実会話事例の機械学習により相手の話し方に合わせた音声応答を行う対話エージェント

平沼英翔†1　　三武　裕玄†1　長谷川　晶一†1

人同士の対話は声の高さや話速などの韻律が話者に合わせて変化している．お互いの韻律が相手に合わせて変化することにより，自然な会話となることが分かっている．しかし，従来の音声対話手法は、用意された韻律を用いたものや，人手で設計されたアルゴリズムに基づくものが多く，自然さが十分ではない．そこで本研究では実会話事例から機械学習によって話者の韻律に応じた韻律を付加することが出来る対話エージェントを作成し，人同士の対話に近い自然な対話の実現を目指す．

Conversational agent that generate voice response according to talker’s locution by machine learning of actual conversation case

HIDETO HIRANUMA†1　SHOICHI HASEGAWA †1
HIRONORI MITAKE†1

Prosody such as voice height and speech speed is changing according to the speaker in human conversation. It is known that a natural conversation is formed by mutual prosody changing according to the talker. However, many voice interaction methods are based on prepared prosody or based on manually designed algorithms. These are not enough nature. In this research, create voice interaction agent that can add prosody according to prosody of speaker by machine learning from actual conversation case and aim to realize natural conversation close to human.

1. はじめに

最近、Siri や Google Assistant、Amazon Echo 等の音声対話システムが一般消費者の間で広く利用されている。対話システムは大きく分けると 2 種類ある。特定のタスクの達成を目的にするものをタスク指向型対話システムと呼び，雑談的に対話を続けるものを非タスク指向型対話システムと呼ぶ[1]。

非タスク指向型の音声対話システムは会話体験そのものを目的とするため，ユーザーである人間と言語情報を通じた会話だけでなく，表情やジェスチャーなど非言語情報も重要である。さらに，対話する相手がどのような見た目，またどのような人物設定であるかということも大きな影響を及ぼす。

そのような経緯から近年、非タスク指向型音声対話システムとして注目されているのが、身体性を持つ対話エージェントである。身体性を持つ対話エージェントとは音声言語に加え，視線の動きや、ジェスチャー等の非言語情報を用いてコミュニケーションを行うキャラクタ型の対話システムのことである。現在、美術館での展示員や介護施設利用者の会話相手等，人と同様の社会的役割を果たすことが期待されている。

しかし、このような対話エージェントに利用されている音声対話システムは、人のような話し方をするように予め用意された感情を機械音声に設定するものが多い。話し方や感情が変化しない機械音声では、状況に合わない応答を取ってしまう場合があるため，人ではなく機械と話している不自然さを感じ，会話を続けにくいと感じさせてしまう．例えば相手が嬉しそうに話しているのに暗い印象を与える低く抑揚のない声で返答したり、また、相手が悲しそうに話しているのに設定された高く強い声で返答したりすると自然な会話は成立しない。これは通常、人間同士の会話では、呼びかける方（話し手）の発話の声の強弱、長短、抑揚などからなる韻律に合わせて、応答者（聴き手）も同様に発話の韻律を変化させることで自然な会話が成立するためである。この様な韻律の変化に合わせて、対話エージェントの話し方を変化させる製品がヤマハ株式会社(ヤマハ)で開発されている。ヤマハの HEARTalk[2]は話し手の韻律をリアルタイムに解析し、その応答に適した自然な韻律を返すことができる製品である。ただし、HEARTalk の音声対話システムは、韻律の変化が話し手の一言の発言からのみのアルゴリズムに基づいたものであり、対話の一連の流れである時系列変化を考慮出来ていない。例えば、暗い内容の話の最後に一言だけ明るい内容を発言した場合、明るい韻律で話し手に応答してしまう。これでは自然な韻律での応答が出来ているとは言い難い。自然な韻律を生成するためには、時系列変化を考慮することが必要になる。

そこで本研究では，対話エージェントが話し手の韻律に応じてリアルタイムに韻律を決定するための韻律決定モデルを，時系列変化を観測できる実会話事例から隠れマルコ

†1 東京工業大学　工学院　情報通信系
Tokyo institute of Technology.

