人工知能学会研究会資料
SIG-SLUD-B503-09

傾聴対話システムのための言語情報と韻律情報に基づく多様な形態の相槌の生成

Generating a Variety of Backchannels Based on Linguistic and Prosodic Features for Attentive Listening Agents

山口 貴史[1*] 井上 昂治[1] 吉野 幸一郎[2] 高梨 克也[1]
Nigel G. Ward[1,3] 河原 達也[1]
Takashi Yamaguchi[1] Koji Inoue[1] Koichiro Yoshino[2] Katsuya Takanashi[1]
Nigel G. Ward[1,3] Tatsuya Kawahara[1]

[1] 京都大学 [2] 奈良先端科学技術大学院大学 [3] テキサス大学エルパソ校
[1] Kyoto University [2] Nara Institute of Science and Technology [3] University of Texas at El Paso

Abstract: There is a growing interest in conversation agents which conduct attentive listening. However, the current conversation agents always generate the same or limited form of backchannels every time, giving a monotonous impression. We have investigated generation of a variety of backchannels according to the dialogue context using the corpus of counseling dialogue. At first, we annotate all acceptable backchannel form categories considering the arbitrary nature of backchannels. Then, we conduct machine learning to predict a backchannel form from the linguistic and prosodic features of the preceding context. This model outperformed the method which always outputs the same form of backchannels and also the method which randomly generates backchannels. Finally, subjective evaluations by human listeners show that the proposed method generates backchannels more naturally giving a feeling of understanding and empathy.

1 はじめに

近年、タスク指向型対話システムに加えて、雑談型対話システムも検討されるようになってきている [1]。雑談型対話システムの機能の一つにユーザの話を聞く傾聴がある。傾聴とは話し手の話に共感を示しつつ、話し手がより多く話せるように手助けをして話を聴くことである [2]。音声対話システムが傾聴を行うことにより、入院患者や高齢者の話し相手となること [3] や、ユーザの話したい、話を聞いてもらいたいといった欲求を満たすこと [4] が期待されている。傾聴を行う際に重要となる対話行為としては、話し手の発話に対して「相槌をうつ」「質問をする」「共感を示す」などが挙げられる。これらのうち、質問や共感では相手の発話の十分な認識・理解が必要であるが、これは技術的に容易ではない。これに対し、相槌は先行発話の韻律や節末のパターンに基づいて生成できる可能性がある。

相槌は会話を円滑に進める上で非常に重要な要素である。相槌は話し手の話を「聞いていること」、「理解していること」、「共感していること」などを表す役割がある [5]。また、相槌をうつことによって会話全体のリズムを生みだすこともできる。近年は相槌をうつ対話システム [6, 7, 8, 9] や、相槌をうつタイミング [10, 11, 12, 13, 14, 15, 16] について研究されている。しかし、多くのシステムはあらかじめ決められた形態の相槌のみをうっており、その形態のバリエーションは乏しい。聞き手のうつ相槌が常に同じ形態の相槌（例えば「うん」）のみであると、話し手は相手が自身の話を聞いているのか、理解してくれているのかと不安や不満を感じる。さらに、会話のリズムに不自然さや単調さが生じる。これに対して、ユーザの話を傾聴するようなシステムにおいては、文脈に応じて多様な形態の相槌をうつ必要がある。そこで、本研究ではそのような相槌の生成を目標として、相槌形態の分析・予測・生成に取り組んでいる [17, 18]。

2 対話コーパスと相槌の認定

本研究では、上里ら [19] の研究で収録された相談対話を用いる。対話のテーマは「日常の簡単な悩みや困りごと」であり、各セッションの対話は、話し役 1 名、聞き役 1 名で行っている。聞き役はスクールカウンセラー 2 名、話し役は大学生 8 名で、合計 8 対話が収録されている。対話時間は 20〜30 分である。この対話は大きく前半と後半に分かれており、前半はカウンセラが相談者の話を聞き、後半は相談者にアドバイスなどをしている。そこで、後述のアノテーションや予測・生成においては、前半部分を主に用いる。（モデルの学習には全データを用いる）

