音声対話システムにおける音声合成のための 対話行為情報を利用した文末音調ラベル推定

Estimating Sentence Final Tone Labels using Dialogue-Act Information for Text-to-Speech Synthesis within a Spoken Dialogue System

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

keywords: sentence final tone labels, dialogue-act, speech synthesis, spoken dialogue systems

Summary

This paper proposes a novel sentence final tone labels estimation method using dialogue-act (DA) information for text-to-speech synthesis within a spoken dialogue system. Estimating appropriate sentence final tone labels is considered essential for communicating the exact system's intentions to users by an utterance. In this paper, we propose to utilize DA features as well as the conventional features, morphological information of the utterance text, to estimate the sentence final tone labels. For this study, we use the speech database with DA tags which we constructed in our previous study. We added sentence final tone labels to this database so that each utterance has all information of utterance text, DA and sentence final tone label. Based on this database, we build the proposed sentence final tone estimation model. We evaluated the proposed method by comparing its performance with the conventional method. The evaluation results show that the proposed method overwhelms the conventional method in accuracy. We also analyze the estimation results to investigate the efficacy and difficulty by the proposed method.

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文末音調を適切に制御するため、対話行為 (dialogue act; DA) を考慮した文末音調ラベルの推定手法を提案する。語用論分野における Grice の理論 [Grice 91] や発話行為論 [Austin 75, Searle 80] では、人と人のコミュニケーションの成立のためには、話し手の発話意図が正確に聞き手へ伝達されることが不可欠であるとされる。人と機械のコミュニケーションにおいても同様に、システムの発話意図が正確にユーザへ伝達されるように、システムの発話を生成することが重要である。ここで、伝達される発話意図は、発話文章や文脈 [Wilson 12] のみでなく、韻律にも依存する [Hellbernd 16, 前川 02]。このため、ユーザの発話意図の理解のために韻律を使用する [Cutler 85, Shriberg 98, Ward 00] だけでなく、音声合成技術により、発話意図の持つ韻律の特徴を再現し、システムの

発話意図をユーザへ正確に伝達する必要があると考えられる。

本研究では、発話意図の持つ韻律の特徴として、特に、文末音調に着目する。文末音調とは、発話の文末付近の F0 軌跡である。文末音調は、しばしば、“無音調”、“上昇下降調”などの類型 (文末音調ラベル) により記述される [郡 15, Maekawa 02, Venditti 05]。日本語の対話音声では、同一の発話文章であっても、異なる文末音調ラベルにより、異なる発話意図が伝達されることが知られている [郡 15, 森山 89, 杉藤 01, 轟木 08, 上村 89]。このため、音声対話システムは、発話意図を考慮し、適切な文末音調ラベルの音声を合成する必要がある。従来研究により、文末音調ラベルが指定されれば、対応する文末音調の音声合成が可能である [Black 96, 郡山 12, 栗原 18]。一方で、音声対話システムでは、小規模なルールに基づくシステムを除けば、システム発話の候補数は膨大である。このため、全ての発話に事前に文末音調ラベルを指

定することは困難である。以上から、正確に発話意図を伝達するためには、発話意図を考慮し自動的に文末音調ラベルを推定する技術が有用であると考えられる。

発話の文末音調ラベルを推定する先行研究には、発話文章の形態素等の特徴量に基づく手法 [中嶋 14, Ross 96] がある。Ross らは、英語のラジオニュースの読み上げ音声について、文末音調ラベルを含む韻律記号を推定する手法を検討した [Ross 96]。中嶋らは、日本語の宣伝口調の音声について、文末を含む、各アクセント句末の音調 (句末音調) のラベルを推定する手法を検討した [中嶋 14]。しかし、いずれも対話音声ではなく独話音声を対象とした研究であるため、データベース中に現れる対話行為は主に情報伝達などであり、種類は限定的であると考えられる。また、推定に使用される情報は発話文章のみであり、発話意図の使用は検討されていない。以上から、本研究では、発話文章に加え、発話意図を考慮した文末音調ラベルの推定を検討する。

発話意図を考慮した文末音調ラベルの推定を行うためには、システムの発話意図に相当する情報を取得する必要がある。本研究では、システムの発話意図に相当する情報として、雑談対話システムで活用される対話行為 (dialogue act; DA) [目黒 12] の使用を検討する。対話行為とは、発話意図の抽象的な表現であり [Stolcke 00]、雑談対話システム [Higashinaka 14] 等の多くのシステムの対話処理部で使用される。対話処理部では、発話生成の処理過程で、まず生成される発話の対話行為が決定され、続いて、対話行為の表現のために適切な発話文章が生成される [Sugiyama 13]。このため、対話行為は、伝達されるべき発話意図の抽象表現とみなすことができ、各発話について発話文章とともに取得可能である。文末音調ラベルの推定のために、発話文章だけでなく対話行為を活用することで、推定精度が向上すると期待される。

もし、発話文章および対話行為に対し、文末音調ラベルが決定的な関係であれば、例えばルール設計などにより、文末音調ラベルは比較的容易に推定可能であると考えられる。しかし、一般に、対話行為は対話制御のために設計されるため、文末音調ラベルを決定するために十分な情報を含むとは限らない。例えば、同一の対話行為タグに対応する発話であっても、タグで表現されるよりも詳細な発話意図の違いにより、異なる文末音調ラベルで発話される可能性がある。したがって、様々な文脈の下で実際に発話された音声を使用し、提案法の効果を評価する必要があると考えられる。このため、本研究では、我々がこれまでに構築した、対話行為情報付き音声データベース [北条 16, 北条 20] を使用する。この音声データベースは、テキストチャットコーパスに含まれる発話について、文脈と対話行為を考慮し発話された音声 が収録されている。本研究では、この音声データベースに文末音調ラベルを付与し、検討に使用する。

