情報処理学会論文誌 Vol.55 No.1 553–562 (Jan. 2014)

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

中嶋 秀治[1,a)] 水野 秀之[1] 吉岡 理[1] 高橋 敏[1]

受付日 2012年12月28日, 採録日 2013年10月9日

概要：表現豊かな音声では，疑問文ではなくても，句末で基本周波数（F_0）の上昇が生じる．このようなF_0の合成にとって，句末音調ラベルのようなトーンラベルの有効性が知られている．従来は句の長さや句の文内位置などの数量主体の特徴量を利用して句末音調ラベルの予測が行われたが，従来の音声学的分析によれば，句の担う意味に応じて句末音調ラベルが異なることが示されている．意味処理の代わりに，本論文では句末の複数個の単語の出現形と品詞と句末のポーズの有無の組合せからなる句末コンテキストの利用を提案する．日本語東京方言の表現豊かな音声を対象として，テキストからの句末上昇ラベルの予測実験を行ったところ，提案のコンテキストによって従来の特徴量と同等以上の性能が得られた．これにより，句末コンテキストの利用の有効性を確認できた．

キーワード：表現豊かな音声合成，テキストからの音声合成，韻律，基本周波数，句末音調，句末コンテキスト

Using Phrase End Context to Predict Phrase Boundary Rise Labels from Text for Expressive Text-to-Speech Synthesis

Hideharu Nakajima[1,a)] Hideyuki Mizuno[1] Osamu Yoshioka[1]
Satoshi Takahashi[1]

Received: December 28, 2012, Accepted: October 9, 2013

Abstract: Expressive speech shows phrase boundary rise of fundamental frequency (F_0) even if it is not an interrogative sentence. To synthesize expressive F_0, tone labels such as phrase boundary tone label are known to be useful. Though conventional label prediction mainly uses many numerical features such as phrase length and phrase location in a sentence, conventional phonetic analysis reveals a complex relationship between phrase semantics and phase boundary tone label. Instead of using semantic processing, this paper proposes the use of phrase end context, which consists of several word surface strings and their part-of-speech in the phrase and the existence or non-existence of pause at the phrase final position. Experiments on Japanese expressive speech, Tokyo dialects, that target phrase boundary rise label prediction show that the proposed phrase end context attains performance equal to or better than that of conventional features, confirming the usefulness of the phrase end context proposal.

Keywords: expressive speech synthesis, text-to-speech synthesis, prosody, fundamental frequency, phrase boundary tone, phrase end context

1. はじめに

ニュース文や情報案内文を対象とした読み上げ口調での音声合成は実際に利用されるようになり，新たな適用領域を目指した研究が行われている．そのような新領域として，e-Commerce や電話自動応対やエンタテイメントのような人に向かって表現豊かに語りかけるさまざまな場面があり，それぞれの場面での自然な口調を備えた音声の合成を目的とした研究が行われている [1]．本論文では，そのよ

[1] 日本電信電話株式会社 NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories, NTT Corporation, Yokosuka, Kanagawa 239–0847, Japan
[a)] nakajima.hideharu@lab.ntt.co.jp

© 2014 Information Processing Society of Japan

うな場面での使用を念頭に置いた口調や音声を「表現豊かな口調/音声（expressive style / speech）」と呼び，議論を進める．表現豊かな音声の再現には韻律特徴が重要な役割を果たすことが知られており，本論文ではそのうちの基本周波数（fundamental frequency, F_0）に焦点をあてる．

表現豊かな音声の F_0 の特徴の 1 つとして，アクセント句の末尾での F_0 の動きの多様性がある．従来の読み上げ口調の音声合成対象の文は情報提供目的の平叙文が多く，日本語の東京方言の場合，アクセント句ごとの F_0 の変化形状はおおむね「へ」の字型になる場合が多かった．一方，表現豊かな音声では，文が疑問文ではない場合でも，F_0 がアクセント核の位置で降下した後に，句末付近で再び上昇する場合がある．このような F_0 の動きにより，発話の継続意思を示すという談話機能や，聞き手に同意を求めるといった話者の意図が伝えられる．このような F_0 の現象は「句末音調」と呼ばれ，これまでにも F_0 変動の記述法の検討 [2], [3] や音調分類の音声学・音韻論的研究 [4], [5] が行われてきた．

F_0 の変化形状を記述する方法の 1 つとして，Tone and Break Indices（ToBI）[2], [3] の Tone label がある．ToBI の各シンボルの意味する F_0 変動が起こる時刻の予測 [6] や，ToBI ラベルからの F_0 時系列の生成については，HMM 音声合成の枠組みを用いるもの [6]，線形回帰モデルを用いるもの [7]，および，指令応答モデルを用いるもの [8] においてすでに効果が確認されている．その一方で，それらの音声処理の入力の一部となる ToBI ラベルのような F_0 の変化形状を記すラベルを予測する言語処理的研究は，従来の日本語の音声合成の研究においては，これまで行われてこなかった．これまでの日本語の読み上げ口調の音声合成の研究では，合成の目標を東京方言の音声にする場合が多く，かつ，合成対象の文は情報案内を目的とした平叙文が大半であった．この場合，おおむね「へ」の字であるという標準的な形状に関する知見と音声合成器の言語解析から得られるアクセント情報の双方から，ToBI で記述されるようなアクセント句内の F_0 の変化形状情報を得ることができるために，F_0 変化形状を表すラベルを予測する必要性が低かったと考えられる．しかし，読み上げ音声に比べて，表現豊かな音声における句末音調の発生頻度は高い [3], [4]．テキストの読み上げを主な用途とするこれまでの音声合成であれば，句末音調は大きな問題にはならなかったが，目的をもって人に語りかけるような用途 [1] での表現豊かな音声合成においては句末音調の再現は重要な要素の 1 つである．従来研究 [4] によれば，句末音調は伝達される対話音声の正しい理解にとって重要な役割を果たしており，表現豊かな音声合成において句末音調を実現することは，句末音調の発生割合の多寡にかかわらず，重要な課題である．句末音調の有無をテキストから予測できれば，既存の F_0 生成系の研究成果 [6], [7], [8], [9] と合わせることによって，表現豊かな音声のテキストからの合成の実現に必要な要素が揃う．

