日本語 HMM 音声合成のコンテキストラベルの改良*

☆橋本浩弥, 鈴木雅之, 広瀬啓吉, 峯松信明 (東大)

1 はじめに

最近のモバイル端末において、音声認識、音声合成を利用したサービスが搭載されるようになり、音声をインターフェースに利用したシステムが注目されている。現在利用されている音声合成は、主に音声波形の素片を繋いでいく、波形接続型のシステムが主流であるが、近年、統計的手法を用いた手法が注目されている。その代表例として、隠れマルコフモデル（Hidden Markov Model; HMM）に基づく音声合成システムがある [1]。このモデルでは、音声分析再合成技術を用いることによって、音声の波形を直接取り扱うのではなく、特徴量ベースで取り扱い、学習用音声コーパスから HMM を学習する。これに適応や変換をかけることにより、従来の波形接続方式と比べて、比較的容易に話者、あるいは感情、発話スタイルを様々に変化させた音声を実現することができることが知られている [2, 3]。この HMM を用いた手法は元々、音声認識において発達してきた手法であり、基本的な要素は共通している。音声認識においては、その目的から主に音韻性のみが注目されるため、該当音素に対し、前後の音素で区別したトライフォン (triphone) が HMM の単位として主に用いられているが[1]、音声合成においては、自然な音声を実現するために、音素に加えて様々なコンテキスト情報を加えたラベルが用いられている。しかし、ラベルの種類を多くしていくと、存在しうるラベルの組み合わせが爆発的に増加していくため、ほとんどのラベルにおいて、データスパースネスの問題が発生する。これに対して、ベイジアンネットワークを用い、ラベル同士の因果関係を見出すことによって、重要なラベルを取捨選択する手法や [4]、従来のアクセント型に基づくコンテキストに代わり、音声の基本周波数 (F0) を音素ごとに量子化したものをコンテキストとして用いる手法 [5] などが提案されているが、あまり効果を上げていない。音声認識では、様々な話者による多様な音声から話者性を取り除き、発話内容を決定する問題であるため、統計的機械学習と非常に相性が良いが、音声合成はその逆問題であるため、自然で多様な音声を実現するためには、音声の特徴を適切に捉えた、品質の良いラベルが重要であると考えられる。そしてそのラベルの種類はなるべく少数であり、テキストから容易かつ安定に推定されるものである必要がある。しかし、従来の日本語音声において用いられているラベルには、次のような問題がある。ラベルに用いられているアクセント句は、定義に曖昧性があり、テキストだけではなく、話者や話者の発話速度、発話スタイルに依存するため、自動推定することが困難である。また、文の長さや文中における位置情報がラベルとして用いられているが、任意の長さの文を生成可能にするためには、非常に多くのラベルの種類を必要とする上、文の一部分だけが変化した場合に、文全体のラベルが変化してしまう。

そこで本稿では、曖昧性の少ないラベルを用い、文の長さに依存せず、ラベルの自動推定が従来に比べて容易になるようなコンテキストラベルを提案する。そして、そのラベルの有効性を聴取実験により確認する。

2 コンテキストラベル

HMM 音声合成では、音素環境だけでなく、テキストから抽出・推定される様々な（コンテキスト）ラベルで分類されたカテゴリ毎に HMM が構築される。まず、代表的な HMM 音声合成システムである HTS[2] で用いられている日本語音声用ラベルについて述べてから、その問題点を指摘し、それを改善するラベルを提案する。

2.1 従来手法のコンテキストラベル

従来用いられてきたコンテキストラベルを表 1 に示す。韻律に関するラベルはアクセント句単位で、定義されていることがわかる。一般に、句頭においてピッチの上昇を伴う場合をアクセント句境界があるとし、ピッチが下降する直前のモーラをアクセント核と呼ぶ。アクセント核は、アクセント句につき高々 1 個のアクセント核があるとすることが多いが、ピッチの上昇を伴わない（少ない）場合、副次アクセントとして定義されることがある。しかし、ピッチの上昇を伴うか、伴わないかは明確に区別できるものではないという問題がある。アクセント句境界は主にテキストのみから推定されることが一般的であるが [6]、本来、話者の発話速度、発話スタイルによって変化するものである。そのため、学習データにおいては、テキストだけではなく、音声の基本周波数も利用して自動推定する研究も提案されてはいるが [7], 現状では手動で抽出することが多いのが実情である。また、従来用いられているラベルにあるアクセント句の位置は、ある同じテキストを読み上げた 2 つの音声について、一部分だけアクセント句の長さが異なる場合、その後続部分が同じ発話構造をもっていたとしても、ラベルが異なったものになってしまうという問題がある。そして、アクセント句や、呼気段落は発話によっ

