対話行為情報を表現可能なDNN音声合成と 発語内行為自然性に関する評価

DNN-based Speech Synthesis using Dialogue-Act Information and Its Evaluation with Respect to Illocutionary Act Naturalness

北条 伸克
Nobukatsu Hojo

NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
nobukatsu.hojo.cd@hco.ntt.co.jp

井島 勇祐
Yusuke Ijima

NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories

杉山 弘晃
Hiroaki Sugiyama

NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

宮崎 昇
Noboru Miyazaki

(同 上)

川西 隆仁
Takahito Kawanishi

(同 上)

柏野 邦夫
Kunio Kashino

(同 上)

keywords: text-to-speech, spoken dialogue system, dialogue-act, illocutionary act naturalness.

Summary

This paper aims at improving naturalness of synthesized speech generated by a text-to-speech (TTS) system within a spoken dialogue system with respect to “how natural the system’s intention is perceived via the synthesized speech”. We call this measure “illocutionary act naturalness” in this paper. To achieve this aim, we propose to utilize dialogue-act (DA) information as an auxiliary feature for a deep neural network (DNN)-based speech synthesis system. First, we construct a speech database with DA tags. Second, we build the proposed DNN-based speech synthesis system based on the database. Then, we evaluate the proposed method by comparing its performance with two conventional hidden Markov model (HMM)-based speech synthesis systems, namely, the style-mixed modeling method and the style adaptation method. The objective evaluation results show that the proposed method overwhelms the style-mixed modeling method in the accuracy of reproduction of global prosodic characteristics of dialogue-acts. They also reveal that the proposed method overwhelms the style adaptation method in the accuracy of reproduction of sentence final tone characteristics of dialogue-acts. The subjective evaluation results also show that the proposed method improves the illocutionary act naturalness compared with the two conventional methods.

1. はじめに

本研究では、音声対話システムで使用される合成音声の自然性を向上させるため、発話意図を考慮する deep neural network (DNN) 音声合成を提案する。人同士の対話では、ある参加者が発話を行うたびに、聞き手がその発話意図を推論する [Grice 91, Wilson 12]。発話意図の表現が不自然であれば、聞き手の推論のために必要な認知負荷の増大、文脈との不整合や誤解の発生など、円滑なコミュニケーションを阻害する要因となる。音声対話システムにおいても同様に、円滑なコミュニケーションを実現するためには、発話意図を自然に表現する必要があると考えられる。伝達される発話意図は、発話音声の韻律的特徴（基本周波数（F0）や音素継続時間長（継続

長）等の傾向）により変化する [Hellbernd 16, 前川 02]。したがって、韻律的特徴を制御することで、発話意図を自然に表現する音声合成技術が必要である。

音声合成分野では、通常、合成音声の自然性は文脈から切り離された音声サンプルの品質（肉声感等）として評価される [ITU-T 94, Bezooijen 97]。一方、上述の理由から、円滑なコミュニケーションを実現するためには、肉声感に加えて、発話意図を表現する音声として自然である必要がある。発話行為論 [Austin 75, Searle 80] の分類を使用すると、前者は音声を発する行為（発語行為）、後者は発話により話者の発話意図を伝える行為（発語内行為）としての自然性と言い換えられる。本研究は、前者を発語行為自然性、後者を発語内行為自然性と定義し、発語内行為自然性の向上を目指す。

発語内行為自然性を向上させるためには、発話意図の韻律的特徴を合成音声に反映する必要があると考えられる。関連研究として、感情または発話スタイルを合成音声で表現する、感情音声合成 [An 17, Barra-Chicote 10, Lorenzo-Trueba 18, Nose 07, Xue 18, Yamagishi 05, Yamagishi 07] がある。感情音声合成によって、“happy”, “sad” などの基本感情を中心に、3~10 個程度の感情を合成音声により表現可能である。音声合成手法として、hidden Markov model (HMM) に基づく手法 [Nose 07, Yamagishi 05, Yamagishi 07] に加え、DNN に基づく手法 [Xue 18] や long short term memory-recurrent neural network (LSTM-RNN) などの時系列モデルに基づく手法 [An 17, Lorenzo-Trueba 18] の有用性が示されている。パラ言語・非言語情報 [Fujisaki 97] を合成音声で表現するという共通点から、これらの音声合成手法は本研究においても有用である可能性がある。

一方で、感情音声合成と本研究には、相違点がある。それは、発話意図の表現では、特に、発話の末尾の音調（文末音調）の制御が重要であるという点である。感情音声の基本的な特徴として、悲しげな音声は一般的に F0 の発話平均が低く、話速が遅い傾向がある [Scherer 77] など、発話文章によらず、発話全体に顕著に韻律的特徴が現れることが知られている。また、発話意図に関しても同様の特徴が示されている [Syrdal 08]。一方で、日本語音声については、文末音調の特徴も重要である [井上 93, 森山 89, 郡 15, 杉藤 01, 轟木 08, 上村 89]。特に、[岩田 17] では、音声の文末音調のみを制御することで、その僅かな変化により、様々な発話意図やニュアンスが表現されることが示されている。ここで、文末音調は、発話意図のみでなく、発話文章の文末形式（終助詞等）により大きく傾向が異なることが知られている [郡 18]。文末音調は、いわば、発話意図だけでなく発話文章にも依存して現れる特徴であり、発話文章によらず発話全体に現れる（大域的な）韻律的特徴と対照的である。上述の相違点から、感情音声合成の手法が発語内行為自然性の向上のために直接適用可能であるかについては、改めて検証される必要がある。

発話意図を考慮する音声合成手法を検討した先行研究には、波形接続音声合成を用いたコールセンタ応答音声の生成 [Syrdal 08], HMM 音声合成を用いたレストランやホテルの情報案内音声の生成 [Tsiakoulis 14] がある。いずれの研究においても、発語行為自然性が評価され、その改善が示されている一方、発語内行為自然性の観点からの評価は行われていない。また、先行研究では、発話意図を表現する音声合成手法として、DNN の使用は検討されていない。DNN は、感情音声合成において HMM に対する有用性が示されており [Xue 18], 発話意図の表現においても有用であることが期待される。以上から、本研究では、発話意図を表現する音声合成について DNN の使用を検討し、発語内行為自然性に関する評価により、HMM に基づく従来法との比較を行う。

本研究では、HMM に基づく従来法として、話者や感情の表現において広く使用される、平均声モデルとモデル適応に基づく手法 [Tachibana 06, Yamagishi 09], およ

びスタイル混合モデル [Yamagishi 05] を使用する。これらの従来法について、各発話意図を 1 つの感情カテゴリとみなしモデルの学習および音声合成を行う。モデル適応に基づく手法では、あらかじめ複数の発話意図の音声をもとに学習した平均声モデルに対して、各発話意図の音声を用いて適応学習を行う。平均声モデルの学習時には、全ての発話意図に対し単一の決定木が構築される。ここで、決定木の構築には、発話文章の言語的情報（音素やアクセントなど）に関する質問のみが使用され、発話意図は考慮されない。このため、発話文章と発話意図の双方に依存するような音声の特徴をモデル化するために最適な木構造が得られるとは限らない。特に、上述の通り、文末音調は、発話意図だけでなく発話文章（文末形式）にも依存する特徴である。このため、モデル適応に基づく手法では、文末音調が十分に再現されない懸念がある。これに対し、スタイル混合モデルは、決定木の構築の際、言語的情報のみでなく発話意図に関する質問を使用する。このため、文末音調の再現に適した木構造が得られると期待される。一方で、各発話意図に関する質問が決定木中で使用されるかどうかは、最小記述長 (minimum description length; MDL) 基準 [Shinoda 97] により決定されるため、韻律的特徴が顕著でない発話意図に関する質問は使用されない可能性がある。その結果、発話意図の持つ韻律的特徴が十分に再現されない懸念がある。以上から、HMM に基づく従来法は、いずれも、決定木に関する制約により、発話意図の持つ韻律的特徴の再現性に課題があると考えられる。

DNN に基づく提案手法では、決定木を使用せず、入力情報（言語的情報と発話意図）と出力情報（音声パラメータ）の関係をニューラルネットワークによりモデル化する。ニューラルネットワークは、入出力の関係について明示的な制約を持たないため、従来研究の課題が解決され、発語内行為自然性が改善されると期待される。ここで、上述の通り、発語内行為自然性の改善のためには、韻律情報の表現力が特に重要であると考えられる。このため、本研究では、DNN に基づく手法として、発話意図を補助特徴量として入力する DNN 音声合成を提案する。同様の構成の DNN は、複数話者の音声合成において、出力層を分岐する構成 [Wu 15] と比べ、F0 の推定誤差が小さいことが示されている [Hojo 18] ため、文末音調を再現する上で有用であると期待される。また、大規模なデータを使用し、より表現力の高い時系列モデル [An 17, Lorenzo-Trueba 18, Skerry-Ryan 18, Xue 18] を検討することも有用であると考えられるが、本研究では、構築された音声データベースのサイズ等の制約から、feed-forward 型の DNN に基づく手法を検討する。

本論文の構成について述べる。2 章では、本研究で構築した発話意図情報付き音声データベースについて述べ、発話意図の持つ韻律的特徴の分析を行う。3 章では、提案法として、発話意図を補助特徴量として入力する DNN 音声合成手法について述べる。4 章では、発語内行為自然性に関する評価実験を行い、提案法の有用性を明らかにする。最後に、5 章では、本研究のまとめを行う。

[34]		
番号	話者	発話
	A	スポーツにはちょっと疎くて、小学生から中学に上がる頃くらいまでテニスを習っていたみたいです。
	A	Bさんは何かされますか？
	B	そうですね。
3401	B	私はバスケをしていました。
	B	テニスも趣味程度ですができます。
	A	チームプレイが求められるスポーツの経験がほとんどないので、
3402	A	承認 バスケが出来るというのは尊敬します。
3403	A	自己開示_事実 スラムダンクも読んでましたし。

図1 音声データベース構築に使用した収録文章例。赤枠で示された最終1発話を音声収録に用いる。

2. 対話行為情報付き音声データベースの構築

2.1 音声データベースの構築

発話意図を考慮する音声合成の検討のためには、発話意図のカテゴリ（タグセット）と、発話意図タグが付与された音声データベースが必要である。本研究では、発話意図のタグセットとして、雑談対話システムにおいて使用される対話行為情報 [目黒 12] を使用し、対話行為タグ付きの音声合成用データベースを構築する。対話行為とは、発話意図の抽象的な表現である [Stolcke 00]。本研究で使用する対話行為タグセットは、雑談対話における発話の発話意図を記述するために設計されたものである。

