and Toshiyuki Takezawa[1]

[1]広島市立大学大学院情報科学研究科

[1]Graduate School of Information Sciences, Hiroshima City University

Abstract: With the previous IVR (Interactive Voice Response) systems, we can tell the computer to do our simply tasks only by giving a short voice command. But in human interaction, we do not rely on only the content of utterance, like the person who said "I'm fine" with sad voice maybe not really fine, for instance. In this paper, we propose an intelligent IVR system which uses not only the content but also the acoustic features of the utterance to make better response. Firstly, the method estimates user's emotion (Acceptance, Anger, Hope, Disgust, Fear, Pleasure, Sad, Surprise and Neutral) by SVM using acoustic features extracted from user's voice. Then the method applies AIML-based matching rules to the user's utterance based on acoustic and linguistic features. As a result of experiment, the proposed system was preferred, was felt as more flexible and got closer to human-like communication.

1．はじめに

近年携帯電話や車に搭載される音声対話システムが普及しつつある．これらの対話システムは，「ドアを開けて」や「電話をかけて」のように何かしらの作業を実施したり，「今日の天気は？」や「世界で一番高い山は？」のように情報を検索したりすることが出来る．例えばアップルの iOS には Siri があり，Android にはしゃべってコンシェルがある．これらは検索エンジンとアシスタントの両方の機能を持っている．

しかし従来の音声対話システムでは“字面でしか判断できない”という問題点がある．例えばシステムからの「元気ですか？」という問いかけに対して，力無く「元気だよ…」と答えても，音声認識処理では「元気だよ」という文字列としてしか情報を得られないため，システムはユーザが元気なものとして対話を進める．しかし人間同士の対話の場合，たとえ「元気だよ」と答えていても，声の調子や表情などのノンバーバル情報からも相手の状態を判断し，「強がっているけど本当は元気じゃないな」というような判断をすることができる．

本研究では音響分析によりユーザの感情を推定し，発話内容と推定感情の両方を考慮した対話を実現することを目指す．音響分析では声の高さと大きさの変化から特徴量を取り出し，機械学習器によって，Plutchik[1]の 8 感情（受容，怒り，期待，嫌悪，恐れ，喜び，悲しみ，驚き）に平静を加えた 9 種類の感情から 1 種類の感情を選択する．発話内容については音声認識結果を AIML 言語に基づいて作成した返答ルールの条件部と照らし合わせて適用するルールを決定する．

2．構築システム

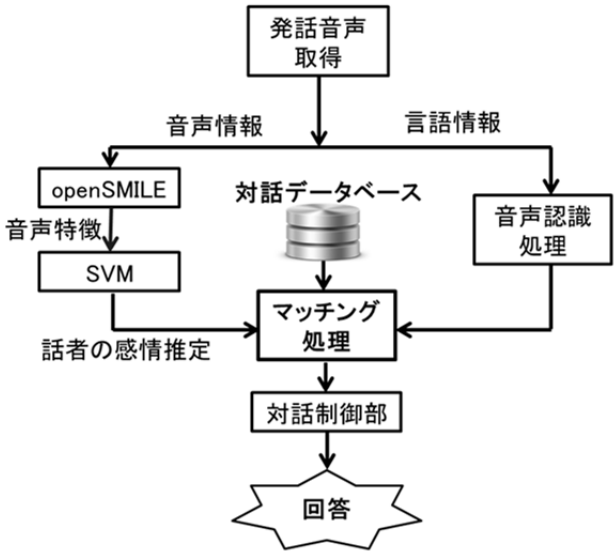


図 1：システムの処理の流れ

システムは入力されたユーザの発話に対して音響的に感情を分析する処理と言語的に発話内容を分析する処理を並列に行う．音声情報処理部ではopenSMILE[2]を利用して音響的特徴を抽出し，この特徴を SVM（Support Vector Machine）に学習させることによってユーザの感情を推定する．言語情報処理部では音声認識結果とデータベースにある返答ルールのマッチングを行う．対話制御部では，ユーザ感情推定の結果と返答ルールから返答発話を生成し，音声合成装置を使って返答を出力する．

２．１．音響的特徴からの話者の感情推定

一般的に話者の感情状態は声の大きさや高さなどといった声の調子に表れる．これまで発話の音響的特徴から話者の感情を推定する研究が数多く行われているが，本研究では，音響的特徴量を入力音声データから算出し，機械学習器に学習させることによって分類を行う．素性値の算出には INTERSPEECH 2009 Emotion Challenge [3]でコンペティションに用いられたオープンソースソフトウェア openSMILE [2]を用いる．

openSMILE で取得可能な特徴量を表 1 および表 2 に示す. 本研究では openSMILE で取得できる特徴量から快不快感情の推定に有効な素性を Forward Stepwise Selection(FSS)手法を用いて絞り込む．