そこで本研究では，合成対象の文の言語表現から，句末での F_0 変化形状を記述する句末上昇ラベル，すなわち，F_0 が句末付近で再上昇するか否か，という 2 値のラベルを予測する言語処理的方式の検討を行った．実際の句末音調は上昇と下降の 2 種類ではなく，上昇後に下降するなどさまざまである．しかし，自然な音声に対しては，ラベリング従事者間で，句末上昇のみ（H%）か上昇後下降か（HL%）といった判定の不一致が増加し，複雑な動きの句末音調になるほど出現数も限られてくる．そこで，まずは 2 つのカテゴリへの予測として本研究では検討した．特に本論文では，この句末上昇ラベルの予測において，句末の複数個の単語の出現形とそれらの品詞で定まる単語種別と句末のポーズの有無をベクトルの形で表現した「句末コンテキスト」を利用する方式の有効性を述べる．この単語種別は単語出現形（字面）と品詞との組を短く表現するために導入した語であり，たとえば，同じ字面の単語出現形「と」には引用助詞と格助詞の別があるが，組にすることによって，“と_引用助詞”と “と_格助詞”が別扱いできるようになる．

以下，2 章で，句末コンテキストを導入する経緯を従来研究との関係の中で述べる．続いて，句末コンテキストを用いて句末上昇ラベルを予測する方法を 3 章で述べる．4 章で，本研究で用いる表現豊かな音声データベースの概要とその特徴について述べ，これを用いた評価実験について 5 章で述べる．

2. 従来研究との関係

日本語の音声を対象とするテキストからの句末音調の予測についての研究は，筆者が調べた限り，これまでになかった．一方，英語のラジオニュース読み上げを対象とする研究は存在した [10]．その研究では，句末の位置でのみ，前の句末での予測結果とは独立に，句末で基本周波数が上がるか下がるかといった句末音調ラベルの予測が行われた．予測には分類木が用いられ，予測のための特徴量（分類木の質問に現れる変数）として，音節，単語，minor phrase（日本語のアクセント句に相当），major phrase（日本語のイントネーション句に相当），文といったさまざまな言語単位の長さやそれらの単位と句末との距離，そして，句末の 2 単語および次の句頭の 1 単語の，辞書参照して得られる品詞や強勢の有無といった約 50 個の数量情報が多数を占める特徴量が利用された．カテゴリカルな情報の 1 つとして品詞が利用されたが，本論文で提案するような単語の出現形と品詞とを組にした単語種別や，句内の複数の単語種別を含む句末コンテキストは利用されなかった．

日本語のイントネーション句境界位置設定（phrasing）についての音声研究 [11] によれば，句末音調はイントネーション句の末尾に生起し，イントネーション句の末尾には

© 2014 Information Processing Society of Japan

同時にポーズが生じることが多いことが分かっているため，ポーズとイントネーション句境界とが共起しやすいことは容易に予測できる．逆にいえば，ポーズが予測できれば，イントネーション句境界が予測できることが示されているといえるが，句末音調が上昇か下降かまで予測できることは示されていない．また，イントネーション句境界とある種の統語構造上の境界（並列節の境界など）との対応関係についての研究はなされている [12]．しかし，統語構造上の境界種別と句末音調種別（上昇か下降か）との対応が明確になっているわけではない．そのため，節や句の言語情報と句末音調の生起との関係の把握が必要となる．

日本語の句末音調についての分析研究 [4] によれば，i) 句末音調と，ii)（文献 [4] などでは「文末詞」で論じられている）単語種別と，iii)（発話権の維持/開放などの）談話機能，の 3 つの関係が整理された．まず，単語が異なれば，それらの単語が同じ品詞を持っていても，句末音調が異なることが示されている．さらに，句末の単語種別が同一であっても，伝えられる談話機能が異なる場合には，句末音調が異なる場合があることが示されている．しかし，談話機能以外の意図などのその他の観点からの整理はまだ行われていないし，どれだけの観点での整理が必要かも明らかではない．ところが，談話機能や意図のような意味情報は品詞では伝達できないので，従来の予測法 [10] で用いられた句境界付近の品詞では句末音調の予測には不十分であることが予想され，句末の複数の単語種別から構成される句末コンテキストを考慮にいれる必要が示唆される．他の句末音調の生成と知覚についての研究 [5] によれば，F_0 の高さ，上昇の形状，上昇のタイミングを制御することにより，さまざまな句末音調を生成でき，それらの音調の違いによって意味情報の違いを表現できることが検討された．しかし，言語表現からの句末音調の予測までは行われていない．

意図や談話機能のような意味情報は，句末音調の決定に限らず，広く言語処理において渇望され，研究が進められてきた [13]．しかし，それらを高い精度で文から自動認識できる装置は現在存在しない．そのような意味情報を表すシンボルの定義や必要最小限のシンボル数すら変わりうる研究段階である．近年の言語処理では，そのような意味情報を表すシンボルの代わりに，単語の組合せに基づくコンテキスト表現を導入し，所望の処理結果に対応するコンテキストとの間でのコンテキストどうしの類似性計算を用いることにより，単語の意味多義解消やテキストの分類などを実現し，処理の実用化が導かれてきた [14]．本研究でも，意味情報が句末のコンテキストによって表現されることを期待し，この句末コンテキストの句末上昇の予測における効果の検討を行う．

3. 句末コンテキストに基づくテキストからの句末上昇ラベルの予測

2 章の状況から，日本語の表現豊かな音声の句末で生じる F_0 の再上昇を示すラベル（ToBI の H%に相当するシンボルラベル）を，テキストから抽出しうる情報に基づいて言語処理的に予測する方式を検討する．特に，予測に用いる特徴量としての句末の数個の単語種別と句末ポーズの有無の組合せからなる句末コンテキストの利用を提案し，その有効性を検討する．このように，句末のポーズの有無は既知として検討を行うので，句境界設定（文献 [12] など）の後段の位置の処理としての検討となる．

