論文 / 著書情報
Article / Book Information

論題(和文)	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クストの拡張
Title(English)	Extension of context set for generating diverse prosodic variations in HMM-based spontaneous conversational speech synthesis
著者(和文)	郡山知樹, 能勢 隆, 小林 隆夫
Authors(English)	Tomoki Koriyama, Takashi Nose, Takao Kobayashi
出典(和文)	電子情報通信学会論文誌, Vol. J95-D, No. 3, pp. 597-607
Citation(English)	, Vol. J95-D, No. 3, pp. 597-607
発行日 / Pub. date	2012, 3
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2012 Institute of Electronics, Information and Communication Engineers.

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クストの拡張

郡山　知樹[†a)]　　能勢　隆[†]　　小林　隆夫[†]

Extension of Context Set for Generating Diverse Prosodic Variations in HMM-Based Spontaneous Conversational Speech Synthesis

Tomoki KORIYAMA[†a)], Takashi NOSE[†], and Takao KOBAYASHI[†]

あらまし　本論文では自発性の高い対話音声の合成において，多様な韻律を生成するための拡張コンテクストを提案する．HMM音声合成では音韻・韻律の変動要因をコンテクストとして考慮し学習・合成を行っているが，従来の読上げ音声のためのコンテクストセットでは対話音声の韻律の多様性を実現することが困難である．そこで，本論文では大規模音声コーパスである日本語話し言葉コーパス (CSJ) に収録されている対話音声を対象とし，CSJに付与されている様々な情報をコンテクストとして追加し拡張コンテクストとした．コンテクストの増加による過学習を避けるための決定木クラスタリングの新たな停止基準を導入し，従来のコンテクストと拡張コンテクストの比較を行った．その結果音素引き延ばし及びX-JToBIのトーン層ラベルに基づく情報がコンテクストとして有効であった．更に実用性を考慮して，合成時に拡張コンテクストの一部を自動的に求める手法の有効性を評価し，正解のコンテクストを用いた場合と同程度の自然性が得られることを確認した．

キーワード　対話音声，話し言葉音声，HMM音声合成，韻律コンテクスト，日本語話し言葉コーパス

1. まえがき

音声合成技術の進歩に伴い，公共交通機関の音声案内やカーナビゲーションシステム，映像コンテンツなど日常の中でコンピュータによって合成された音声を耳にする機会が増えている．音声合成のさらなる応用として，ヒューマン・マシン間のコミュニケーションへの適用が期待されているが，そのためには読上げ調の音声だけでなく，実際に人間と対話しているかのような話し言葉調の音声も生成可能なシステムが望ましい．しかし，自発性の高い話し言葉音声は読上げ調の音声に比べ韻律の変動が大きく，従来の読上げ調の音声合成手法をそのまま適用するだけでは十分な品質を得ることは難しい．

対話音声合成に対する取組みとして，非常に長い時間収録された対話音声コーパスが利用可能であれば，素片選択型の音声合成手法を用いることで自然な音声を合成することができることが示されている [1]．また，比較的少量の音声コーパスを用いて話し言葉音声を合成する手法として．HMM音声合成 [2] に基づく手法がいくつか提案されている [3]～[6]．文献 [3] では，日本語話し言葉コーパス (CSJ) [7] を対象とし，話し言葉音声の基本周波数 (F0) 曲線と音素継続長といった韻律情報を数量化I類によってモデル化 [8] している．文献 [4] ではF0を多空間確率分布HMM (MSD-HMM) [9] でモデル化することで話し言葉音声の韻律を表現している．このとき，対話音声と読上げ音声との混合データを用いることで合成音声の自然性を維持したまま，話し言葉らしさを再現できることが報告されている．一方で，読上げ音声と話し言葉音声のパラレルデータを学習データとして，読上げ調のモデルのスペクトル及び継続長を変換することにより話し言葉らしさを再現する手法も提案されている [5]．また，我々もこれまでに任意の話者の対話音声データの量を減らすための平均声に基づく手法が有効であることを示した [6]．

† 東京工業大学大学院総合理工学研究科，横浜市
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama-shi, 226–8502 Japan
a) E-mail: koriyama.t.aa@m.titech.ac.jp

これらの手法によって合成音声の自然性が向上しているとはいえ，実際に人間が発声した音声との差異は依然として大きい．その原因の一つとして韻律の多様性が十分に実現されていないことが挙げられる．HMM音声合成では音韻・韻律の変動要因を考慮したコンテクスト依存モデルを学習し合成を行っているが，従来の読上げ調の音声を前提としたコンテクストだけでは対話音声の多様性を表現することが困難である．一方で，複数の感情や発話様式といったスタイルを発話全体のコンテクストとして学習したスタイル混合モデルを用いることで，多様なスタイルの合成音声を生成可能である [10]．また，発話中の強調されている区間にコンテクストを加えることにより，部分的に強調された発話が合成可能であることも報告されている [11]．同様にして，発話内の対話音声特有の表現に対しても適切なコンテクストを使用することにより多様な表現を生成できると考えられ，自動でのラベリングが可能な発話単位の音響特徴量に基づくコンテクストを使用する手法を検討した [12] が，十分ではなかった．