*連絡先：京都大学情報学研究科知能情報学専攻 河原研究室
京都市左京区吉田本町
E-mail: takashi@ar.media.kyoto-u.ac.jp

2.1 発話単位と相槌の認定

本研究では発話単位として、間休止単位（IPU）と節単位の 2 つを用いる。

間休止単位（IPU）：間休止単位（IPU）は、笑いや咳などの「非言語発話」を除いた 200ms 以上の休止で挟まれた区間ごとに設定される発話単位である。

節単位：節単位は『日本語話し言葉コーパス』（CSJ）で定義された節境界 [20, 21] を区切りとした発話単位である。節境界には、その直後の構造的な切れ目の大きさという観点から切れ目の強さの順に絶対境界、強境界、弱境界の 3 種類があるが、発話単位の認定では、これら 3 種類を区別せずに扱う。

相槌については様々な定義がなされているが、本研究ではメイナード [22] の「話し手が発話権を行使している間に聞き手が送る短い表現」を採用し、相槌の形態ごとの分類には伝 [23] による、「うん」や「ふんふん」といった促しや受容を表す応答系感動詞と、「あー」や「はー」といった興味や関心・共感を表す感情表出系感動詞の 2 つを用いた。また分析対象とする相槌には、先行発話末 500ms 以内でオーバーラップするものも含めた。ここでは、応答系の「うん」「ふん」を同種として扱い、その繰り返し回数によって、応答系 1 回、応答系 2 回、応答系 3 回以上というカテゴリに分類した。また、「あー」、「はー」、「へー」は応答系と振る舞いが違うため [24]、感情表出系として一つのカテゴリとした。つまり、対象とする相槌を応答系 1 回の相槌、応答系 2 回の相槌、応答系 3 回以上の相槌、感情表出系の相槌の 4 つにカテゴリ化した [17]。

2.2 相槌アノテーション

一般に、ある発話に対してうつことのできる相槌形態は一つのものだけに限られない。たとえば、強い同意を示すためには「あー」だけでなく「うんうんうん」なども可能であるかもしれない。さらに、相槌をうつことができる箇所も個人によってさまざまである。そのような相槌の任意性に対処するために、相槌カテゴリを複数の作業者によってアノテーションし、うつことができる相槌形態を拡張することとした [18]。具体的には、カウンセラが相談者の話を聞いて、頻繁に相槌をうっている前半部分に対して、次の 2 つの時点でこのアノテーションを行った。アノテーションした対話は、上里ら [19] の実験で使用した 4 対話である。

(A) 相槌がうたれている節末：先行発話の節末でカウンセラが実際に相槌をうった箇所である。ここでは、どの相槌カテゴリがうてるかどうかのアノテーションを行った。

(B) 相槌がうたれていない節末と、節末に該当しない IPU 末：節末かつ IPU 末で相槌が生起している箇所は (A) に含まれているため、(A) と (B) は排他的である。ここでは、応答系 1 回の相槌か応答系 2 回の相槌が「うてる」か「うてない」かの 2 値でアノテーションした。

ここでアノテーションされた相槌カテゴリは、3 章と 4 章での予測実験の際の正解ラベルとして用いる。

3 言語と韻律情報に基づく相槌形態の予測

先の報告 [17] において、話し手の発話の特徴から、相槌形態を予測できるか調べるために、先行発話の持つ特徴と後続する相槌との関係を分析した。[17] では、話し手の発話の節境界 [20, 21] にうたれる相槌に焦点を当て、先行発話の節境界の種類や構文構造と 2 章で定義した相槌カテゴリとの関係について分析し、節境界の種類は感情表出系と応答系の区別に、構文構造は応答系 1 回と 2 回の区別に、それぞれ有用であるという知見を得た。

こうした知見に基づき、先行発話から得られる特徴を用いて相槌形態の予測を機械学習により行う。ただし、ここでは、相槌をうつタイミングは所与とし、相槌がうたれている節末のみを予測対象位置とする。先の報告 [18] では、[17] の分析に基づいて、先行発話から得られる言語的特徴から 2 章で定義した相槌カテゴリを予測した。本稿では、素性に韻律的特徴を加えて予測を行う。