3.1 分類タスクとしてのラベル予測

本研究では，前述のように，句末上昇ラベルの予測問題を句末で F_0 が再上昇するか否かという 2 つのカテゴリへの分類問題として設定する．二値分類ができなければ細分類ができるとは考えにくいので，まずは二値分類から取り組んだ．しかし，二値分類であっても，つねに一定の下降調の音声しか合成できない場合と比較すれば，上昇が再現できることは有効と考えられる．

そして，この分類処理を各アクセント句の末尾で行うこととする．その理由は，従来の音調分類の研究 [4] の知見を活かし，従来の予測法の研究 [10] との比較を通して，句末上昇予測における句末コンテキストの効果の確認に焦点を当てるためである．これとは異なる課題設定の方法として，ToBI ラベルのようなシンボルラベル系列を，句よりも小さい音節や音素やさらにそれよりも小さなフレーム単位に合わせて，句頭または句末から順次予測するという課題設定もあげられるが，これはラベル予測に寄与する要因が明らかになった後の実装法の課題として別途検討することが可能である．

3.2 予測のための特徴量としての句末コンテキスト

表現豊かな日本語音声での句末上昇の予測のための特徴量の 1 つとして，単語の出現形とその品詞との組からなる「単語種別」を用いる．そして，句末の複数個の単語種別と句末のポーズの有無からなる句末コンテキストを用いる．以下の議論では単語種別を “word_POS” のように記す．

より長い句末コンテキスト，すなわち，句の末尾から句の先頭に向かって多くの単語を取り入れた句末コンテキストを用いることで句末音調の生起文脈を細分化できるが，その一方で，データスパースネスの問題を招く危険がある．本研究では，後述のように，分類器として，“Classification And Regression Tree（CART）”[15] や “Support Vector Machine（SVM）”[16] を予測に用いる．また，単語の出現形と品詞とを句末コンテキストを表すベクトルの別々の要素として与える．単語の出現形と品詞とで 1 つの

© 2014 Information Processing Society of Japan

新たなラベルとなるわけではないので，句末コンテキストが部分的に学習時の句末コンテキストと似ていれば，その似ている部分が予測に寄与する．たとえば，SVM では，句末コンテキストをベクトル表現する場合，サポートベクトルと入力ベクトルとの部分一致がカーネル計算において寄与し，CART では，分類力の高い句末コンテキストのいずれかの要素が木の根に近いノードに配置され，要素別に活用されるからである．そのため，希な固有表現の出現形が句末コンテキストに含まれる場合でも，その単語の品詞や周囲に現れる単語からの補償により，データスパースネスによる精度劣化が生じにくいと考えた．

3.3 モデルの予測範囲

本研究ではモデルの予測範囲を句末上昇の有無のみに限定する．意味情報や話者の特定は行わない．しかし，以下のような依存性が予想されるため，場面依存，かつ，話者依存のモデルを用いる．

場面依存性については，たとえば，日本語教育の文献 [17] によれば，「呼びかけ」という同じ意図を持つ発話でも，場面によって用いられる句末音調が異なる下記の事例がある．

- 句末上昇：「あなた ／，傘を忘れていますよ！」
- 句末上昇下降：「あなた ／＼，起きてください．」

このように，同じ「あなた」という語が呼びかけの意図で発話されている場合でも，句末音調は異なる．前者は屋外の騒がしい場面で去っていく人物に確実に声が届くように発話され，後者は家庭内のような静かな場面で発話される．このことは，逆に，意図単独では句末音調の決定要因にならないことも示している．合成対象の文の言語情報から場面を特定することも難しい．本研究では，予測モデルを場面ごとに持つことにより，句末コンテキストの有効性確認に焦点を当てるうえでの障害になりうる場面という要素を排除することにした．

また，予測対象は F_0 形状を表すシンボルであるが，シンボルの意味する句末音調の上げ下げ自体には話者の個人性が関係すると予想される．発話内容にかかわらず，ほぼすべてのアクセント句末を上昇口調で話す一部の女子学生や若者の話し方はその一例である．本研究で用いる文（発話内容）は 4 章で述べるが，童話，顧客への対話，商品宣伝といった内容であり，意味解釈の曖昧性を含む可能性が小さい素材である．そのため，同じ話者内であれば，意味解釈とそれによって用いられる句末音調の一貫性も期待できる．本研究では，予測モデルを話者ごとに持つことにより，句末コンテキストの有効性確認に焦点を当てるうえでの障害になりうる話者性という要素も排除することにした．

このように，場面とあわせて，場面依存，かつ，話者依存のモデルを用いることにした．これらにより，句末音調ラベルの予測への場面や話者や話者の意味解釈という要素の関与を排除し，本研究の焦点を句末コンテキストの有効性にあてることにした．本研究では，前述のような条件設定や後述するように従来の音声合成研究に比べて多くないデータ量での検討を行ってはいるが，最近の音声合成の研究では，話者と場面が限られた音声データや HMM 音声合成のコンテキストを用いた話者適応やコンテキスト情報の抽出の検討（文献 [18] など）がさかんに行われている．そのため，話者や場面が限られた少量の音声データの利用を前提とした句末音調の予測の実現は実用面においても有効であると考えられる．

4. 表現豊かな音声データベース

4.1 音声の発話場面とデータサイズ

本研究では，童話の語り聞かせ（FT），コールセンタオペレータの電話応対（OP），マスメディアを通した商品アピール（AP）といった 3 つの場面において自然な韻律で発話された音声合成の研究用の音声データベース [19] を用いる．どれもそれぞれの場面の実際の文を人に向かって語りかけるように発話された音声であり，FT 場面に見られる童話の場面を想像させるような音声，AP 場面に見られる明かるく元気のよい印象の音声，OP 場面に見られる楽しい印象の感謝や真剣な印象の謝罪などの音声を含む点から，表現豊かに発話された音声データベースとなっている．1 章で述べたように，目的をもって人に語りかけるような音声合成の実現を目的とするため，意図しない揺れや意味の曖昧性を含む発話とならないように，それらの曖昧性などの要因を含む文は音声収録前の文選択の段階で可能な限り排除している．