そこで，本論文では対話音声合成の韻律の多様化の実現を目的とし，HMM音声合成の枠組みにおいて対話音声合成のための新たな音韻・韻律コンテクストを提案する．具体的には，CSJに収録されている対話音声を対象とし，CSJのアノテーションデータをもとに新たなコンテクストを導入する．これにより，対話音声合成において具体的にどのようなコンテクストを使用すべきかについて検討を行う．また，提案法ではコンテクストの追加に伴いモデルの過学習の影響が無視できないため，これを緩和するために決定木に基づくコンテクストクラスタリングにおいて従来使用されてきたMDL基準 [13] に加え，新たな分割停止基準を導入しその有効性を評価する．更に，合成時の実用性を考慮し，新たなコンテクストの一部を自動的に予測した場合についても検討する．

2. 対話音声のための拡張コンテクスト

2.1 ベースラインコンテクスト

HMM音声合成において，コンテクストはスペクトル，F0，継続長といった音響的特徴を表す変動要因として音素単位に付与され，コンテクストの組合せに対しコンテクスト依存HMMが学習される．通常の読上げ音声合成では以下のコンテクストが主に使用されており [2]，ここではこれらをBASELINEコンテクストと呼ぶことにする．

- 音素

{ 先行，当該，後続 } 音素の種類

- モーラ

アクセント句内のモーラ位置

- アクセント

{ 先行，当該，後続 } アクセント句内の長さ，アクセント型，位置，前後のポーズの有無

- 呼気段落

{ 先行，当該，後続 } 呼気段落の長さ，位置

- 発話長

2.2 拡張コンテクスト

対話音声の変動要因を読上げ音声のためのBASELINEコンテクストのみで表すことは困難なため，本論文では表1のF〜Lに示す七つのコンテクストカテゴリーを追加し，その有効性を検討する．これらのコンテクストはCSJのコアデータに収録されているラベルから自動的に決定されるものであり，ベースラインコンテクストと同様音素単位に付与される．その詳細を以下に示す．

- 音素引き延ばし

音素引き延ばしとは，母音あるいは子音が通常より長い時間発話される現象のことであり，考えながら話しているときや驚いたとき，強調するときなどに現れる．書き起こしテキスト上では「スゴ <H> イ」「カイ <Q> セキ」のように表記される．ここで <H> と <Q> はそれぞれ母音・子音の引き延ばしであり，つまり「スゴーイ」「カイッセキ」のように発話される．実際の音素引き延ばしの例を図1に示す．音素引き延ばしを伴う図中の (a) の発話では「駅（エキ）」の「i」が，(b) に比べ非常に長く発音されている．

- 発話スタイル

CSJでは書き起こしテキストに特定の発話スタイルによって発話された区間にラベルが付けられている．本論文では笑いながら，ささやきながらの発話や不明瞭な発話を発話スタイルのコンテクストとして使用

表1 コンテクストカテゴリー
Table 1 Context categories.

BASELINE	ADDITIONAL
A 音素	F 音素引き延ばし
B モーラ	G 発話スタイル
C アクセント	H トーンラベル
D 呼気段落	I 非流暢性
E 発話長	J 音素付加情報
	K 単語単位
	L 節

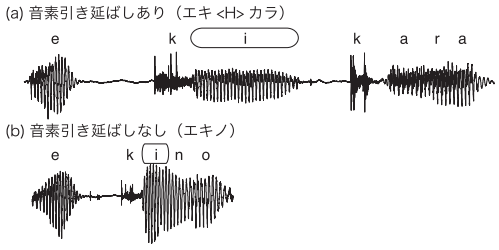


図 1　音素引き延ばしの例
Fig. 1　Example of phone prolongation.

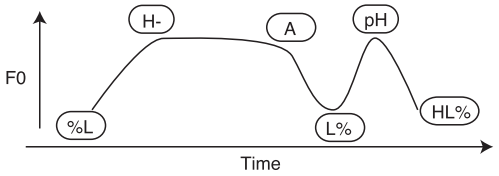


図 2　上昇下降調のアクセント句の F0 曲線とそれに対応する X-JToBI トーン層ラベルの概略
Fig. 2　Schematic example of relationship between X-JToBI tone tier labels and the F0 contour of an accent phrase which ends with rise-fall.

する．

- トーンラベル

日本語のアクセント句は複数の語から構成され，アクセント句内の F0 の相対的な変動は 0 型，1 型などのアクセント型によって表される．読上げ調の音声ではアクセント型によって F0 曲線の概形を表すことが可能であるが，話し言葉では F0 の変化がより複雑になり，アクセント型のみでは F0 曲線を表すことは困難となる．その例の一つに上昇調や上昇下降調，下降上昇調といった句末境界音調がある．同じ語彙であっても句末境界音調を変化させることで，問いかけや確認など様々な発話意図を表現することができる．この句末境界音調を含む複雑な F0 曲線を記述するために，CSJ では ToBI を日本語の話し言葉音声のために拡張したラベリングスキームである X-JToBI [14] を用いている．X-JToBI のトーン層のラベルは，図 2 の上昇下降調のアクセント句に示すように F0 曲線の変曲点に対して付与される．

本論文では表 2 に示すトーン層のラベルに基づいた情報をコンテクストとして使用する．具体的には，句末境界音調，ラベル「FL」「FH」で表される韻律的フィラー，アクセント句が句頭・句末ラベルで正常に開始・終了しない韻律的語断片，といった F0 変動の種類をアクセント句のコンテクストとする．また．各

表 2　トーン層ラベルの一覧
Table 2　List of tone tier labels.

