6M-04

非言語音響情報を利用した対話破綻検出に関する検討

阿部 元樹† 栂井 良太‡ 綱川 隆司†‡ 西田 昌史†‡ 西村 雅史†‡
静岡大学 情報学部† 静岡大学 総合科学技術研究科‡

1. はじめに

近年，雑談対話システムが注目を集めており，当該分野の研究が積極的に行われている[1]．しかし，ドメインが限定されない雑談対話は制御自体が難しい．加えて音声対話の場合は音声認識誤りも生じることから，結果としてシステムがユーザに対して不適切な応答を返してしまう「対話破綻」という状況が生じやすい．一方，事後的ではあるが，システム自体が対話破綻状況を自動検知できれば，素早いリカバリが可能になると期待される． すでに，「対話破綻検出チャレンジ [2][3]」といった研究課題が提唱され，多くの関連研究が行われているが，これらはテキストベースの対話を対象としたものであった．

本研究では対話破綻時に生じると考えられる，ユーザ音声の音響的変化に着目する．先行研究[4]で収集した雑談音声対話データを用い，テキストログと録音の 2 種類のデータに対して対話破綻に関する主観評価を行って，両者の差異を分析した．また，主観評価結果を正解データとし，ポーズ長や発話長などの 10 個の音響的素性を用いた破綻検出器の性能評価も行った．

2. 対話破綻評価実験

2.1. 対話収録[4]

雑談音声対話データの収集は Wizard of Oz(WoZ)法による雑談音声対話システムを用いた．被験者 10 人に対し，各人，1セッションが 15 ターンからなる対話を 3 セッション，合計 450 ターン分の発話（対話音声の録音及びその書き起しテキストログ）の収集を行った．

2.2. アノテーション

3 名のアノテータが，表 1 に示す破綻評価基準に基づいて破綻の分類を行う．収録した 450 ターン分の発話をデータセット A,B（以下それぞれ A,B と表記）に分割し，2 回に分けてアノテーションを実施した．1 回目は A に対してテキストログでの分類を，B に対して録音での分類を行い，数日空けて 2 回目は B に対してテキストログでの分類を，A に対して録音での分類を行った．

表 1：破綻評価基準

分類	基準
○	破綻ではない（当該システム発話の後対話を問題なく継続できる）
△	破綻と言い切れないが，違和感を感じる発話（当該システム発話の後対話をスムーズに継続することが困難）
×	明らかにおかしいと思う破綻した発話（当該システム発話の後対話を継続することが困難）

2.3. 結果

各ターンのテキストログ，録音のそれぞれに対して，3 人のアノテータの評価結果の多数決をとり，それを破綻の評価結果とする．なお，△については破綻側と非破綻側に属する 2 パターンの扱いとした．表 2，表 3 に結果を示す．表 2 から，テキストだけの場合よりも，録音を聴いた場合には破綻と判断する割合が若干減っていることがわかる．また，表 3 より，同じ発話に対するテキストログと録音間の分類結果の一致度を見てみると，3 者共に約 20%の不一致が生じていることが分かる．加えて，各アノテータに対して実験完了後に「テキストログと録音とで評価時にどのような違いを感じたか？」を尋ねたところ，「音声で聞くとユーザがどう解釈して受け取ったかが分かりやすい」や「ユーザの声の調子や淀み具合が判定に影響した気がする」などのコメントが得られ，非言語音響情報が対話破綻検出に関して重要な特徴量になりうることが分かった．

表 2：評価結果
（上段は破綻：非破綻の数，下段は破綻率）

方法	(○+△)：×	○：(△+×)
テキストログ	365:85 (0.189)	306:144 (0.320)
録音	368:82 (0.182)	315:135 (0.300)

A Study on Dialogue Breakdown Detection Using Non-verbal Acoustic Features
Motoki Abe†, Ryota Togai‡, Takashi Tsunakawa†‡, Masafumi Nishida†‡ and Masafumi Nishimura†‡
† Factulty of Informatics, Shizuoka University
‡ Graduate School of Integrated Science and Technology, Shizuoka University

表3：各アノテータのテキストログ，録音での評価の一致度

アノテータ	a	b	c
テキストログ ○：△：×	304:73:73	331:43:76	276:64:110
録音 ○：△：×	324:64:63	334:38:78	248:100:102
一致度	0.756	0.858	0.733

3. 対話破綻検出実験

3.1. 破綻識別器

今回，システム発話直後のユーザ発話に着目し，表4に示す10個の音響的素性を用いて破綻の自動識別を試みる．これらの素性は筆者による予備検討や近年の研究で感情認識に有効ではないかと指摘されたものである[5]．言い淀み・吃音・笑い・息を除く6個の素性は式(1)によって正規化された値を用い，残り4個の素性については，その事象が発生した場合は1，発生しなかった場合は0として処理する．

$$\tilde{x} = (x_i - \bar{x})/SDx \quad \cdots \quad (1)$$

x_i：各ユーザの任意の1発話における素性の値
$\bar{x}$：各ユーザの全45発話における素性の平均値
SDx：各ユーザの全45発話における素性の標準偏差

識別器としてはランダムフォレストを用いた．正解は表2に示した，録音に対する2種類の主観評価結果とし，それぞれ10分割交差検証による性能評価を行った．

表4：音響的素性

素性	概要
ポーズ長	発話を開始するまでの時間
発話長	発話区間の時間長
ポーズ・発話長比	上記2つの比率
無音区間長	発話区間における無音区間長
有音率	発話区間における有音率
ピッチレンジ	第一文節のピッチレンジ
言い淀み	滞った発話
吃音	円滑に発せられない発話
笑い	笑い声
息	呼吸音

3.2. 性能評価結果

適合率・再現率・F値の3つの指標に基づいて性能評価を行った結果を表5に示す．

結果として高い検出性能を示すことは出来なかった．原因として，例えば疑問文応答のように破綻でないにもかかわらず破綻に近い音響的素性の変化が発生するものや，反対に破綻にも関わらず非破綻と同じように音響的変化を伴わないものも多数存在したことが挙げられる．

表5：破綻検出器の性能評価結果

	適合率	再現率	F値
録音 (○+△)：×	0.250	0.024	0.044
録音 ○：(△+×)	0.400	0.119	0.183

4. おわりに

本研究では対話破綻検出に対して音響情報の有用性の検証を行い，非言語音響情報が対話破綻検出に有効な素性であることが確認できた．今後，より有用な音響的素性を見出し，言語情報との組合せによる破綻検出器の構築を行いたい．

謝辞

本研究の一部はJSPS科研費16K01543の助成を受けたものである．

参考文献

[1] 小林峻也，萩原将文：ユーザの嗜好や人間関係を考慮 する非タスク指向型対話システム，人工知能学会論文誌，Vol.31，No.1，pp.DSF-A_1-10（2016）．

[2] 対話破綻検出チャレンジ (https://sites.google.com/site/dialoguebreakdowndetection/)

[3] 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）．

[4] 阿部元樹他：対話破綻検出のための音響情報分析，WiNF2016，A-13X(2016).

[5] Johanna D. Moore, Leimin Tian, Catherine Lai: Word-Level Emotion Recognition Using High-Level Features, Computational Linguistics and Intelligent Text Processing, Vol.8606, pp.17-31 (2014).