

HMM 音声合成におけるアクセントラベリング基準が 合成音声に与える影響の分析

高橋 遼太^{1,a)} 能勢 隆^{b)} 伊藤 彰則^{c)}

概要：本論文では、従来の HMM 音声合成において曖昧であったアクセントラベリング基準について検討を行い、合成音声への影響を調べる。具体的には、アクセント型の表現およびアクセント句境界の基準について検討する。アクセント型については、尾高型が 0 型とモーラ長型の 2 通りの表現があることに着目し、それらを用いた場合に合成音声の F0 がどのような影響を受けるかについて客観評価を行う。また、2 段階クラスタリングを用いる効果についても検証する。アクセント句境界については、アクセント句によっては 0 型と 1 型の 2 つのアクセント句で表現する場合と、それらを結合し 1 つのアクセント句として表現する場合があります。これらの違いが合成音声に与える影響を調べる。またこれらの評価において、日本語アクセントの高低の誤りを客観的指標として導入し、この指標の有効性について分析を行う。

1. はじめに

日本語音声においては、「箸」と「橋」をアクセントで区別するように、アクセントの誤りにより正しく意味が伝わらない場合が存在する。また、アクセントの誤りは音声の自然性の低下にもつながるため、音声対話システムにおいてアクセントの精度が低い場合、対話相手にとってストレスとなる可能性もある。このように、日本語テキスト音声合成において、アクセント情報を正しく表現・再現することは非常に重要である。

近年急速に利用が拡大している隠れマルコフモデル (HMM) に基づく音声合成は、統計的パラメトリック音声合成の一種であり、従来の波形接続型の方式に比べ、目標話者の音声データが比較的限られている場合でもその話者の特徴を反映した自然で不連続感の少ない合成音声を生産できるという利点がある。特に、基本周波数 (F0) については、プロのナレーターなどが発話した音声であれば、高い精度で再現できることがわかっており、波形接続型の音声合成システムの一つとして知られる XIMERA[1] においても、ターゲットの韻律特徴を予測する際に利用されている。

HMM 音声合成において、モデルを学習する際やテキストから音声を作成するには音韻や韻律情報を記述したラベルが必要であり、このラベルの精度が合成音声に影響を

及ぼす。精度の高い韻律モデルを学習するためには、学習用の音声データに対して正しいアクセントをラベリングする必要がある。この際、ラベリング基準を統一しておく必要があるが、実際には、アクセント型の表現およびアクセント句境界の決定方法に若干の曖昧性が存在するため、ラベリング基準をどのように統一すべきかについては検討が必要である。

アクセント型については、尾高型を表現する際に、0 型と N 型 (N はモーラ長) の 2 通りの表現があり、本来単語単位でのアクセント型を議論する際には、続く助詞・助動詞との組み合わせによって 0 型とモーラ長型を区別する。しかし、例えばオープンソースの音声合成ソフトウェアである Open JTalk[2] では、0 型を用いずにモーラ長型により表現している。アクセント句の決定基準については、アクセント句によっては 0 型 (またはモーラ長型) と 1 型の 2 つのアクセント句で表現する場合と、それらを結合し 1 つのアクセント句として表現する場合があります。これらの基準を統一していない場合、合成音声に悪影響を及ぼす可能性も考えられる。

本論文では、上記で述べた、アクセント型の表現およびアクセント句境界の決定基準の曖昧性に着目し、より精度の良い韻律モデルを学習するために、客観評価基準を用いて評価・分析を行う。アクセント型についてはまず 0 型とモーラ長型のどちらがより適切かについて実験を行う。0 型を用いた場合には、HMM の学習における状態のクラスタリングにおいて 0 型と 1 型が同じ状態として共有されてしまう問題がある。これについては 2 段階クラスタリング