3.1 予測に用いる韻律的特徴

相槌のタイミングに関する多くの研究 [13, 14] で先行発話末から得られる F0 やパワーなどの韻律情報が用いられている。そこで、このような韻律情報から得られる特徴についても検討し、言語情報のみの場合と比較する。韻律に関する情報は言語情報に比べて計算コストが低く、リアルタイムで計算することが可能である [13]。具体的には、西村ら [13] にならい、先行発話末 150ms から得られる F0 とパワーについて 1 次回帰係数やレンジを特徴に含めた。F0 とパワーは、Tsanas ら [25] の報告で抽出性能が最も有用とされる STRAIGHT [26, 27] に基づく帯域毎に正規化された自己相関と瞬時周波数を利用した NDF [28] を用いて 5ms ごとに抽出した。ただし、F0 は 10 を底とする対数 F0 とした。さらに、発話長や先行発話末 150ms の話速と軋みさ（Creakiness）も特徴として検討した。発話長は節の先頭から節末までの発話の時間である。軋みさ（Creakiness）は末尾 150ms の間で、F0 におけるジッタが起こった回数をカウントしたものである。

3.2 実験条件

先行発話の韻律的特徴と言語的特徴から相槌カテゴリを機械学習により予測する。相談対話 [19] 8 対話のうち、1 対話を評価用、残りの 7 対話を学習用とした交差検定を行った。ただし、評価データは 2.2 節の (A) で追加形態のアノテーションを行った 4 対話（カウンセラが頻繁に相槌をうっている対話の前半部分）とする。分類器には ロジスティクス回帰（Liblinear [29]）を使用した。

評価には、適合率と再現率、これらの調和平均である F 値を使用し、各対話ごと、各カテゴリごとで算出した。3 章で述べたように、評価の際に、元形態だけでなく、2.2 節の (A) のアノテーションで追加形態と認定

表 1: 相槌形態の予測結果（相槌カテゴリごと）

		応答系 1 回	応答系 2 回	応答系 3 回以上	感情表出系	平均
提案モデル（言語＋韻律的特徴）	適合率	0.982 (54/55)	0.826 (19/23)	0.488 (63/129)	0.654 (53/81)	0.656 (189/288)
	再現率	0.460 (23/50)	0.484 (31/64)	0.830 (83/100)	0.703 (52/74)	0.656 (189/288)
	F 値	0.626	0.611	0.615	0.678	0.656
提案モデル（言語的特徴）	適合率	0.980 (48/49)	0.714 (15/21)	0.468 (65/139)	0.646 (51/79)	0.622 (179/288)
	再現率	0.400 (20/50)	0.391 (25/64)	0.840 (84/100)	0.676 (50/74)	0.622 (179/288)
	F 値	0.568	0.505	0.601	0.660	0.622
ランダム生成	適合率	0.777	0.750	0.414	0.389	0.535
	再現率	0.334	0.394	0.707	0.521	0.535
	F 値	0.464	0.513	0.521	0.458	0.535

された相槌カテゴリも正解としている。そのため、再現率と再現率は、分子を元形態か追加形態かのいずれかが当たった方としている。ある予測位置において、元形態ではなく、その追加形態を予測した場合には、再現率において追加形態ではなく、元形態に対する正解としてみなしている。

使用する特徴を以下にまとめる。ただし、言語的特徴は [18] で使用した特徴に、さらに二つ前の節境界の種類と相槌カテゴリを加えている。

言語的特徴：節境界の種類、末尾語、末尾語の品詞、節境界ラベル、文節数、構文木の深さ、係り受けの数、一つ前の節境界の種類、一つ前の相槌カテゴリ、二つ前の節境界の種類、二つ前の相槌カテゴリ

韻律的特徴：発話長、対数 F0 の 1 次回帰係数、パワーの 1 次回帰係数、対数 F0 のレンジ、パワーのレンジ、話速、軋みさ（Creakiness）

なお、比較のためのベースラインとしては、学習データにおける相槌カテゴリの度数分布にしたがって相槌形態をランダムに決める「ランダム」を採用した。ただし、ランダム生成の結果 (適合率・再現率・F 値) は 1000 回試行の平均とした。