*Improvement of context labels in HMM-based speech synthesis for Japanese by Hiroya HASHIMOTO, Masayuki SUZUKI, Keikichi HIROSE, and Nobuaki MINEMATSU (The University of Tokyo)

[1]前後 2 つまで考慮したクインフォン (quinphone) も広く用いられている

[2]HTS, http://hts.sp.nitech.ac.jp/

ては、文の長さと同様、明確な上限がないため、任意の文章を生成可能にするためには、非常に多くのラベル数を必要とし、可能なラベルの組み合わせが爆発的に増加するという問題がある。

2.2 提案手法によるコンテキストラベル

前節で指摘した問題点を踏まえて、設計方針としては、発話スタイルによって長さが変わってしまう情報や、絶対的な位置情報（呼気段落におけるアクセント句の位置や、文中における呼気段落の位置）を用いず、可能な限り相対的な（直前直後の）情報を用いることにより、1 文の長さに、ラベルの種類が依存しないようにする。

提案手法によるコンテキストラベルを表 2 に示す。提案手法の特徴として、次のようなものが挙げられる。

- アクセント句の代わりに、文節を用いている。

文節は、アクセント句に比べて、話者性に依存せず、言語情報のみから一意に決定されるものであるため、曖昧性が少ないという利点がある。文節境界は、名詞連続の場合を除いて、ほぼ正確に自動推定することができる。ただし、「…、という…」のようなケースは読点直前と読点直後の「と」を含めて 1 つの文節とされることが多いが、それでは 1 つの文節句中に、休止が入ってしまうため、ここでは、読点は必ず文節句境界があるとし、直後の「と」は自立語を持たない単独の文節句であるとして取り扱う。

- 文節を基本単位としては最長の単位とすることにより、その長さを高々20 モーラとすることができる（名詞連続を除く）。
- 単語や、文節において、文や呼気段落における位置情報を用いるのでなく、直前直後の相対的な情報を用いる。

これにより、ラベルが文の長さに依存しないため、ラベルの数を従来に比べて大幅に抑制することができる。また、今回は直前直後の情報のみを用いているが、音声認識で用いられている quinphone と同様に、学習データが十分にあれば、前後 2 つまで考慮しても良いと考えられる。

- アクセントを高低の 2 値のみで表現している。

アクセント句を用いていないため、アクセント型の代わりに、アクセントを H(High) と L(Low) の 2 値で表現している。副次アクセントは通常のアクセントと区別せず、その単語にアクセントがあるものとしている (1 型を除いて、1 モーラ目が L、2 モーラ目が H とする)。副次アクセントは、「ある」、「とき」などの付属語としての役割が強い語句で多くみられるが、これは、これらの語句の文中での役割が自立語と比較して小さいため、明確なアクセントとして表現されないためと考えられる。実際、強調が置かれた時にはアクセントが明確に現れる。そして、このラベルではアクセント句境界を推定する必要がなく、単語アクセントのみを推定すれば良いことを示している。例えば従来は、「東京」と「大学」がそれぞれ単語単独では 0 型のアクセントであるが、それが「東京大学」になるとき 5 型のアクセントになるとされる。しかし、これは「大学」が 1 型のアクセントに変化したと考えることもできる。このように考えることにより、「東京」にのみ強調を置くことが容易になるというメリットがある。また、0 型が連続する場合、途中の L が消失しているように聴こえることが多いが、これは、アクセント結合によるアクセントの変化ではなく、副次アクセントと同様に、アクセントが明確に現れていないだけであると考えられる。実際、ゆっくりと明瞭に読み上げる時には、アクセントが明確に表れることからも、アクセント結合とは異なる現象であると考えられる。