¹ 東北大学 大学院工学研究科
Graduate School of Engineering, Tohoku University

a) ryota.takahashi.s7@dc.tohoku.ac.jp

b) tnose@m.tohoku.ac.jp

c) aito@spcom.ecei.tohoku.ac.jp

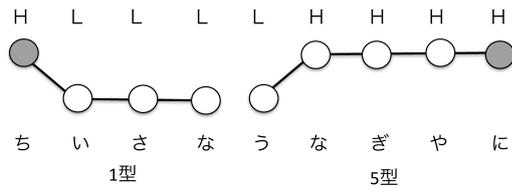


図 1 日本語におけるアクセント句とアクセント型
Fig. 1 Accent phrase and accent type in Japanese

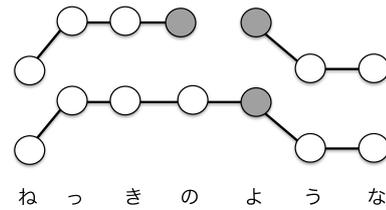


図 2 アクセント句境界の曖昧性
Fig. 2 Ambiguity of accent phrase boundary

を用い、その効果を検証する。アクセント句境界については、0型と1型を連結した場合と、個別に扱った場合について、F0の再現性がどのように変化するかを調べる。これらの評価には客観的指標として原音声と合成音声の対数F0のRMS誤差を用いるが、これのみではアクセントの精度を直接的に評価することは困難である。そのため、日本語アクセントの高低の誤りについての新たな客観的指標を導入し、この指標の妥当性・有効性について評価・分析を行う。

2. 日本語のアクセント

2.1 アクセント句とアクセント型

日本語単語アクセントは、複雑な時間変化を占めるピッチパターンを扱うことなく、点ピッチパターンに十分その性質を表現できることが明らかとなっている [3]。発話中には文法的、意味的なまとまりとして、1つあるいは複数の単語を連結したまとまりにアクセントが1つ付く傾向があり、これをアクセント句と呼ぶ。各アクセント句にはアクセント型が定義され、東京方言におけるアクセント型はアクセント句内のアクセント核 (語のアクセントが下がる箇所) の場所として定義され、図1で見ると色塗りの部分が該当する [4]。ここでアクセントがない、すなわちピッチが高から低に変わる箇所がないアクセント句 (例えば図1の「うなぎやに」) の場合、0型あるいはモーラ長型として表現することが可能である*1。通常、モーラ長型は後に来る助詞・助動詞を含めない場合の表現であり、これらも含めた場合の平板型は0型として表現されることが多いが、Open JTalk で使われている解析器では0型を用いず、モーラ長型で統一されている。

アクセント句については、0型 (あるいはモーラ長型) と1型を一つにまとめて単一のアクセント句として表現することもできるため、アクセント句境界についても曖昧性が存在する (図2)。

2.2 アクセント精度の客観的指標

HMM 音声合成において、生成されたF0系列の再現性を測る基準として、原音声と合成音声間の対数F0のRMS

*1 アクセント辞書などにおける日本語の単語のアクセント型としては、後続モーラが「高」となる「平板型」を0型と、「低」となる尾高型をモーラ長型とし使い分けられて定義されている。 [5]