ラベル	概要
%L	句頭境界．F0 上昇の開始
H−	上昇の終端
A	下降の開始．アクセント
L%	下降の終了．下降調の終端
H%	上昇調
LH%	下降上昇調
HL%	上昇下降調
HLH%	上昇下降上昇調
pL	ローポインタ．LH%，HLH%に伴う
pH	ハイポインタ．HL%，HLH%に伴う
FL	F0 が低く変化の少ないフィラー句
FH	F0 が高く変化の少ないフィラー句

トーン層ラベルの付与されているモーラに対する当該モーラの相対的なモーラ位置をモーラのコンテクストとして使用する．

- 非流暢性

自発性の高い音声には読上げ音声には現れない非流暢成分が含まれる．CSJ では，言語上意味をほとんどもたないフィラー，言いよどみによる語断片，助詞の言い直しの 3 種類の非流暢成分にラベルが付与されており，これらの成分をアクセント句のコンテクストとして使用する．

- 音素付加情報

CSJ では音素情報が詳細に付与されており，一般的な音素セットには分類されない発音に対してもタグが付与されている．ここでは，出現頻度の比較的高い 2 種類のタグに注目する．具体的には各音素に対して，タグ <sv> で示される母音後の声帯振動や，タグ <cl> で示される子音前の無音区間が含まれるかどうかの真偽をコンテクストとして使用する．

- 単語単位

対話音声にはその自発性の高さゆえに語の融合や省略，音便といった語彙に関する特有の現象が観察される．そのような現象や品詞などの形態素は CSJ では短単位，長単位という 2 種類の単語に対して付与されている．短単位単語は日本語の辞書の語彙とおおよそ一致する語であり，長単位単語はいくつかの短単位単語の複合語である．本論文では長単位，短単位の両単位の単語の情報をコンテクストとして使用する．

- 節

節は主語と述語からなる文法上の単位であり，その境界は CSJ の書き起こしデータから自動的に決定される．本論文では節の種類と節中のモーラ位置をコン

テクストとして使用する．

3. コンテクスト決定木の停止基準

HMM 音声合成では，モデルパラメータの推定精度を上げ，また学習データに含まれないコンテクストのパラメータを予測するために音韻・韻律コンテクストに基づく決定木クラスタリングが行われる [2]．分割前のモデル λ に対し，リーフノード S_{m_p} を S_{m_y} と S_{m_n} に分割したモデルを λ' とすると，ゆう度の変化量 $\Delta\mathcal{L}$ を式 (1) で表すことができる [15]．

$$\begin{aligned}\Delta\mathcal{L} =& \mathcal{L}(\lambda') - \mathcal{L}(\lambda)\\ =& \frac{1}{2}(\Gamma_{m_y}\log|\Sigma_{m_y}| + \Gamma_{m_n}\log|\Sigma_{m_n}| \\ & - \Gamma_{m_p}\log|\Sigma_{m_p}|)\end{aligned} \tag{1}$$

ただし，Γ_m，Σ_m はそれぞれノード S_m における状態占有頻度 (state occupation count) と出力分布の共分散行列を表す．状態占有頻度 Γ_m は，リーフノードに含まれる状態の占有確率の総和で求められる．

$$\Gamma_m = \sum_{c\in C(m)}\sum_{n=1}^{N}\sum_{t=1}^{T_n}\gamma_t^n(c) \tag{2}$$

ここで，$\gamma_t^n(c)$ は n 番目のサンプルの t フレーム目における状態 c の占有確率であり．$C(m)$ はリーフノード S_m に含まれる状態の集合を，N は学習に使用されたサンプルの総数，T_n は n 番目のサンプルの総フレーム数をそれぞれ表す．クラスタリングの際には $\Delta\mathcal{L}$ が最大となるような質問によってリーフノードが分割される．このとき，決定木の分割停止基準には，モデルパラメータ数と学習データの量から決定される最小記述長 (MDL) が有効であることが報告されている [13]．

しかし，**2.** で説明した拡張コンテクストを対話音声合成に用いる場合，コンテクストの種類が増えることで，学習データ不足による過学習に陥り，MDL 基準が有効に作用しない場合が生じる．CSJ に収録されている女性話者 (ID=19) の 22.5 分程度の対話音声を用いてモデルを学習した場合の，リーフノード内の各コンテクストに対して割り当てられる学習データ数の総和（以下観測サンプル数）の分布を図 3 に示す．図から，観測サンプル数が数個しかないリーフノードが大量に存在しており，これらのノードでは過学習が起きている可能性が高いことが分かる．

この過学習の問題を解決するため，各リーフノードの最小状態占有頻度と最小観測サンプル数 (minimum number of observations) の 2 種類の停止基準を導入し MDL 基準と併用する．最小状態占有頻度はリーフノードを構成する観測フレームの総数を制限するパラメータである [15]．しかし，対話音声には音素引き延ばしなど長い音素が多く現れるので，一つのリーフノードが非常に長い数個のサンプルで構成されてしまう可能性がある．そのような場合には，リーフノードで観測されるサンプルの総数を制限する最小観測サンプル数を用いた方がより効果的であると考えられる．