本研究で対象とする句末音調は，F_0 がアクセント句内のアクセント核の位置で 1 度下降した後で，アクセント句の末尾に向けて再上昇する現象である．これを意味するラベルが本データベースには付与されている．たとえば，

- FT：「六兵衛よ ／」と呼びかける発話
- OP：「盗難補償なんですが ／」と話題を提示する発話
- AP：「簡単でしょう ／」と同意を求める発話

などにおいてみられる．

このデータベースでは，前記の 3 つの場面の実際の文を，3 人の話者（全員東京方言の女性声優）が，各場面において自然で表現豊かな口調（expressive style）と，淡々とした読み上げ口調（reading style）の 2 種類の発話口調で発話している．収録発話数（文数）は，各話者のリテイク回数の違いなどにより，異なった．本研究では，口調間，および，話者間での比較検討を行うために，各場面において，2 つの口調の両方を，3 話者全員から収録できた発話のみを用いる．それらについての本音声データベースの諸元を**表 1** に記す．なお，読み方が複数ある語のルビや句読点が書かれたスクリプトを話者に提示したが，収録時には，アクセント句境界やポーズの位置の置き方などの発話についての韻律面での細かな指示は与えなかった．そのため，

© 2014 Information Processing Society of Japan

表 1 表現豊かな音声データベースの諸元
Table 1 Summary of expressive speech database.

	FT	OP	AP
♯ of sentences	64	104	152
average ♯ of accent phrases	684	1,061	1,550
♯ of speakers	3	3	3
Total length (average) [min.]	13	14	20

表 2 句末上昇ラベルの発生率の分布
Table 2 Occurrence distribution of phrase boundary rise labels.

(a) Expressive speech database.
expressive style (e), reading style (r)

[%]	FT		OP		AP	
	e	r	e	r	e	r
w/o pause	0.3	0.2	0.7	0.1	0.1	0.1
with pause	1.1	0.8	16.6	4.2	6.4	2.7
all	0.6	0.4	6.8	1.8	2.7	1.2

(b) CSJ database.

[%]	Int. Aca.	Int. Sim.	T.O.D.
w/o pause	0.07	0.05	0.03
with pause	21.6	18.5	18.1
all	15.8	13.4	12.9

表 1 のアクセント句数と音声データの合計の時間長は 3 話者で異なるので，平均値を記した．全部の句末を上昇させるような極端な話者もいなかった．

発話文を日本語形態素解析器 JTAG [20] を用いて単語に分割し品詞を付与した．句末コンテキストの有効性の検証に研究の焦点を当てるため，単語分割と品詞付与の結果を人手で修正した結果を用いて以後の実験を行う．品詞の数はおよそ 50 であった．

4.2 句末上昇ラベルの発生率

表 1 の分量の本音声データベース内での句末上昇の発生率 [%] を**表 2** (a) に示す．FT，OP，AP といった場面間でアクセント句数が異なる状況下での比較のため，頻度ではなく，発生率を記した．‘e’ の列は expressive style の略で表現豊かな音声，‘r’ の列は reading style の略で読み上げ音声である．最下の “all” の行は全アクセント句で計算した句末上昇の発生率である．“with pause” と “w/o pause” の各行は末尾にポーズをともなう，および，ともなわないアクセント句での発生率である．FT では口調によらず発生率は 1%以下で，ほとんど発生しなかった．さまざまな情景描写の文が発生率を下げたと考えられる．会話部分が増えると FT での発生率が変わってくると考えられるが，会話部分については他の OP や AP を対象とした検討結果から補うことが可能であると考えられる．よって，以後は OP と AP のみを対象とする．

表 2 (a) の all の行のように，OP と AP では全体の数%のアクセント句で句末上昇が発生した．e 列と r 列との比較から，表現豊かな口調では句末上昇の発生率が高い様子が確認された．OP の “with pause” における r 列と e 列における発生率が 10 ポイント以上も異なる点からも，表現豊かな音声のテキストからの合成における句末上昇の再現は重要な問題であると確認できる．また，表 2 (a) の “with pause” と all の行との比較から，末尾にポーズをともなうアクセント句での発生率は全体で見た場合の 2 から 3 倍に高まる．この結果は，句末のポーズの有無が予測のための 1 つの特徴量として有効である可能性を示唆する．これはこの特徴量に相当する情報が “depth of phrase break” として従来研究でも利用されたことと符合する [10]．しかし，pause をともなうすべての句末で句末上昇が発生しているわけではないため，従来のテキストからの句読点予測の問題とは問題の性質が異なる．

一般に入手可能な日本語話し言葉コーパス（CSJ コーパス）[21] の 3 種の対話音声（学術講演インタビュ（Int. Aca.），模擬講演インタビュ（Int. Sim.），タスク指向対話（T.O.D.））での句末上昇ラベルの発生率を表 2 (b) に記す．この CSJ コーパスと表 1 の電話応対場面の音声データ（OP）との間で，ポーズを句末にともなう場合の句末上昇ラベルの発生率は表 2 (a) の値とほぼ近い値（16〜21%）になっており，本音声データベース [19] の電話応対対話（OP）での句末上昇の発生率が特殊ではなさそうであることが確認された．

4.3 特徴量の分類力

ここでは，分類のための特徴量の有用性，すなわち，分類力を確認する．分類力を計る尺度として，それぞれの特徴量を用いることによって得られる情報利得（相互情報量）を計算した．ここで，Y を句末で再上昇するか否かを示す変数とし，X を分類のための特徴量とするとき，$H(Y)$ が Y の Entropy，$H(Y|X)$ が条件付き Entropy である．このとき，情報利得 IG は

$$IG = H(Y) - H(Y|X)$$

と定義される．特徴量 X の導入による情報利得 IG が大きいほど，X の分類力が高いことを意味する．X には句内の各位置の単語の出現形，品詞，句末ポーズの有無などの 1 つ 1 つの特徴量が該当する．

品詞（POS），単語出現形（word），および，単語の出現形と品詞の組である単語種別（word_POS）を特徴量として，話者ごとに算出した情報利得を**表 3** に記す．話者間で発話数（文数）は同一であっても句末上昇のラベル数が話者ごとに異なるので，特徴量で条件付けする前の Entropy（$H(Y)$）の値が話者ごとにそもそも異なる．表 3 の word や word_POS の行のように，単語の出現形（表 3 の word

© 2014 Information Processing Society of Japan

表 3 単語，品詞，単語種別ごとのエントロピーと情報利得（分類力）
Table 3 Entropy and information gain (as classification power) by word, POS, and word_POS.