3.3 実験結果

予測結果を表 1 に示す。ランダム生成の結果（適合率・再現率・F 値）は 1000 回試行の平均であるため、表中に個数は載せていない。韻律的特徴を加えることによって、すべてのカテゴリで予測精度が向上した。この結果から、今回用いた韻律情報は相槌形態を区別するのに有効であることがわかる。

相槌形態ごとに詳しく見ると、特に応答系 1 回と 2 回の F 値が大きく向上していることがわかる。また、応答系 3 回の適合率と応答系 1 回と 2 回の再現率が上がっていることから、韻律情報を加えることによって、応答系 3 回と予測されていたものが応答系 1 回や応答系 2 回に予測されていると推測される。

これらの各相槌カテゴリに対応する先行発話末の韻律に応じて相槌が使い分けられている可能性がある。本研究で扱っている相談対話は、話し手の各ターンが長い。各発話末の韻律的特徴に応じた応答系相槌により、話し手は長くターンを保持することができ、自身の相談内容を十分に話すことができると考えられる。

4 多様な形態の相槌の生成

前章では、多様な形態の相槌の中から適しているものを予測できるかを確かめるため、対象となる生起位置をカウンセラが相槌をうった節末のみに限定し、そこで 4 クラス分類を行った。しかし、システムによる相槌の生成のためには、相槌を「うつ」か「うたない」かも含めた予測を行う必要がある。そこで本章では、節末であるかどうかや、相槌の有無にかかわらず、相槌が生起可能なすべての候補位置で「応答系 1 回の相槌」、「応答系 2 回の相槌」、「応答系 3 回以上の相槌」、「感情表出系の相槌」、「うたない」の 5 つのカテゴリのどれかの予測を行う。候補位置としては、音響的区切りである IPU 末を用いる。すなわち、カウンセラによる相槌の生起に関わらず、すべての節末と IPU 末を予測対象位置とする。

4.1 予測モデル

本研究では、5 クラスのどれかを予測するために以下の二通りのモデルを検討する。

5 クラス分類モデル：5 クラスのいずれかを予測するモデル「5 クラス分類モデル」は、第 5 章の 4 クラスの予測モデルに、「うたない」を追加して 5 クラスの予測に拡張したものである。

2 クラス分類モデル：4 つの相槌カテゴリそれぞれについて、そのカテゴリを「うつ」か「うたない」かの 2 クラスを予測するモデルを組み合わせた「2 クラス分類モデル」では、出力する相槌カテゴリを 2 段階で選ぶ。まず、先行発話から得られる特徴をそれぞれの相槌カテゴリのモデルに与え、それぞれの相槌カテゴリを「うつ」確率値と「うたない」確率値を求める。次に、各カテゴリの「うつ」確率値が閾値以上なら、それを「うてる相槌カテゴリ」とみなし、どのカテゴリも閾値未満なら「うたない」を予測結果として出力する。ここで、「うてる相槌カテゴリ」が 1 つならば、その相槌カテゴリを予測結果とする。また、「うてる相槌カテゴリ」が 2 つ以上ならば、その中で最も確率値が高いものを予測結果として出力する。閾値は、0.5 から順に下げていき、「相槌をうつ」と「うたない」の頻度分布が、テストデータの頻度分布に最も近くなったときの値を採用する。その結果、閾値は、0.275 とした。

表 2: 相槌の生成及び形態の予測結果（相槌カテゴリごと）

		応答系 1 回	応答系 2 回	応答系 3 回	感情表出系	うたない	平均
5 クラス分類	適合率	0.676 (25/37)	0.750 (15/20)	0.293 (27/92)	0.357 (5/14)	0.648 (468/722)	0.610 (540/885)
	再現率	0.189 (20/106)	0.225 (20/89)	0.311 (38/122)	0.173 (13/75)	0.911 (449/499)	0.610 (540/885)
	F 値	0.295	0.346	0.302	0.223	0.757	0.610
2 クラス分類	適合率	0.657 (67/102)	0.820 (41/50)	0.333 (64/187)	0.342 (20/56)	0.769 (377/490)	0.643 (569/885)
	再現率	0.311 (33/106)	0.382 (34/89)	0.672 (82/122)	0.467 (35/75)	0.775 (385/499)	0.643 (569/885)
	F 値	0.422	0.521	0.454	0.405	0.772	0.643
ランダム生成	適合率	0.510	0.464	0.156	0.125	0.584	0.431
	再現率	0.170	0.195	0.281	0.232	0.597	0.431
	F 値	0.253	0.272	0.156	0.160	0.591	0.431