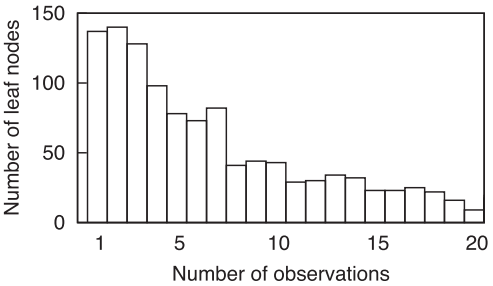


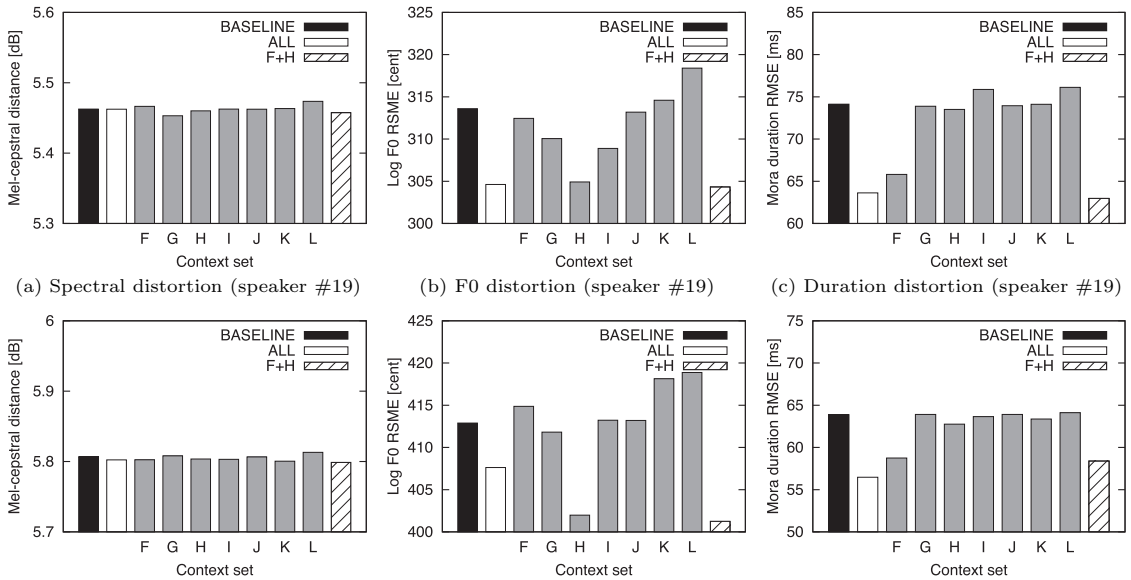
図 3 各リーフノードに含まれる観測サンプル数のヒストグラム
Fig. 3 Histogram of the number of observations contained in one leaf node.

4. 客観評価実験

4.1 実験条件

CSJ コアデータに含まれる対話音声を学習及び評価に用いた．コアデータには男女各 3 名の対話音声が含まれており，使用可能なデータ量は話者によって異なる．比較評価の使用に耐える品質の対話音声を合成できる話者依存モデルの学習には，ある程度のデータ量が必要である [6] ことを考慮して，本研究では 25 分以上の対話音声データが使用可能な女性話者 2 名 (ID=19，514) を選択した．対話音声は「学会講演インタビュー」「模擬講演インタビュー」「課題指向対話」の 3 対話からなり，各話者に対し約 25 分の発話データを使用した．

音声の特徴ベクトルは，16 kHz でサンプリングされた音声信号に対し STRAIGHT [16] で抽出した平滑化スペクトルと F0，非周期性指標から求められる 0 次から 39 次のメルケプストラム，対数 F0 と 5 次元非周期性指標，及びそれらの Δ, Δ^2 パラメータから成る 138 次元ベクトルとした．なお，5 次元非周期性指標は STRAIGHT で抽出される非周期性指標を



(a) Spectral distortion (speaker #19) (b) F0 distortion (speaker #19) (c) Duration distortion (speaker #19)

(d) Spectral distortion (speaker #514) (e) F0 distortion (speaker #514) (f) Duration distortion (speaker #514)

図 4　コンテクストセットとスペクトル，F0，モーラ継続長ひずみの関係
Fig. 4　Distortions of acoustic features with different context sets.

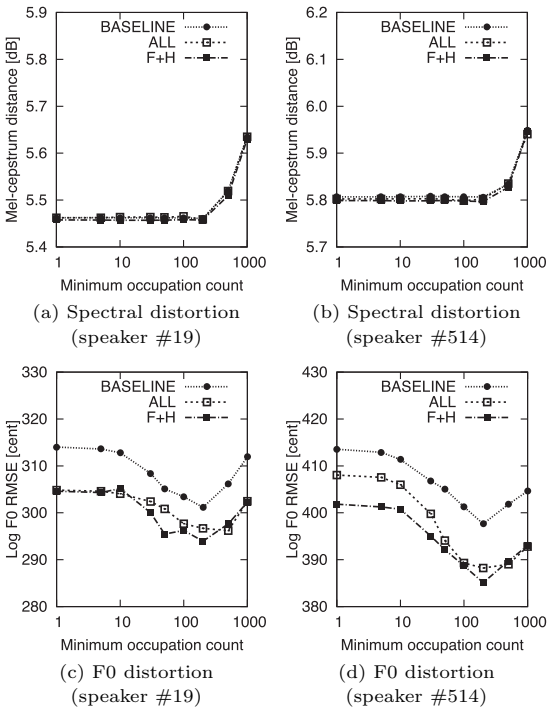
0-1, 1-2, 2-4, 4-6, 6-8 kHz のそれぞれの帯域で平均化することで求められる．フレーム周期は 5 ms とした．音響モデルには，HMM に状態継続長を明示的に組み込んだ隠れセミマルコフモデル (HSMM) [17] を用いた．HSMM は対角共分散単一混合ガウス分布をもつ 5 状態の left-to-right スキップなしモデルとした．決定木の停止基準には MDL 基準に加え，**3.** で述べた停止基準を用いた．学習及び評価には CSJ に付与されているラベルから自動的に作成したコンテクストラベルを使用した．自発音声には呼気段落，アクセント句の区切りが明確に存在しない場合があるが，本論文では X-JToBI の BI 層ラベルのうち，イントネーション句境界とフィラー句境界で区切られた区間を呼気段落とし，アクセント句境界と語断片境界で区切られた区間をコンテクスト上のアクセント句とした．

