E-001

適応的な相槌生成のための複数の識別器の構築

赤井 元紀 † 武田 龍 ‡ 駒谷 和範 ‡

† 大阪大学 工学部電子情報工学科 ‡ 大阪大学 産業科学研究所

1. はじめに

近年，雑談など対話そのものを目的とした非タスク指向型対話システムの必要性が高まっている．そのような対話システムでは，応答の内容だけでなく，そのタイミングも重要である．これまでにもユーザからの音声入力に対して，相槌を打つシステムが構築されている [1, 2, 3].

このような従来研究の問題として，どのようなユーザや状況に対しても同様に相槌を打つ点がある．ユーザによって話すペースは異なるため，相槌を打つかどうかの判定も本来そのユーザの特性に依存するはずである．また，同じユーザでも，例えば現在の話題が盛り上がっているか等の状況に応じて，適切な相槌の頻度やタイミングは異なる．

本研究では，ユーザや状況に対応する複数の識別器を使用することで，適応的な相槌生成を目指す．この際に用いる識別器を，ユーザや状況に対応させて複数構築する．この識別器の学習データは，複数のアノテータにより付与されたラベルの一致数に基づいてラベルを生成することで得る．そもそも相槌を打つかどうかの判定はアノテータの主観に依存することから，付与ラベルの揺れを利用して状況を定める．本稿では，人手で与えた対話の状況を用いて，構築した複数の識別器を選択することで，システムが取るべき行動をより正しく出力できることを示す．

2. 対話の状況に応じた複数の識別器の構築

2.1 複数識別器を用いた発話頻度の対話状況適応

システム実行時における対話の状況に応じた相槌生成について説明する．図 1 に本相槌システムの実行時の処理フローと本研究のスコープを示す．入力は無音区間で区切られた音声信号と対話の状況，出力はそれに対するシステムが取るべき行動である．行動を表すラベルは，対話に不可欠な silent（黙る），nod（相槌を打つ），talk（ターンを取得する）の 3 つである．まず，入力音声信号から特徴量を抽出し，システム内部に持つ複数の識別器へ同時に入力する（図中では 3 つ）．各識別器はそれぞれある対話の状況を想定して構築されており，独立に行動ラベルを出力する．本稿では，対話の状況に対応する識別器からの出力を選択することで，状況に応じた相槌生成を実現する．

複数の識別器は，想定する対話の状況に適合した頻度で行動ラベルを出力するよう学習する必要がある．本研究では，対話の状況として，話の盛り上がりという場の様子（frequently），思考中や言い淀みというユーザの状態（occasionally），どちらでもない（sometimes）を取り上げる．例えば，場の盛り上がりを想定している識別器（frequently）であれば，相槌やシステム発話をより多く生成し，ユーザ発話を促すことが期待される．反対に，ユーザの思考中や言い淀みを想定している識別器（occasionally）であれば，発話間での不要な相槌を避け，聞き役に徹するべきである．つまり，frequently 用の識別器は nod や talk という行動ラベルを一番出力しやすく，occasionally 用の識別器は一番 silent を出力しやすいように学習されるべきである．本研究では学習用の行動ラベル生成を工夫することで，所望の識別器を学習する．