このとき、各発話に対して「うてる」と判断された相槌カテゴリの個数は平均 2.3 個であった。

4.2 実験条件

先行発話の特徴から、「うたない」を含めた 5 つの相槌カテゴリのいずれかを機械学習により予測する。実験条件は前章の 4 クラス予測と同様である。ただし、本章では予測対象位置をすべての節末と IPU 末に拡大しているため、2.2 節の (A) だけでなく (B) のアノテーションも用いる。

使用する特徴は 3.2 節と同様である。ただし、本実験では、3.2 節と比べて予測位置が増えていることから、文脈情報をより細かく得るために一つ前の IPU の末尾語の品詞を加えた。

比較のためのベースラインとしては、3 章と同じく、ランダムに決める方法を用いる。

4.3 実験結果

予測結果を表 2 に示す。提案手法である 5 クラス分類モデルと 2 クラス分類モデルともに、ランダム生成と比べて有効に予測できていることがわかる。この結果は、提案手法の有用性を示すものである。

次に、提案手法同士の比較では、すべてのカテゴリで 2 クラス分類モデルの方が結果がよいことがわかる。特に 5 クラス分類モデルは、再現率が「うたない」以外の 4 カテゴリで 2 クラス分類モデルよりも大幅に低く、そもそも相槌をうつと予測できている箇所自体が少ない。これに対して、2 クラス分類モデルはこの点が大幅に改善されている。

さらに、相槌形態ごとに詳しく見ると、2 クラス分類モデルは 5 クラス分類モデルに比べて、適合率が高いことから、応答系 1 回や応答系 2 回を適切な箇所で予測できていることがわかる。次に、応答系 3 回や感情表出系では、いずれの手法も適合率が低い。これは、節末でない IPU 末に対して、(B) のアノテーションでは応答系 1 回や応答系 2 回が「うてる」かどうかだけをアノテーションしたことによるためと考えられる。なお、ランダム生成では、感情表出系の適合率が非常に低い。感情表出系は出現の頻度がそれほど多くなく、そのパターンも限定されているためと考えられる。

5 音声を用いた印象評価実験

5.1 音声サンプル

本章では、前章の予測に基づいて生成された相槌について、音声データを用いた印象評定実験による評価を行う。

4 章で最も結果がよかった 2 クラス分類モデルの予測結果に基づいて相槌を生成し、音声データを作成した。聴取する被験者は２０代の男女９名（男性５名 女性４名）である。各被験者はこの音声データを聴取して、相槌の印象を評価する。

相槌を挿入する位置は 4 章と同様、節末もしくは IPU 末である。比較のため、相談対話の音声に対し、以下の 3 条件の相槌音声データを挿入した提示刺激を作成した。相槌音声はアンドロイド ERICA の TTS（HOYA 音声合成ソフトウェア VoiceText ERICA） [30] のものを使用した。

1. ベースライン条件：ランダムに生成されたカテゴリ（4 章と同じ）
2. 予測条件：提案手法の 2 クラス分類モデルにより生成された相槌カテゴリ
3. カウンセラ条件：もとの対話収録時にカウンセラがうったのと同じ相槌カテゴリ（ただし、相槌音声は元のカウンセラのものではなく、他の条件と同様、TTS のものを使用）

相槌音声の形態としては、応答系として「うん」「うんうん」「うんうんうん」、感情表出系では「あー」「はー」を使用する。ただし、予測条件とベースライン条件において、感情表出系をうつ場合には、 [18] のアノテーションにおいて「あー」よりも許容性があるとされた「はー」を使用した。相槌音声は３条件とも同じ音声を使用した。各形態の音声は、相談対話にあうように、なるべく控えめなものを選んだ。

使用した対話音声は 3 章と 4 章で用いた 4 対話から 1 分半〜2 分程度連続した箇所を２箇所ずつ抜粋した合計 8 対話セグメントである。抜粋箇所は、「話の内容がまとまっている」、「カウンセラが比較的話しかけたりせず相槌をうっている」、「カウンセラの相槌をうつ際の癖が現れていない」などを基準とした。