このように考えることで、多くのケースにおいて

Table 1　従来手法によるコンテキストラベル

先行音素
当該音素
後続音素
アクセント句内モーラ位置 (単位: モーラ)
アクセント型とモーラ位置との差 (単位: モーラ)
先行品詞 ID
先行品詞の活用形 ID
先行品詞の活用型 ID
当該品詞 ID
当該品詞の活用形 ID
当該品詞の活用型 ID
後続品詞 ID
後続品詞の活用形 ID
後続品詞の活用型 ID
先行アクセント句の長さ (単位: モーラ)
先行アクセント句のアクセント型
先行アクセント句と当該アクセント句の接続強度
先行アクセント句と当該アクセント句間のポーズの有無
当該アクセント句の長さ (単位: モーラ)
当該アクセント句のアクセント型
先行アクセント句と後続アクセント句の接続強度
当該呼気段落でのアクセント句の位置
疑問文かそうでないか
後続アクセント句の長さ (単位: モーラ)
後続アクセント句のアクセント型
後続アクセント句と当該アクセント句の接続強度
後続アクセント句と当該アクセント句間のポーズの有無
先行呼気段落の長さ (単位: モーラ)
当該呼気段落の長さ (単位: モーラ)
文中での当該呼気段落の位置
後続呼気段落の長さ (単位: モーラ)
文の長さ (単位: モーラ)

Table 2　提案手法によるコンテキストラベル

先行音素
当該音素
後続音素
先行モーラのアクセント （0:Low, 1:High）
当該モーラのアクセント （0:Low, 1:High）
後続モーラのアクセント （0:Low, 1:High）
単語内における位置の正順 (単位: モーラ)
単語内におけるモーラ位置の逆順 (単位: モーラ)
文節内におけるモーラ位置の正順 (単位: モーラ)
文節内におけるモーラ位置の逆順 (単位: モーラ)
先行単語のモーラ数
当該単語のモーラ数
後続単語のモーラ数
先行文節のモーラ数
当該文節のモーラ数
後続文節のモーラ数
先行単語の品詞 ID1
当該単語の品詞 ID1
後続単語の品詞 ID1
先行文節における自立語の品詞 ID1
当該文節における自立語の品詞 ID1
後続文節における自立語の品詞 ID1
先行単語の品詞 ID2
当該単語の品詞 ID2
後続単語の品詞 ID2
先行文節における自立語の品詞 ID2
当該文節における自立語の品詞 ID2
後続文節における自立語の品詞 ID2
単独で 1 モーラの母音であるか (0:No, 1:Yes)
当該モーラが長母音を含むか (0:No, 1:Yes)

曖昧性をなくすことができる。残る問題として、「強ければ」のように 1 型でも 2 型でも良い場合は、その可能な候補をテキストから推定し、発話速度等を加味して決定するようなシステムが必要であると考えられる。また、”形容詞”+”名詞”はアクセント結合をしてもしなくても良い場合が多く、同様にどちらにおいても対応可能にする必要がある。無論、3 単語以上の名詞連続は大きな課題である。

- 単母音、長母音を明示化している。

「… のお客 … 」と「能力」は初めの 3 音素が/noo/であり、同じ音素列になってしまうため、これを明示的に区別するラベルを加えている。尚、今回は音素ラベルに長母音を含んだものを用いていないため、”長母音を含むか”といったラベルを加えているが、勿論、音素ラベルに長母音を加えることも可能である。

その他、品詞 ID1 とは、”動詞”,”名詞”,”形容詞”,”形状詞”,”連体詞”,”副詞”,”接続詞”,”代名詞”, ”感動詞”,”助詞”,”助動詞”, ”接頭辞”,”接尾辞”,”文頭”,”休止”,”文末”であり、品詞 ID2 とは、”自立可能”,”非自立可能”,”一般”,”普通名詞”,”数詞”,”固有名詞”,”名詞的”, ”動詞的”,”形容詞的”,”形状詞的”, ”格助詞”,”準体助詞”,”副助詞”,”接続助詞”,”係助詞”,”終助詞”,”助動詞語幹”,”タリ”,”フィラー”である。これらは、Unidic[3]に基づくものであり、品詞 ID1 については、”文頭”,”休止”,”文末”を品詞として追加している。文頭、文末、文中の休止区間（ショートポーズ）を品詞扱いしておくことにより、単語や文節単位でみたときに、前後に休止があるのかどうかという情報が組み込まれている。