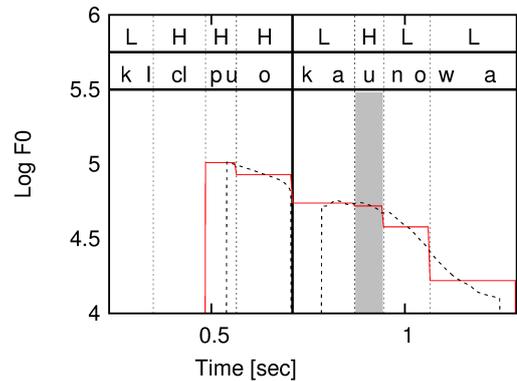


図 3 対数 F0 系列と高低誤り
Fig. 3 HL error with Logarithm F0 series

誤差がよく用いられる。これを用いることにより、アクセントの精度についてもある程度予測することが可能であるが、誤差が大きい場合に必ずしもアクセントが誤っているとは限らない。F0としては誤差が大きいですが、実際のアクセントとしては正しいアクセントで生成されているということも有り得る。このように、RMS 誤差のみでアクセントの精度について客観的に評価するには限界があるため、新しい指標が必要となる。本研究では、アクセントによる相対的なピッチの高低に対しラベルの高低が致命的に誤った場合を考慮して、図3に示すような「HL 誤り」を導入する。図では「切符を買うのは」という音声に対し、「切符を (0型)」「買うのは (2型)」というアクセント情報がラベリングされている。図には音声の対数F0系列と、そのモーラ毎の平均値を併せて示している。高を「H」、低を「L」を表すと、本来「かう」の部分はラベルでは「LH」となっているのに対し、対数F0の平均値は「か」から「う」で減少している。このような場合はアクセントを結合し「か」を「H」とする「切符を買うのは (6型)」とすべきと考えられる。このように、HL 誤りにより音声とラベルとの不一致を調べることができる。本研究ではこれをHL 誤りとして客観評価指標として扱う。測定方法としては、前後のモーラとの平均F0の高低を調べ、それがアクセント型が示す高低と比較し、一致しない場合1とカウントする。「HH」や「LL」のようにアクセントとして高低が変化しない箇所についてはこの判定は行わない。

3. HMM 音声合成におけるアクセントラベリング

HMM 音声合成に限らず，コーパスベースの音声合成方式ではあらかじめ収録した音声に対し，音素などの音韻情報やアクセントなどの韻律情報をラベルとして付与する必要がある．このうち，音韻情報は，テキスト解析により自動的にラベル付けすることができる．一方でアクセントについては，同じテキストであっても必ずしも同じアクセントで発話されるとは限らないため，正しいアクセントを付与するには音声を耳で確認して手作業により修正を行う必要があるが，これには大きな人的コストが必要となる．一方で，アクセント情報を自動で付与する研究も進められているが，未だ人手によるラベルを超える結果は得られていない [6]．さらに，既に述べたとおり，アクセント型やアクセント句境界の決定については文字情報のみから完全に決定することは難しく，現行のモデルではアクセント結合規則という規則で分割されることが多い [7][8][9]．加えて明確な規則があるわけではないため，この曖昧性により合成音声における F0 の再現性，およびアクセントの精度がどのように変化するかについてはこれまで十分な検討は行われていない．

4. アクセント型表現の統一

4.1 同一アクセントに対する異なるアクセント型表現

2.1 節で述べたように，通常 0 型として表現されるアクセント句は Open JTalk ではモーラ長型で表現されている．モーラ長型ではアクセント句の長さ（モーラ長）に応じてアクセント型のバリエーションが増えるのに対し，0 型ではそれらを一つのアクセント型として表現するため，特に学習データが少ない場合などにおいて精度よく F0 をモデル化できると考えられる．6 節の実験で用いた男性話者 MHT の 503 文において，0 型およびモーラ長型を用いた場合のアクセント型の頻度分布を図 4 に示す．図から，0 型については学習サンプル数が比較的多いのに対し，モーラ長型を用いた場合にはこれらのサンプルは各モーラ数に分配されていることが確認できる．一方，欠点としては，0 型と 1 型は数としては隣り合っているが，モーラ毎の高低は正反対である．HMM 音声合成では決定木などを用いて状態共有を行うため，クラスタリングの仕方によっては 0 型と 1 型が同じ状態として共有される場合がある．したがって，この共有を避けるためあらかじめ 2 段階クラスタリングを用いて木を分けておく必要があるが，このような明示的な木の分割を用いた場合，制約を用いない場合に比べ学習に影響がでてしまう可能性がある．に一方で，モーラ長型を使った場合には，同じアクセント句長の場合，N - 1 型と N 型では高低は一つしか異なるため，上記のような問題は起こらない．

