2T-08

非言語情報を用いた対話システムにおける対話破綻の検出

秋水紫苑† 入部百合絵† 北岡教英‡

愛知県立大学情報科学部† 徳島大学大学院社会産業理工学研究部知能情報系‡

1 はじめに

人工知能技術の進展に伴い対話システムが身近なものになっているが，人間同士のようなスムーズな対話ができているとは言い難い．例えば，対話の流れに沿わない応答や，急な話題転換を行うといった対話破綻が頻発しているのが現状である．しかし，対話破綻の検出を行うことができれば，対話破綻を回避するための対話シナリオに切り替える，あるいは破綻しても破綻の検出を行うことで破綻からの回復を行うといったことが可能となる．

従来の対話破綻の研究では，テキストチャットを対象とした言語情報からの破綻検出が数多く研究されている[1][2]．また，音声対話システムにおける雑談音声対話を対象とした音響情報からの破綻検出の研究も行われている[3]．本研究では，雑談音声対話の動画を観察した結果，視線情報も対話破綻検出の判断材料として有用であることが確認されたため，本研究では音響情報と視線情報を用いた対話破綻の検出を行う．

2 収集した雑談対話音声

本研究では，人間が対話システムの振りをして対話を行う Wizard-of-Oz（WoZ）法を用いて収録された雑談対話音声を使用する．被験者数は 10 人であり，1 人につき 6 セッション(被験者 5 のみ 4 セッション)の対話音声を収録した．被験者には 1 セッションにつき 10 発話以上の対話を行うよう指示し，セッション終了が分かるように対話の終了時には「さようなら」と発話してもらった．収録されたデータに対し，破綻のラベルを付与した．システムが破綻した最初の発話に破綻ラベルを付与する作業をセッション毎に行い，破綻直前のユーザ発話を破綻前，直後のユーザ発話を破綻後とした（図 1）．ラベルを付与した結果，破綻前のユーザ発話と破綻後のユーザ発話のセットを 42 セット得た．

Detection of dialog breakdown in dialog system using nonverbal information

†Shion Akimizu, Yurie Iribe, School of Information Science and Technology, Aichi Prefectural University

‡Norihide Kitaoka, Graduate School of Technology, Industrial and Social Sciences, Tokushima University

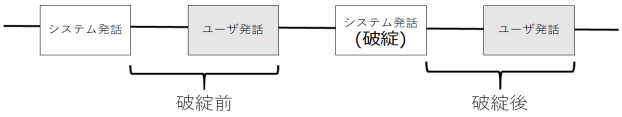

図 1：破綻前後の音声の分析範囲

3 対話破綻検出に用いる特徴量

破綻前と破綻後のユーザ発話から音響的特徴量と視線情報を抽出し，破綻前後で有意な差が現れる特徴量を抽出する．そして有意差の認められた特徴量をもとに対話破綻を検出する．

3.1 音響的特徴量の抽出

音響的特徴量の抽出には音響解析ツールである OpenSMILE を用いて，384 次元の特徴量を抽出した．抽出した音響的特徴量の分析を行ったところ，162 次元音響的特徴量に t 検定による有意差が認められた．紙面上の関係で，Zero-crossing rate/Voice Probability/F0 の平均，Voice Probability/F0 の標準偏差の分析結果のみ表 1，2 に示す． Zero-crossing rate に有意差が認められた理由として，破綻後は笑い声や言い淀み等が破綻前に比べ発生しやすいことが要因であると考えられる．また，システムの対話破綻後のユーザは次の発話を躊躇するため，ユーザの発話前の無音区間が長くなる，あるいは言い淀みが増加することから Voice Probability に有意差が認められたと考えられる．一方，破綻前後の F0 の変化率を算出した結果，破綻前に比べ破綻後の F0 値は約 35%下がることが確認された．

3.2 視線情報の抽出

実験中，被験者は MMDAgent のエージェントを対話相手として視線を向けることから，破綻前後の視線の変化にも注目する．視線抽出にはアイトラッカーを用いて視線の x，y 座標を出力する．座標はエージェントを映すモニターの左上を原点(0,0)，右下(1,1)として算出した．また，視線情報の抽出間隔は 40fps とし，29 セッション