本研究では、提案法の効果の検証に加えて、文末音調

ラベル推定の観点から対話行為タグセットの設計に関する考察を行う。上述の通り、対話行為タグは、一般に対話制御のために設計されるため、文末音調ラベルを決定するために十分な情報を含むとは限らない。このため、文末音調ラベルの推定精度を改善するためには、対話行為タグの数を増やすことが考えられる。しかし、不要な対話行為タグが追加された場合、対話処理部で使用される変数の次元が増加するため、複雑化する。このため、十分な品質の発話文章を生成するために、より多くの学習データが必要となる。もしくは、学習データ量が限られている場合には、過学習が生じ、システム発話の品質が劣化する弊害がある。したがって、提案法によって文末音調ラベルが正しく推定されない発話を調査し、推定結果の改善のために必要な対話行為タグを明らかにすることは有用である。これにより、将来的に、対話制御だけでなく、音声合成および文末音調の制御の観点からも有用な情報を持つ対話行為タグセットを設計し、音声対話システムで使用することが可能となる。

これまでに、我々は上述の音声データベースを使用し、音声合成用の DNN 音響モデルに対話行為情報を入力し、韻律を生成する手法を検討した [北条 16, 北条 20]。この検討では、文末音調ラベルを使用せず、合成音声の F0 やスペクトル等の特徴量の推定のために対話行為を使用した。文末音調に関する評価として、収録音声の文末 F0 との相関による客観評価を行った。客観評価の結果、対話行為を考慮することで、文末 F0 の相関が改善される傾向が示された。しかし、文末 F0 の相関は、文末の音素や引き伸ばしの有無など、様々な要因によって変動する。このため、「発話意図の伝達のため適切な文末音調が合成されているか」について評価を行い、課題を分析することは困難であった。一方、発話意図の伝達精度を評価するためには、主観評価実験を実施することも考えられる。しかし、発話意図の伝達精度に関する主観評価実験手法は研究途上であり [Hojo 19]、十分に信頼性の高い評価結果が得られるとは限らない。また、聴取実験により、様々な文脈、発話文章、対話行為に対応する多数の音声を評価することは、実験コストの観点からも難しい。本研究では、評価対象が文末音調ラベルの正誤に限られるものの、多数の発話について、伝達される発話意図に関する客観評価を行うことが可能である。これにより、様々な条件下で発話された音声を使用し、提案法の効果や課題を分析することが可能となる。5 章では、提案法の効果や課題について考察する。

以上で述べた通り、本研究では、対話行為を考慮した文末音調ラベルの推定について検討する。また、推定結果の正誤に基づき、提案法の効果や今後の課題を分析する。以下、2 章では、本研究で使用した発話意図および文末音調アノテーション付き音声データベースについて述べる。3 章では、発話意図を使用した文末音調推定手法について述べる。4 章では、評価実験について述べ、5

[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きりです。
	A	Bさんは何かされますか？
	B	そうですね。
	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。
3401	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	バスケが出来るというのは尊敬します。
3403	A	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。

章で実験結果の考察を行う。

2. 音声データベース

1 章で述べた通り、本研究では、我々が過去に構築した対話行為付き音声データベースに対し、文末音調ラベルを付与し、検討に使用する。2.1 節では、まず、対話行為情報付き音声データベースの構築について述べる。続いて、2.2 節では、文末音調ラベルのアノテーションについて述べる。

2.1 対話行為情報付き音声データベース [北条 16, 北条 20]

1 章で述べた通り、本研究では、対話処理部から獲得される対話行為を伝達されるべき発話意図の抽象表現とみなし、文末音調ラベル推定に使用する。本研究では、対話行為を表現する音声を収録し、データベースとして使用する。

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 [Higashinaka 14] 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。各発話について、人手により対話行為が付与されている。対話行為として、先行研究 [目黒 12] のタグセットに“質問_自問”を追加した、合

表 1: 音声データベースの統計。

対話行為	L%	LH%	H%	HL%	合計
自己開示_予定	88	1	17	0	106
自己開示_事実	105	3	20	1	129
自己開示_欲求	83	1	26	0	110
自己開示_経験	78	1	9	3	91
自己開示_習慣	78	4	21	0	103
自己開示_評価+	113	5	72	3	193
自己開示_評価-	106	2	63	14	185
自己開示_評価 0	67	2	44	6	119
情報提供	80	1	36	7	124
共感・同意	66	6	63	4	139
非共感・非同意	135	8	21	5	169
承認	73	0	94	8	175
繰り返し	132	9	16	0	157
言い換え	32	0	89	4	125
提案	67	19	8	0	94
確認	18	44	25	4	91
質問_予定	14	75	0	1	90
質問_事実	12	110	6	1	129
質問_欲求	10	58	5	1	74
質問_経験	9	91	1	0	101
質問_習慣	3	101	1	2	107
質問_評価	7	74	19	3	103
質問_情報提供要求	16	76	6	15	113
質問_自問	51	13	12	31	107
フィラー	22	0	2	20	44
相槌	26	2	14	2	44
感嘆	20	9	10	3	42
挨拶	69	0	4	0	73
感謝	26	0	0	0	26
謝罪	70	0	0	0	70
合計	1676	715	704	138	3233

計 33 種類のタグセットが使用されている。タグの付与作業では、92 発話について 3 名の作業者が対話行為を付与している。これらの付与結果について、高い一致係数 ($\kappa = 0.718$) が確認されている。残りの発話については、3 名の作業者が分担してタグを付与している。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]。

音声収録では、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、これらの重複を除外した。また、「はい」、「うーん」等の独立詞のみからなる発話については、独自に音調を捉えることが適切であると考えられる [国立 60] ため、文末音調を定義せず、データベースから除外した。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。以上から、本研究では全 3233 発話を研究対象とした。本研究の音声データベースの統計を表 1 に示した。

2.2 文末音調ラベル

先行研究では、日本語における文末音調の分類には、カテゴリ数や定義の異なるいくつかの方法が存在する。音声合成システムで使用する文末音調ラベルの規格 [赤羽 11] や句末音調ラベル推定 [中嶋 14] では、2 種類のラベル（「通常」「疑問」等）が使用されている。一方で、対話音声の文末音調の機能はより多様であると考えられており、しばしば、より多数のカテゴリからなるラベルが使用される。例えば、JToBI [Venditti 05] では 4 種類、X-JToBI [Maekawa 02] では 5 種類のラベルが使用される。本研究では、JToBI [Venditti 05] で使用される 4 種類の文末音調