通常の音声合成用の音声収録では、例えば ATR503 文 [Kurematsu 90] などの音素バランス文を使用し、各文章を個別に読み上げ、音声を収録する。一方、本研究で使用する音声データベースの音声収録では、発話意図の自然な表現を行うため、発話意図に合致した文脈および収録文章を使用することが望ましいと考えられる。そこで、本研究の音声収録では、図1に例示する収録文章を使用する。音声収録時には、図1中の赤枠で示された最終ターンの発話のみが発声される。発声される発話より以前の発話については、対話の文脈に合わせた自然な発話を行うために掲載される。また、音声収録時には、対話の文脈と対話行為に応じ、自然な音声を発声するように指示を行う。

本研究では、収録文章の元となるテキストデータとして、テキストチャットコーパス [目黒 12] を使用する。このコーパスにはテキスト対話が収録されており、そのテーマは12個で、食べ物、旅行、映画などである。2名による対話形式であり、全体で12名の実験参加者による対話が収集されている。また、各発話に対して、専門家2名により、対話行為が付与されている。このコーパス中の発話から、各対話行為の出現回数、および各対話行為における各音素の出現回数のバランスを考慮して発話を選択することで、収録文章を設計する。各対話行為の出現回数、および各対話行為における各音素の出現回数のバランスの尺度については、エントロピーを使用する [Nose 17]。

本研究では、プロの女性声優1名による発話を収録した。以上の手続きにより、合計5177発話の音声を収録した。収録された5177発話には、発話文章および対話行為が同一である発話が含まれるため、重複する発話を除外し、3410発話（音声長の合計約140分）を音声デー

表1 対話行為別の音声データベースの統計。[#]により発話数を、[sec]により発話時間長（秒）を示した。

対話行為		発話量		対話行為		発話量	
		[#]	[sec]			[#]	[sec]
情報提供	Info	120	293	質問_習慣	Q_h	108	255
自己開示_事実	Sd_f	127	324	質問_予定	Q_pl	92	231
自己開示_経験	Sd_e	91	298	質問_欲求	Q_d	75	237
自己開示_習慣	Sd_h	101	310	質問_評価	Q_pr	108	260
自己開示_予定	Sd_p	108	275	確認	Conf	120	273
自己開示_欲求	Sd_d	113	298	提案	Prop	99	234
自己開示_評価+	Sd_+	184	530	質問_自問	Q_s	111	285
自己開示_評価-	Sd_-	186	490	繰り返し	Rept	149	277
自己開示_評価0	Sd_0	109	312	言い換え	Para	121	291
共感・同意	Sym	146	321	相槌	Ack	83	165
非共感・非同意	N-Sym	176	416	フィラー	Fill	71	130
承認	Apro	173	400	感嘆	Admn	130	193
質問_情報提供要求	Q_i	122	301	挨拶	Grt	74	140
質問_事実	Q_f	111	267	感謝	Thks	26	81
質問_経験	Q_e	100	262	謝罪	Apol	70	196
				合計		3410	8353

図2 対話行為別の平均 F0、平均話速の散布図。

タベースとして使用した。また、対話行為として、全33種類からなる対話行為から、“その他”などの対話行為を除いた全30種類を研究対象とした。音声データベースの統計情報を表1に示した。全ての発話について、後述する音声分析、音声合成モデルの学習時に使用するため、音素境界の時間情報、アクセント型、アクセント句境界等のアクセント情報を人手で付与した。

2.2 音声分析

各対話行為について、その韻律的特徴を明らかにするために、2.1節で構築された音声データベースを使用した音声分析を行う。

§1 大域的な韻律的特徴の分析

まず、各発話について算出された話速と F0 について、対話行為ごとに平均を算出し、図2に散布図で示した。平均話速の算出には人手で付与した音素境界情報を使用した。F0の分析には STRAIGHT [Kawahara 99] を使用した。“感嘆 (Admn)”、“フィラー (Fill)”、“感謝 (Thks)”の3つの対話行為では、韻律的特徴が他と大きく異なる傾向が明らかとなった。また、“自己開示_評価+ (Sd_+)”、“承認 (Apr)”、“共感・同意 (Sym)”などの対話行為では、平均 F0 が他よりも高い傾向が明らかとなった。これらの対話行為では、文章が全体として明るい口調で発話された

図3 平均文末 F0 の算出方法。図は、文末形式：“か”，対話行為：“質問_経験”の平均文末 F0 の算出方法を例示した。

ためであると考えられる。一方，“自己開示_評価-(Sd-)”では、平均 F0 が他よりも低い傾向が明らかとなった。文章が全体として暗い口調で発話されたためであると考えられる。また，“繰り返す (Rept)”，“質問_自問 (Q_s)”，“承認 (Apro)”，“共感・同意 (Sym)”，“自己開示_評価-(Sd-)”などの対話行為では、平均話速が遅い傾向が明らかとなった。音声合成時に対話行為を考慮することで、これらの大域的な韻律的特徴が再現され、その対話行為の発話音声として自然な音声合成が可能となると期待される。また、上述の対話行為で大域的な韻律的特徴が確認された一方で、他の多くの対話行為では、平均話速、平均 F0 には大きな違いが見られなかった。

§2 文末音調の分析

本節では、対話行為の持つ文末音調の傾向を分析する。先行研究によって、同じ発話文章でも、その発話の発話意図に依存して文末音調が変化する傾向が知られている [前川 02, 郡 18]。本研究で使用する音声データベースにおいても、同様の傾向が見られるかを確認する。文末音調は、対話行為のみでなく、発話文章、特にその文末形式（終助詞等）により傾向が異なる [郡 18]。このため、文末音調の分析は、文末形式ごとに行う。ただし、文末形式は多数存在し、その全ての場合について結果を示すことは困難であるため、本節では、音声データベースにおける出現頻度の高い上位 3 つの文末形式（“文末詞なし”（「～です」、「～ます」、「～でした」等、終助詞のつかない文章），“～ね”，“～か”）について結果を例示する。また、以下では、文末音調の分類として、郡の分類（疑問型上昇調 ↗、強調型上昇調 ↑、平坦調 →、上昇下降調 ↘、急下降調 ↓、無音調 _） [郡 15] を使用する。

図 3 に分析方法を例示した。まず、ある文末形式および対話行為に対応する発話の最終 1 音節の F0 軌跡（文末 F0）を抽出する（図 3(a) の点線）。続いて、各発話で、文末 F0 の時間長を正規化する（図 3(b) の点線）。正規化された文末 F0 から、平均を算出する（図 3(b) の実線）。また、別途最終 1 音節の平均時間長を算出する。最後に、時間長が正規化された平均文末 F0 が、最終 1 音節の平均時間長となるよう時間方向に伸縮し、平均文末 F0 を得る。平均文末 F0 の全体の形状は保ったまま、始端の高さが 0 となるように周波数方向に平行移動し、平均文末 F0 を図示する（図 3(c) の実線）。プロットの単純化のため、“自己開示 (Sd)”，“質問 (Q)” のサブカテゴリはそれ

(a) 文末形式：“文末詞なし”

(b) 文末形式：“ね”

(c) 文末形式：“か”

図4 文末形式・対話行為別の平均文末 F0。

ぞれ区別せず、カテゴリ全体の平均を表示した。ただし、“質問_自問 (Q_s)”については、他の“質問 (Q)”の対話行為の文末音調と異なる傾向が得られたため、区別して表示した。

図 4 に、文末形式別および対話行為別の平均文末 F0 を示した。図 4(a) から、“文末詞なし”の発話について、“情報提供 (Info)”や“自己開示 (Sd)”などの対話行為では、文末 F0 はおよそ 500 cent 下降する一方、“質問 (Q)”の対話行為では、文末 F0 はおよそ 200 cent 上昇することがわかる。これは、“文末詞なし”の文末は、“情報提供 (Info)”や“自己開示 (Sd)”の対話行為では無音調で発話された（例：「結構有名です _」）一方、“質問 (Q)”の対話行為では疑問型上昇調で発話された（例：「パクチーの日本名って知ってます ↗」）ためであると考えられる。

次に、図 4(b) から、文末形式が“ね”である発話については、“フィラー (Fill)”ではおよそ 400 cent の下降が見られたが、他の対話行為では同様の傾向は確認され

図5 DNNに基づく音声合成の概略図。

なかった。これは、文末形式が“ね”である発話は、“フィラー (Fill)” 以外の対話行為では強調型上昇調で発話された (例:「そうなんです↑」) 一方、“フィラー (Fill)” の対話行為では上昇下降調で発話された (例:「そうですね↘」) ためであると考えられる。

続いて、図4(c)から、文末形式が“か”である発話については、“相槌 (Ack)”, “繰り返し (Rept)”, “言い換え (Para)”, “感嘆 (Admin)”, “質問_自問 (Q_s)” などの対話行為では、文末音調が下降するが、“質問 (Q)” の対話行為ではおよそ 600cent, “確認 (Conf)” の対話行為ではおよそ 200cent, 文末音調が上昇することがわかる。これは、文末形式が“か”である発話は、“相槌 (Ack)”, “繰り返し (Rept)”, “言い換え (Para)”, “感嘆 (Admin)”, “質問_自問 (Q_s)” などの対話行為では多くの発話において無音調で発話された (例:「国内中心に一人旅をします」—「一人旅ですか↘」) 一方、“質問 (Q)”, “確認 (Conf)” などの対話行為では疑問型上昇調で発話された (例:「小学校の宿題ですか↗」) ためであると考えられる。

以上より、本研究で使用する音声データベースにおいて、文末音調が対話行為に依存する傾向が確認された。また、“文末詞なし”, “ね”, “か” の文末形式においてそれぞれ異なる傾向が得られており、文末音調は対話行為と文末形式の双方に依存する傾向が確認された。また、例えば、“自己開示 (Sd)” と“質問 (Q)” のように、図2では大域的な韻律的特徴は顕在的でないものの、文末音調の特徴は顕在的である対話行為が確認された。

以上の音声分析の結果から、図2で示された大域的な韻律的特徴と、図4で示された文末音調の特徴の双方を合成音声に再現することで、対話行為に応じた自然な音声合成が可能であると期待される。

3. 音声合成手法

3.1 DNNに基づく音声合成 (DNN-BASELINE)

従来の DNN 音声合成 [Zen 13] の概略図を図5に示した。従来法では、入力された言語特徴量ベクトルから、継続長を回帰するモデル (継続長モデル), 音響特徴量ベクトルを回帰するモデル (音響モデル) として DNN を使用する。まず、入力テキストは言語特徴量ベクトルへ変換される。言語特徴量ベクトルは、言語特徴量に関する質問に対する答えに対応するバイナリ変数と、数値変数

図6 提案法の概略図。継続長モデル、音響モデルとも、同様の構成を用いる。

により構成される。次に、継続長モデルにより言語特徴量ベクトルから各音素の継続長を推定する。継続長をもとにフレーム特徴 [Wu 16] を計算し、言語特徴量ベクトルとともに音響モデルへ入力することで、音響特徴量ベクトルを推定する。音響特徴量ベクトルは、スペクトル特徴量や F0, およびその動的特徴を連結して構成する。推定された音響特徴量ベクトルをもとに音声波形生成を行い、出力音声を得る。