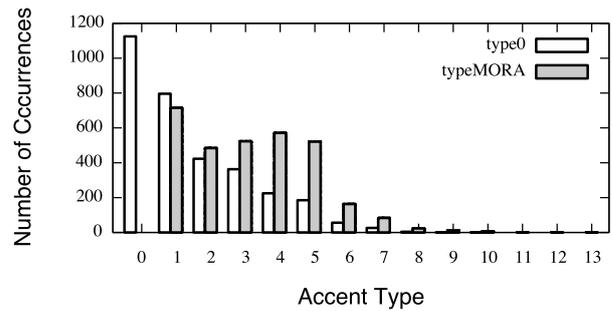


図 4 0 型統一とモーラ長型統一におけるアクセント型分布
Fig. 4 The accent type distribution in type0 and type of Mora length unification

表 1 0 型と 1 型が混在したリーフノードの割合
Table 1 Percentage of leaf node type 0 and type 1 are mixed

話者	MHT	MMY	FKS	FYM	MSH
リーフノード数	1864	1694	1589	1800	1544
混在したノード数	556	579	749	655	654
割合 (%)	29.8	34.2	47.1	36.4	42.4

4.2 0 型表現におけるアクセント精度の低下

4.1 節で述べたように，アクセント型の表現として 0 型を用いた場合には，HMM の学習時に 0 型と 1 型が決定木の同じノードに振り分けられ，結果としてモデルにおけるアクセントの精度が低下するという問題がある．

表 1 はアクセント型として 0 型を用いた場合について，6 の実験で用いた 5 名の話者のモデルの学習時に構築された決定木のリーフノードにおいて，どの程度 0 型と 1 型が混在しているかをパーセンテージで示している．表より，いずれの話者についてもリーフノードにおいて 0 型と 1 型が一定数混在していることが確認できる．この問題を解決するために，次節で 2 段階クラスタリングを導入する．

4.3 2 段階クラスタリングによる HL 誤りの軽減

アクセント型として 0 型を用いた場合には，アクセント型のバリエーションを抑えられるという利点がある反面，4.2 節で述べたように，0 型と 1 型が学習の過程で同じモデルとして学習され，これは本来高低の異なるアクセントであるので合成の際に高低に悪影響を及ぼすと考えられる．

これを解消するため，2012 年に三井らによって提案された 2 段階クラスタリング [10] を採用する．これは従来のクラスタリングを 2 回に分けて行うというものであり，今回は最初の質問を「0 型か，1 型か，2 型以上か」の三択にして振り分け，0 型と 1 型を同じ状態として共有されないようにする．しかし 0 型と 1 型がそれぞれ単独で学習されるため，学習効率が悪くなるのではないかと懸念がある．そのことから，学習文章数を順に増やしていき推移を見る必要がある．

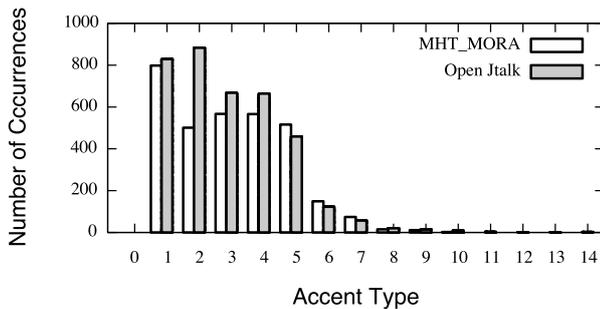


図 5 アクセント型の分布の比較

Fig. 5 Comparison of the distribution of accent type

5. アクセント句境界基準の統一

現在用いられているアクセント句には曖昧さが存在する。ATR503 文における手動ラベルと Open JTalk の解析器でつけた自動ラベルのアクセント型の分布を 5 に示す。図より Open JTalk では 2 型, 3 型が多く見られることからラベリングにおいてアクセント句の決め方になんらかの基準の違いが存在すると考え、その違いと影響を検討する。