以下の評価ではいずれも，各話者 25 分の実験データをそれぞれ約 2.5 分のデータセットに 10 分割し，1 セットを評価データ，残りの 9 セット，22.5 分を学習データとして用いて，10 セットの結果の平均を求めることによって評価を行う 10-fold クロスバリデーションにより客観評価を行った．

4.2　拡張コンテクストの評価

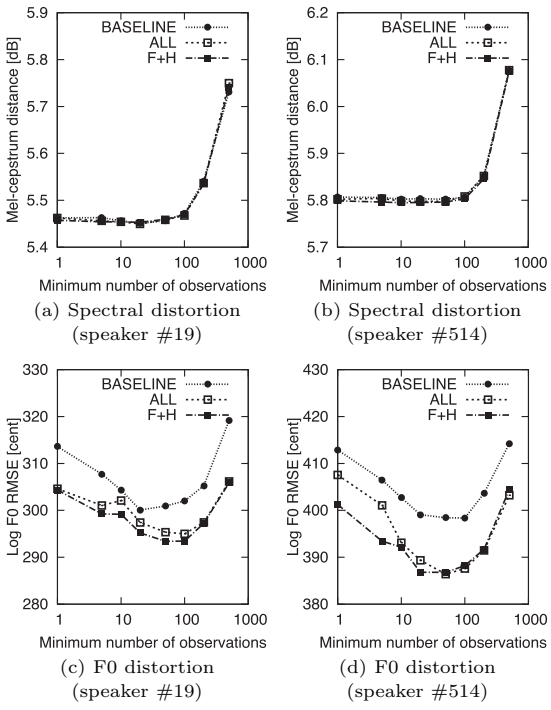
新たなコンテクストを追加した拡張コンテクストの有効性を評価するために，合成音声の原音声に対するスペクトル，F0，モーラ継続長のひずみを求めた．図 4 はコンテクストセットごとの平均メルケプストラム距離，対数 F0，モーラ継続長の RMS 誤差をそれぞれ表している．このとき，最小状態占有頻度は 5.0 とし，最小観測サンプル数に関してはしきい値を設けなかった．図中の BASELINE は従来のコンテクスト，ALL は **2.2** で述べた拡張コンテクストをすべて従来のコンテクストに加えたコンテクストを示している．対数 F0 とモーラ継続長の RMS 誤差は，拡張コンテクストを用いることで，従来のコンテクストセットに比べ減少していることが分かる．メルケプストラム距離については BASELINE と ALL の差はほとんど見られなかった．

それぞれのコンテクストカテゴリーの有効性を評価するために，表 1 の F から L までのカテゴリーを個別に BASELINE に加えてそれぞれのひずみを求めた．その結果を図 4 の F〜L に示す．F0 ひずみに関してはトーンラベル (H) を用いることで最もひずみが小さくなった．このことから F0 曲線の変曲点や句末境界音調といったトーンの情報が自然な F0 パターンを生成するための重要な情報であると考えられる．また，発話スタイル (G) を加えることによっても対数 F0 の RMS 誤差がわずかではあるが減少した．一方で節 (L) を加えた場合には F0 ひずみが大きくなってしまった．



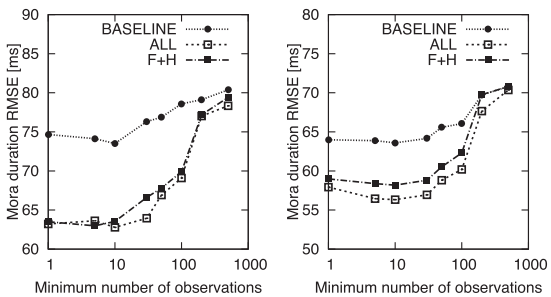
(a) Spectral distortion (speaker #19)
(b) Spectral distortion (speaker #514)
(c) F0 distortion (speaker #19)
(d) F0 distortion (speaker #514)

図 5　最小状態占有頻度とスペクトル，F0 ひずみの関係
Fig. 5　Spectral and F0 distortions as a function of the minimum occupation count.



(a) Spectral distortion (speaker #19)
(b) Spectral distortion (speaker #514)
(c) F0 distortion (speaker #19)
(d) F0 distortion (speaker #514)

図 6　最小観測サンプル数とスペクトル，F0 ひずみの関係
Fig. 6　Spectral and F0 distortions as a function of the minimum number of observations.