3.2 対話行為コードを入力する DNN 音声合成 (DNN-DACODE)

DNN 音声合成において、対話行為の持つ韻律的特徴を再現した音声合成するための最も単純な方法は、対話行為ごとに個別のモデルを学習し、音声合成時には、合成したい対話行為に応じて使用するモデルを変更する方法である。しかし、この方法により高品質な合成音声を得るためには、各対話行為について十分な量の学習データが必要であり、データ整備のためのコストが上昇する。

提案法では、言語特徴量と同様に、対話行為を1つの入力情報とみなし、1つの DNN により全ての対話行為の音声をモデル化する。対話行為を入力情報として考慮し DNN を学習することで、音声合成時には、指定された対話行為の韻律的特徴を再現可能であると考えられる。また、言語特徴量ベクトルから音響特徴量ベクトルを回帰する部分のネットワークについては、全ての対話行為のデータを活用して学習することが可能となるため、一つ一つの対話行為の学習データは少量である場合も、高品質な音声合成が可能であると期待される。

また、DNN に基づく提案法は、モデル化される入出力ベクトル間の関係に明示的な制約を持たない。このため、言語特徴量ベクトル、対話行為および音響特徴量の間の複雑な依存関係がモデル化可能であると期待される。一方、1章で述べたとおり、従来のモデル適応に基づく手法 [Tachibana 06, Yamagishi 09] では、平均声モデルの決定木の構築時に対話行為に関する質問は使用されないという制約を持つ。このため、文末形式と対話行為の双方に依存する特徴である文末音調に関して、それを再現するための最適な木構造が得られるとは限らない。このため、文末音調が十分に再現されない懸念がある。また、従来のスタイル混合モデル [Yamagishi 05] では、決定木の構築時に、対話行為に関する質問が使用されるかどうかは、MDL 基準により決定される。このため、韻律的特

表 2 発話文章類型と類型化基準.

類型名	類型化基準			サブカテゴリ	発話例
完全文	非定型文	独立詞以外の品詞を含む	述語あり	12 (文末詞なし, か, な, ね, や, よ, かな, かね, け, なの, かしら, ください)	ちょうど良いですよ
言いさし文			主節なし	15 (か, が, し, て, で, と, に, かも, から, けど, たら, って, では, ので, けども)	偶然というか
述語省略文			述語なし		-
独立詞文		独立詞のみ		-	えー
定型文	定型文			3 (挨拶定型文, 感謝定型文, 謝罪定型文)	ありがとうございます

表 3 評価実験で使った対話行為類型.

類型名	対話行為
情報伝達	情報提供, 自己開示_{事実, 習慣, 経験, 予定, 評価+, 評価-, 評価 0, 欲求}, 共感・同意, 非共感・非同意, 承認
質問	質問_{情報提供要求, 事実, 習慣, 経験, 予定, 評価, 欲求}, 確認, 提案
自問	質問_自問
繰り返し/言い換え	繰り返し, 言い換え
相槌	相槌
フィラー	フィラー
感嘆	感嘆
定型文	挨拶, 感謝, 謝罪

徴が顕著でない発話意図に関する質問は使用されず, その特徴が十分に再現されない懸念がある. 従来法は, いずれも決定木に関する制約により, 韻律的特徴の再現性に課題があると考えられる. 提案法により, このような従来法の課題が解決可能であると期待される.

提案法の概略を図 6 に示す. 提案法の DNN 音響モデルは, 話者情報ではなく対話行為を用いる点を除き, 話者コードを用いる DNN 音声合成 [Hojo 18] と同様である. DNN の学習に使用する対話行為のカテゴリ数が M 個の場合, 対話行為 c ($c = 1, \dots, M$) について, 対話行為コード $z_c = [z_{1c}, \dots, z_{Mc}]$ は次式で表現される.

$$z_{mc} = \begin{cases} 1 & (m = c) \\ 0 & (m \neq c) \end{cases} \quad (1)$$

対話行為コード z_c は, 重みパラメータにより回帰され, 隠れ層へ入力される. モデル学習時には, 学習データに含まれる全ての対話行為の発話について, 言語特徴量ベクトル, 対話行為コードおよび音響特徴量ベクトルを使用し, ネットワーク全体のパラメータを学習する. 音声合成時には, 合成する対話行為に対応する対話行為コードと言語特徴量ベクトルを DNN へ入力し, 順伝播を行うことで, 音響特徴量を出力する.

4. 評価実験

4.1 評価セットの設計

本節では, 本研究の評価実験で使用する発話集合 (評価セット) の設計方法について述べる. 従来研究では, 評価セットとして, 音声データベースからランダムに選択された発話を使用する. しかし, 2.2 節で確認した通り, 対話行為の韻律的特徴は, 対話行為のみでなく発話文章の文末形式にも依存する. したがって, 従来と同様の方法で評価セットを設計した場合, 得られる評価結果は, ランダムに選択される発話の対話行為や文末形式に強く依存し, 再現性が損なわれる懸念がある. そこで, 本研究では, 対話行為や文末形式を統制した評価セットの設計を行う. ただし, 対話行為には 30 パターンが, 文末形式にも多数のパターンが考えられるため, その組み合わせの全ての場合を含む評価セットを使用することは, 評価実験コストの観点から困難である. したがって, 本研究では,

- (1) 発話文章および対話行為について, 少数のパターンに類型化する (4.1.1 節, 4.1.2 節).
- (2) 対話行為の各類型について, 使用される頻度の高い発話文章の類型を調査し, 頻度の上位 3 位までを評価対象とする (4.1.3 節).

特に, (2) により, 使用される頻度の高いパターンを評価対象とすることで, 実際に音声対話システムが使用する頻度の高い対話行為および文末形式に関する評価を行う.

§1 発話文章の類型化

2.2 節で確認した通り, 対話行為の韻律的特徴は発話文章の文末形式に依存するため, 文末形式に着目した発話文章の類型化を行う. ここで, 対話音声の発話文章には, 例えば「ちょうどいいですよ (自己開示_評価+)」のように, 主節や文末詞を含む発話のみでなく, 例えば「偶然というか (自己開示_評価 0)」, 「香川が (感嘆)」, 「えー (フィラー)」など, 主節を含まない発話も含まれる. このような発話も評価対象に含めるため, 本研究では, 表 2 に示す発話文章の類型を使用し, 評価セットを設計する. 主節のある発話文章は“完全文”, 主節のない発話は, 述語の有無, 独立詞以外の品詞の有無を基準に, “言いさし文 [白川 09]”, “述語省略文”, “独立詞文” にそれぞれ類型化した. “挨拶” (「よろしくお願ひします」, 「こんにちわ」など), “感謝” (「ありがとうございます」, 「ありがとうございました」など), “謝罪” (「すみません」, 「失礼しました」など) の発話については, 対話行為と発話

表 4 テキストチャットコーパス [目黒 12] における各対話行為類型・発話文章タイプの頻度.

対話行為類型	頻度 1 位			頻度 2 位			頻度 3 位			頻度 4 位以降		合計
	類型名	発話数	割合	類型名	発話数	割合	類型名	発話数	割合	発話数	割合	
情報伝達	完全文_文末詞なし	34168	38%	完全文_ね	22897	25%	言いさし文_が	5745	6%	27368	30%	90178
質問	完全文_か	14191	69%	完全文_ね	1990	10%	完全文_文末詞なし	1531	7%	2860	14%	20572
自問	完全文_文末詞なし	297	21%	完全文_かな	261	19%	完全文_かね	159	11%	674	48%	1391
繰り返し/ 言い換え	完全文_ね	1111	35%	完全文_か	1001	31%	述語省略文	477	15%	606	19%	3195
相槌	独立詞文	1926	43%	完全文_か	1403	32%	完全文_ね	834	19%	269	6%	4432
フィラー	独立詞文	596	62%	完全文_ね	309	32%	述語省略文	27	3%	33	3%	965
感嘆	独立詞文	1624	91%	述語省略文	53	3%	完全文_か	49	3%	55	3%	1781
定型文	挨拶定型文	8652	78%	感謝定型文	2091	19%	謝罪定型文	368	3.3%	0	0%	11111

文章に決まった対応関係が存在するため、それらを“定型文”として区別した。“完全文”は、さらに文末形式により区別し、“完全文_文末詞なし”、“完全文_ね”、“完全文_か”などに類型化した。また、“言いさし文”については、その末尾の接続助詞等（「けど」、「のに」、「て」等）に基づき、“言いさし文_けど”、“言いさし文_のに”、“言いさし文_て”などに類型化した。以上の類型化により、本研究では、発話文章を 32 種類に類型化した。

§2 対話行為の類型化

対話行為の類型化の際は、まず、“フィラー”、“感嘆”、“自問”など、2・2・1 節において、大域的な韻律的特徴が顕著であった対話行為についてそれぞれ 1 つの類型を割り当てた。また、本研究では、“自己開示”および“質問”（ただし、“質問_自問”を除く）の対話行為に関するサブカテゴリは区別せず、それぞれ 1 つの類型とした。同様に、対話行為間の意味的な類似性に着目し、8 種類に類型化を行った。以上から、本研究では表 3 に示す対話行為類型を使用した。ただし、今回の実験では、評価実験コストの制約から“自己開示_評価+”と“自己開示_評価-”など、図 2 において異なる韻律的特徴が確認された対話行為について、同一の類型を割り当てた。今後の課題として、評価セットのサイズを増加させ、各対話行為について詳細な評価を行うことが挙げられる。

§3 各類型の発話頻度の調査

各対話行為類型について、使用される頻度の高い発話文章類型を調査した。調査には、テキストチャットコーパス [目黒 12] を使用した。表 4 に調査結果を示した。発話文章の上位 3 位までを評価セットとして使用することで、“自問”の 5 割以上、“情報伝達”の 7 割以上、それ以外の対話行為タイプのそれぞれ 8 割以上の頻度の発話が評価対象に含まれることがわかる。このように、各対話行為類型において出現頻度の高い発話文章類型を評価に使用することで、一般性の高い評価結果を得ることが可能であると期待される。以上により、評価対象となる対話行為類型・発話文章タイプのパターン数は合計で 24 (8×3) となる。客観評価実験の評価セットとして、24 種類のそれぞれについて 10 発話 (合計 240 発話) を音声データベースからランダムに選択した。主観評価実験の評価セットとして、24 種類のそれぞれについて 5 発話 (合計 120 発話) をテキストチャットコーパスから選択した。主観評価実験の評価セットの選択時は、テキストチャットコーパス中の発話のうち、文脈や発話文章から会話の状況が