ⓒ2017 Information Processing Society of Japan

フモデル(HMM)を用いた機械学習によって獲得することを目的とする．実会話における対話の韻律を記録し，それを再現するような韻律決定モデルを機械学習により獲得することで，韻律における複雑なパターンの検出が可能になると考えられ、予め感情を設定する従来の方法よりも自然な対話応答を実現することが期待できる．様々な対話内容の韻律を生成することが理想であるが､今回は｢あいずち｣における韻律の変化に着目して、学習や再現を行った．

2. 関連研究

HMM を用いて韻律を考慮した音声合成を行う研究は複数ある。吉村ら[3]の研究では、HMM により動的特徴量を考慮することで, 滑らかで自然性の高い音声スペクトル系列, ピッチパターンが得られ、データ発話者の個人性をよく再現した自然な音声が得られることが確認されている。しかし、対話システムとして用いる場合は、学習データ発話者の話し方を対話エージェントに設定することになるため、話者の韻律に対してのエージェントの韻律変化はない。

また、1 章でも述べたように、ヤマハの HEARTalk は話し手の韻律をリアルタイムに解析し、その応答に適した自然な韻律を返すことができる対話システムである。ただし、HEARTalk の音声対話システムは、高い声で話しかけるとエージェントが高い声で応答する様な韻律の同調的特徴のみを扱っている。また、韻律の変化が話し手の一言の発言からのみのアルゴリズムに基づいたものであり、対話の一連の流れである時系列的な変化を考慮出来ていない。

3. 予備実験

実会話の返答における韻律には，数多くのパターンが存在すると考えられる．それを一つ一つ見つけルールとして設計するには，手間と時間がかかる。また観察によって明示的なパターンとして発見できるものばかりとも限らない。そこで実会話事例から機械学習を用いて韻律決定モデルを獲得する． 機械学習には HMM を用いる. HMM を用いることで、対話の一連の流れである時系列的な変化をモデル化することが可能になり、より複雑なパターンの検出が可能になると考えられる。今回予備実験として、まず機械学習したデータが西村ら[4]の研究にも示されている話し手と聴き手の韻律の同調が見られるか確認する。また、機械学習したデータを用いて、話者の発話の強弱・長短・高低から対話エージェントの音声合成に韻律を付加する。システムの全体像を図 1 に示す。

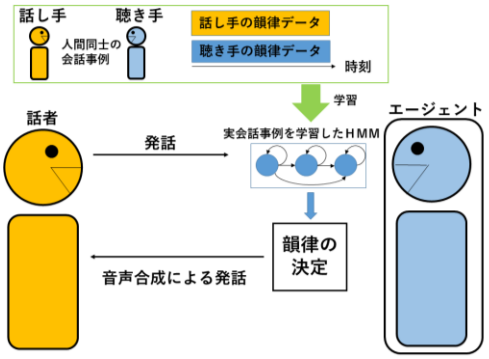


図 1 システム全体像

3.1 実会話事例の収集

対話エージェントの韻律変化の決定の規範となる，人間同士の対話の韻律を記録する実験を行う． 図 2 の様に 1 名の話し手とエージェント役として 1 名の聴き手に対面会話を行ってもらい，その韻律データをそれぞれ記録する．今回の実験では，聴き手の韻律により変化が出る様に話し手は 3 種類の異なるストーリーを話し、聴き手は任意のタイミングで「なるほど」と発言しあいずちをうつ形態で対話を行う。