1. 無音調 (L%)
2. 疑問型上昇調 (L%LH%)
3. 強調型上昇調 (L%H%)
4. 上昇下降調 (L%HL%)

を使用する。以下では簡略化し、疑問型上昇調は、“LH%”，強調型上昇調は“H%”，上昇下降調は“HL%”でそれぞれ表す。X-JToBI では、上記に加え、上昇下降上昇調 (L%HLH%) 等も使用されるが、本研究で使用するデータベースには含まれなかったため、本研究では研究対象としない。また、普通の上昇調/反問の上昇調/つり上げ調の区別や、強めの上昇調/浮き上がり調の区別 [川上 63] も提案されているが、同様に本研究の研究対象には含めない。

本研究では、対話行為情報付き音声データベースの各発話について、上記の 4 種類の文末音調ラベルを付与し、使用する。2 名のアノテータがラベルを付与した。アノテータ間のラベルの一致率 (Cohen の κ [Cohen 60]) は 0.84 と高い一致率であった。X-JToBI のアノテーションを付与した先行研究 [菊池 03] と比較すると、本研究では高い一致率が得られた。自然発話音声を対象とした先行研究と比較し、演技音声を対象とした本研究では、発話の際に文末音調が明確に区別されたため、音調ラベルの判断が比較的容易であったと考えられる。アノテータ間で異なるラベルが付与された発話については、アノテータ間で協議し、正解ラベルを決定した。また、各発話について、形態素および品詞の情報を人手で付与した。音声データベース中の対話行為別の文末音調の統計を表 1 に示した。文末音調ラベルにより発話数に偏りがあるものの、各ラベルについて 130 発話以上がデータベースに含まれていることがわかる。

3. 文末音調推定手法

3.1 ベースライン手法

本研究では、ベースライン手法として、発話文章から得られる形態素や品詞等の情報に基づき文末音調を推定する従来法を使用する。従来法には、英語のラジオニュース読み上げ音声に関する検討 [Ross 96] と日本語の宣伝

図 2: ベースライン手法および提案法の概略図。文末コンテキストの形態素数 $N = 2$ の場合を示した。

表 2: 音声データベースにおける文末音調のエントロピーと、各特徴量と文末音調の相互情報量。文末コンテキスト (word-pos) の形態素数 $N = 1$ の場合を示した。

$I(Y; \text{word-pos})$	0.78
$I(Y; \text{DA})$	0.43
$I(Y; \text{word-pos} + \text{DA})$	0.99

文章の読み上げに関する検討 [中嶋 14] がある。日本語の宣伝文章の読み上げに関する検討では、句末の複数個の形態素の出現形および品詞 (文末コンテキスト) が特徴量として使用され、Ross らの手法よりも高い推定精度が得られることが確認されている。本研究では、同じ日本語音声に関する検討を行った手法 [中嶋 14] をベースライン手法として使用する。

図 2 (a) にベースライン手法の概略図を示した。まず、入力された発話文章を形態素解析し、形態素の出現形および品詞の情報を得る。発話文章末尾の N 形態素について、その出現形および品詞 (文末コンテキスト) を抽出する。文末コンテキストを特徴量とし、文末音調ラベル推定器に inputs する。文末音調ラベル推定器として、classification and regression tree (CART) [Breiman 84] や SVM [Vapnik 98] を使用する。推定された文末音調ラベルは、発話文章とともに音声合成器に inputs され、文末音調が制御される。

3.2 提案法

対話音声の文末音調ラベルは、同じ発話文章であっても、発話意図によって変化する場合があります。このような文末音調ラベルの変化は、発話文章の形態素や品詞等の情報のみから推定することは困難であると考えられる。したがって、システムの発話意図に相当する情報が取得可能であれば、文末音調ラベルの推定精度が向上すると期待される。2 章で述べた通り、対話行為は多くの対話システムの対話処理部から取得可能な情報であり、システムの発話意図の抽象表現とみなす

ことが可能である．以上から，提案法は，従来の文末コンテキストに加え，対話行為を特微量として使用する．図 2 (b) に提案法の概要図を示した．対話処理部から，対話行為を取得し，文末音調ラベルの推定に使用する点がベースライン手法と異なる．

ここで，文末音調ラベルの推定のために対話行為を特微量として使用する効果について，相互情報量を用いて予備的に調査する．相互情報量 $I(Y, X)$ は，ある確率変数 X, Y について次式で求められる．

$$I(Y, X) = \sum_{X, Y} P(X, Y) \log \frac{P(X, Y)}{P(X)P(Y)}$$

相互情報量は，変数 X が与えられた際に，文末音調ラベル Y に対して得られる情報量を表す指標であり， Y を推定する上で，変数 X がどの程度重要な特微量であるかを示すものと解釈できる．表 2 に各特微量の相互情報量を示した．ただし，文末音調ラベルを “ Y ”，文末コンテキストを “word_pos”，対話行為を “DA”，文末コンテキストおよび対話行為の双方を使用する提案法に対応する場合を “word_pos + DA” でそれぞれ表した．まず， $I(Y; \text{word_pos}) > I(Y; \text{DA})$ であることから，文末コンテキストと対話行為を比較すると，文末コンテキストがより重要な特微量であると考えられる．次に， $I(Y; \text{word_pos}) < I(Y; \text{word_pos} + \text{DA})$ であることから，文末コンテキストと合わせて対話行為を使用することで，文末音調ラベルの推定精度が上昇すると考えられる．以上から，提案法により，文末コンテキストと対話行為の両方を特微量として使用することで，従来法に比べ文末音調の推定精度が向上すると期待される．