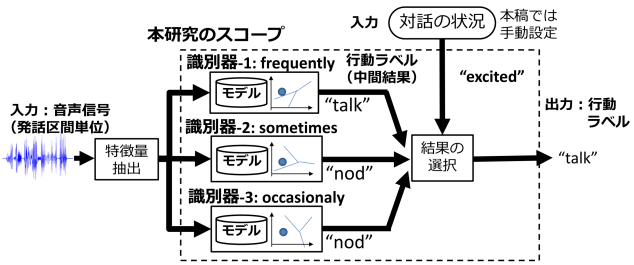

図 1: 実行時の処理フローと研究のスコープ

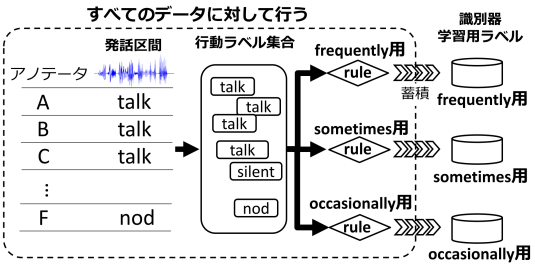

図 2: 各識別器用の学習用行動ラベル生成

2.2 発話頻度を想定した学習用行動ラベルの生成

本研究では，複数のアノテータにより付与された行動ラベルを用いて，各発話頻度を想定した識別器の学習用行動ラベルをルールベースで生成する．図 2 に行動ラベル生成法の全体像を示す．まず，学習用入力データ中のある音声信号に対して，複数のアノテータ（e.g., A, B, ..., F）により行動ラベルの付与を行う．ここでは，その発話までの対話の流れを元にアノテータが最も自然だと思うラベルを付与する．通常，付与ラベルは必ずしも一致せず，必ず揺れが生じる．次に，発話頻度を想定した各ルールに，これらが付与されたラベル集合を入力する．このルールは，ラベル集合の情報をもとに，1 つの行動ラベルを選択するように設計される．これらを学習用入力データすべてに対して行うことで，各識別器用の行動ラベルを収集する．

ルールは，付与されたラベルの頻度と想定する発話頻度を反映した閾値を用いて設計される．発話頻度 $x \in \{\text{frequently}, \text{sometimes}, \text{occasionally}\}$ を想定した学習用行動ラベルは次のルールで決定される．(1) talk のラベル数が $T_{x,\text{talk}}$ 以上なら talk，(2) nod または talk のラベル数が $T_{x,\text{nod}}$ 以上なら nod，(3) いずれも満たさないなら silent とする．ここで，閾値 $T_{x,\text{talk}}, T_{x,\text{nod}}$ は発話頻度に依存するパラメータである．各対話の状況を想定して閾値を調整することで，全体として傾向の異なる学習用の行動ラベルが得られる．

Constructing Multiple Classifiers towards Adaptive Backchannel Generation: *Genki Akai, Ryu Takeda, and Kazunori Komatani (Osaka Univ.)*

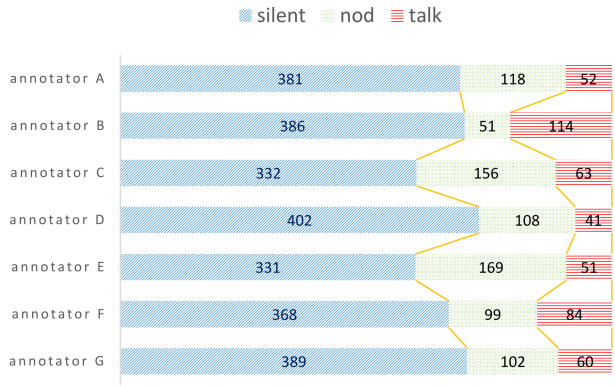

図 3: 各アノテータが付けた行動ラベルの数

表 1: 学習用の行動ラベルの数の合計

	silent	nod	talk
frequently	2156	1343	358
sometimes	2583	860	414
occasionally	2793	553	511

3. 評価実験

3.1 データとラベルの収集

研究室の学生 7 名から，システムとの対話音声を収集した．データ収集は (1) システムが事前に用意した質問文の音声を出力，(2) ユーザがそれに対し回答，という手順を 10 個の質問文に対し繰り返すことにより行った．質問文は「あなたは普段朝食はご飯とパンのどちらを食べていますか? 理由とともにお答えください.」のような，ユーザがなるべく多く発話しやすいような文とした．

収録したデータへの行動ラベルの付与を依頼した．アノテータは，データ収集時と同じ学生 7 名である．本実験では，音声を発話区間ごとに分割するために Julius 付属の adintool[*1] を用いた．分割された発話区間の数は 551 であった．図 3 に各アノテータが付けた行動ラベルの数を示す．行動ラベルの数の合計は，silent が 2589，nod が 803，talk が 465 となった．