3.1.1 実験装置

対話中、話し手と聴き手はマイク「HYP-190H」[5]を装着する。このマイクを用いて、話し手と聴き手の声の大きさ、高低、速度を検出する。

3.1.2 収集データ

約180秒の対話を 3 回行い、計約540秒の対話の韻律データを0. 02秒毎に記録する。韻律データは話し手と聴き手のマイクに入力された声の音量、高低である。

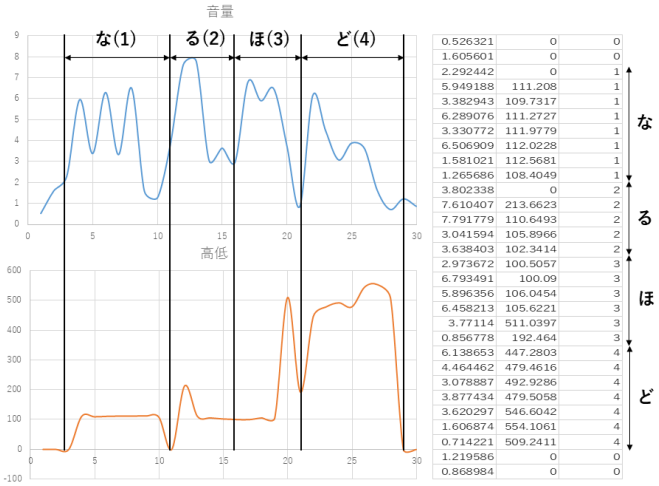
図 2 実会話記録中の様子

3.2 学習

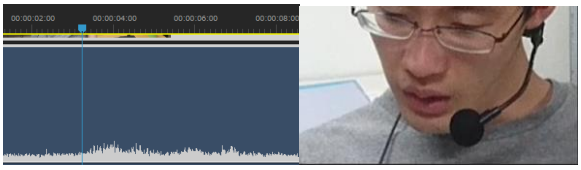
機械学習の過程について説明する。以下の様な流れで収集データを学習させる。

(1) 図 1 (a)の様に収集したデータにおいて、聴き手が発言した「なるほど」という言葉の声の音量、高低のデータを発言ごとにグラフ化し、図 1(b)スローモーションの動画と比較しながら、音素ごとに区切る。それぞれの音素にラベルとして数字を割り当て、音素ごとの音量・高低・長さを明確にする。

ⓒ2017 Information Processing Society of Japan



(a)　「なるほど」における韻律データ



(b)　スローモーション動画による音素の確認

図 1　「なるほど」の各音素の韻律データの取得

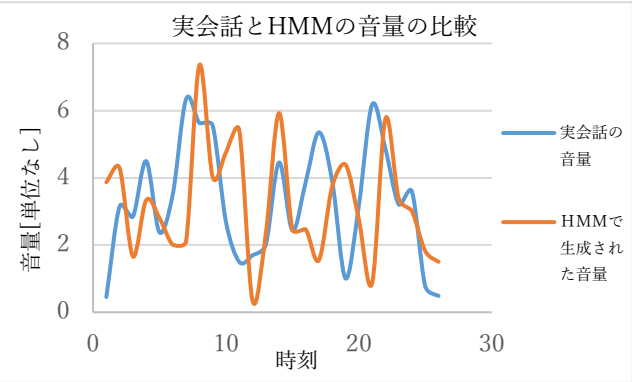
(2)　音素ごとのラベルデータと、話し手利き手の声の音量・高低の計 5 次元のデータを Blaum-Welch Learning のアルゴリズムを用いて HMM を生成する。

(3)　生成した学習データにおいて、「なるほど」の韻律として成り立たないデータが生成されるため、外れ値として除外する。

機械学習には Accord.NET[6]の HMM クラスを用いる。

3.3　学習データの確認

図 2 に HMM で生成された聴き手の「なるほど」の韻律データと実会話で記録した「なるほど」の韻律データを示す。図 2(b) の実会話における「なるほど」の韻律に近いデータとなっていることが確認できる。また図 3 に示す様に話し手が話している際は、聴き手の韻律に変化がないことから、HMM で生成されたデータに話し手と聴き手の発言する順序が考慮されていることが確認できる。