4. 実 験

4.1 評 価 手 法

本研究では，従来研究 [中嶋 14, Ross 96] と同様に，推定された文末音調ラベルの正誤による客観評価実験を行う．本節では，まず，客観評価を行う理由について述べる．音声対話システムにおける音声合成の従来研究では，しばしば，主観評価が行われる．例えば，対話行為を考慮する音声合成の検討では，合成音声の自然性が評価され，その改善が示されている [北条 20, Tsiakoulis 14, Syrdal 08]．また，システムと対話形式の音声に対する主観評価では，読み上げ口調に比べ，対話口調を使用した場合に自然性が高いことが示されている [Sugiura 15]．主観評価では，合成音声や対話システムの印象を直接評価可能である．本研究では，システムの発話意図が正確に伝達されたかどうかを直接評価するため，合成音声のサンプルを作成し，実験参加者に「合成音声から，参照意図（例：フィラー）が感じられるか？」と質問する評価方法 [Hellbernd 16] を適用することが考えられる．

一方で，評価に使用する参照意図によっては，実験参加

表 3: 手法名と実験条件．

手法名	識別器	特微量
SVM-BASELINE	SVM	word_pos
SVM-DA	SVM	word_pos + DA
CART-BASELINE	CART	word_pos
CART-DA	CART	word_pos + DA

者が参照意図の定義を正確に理解することが困難であり，主観評価結果の信頼性が損なわれる場合があるという課題がある [Hojo 19]．この課題を解決するためには，主観評価ではなく，対話者の反応による評価 [Dall 17, Misu 11] を行うことも考えられる．しかし，対話者の反応による評価方法が提案されているのは，ユーザの相槌の喚起 [Misu 11] やフィラーの音声合成 [Dall 17] など，一部の発話機能や発話意図についてである．このため，本研究が対象とする全ての対話行為タグに関する評価手法は確立されていない．また，発話意図の伝達とは異なる，自然性や親しみやすさ等の評価尺度を使用するとも考えられる [Sugiura 15]．しかし，対話シナリオによっては，実際にはシステムの発話意図が対話者に誤解されているものの，実験参加者がそれに気づかず，好印象な対話システムであると評価される場合がある．このため，本研究の目的に対応した評価結果が得られるとは限らない．以上から，主観評価では，本研究について十分な信頼性の評価結果が得られないという懸念がある．

本研究では，推定された文末音調ラベルの正誤による客観評価を行う．先行研究により，文末音調ラベルと発話意図の間には，直接的な関係が示されていることから，客観評価のみを行った場合も，妥当性は大きく損なわれないと考えられる．また，客観評価では，主観評価に比べ評価コストが低いため，交差検定により音声コーパスの全ての発話をテストセットとして使用することができる．これにより，例えば，5.2 節における提案法の課題分析では，様々な文脈における発話を使用し考察を行うことが可能となる．

4.2 実 験 条 件

提案法の効果を検証するため，文末音調の推定実験を行った．分類器として，ベースライン手法と同様に，SVM と CART を使用した．特微量として，文末コンテキストを使用する従来法 (“word_pos”) と，文末コンテキストおよび対話行為を使用する提案法 (“word_pos + DA”) を比較した．以上から，表 3 に記載された 4 条件 (SVM-BASELINE, SVM-DA, CART-BASELINE, CART-DA) について，文末音調ラベルの推定精度を比較した．

SVM-BASELINE および SVM-DA では，RBF カーネルを使用した．コスト C は 1.0 に，RBF カーネルのハイパーパラメータ γ は 0.2 に設定した．CART-BASELINE および CART-DA では，分類基準として，Gini 不純度を使用した．過学習を回避するため，決定木の深さの最大値

図 3: 文末コンテキストとして使用する形態素数 N に対する一致係数 κ 。エラーバーは 95%信頼区間を示す。

は 16 に設定した。文末コンテキストで使用する形態素および品詞情報は、人手で付与したものを使用した。全ての手法で、学習データ中の出現頻度が 4 回未満である形態素を特徴量化する際は、品詞情報のみを使用し、“UNK_名詞”，“UNK_終助詞”などのように変換した。また、すべての手法で、文末コンテキストの形態素数 N を 1 から 4 まで変化させ、手法間の比較を行った。

評価は音声データベースを 10 分割し、交差検定により行った。評価尺度として、Cohen の一致係数 κ [Cohen 60] を使用した。一致率ではなく、Cohen の一致係数 κ を使用することで、正解の文末音調ラベルの偏りを考慮し、偶然の一致の影響を排除することが可能である。一致係数 κ は 0.6 以上のとき、正解と予測結果の間には「実質的な一致がある」と解釈される尺度である [Landis 77]。

4.3 実験結果

実験結果を図 3 に示す。まず、各手法について、特徴量として使用する文末コンテキストの形態素数 N を変化した場合を比較する。各手法で共通する傾向として、 $N = 3$ のとき一致係数 κ が最大となることがわかる。このため、今回の実験条件では、文末コンテキストの形態素数として、 $N = 3$ を最適値とし、手法間の比較を行う。

続いて、対話行為を考慮する提案法の効果を検証する。SVM-DA の一致係数 κ は、SVM-BASELINE に比べ、約 0.12 改善した。同様に、CART-DA の一致係数 κ は、CART-BASELINE に比べ、約 0.11 改善した。 χ^2 検定の結果、SVM-DA と SVM-BASELINE、CART-DA と CART-BASELINE の間には、それぞれ、有意水準 5% で有意差が確認された。以上から、雑談対話システムで利用される対話行為タグセットを使用し、文末音調ラベルを推定する提案法の効果が示された。上記の提案法の効果を検証するために、さらに主観評価実験を行うことも考えられる。しかし、4.1 節で述べた通り、信頼性の高い

表 4: 従来法 (SVM-BASELINE) および提案法 (SVM-DA) による正解発話数の内訳。提案法の正解発話数と従来法の正解発話数の差が大きい順に示した。

文末コンテキスト (文末音調ラベル)	SVM- BASELINE	SVM- DA	正解発話数 の増分
終助詞:か (L%)	82	196	114
終助詞なし (LH%)	10	110	100
終助詞:ね (HL%)	2	18	16
終助詞:かね (HL%)	7	21	14
その他	2710	2707	-3
合計	2811	3052	241