3　実験

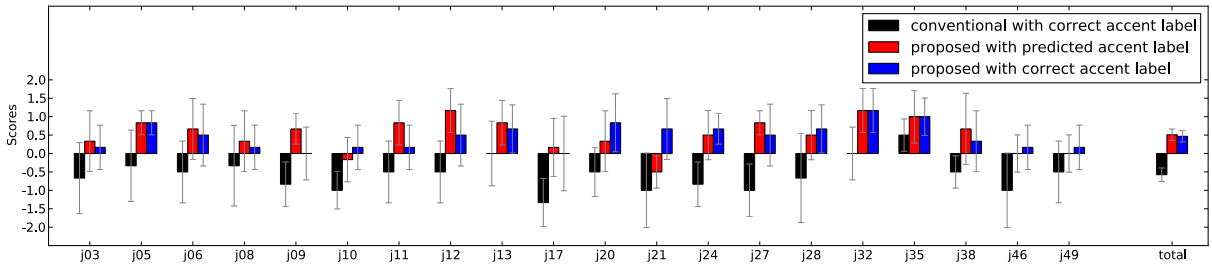
従来手法によるラベルと提案手法によるラベルのそれぞれを用いて HMM を学習し、音声を合成した。そして、主観評価実験により音声の自然性を比較した。

3.1　実験条件

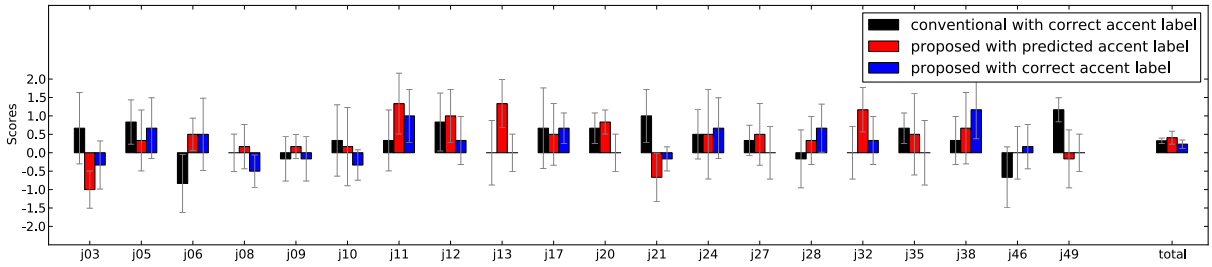
音声データは ATR 日本語音声データベース [8] の B セットの中から、男性話者 MMI と女性話者 FTY を選択した。各話者について、全 503 文のうち、サブセット A から I までの 450 文で HMM を学習し、サブセット J の 53 文を合成した。音声の分析は STRAIGHT を用いて [9]、F0、スペクトル包絡特徴量、非周期性指標を抽出した。フレーム周期は 5 [msec]、F0 は、女性話者 FTY は最小値 120 [Hz]、最大値 400 [Hz] で、男性話者 MMI は最小値 60 [Hz]、最大値 250 [Hz] でそれぞれ抽出した。HMM に用いた特徴量は、0 から 39 次元までのメルケプストラムと 0-1、1-2、2-4、4-6、6-8 [kHz] の 5 帯域の平均非周期性指標、対数 F0、およびそれらの Δ、Δ^2 を含めた 138 次元のベクトルとした。メルケプストラムと平均非周期性指標は、スペクトル包絡特徴量と非周期性指標からそれぞれ SPTK [4]を用いて求めた。HMM は HTS-2.1 を用いて構築した。状態継続長分布を明示的に含んだ 5 状態 left-to-right HSMM を用い、各状態の出力は単一の対角共分散ガウス分布とし、決定木によるコンテキストクラスタリングを行い、木の停止基準には MDL 基準を用いた。

