顔の表情変化と音響情報に基づいた対話破綻検出

情報科学科　鈴木 海斗　　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となってきている．しかし，システムの発話には文として意味の通らない応答や対話の流れに沿わない発話が多く，対話破綻が頻発している．対話破綻を検出する技術が実現されれば，対話破綻の回避を行う，または対話破綻をしたとしても破綻から回復するような対話戦略を立てることができる

これまでに言語情報や音響情報を用いた対話破綻検出の研究が進められている [1]．しかし，人間 - システム間での対話動画を観察すると，システムの対話破綻後にユーザの表情に変化が現れていることが多く，顔の表情変化も対話破綻検出に有効であると考えられる．顔の表情変化と音響の一方のみに特徴が現れることも考えられるため，本研究では顔の表情変化に着目するとともに，音響情報を併用した対話破綻の検出を行う．

2　使用した雑談対話動画

本研究ではシステムのふりをした人間が，ユーザと対話する WoZ (Wizard-of-Oz)法を用いて収録された雑談対話をもとに，顔表情と音響情報を解析する．被験者は 7 人である． 2 人のアノテータがシステム発話に対し破綻/非破綻のタグ付けを行った結果，破綻とタグ付けされた発話，非破綻とタグ付けされた発話はそれぞれ 54 発話ずつであり，本研究では合計 108 発話を使用する．

3　対話破綻検出に用いる特徴量

システム発話後のユーザから音響的特徴量と顔特徴量をそれぞれ抽出する．そして，ラッパー法により次元圧縮された特徴量をもとにシステムの対話破綻を検出する．

3.1　顔表情の変化を示す特徴量

機械学習ライブラリ Dlib を用いて対話動画から顔の特徴点を抽出する．表情表出時には目や口といった特定の点が動くと考えられるため，特徴点間の距離や面積といった顔特徴量 112 次元を抽出した．抽出した特徴量に対して特徴量選択を行った結果，目と眉の距離，鼻と口角の距離，目の見開きといった 6 次元の特徴量が選択された．システム発話が破綻した場合，驚く，困惑する，笑うといった反応がみられるため，このような特徴量が選択されたと考えられる．

3.2　音響的特徴量

システム発話終了からユーザ発話終了までを分析区間とし，音響解析ツール OpenSMILE を用いて対話音声から 384 次元の音響的特徴量を抽出した．抽出した特徴量に対して特徴量選択を施した結果，MFCC (Mel-Frequency Cepstral Coefficients)や ZCR(Zero-crossing rate)など 10 次元の特徴量が選択された．MFCC が選択された理由として，破綻をした場合は暗く返答するといった声質の変化が挙げられる．また，システム発話が破綻した場合は笑い声や言い淀みが発生しやすいことから ZCR に変化が現れたのだと考えられる．

3.3　音響的特徴量と顔特徴量

音響的特徴量と顔特徴量の両方を用いた 496 次元の特徴量に対して特徴量選択を行った結果，MFCC，ZCR，顔の移動距離といった 11 次元が特徴量として選択された．

表 1：分析区間による比較結果

分析区間	Accuracy	Precision	Recall	F-measure
I	0.694	0.706	0.667	0.686
II	0.796	0.767	0.852	0.807
III	0.852	0.852	0.852	0.852
IV	0.685	0.685	0.685	0.685
V	0.676	0.679	0.667	0.673
VI	0.787	0.782	0.796	0.789

表 2：特徴量毎の識別結果

特徴量	Accuracy	Precision	Recall	F-measure
音響	0.750	0.787	0.685	0.733
顔	0.731	0.778	0.648	0.707
音響＋顔	0.852	0.852	0.852	0.852

4　評価結果

4.1 分析区間の比較

対話破綻後に顔の表情変化が現れる適切な区間を明らかにするために，顔の表情変化を抽出する区間を変更し，識別結果を比較した．結果を表 1 に示す．表情の特徴量を抽出した分析区間はシステム発話終了後を対象にしており，表の Ⅰ ～Ⅴは終了後の分析区間を 1 秒刻みで変更した場合（1 秒間～5 秒間），Ⅵはシステム発話終了直後からユーザが発話を開始するまでを分析区間とした．

表 1 より分析区間をシステム発話終了後 3 秒間としたⅢの検出率が最も高い結果となった．この結果より，対話破綻により生じる表情変化は破綻発話終了から 3 秒の間に表れやすいことが示唆された．

4.2 音響的特徴量と顔特徴量の比較

次に特徴量毎の比較実験を行い，識別結果を表 2 に示す．本研究では複数の識別器の中で最も検出率の高かった Random Forest を採用し，10-分割交差検証で評価した．

表 2 より，音響的特徴量と顔特徴量をそれぞれ単独で用いる場合よりもそれらを組み合わせることによって，識別率が向上することが明らかとなった．しかしながら，被験者や破綻の種類により，顔や音響の変化の仕方や反応が現れる区間が異なる可能性も考えられるため，それらの違いを考慮した対話破綻の検出手法が必要である．

5　おわりに

本研究では，システム発話の破綻と非破綻に対するユーザの反応を比較分析することで，対話破綻検出に有効な音響的特徴量と顔特徴量を抽出し，対話破綻検出を行った．実験結果より，85%の精度でシステム対話の破綻を検出することができた．今後の課題として，被験者や破綻の種類による反応の違いを考慮した対話破綻検出が挙げられる．

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).