表 3: 被験者による相槌の印象評価結果

評価項目	ベースライン条件	予測条件	カウンセラ条件
Q1：全体を通して相槌は自然でしたか t	**-0.42**	**1.04****	0.79
Q2：全体を通してテンポよく進んでいましたか	**0.25**	**1.29***	1.00
Q3：全体を通して真面目に聞いてくれていると感じましたか	0.25	1.04	1.08
Q4：全体を通して集中して聞いてくれていると感じましたか	0.50	1.29	0.96
Q5： 全体を通して積極的に聞いてくれていると感じましたか	0.63	1.21	1.08
Q6：全体を通して親身に聞いてくれていると感じましたか	0.33	1.25	0.96
Q7：全体を通して理解してくれていると感じましたか	**-0.13**	**1.17****	0.79
Q8：全体を通して関心を持ってくれていると感じましたか	**0.21**	**1.21**	1.04
Q9：全体を通して共感してくれていると感じましたか	**0.13**	**1.04***	0.46
Q10：このカウンセラと話したいと思いましたか	**-0.33**	**0.96****	0.29
		* $p < 0.05$	** $p < 0.01$

5.2 印象評定

実験に際して、抜粋した音声は対話の途中から始まるため、被験者が話の内容を理解しやすくなるよう、各対話セグメントの音声を聞いてもらう前にそこまでの対話内容の概略を口頭で説明した。

8 対話セグメント × 3 条件で合計 24 のサンプルがあるため、各サンプルについて、3 名の異なる被験者から回答が得られるようにする。そのため、対話セグメント間で同じ被験者の組み合わせの重複が最小限になるようにした。また、サンプルを聞く順番も被験者ごとに異なっている。

評価項目は堀口 [5] の相槌の機能を参考にして作成した。被験者はそれぞれのサンプルに対して、これらの評価項目について 7 件法（-3：全くそう思わない 〜 3：非常にそう思う）で評定する。

5.3 評価結果

各被験者の評価値の平均を表 3 に示す。どの項目においてもベースライン条件が最も低い評価値であった。ランダムに生成されるベースライン条件と提案法である予測条件との間に有意差があるかを t 検定で検定したところ、有意水準 1%で項目 Q1、Q7、Q10、有意水準 5%で項目 Q2、Q8、Q9 で有意差が認められた（表の太字）。このように、提案法では、半数以上の評価項目でベースライン条件よりも高い評価を得た。また、予測条件とカウンセラ条件の間でも t 検定で検定したところ、すべての項目で有意差がなかった。この結果から、カウンセラと同程度の評価値を得ていると考えられる。

各項目ごとに考察を行う。まず、項目 Q1 の「自然さ」では、カウンセラ条件や予測条件と比べて、ベースライン条件の評価が大きく低いことから、相槌形態やタイミングがランダムに生成されたものだと不自然な印象を与えることがわかった。また、項目 Q10 の「カウンセラと話したいか」といった項目に関しても、ランダム生成だと、話したいという印象を与えることができないことがわかる。

項目 Q2 の「テンポの良さ」では、予測条件やカウンセラ条件の方がベースライン条件と比べて評価値が非常に高い。ベースライン条件では対象位置で相槌をうつかどうか自体がランダムであるのに対して、予測条件では相槌をうつタイミングがより適切であるため、テンポよく会話が進んでいるという印象を与えていると考えられる。

ここでは、相槌が生起可能なすべての位置で予測しているため、相槌を「うつ」か「うたない」かの判断が被験者の印象に大きく影響を与えると考えられる。実際に、4 章での予測実験で、ランダム生成での「うたない」の適合率や再現率が低いことから、うつべきでない位置でうっていると考えられる。一方、予測条件では、「うたない」の適合率や再現率はベースライン条件と比べれば高いものの、十分に高いとまではいえないが、テンポに関する項目 Q2 の評価はよい。このことから、タイミングに関しては、ある程度以上の精度があれば、必ずしも悪い印象を与えることはないと考えられる。