(a) Duration distortion (speaker #19)
(b) Duration distortion (speaker #514)

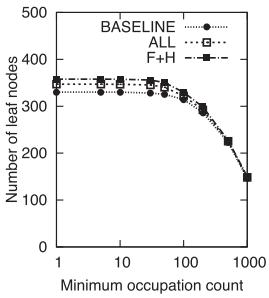
図 7　最小観測サンプル数とモーラ継続長ひずみの関係
Fig. 7　Mora duration distortions as a function of the minimum number of observations.

これは学習データ量不足によるモデルの過学習が起きてしまっているためと考えられる．モーラ継続長に関しては，音素引き延ばし (F) がひずみの減少に対して有効であり，他のカテゴリーはあまり有効でないという結果となった．

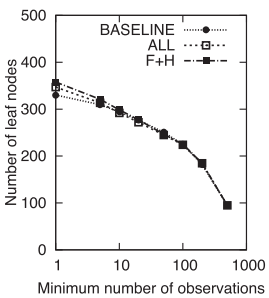
これらの結果を踏まえて，継続長及び F0 の再現性の向上にそれぞれ効果を示した音素引き延ばし (F) とトーンラベル (H) の，二つのコンテクストカテゴリーを BASELINE に加えた場合についても評価を行った．図 4 中の F+H はその結果を示しており，両コンテクストカテゴリーを追加するだけでもひずみが BASELINE から ALL の場合と同程度あるいはそれ以上に減少するという結果となった．

4.3　分割停止基準の評価

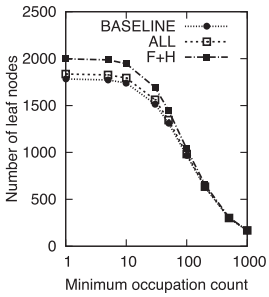
提案した分割停止基準の有効性を示すために，最小状態占有頻度と最小観測サンプル数を変化させたときの合成音声の原音声に対するひずみの変化を求めた．メルケプストラム及び対数 F0 に対する決定木構築において，最小観測サンプル数を制限せずに最小状態占有頻度を変化させたとき，あるいは最小状態占有頻度を 5.0 に固定した上で最小観測サンプル数を変化させた．そのときの平均メルケプストラム距離及び対数 F0 の RMS 誤差の変化を図 5，図 6 にそれぞれ示す．また，図 7 は状態継続長のクラスタリングにおいて，最小観測サンプル数を変化させたときのモーラ継続長の RMS 誤差の変化を表している．スペクトル，F0 と異なり，モーラ継続長ひずみと最小状態占有頻度の関係を示していないのは，状態継続長がフレームに対する



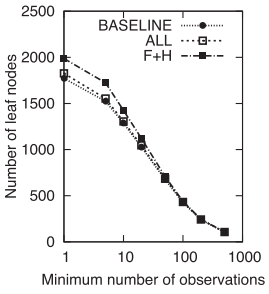
(a) Tree for mel-cepstrum using minimum occupation count



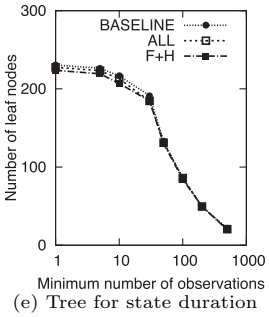
(b) Tree for mel-cepstrum using minimum number of observations



(c) Tree for log F0 using minimum occupation count



(d) Tree for log F0 using minimum number of observations



(e) Tree for state duration using minimum number of observations

図 8　話者 ID=19 における最小状態占有頻度及び最小観測サンプル数と平均リーフノード数の関係

Fig. 8　Relationship between tree size and stopping criteria.

特徴量ではないため，状態占有頻度が定義されないからである．更に図 8 は話者 ID=19 において停止基準を変化させたときの平均リーフノード数を表している．

図 5 及び 6 から，最小観測サンプル数 100 以下ないしは最小状態占有頻度 200 以下の場合はメルケプストラム距離にほとんど変化がないことが分かる．よって，メルケプストラムに関するコンテクストクラスタリングの際には，新たな分割停止基準を導入する必要性が低いと考えられる．一方で F0 ひずみに関しては，最小状態占有頻度，あるいは最小観測サンプル数を導入することでひずみが減少した．つまり，クラスタリングにおける新たな分割停止基準を導入することで，過学習を軽減できると考えられる．モーラ継続長ひずみに関しては，スペクトルひずみと同様に最小観測サンプル数が小さいときにモーラ継続長の RMS 誤差の変化が小さく，新たな分割停止基準の効果があまり見られないことが図 7 から分かる．また，図 8 において最小状態占有頻度が小さいときはリーフノード数がほとんど変化していないため，分割するノードの決定への影響が小さいものと考えられる．

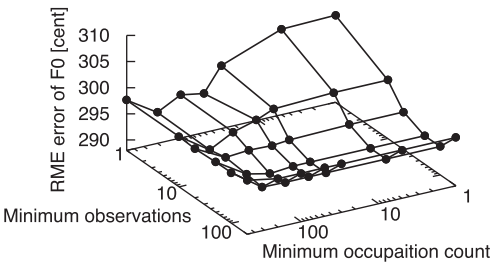


図 9　最小状態占有頻度及び最小観測サンプル数と F0 ひずみの関係

Fig. 9　F0 distortions as a function of the minimum occupation count and the minimum number of observations.