5.1 アクセント句境界の曖昧性

「熱気のような」というフレーズに対し、これを一つのアクセント句として学習することが可能であるが、「熱気」と「ような」というように 2 つのアクセント句として分けて考えることも可能である。現状ではどちらが適切かは決められておらず、混在している。手動でラベリングする際には長いフレーズが見られ、Open JTalk の自動ラベリングにおけるアクセント結合規則では比較的短い句で決められていることが確認できた。このような現象が考えられるのは 0 型と 1 型が並んでいるときにのみ考えられる。

5.2 アクセント句境界の統一

統一基準としては、手動ラベリングを行ったアクセント句に対して、Open JTalk の結合基準において 0 型と 1 型に分離されている箇所を分割する。これにより手動によるアクセントの精度を維持したまま分割基準に適應させることが可能である。また、0 型と 1 型が並んでいるときに必ずしも手動ラベルで結合されている訳ではない。そのため手動ラベルにおいて同様の並びがあった場合にそれらを全て結合した際の影響を考える必要がある。以上のことから、分割基準、従来手動基準、結合基準の 3 通りについて検証を行う。

6. 実験

6.1 実験条件

実験には ATR 日本語音声データベースセット B の話者

5 名による音素バランス文 503 文を用いた。話者は MHT, MMY, MSH, FKS, FYM を対象とした。サブセット J の 53 文を評価用とし、残りの 450 文を対象に学習文を用いた。音声信号を 16kHz でサンプリングしフレーム周期 5ms で音声特徴量を抽出した。スペクトル特徴量として STRAIGHT 分析により得られた平滑化スペクトルから求めた 0 から 39 次までのメルケプストラムを、音源特徴量として対数 F0 及び 5 次元の非周期性指標を用い、それらの一次及び二次の動的特徴量を加えた 138 次元の特徴量ベクトルを使用した。音響モデルとして 5 状態の left-to-right HSMM を用い、出力分布の共分散行列は対角を仮定した。評価には 2 つの客観評価に基づいて行い、一つは発話話者の実際の音声との距離をとる RMS 誤差 [cent] と、もう一つは HL 誤りカウントを用いる。

6.2 アクセント型表現の比較

以下の 3 通りの基準で学習・合成を行って比較をする。

- (1) type0 0 型に統一したもの
- (2) typeMORA モーラ長型に統一したもの
- (3) type0 0 型に統一して 2 段階クラスタリングを行ったもの

評価としては話者ごとに評価値の平均をとり、それを全話者について平均した。横軸を学習文章数、縦軸を評価値とした結果を図 6 に示す。RMS 誤差では学習文章数が比較的少ない範囲では 0 型に統一したものが優位性が見られるが、文章数を増やすとその差は殆ど見られなくなることがわかる。一方で HL 誤りについては学習文章数を増やしても 0 型においては相対的に多く検出されたままである。2 段階クラスタリングを行うことでモーラ長型同様 HL 誤りが減少していることから、クラスタリングにおいて 0 型と 1 型が同じリーフノードに入ることによってアクセントの高低に悪影響を及ぼしていたことが示された。これらのことから、2 段階クラスタリングを行わないのであればモーラ長型で扱うのが適切だと言える。

6.3 アクセント句境界基準の比較

第 5 節で述べたような曖昧さのあるアクセント句に対して、以下の 3 通りの手法で学習・合成を行って比較する。

- (1) conventional 手動による従来基準のアクセント句境界
- (2) separate Open JTalk 基準に基づく分割基準
- (3) combine 0 型と 1 型の繋がりを全て一つのアクセント句とみなす結合基準