Domain			speaker id		
		X	#1	#2	#3
OP	Entropy ($H(Y)$)		0.48	0.27	0.30
	IG	POS	0.22	0.15	0.16
		word	**0.27**	**0.20**	**0.22**
		word_POS	**0.30**	**0.21**	**0.22**
AP	Entropy ($H(Y)$)		0.17	0.20	0.18
	IG	POS	0.12	0.17	0.13
		word	**0.14**	**0.18**	**0.15**
		word_POS	**0.15**	**0.19**	**0.15**

表 4 話者間で共通の位置で句末上昇が発生する割合
Table 4 Occurrence ratios [%] of common phrase boundary rise label among speakers.

speaker id	#1	#2	#3
OP	37.3	78.8	66.1
AP	92.1	77.8	83.3

の行）や単語種別（word_POS の行）のほうが，従来法 [10] で用いられた品詞（POS の行）に比べて情報利得が大きく，分類力が高いと予想される．これにより，従来法 [10] では用いられていない特徴量である単語や単語種別を用いることが有望であることが確認された．従来法 [10] で用いられた句の長さや句の文内での位置などの特徴量についても同様に情報利得を算出したが，単語や品詞の示す表 3 の情報利得に比べると，はるかに小さい値（表 3 にあげた情報利得の半分にも至らないわずかな値）であった．

4.4 句末上昇発生率の個人性

表 3 の Entropy や IG の値が個人ごとに異なることから，句末上昇の生起する位置が個人ごとに異なることが予想される．そこで，3 話者で共通する句末上昇位置の割合を話者ごと，かつ，場面ごとに調査した．結果を**表 4** に記す．

表 4 のように，場面 AP の話者#1 の場合のように，他の 2 人の話者と同じ位置において高い割合で句末上昇を発話した場合もあったが，他の 5 つの場合のように，一致割合は 37〜83%とさまざまで，全話者が共通の箇所で句末上昇の発話を行っているとはいい難い．この結果からも，話者ごとのモデルを用いることの妥当性が支持される．

5. 実験

5.1 方法

句末上昇ラベル予測への句末コンテキストの有効性を確認する．1 章で述べたように，句末音調を含む F_0 変化形状を表すシンボルから F_0 を生成できることはすでに確認されている [6], [7], [8], [9]．そのため，ここではテキストから各句末において F_0 が再上昇するか否かのシンボルの予測のみに焦点を当てて評価を行う．また，この目的のため，正確な単語境界，品詞，句末のポーズの有無の情報を入力とする．

表 5 実験設定
Table 5 Experimental configuration.

configuration id	prediction feature	classifier
c_1	Phrase end context	SVM
c_2	Phrase end context	CART
c_3	Conventional feature	CART

ここでは**表 5** の 3 つの設定での実験を行う．まず，本論文で提案する句末コンテキストと分類器としての SVM を用いる設定 c_1 での実験を行う．次いで，従来法との比較においての条件を揃えるため，本提案の句末コンテキストと分類器としての CART を用いる設定 c_2 での実験を行う．最後に，従来法の特徴量と分類器を用いる設定である c_3 を行う．

設定 c_1 と c_2 では本論文で検討してきた句末コンテキストを特徴量とする．句末コンテキストは，

(1) アクセント句末にポーズが有るか否かを示すラベル
(2) アクセント句の末尾から句の頭に向かってとった N 個の単語の出現形とそれらに対応する品詞

から構成される．句の中の単語数が N に満たない場合には，単語がないことを示すラベルを不足数だけ用いる．本実験では CART には Wagon [22] を，SVM には TinySVM [23] を用いる．

設定 c_3 は，従来法 [10] の特徴量と分類器（CART）とを用いる場合である．文献 [10] の対象は英語であったので，特徴量が元々英語用であった．本実験の設定 c_3 ではそれらの特徴量の一部を日本語用に置き換えたものを利用する．特徴量の置き換えの詳細は付録に記した．

それぞれの設定で，各場面での各話者の全データを 5 つに分け，5 分割交差実験を行う．

5.2 評価尺度

本研究の予測課題に対しては，句末での音調が文脈に応じて上がるべきところで上がり，下がるべきところで下がる必要がある．句末の上がり下がりを示す正解と予測結果との一致率を計るのであるが，表 2 (a) のように，句末で F_0 の再上昇の起こらない句の数のほうが圧倒的に多い．そのため，頻度の偏りを考慮に入れない正解率を用いると，圧倒的に数の多い下降音調に埋もれて，数の少ない句末上昇の適切な評価が行えなくなる．そこで，評価尺度として Cohen の *kappa* [24] を用いることにより，句末での F_0 再上昇が生じる場合と生じない場合との間での出現頻度の偏りの影響を排除し，かつ，再上昇が生じる場合と生じない場合の両方の評価を行う．この尺度は -1 から 1 までの値

© 2014 Information Processing Society of Japan

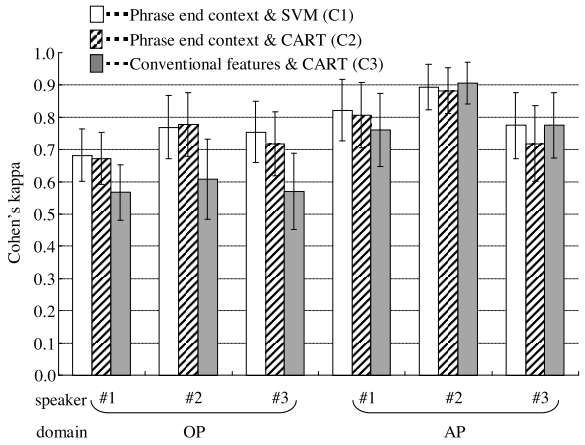

図 1 *Cohen* の *kappa* での性能比較（バーは 95%信頼区間）
Fig. 1 Performance comparison with *Cohen's kappa* coefficients (Small bars mean 95% confidence intervals).