主観評価の実施には課題があるため、今後の課題とする。

最後に、分類器間の比較を行う。各実験条件において、SVM に基づく手法は、CART に基づく手法に比べ、同程度かより高い一致係数を示すことがわかる。本研究の実験条件では、CART よりも SVM に基づく手法の分類精度が高いことが示された。このため、5 章では、SVM に基づく分類結果に基づく考察を行う。

5. 考察

本章では、雑談対話システムで使用される対話行為タグを用いて文末音調ラベルを推定する提案法により、対話行為を考慮する効果と課題を分析する。具体的には、提案法の効果として、どのような文末コンテキスト、文末音調ラベルおよび対話行為において、対話行為を考慮することで正解の発話数が増加したのかを分析する (5.1 節)。また、提案法の課題として、提案法が不正解であった発話を同様に分析する (5.2 節)。

5.1 提案法による効果の分析

図 4 に、従来法 (SVM-BASELINE) および提案法 (SVM-DA) について、文末コンテキストおよび文末音調ラベル別の正解発話数 (上段)、不正解発話数 (下段) を示した。ただし、終助詞のない発話については、文末の形態素や品詞の種類によらず、“終助詞なし”とした。また、発話数が少量である終助詞については、簡単のため、“終助詞:その他”に統一した。まず、図 4 において、従来法に比べ提案法の正解発話数が特に大きく増加した文末コンテキストおよび文末音調ラベルを明らかにする。従来法と提案法の正解発話数の内訳を表 4 に示した。提案法により正解発話数が増加したのは、主に、

1. 終助詞:か (L%)
2. 終助詞なし (LH%)
3. 終助詞:ね (HL%)、終助詞:かね (HL%)

の発話であった。以下、本節では、これらの三つの場合について、推定結果の改善に寄与した対話行為や発話の例を確認する。

図 4: 各手法による推定結果について、正解の発話数（上段）、不正解の発話数（下段）のプロット。

表 5: “終助詞:か (L%)” の正解発話数の内訳。提案法の正解発話数と従来法の正解発話数の差が大きい順に示した。

対話行為	SVM-BASELINE	SVM-DA	正解発話数の増分
繰り返し	0	68	68
相槌	0	16	16
言い換え	0	14	14
感嘆	0	10	10
その他	82	88	6
終助詞:か (L%) の合計	82	196	114

§1 “終助詞:か (L%)” の推定における対話行為の効果
 “終助詞:か (L%)” の推定に寄与した対話行為を明らかにするため、正解発話数の内訳を表 5 に示した。考慮する効果が大きい対話行為は、主に、

繰り返し、相槌、言い換え、感嘆 (1)

であった。例えば、

- 「ビーズですか (L%)」[繰り返し]
- 「あ、そうだったのですか (L%)」[相槌]
- 「おお、10 周年ですか (L%)」[言い換え]
- 「そんなに持つんですか (L%)」[感嘆]

等の発話において、対話行為を考慮することにより、文末音調ラベルの推定結果が改善することが確認された。

音声データベース全体では、“終助詞:か” の発話は、多くの場合 “LH%” で発話された。その場合の対話行為は (1) とは異なり、例えば、

- 「手に持っているのは、ビーズですか (LH%)」[確認]

のように、“質問”、“確認” など、聞き手に何かを問いかける対話行為であった。対話行為を考慮する提案法により、上記の 2 種類の文末音調を区別することが可能となり、文末音調ラベルの推定結果が改善されたと考えられる。

表 6: “終助詞なし (LH%)” の正解発話数の内訳。従来法と提案法の正解発話数の差が大きい順に示した。

対話行為	SVM-BASELINE	SVM-DA	正解発話数の増分
質問_事実	1	25	24
質問_経験	0	18	18
質問_情報提供要求	1	18	17
確認	1	15	14
質問_予定	4	13	9
質問_習慣	2	9	7
質問_評価	1	8	7
質問_欲求	0	4	4
その他	0	0	0
終助詞なし (LH%) の合計	10	110	100

“終助詞:か (L%)” の発話は、「疑問文を無音調で発話し、話し手の納得や了解などを伝達する」という特徴から、日本語モダリティ研究における“自問納得”[仁田 91] の用法に相当すると考えられる。提案法では、(1) の対話行為を考慮することで、“自問納得”の表現のために適切な無音調の文末音調ラベルが推定可能となったと考えられる。

§2 “終助詞なし (LH%)” の推定における対話行為の効果

続いて、“終助詞なし (LH%)” の推定に寄与した対話行為を明らかにするため、対話行為に関する内訳を表 6 に示した。考慮する効果が大きい対話行為は、主に、

質問（質問_自問を除く）、確認 (2)

であった。例えば、

- 「使ったことあります (LH%)」[質問_経験]
- 「どっちも windows (LH%)」[質問_事実]
- 「と云いますと (LH%)」[確認]

等の発話において、対話行為を考慮することで推定結果が改善することが確認された。

表 7: “終助詞:ね (HL%)” および “終助詞:かね (HL%)” の正解発話数の内訳. 従来法と提案法の正解発話数の差が大きい順に示した.

対話行為	SVM-BASELINE	SVM-DA	正解発話数の増分
フィラー	2	19	17
質問_自問	2	12	10
その他	5	8	3
終助詞:ね/かね (HL%) の合計	9	39	30

音声データベース全体では, “終助詞なし” の発話文章は, 多くの場合, “L%” で発話された. その場合の対話行為は (2) とは異なり, 例えば,

- 「私は, 使ったことがあります (L%)」 [自己開示_経験]

のように, “自己開示” などの対話行為で使用された. 対話行為を考慮する提案法により, 上記の二つの用法が区別され, 文末音調ラベルの推定結果が改善されたと考えられる.

これらの発話は, 疑問文で使用される終助詞 (“か”, “の” など) を伴わず, 上昇イントネーションにより疑問の意図を表現することから, 日本語モダリティ研究における “上昇イントネーションによる疑問表現” [仁田 91] に相当すると考えられる. 以上から, 提案法では, (2) の対話行為を考慮することで, “上昇イントネーションによる疑問表現” のために適切な疑問型上昇調の文末音調ラベルが推定可能となったと考えられる.