理解しにくいものを除外した上で、ランダムに選択した。

4.2 実験条件

音声のサンプリング周波数は 22.05 kHz、分析周期は 5 ms とし、スペクトル特徴量として、STRAIGHT 分析 [Kawahara 99] から得られる 40 次メルケプストラムを使用した。また、5 次元の非周期性指標を使用した。

本研究では、次の 5 つの手法

- HMM-BASELINE: 対話行為を用いない HMM 従来法 [Tokuda 00]
- HMM-MIXED: 対話行為をスタイル混合モデルによりモデル化する HMM 従来法 [Yamagishi 05]
- HMM-ADAPT: 対話行為を平均声モデルとモデル適応によりモデル化する HMM 従来法 [Tachibana 06, Yamagishi 09]
- DNN-BASELINE: 対話行為を用いない DNN 従来法 [Zen 13]
- DNN-DACODE: 対話行為を用いる DNN 提案法

について比較を行った。対話行為を考慮する従来法に対する提案法の有用性を検証するため、HMM-MIXED および HMM-ADAPT と DNN-DACODE を比較した。また、対話行為を考慮することの有用性を検証するため、DNN-BASELINE と DNN-DACODE を比較した。また、発語内行為自然性の評価手法として、本研究の主観評価実験方法が妥当であるかを検証するため、HMM-BASELINE と DNN-BASELINE を比較した。対話音声の自然性に関する主観評価結果には、合成音声の発語行為自然性（肉声感等）と発語内行為自然性の両方が影響する可能性がある。特に発語内行為自然性に関する評価を行うためには、肉声感による評価への影響を抑制する必要がある。このため、本研究では、肉声感を評点に反映しないように実験参加者へ教示を行う (4.4.4 節)。ここで、DNN-BASELINE と HMM-BASELINE を比較すると、モデル構造の違いから、DNN-BASELINE がより高い肉声感を示すと考えられる。一方、同一の音声データベースから構築される、対話行為を考慮しない手法であるという共通点から、両者の発語内行為自然性には大きな差はないと考えられる。したがって、これらの 2 つの手法に対する評点が同等であれば、肉声感による評点への影響が抑制されており、発語内行為自然性の評価として本研究の評価手法が妥当であることが検証されることが考えられる。

DNN-BASELINE, DNN-DACODE では、継続長の DNN

モデルの入力ベクトルとして、486 次元の言語情報ベクトルを使用した。音響特徴量の DNN モデルの入力ベクトルでは、フレーム特徴を加え、489 次元の言語情報ベクトルを使用した。出力ベクトルとして、メルケプストラム、対数 F0、非周期性指標、それらの動的特徴、および有声/無声を表すバイナリ変数を連結した 139 次元のベクトルを使用した。入力ベクトルは、0.01 から 0.99 の範囲に正規化し、出力ベクトルは、標準正規化した。継続長の DNN モデルでは、隠れ層数は 2、ユニット数は 256 とした。音響特徴量の DNN モデルでは、隠れ層数は 4、ユニット数は 256 とした。DNN-DACODE では、30 次元の対話行為コードを全ての隠れ層に入力した。隠れ層の活性化関数としてシグモイド関数を、出力層の活性化関数として線形関数を使用した。

DNN の重みパラメータは、乱数により初期化を行った。乱数として、入力ベクトルの次元を D としたとき、平均 0、分散 $\frac{1}{D}$ の正規乱数を使用した。DNN のバイアスパラメータの初期値は、0 とした。重みパラメータおよびバイアスパラメータは、学習データと出力ベクトルの平均二乗誤差を目的関数とし、Adam アルゴリズム [Kingma 14] による誤差逆伝播により学習した。Adam アルゴリズムのパラメータは、音響モデルについて $\alpha = 0.00015$ 、継続長モデルについて $\alpha = 0.001$ 、双方のモデルについて $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8$ とした。

HMM-BASELINE, HMM-MIXED, HMM-ADAPT として、5 状態の left-to-right 型のスキップなしの多空間確率分布隠れセミマルコフモデル (multi-space probability distribution hidden semi-Markov model; MSD-HSMM) [Zen 07] を使用した。観測ベクトルは 40 次メルケプストラム、対数 F0、5 次元の非周期性指標、およびそれらの動的特徴量を連結した 138 次元のベクトルとした。状態出力分布は対角共分散行列を持つガウス関数とした。モデル学習時の MDL [Shinoda 97] のパラメータは $\alpha = 1.0$ とした。

HMM-MIXED では、言語的情報に関する質問に加え、対話行為に関する質問を使用した。従来法 [Yamagishi 05] と同様に特定の発話行為であるかという質問 (30 個) を使用した。また、本研究では、4.1.2 節で示した対話行為類型に関する質問 (8 個) を追加し、合計 38 個の質問を使用した。

HMM-ADAPT では、まず全ての対話行為の音声データを使用して平均声モデルを学習した。平均声モデルの学習時には、各対話行為の大域的な韻律的特徴の差異が木構造に影響するのを防ぐため、speaker adaptive training (SAT) を用いた。続いて、各対話行為に対応する音声データを使用し、平均声モデルを適応することで各対話行為のモデルを学習した。モデル適応アルゴリズムとして、constrained structural maximum a posteriori linear regression (CSMAPLR) と MAP 適応の組み合わせ手法 [Yamagishi 09] を使用した。

評価を行う 5 手法について、全ての対話行為に関する発話を学習データとして使用した。ここで、本研究では、音声データベース全体に含まれる発話数 (3410 発話) に対し、評価セットの発話数 (240 発話) が比較的多数であ

図7 対話行為類型別の、収録音声の平均 F0 および平均話速の散布図。

る。このため、通常の音声合成研究と同様に、評価セットを音声データベースから除外し、学習データを構成する場合、学習データの総量が大きく減少してしまう。そこで、本研究では、評価セットを 10 個のサブセットに分割し、各サブセットについて、モデルの学習および評価用サンプルの生成を行った。各サブセットに関するモデル学習の際には、音声データベースから当該サブセットの発話を除外した全ての音声データを使用した。評価セットをサブセットへ分割する際は、対話行為類型および発話表層類型のバランスを取った。最後に、各サブセットから得られる評価用サンプルを統合し、評価セットとして使用した。

DNN に基づく手法では、学習データの 5% を開発セットとして使用した。また、評価を行う 5 手法について、出力ベクトルに対し最尤パラメータ生成 (maximum likelihood parameter generation; MLPG) [Tokuda 00] を適用し、音響特徴量の時系列を生成した。DNN に基づく手法では、学習データから計算された出力ベクトルの分散を MLPG のために使用した。評価を行う 5 手法すべてについて、生成されたメルケプストラムの時系列に対し、アフィン変換によるポストフィルタ [Silén 12] を適用した。

4.3 客観評価実験

本節では、大域的な韻律的特徴および文末音調の再現性を客観評価する。客観評価尺度として、従来使用される対数 F0 の RMS 誤差、継続長の RMS 誤差に加え、“発話平均対数 F0 の RMS 誤差”、“発話平均継続長の RMS 誤差”を使用する。これらの発話平均に対し算出される RMS 誤差 $D_{utterance}$ は、

$$D_{utterance} = \sqrt{\frac{1}{N} \sum_{n=1}^N (\bar{x}_n - \bar{y}_n)^2} \quad (2)$$

で表される。ここで、評価セット中の発話数を N 、発話の番号を $n (= 1, \dots, N)$ 、発話単位で算出される合成音声の平均特徴量 (平均対数 F0 または平均継続長) を $\bar{x}_n$ 、自然音声の平均特徴量を $\bar{y}_n$ とした。 $D_{utterance}$ により、発話全体の平均として、声の高さ/低さや、話速の速さ/遅さが再現されたかを評価することができる。文末音調の

図 8 発話平均対数 F0 および発話平均継続長の RMS 誤差。

再現性の尺度としては、収録音声との文末（最終 1 音節）F0 の相関を使用する。文末 F0 の相関は対話行為のみでなく文末形式にも依存すると考えられるため、発話文章類型ごとに結果を示す。F0 に関する客観評価では、収録音声の継続長を使用し生成される F0 を使用する。以下では、手法間で大きな傾向差が見られた対話行為類型・発話表層類型のみにについて考察を行う。

§1 発話平均対数 F0 および発話平均継続長の RMS 誤差

まず、大域的な韻律的特徴が顕著な対話行為を確認するため、図 7 に対話行為類型別に算出される、収録音声の平均 F0、平均話速を示した。

- 自問、フィラー：F0 が低い、話速が遅い
- 感嘆：F0 が高い、話速が遅い
- 相槌：話速が遅い

という傾向が特に顕著であることがわかる。これらの傾向は、図 2 の分析結果と同様である。本節では、特に、これらの対話行為類型における傾向が再現されるかを確認する。

図 8 に、発話平均対数 F0 および発話平均継続長の RMS 誤差を示した。まず、対話行為を考慮する従来法に対する提案法の有用性を確認するため、HMM-MIXED、HMM-ADAPT と DNN-DACODE を比較する。全対話行為類型の平均では、発話平均対数 F0、発話平均継続長の RMS 誤差は、

- HMM-MIXED > HMM-ADAPT $\approx$ DNN-DACODE となる。ただし、発話平均 F0 について 7.0 cent、発話平均継続長について 1.0 ms 以上の差を不等号で、それ以下の差を $\approx$ で表した。ここで、対話行為類型別の比較を行うと、発話平均 F0 の RMS 誤差は、

- HMM-MIXED $\approx$ HMM-ADAPT > DNN-DACODE (自問)
- HMM-MIXED > HMM-ADAPT $\approx$ DNN-DACODE (フィラー、感嘆)

となり、発話平均継続長の RMS 誤差は、

- HMM-MIXED > HMM-ADAPT $\approx$ DNN-DACODE (自問、フィラー、感嘆)

となる。以上から、発話平均の RMS 誤差について、提案法は HMM-MIXED よりも優位であ

り、HMM-ADAPT と同等であることが示された。HMM-MIXED では、合成する対話行為に関する質問が決定木に現れない場合があるため、合成する対話行為の大域的な韻律的特徴が十分に再現されなかったと考えられる。

続いて、対話行為を考慮することの有用性を評価するため、DNN-BASELINE と DNN-DACODE を比較する。全対話行為類型の平均を比較すると、発話平均 F0、発話平均継続長の RMS 誤差は、

- DNN-BASELINE > DNN-DACODE (自問、フィラー、感嘆)

となる。このことから、大域的な韻律的特徴の再現性について、提案法は DNN-BASELINE よりも優位であることが示された。これにより、対話行為を考慮する有用性が示された。

一方、収録音声について、“相槌”の話速が遅い特徴が確認されたが、全ての合成音声手法について、平均継続長の RMS 誤差は同程度であった。対話行為を考慮しない音声合成手法においても継続長の傾向が再現された理由は、この傾向が発話文章に由来するためであると考えられる。“相槌”で使用する発話文章は、例えば、「あー」、「ねえ」など、長母音を使用する発話が多い。継続長の傾向は、この発話文章の特徴に由来するものであるため、対話行為を考慮しない音声合成手法においても再現されたと考えられる。

本節の実験結果から、対話行為の持つ大域的な韻律的特徴の再現性について、提案法は HMM-MIXED に対し優位であり、HMM-ADAPT と同等であることが示された。また、提案法は DNN-BASELINE に対し優位であり、平均 F0 および平均継続長の再現性の観点から、対話行為を考慮することの有用性が示された。

§2 対数 F0 および継続長の RMS 誤差の評価結果

図 9 に対数 F0 および継続長の RMS 誤差を示した。まず、対話行為を考慮する従来法に対する提案法の有用性を確認するため、HMM-MIXED、HMM-ADAPT と DNN-DACODE を比較する。全対話行為類型の平均では、対数 F0 の RMS 誤差について、

- HMM-MIXED > HMM-ADAPT $\approx$ DNN-DACODE 継続長の RMS 誤差について、

図9 対数 F0 および継続長の RMS 誤差.