分割は Open JTalk で用いている自動ラベリングには誤りがあることを考慮し手作業で行った。話者は MHT と FKS の 2 名に対して行い、50, 100, 200, 400 文になるようデータセットを全 400 文から重複なく全ての文を選択するよう順に 8, 4, 2, 1 通り選択し、それぞれの平均をとっ

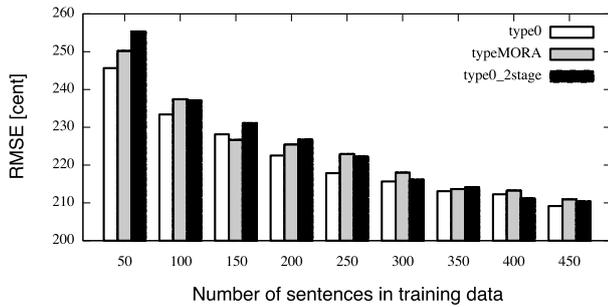


図 6 アクセント型表現の比較結果 (F0 歪み)

Fig. 6 Comparison result of accent type representation (F0 distortion)

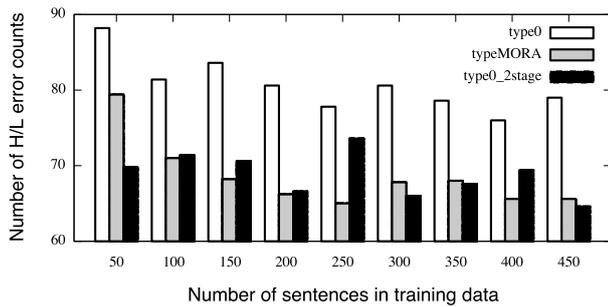


図 7 アクセント型表現の比較結果 (HL 誤り)

Fig. 7 Comparison result of accent type representation (HL error)

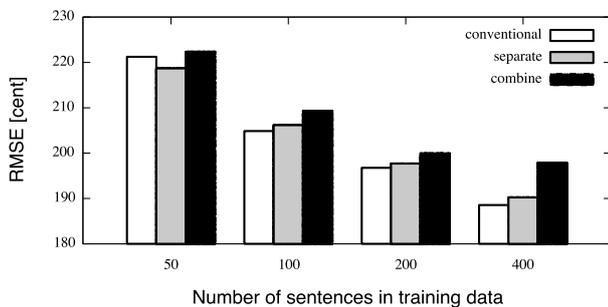


図 8 アクセント句境界基準の比較結果 (F0 歪み)

Fig. 8 Comparison result of the accent phrase boundary criteria (F0 distortion)

た．結果を図 8 に示す．結合を行った基準においてやや大きく RMS 誤差が生じていることがわかる．原因としては文脈を無視した結合規則により学習における話者性が損なわれたのが考えられる．一方で HL 誤りの方では結合基準，分割基準の順に誤り数が少ない．結果として，どちらかの基準に統一することはアクセント通りの韻律で音声合成できることがわかった．以上から，HL 誤りが減少し RMS 誤差が比較的少ない分割基準，即ち短いアクセント句を選択した方が良いと言える．

7. アクセント精度の客観評価指標の分析

本研究で用いた HL 誤りでは，聴覚的に誤りだと感じる

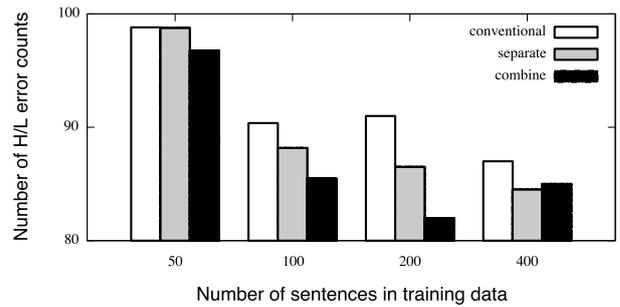


図 9 アクセント句境界基準の比較 (HL 誤り)