提案手法で学習用の行動ラベルを生成するための閾値 $T_{x,\mathrm{talk}}$ および $T_{x,\mathrm{nod}}$ は人手で決定した．それぞれ $x =$ frequently, sometimes, occasionally の順に 4, 3, 2 および 2, 3, 4 とした．

評価実験における対話の状況ラベルは筆頭著者が自身の感覚に基づいて付与した．対話の状況ラベルは発話頻度の多い順に excited，normal，thinking の 3 種類とした．対話の状況ラベルの数は excited が 44，normal が 470，thinking が 37 となった．以降では，excited と frequently，normal と sometimes，thinking と occasionally がそれぞれ対応するとみなして実験を進める．

3.2 実験条件

発話区間から抽出する特徴量には INTERSPEECH 2013 Computational Paralinguistics Challenge [4] で使用された，F0，パワー，MFCC，ゼロ交差率などをもとに導出された 6373 次元の韻律に関する特徴量を用いた．これらの特徴量は openSMILE[*2] を用いて抽出した．識別器には Weka[*3] の SVM を用いた．

識別器を 1 つだけ構築した場合 (ベースライン) と識別器を複数構築する場合 (提案手法) のそれぞれについて正解率の平均で評価する．まず，学習セットを基に識別器の学習を行った．次に，発話区間ごとに推定を行い，得られた行動ラベルをテストセットと比較した．そして，行動ラベルが一致した場合正解とした．ベースラインの学習セットはアノテータ 1 名の行動ラベルとした．提案手法の学習セットはアノテータ 6 名の行動ラベルから生成した学習用の行動ラベルとした．テストセットは共に学習セットとは別のアノテータ 1 名の行動ラベルとした．本実験では，考えられる学習セットとテストセットの組み合わせを全て試す．つまり，ベースラインは全 $6 \times 7 = 42$ パターンについて，提案手法は全 7 パターンについて正解率を計算し，その平均で評価する．

表 1 に提案手法で全 7 パターンのそれぞれについて得られた，対話の状況ごとの学習用の行動ラベルの数の合計を示す．silent は occasionally が最も多く，nod は frequently が最も多いことから各対話の状況に適した学習用の行動ラベルが得られたことがわかる．

表 2: 全体の正解率の平均 (%)

マジョリティーベースライン	ベースライン	提案手法
67.1	73.9	78.3

表 3: 識別器ごとの正解率の平均 (%)

frequently	sometimes	occasionally	全体
64.3	78.4	93.8	78.3

3.3 実験結果と考察

ベースラインと提案手法の正解率の平均を表 2 に示す．ここで，マジョリティーベースラインとは全ての発話区間に対し silent を出力した場合の正解率である．提案手法を用いるとベースラインより約 4 ポイント正解率が上昇しており，提案手法が有効だとわかる．

また，提案手法における識別器ごとの正解率の平均を表 3 に示す．sometimes と occasionally の正解率の平均がベースラインより高いことから，対話の状況を適切に与えることができれば正解率が上がることがわかる．ただし，上記の結果は学習セットとテストセットで同じ入力が与えられている点で問題がある．今後，発話区間に関して cross validation を行う予定である．

今後は，ラベルの一致数の閾値の最適化や対話の状況の自動推定などを行うことで，システムの適応的な相槌の実現を目指す．

参考文献

[1] 大須賀 智子，堀内 靖雄，西田 昌史，市川 熹: 音声対話での話者交替/継続の予測における韻律情報の有効性．人工知能学会論文誌, Vol.21, No.1, pp.1-8, 2006.

[2] 小川 翼，伊藤 敏彦: リアルタイム発話継続/交替予測システムの構築．HAI シンポジウム 2014，Vol.43，G-12，2014.

[3] 西村 良太，中川 聖一: 応答タイミングを考慮した音声対話システムとその評価．情報処理学会研究資料，Vol.2009-SLP-077，No.22，2009.

[4] B. Shuller et al.: The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conict, Emotion, Autism. INTERSPEECH, pp.148-152, 2013.

[*1] http://julius.sourceforge.jp/juliusbook/ja/adintool.html
[*2] http://audeering.com/technology/opensmile/
[*3] http://www.cs.waikato.ac.nz/ml/weka/