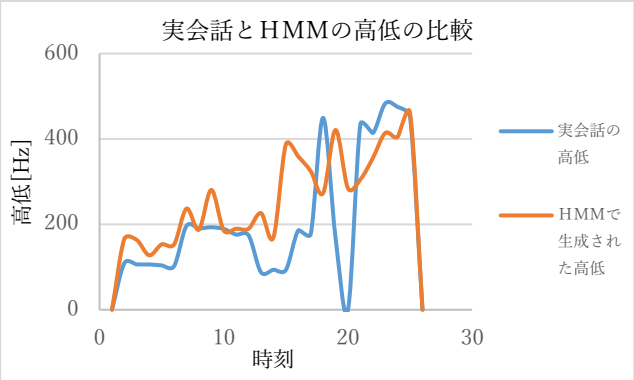


図 2　HMM と実会話の韻律データの比較

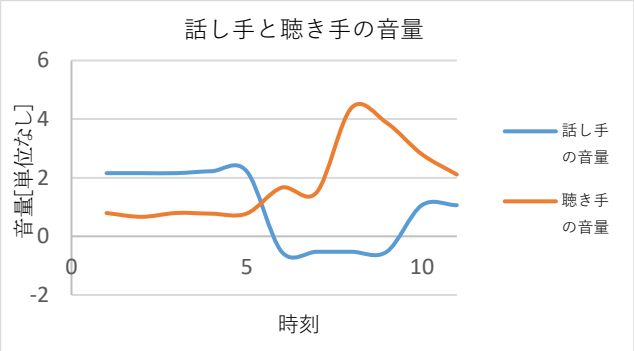


図 3　HMM で学習された話し手と聴き手の韻律データ

また時系列的に韻律の変化を見てみる。HMM で学習したデータにおいて話し手の韻律に聴き手の韻律が対応しているかを確認するために、話し手の 2 秒分の韻律（100 個分の時系列データ）の平均を取り、聴き手のあいずちの韻律と比較すると、韻律の同調が確認できる。結果を表 1、図 4 に示す。

表 1

	話し手		聴き手
音量	9.366357		3.404576
	8.788008		3.001936
高低[Hz]	279.1338		246.1
	225.2008		232.3277

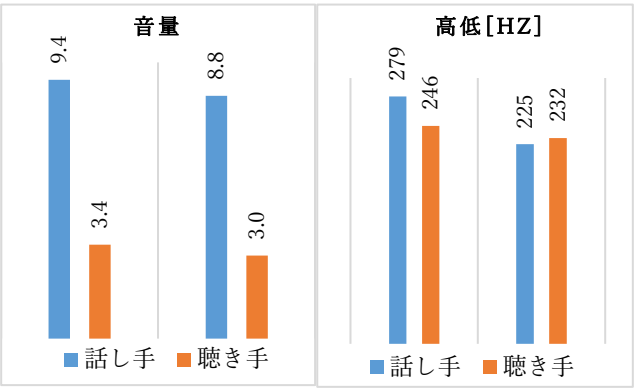


図 4　話し手と聴き手の韻律の変化

ⓒ2017 Information Processing Society of Japan

3.4 音声合成

HMM により学習された「なるほど」の韻律データを Unity with Vocaloid[7]を用いて音声合成する。今回、音声合成エンジンには音素ごとに韻律を付加することが可能である Unity with Vocaloid を用いる。音素ごとの音量・高低・長さを設定し音声合成を行う。音量については音素ごとの連続値の平均を適用し、高低に関しては音素ごとの最大値を適用した。主観だが、滑らかな音声合成が出来ていないため、システムの改良が必要だと感じられる。

4. 提案手法

予備実験では連続値の 5 次元の時系列データを学習し HMM を作成した。しかし、多次元の連続量を単一の HMM で学習する場合、状態数を十分多くする必要があり、計算時間が多くかかってしまう。これに対して HMM を階層的に用いる手法[8]が提案されており、長時間の連続値データからパターンを発見することに適している事から、提案手法として利用を検討する。

4.1 階層型 HMM を用いた学習

2 階層隠れマルコフモデルについて説明する．処理の流れは以下の通りである．処理の概要図を図 5 に示す。

(1) 話し手の時系列の韻律データを任意のフレーム数 ω 毎に区切る。聴き手の韻律データはあいずち毎に任意のフレームで区切る。
(2) 話し手・聴き手の各フレームの部分時系列データについて，Blaum-Welch-Learning のアルゴリズムを用いて任意の状態数の μ HMM を生成する。
(3) 話し手と聴き手の生成した各 μ HMM をラベル付けする。
(4) 話し手と聴き手の μ HMM を時系列に基づき、1 つの状態として，Blaum-Welch-Learning のアルゴリズムを用いて全体の HMM を生成する。