§3 “終助詞:ね (HL%)” および “終助詞:かね (HL%)” の推定における対話行為の効果

“終助詞:ね (HL%)” および “終助詞:かね (HL%)” の推定に寄与した対話行為を明らかにするため, 対話行為に関する正解発話数の内訳を表 7 に示した. 考慮する効果が大きい対話行為は, 主に,

フィラー, 質問_自問 (3)

であった. 例えば,

- 「あとはどう調理するかですね (HL%)」 [質問_自問]
- 「なんですかね (HL%)」 [フィラー]

等の発話において, 対話行為を考慮する効果が確認された.

一方, 音声データベース全体では, “終助詞:ね” や “終助詞:かね” の発話文章は, 多くの場合 “H%” で発話された. その場合の対話行為は (3) とは異なり, 例えば,

- 「あとはどう調理するかですね (H%)」 [確認]

のように, “自己開示”, “確認” などの対話行為で使用された. 対話行為を考慮する提案法により, (3) の対話行為において, 上昇下降調で発話される傾向が再現されたと考えられる.

本節の考察から, 雑談対話システムで使用される対話

表 8: 文末コンテキスト・文末音調ラベル別の, 提案法による不正解発話数 (分子) および全発話数 (分母). 提案法 (SVM-DA) による不正解発話数が大きい順に示した.

文末 コンテキスト	文末音調 ラベル	SVM-BASELINE	SVM-DA
終助詞:ね	HL%	72/74	56/74
終助詞なし	LH%	122/132	22/132
その他		228/3027	103/3027
合計		422/3233	181/3233

表 9: “終助詞:ね” の不正解発話数 (分子) および全発話数 (分母) の内訳. 対話行為は提案法による不正解発話数が多い順に示した. 無音調 (L%) の発話は音声データベースに含まなかった.

対話行為	LH%	H%	HL%	合計
自己開示_評価-承認	0/0	0/45	11/11	11/56
自己開示_評価 0	0/0	0/34	6/6	6/40
情報提供	0/0	0/31	6/6	6/37
その他	13/13	4/335	26/44	43/392
合計	13/13	4/532	56/74	73/619

行為タグセットを使用し, 雑談対話中の発話の文末音調ラベルを推定する提案法では,

1. 自問納得
2. 上昇イントネーションによる疑問表現
3. フィラー, 自問

の発話意図の伝達のために適切な文末音調ラベルの推定が可能となったと考えられる.

5.2 課題分析

続いて, 提案法による推定結果が不正解であった発話に着目し, 提案法の課題の分析を行う. 図 4 において提案法による不正解の発話数が特に多い文末コンテキストおよび文末音調ラベルを明らかにするため, 提案法による不正解発話数の内訳を表 8 に示した. 雑談対話システムで使用される対話行為タグセットを使用し, 雑談対話中の発話の文末音調ラベルを推定する提案法では, 不正解となる発話は, 主に,

1. 終助詞:ね (HL%)
2. 終助詞なし (LH%)

であった. “その他” にも, 不正解の発話は 103 発話含まれるが, 一つの文末コンテキスト・文末音調ラベルに対応する不正解発話数は 13 発話以下であった. 以下, 本節では, 発話の事例数が多く, 一般性の高い知見を得ることが可能であると考えられることから, 上記の二つの場合についてのみ考察する.

§1 “終助詞:ね (HL%)” の推定に関する課題分析

“終助詞:ね (HL%)” の推定が困難であった対話行為を明らかにするため, “終助詞:ね” の不正解発話数の内

訳を表9に示した。5・1・3節で効果が確認された“質問・自問”と“フィラー”以外の多くの対話行為（“自己開示・評価-”，“承認”，“自己開示・評価0”，“情報提供”など）において，“HL%”が適切に推定されない事例が確認された。

これらの対話行為で文末音調ラベルの推定が困難であった理由を明らかにするため、表9では、データベース中の全発話数についての統計を合わせて示した。全発話数についての統計から、同じ文末コンテキスト（“終助詞:ね”）と対話行為に対し，“H%”と“HL%”の両方の発話が含まれることがわかる。このため、文末コンテキストと対話行為に基づき文末音調ラベルを推定する提案法では，“H%”と“HL%”の発話を区別することが困難であると考えられる。

“終助詞:ね（H%）”と“終助詞:ね（HL%）”の発話意図の違いは、先行研究において、「平調・上昇下降調を用いた場合、強調型上昇調を用いた場合と異なり、話者が命題内容に対して強い感情を抱いていることが表明される」[大島13]とされる。言い換えるならば，“H%”では発話文章で言及される情報の伝達が，“HL%”では発話文章で言及される情報に対する、話者の何らかの感情の伝達が意図されると考えられる。本研究の音声データベース中の発話例では、

話者A: 昨日は地震がありましたね

話者B: けっこう大きかったですね（H%）[自己開示・評価-]

における話者Bの発話では、情報（「昨日の地震は大きかった」という話者Bによる評価）の伝達が意図されたため，“H%”で発話されたと考えられる。一方、

話者A: なので夜中によくインスタントのものを食べてしまったり、それがかえってよくないというか

話者B: 太りますね（HL%）[自己開示・評価-]

では、情報（「話者Aが太りそうだ」という話者Bによる評価）よりも、むしろ話者Bの感情（「話者Aが太りそうだ」という状況に対する同情など）の伝達が意図されたため，“HL%”で発話されたと考えられる。上記の例のように、“終助詞:ね”の発話では、対話行為のタグは同一であっても、文末音調の違いにより、異なる発話意図が伝達されると考えられる。以上から、“終助詞:ね（HL%）”を適切に推定するためには、上記の発話意図を区別可能な対話行為タグセットを新たに設計し使用することが有用である。もしくは、文脈を考慮し、発話文章で言及される情報と、その情報に対する話者の感情の、どちらを伝達する発話意図がより適切であるかを推定した上で、文末音調ラベルを推定することが有用である。後者の手法のためには、提案法で使用した文末コンテキストや対話行為の情報のみではなく、文脈や発話文章の意味を考慮する手法が必要であると考えられる。