Fig. 9 Comparison result of the accent phrase boundary criteria (HL error)

もの以外もカウントしてしまい，この HL 誤りを実際のアクセント通りに読み上げた自然音声に適応させても誤りの数は 0 にならない．従って，これらの誤りパターンについて分析を行う必要がある．聴覚的な誤りではないと考えられるものを 3 通りに分類し，以下の節で説明する．

- (1) 句境界を跨ぐケース
- (2) F0 平均の差が微小なケース
- (3) F0 のピークずれのケース

7.1 HL 誤りの分類：1. 句境界を跨ぐケース

一つ目はアクセント句境界を跨ぐときの誤りである．アクセント句境界を決める際に H と L が隣接することがあるが，実際の音声では高低に大きな差を付けずに発声することが往々にしてある．そのため，聴覚的に誤りと認識しないがラベルの HL とは一致しないという問題である．図 10 では「頭を」と「下げた」の間にアクセント句境界があるが，このラベルを「頭を下げた (2 型)」として結合したラベルに修正する必要があると考えられる．

7.2 HL 誤りの分類：2. F0 平均の差が微小なケース

二つ目は隣接する F0 平均の差が微小な場合である．図 11 の「崩れてしまう」のように文脈において抑揚の少ないフレーズである際に，F0 によっては聴覚上よりも誤りとして多くカウントされてしまうことがある．そのためこのように前後との差が微小であるものに対しては閾値を設けるような形でカウントしないなどの対応が必要がある．

7.3 HL 誤りの分類：3. F0 のピークずれのケース

最後は F0 のピークがずれているケースである．HL 誤りは連続した F0 の平均をとっているのので，聴覚的に影響のない範囲で F0 のピークがずれることで HL が変わることがある．図 12 のように聴覚上では「じ」にアクセント核があるように聞こえるが，F0 の平均をとると「ん」が高い音だと判定されアクセントの核がずれたような形になる．従って今回用いた HL 誤りを主観評価基準に近づけるのであれば，聴覚的にこの F0 ピークのズレが同じ 1 モー

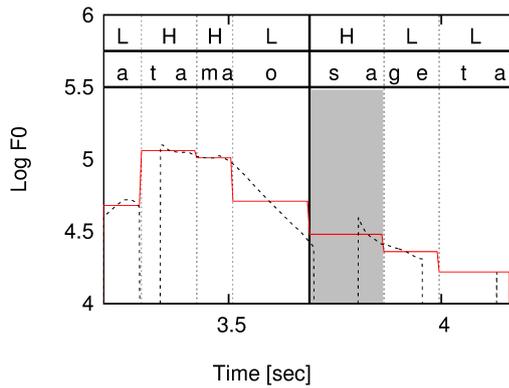


図 10 HL 誤りの分類: 1. 句境界を跨ぐケース

Fig. 10 HL error classification: Case.1 Straddles the clause boundary

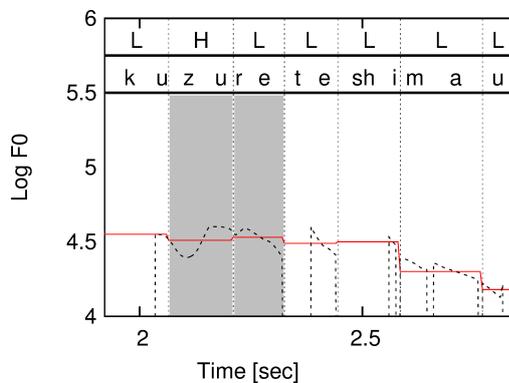


図 11 HL 誤りの分類: 2. F0 平均の差が微小なケース

Fig. 11 HL error classification: Case.2 Difference between the average F0 is a little