次に、相槌の機能に関する項目について、項目 Q3〜Q6 の「聞いてくれていると感じるか」といった項目では、ベースライン条件と予測条件で有意な差がみられなかった。このことからは、単に聞いてくれているという印象を与えるだけなら、どのような形態の相槌でも十分であるという可能性もうかがえる。しかし、「理解」や「共感」に関する項目 Q7、Q8、Q9 では、いずれの項目でも予測条件はベースライン条件を有意に上回るとともに、カウンセラ条件とも差がない。

一般に、聞いていないと理解できず、理解できていないと共感できないように、聞き手の理解には深さの度合いがあり、これに応じて聞き手はこの理解の深さを反応の強さによって表すと考えられる [31]。4 章での予測実験でランダム生成での感情表出系の適合率が非常に低かったことから、特に「理解」や「共感」を表す機能を持つ感情表出系が不適切な箇所で出現していたことの影響が大きいと考えられる。

4 章の 2 クラス分類モデルの結果では、「うたない」以外の 4 カテゴリの精度はあまり高くないように見える。しかし、予測された結果は、カウンセラ条件と同程度の評価を得ている。予測精度が十分でなくても、ユーザに対して一定の印象を与えられる可能性がある。

ただし、カウンセラ条件も含めて評価値は必ずしも高くない。その理由として形態以外の要因が考えられる。具体的には、上里ら [19] の研究のように韻律の調整を適応的に行うことや、タイミングについても先行する発話へのオーバーラップを含めて精密に調整することが考えられる。

6 まとめ

本研究では、言語的特徴だけでなく、韻律的特徴も用いて、相槌形態の機械学習による予測を行った。その結果、ランダムに生成する方法と比べて効果的に予測でき、韻律的特徴が相槌形態を区別するのに有効であることが示された。次に、相槌の生起位置の予測も含めた生成を目指して、相槌が生起可能なすべての位置を対象とした予測実験を行った結果、ランダムに生成する方法と比べて有効に予測できることが示された。さらに、このシステムで生成した相槌の音声を用いた印象評価実験を行った結果、ランダムに生成する方法と比較して、相槌の自然さや「理解」「共感」などの表現において有意に高い評価を得ることができた。

今後の課題として、上里ら [19] の知見を用いて、相槌を生成する際の韻律的特徴を調整することにより、相槌生成モジュールを自律型アンドロイド ERICA の音声対話システム [30] に実装することが考えられる。

謝辞：アノテーションと印象評価実験にご協力いただいた皆様に感謝いたします。本研究は JST ERATO 石黒共生ヒューマンロボットインタラクションプロジェクトによる。

参考文献

[1] 河原達也. 音声対話システムの進化と淘汰 : 歴史と最近の技術動向. 人工知能学会誌, Vol. 28, No. 1, pp. 45–51, 2013.

[2] 楡木満生. 積極的傾聴法. 医学教育, Vol. 20, No. 5, pp. 341–346, 1989.

[3] 山本大介, 小林優佳, 横山祥恵, 土井美和子. 高齢者対話インタフェース :『話し相手』となって、お年寄りの生活を豊かに. 電子情報通信学会技術研究報告. HCS, Vol. 109, No. 224, pp. 47–51, 2009.

[4] 目黒豊美, 東中竜一郎, 堂坂浩二, 南泰浩. 聞き役対話の分析および分析に基づいた対話制御部の構築. 情報処理学会論文誌, Vol. 53, No. 12, pp. 2787–2801, 2012.

[5] 堀口純子. コミュニケ-ションにおける聞き手の言語行動. 日本語教育, No. 64, pp. p13–26, 1988.

[6] 下岡和也, 徳久良子, 吉村貴克. 音声対話ロボットのための傾聴システムの開発. 人工知能学会研究会資料. SLUD, 言語・音声理解と対話処理研究会, Vol. 58, pp. 61–66, 2010.

[7] 横山祥恵, 山本大介, 小林優佳, 土井美和子. 高齢者向け対話インタフェース : 雑談継続を目的とした話題提示・傾聴の切替式対話法. 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2010, No. 4, pp. 1–6, 2010.

[8] 大竹裕也, 萩原将文. 評価表現による印象推定と傾聴型対話システムへの応用. 日本知能情報ファジィ学会論文誌, Vol. 26, No. 2, pp. 617–626, 2014.