表10: “終助詞なし”の不正解数（分子）および全発話数（分母）の内訳。対話行為は提案法による不正解数が多い順に示した。上昇下降調（HL%）の発話は音声データベースに含まれなかった。

対話行為	L%	LH%	H%	合計
質問・自問	0/32	10/10	0/0	10/42
繰り返し	0/63	5/5	0/0	5/68
感嘆	0/10	4/4	1/1	5/15
その他	9/1300	3/113	1/11	13/1424
合計	9/1405	22/132	2/12	33/1549

§2 “終助詞なし（LH%）”の推定に関する課題分析

同様に、“終助詞なし（LH%）”の推定が困難であった対話行為を明らかにするため、“終助詞なし”の不正解発話数の内訳を表10に示した。5・1・2節で効果が確認された“質問”と“確認”以外の対話行為（“質問・自問”，“繰り返し”，“感嘆”など）において，“LH%”が適切に推定されない事例が確認された。

これらの対話行為で文末音調ラベルの推定が困難であった理由を明らかにするため、前節と同様に、不正解の発話数および全発話数の統計を表10に示した。“終助詞なし”の発話では、各対話行為に対し，“L%”と“LH%”の両方の発話が含まれ、かつ，“L%”の頻度がより高いことがわかる。したがって、前節と同様の理由により、提案法では、各発話に対し、頻度がより高い“L%”が推定されたため、“LH%”が適切に推定されなかったと考えられる。

“終助詞なし（L%）”と“終助詞なし（LH%）”の発話意図の違いは、話者が驚きや意外性等の感情の表現を意図するかどうかであると考えられる。本研究の音声データベース中の発話例では、

話者A: 今4年生ですが、いろいろと大変です

話者B: 4年生（L%）[繰り返し]

では、話者Bは「話者Aが今4年生であること」について、驚きや意外性等を感じなかったため、“L%”で発話されたと考えられる。一方で、

話者B: 何のサークルなんですか

話者A: スポーツ雪合戦のサークルです

話者B: スポーツ雪合戦（LH%）[繰り返し]

では、「スポーツ雪合戦」というサークルに対し、話者の驚きや意外性等を感じたため、“LH%”で発話されたと考えられる。上記の例から、対話行為のタグは同一であるものの、文末音調の違いにより、実際には異なる発話意図が伝達されると考えられる。以上から、“終助詞なし（LH%）”を適切に推定するためには、文脈を考慮し、発話文章で言及される情報に対する、話者の驚きや意外性等の感情を推定する必要があると考えられる。このためには、前節と同様に、新たな対話行為タグセットの設計か、文脈や発話文章の意味を考慮する文末音調ラベルの推定手法が有用であると考えられる。

本節の課題分析から，“終助詞:ね (HL%)”，“終助詞なし (LH%)”の推定精度の向上のためには，共通して，新たな対話行為タグセットの設計か，文脈や発話文章の意味を考慮可能な文末音調ラベル推定器の必要性が示唆された．一方，本研究では，表 8 の“その他”に含まれる多くの発話については，文末コンテキストや対話行為の観点から分類した場合，発話の事例数が少なく，一般性の高い考察を行うことが難しかった．このため，文末コンテキスト，文末音調ラベルおよび対話行為のバランスが取られた音声データベースを構築することで，より多様な発話に関する文末音調推定の検討や考察が可能となると考えられる．

6. ま と め

本研究では，音声対話システムの発話音声を作成する上で，適切な文末音調を推定するため，対話行為を考慮する文末音調推定を提案した．我々が先行研究において構築した，対話行為付き音声データベースに対し，文末音調ラベルのアノテーションを追加し，本研究の検討に使用した．文末音調ラベルの推定手法として，従来の発話文章の形態素や品詞の情報のみでなく，対話行為を雑談対話システムの対話処理部から取得し，推定のために使用する手法を提案した．文末音調ラベルの推定実験により，提案法による Cohen の一致係数 κ は，従来法に比べ約 0.12 改善することが示された．これにより，雑談対話システムで使用する対話行為タグセットを使用することで，文末音調ラベルの推定精度が改善することが示された．また，推定結果に対する考察により，提案法では，

1. 自問納得
2. 文末イントネーションによる疑問表現
3. 自問・フィラー

の発話意図を表現するために適切な文末音調が推定されたと考えられる．また，提案法に対する課題分析により，文末音調ラベルの推定精度を向上させるために必要な対話行為タグに関する考察を行った．考察の結果，“自己開示”等の対話行為について「自己開示等で言及される命題内容に，話者が強い感情を抱いていることが表明されるか」を区別すること，および，“繰り返し”等の対話行為について「話者が驚きや意外性を感じているか」を区別することが重要であることが明らかとなった．

本研究では，一つの文脈，発話文章および対話行為に対し，一つの文末音調ラベルを正解データとして定めた．一方で，文末音調の選択は，人間の発話者の間でも一致しない場合があると考えられる．このため，本研究の正解データは，一つの収録結果であり，一般的な発話の傾向を再現しているとは限らない．この場合，ラベルの推定精度が，一般的な発話の傾向の再現性と直接対応しな

いという懸念がある．今後の課題として，複数の発話者の音声を取録し評価実験を行うことで，より一般性の高い知見を得ることがあげられる．

また，本研究で確認された提案法の効果が，発話意図の正確な伝達や対話の自然性の向上に寄与するかを検証するためには，主観評価や対話者の反応による評価を行うことが考えられる．ただし，現段階では，発話意図の伝達に関し，これらの評価方法により信頼性の高い評価結果を得ることは困難である．このため，主観評価および対話者の反応による評価に関する評価手法の設計を含め，今後の課題とした．