をとり，1 に近いほど正解との一致度が高いことを意味する．値が 0.6 以上であるとき，正解と予測結果の 2 つのラベル系列間には「実質上の一致がある」と解釈される [25].

5.3 結果と考察

設定 c_1 の SVM には多項式カーネルを用い，句末コンテキストに導入する単語の数 N を 1 から 5 まで，コンテキストの要素間の組合せの数を表すカーネルの次元を 2 から 4 まで変化させたところ，場面 OP では，N が 4，かつ，次元が 2 または 3 の場合が，また，場面 AP では，N が 2 または 3，かつ，次元が 2 の場合が *kappa* の最大値を示した．

設定 c_2 や c_3 のように，CART のような分類木をデータから構築する場合には，一般に，（事例数に基づく基準よりも）情報量基準が用いられ，さらに，過学習を防ぐ目的で木の枝刈りが用いられる．本実験でもこれらを試したが，事例数に基づく基準，および，枝刈りを用いない場合に比べて，予測性能が著しく劣化した．このため，CART を用いる本実験の設定 c_2 と c_3（表 5）では，事例数基準で，かつ，枝刈りを行わない実験条件で分類木を構築した．また，設定 c_2 での句末コンテキストに含まれる単語の数 N は，条件を揃えるために，設定 c_1 の場合の値と揃えた．

kappa 値を**図 1** に示した．左から順に，白い棒グラフが表 5 の設定 c_1，つまり，句末コンテキストと SVM，真ん中の斜線のかかった棒グラフが設定 c_2，つまり，句末コンテキストと CART，であり，どちらも本提案の句末コンテキストを用いる場合の結果である．右の灰色の棒グラフは設定 c_3，つまり，従来法の特徴量と CART での結果である．すべて，棒グラフの高さが *kappa* 値で，各棒グラフの上部で上下に飛び出た細いバーが各 *kappa* 値の 95%信頼区間である．左側の 9 つの棒グラフが場面 OP での，右側の 9 つが場面 AP での結果であり，話者ごとに異なる棒グラフで示している．

5.3.1 特徴量の比較：設定 c_2 対 c_3

特徴量の比較のため，設定 c_2（斜線の棒グラフ）と c_3（灰色の棒グラフ）との間での結果を比較する．

場面 OP では，提案の句末コンテキスト（設定 c_2）のほうが，従来の特徴量（設定 c_3）に比べて，*kappa* 値が高く，また，この差は有意であった（χ^2 値を用いた独立性の検定・5%水準で）．さらに，従来の特徴量では，*kappa* 値が（実質上の一致があると解釈される）0.6 を超えない話者もあったが，句末コンテキストを用いる場合はどの話者でも 0.6 以上となり，正解との実質上の一致と見なせる性能が得られた．表 3 の情報利得が示唆したように，特徴量に単語の出現形を導入した句末コンテキストを用いることによって，場面 OP での大きな改善が得られた．一方，場面 AP では，どちらの特徴量も高い *kappa* 値を示し，統計的な有意差もなく，ほぼ同等であった．場面 AP においては，表 3 のようにそもそもの Entropy が低く，効果が場面 OP ほどは大きく現れなかった．

以上の実験結果から，本提案の単語の出現形，品詞，ポーズの有無からなる句末コンテキストによって，句末上昇の起こる文脈と近い文脈を捉えることによって，従来の特徴量と同等以上の性能が得られたと考えられる．

5.3.2 分類器の比較：設定 c_1 対 c_2

どちらも同じ句末コンテキストを用いた実験である．設定 c_1 は SVM を用いた実験結果で，得られた *kappa* 値は図 1 の白い棒グラフである．設定 c_2 は CART を用いた実験結果で，各 *kappa* 値は設定 c_1（白い棒グラフ）とほぼ同等であり，どの場合も統計的な有意差はなかった．よって，句末コンテキストを特徴量とする場合，分類器の違いによる差はないと考えられる．

5.3.3 実験結果から見た本提案の特徴

句末上昇の有無を判定するアクセント句からとることのできる句末コンテキストと上昇／下降という句末音調との共起に基づいて予測を行う方法を検討した．合成対象の文から抽出しにくい点で句末コンテキストの有効性を確認するうえでの障害になりうる場面と話者の要因を排除する目的で，場面依存かつ話者依存のモデルとして構成した．

句末コンテキストに取り込む単語の数に関しては，実験結果の冒頭に示したように，場面 OP では，場面 AP に比べて，長い単語数 N を要した．OP においては，たとえば，推察であることを伝える「程度か ／ と / （思います）」のように，分類に重要そうな助詞「か」が句末（'/' が句境界）直前に限らず，句末よりも文頭側の位置にくる事例が見られた．その結果，さらに前の単語も含めた類似性比較が必要であり，単語数 N が大きくなったと予想される．

句末上昇すべきところで上昇し，下降すべきところで下降することを決定できているかどうかを測る尺度として Cohen の *kappa* を用いて，句末コンテキストから決定可能な割合を図 1 に示した．OP では，「～なんですが」，「え

© 2014 Information Processing Society of Japan

えとですね」,「～でしょうか（疑問符なしで実験)）」などの表現で，APでは，「～いるんですね」,「～ますね」などの表現で句末上昇を正しく予測できていた．一方，誤りを分析した結果，似通った句末コンテキストに対して学習データと評価データの間で句末音調が異なるという未学習になっている場合が誤りの半数近く存在した．たとえば，OPでは，「～ように」,「～ても」,「～という」などの表現や，APでは，「〔形容詞〕で」という表現で起こっていた．この原因が話者内の揺れであれば，学習データを増やすことで，そのような対立が学習データに含まれてくるので，学習時の素性編成の段階で，学習データの空間の細分化につながる長い句末コンテキストが選択されることで，改善に向かう見込みがある．一方，誤りの原因が学習データの空間の細分化につながる要素の不足であれば，今回は大きな効果がなく用いなかった前後のアクセント句の単語（や品詞）の導入があげられる．たとえば,「～ように」の例では，発話末の場合は上昇し，次のアクセント句に「してください」などがある場合には下降する例であり，改善が見込まれる．しかし，今回大きな効果が得られなかったように，導入の仕方には工夫が必要である．これらは今後の課題とした．