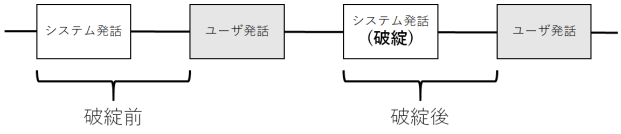

図 2：破綻前後の視線の分析範囲

表 1：各特徴量（平均）の t 検定結果

特徴量	平均		p 値
	破綻前	破綻後	
Zero-crossing rate	0.155	0.171	1.42E-05**
Voice Probability	0.197	0.179	0.00119**
F0	9.78	6.62	0.0240*

*p<0.05, **p<0.01

表 2：各特徴量（標準偏差）の t 検定結果

特徴量	標準偏差		p 値
	破綻前	破綻後	
Voice Probability	0.120	0.106	0.00275**
F0	33.37	26.20	0.0439*

*p<0.05, **p<0.01

分のデータを用いた(視線情報は発話の開始時間の対応がとれないものがあったため)．視線の分析対象を図 2 に示す．システムの破綻発話の直後から視線の変化は顕著になることを仮定し，音響的特徴量の場合とは異なりシステム発話からユーザ発話開始までを分析対象とした．抽出した特徴量は x，y 各座標の平均と標準偏差，平均の変化量，標準偏差の変化量，各変化量の最大値，標準偏差の最大値である．上記の特徴量をもとに求めた視線情報の算出方法を以下に示す．

①破綻前：セッション毎の平均の変化量
　破綻後：5 フレーム毎に算出した平均変化量の内の最大値
②破綻前：セッション毎の標準偏差の変化量
　破綻後：5 フレーム毎の標準偏差の変化量の最大値
③破綻前：セッション毎の標準偏差
　破綻後：5 フレーム毎の標準偏差の最大値

破綻後に微小に変化する視線を適切に捉えるために，分析範囲全体ではなく，短い区間毎に平均や標準偏差を算出し，その変化幅の大きい箇所を検出することを試みた．また，破綻後の視線の変化を捉えるのに適した間隔を確認するため，①から③の破綻後で求めたフレーム間隔を 5 だけではなく 10，15 に変更した値もそれぞれ求めた．紙幅の関係上，t 検定で有意差が認められた視線情報のうち，5 フレームに関する結果のみ表 3 に示す．

表 3：視線情報の分析結果

				p 値
		破綻前	破綻後	
①	y	0.00119	0.178	0.00336**
②	x	-0.0107	-0.0448	0.0166*
	y	-0.0222	-0.100	0.00681**
③	x	0.0649	0.0928	0.00733**
	y	0.144	0.185	0.0353*

*p<0.05, **p<0.01

表 4：破綻検出結果

	Precision	Recall	F-measure
破綻	0.929	0.897	0.912
非破綻	0.900	0.931	0.915
加重平均	0.914	0.914	0.914

表 5：混同行列

		識別クラス	
		破綻	非破綻
真のクラス	破綻	26	3
	非破綻	2	27

4 評価実験

有意差の認められた 162 次元音響的特徴量と 12 次元の視線の特徴量を用いて，対話破綻の識別を行った．識別器は Random Forest を用い，10-分割交差検証により評価した．評価指標には破綻前，破綻後，加重平均に対する Precision，Recall，F-Measure を用いる．その結果を表 4 に示す．また，混同行列を表 5 に示す．

表 4，5 から，非常に高い結果を得ることができ，本研究で識別に用いた特徴量が破綻の検出に有効であることが示された．

5 おわりに

本研究では対話破綻前後の対話音声と視線情報を分析し，破綻の識別に有用な音響的特徴量と視線情報を用いて対話破綻の検出を行った．評価実験より本研究で用いた特徴量が破綻検出に有用であることを確認した．今後の課題は，引き続き対話音声の収録を行うとともに，対話破綻の種類による音響言語情報および非言語情報の特徴を分析することである．

参考文献

[1] 東中竜一郎，船越孝太郎，荒木雅弘，塚原裕史，小林優佳，水上雅博：テキストチャットを用いた雑談対話コーパスの構築と対話破綻の分析，自然言語処理，Vol.23，No.1，pp.59-86 (2016).
[2] 堀井朋，荒木雅弘：類型毎の検出手法の組合せによる雑談発話破綻検出の検討，人工知能学会研究資料，pp.33-36 (2015).
[3] 阿部元樹，梅井良太，狩野芳伸，綱川隆司，西田昌史，西村雅史：音響情報を利用した音声対話システムにおける破綻検出，言語・音声理解と対話処理研究会，Vol.81，pp.102-103 (2017).