マルチモーダル情報を用いた音声対話システムにおける対話破綻検出

情報科学科　秋水 紫苑　　指導教員：入部 百合絵

1　はじめに

近年対話システムが身近なものになっているが, 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. 一方で, 対話破綻の検出を行うことができれば, 対話シナリオを切り替えて対話破綻を回避する, 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, 言語情報や音響情報からの破綻検出が行われている[1]. 本研究では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をそれぞれ 42 発話ずつ得た.

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特定した. 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した.

3.1　音響的特徴量の抽出

対話音声から 384 次元の音響的特徴量を抽出した. 抽出した特徴量のうち, 162 次元の音響的特徴量に t 検定による有意差が認められた. 紙面上の関係で, 有意差が確認された Zero-crossing rate/Voice Probability/F0 のみ考察を記す. 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挙げられる. 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. F0 は声の高さを示しているため, 破綻後は声の高さが低くなる傾向が示唆された.

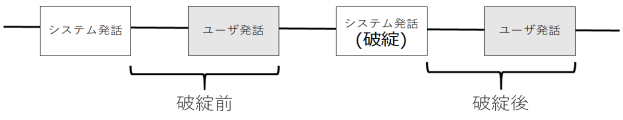

図 1：破綻前後の音声の分析範囲

表 1：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

3.2　視線情報の抽出

収録中, 被験者は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, 視線情報の抽出間隔は 40fps とした. 収録時の都合上, 時間情報の対応がとれない視線情報があったため, 分析にはそれを除く 29 セッション分の音声データを用いた. また, 分析範囲はシステム発話開始からユーザ発話開始までとした. 座標値を用いて, 以下の 3 つの方法により視線情報を算出した.

Ⅰ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の変化量の最大値
Ⅱ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
Ⅲ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, ⅠからⅢの破綻後で求めた視線情報のフレーム間隔を 5 だけではなく 10, 15 に変更した値もそれぞれ求めた.

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. 結果を表 1 に示す.

表 1 から, 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. しかし, 破綻を非破綻と誤認識してしまう傾向があった. 破綻には様々なパターンがあるのに対し, 本研究では区別なく検出する方法を試みたため, 様々な破綻要因に適応化した検出方法が必要である.

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 90%以上の精度で破綻検出を実現することができた. 今後の課題は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] 東中竜一郎他：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).