FSS 手法とは，最初に 1 素性のみを使って評価を行い，最も正解率の高いものを選ぶ．次に，残りの素性から 1 つずつを選び，最初に選択した素性と組み合わせた 2 素性による最適の組み合わせを探す．そしてその次は選択した 2 素性＋他の 1 素性の組み合わせを探し…という流れで，有効な素性の組み合わせを調べる手法である．最適ではない可能性が高いものの，FSS 手法を用いることで，大量の素性から感情推定に有効なものを簡便に選択できる．

学習データとしては,「感情評定値付きオンラインゲーム音声チャットコーパス[4]」のタグ付き演技音声のうち，受容，怒り，期待，嫌悪，恐れ，喜び，悲しみ，驚き，平静の音声各 30 個を用いた．

FSS 手法の結果，表 3 の素性の組み合わせが最も正解率が高かったため，本手法でも表 3 の素性を感情推定に用いる．

２．２．言語情報のマッチング処理

本研究はユーザ発話を音声認識エンジンを使ってテキストに変換し，対話データベースに書いてある返答ルールの条件部とマッチングする．これまで対話システムに関する研究が数多く行われている．例えば，AIML 言語によって作成された ALICE システム[5]は、話題やユーザの情報やユーザの直前の発話を取得して，ユーザに返答する．本研究はユーザの発話とその発話に対する返答データを単純にマッチングすることによって返答を出すシステムだけではなく，音響分析によって得られるユーザの感情とユ

表 1　openSMILE で取得可能な特徴量

特徴	説明
RMSenergy	音量の二乗平均平方根値
MFCC	1 次～12 次メル周波数ケプストラム係数
F0	基本周波数
voiceProb	その時点での音が声である確率
pcm_zcr	波形のゼロ交差率
（および上記各特徴量の一次微分）	

表 2　openSMILE で取得可能な素性値

素性値	説明
max	データ中の最大値
min	データ中の最小値
range	最大値と最小値の差分
maxPos	最大値を出力した位置
minPos	最小値を出力した位置
amean	算術平均
linregc1	線形近似の勾配度
linregc2	線形近似のオフセット
linregcerrQ	線形近似の二乗誤差
stddev	標準偏差
skewness	歪度
kurtosis	尖度

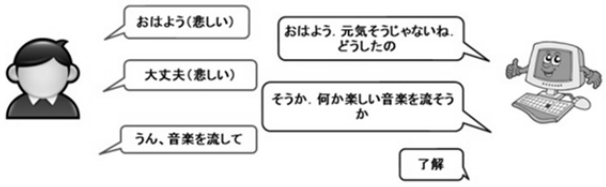
表 3　感情推定に有効な音響的特徴

順位	音響的特徴
1	Δ RMSenergy-max
2	RMSenergy-linregc2
3	MFCC[1]-linregc2
4	Δ MFCC[10]-max
5	Δ MFCC[10]-min
6	Δ MFCC[8]-linregc1
7	Δ MFCC[9]-linregc2
8	Δ voiceProb-maxPos
9	Δ MFCC[12]-linregc2
10	Δ MFCC[4]-linregc1
11	ZCR-amean

ーザの直前の発話内容を条件としてマッチングを行う．

本手法で用いる返答ルールは「質問#条件 1#条件 2#返答」のような形式で作成される．“質問”は直前のユーザ発話を表す．“条件 1”はユーザ発話の前にシステムが出力した発話を表す．条件 1 を参照することで，一問一答型より一段階前まで話の流れを考慮した対話を実現できる．“条件 2”はユーザ発話を音響分析して得られる感情の種類を表す．これにより，同じ字面のユーザ発話に対しても，口調に合わせて異なるリアクションをとることができる．“返答”は条件に合致した場合にシステムが出力する返答の文字列を表す．

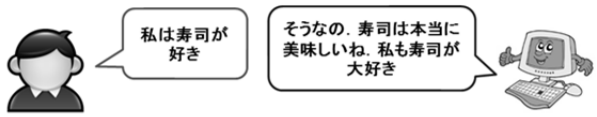
例えば「おはよう# #楽しい#おはようございます」と「おはよう# #悲しい#おはよう．元気そうじゃないね．どうしたの」との 2 つのルールがある場合，ユーザがシステムに「おはよう」と言ったら，システムはユーザの感情を推定して適切な返答を行う．