● HMM-MIXED $\equiv$ HMM-ADAPT $>$ DNN-DACODE となることわかる。ただし、4.3.1 節と同様に、対数 F0 について 7.0 cent, 継続長について 1.0 ms 以上の差を不等号で、それ以下の差を $\equiv$ で表した。RMS 誤差の観点から、従来法に対する提案法の優位性が示された。

続いて、提案法により対話行為を考慮する有用性を確認するため、DNN-BASELINE と DNN-DACODE を比較する。図 9 から、F0 および継続長の RMSE は、“繰り返し/言い換え”の F0 を除き、

● DNN-BASELINE $>$ DNN-DACODE

となることわかる。F0 および継続長の生成誤差の観点から、提案法により対話行為を考慮する有用性が示された。

本節の実験結果から、F0 および継続長の生成誤差に関して、HMM-MIXED および HMM-ADAPT に対する提案法の優位性が示された。また、F0 および継続長の生成誤差に関して、DNN-BASELINE に対する提案法の優位性が示された。これにより、F0 および継続長の生成誤差の観点から、提案法により対話行為を考慮する有用性が示された。

4.3.1 節、4.3.2 節の結果から、大域的な韻律的特徴の再現性について、提案法は、HMM-MIXED よりも優位であった。また、提案法は、HMM-ADAPT と比べると、発話平均 F0 および F0 の RMS 誤差に関しては同等であった。一方、継続長の RMS 誤差は改善されたものの、発話平均継続長の RMS 誤差は同等であったため、継続長に関する大域的な韻律的特徴の再現性についても、両者は同等とする。以上から、大域的な韻律的特徴の再現性について、HMM-ADAPT と提案法は同等であると結論づける。また、提案法は DNN-BASELINE よりも優位であり、大域的な韻律的特徴の再現性の観点から、提案法により対話行為を考慮する有用性が示された。

§3 文末 F0 の相関の評価結果

表 5 に、合成音声と収録音声の文末 F0 の相関を示した。まず、対話行為を考慮する従来法に対する提案法の有用性を確認するため、文末 F0 の相関を評価セット全体で平均し、比較すると、

● HMM-ADAPT $<$ HMM-MIXED $\equiv$ DNN-DACODE となることわかる。ただし、0.05 以上の相関の差を不

等号で表し、それ以下の差を $\equiv$ で表した。文末 F0 の再現性について、提案法は HMM-ADAPT に対して優位であり、HMM-MIXED と同等であることが示された。HMM-ADAPT では、文末音調の傾向を再現するのに適した木構造が得られないため、文末音調が十分に再現されなかったと考えられる。

特に、対話行為類型・発話表層類型別の傾向を確認すると、提案法の文末 F0 の相関が HMM-ADAPT に比べ 0.2 以上改善したのは、

- 質問-3 (文末詞なし)
- 相槌-2, 感嘆-3 (完全文-か)

であった。ただし、“質問の発話表層類型 3”を“質問-3”などのように省略した。これらの類型では、HMM-ADAPT に比べ、提案法の発話内行為自然性が改善されると期待される。

続いて、文末音調の再現性の観点から、対話行為を考慮する有用性を確認するため、DNN-DACODE と DNN-BASELINE を比較する。評価セット全体の文末 F0 の相関が改善されたことから、提案法により対話行為を考慮することの有用性が示された。また、対話行為類型・発話表層類型別の傾向を確認すると、DNN-DACODE により文末 F0 の相関が 0.3 以上改善されたのは、

- 質問-3 (文末詞なし)
- フィラー-2 (完全文-ね)
- 繰り返し/言い換え-2, 相槌-2, 感嘆-3 (完全文-か)

であった。これらの類型では、DNN-BASELINE に比べ、提案法の発話内行為自然性が改善されると期待される。

本節の実験結果から、文末 F0 の再現性について、提案法は、HMM-ADAPT に対し優位であり、HMM-MIXED と同程度であることが示された。また、提案法は DNN-BASELINE に対し優位であり、文末 F0 の再現性の観点から、提案法により対話行為を考慮する有用性が示された。

4.3.1 節、4.3.2 節、4.3.3 節の客観評価実験結果から、提案法は、HMM-MIXED に対し、大域的な韻律的特徴の再現性について優位であり、文末音調の再現性について同程度であることが示された。一方、提案法は、HMM-ADAPT に対し、大域的な韻律的特徴の再現性について同程度であり、文末音調の再現性について優位であることが示された。また、提案法は、DNN-BASELINE

表 5 文末 1 音節の F0 相関
(a) HMM-MIXED (括弧内は HMM-BASELINE からの差分)

対話行為 類型	発話文章類型			
	類型 1	類型 2	類型 3	
情報伝達	0.56 (+0.06)	0.96 (+0.01)	0.79 (+0.04)	
質問	0.58 (+0.07)	0.78 (+0.00)	0.86 (+0.74)	
自問	0.41 (-0.01)	0.74 (+0.09)	0.71 (+0.01)	
繰り返し/ 言い換え	0.95 (+0.05)	0.72 (+0.58)	0.77 (+0.38)	
相槌	0.79 (-0.06)	0.81 (+0.93)	0.88 (+0.01)	
フィラー	0.69 (-0.02)	0.76 (+0.46)	0.80 (+0.29)	
感嘆	0.62 (-0.18)	0.33 (+0.10)	0.25 (+0.90)	
定型文	0.70 (+0.09)	0.88 (+0.25)	0.83 (+0.39)	
平均		0.72 (+0.22)		

(b) HMM-ADAPT (括弧内は HMM-BASELINE からの差分)

対話行為 類型	発話文章類型			
	類型 1	類型 2	類型 3	
情報伝達	0.68 (+0.18)	0.96 (+0.01)	0.78 (+0.03)	
質問	0.46 (-0.06)	0.83 (+0.05)	0.35 (+0.23)	
自問	0.41 (-0.01)	0.58 (-0.06)	0.78 (+0.08)	
繰り返し/ 言い換え	0.95 (+0.04)	0.87 (+0.73)	0.64 (+0.25)	
相槌	0.86 (+0.01)	0.56 (+0.67)	0.92 (+0.05)	
フィラー	0.67 (-0.04)	0.68 (+0.38)	0.77 (+0.26)	
感嘆	0.69 (-0.11)	0.43 (+0.21)	-0.27 (+0.38)	
定型文	0.73 (+0.13)	0.90 (+0.27)	0.85 (+0.41)	
平均		0.67 (+0.17)		

(c) DNN-DA (括弧内は DNN からの差分)

対話行為 類型	発話文章類型			
	類型 1	類型 2	類型 3	
情報伝達	0.60 (-0.00)	0.94 (+0.02)	0.77 (+0.32)	
質問	0.64 (-0.01)	0.89 (+0.06)	0.57 (+0.40)	
自問	0.56 (+0.11)	0.61 (+0.16)	0.76 (+0.05)	
繰り返し/ 言い換え	0.96 (+0.04)	0.93 (+0.52)	0.73 (+0.11)	
相槌	0.81 (+0.02)	0.92 (+1.07)	0.91 (-0.03)	
フィラー	0.65 (-0.12)	0.61 (+0.37)	0.71 (+0.28)	
感嘆	0.82 (+0.07)	0.37 (-0.00)	0.18 (+0.79)	
定型文	0.72 (+0.03)	0.93 (+0.01)	0.72 (+0.16)	
平均		0.73 (+0.19)		

に対し、大域的な韻律的特徴の再現性および文末音調の再現性について優位であり、提案法により対話行為を考慮する有用性が示された。

4.4 主観評価実験

§1 発語内行為自然性の評価方法

各手法で合成された音声について、発語内行為自然性を評価するため、主観評価実験を行った。発話意図は、文脈、発話文章、韻律などを総合して推論される情報である。このため、発語内行為自然性を評価するためには、評価を行う各発話について、文脈を提示する必要があると考えられる。本研究では、図 10 に示す GUI を使用し、主観評価実験を行う。まず、実験参加者に対し、ロボットと対話している場面を想像し、提示された文脈を読むように指示する。続いて、実験参加者は、文脈に続くロボットの発話を想定し音声サンプルを聴取する。最後に、提示された発話意図に対応する音声として自然であるかについて評価を行う。実際の音声対話では、発語内行為自然性は、ユーザ発話やシステムの過去の発話のパラ言語・非言語情報にも影響されると考えられるが、本研究では、考慮する実験条件数を減少させるため、テキストによる

インストラクション：

1. あなたは、りこさんというロボットと会話をしています。ちょうど下記の対話を終えたところを想像してください。
2. 対話に続くりこさんの発話として、音声を再生して聞いてください。
3. 下記の質問に答えてください。

対話文章：(部屋で雑談をしているところです。)

りこさん： 私は一風堂が好きです。
あなた： 私もです。一蘭は好きですか？

▶ 0:00 ● 〴〵

質問：

下記の気持ちで発話した音声として、自然に感じますか？

※ 合成音のような肉声感の低い音声もありますが、肉声感ではなく、話し方（声の高低や話す速度など）について評価してください。

りこさんは、フィラーを発話した。
(発話内容を考えている。また、続けて何かを話そうとしている。)

○1 (不自然) ○2 (やや不自然) ○3 (どちらとも言えない) ○4 (やや自然) ○5 (自然)

戻る

次へ

図 10 主観評価実験で使った GUI の例 (評価サンプルの発話文章：「そうですね」)。

表 6 評価セット全体に対する発語内行為自然性の MOS 値と 99% 信頼区間。

手法	MOS 値
HMM-BASELINE	3.01 ± 0.07
HMM-MIXED	3.36 ± 0.07
HMM-ADAPT	3.48 ± 0.07
DNN-BASELINE	3.02 ± 0.07
DNN-DACODE	3.63 ± 0.06