6. おわりに

テキストからの表現豊かな音声合成の実現に向けて，表現豊かな音声においてさまざまな動きを示す F_0 の句末上昇の有無を言語処理的に予測する際の特徴量として，合成対象の文から得られる句末の数個の単語の出現形，品詞，および，句末のポーズの有無からなる句末コンテキストを提案した．この句末コンテキストの句末上昇の予測における利用は新しく，従来の句の長さや句の文内の位置などの数量を主な要素とした特徴量とは異なる．句末音調の音声学的分析によって示された音調に対応する意味解釈を，句末コンテキストによってとらえることを期待して導入した．情報利得に基づく予測力の点では，従来の特徴量に比べると，本提案の句末コンテキストのほうが句末上昇ラベルの予測力が高いことを確認できた．また，テキストからの句末上昇ラベル予測実験の点では，正解との一致率（*Cohen's kappa*），および，その統計的検定により，商品宣伝アピール場面（AP）で従来の特徴量とほぼ同等，対話場面（OP）で従来の特徴量からの改善が確認された．以上の結果から，単純な特徴量ではあるが，句末コンテキストの有効性を確認できた．HMM音声合成方式での表現豊かな F_0 の生成においては，句末上昇ラベルの導入により F_0 生成がある程度改善することはすでに確認されている [6], [9]．本提案は，少なくとも上記の文脈において，テキストからの表現豊かな音声の F_0 生成の実現に寄与する．以上のように，話者や場面への依存性や発話ごとの揺れによって句末音調は言語的情報のみから完全に予測できるものではないものの，まずは二値分類問題として仮定して定式化することで工学的には有用であることを確認できた．

本研究では上昇から始まる句末音調のすべてを句末上昇として扱ったが，句末での上昇と下降を繰り返す複雑な変動を示す場合もある．本研究により句末の最初の再上昇は予測できるが，その後の変動の細分類は今後の課題の1つである．また，本研究は音声合成の実現に向けたテキスト解析の一機能としての句末上昇の予測を検討し，句末コンテキストの有効性を確認した．表現豊かでも意味解釈の揺れる音声を合成する目的は想定しにくいため，表1の曖昧性のほとんどないデータで実験を行った．さらに自然な対話での検討を行う場合には，CSJコーパスでの評価が有効と考えるが，今後の課題とした．

参考文献

[1] Tang, H., Zhou, X., Odisio, M., Hasegawa-Johnson, M. and Huang, T.S.: Two-stage Prosody Prediction for Emotional Text-to-speech Synthesis, *Proc. Interspeech*, pp.2138–2141 (2008).

[2] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J.: ToBI: A Standard for Labeling English Prosody, *Proc. ICSLP*, pp.867–870 (1992).

[3] 前川喜久雄，菊池英明，五十嵐陽介：XJToBI：自発音声の韻律ラベリングスキーム，情報処理学会研究報告音声言語情報処理，39-23, pp.135–140 (2001).

[4] 石井カルロス寿憲，ニック キャンベル：句末音調の機能的役割：談話機能を中心に，日本音響学会2004年春季研究発表会講演論文集，Vol.I, pp.235–236 (2004).

[5] Venditti, J.J., Maeda, K. and van Santen, J.P.H.: Modeling Japanese Boundary Pitch Movements for Speech Synthesis, *3rd ESCA Workshop on Speech Synthesis* (Jenolan Caves, Australia), pp.317–322 (1998).

[6] 郡山知樹，能勢 隆，小林隆夫：対話音声合成のためのイントネーションラベルのタイミング予測，日本音響学会2011年秋季研究発表会講演論文集，pp.333–334 (2011).

[7] Black, A.W. and Hunt, A.J.: Generating F_0 Contours from ToBI Labels Using Linear Regression, *Proc. ICSLP*, pp.1385–1388 (1996).

[8] 平井俊夫，樋口宣男：韻律ラベリングシステムJ_ToBIのラベル情報を用いた重畳型基本周波数制御モデルパラメータの自動抽出，電子情報通信学会論文誌，D-II, Vol.J81-D-II, No.6, pp.1058–1064 (1998).

[9] 前野 悠，能勢 隆，小林隆夫，井島悠介，中嶋秀治，水野秀之，吉岡 理：多様な発話様式によるHMM音声合成のための韻律コンテキストの検討，日本音響学会2011年春季研究発表会講演論文集，pp.385–386 (2011).

[10] Ross, K. and Ostendorf, M.: Prediction of Abstract Prosodic Labels for Speech Synthesis, *Computer Speech and Language*, pp.155–185 (1996).

[11] Venditti, J.J.: The J_ToBI Model of Japanese Intonation, *Prosodic typology: The phonology of intonation and phrasing*, pp.172–200 (2005).

[12] Selkirk, E.: On Clause and Intonational Phrase in Japanese: The Syntactic Grounding of prosodic constituent structure, 言語研究，Vol.136, pp.35–73 (2009).

[13] 石崎雅人，伝 康晴：談話と対話（言語と計算），東京大学出版会 (2001).