1	おはよう# #楽しい#おはよう．今日は良い天気だね
2	おはよう# #悲しい#おはよう．元気そうじゃないね．どうしたの
3	試験に落ちたよ#おはよう．元気そうじゃないね．どうしたの#悲しい#今度もっと頑張ろう
4	大丈夫###何が大丈夫なの
5	大丈夫#おはよう．元気そうじゃないね．どうしたの#楽しい#良い日を過ごして下さい
6	大丈夫#おはよう．元気そうじゃないね．どうしたの#悲しい#そうか．何か楽しい音楽を流そうか
7	*音楽を流して#そうか．何か楽しい音楽を流そうか##了解

図 2：対話例

ユーザ発話のマッチング条件を緩和するため，返答ルールにはワイルドカードを利用できる．例えばユーザの発話として「おはよう」，「おはようございます」，「おはようさん」などさまざまな表現形態があるが，これらそれぞれに対して別々に返答ルールを作成するのは現実的ではない．そこでワイルドカードを使って「おはよう*# # #おはようございます」というような返答ルールが有ればその 3 つのユーザ発話がいずれもマッチングできる．さらにワイルドカードを使うことによってユーザの発話に含まれる重要な内容を取得し，話の内容を考慮した返答を作成できる（図 3 参照）．



私は*が好き# ##get_wildcard>そうなの．*は本当に美味しいね．私も*が大好き

図３：ワイルドカードを用いた返答生成

３．評価実験

３．１．実験システム

本手法に基づいて話者感情を考慮した対話システムを構築した．音声認識装置には Julius[6]，音響分析装置には openSMILE[2]，機械学習器としてオープンソースの統計解析システム R[7]上で SVM を実行した．登録されている返答ルールの数は 7 個である．

この対話システムを用いて話者感情を考慮したシステムと話者感情を考慮しない従来のシステムを作成し，評価実験を行った．

３．２．質問紙調査

本論文では，自然言語対話システムに話者感情推定機能の有る／無しの 2 種類の振る舞いをさせ，同じ被験者にその印象を評定させることで，自然言語対話システムにおける話者感情推定処理の有効性を明らかにする．

実験手法としては，話者感情を考慮した対話システムの動作を収録した動画と話者感情を考慮していない対話システムの動作を収録した動画を連続で視聴し，それぞれの振る舞いについて，大学生 10 名（男性 8 名，女性 2 名）の被験者がどう感じるかを調査した．表 4，5 に各対話内容を示す．動画の長さは約 1 分であり，2 つの動画を視聴し，2 つ見終わった時点でそれぞれの対話システムに対する印象を質問紙により回答してもらった．なお順序効果を無くすため，被験者を 2 群に分け，それぞれ感情推定有り→無しと感情推定無し→有りの順序で動画を提示した．

印象の評価は，エージェントの印象を表す評価項目に対する 5 段階評価（2：そう思う，1：ややそう思う，0：普通，-1：あまりそう思わない，-2：そう思わない）によって行った．評価項目は，高吉がロボットの印象調査に用いた手法[8]を参考に，“好感因子”，“性格的要因に関する項目”，“機械らしさ・生物らしさに関する項目”に属する形容詞 15 個，さらに本実験向けにユーザが対話システムに対して感じる“使用感に関する項目”として，「楽しめそう」，「イライラする」，「規則的」の 3 つの項目についても調査し，分析した（表 6）．

表 4　話者感情を考慮しない対話事例

システム「システム「こんにちは」 ユーザー「こんにちは」 システム「元気ですか？」 ユーザー「元気だよ（元気なく）」 システム「それは良かった．今日も一日がんばりましょう」

表 5　話者感情を考慮した対話事例

システム「こんにちは」 ユーザー「こんにちは」 システム「元気ですか？」 ユーザー「元気だよ（元気なく）」 システム「本当ですか？声に元気がないですよ.」 ユーザー「疲れたよ」 システム「では，あなたの好きな曲をおかけしましょう」

表 6　質問紙調査の評価項目

グループ	形容詞の項目
好感因子	外交的，真似したい，格好いい，感じがいい，仲良くなれそう，また会いたい，優しそう，共感が持てる，気が利く
性格的要因に関する項目	真面目そう，おとなしい，やんちゃ，不真面目
機械らしさ･生物らしさに関する項目	機械らしい，生物らしい
使用感に関する項目	楽しめそう，イライラする，規則的

３．３．結果

好感因子および各項目の平均得点（Mean）と標準偏差（Standard Deviation）を表 7 に示す．なお好感因子の得点は，各因子内の項目の得点の和となっている．

好感因子について：好感因子が高いことは，直接的に対話システムへの親しみやすさにつながる．本手法の平均得点が有意に高いことがわかった．これは，同じ内容の発話でも音響的に表出された感情によって反応に変化があるため，より相手の気持ちを理解しているように感じた結果と考えられる．特に「優しそう」と「気が利く」の項目での差が著しかった．