文脈の提示を行った。発話意図を提示する際は、その名称 (例：「フィラー」) のみを提示した場合、その定義が実験参加者に十分に理解されず、評価結果の信頼性を損ねる懸念がある [Hojo 19] ため、発話意図の定義の説明文 (例：「発話内容を考えている。続けて何かを話そうとしている」) を合わせて提示する。また、HMM に基づく手法と DNN に基づく手法の肉声感 (発語行為自然性) の違いによる評価結果への影響を抑制するため、「音声の肉声感ではなく、話し方 (声の高低や話す速度など) について評価してください。」という指示を画面に表示した。また、文脈や発話意図に対するセリフ (発話文章) の自然さではなく、話し方について評価するよう教示した。

評価用文章として、対話行為類型、発話文章類型ごとに 5 発話、合計 120 発話を使用した。本研究の音声データベースを構築する際に使用したテキストチャットコーパスから、各発話の先行発話を数発話抽出し、提示する文脈として使用した。20 名の健聴の日本語話者 (男性 10 名、女性 10 名) が評価を行った。実験参加者は、防音室内でヘッドホンにより音声サンプルを受聴した。評価時間に関する制約は設けなかった。発語内行為自然性について、5 段階 (1：不自然、2：やや不自然、3：どちらとも言えない、4：やや自然、5：自然) の MOS 試験を行った。音声サンプルは、ランダムな順序で提示した。

表 6 に主観評価実験結果を示した。また、図 11 に発話行為類型・発話文章類型別の評価結果を示した。各対話行為類型・発話文章類型別に t-検定を行った結果を表 7 に示した。対話行為類型・発話文章類型別の統計検定は、多重比較であるため、通常の音声自然性評価と同様に有意水準 5% の統計検定を行った場合、偽陽性の有意差が得られる可能性が高い。そこで、本研究では、有意水準を 1% に設定した上で、統計検定の結果を客観評価実験結果と対応づけて考察することで、その妥当性を確認する。表 7 に

図 11 発語内行為自然性に関する主観評価実験結果. 99%信頼区間をエラーバーで示した.

は, a) ベースライン手法間の比較 (HMM-BASELINE と DNN-BASELINE), b) 対話行為を考慮する従来法と提案法の比較 (HMM-MIXED と DNN-DACODE, HMM-ADAPT と DNN-DACODE) c) 対話行為の考慮の有無による比較 (DNN-BASELINE, DNN-DACODE) を行った結果を示した.

§2 評価セット全体の平均評点による比較

表 6 から, 評価セット全体での MOS 値は,

- HMM-BASELINE $\approx$ DNN-BASELINE < HMM-MIXED < HMM-ADAPT < DNN-DACODE

となることがわかる. 有意水準 1% の統計検定により, HMM-BASELINE と DNN-BASELINE との間には有意差が確認されず, その他の手法の組み合わせには有意差が確認された. ここで, HMM-BASELINE と DNN-BASELINE との間には, 合成音声の発語行為自然性 (肉声感) に差があると考えられるものの, MOS 値には統計的有意差が確認されなかった. このことから, 本研究の主観評価実験で使用した質問や教示により, 合成音声の肉声感による MOS 値への影響は十分に抑制されたと考えられる. したがって, 提案法による MOS 値の改善は, DNN により合成音声の発語行為自然性 (肉声感) が改善

されたためではなく, 発話意図がより自然に感じられたためであると考えられる. 以上から, 提案法は, HMM-MIXED および HMM-ADAPT に比べ, より高い発語内行為自然性を示すことが示された.

§3 ベースライン手法間の比較

ここで, 表 7 の結果から, 対話行為類型別・発話文章類型別に HMM-BASELINE と DNN-BASELINE を比較すると, “質問-2” において, DNN-BASELINE の MOS 値がより高いことがわかる. ただし, この MOS 値の差は, 発語内行為自然性の改善ではなく, 長い発話文章を自然に合成する上で, DNN が有用であったためであると考えられる. まず, 発話文章の長さ (モーラ数) と MOS 値の相関を分析した結果を表 8 に示した. 分析には, 対話行為による MOS 値への影響が比較的小さいと考えられる, “情報伝達” および “質問-1” の 20 発話のみを使用した. 分析結果から, HMM に基づく手法では負の相関が得られた一方, DNN に基づく手法では高い相関は得られなかった. 次に, 表 A.1 の評価セットを確認すると, “質問-2” では, 他の類型に比べ, 比較的長い発話文章が評価セットに含まれる (「じゃあトレーニングも本格的にやって, からだをつくらないといけないですね」, 「そのうえお

表7 MOS 値の t 検定による比較 ($\alpha = 0.01$)。手法名の略称として, HMM-B : HMM-BASELINE, HMM-M : HMM-MIXED, HMM-A : HMM-ADAPT, DNN-B : DNN-BASELINE, DNN-D : DNN-DACODE を使用した。a) : ベースライン手法間の比較 (HMM-B と DNN-B)。b) : 対話行為を考慮する手法間の比較 (HMM-M と DNN-D, HMM-A と DNN-D)。c) : 対話行為の考慮の有無による比較 (DNN-B と DNN-D)。

対話行為 類型	発語文章類型		
	類型 1	類型 2	類型 3
情報伝達	-	-	-
質問	-	a) HMM-B < DNN-B b) HMM-M < DNN-D b) HMM-A < DNN-D	b) HMM-A < DNN-D c) DNN-B < DNN-D
自問	b) HMM-M < DNN-D c) DNN-B < DNN-D	c) DNN-B < DNN-D	c) DNN-B < DNN-D
繰り返し/言い換え	-	c) DNN-B < DNN-D	-
相槌	-	b) HMM-A < DNN-D c) DNN-B < DNN-D	-
フィラー	b) HMM-M < DNN-D c) DNN-B < DNN-D	b) HMM-M < DNN-D c) DNN-B < DNN-D	b) HMM-M < DNN-D c) DNN-B < DNN-D
感嘆	c) DNN-B < DNN-D	c) DNN-B < DNN-D	b) HMM-M < DNN-D b) HMM-A < DNN-D c) DNN-B < DNN-D
定型文	b) HMM-M < DNN-D	b) HMM-M < DNN-D c) DNN-B < DNN-D	b) HMM-M < DNN-D

表8 “情報伝達” および “質問-1” の評価文章 20 発話に関する、モータ数と MOS 値の相関分析結果。

手法	相関係数	p 値
HMM-BASELINE	-0.29	0.21
HMM-MIXED	-0.33	0.16
HMM-ADAPT	-0.55	0.01
DNN-BASELINE	0.10	0.69
DNN-DACODE	-0.14	0.55

子さんがいらっしやると、生活がかなり制限されますね」など)。以上から、“質問-2”において、DNN-BASELINE の MOS 値が HMM-BASELINE よりも改善された理由は、長い発話文章を合成する上で HMM よりも DNN が有用であり、かつ、その有用性が“質問-2”の評価セットにおいてより顕著に現れたためであると考えられる。つまり、“質問”の意図を伝達する上での発語内行為自然性が向上したためではなく、本実験の評価セットの設計に依存した問題であると考えられる。“質問-2”以外の類型では、DNN-BASELINE と HMM-BASLINE の間に MOS 値の有意な差は見られなかったため、上述の傾向は、発語内行為自然性の評価結果としての妥当性を損なうものではないと考えられる。今後の課題として、発話文章のみでなく、発話文章の長さを統制した評価セットの設計が考えられる。

§4 提案法と対話行為を考慮する従来法の比較

対話行為を考慮する従来法に対する提案法の有用性を確認するため、HMM-MIXED, HMM-ADAPT および DNN-DACODE の評価結果を対話行為類型・発話文章類型別に比較する。

まず、HMM-MIXED と DNN-DACODE を比較する。表7の結果から、各対話行為類型および発話文章類型において、DNN-DACODE は HMM-MIXED と同等か、より高い自然性を示すことがわかる。HMM-MIXED に対する提案法の有用性が示された。提案法により自然性が向上した対話行為類型・発話文章類型は、

- (1) 質問-2
- (2) 自問-1, フィラー-{1, 2, 3}, 感嘆-3
- (3) 定型文-{1, 2, 3}

であった。ここで、(1)において提案法の MOS 値が改善

した理由は、4.4.3 節と同様に、DNN により長い発話を自然に合成することが可能となったためであり、発語内行為自然性の向上によるものではないと考えられる。また、(2)において提案法の MOS 値が改善した理由は、4.3.1 節の分析結果から、提案法により、“自問”、“フィラー”、“感嘆”の対話行為類型において、大域的な韻律的特徴の再現性が改善されたためであると考えられる。ただし、

(i) 自問-{2, 3}

(ii) 感嘆-{1, 2}

では、(2)と同様に大域的な韻律的特徴が改善されたものの、MOS 値の改善は見られなかった。このうち、(i)については、DNN-DACODE の MOS 値が上回る傾向が同様に確認されたものの、統計的有意差は確認されなかった。一方、(ii)については、DNN-DACODE と HMM-MIXED の MOS 値の間に一貫した傾向が確認されなかった。“感嘆-1 (独立詞文)”および“感嘆-2 (述語省略文)”では、発話文章のみから、感嘆の発話であることが推論可能である(例えば、「あははは」「びっくり」など)一方、“感嘆-3 (完全文_か)”の発話文章は、“感嘆”の韻律的特徴が再現されなければ、他の対話行為(“質問”など)の発話として推論される可能性が高い(例えば、「駅二つですか」など)。このため、“感嘆-{1, 2}”では、“感嘆-3”に比べ、“感嘆”の意図を自然に表現するために、大域的な韻律的特徴を顕著に再現する必要性が小さかったと考えられる。一方、“感嘆-{1, 2}”においても、伝達されるべき感情の強さなどを指定して発語内行為自然性を評価する場合、DNN-DACODE の自然性がより高く評価される可能性がある。今後の課題として、感情表出の強さの観点を含めた評価が挙げられる。

(3)については、本研究の客観評価実験結果との明確な対応関係が得られなかった。今後の課題として、定型文の発語内行為自然性に寄与する要因の分析が挙げられる。

以上の提案法と HMM-MIXED の比較により、各対話行為類型・発話文章層において、提案法は、HMM-MIXED と比べ、同等かより高い発語内行為自然性を示すことが明らかとなった。客観評価実験結果との対応から、大域的な韻律的特徴の再現性が改善されたためであると考え

られる。

続いて、DNN-DACODE と HMM-ADAPT を比較する。表 7 の結果から、各対話行為類型・発話文章類型において、DNN-DACODE は HMM-ADAPT と同等か、より高い MOS 値を示すことがわかる。HMM-ADAPT に対する提案法の有用性が示された。提案法の MOS 値が改善したのは、