更に最小状態占有頻度と最小観測サンプル数を同時に変化させたときの F0 ひずみの変化を調べた．話者 ID=19 に対しコンテクストセット F+H を用いた場合における対数 F0 の RMS 誤差の変化を図 9 に示す．図から，一方しか使用しない場合に比べ双方の基準値を同時に用いることによるひずみの減少はほとんど見られず，基準値のどちらかを一方を使用することでひずみを抑えられることが分かる．また，図 5 と図 6 の F0 ひずみを比較すると，最小観測サンプル数を停止基準に用いた方が，最小状態占有頻度に比べひずみが基準値の変化の影響をやや受けにくいものと考えられる．

5.　ラベル位置コンテクストの予測

実際の音声合成で拡張コンテクストを利用する場合，全てのコンテクストを合成時にユーザが入力することは好ましくない．ここで F+H は ALL に近い結果を得られているため，F+H の追加コンテクストに注目する．F+H の追加コンテクストは，音素引き延ばし，

X-JToBI トーン層ラベルの相対的なモーラ位置，F0 変動の種類（句末境界音調，韻律的フィラー，韻律的語断片）である．実際の入力インタフェースを考えると，音素の引き延ばしには長音や促音の入力，F0 変動の種類には句末境界音調や韻律的フィラー・語断片を表す記号の入力を行うだけでよいので，ユーザの入力は大きく煩雑にはならないと考えられる．一方で，X-JToBI トーン層ラベルの相対的なモーラ位置をコンテクストとして使用し合成するためには，ラベルの位置情報が必要となる．しかし，一つひとつのアクセントに対し数種類のラベルの有無を選び，その位置を手作業で付与することは実用的ではないため，ラベルの位置は自動的に求められることが望ましい．実際のところ，アクセント型や句末境界音調が与えられればF0 曲線の概形が決まるため，そこからラベルの有無及び位置を求めることが可能であると考えられる．

本実験では，アクセント型や句末境界音調が与えられた状態でラベルの有無及び位置を求める方法として，規則によって決定する手法と決定木を用いて予測する手法の 2 種類の手法を検討した．

5.1 規則によるラベル位置の決定

X-JToBI のトーン層ラベルは F0 の変化に対し一定のラベリングスキームによって付与されている．例えば句頭音調「H–」は 2〜3 モーラ付近に，上昇調における「L%」「H%」は最終モーラの始端付近と終端付近にそれぞれ付与されるようになっている．よって，このスキームに基づいた規則からラベルの位置をある程度予測することが可能である．そこで，本実験では表 3 に示す規則を用いてラベルの位置を求めた．ただし，1 型アクセントの場合には「H–」のラベルが，0 型アクセントの場合には「A」のラベルが，それぞれアクセント句中に存在しないものとした．規則によって求めたラベル位置の実際の位置に対する正解率を表 3 に示す．

5.2 決定木によるラベル位置の予測

規則によるラベル位置の決定では，実際のラベルに即していない規則を指定してしまう可能性がある．ここで，合成時には HMM の学習に使用したデータを先見情報として使用することができるため，HMM の学習データを用いて合成時のラベル位置の予測を行う．本研究では有無及び位置の識別に WEKA [18] に実装されている決定木の C4.5 を用いた．説明変数には従来のコンテクストであるアクセント情報（アクセント句の長さやアクセント型）と F0 変動の種類を用い，目標変数には各ラベルに対し，アクセント句の先頭，知覚上のアクセント核，句の終端という 3 種類の位置を基準とした +1，+2，−1，−2 などの相対的なモーラ位置を使用した．なお，知覚上のアクセント核とは手作業やアクセント型予測により求められるアクセント核であり，0 型アクセントのアクセント句の場合には，先頭モーラの一つ前の位置，すなわち 0 番目のモーラにあたる位置を基準位置とした．前節と同様に 10-fold クロスバリデーションにより評価を行った．

ラベル位置の予測精度を表 4 に示す．表から，最もスコアの高くなる基準はラベルによって異なり，句頭境界や上昇の終端などの句の先頭に近いラベルでは先頭からの位置で，句末境界やハイ/ローポインタなどの終端に近いラベルでは終端からの位置でそれぞれ精度が高くなっていることが分かる．規則を用いた場合と比べると，「H–」の精度が大きく増加している．これは「H–」に関する規則が不十分であったためと考えられる．

5.3 予測結果に基づく合成

以上の結果に基づき，規則及び決定木から求められた位置情報から得られるコンテクストを用いて音声を合成した．前節と同様に各話者約 25 分の対話音声データを使用し，10-fold クロスバリデーションにより評価した．決定木を用いた場合は表 4 において太字となっ

表 3 ラベル位置の決定規則と正解率 [%]
Table 3 Rule and accuracy of label positions [%].

ラベル	位置	正解率 [%]
句頭境界 (%L)	先頭モーラ	88.51
上昇の終端 (H–)	2 番目のモーラ	63.96
下降の開始 (A)	アクセント核モーラ	86.80
下降の終了 (L%)	最終モーラ	83.96
句末境界 (*%)	最終モーラ	89.71
ローポインタ (pL)	最終モーラ	99.98
ハイポインタ (pH)	最終モーラ	99.76
フィラーポインタ (FL or FH)	中央のモーラ	97.84

表 4 ラベル位置の予測精度 [%]
Table 4 Accuracy of predicted label positions [%].