従来手法によるコンテキストラベルは、手動抽出されたものを用いた。提案手法において、アクセントに関するラベルは鈴木らの手法によって自動推定したラベルと [6]、手動抽出されたラベルの 2 種類を用意した。形態素解析は Mecab[5]を用いているが、読み誤り、及びそれに起因すると思われるアクセント誤りについては手動で修正している。合成された音声 53 文の内、無作為に 20 文選び、それぞれについて、従来手法による合成音声と、提案手法による 2 種類の合成音声の合計 3 種類、全体で 60 文の音声を用意した。そして、音声の自然性を 6 人の被験者が主観評価した。評価は 5 段階であり、明らかに品質が良いと

[3] Unidic, http://www.tokuteicorpus.jp/dist/
[4] SPTK, http://sp-tk.sourceforge.net/
[5] Mecab, https://code.google.com/p/mecab/



(a) FTY



(b) MMI

Fig. 1　主観評価実験

評価されたときを 2 とし、品質が良いと評価されたときを 1 とし、どちらともいえないと評価されときを 0 とし、品質が悪いと評価されたときを −1 とし、明らかに品質が悪いと評価されたときを −2 とした。

3.2　結果

結果を図 1 に示す。それぞれのバーは、被験者の平均値とその 95%信頼区間を表示している。横軸は J セット 5 3 文から無作為に抽出された文番号を示しており、最後は 20 文全体のスコアの平均である。20 文全体でのスコアは、それぞれ、FTY の従来手法は -0.59 ± 0.18、提案手法（推定されたアクセントラベル）は 0.50 ± 0.16、提案手法（手動抽出したアクセントラベル）は 0.47 ± 0.15 であり、MMI の従来手法は 0.33 ± 0.07、提案手法（推定されたアクセントラベル）は 0.41 ± 0.16、提案手法（手動抽出したアクセントラベル）は 0.23 ± 0.11 であった。

これらをまとめると、MMI については有意な差がでなかったが、FTY については提案手法の方が有意に優れている結果となった。傾向としては、提案手法の方が音質が優れている傾向にあるが、一部の文においてイントネーションがうまく再現されず、大きく評価を落としていた。

4　おわりに

本稿では、HMM 音声合成におけるコンテキストラベルの新しい提案をした。そして、その有効性を聴取実験により確認した。さらに、このラベルは文の長さに依存していないため、任意の長さの文に対して、安定して音声を合成ができることが期待される。今回は、学習と評価に用いたコーパスが ATR 日本語音声データベースであるため、比較的短い文しかなかったが、より一文が長い大規模なコーパスを用いることにより、提案手法の有効性が期待される。また、定義が不明確で、自動抽出が困難なアクセント句を必要としないため、ラベルの作成コストを削減できることが期待される。多様な音声合成を実現するためには、より多くのパラメータを必要とするため、スパース性は避けて通れない問題である。そのため、音声の性質を適切に捉えた質の良いラベルは必要不可欠であり、また、そのラベルは安定して自動抽出可能なものが望ましい。今回提案したラベルを更に改良し、かつ自動抽出可能なシステムを構築することで、発話スタイルや感情などの様々な音声を高品質で合成することを目指したい。

参考文献

[1] T. Yoshimura, *et al,* Proc. EUROSPEECH, pp. 2523–2526, 1997.
[2] J. Yamagishi, *et al,* IEICE Trans. Inf. & Syst., vol. E88-D, no. 3, pp. 503–509, 2005.
[3] T. Nose, *et al,* Proc. ICASSP, pp. 833–836, 2007.
[4] Heng Lu, *et al,* Proc. INTERSPEECH, 2012.
[5] 大木 康次郎, *et al,* 電子情報通信学会技術研究報告. SP, 音声, vol.109, no. 356, pp. 141–146, 2009.
[6] 鈴木 雅之, *et al,* 日本音響学会秋季講演論文集, 2-2-12, pp. 299–302, 2012.
[7] 山本 麻美, *et al,* 電子情報通信学会技術研究報告, SP2010–109, pp. 37–42, 2011.
[8] A. Kurematsu, *et al,* Speech Communication, vol. 9, pp. 357–363, 1990.
[9] H. Kawahara, *et al,* Speech Communication, vol. 27, no. 3–4, pp. 187–207, 1999.