(1) 質問-2

(2) 質問-3, 相槌-2, 感嘆-3

であった。ここで、(1) で提案法の MOS 値が改善した理由は、4.4.3 節や HMM-MIXED と提案法の比較と同様に、DNN により長い発話を自然に合成することが可能となったためであり、発語内行為自然性の向上によるものではないと考えられる。また、4.3.3 節の評価結果から、(2) は、提案法の文末 F0 の相関が HMM-ADAPT よりも 0.2 以上高い類型と一致する。(2) では、文末 F0 の相関の改善により、提案法の発語内行為自然性が改善されたと考えられる。

以上の提案法と HMM-ADAPT の比較から、各対話行為類型・発話表層類型において、提案法は、HMM-ADAPT と比べ、同等かより高い発語内行為自然性を示すことが明らかとなった。客観評価実験結果との対応から、文末音調の再現性が改善されたためであると考えられる。

本節の実験結果から、提案法は、各対話行為類型・発話表層類型において、HMM-MIXED および HMM-ADAPT に比べ、同等かより高い発語内行為自然性を示すことが明らかとなった。客観評価実験結果との対応から、HMM-MIXED に比べ大域的な韻律的特徴の再現性が改善されたことと、HMM-ADAPT に比べ文末音調の再現性が改善されたことが寄与したと考えられる。

§5 対話行為の考慮の有無による比較

最後に、提案法により対話行為を考慮することの有用性を確認するため、DNN-BASELINE と DNN-DACODE を比較する。表 7 の結果から、各対話行為類型および発話文章類型において、DNN-DACODE は DNN-BASELINE と同等か、より高い MOS 値を示すことがわかる。DNN に基づく音声合成において、対話行為を考慮することで発語内行為自然性が改善されることが示された。提案法により発語内行為自然性の改善が示されたのは、

(1) 自問-{1, 2, 3}, フィラー-{1, 2, 3}, 感嘆-{1, 2, 3}

(2) 質問-類型 3, 繰り返し/言い換え-2, 相槌-2

(3) 定型文-2

であった。4.3.1 節の評価結果から、(1) では、DNN-DACODE により、平均 F0 および平均継続長の傾向が再現されることがわかる。これらの対話行為類型では、大域的な韻律的特徴が再現されることで、発語内行為自然性が向上したと考えられる。また、4.3.3 節の評価結果から、(1) のフィラー-2, 感嘆-3 および (2) では、DNN-DACODE の文末 F0 の相関は DNN-BASELINE と比べ、いずれも 0.3 以上改善することがわかる。これらの対話行為類型・発話文章類型では、文末音調の傾向が再現されることで、発語内行為自然性が改善されたと考えられる。(3) については、4.4.4 節における提案法と HMM-

MIXED との比較と同様に、本研究の客観評価実験結果との明確な対応関係が得られなかった。今後の課題として、定型文の発語内行為自然性に寄与する要因の分析が挙げられる。

4.4 節の主観評価実験結果から、提案法は、HMM に基づく従来法に比べ、より高い発語内行為自然性を示すことが示された。主観評価実験結果と客観評価実験結果との対応から、提案法は、HMM-MIXED に比べ、大域的な韻律的特徴の再現性が改善されたため、発語内行為自然性が向上したと考えられる。一方、提案法は、HMM-ADAPT に比べ、文末音調の再現性が改善されたため、発語内行為自然性が向上したと考えられる。また、提案法は、DNN-BASELINE に比べ、より高い発語内行為自然性を示すことが明らかとなった。提案法により、対話行為を考慮することで発語内行為自然性が改善することが示された。

5. 結 論

本研究では、音声対話システムから生成される対話音声について、「話し手の発話意図が聞き手に自然に感じられるか」を発語内行為自然性と定義し、発語内行為自然性を向上させる音声合成技術を検討した。発話意図として、雑談対話システムで使用される対話行為を使用し、対話行為が付与された発話音声により構成される音声データベースを構築した。DNN に基づき対話行為を表現する音声合成手法を提案し、HMM に基づく 2 つの従来法(スタイル混合モデルおよびモデル適応手法)と比較した。主観評価実験の結果から、提案法は、HMM に基づく従来法に比べ、より高い発語内行為自然性を示すことが明らかとなった。主観評価実験結果と客観評価実験結果との対応から、提案法は、スタイル混合モデルに比べ、大域的な韻律的特徴の再現性が改善され、発語内行為自然性が向上したと考えられる。一方、提案法は、モデル適応手法に比べ、文末音調の再現性が改善され、発語内行為自然性が向上したと考えられる。また、提案法と対話行為を考慮しない DNN 音声合成手法を比較すると、主観評価実験の結果から、提案法の発語内行為自然性が改善されることが明らかとなった。提案法により、対話行為を考慮することで発語内行為自然性が改善することが示された。今後の課題として、“感嘆”などの対話行為において、その表出の強さの制御と評価が挙げられる。また、本研究で考慮した文末形式以外の特徴が発語内行為自然性の評価結果に影響することが考えられるため、より再現性の高い評価結果を得るためには、様々な発話文章の特徴と発語内行為自然性の関係を考慮した評価セットの設計が必要であると考えられる。

◇ 参 考 文 献 ◇

- [An 17] An, S., Ling, Z., and Dai, L.: Emotional statistical parametric speech synthesis using LSTM-RNNs, in *Proc. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1613–1616 (2017)
- [Austin 75] Austin, J. L.: *How to Do Things with Words*, Oxford University Press (1975)

- [Barra-Chicote 10] Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., and Macias-Guarasa, J.: Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech, *Speech Communication*, Vol. 52, No. 5, pp. 394–404 (2010)
- [Bezooijen 97] Bezooijen, van R. and Heuven, van V. J.: Assessment of synthesis systems, *Handbook of Standards and Resources for Spoken Language Systems*, pp. 481–563 (1997)
- [Fujisaki 97] Fujisaki, H.: Prosody, models, and spontaneous speech, in *Computing Prosody*, pp. 27–42, Springer (1997)
- [Grice 91] Grice, H. P.: *Studies in the Way of Words*, Harvard University Press (1991)
- [Hellbernd 16] Hellbernd, N. and Sammler, D.: Prosody conveys speaker's intentions: Acoustic cues for speech act perception, *Journal of Memory and Language*, Vol. 88, pp. 70–86 (2016)
- [Hojo 18] Hojo, N., Ijima, Y., and Mizuno, H.: DNN-based speech synthesis using speaker codes, *IEICE Transactions on Information and Systems*, Vol. 101, No. 2, pp. 462–472 (2018)
- [Hojo 19] Hojo, N. and Miyazaki, N.: Evaluating intention communication by TTS using explicit definitions of illocutionary act performance, *Proc. Interspeech 2019*, pp. 1536–1540 (2019)
- [ITU-T 94] ITU-T.: A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices (1994)
- [Kawahara 99] Kawahara, H., Masuda-Katsuse, I., and De Cheveigné, A.: 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)
- [Kingma 14] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014)
- [Kurematsu 90] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., and Shikano, K.: ATR Japanese speech database as a tool of speech recognition and synthesis, *Speech Communication*, Vol. 9, No. 4, pp. 357–363 (1990)
- [Lorenzo-Trueba 18] Lorenzo-Trueba, J., Henter, G. E., Takaki, S., Yamagishi, J., Morino, Y., and Ochiai, Y.: Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis, *Speech Communication*, Vol. 99, pp. 135–143 (2018)
- [Nose 07] Nose, T., Yamagishi, J., Masuko, T., and Kobayashi, T.: A style control technique for HMM-based expressive speech synthesis, *IEICE Transactions on Information and Systems*, Vol. 90, No. 9, pp. 1406–1413 (2007)
- [Nose 17] Nose, T., Arao, Y., Kobayashi, T., Sugiura, K., and Shiga, Y.: Sentence selection based on extended entropy using phonetic and prosodic contexts for statistical parametric speech synthesis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 5, pp. 1107–1116 (2017)
- [Scherer 77] Scherer, K. R. and Oshinsky, J. S.: Cue utilization in emotion attribution from auditory stimuli, *Motivation and emotion*, Vol. 1, No. 4, pp. 331–346 (1977)
- [Searle 80] Searle, J. R., Kiefer, F., Bierwisch, M., et al.: *Speech Act Theory and Pragmatics*, Vol. 10, Springer (1980)
- [Shinoda 97] Shinoda, K. and Watanabe, T.: Acoustic modeling based on the MDL criterion for speech recognition, in *Proc. Eurospeech 1997*, pp. 99–102 (1997)
- [Silén 12] Silén, H., Helander, E., Nurminen, J., and Gabbouj, M.: Ways to implement global variance in statistical speech synthesis, in *Proc. 13th Annual Conference of the International Speech Communication Association* (2012)
- [Skerry-Ryan 18] Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R., and Saurous, R. A.: Towards end-to-end prosody transfer for expressive speech synthesis with tacotron, *arXiv preprint arXiv:1803.09047* (2018)
- [Stolcke 00] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech, *Computational Linguistics*, Vol. 26, No. 3, pp. 339–373 (2000)
- [Syrdal 08] Syrdal, A. K. and Kim, Y.-J.: Dialog speech acts and prosody: Considerations for TTS, in *Proc. Speech Prosody*, pp. 661–665 (2008)
- [Tachibana 06] Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T.: A style adaptation technique for speech synthesis using HSMM and suprasegmental features, *IEICE Transactions on Information and Systems*, Vol. 89, No. 3, pp. 1092–1099 (2006)
- [Tokuda 00] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T.: Speech parameter generation algorithms for HMM-based speech synthesis, in *Proc. ICASSP 2000*, Vol. 3, pp. 1315–1318 (2000)
- [Tsiakoulis 14] Tsiakoulis, P., Breslin, C., Gasic, M., Henderson, M., Kim, D., Szummer, M., Thomson, B., and Young, S.: Dialogue context sensitive HMM-based speech synthesis, in *Proc. ICASSP 2014*, pp. 2554–2558 (2014)
- [Wilson 12] Wilson, D. and Sperber, D.: *Meaning and Relevance*, Cambridge University Press (2012)
- [Wu 15] Wu, Z., Swietojanski, P., Veaux, C., Renals, S., and King, S.: A study of speaker adaptation for DNN-based speech synthesis, in *Proc. Interspeech 2015* (2015)
- [Wu 16] Wu, Z., Watts, O., and King, S.: Merlin: an open source neural network speech synthesis system, in *Proc. 9th ISCA Speech Synthesis Workshop*, pp. 202–207 (2016)
- [Xue 18] Xue, L., Zhu, X., An, X., and Xie, L.: A comparison of expressive speech synthesis approaches based on neural network, in *Proc. the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and 1st Multi-Modal Affective Computing of Large-Scale Multimedia Data*, pp. 15–20 (2018)
- [Yamagishi 05] Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T.: Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis, *IEICE Transactions on Information and Systems*, Vol. 88, No. 3, pp. 502–509 (2005)
- [Yamagishi 07] Yamagishi, J., Kobayashi, T., Tachibana, M., Ogata, K., and Nakano, Y.: Model adaptation approach to speech synthesis with diverse voices and styles, in *Proc. ICASSP 2007*, Vol. 4, pp. 1233–1236 (2007)
- [Yamagishi 09] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J.: Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm, *IEEE Transactions on Audio, Speech, and Language Process.*, Vol. 17, No. 1, pp. 66–83 (2009)
- [Zen 07] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T.: A hidden semi-Markov model-based speech synthesis system, *IEICE Transactions on Information and Systems*, Vol. 90, No. 5, pp. 825–834 (2007)
- [Zen 13] Zen, H., Senior, A., and Schuster, M.: Statistical parametric speech synthesis using deep neural networks, in *Proc. ICASSP 2013*, pp. 7962–7966 (2013)
- [井上 93] 井上 優: 発話における「タイミング考慮」と「矛盾考慮」: 命令文・依頼文を例に, 国立国語研究所報告 105 研究報告集, Vol. 14, pp. 333–360 (1993)
- [岩田 17] 岩田 和彦, 小林 哲則: 対話音声合成の表現力向上に向けた文末詞と音調の組合せによる発話意図の表現に関する実験的検討, 電子情報通信学会論文誌 D, Vol. 100, No. 11, pp. 938–948 (2017)
- [郡 15] 郡 史郎: 日本語の文末イントネーションの種類と名称の再検討, 言語文化研究, Vol. 41, pp. 85–107 (2015)
- [郡 18] 郡 史郎: 終助詞類のアクセントとイントネーション: 「よ」「か」「の」「な」「でしよう」「じゃない」, とびはね音調の「ない」, 言語文化共同研究プロジェクト, Vol. 2017, pp. 13–26 (2018)
- [轟木 08] 轟木 靖子: 東京語の終助詞の音調と機能の対応について-内省による考察, 音声言語, pp. 5–28 (2008)
- [上村 89] 上村 幸雄: 日本語のイントネーション, ことばの科学, Vol. 3, pp. 193–220 (1989)
- [森山 89] 森山 卓郎: 文の意味とイントネーション, 講座 日本語と日本語教育, Vol. 1, (1989)
- [杉藤 01] 杉藤 美代子: 終助詞「ね」の意味・機能とイントネーション, 文法と音声, pp. 3–16 (2001)
- [前川 02] 前川 喜久雄, 北川 智利: 音声はパラ言語情報をいかに伝えるか, 認知科学, Vol. 9, No. 1, pp. 46–66 (2002)
- [白川 09] 白川 博之: 「言いさし文」の研究, くろしお出版 (2009)
- [目黒 12] 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)