[9] D. DeVault, R. Artstein, G. Benn, T. Dey, G. E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proc. AAMS*, pp. 1061–1068. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[10] 岡登洋平, 加藤佳司, 山本幹雄, 板橋秀一. 韻律情報を用いた相槌の挿入. 情報処理学会論文誌, Vol. 40, No. 2, pp. 469–478, 1999.

[11] N. Ward and W. Tsukahara. Prosodic features which cue backchannel responses in English and Japanese. *J. Pragmatics*, Vol. 32, No. 8, pp. 1177–1207, 2000.

[12] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech*, Vol. 41, No. 3, pp. 295–321, 1998.

[13] 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. 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2009, No. 22, pp. 1–6, 2009.

[14] N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa. Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems. *J. Japanese Society for Artificial Intelligence*, Vol. 20, pp. 220–228, 2005.

[15] Y. Kamiya, T. Ohno, and S. Matsubara. Coherent back-channel feedback tagging of in-car spoken dialogue corpus. In *Proc. SIGdial*, pp. 205–208, 2010.

[16] D. Ozkan and L.-P. Morency. Modeling wisdom of crowds using latent mixture of discriminative experts. In *Proc. ACL/HLT*, 2011.

[17] 山口貴史, 井上昂治, 吉野幸一郎, 高梨克也, 河原達也. 傾聴対話における相槌形態と先行発話の統語構造の関係の分析. 人工知能学会研究会資料. SLUD, 言語・音声理解と対話処理研究会, Vol. 73, pp. 21–26, 2015.

[18] 山口貴史, 井上昂治, 吉野幸一郎, Nigel G. Ward, 高梨克也, 河原達也. 多様な相槌をうつ傾聴音声対話システムのための相槌形態の予測. 人工知能学会研究会資料. SLUD, 言語・音声理解と対話処理研究会, Vol. 75, pp. 1–6, 2015.

[19] 上里美樹, 吉野幸一郎, 高梨克也, 河原達也. 傾聴対話における相槌の韻律的特徴の同調傾向の分析. 人工知能学会研究会資料. SLUD, 言語・音声理解と対話処理研究会, Vol. 70, pp. 7–13, 2014.

[20] 丸山岳彦, 高梨克也, 内元清貴. 節単位情報『日本語話し言葉コーパス構築法』, pp. 255–322. 国立国語研究所, 2006.

[21] 高梨克也, 内元清貴, 丸山岳彦. 『日本語話し言葉コーパス』における節単位認定. 『日本語話し言葉コーパス』同梱マニュアル, 2004.

[22] 泉子・K・メイナード. 会話分析. くろしお出版, 1993.

[23] 伝康晴. 対話への情報付与, 小磯花絵 (編)『講座 日本語コーパス 3: 話し言葉コーパス—設計と構築—』. 朝倉書店, 2015.

[24] 常志強, 高梨克也, 河原達也. ポスター会話におけるあいづちの韻律的特徴に関する印象評定. 人工知能学会研究会資料. SLUD, 言語・音声理解と対話処理研究会, Vol. 56, pp. 31–36, 2009.

[25] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford. Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive kalman filtering. *The Journal of the Acoustical Society of America*, Vol. 135, No. 5, pp. 2885–2901, 2014.

[26] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, Vol. 27, No. 3, pp. 187–207, 1999.

[27] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. TANDEM-STRAIGHT : A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation. In *Proc. ICASSP*, pp. 3933–3936, 2008.

[28] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino. Nearly defect-free f0 trajectory extraction for expressive speech modifications based on STRAIGHT. In *Proc. INTERSPEECH*, pp. 537–540, 2005.

[29] R. E. Fan, K. E. Chang, C. H. Hsieh, X. R. Wang, and C. J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874, 2008.

[30] 井上昂治, 河原達也. 自律型アンドロイド Erica のための音声対話システム. 人工知能学会研究会資料. SLUD, 言語・音声理解と対話処理研究会, Vol. 75, pp. 21–24, 2015.

[31] J. Allwood, J. Nivre, and E. Ahlsén. On the semantics and pragmatics of linguistic feedback. *The Journal of semantics*, Vol. 9, No. 1, pp. 1–26, 1992.