◇ 参 考 文 献 ◇

- [Austin 75] Austin, J. L.: *How To Do Things With Words*, Oxford University Press (1975)
- [Black 96] Black, A. and Hunt, A.: Generating F_0 contours from ToBI labels using linear regression, in *Proc. Fourth International Conference on Spoken Language Processing*, Vol. 3, pp. 1385–1388 (1996)
- [Breiman 84] Breiman, L., Friedman, J., Stone, C., and Olshen, R.: *Classification and Regression Trees*, CRC Press (1984)
- [Cohen 60] Cohen, J.: A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 37–46 (1960)
- [Cutler 85] Cutler, A. and Pearson, M.: On the analysis of prosodic turn-taking cues, in *Intonation in Discourse*, pp. 139–155, Croom Helm (1985)
- [Dall 17] Dall, R.: *Statistical Parametric Speech Synthesis using Conversational Data and Phenomena*, PhD thesis, The University of Edinburgh (2017)
- [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)
- [Higashinaka 14] Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing, in *Proc. COLING 2014*, pp. 928–939 (2014)
- [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)
- [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)
- [Landis 77] Landis, J. R. and Koch, G. G.: The measurement of observer agreement for categorical data, *Biometrics*, pp. 159–174 (1977)
- [Maekawa 02] Maekawa, K., Kikuchi, H., Igarashi, Y., and Venditti, J.: $\mathcal{X}$ -JToBI: An extended J-ToBI for spontaneous speech, in *Proc. Seventh International Conference on Spoken Language Processing* (2002)
- [Misu 11] Misu, T., Mizukami, E., Shiga, Y., Kawamoto, S., Kawai, H., and Nakamura, S.: Toward construction of spoken dialogue system that evokes users' spontaneous backchannels, in *Proc. SIGDIAL 2011*, pp. 259–265 (2011)
- [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)
- [Ross 96] Ross, K. and Ostendorf, M.: Prediction of abstract prosodic labels for speech synthesis, *Computer Speech & Language*, Vol. 10,

- No. 3, pp. 155–185 (1996)
- [Searle 80] Searle, J. R., Kiefer, F., Bierwisch, M., et al.: *Speech Act Theory and Pragmatics*, Vol. 10, Springer (1980)
- [Shriberg 98] Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., Taylor, P., Ries, K., Martin, R., and Van Ess-Dykema, C.: Can prosody aid the automatic classification of dialog acts in conversational speech?, *Language and Speech*, Vol. 41, No. 3-4, pp. 443–492 (1998)
- [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)
- [Sugiura 15] Sugiura, K., Shiga, Y., Kawai, H., Misu, T., and Hori, C.: A cloud robotics approach towards dialogue-oriented robot speech, *Advanced Robotics*, Vol. 29, No. 7, pp. 449–456 (2015)
- [Sugiyama 13] Sugiyama, H., Meguro, T., Higashinaka, R., and Minami, Y.: Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures, in *Proc. SIG-DIAL 2013*, pp. 334–338 (2013)
- [Syrdal 08] Syrdal, A. K. and Kim, Y.-J.: Dialog speech acts and prosody: Considerations for TTS, in *Proc. Speech Prosody 2008*, pp. 661–665 (2008)
- [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)
- [Vapnik 98] Vapnik, V.: *Statistical Learning Theory*, Wiley New York (1998)
- [Venditti 05] Venditti, J. J.: The J.ToBI model of Japanese intonation, *Prosodic Typology: The Phonology of Intonation and Phrasing*, pp. 172–200 (2005)
- [Ward 00] Ward, N. and Tsukahara, W.: Prosodic features which cue back-channel responses in English and Japanese, *Journal of Pragmatics*, Vol. 32, No. 8, pp. 1177–1207 (2000)
- [Wilson 12] Wilson, D. and Sperber, D.: *Meaning and Relevance*, Cambridge University Press (2012)
- [菊池 03] 菊池 英明, 前川 喜久雄, 五十嵐 陽介, 米山 聖子, 藤本 雅子: 日本語話し言葉コーパスの音声ラベリング (<特集> 音声学と音声工学), 音声研究, Vol. 7, No. 3, pp. 16–26 (2003)
- [栗原 18] 栗原 清, 清山 信正, 熊野 正, 今井 篤: 読み仮名と韻律記号を入力とする日本語 end-to-end 音声合成の音質評価, 情報処理学会研究報告. SLP, 音声言語情報処理, 第 19 巻, pp. 1–6 一般社団法人情報処理学会 (2018)
- [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)
- [郡山 12] 郡山 知樹, 能勢 隆, 小林 隆夫: HMM に基づく対話音声合成における多様な韻律生成のためのコンテキストの拡張, 電子情報通信学会論文誌 D, Vol. 95, No. 3, pp. 597–607 (2012)
- [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 VI, pp. 5–28 (2008)
- [国立 60] 国立国語研究所: 話し言葉の文型 (1) -対話資料による研究-, 秀英出版 (1960)
- [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)
- [仁田 91] 仁田 義雄: 日本語のモダリティと人称, 第 4 巻, ひつじ書房 (1991)
- [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)
- [赤羽 11] 赤羽 誠, 水野秀之, 平沢純一, 中嶋信弥, 田中和世: 日本語テキスト音声合成用記号の標準化, 電子情報通信学会誌 D, Vol. 94, No. 12, pp. 1060–1066 (2011)
- [川上 63] 川上 泰: 文末などの上昇調について, 国語研究, No. 16 (1963)
- [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)
- [大島 13] 大島 デヴィッド義和: 日本語におけるイントネーション型と終助詞機能の相関について, 国際開発研究フォーラム, Vol. 43, pp. 47–63 (2013)

- [中嶋 14] 中嶋 秀治, 水野 秀之, 吉岡 理, 高橋 敏: テキストからの表現豊かな音声合成のための句末上昇ラベル予測における句末コンテキストの利用, 情報処理学会論文誌, Vol. 55, No. 1, pp. 553–562 (2014)
- [北条 16] 北条 伸克, 井島 勇祐, 杉山 弘晃: 対話行為情報を表現可能な音声合成の検討, 2016 年度人工知能学会全国大会 (第 30 回) 論文集 (2016)
- [北条 20] 北条 伸克, 井島 勇祐, 杉山 弘晃, 宮崎 昇, 川西 隆仁, 柏野 邦夫: 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, 人工知能学会論文誌, Vol. 35, No. 2, pp. A-J81.1 (2020)
- [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 10 月 15 日 受理

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了。2015 年 同研究科院博士課程修了。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。