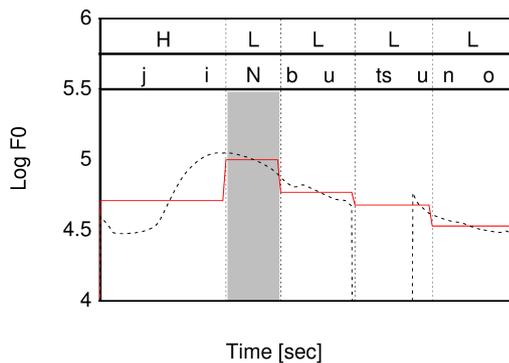


図 12 HL 誤りの分類: 3. F0 のピークずれのケース

Fig. 12 HL error classification: Case.3 Peak of F0 shifts

ラの範囲でもどこまで許容されるのかも考えないと正確な判断はできない。以上で課題に挙げた点を客観評価の改善のため今後も検討を行う。

8. おわりに

本稿では、HMM 音声合成におけるアクセントラベリングの基準が日本語合成音声の品質に与える影響を検証し、

アクセントラベルの基準における 2 つの曖昧さについて解消した。実験結果から、

- (1) 0 型よりモーラ長型のほうが良い
- (2) Open JTalk で決定されるアクセント句^{*2}を用いることで他の基準による手法よりもアクセントの誤りを抑えつつ F0 の再現性を高めることができる

以上のことが判明した。また、クラスタリングにおいて 0 型と 1 型が同じ状態として共有されてしまうことが HL 誤りを生じさせること、アクセント句の定め方により HL 誤りの少ない音声の合成が出来ることが確認できた。しかし、本結果では RMS 誤差で測定している話者性と HL 誤りの結果はトレードオフの関係にある、今よりも自然な音声を作成するためには両方の評価を改善した音声を作ることが求められるので、今後も引き続きこの側面から検討を行う必要がある。また、HL 誤りに関しては単なる評価基準としてだけでなく現在手動で行っている学習ラベルを自動化するような指標になると考えられるので機械学習などを踏まえ今後の課題とする。

参考文献

- [1] 河井, 戸田, 山岸, 平井, 倪, 西澤, 津崎, 徳田, : “大規模コーパスを用いた音声合成システム XIMERA”. 電子情報通信学会論文誌 D, Vol.J89-D, No.12, pp.2688-2698. (2006)
- [2] 入手先 (<http://open-jtalk.sp.nitech.ac.jp/>)
- [3] 橋本 新一郎: “日本語単語アクセントの諸性質”, 電子通信学会論文誌 D 56(11), 654-661, 1973-11. (1973)
- [4] NHK 放送文化研究所 (編): “NHK 日本語アクセント辞典”, NHK 出版, (1998).
- [5] 斎藤純男, “日本語音声学入門 改訂版”, 三省堂. (1997)
- [6] 大西浩之, 能勢隆, 郡山智樹, 小林隆夫: “HMM 音声合成における正規化学習を用いたアクセント誤り削減の検討”. 日本音響学会 2014 年春季研究発表会講演論文集, 1-R5-16, pp.411-412. (2014)
- [7] 匂坂, 佐藤: “日本語単語連鎖のアクセント規則” 信学論 (D) vol. J66-D, no. 7, pp. 849-856. (1983)
- [8] 宮崎: “単語間の意味的結合関係を用いた複合アクセント句の自動抽出法”, 信学論 (D) vol. J68-D, no. 1, pp. 25-32. (1985)
- [9] 喜多, 峯松, 広瀬: “日本語テキスト音声合成を目的としたアクセント結合規則の構築と改良”, 信学技報 SP-102(108), 13-18, 2002-05-24. (2002)
- [10] 三井康行, 近藤玲史, 加藤正徳: “二段階クラスタリングを用いた HMM に基づく韻律生成”. 信学技報 IEICE Technical Report SP2012-80 (2012)
- [11] 入手先 (<http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>)

*2 MeCab とアクセント辞書を用いてアクセント結合して決定される