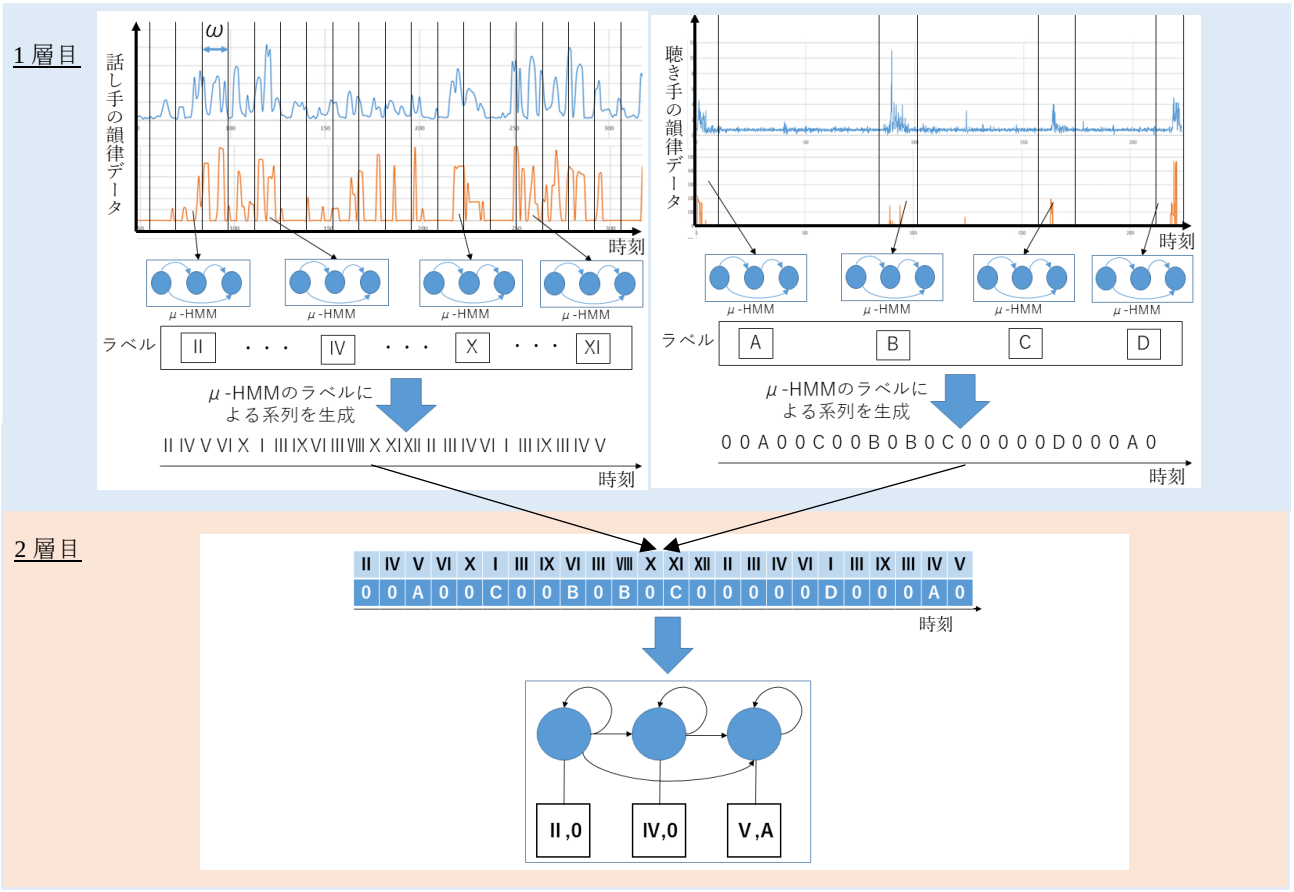


図 5　2 階層 HMM の概要図

ⓒ2017 Information Processing Society of Japan

5. 今後の展望

今後、4 章で提案した 2 階層 HMM による学習と話し手の発言に対応したエージェントの音声合成を行うインタラクティブシステムの開発を行う予定である。また音声合成においては現在は主観での評価しか行っていないため、被験者を用意し、評価実験を行う予定である。

謝辞

本研究は科研費（17K17713）の助成を受けたものである。

参考文献

[1] 狩野 芳伸. “コンピューターに話が通じるか.対話システムの現在” 情報管理 (2017.1. vol. 59 no. 10)

[2] “心が感じられる音声対話システム” HEARTalk™” [http://www.y2lab.com/project/heartalk/]

[3] 吉村 貴克 “HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の 同時モデル化” 電子情報通信学会論文誌 2000/11 Vol. J83–D–II No. 11

[4] 西村 良太 “音声対話における韻律変化をもたらす要因分析” (2009)

[5] オーディオテクニカ 「HYP-190H」 [https://www.audio-technica.co.jp/mi/show_model.php?modelId=2531]

[6] Accord.NET Machine Learning Framework [http://accord-framework.net/]

[7] Unity with Vocaloid [http://business.vocaloid.com/unitysdk/]

[8] 道木加絵,“HMM 間の遷移関係と統計処理に基づく人間の行動モデルの生成”,JSMENo.15-2 Proceeding,2015

ⓒ2017 Information Processing Society of Japan