[14] Manning, C. and Schuetze, H.: *Foundations of Statisti-*

© 2014 Information Processing Society of Japan

cal Natural Language Processing, MIT Press (1999).
[15] Breiman, L., Friedman, J., Olshen, R.A. and Stone, C.J.: *Classification and Regression Trees*, Wadsworth & Brooks (1984).
[16] Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer (1995).
[17] 窪園晴夫，田中真一：日本語の発音教室　理論と練習，くろしお出版 (2001).
[18] Maeno, Y., Nose, T., Kobayashi, T., Koriyama, T., Ijima, Y., Nakajima, H., Mizuno, H. and Yoshioka, O.: HMM-based Expressive Speech Synthesis Based on Phrase-level F0 Context Labeling, *ICASSP*, pp.7859–7863 (2013).
[19] Nakajima, H., Miyazaki, N., Yoshida, A., Nakamura, T. and Mizuno, H.: Creation and Analysis of a Japanese Speaking Style Parallel Database for Expressive Speech Synthesis (2010), *Oriental COCOSDA*, paper-id=30, available from ⟨http://desceco.org/O-COCOSDA2010/proceedings/paper_30.pdf⟩ (accessed 2012-12-28).
[20] Fuchi, T. and Takagi, S.: Japanese Morphological Analyzer Using Word Co-occurrence – JTAG, *Coling-ACL*, pp.409–413 (1998).
[21] 前川喜久雄，籠宮隆之，小磯花絵，小椋秀樹，菊池英明：日本語話し言葉コーパスの設計，音声研究，Vol.4, No.2, pp.51–61 (2000).
[22] Edinburgh Speech Tools Library, available from ⟨http://festvox.org/docs/speech_tools1.2.0⟩ (accessed 2012-12-28).
[23] TinySVM: Support Vector Machines (2012), available from ⟨http://chasen.org/˜taku/software/TinySVM⟩ (accessed 2012-12-28).
[24] Cohen, J.: A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, Vol.XX, No.1, pp.37–46 (1960).
[25] Landis, J.R. and Koch, G.G.: The Measurement of Observer Agreement for Categorical Data, *Biometrics*, Vol.33, No.1, pp.159–174 (1977).

付　　録

本実験の設定 c_3 では，Ross ら [10] の用いた特徴量を，本質を損なわないよう配慮しつつ，日本語用に置き換えた．その詳細を文献 [10] の末尾のリストに対応付けて以下に記す．全般的な点では，Ross らは句や単語の長さや位置を測る単位としてシラブルを用いたが，本研究ではモーラに置き換えた．また，minor phrase をアクセント句，major phrase をポーズにはさまれたアクセント句の系列（ほぼイントネーション句相当）に置き換えた．

A.1 Dictionary information

- Dictionary stress：英語では強勢の位置を利用していたが，対象が日本語であるためアクセント核の位置に置き換えた．ここでは平板型で核がないか，起伏型で核がくる位置が長母音上か短母音上か，機能語か，の 4 つのカテゴリとした．

A.2 Part-of-speech

- 品詞の種類：Ross らは当時のシステム上の制限から 8 種類に限定したが，今回用いたツールには制限がないので，データに現れた全種の品詞を用いた．
- 単語クラスのカテゴリ：Ross らの用いた *less/more-accentable function word* は用いず，機能語，内容語，固有名詞の 3 つのカテゴリとした．

A.3 Prosodic phrase structure

- phrase break size：アクセント句末でのポーズの有無を示す 2 カテゴリとした．

A.4 New/Given status

- 本音声データベースは音素バランスを優先して収集されたので，*story* や *paragraph* は存在しない．OP 場面の収録で，対話状況を示すために話者に提示した前発話と当該の発話との間のみでの情報の新旧を示すバイナリの情報を用いた．

A.5 Paragraph structure

- 上と同様の理由で，paragraph 内での句や文の位置や長さについての情報は用いなかった．

A.6 Label of other units

- pitch accent：Ross らの論文では，“none, high, downstepped, low” の 4 種であったが，アクセント句のアクセント型とした．
- boundary tone on word：本研究ではこれが予測対象であるので，用いていない．
- previous boundary tone：未使用との記載が文献 [10] 内にあったので用いていない．
- number of syllables from the last prominence：句のアクセント核の位置から句末までの距離とした．平板型の場合は 0 とした．8 以上は 8 とした．
- last preceding pitch accent type in the phrase：句のアクセント型か平板型であれば，‘H’，その他は ‘L’ とした．
- number of pitch accents since last minor break：句のアクセント型が平板型であれば 0 個，起伏型であれば 1 個とした．
- number of pitch accents since last major break：上記の累積数とした．

© 2014 Information Processing Society of Japan

中嶋 秀治 （正会員）

1990 年徳島大学工学部情報工学科卒業．1992 年徳島大学大学院工学研究科情報工学専攻修了．同年日本電信電話株式会社入社．1997～2002 年国際電気通信基礎技術研究所（ATR）へ出向．2002 年 4 月 NTT へ復帰．2010 年 3 月早稲田大学大学院博士後期課程修了．博士（国際情報通信学）．音声翻訳対話に関わる音声言語情報処理の研究開発，現在は音声合成のテキスト解析処理の研究開発に従事．現在，NTT メディアインテリジェンス研究所音声言語メディアプロジェクト研究主任．電子情報通信学会，日本音響学会，日本音声学会，言語処理学会，日本認知科学会各会員．

水野 秀之

1986 年名古屋大学工学部電気電子学科卒業．1988 年名古屋大学大学院工学研究科情報工学専攻修了．同年日本電信電話株式会社入社．以来，声質変換，音声合成の研究開発に従事．博士（工学）．現在，NTT メディアインテリジェンス研究所音声言語メディアプロジェクト主任研究員．電子情報通信学会，日本音響学会各会員．1993 年日本音響学会技術開発賞受賞．

吉岡 理

1990 年三重大学工学部電子工学科卒業．1992 年三重大学大学院工学研究科電気工学専攻修了．同年日本電信電話株式会社入社．以来，音声認識を中心とした音声言語情報処理，音声対話システムの研究開発に従事．現在，NTT メディアインテリジェンス研究所音声言語メディアプロジェクト主幹研究員．電子情報通信学会，日本音響学会各会員．

高橋 敏 （正会員）

1987 年早稲田大学理工学部電気工学科卒業．1989 年早稲田大学大学院理工学研究科電気工学専攻修了．同年日本電信電話株式会社入社．以来，音声認識を中心とした音声言語情報処理の研究開発に従事．現在，NTT メディアインテリジェンス研究所音声言語メディアプロジェクト主席研究員．博士（情報科学）．電子情報通信学会，日本音響学会各会員．1993 年日本音響学会粟屋潔学術奨励賞受賞．

© 2014 Information Processing Society of Japan