[担当委員: 岡田 将吾]

2019 年 8 月 5 日 受理

◇ 付 録 ◇

表 A.1 主観評価実験で使用した発話文章および対話行為.

対話行為類型	発話文章類型	発話文章
情報伝達	完全文.文末詞なし	けど、好物なためかうどんはよく作ります (自己開示.習慣) ただ辛いだけかなという気もしました (自己開示.評価 0) 掃除機だけ嫌いという感じです (自己開示.評価-) 行ったのは一回だけですが、そのサービスは利用しました (自己開示.経験) 今度試してみます (自己開示.予定)
	完全文.ね	飲み会はやりませんが、バイトではやったことありませんね (自己開示.経験) 冬なんかはだいたい温かい日本茶ですね (自己開示.習慣) アクティブですね (承認) 私は海外だったら、取り合えず英語圏で食べ物おいしいところに行きたいですね (自己開示.欲求) それは実用的ですね (自己開示.評価+)
	言いさし文.が	全然無理でしたが (自己開示.評価-) 米沢牛でもいろいろ種類があったんですが (情報提供) 対戦が面白いんですが (自己開示.評価+) 特定のアーティストが好きというよりは、その時々で気に入った曲を聴くという感じなんですが (非共感・非同意) キャンプ場のアルバイトだったのですが (自己開示.事実)
質問	完全文.か	紅葉とか見に行く予定はありますか (質問.予定) お笑いの話はいかがでしょうか (提案) 他に気になる水族館や動物園はありますか (質問.欲求) 魚好きですか (質問.評価) どんな絵なんですか (質問.情報提供要求)
	完全文.ね	では東京に詳しいのですね (質問.事実) じゃあトレーニングも本格的にやって、体をつくらないといけないですね (質問.予定) モンブランとか買って食べたいですね (質問.欲求) なんで導入しないんでしょうね (質問.情報提供要求) そのうえおさんがいらしゃると、生活がかなり制限されますね (質問.習慣)
	完全文.文末詞なし	どこか行きたいところあります (質問.欲求) ご飯とかどうする予定です (質問.予定) 今年の夏は夏らしい事しました (質問.経験) カフェはどこによく行きます (質問.習慣) 他に以前読んだおもしろかった本はありましたか (質問.評価)
自問	完全文.文末詞なし	5 兆なくなっちゃいそうです (質問.自問) ルンバでしたっけ (質問.自問) だれだろう (質問.自問) どうでしょう (質問.自問) なんでしょう (質問.自問)
	完全文.かな	今日ほこれから降るかな (質問.自問) それかバスターかな (質問.自問) 次の日だったかな (質問.自問) あとは何があったかな (質問.自問) 近くのスポーツクラブは、確か 22 時とか 23 時くらいまでだったかな (質問.自問)
	完全文.かね	要は相性ですかね (質問.自問) そうなんですかね (質問.自問) 子供ってなんであんなずーっと集中して観れるんですかね (質問.自問) このままだとすぐに冬ですかね (質問.自問) 設置アンテナの数が電話の数に対応出来るのでしょうかね (質問.自問)
繰り返し/言い換え	完全文.ね	スバルですね (言い換え) ブラッディマリーですね (繰り返し) ピタラですね (繰り返し) 呑み屋をやってらしたんですね (繰り返し) 海は見ただけで涼しくなりそうですね (言い換え)
	完全文.か	完全におもちゃ扱いではないですか (言い換え) キングコングでしたか (繰り返し) 公園の木ですか (繰り返し) そんなにするんですか (言い換え) ロックですか (繰り返し)
	述語省略文	妊婦さんはお金かかりそう (言い換え) 九州 (繰り返し) 見た目すごい赤そう (言い換え) フカミは日本と韓国だけ (繰り返し) そんな優しいところ (言い換え)
相槌	独立詞文	ああなるほど (相槌) あー (相槌) ねえ (相槌) うーん (相槌) おっ (相槌)
	完全文.か	本当ですか (相槌) そうだったのですか (相槌) あそいでしたか (相槌) あっそうなんですか (相槌) そんなものがあるんですか (相槌)
	完全文.ね	どうなんでしょうかね (相槌) なるほどそうなんですね (相槌) そうですね (相槌) そうなのですね (相槌) いろいろあるんですね (相槌)
フィラー	独立詞文	あー (フィラー) いやー (フィラー) うーん (フィラー) うーむ (フィラー) いやあ (フィラー)
	完全文.ね	うーん悩めますね (フィラー) うーんどうでしょうね (フィラー) どうなんでしょうね (フィラー) なんかですね (フィラー) んーそうですね (フィラー)
	述語省略文	ランチは (フィラー) 普通に知名度があるのは (フィラー) なんか (フィラー) 何の (フィラー) その (フィラー)
感嘆	独立詞文	あら (感嘆) お (感嘆) ありや (感嘆) あはは (感嘆) まあ (感嘆)
	述語省略文	びっくり (感嘆) それそれ (感嘆) 香川が (感嘆) そんなに (感嘆) 無理ゲー (感嘆)
	完全文.か	そうなんですか (感嘆) 買ったんですか (感嘆) そうですね (感嘆) 駅二つですか (感嘆) 黒いんですか (感嘆)
定型文	挨拶定型文	おはようございます (挨拶) ではここで (挨拶) まあ後は (挨拶) こちらこそ (挨拶) 今日も宜しくお願いします (挨拶)
	感謝定型文	それではありがとうございました (感謝) いろいろ情報ありがとうございました (感謝) はい 4 日間ありがとうございました (感謝) ありがとうございましたまた後ほど (感謝) 本日はありがとうございました (感謝)
	謝罪定型文	何か話題変わってすみません (謝罪) すみませんそろそろ (謝罪) 名前間違えて失礼しました (謝罪) ちょっと語弊があったようで申し訳ないです (謝罪) すいません、わかりにくくて (謝罪)

——— 著 者 紹 介 ———

北条 伸克

2012 年東京大学工学部物理工学科卒業。2014 年同大学大学院情報理工学系研究科システム情報学専攻修士課程修了。同年、NTT 入社。以来、音声合成の研究開発に従事。日本音響学会、電子情報通信学会各会員。

井島 勇祐

2007 年八代工業高等専門学校専攻科修了。2009 年東京工業大学大学院総合工学研究科物理情報システム専攻修士課程修了。2015 年同研究科博士課程修了。博士（工学）。2009 年に NTT 入社。以来、音声合成の研究開発に従事。電子情報通信学会、日本音響学会各会員。

杉山 弘晃(正会員)

2007 年東京大学工学部機械情報工学科卒業。2009 年同大学院知能機械情報学専攻修士課程修了。2015 年奈良先端科学技術大学院大学情報科学研究科博士課程修了。博士（工学）。2009 年に NTT 入社。人と自然な対話を行う雑談対話システムの研究に従事。IEEE、電子情報通信学会、情報処理学会各会員。

宮崎 昇(正会員)

1995 年東京工業大学工学部情報工学科卒業。1997 年同大学院総合理工学研究科知能科学専攻修士課程修了。同年、NTT 入社。電子情報通信学会、日本音響学会各会員。

川西 隆仁

1996 年京都大学工学部情報工学科卒業。1998 年奈良先端科学技術大学院大学情報科学研究科修士課程修了。同年、NTT 入社。以来、メディア認識・探索・配信の研究開発に従事。博士（工学）。電子情報通信学会シニア会員、情報処理学会、日本応用数理学会各会員。2003 年情報処理学会 FIT ヤングリサーチ賞、2012 年日本応用数理学会ベストオーサー賞、2017 年電子情報通信学会 情報・システムソサイエティ活動功労賞を受賞。

柏野 邦夫(正会員)

1990 年東京大学工学部卒業。1995 年同大学院工学系研究科博士課程修了。博士（工学）。同年、NTT に入社。現在、同社上席特別研究員。東京大学大学院情報理工学系研究科客員教授を兼務。メディア情報の探索・認識・生成に興味をもつ。電子情報通信学会、情報処理学会、日本音響学会、IEEE、ACM の各会員。