表 7　各因子・項目の平均得点および標準偏差

			感情を考慮したシステム	従来のシステム
		N	10	10
好感因子	外交的	Mean	1.20	0.40
		SD	0.97	0.66
	真似したい**	Mean	0.60	-0.50
		SD	0.49	0.81
	格好いい*	Mean	0.10	-0.70
		SD	1.22	0.78
	感じがいい*	Mean	1.30	-0.30
		SD	1.00	1.10
	仲良くなれそう**	Mean	1.30	-0.20
		SD	0.64	0.87
	また会いたい**	Mean	0.90	-0.30
		SD	0.83	0.64
	優しそう**	Mean	1.60	0.10
		SD	0.49	0.94
	共感が持てる**	Mean	1.10	-0.40
		SD	0.54	0.92
	気が利く**	Mean	1.60	-0.50
		SD	0.66	1.12
性格的要因	真面目そう	Mean	0.70	1.20
		SD	0.64	0.75
	おとなしい	Mean	-0.2	0.60
		SD	0.98	0.66
	やんちゃ	Mean	-0.50	-1.20
		SD	1.20	0.75
	不真面目	Mean	-1.30	-0.90
		SD	0.64	1.04
機械・生物	機械らしい**	Mean	-0.80	1.00
		SD	0.87	1.10
	生物らしい**	Mean	0.80	-0.80
		SD	0.60	0.87
使用感	楽しめそう**	Mean	1.00	-0.50
		SD	0.63	0.81
	イライラする	Mean	1.00	-0.60
		SD	1.00	0.92
	規則的**	Mean	-0.80	0.90
		SD	0.40	0.70

**p<.01　　*p<.05

注：N:被験者の数，p: t 検定の確率

機械らしさ・生物らしさに関する項目：従来のシステムがより機械らしく，提案システムがより生物ら

しいという回答が得られた．これは従来システムの「口調に関わらず同じ回答を返す」という特徴が単調な機械処理を連想させたことに比べ，提案システムが口調を考慮して反応を変えたことがより人間同士のコミュニケーションに近い印象を与えたためと考えられる．

使用感に関する項目:「楽しめそう」の項目において，本手法のエージェントが高い平均得点を得ている．これは，話し方を考慮して応答が変化するため，対話システムの反応を楽しむような被験者の気持ちによるものと思われる．逆に，「規則的」の項目では従来手法のエージェントが高い平均値を得ている．これは，従来手法では返答が固定なため，被験者がエージェントの反応に対し規則性を感じた結果と考えられる．

４．まとめ

本研究では音声に含まれる感情を考慮した自然言語対話システムを提案した．より人に近い対話システムを実現するためユーザの発話の音声特徴からユーザの感情を推定し，返答プラン決定の条件として用いた． 11 種類の音響的特徴を機械学習器 SVM に学習させることで，9 種類のユーザ感情を推定できるようにした．また AIML に基づいた返答ルールを使用した．

評価実験の結果，提案手法の好感因子の得点が有意に高かった．また提案手法はより生物らしいと感じられており，使用感の面にも従来手法より楽しめそうだと思われる．

今後の課題としては，より知的で柔軟な対話システムを作成するため，返答ルールのデータベースを拡大し，またユーザ感情推定の正解率を高める予定である．

謝辞

本研究の一部は JSPS 科研費 50285425 の助成を受けたものです．

参考文献

[１] Plutchik, R., “The emotions”, New York: Random House (1962)

[２] F. Eyben, M.Wöllmer, B. Schuller, “openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor,” ACM Multimedia Conference – MM, pp.1459-1462 (2010).

[３] B. Schuller, S. Steidl, A. Batliner, “The INTERSPEECH 2009 Emotion Challenge,” Proc. of INTERSPEECH2009, pp.312-315 (2009).

[４] 有本泰子，河津宏美，音声チャットを利用したオンラインゲーム感情音声コーパス," 日本音響学会 2013 年秋季研究発表会講演論文集, 1-P-46a (2013).

[５] ALICE and AIML Documentation http://www.alicebot.org/documentation/

[６] 李晃伸，河原達也，鹿野清宏． 「２パス探索アルゴリズムにおける高速な単語事後確率に基づく信頼度算出法」 情報処理学会研究報告，2003-SLP-49-48 (2003).

[７] R. Ihaka and R. Gentleman, “R: A language for data analysis and graphics,” Journal of computational and graphical statistics, Vol.5, pp.299-314 (1996).

[８] 高吉幸治，田中俊也，“ロボットの振る舞いと知性・性格の印象の関係，”情報処理学会研究報告（CVIM），コンピュータビジョンとイメージメディア，Vol.2007, No.87, pp.43-48 (2007).