ラベル＼基準	先頭	アクセント核	終端
句頭境界 (%L)	**90.45**	90.19	89.80
上昇の終端 (H–)	**77.58**	77.58	76.72
下降の開始 (A)	88.59	**88.64**	87.77
下降の終了 (L%)	82.84	80.89	**83.97**
句末境界 (*%)	89.28	86.71	**89.74**
ローポインタ (pL)	99.52	99.45	**99.93**
ハイポインタ (pH)	99.62	98.86	**99.73**
フィラーポインタ (FL or FH)	**97.80**	97.80	97.68

表 5　トーンラベルの位置の自動予測による F0 ひずみ
Table 5　F0 distortions using predicted positions of tone labels.

	F0 RMS 誤差 [cent]	
コンテクストセット	話者 ID=19	話者 ID=514
F+H CORRECT	293.4	386.8
F+H RULE	298.4	390.9
F+H PREDICTED	299.3	391.3
F+H W/O POSITION	300.9	396.0
BASELINE	300.9	398.4

表 6　トーンラベルの位置の自動予測によるモーラ継続長ひずみ
Table 6　Mora duration distortions using predicted positions of tone labels.

	モーラ継続長 RMS 誤差 [ms]	
コンテクストセット	話者 ID=19	話者 ID=514
F+H CORRECT	63.1	58.6
F+H RULE	62.9	58.6
F+H PREDICTED	62.9	58.8
F+H W/O POSITION	65.9	58.7
BASELINE	74.0	64.0

ている予測結果を使用した．前節の客観評価の結果から音響モデルのリーフノードにおける最小観測サンプル数は 50 とした．表 5 及び表 6 に合成音声の原音声に対するひずみを示す．表の F+H CORRECT は CSJ に付与されたラベルをそのまま用いた場合であり，前節の F+H に相当する．また，F+H W/O POSITION はラベルの位置の情報を使用しないコンテクストセットであり，F+H RULE，F+H PREDICTED はそれぞれラベルの位置を規則，決定木で求めたコンテクストセットである．F+H CORRECT と F+H W/O POSITION を比較すると，F0，モーラ継続長においてある程度の差が存在し，ラベルの位置の情報がひずみを小さくする効果があることが分かる．また，ラベル位置を予測した F+H RULE 及び F+H PREDICTED では，話者 ID=19 のモーラ継続長ひずみが F+H CORRECT に近くなり，話者 ID=514 の F0 ひずみが F+H W/O POSITION から 5 cent 程度減少した．F+H RULE と F+H PREDICTED の差はわずかであり，規則を用いても決定木による予測と同程度の性能を実現することができる．話者によって効果は異なるが，結果としてトーンラベルの位置を予測したものであっても，ある程度合成音声の再現性を保つことができると考えられる．

6. 自然性の評価

以上の結果を踏まえ，MOS 評価による合成音声の主観評価を行った．コンテクストセットには BASELINE, ALL，F+H CORRECT 及び，F+H PREDICTED を用い，客観評価結果から ALL と F+H CORRECT, F+H PREDICTED では最小観測サンプル数を 50 に制限した．10 名の被験者は合成音声の自然性を 5 段階 (5: excellent, 4: good, 3: fair, 2: poor, and 1: bad) で評価した．話者 ID=19, 514 それぞれに対して，客観評価に使用した 10 モーラ以上の 294 発話，323 発話からランダムに選んだ各 10 発話を各被験者の評価音声とした．評価に使用した発話の合計は (2 話者) × (10 被験者) × (10 発話) の計 200 発話であり，複数被験者に重複して評価された発話を考慮すると異なる 168 種類であった．そのうち下降調 (L%) 以外の上昇調 (H%) や上昇下降調 (HL%) などの対話特有の句末音調が含まれる発話は 126 種類，韻律的フィラー (FL か FH) が含まれる発話は 86 種類，音素引き延ばしの含まれる発話は 88 種類であった．結果の平均スコアを図 10 に示す．ALL，F+H CORRECT 及び F+H PREDICTED は BASELINE に比べ高い平均スコアを得ることができ，その差は有意水準 0.05 において有意であった．また，F+H CORRECT と ALL を比較すると自然性にほとんど差がなく，トーンと音素引き延ばしの情報が拡張コンテクストの中で重要な要素であることが分かった．更にラベルの位置の自動予測を行った F+H PREDICTED においても CSJ のラベルをそのまま用いた F+H CORRECT とほぼ同等の自然性が得られた．

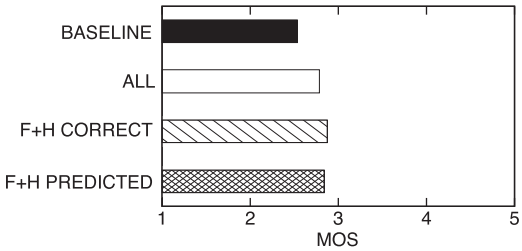


図 10　合成音声の自然性に関する MOS 評価の結果
Fig. 10　Result of a MOS test on the naturalness of synthetic speech.

7. むすび

多様な韻律をもつ対話音声の合成に向けて，CSJ のアノテーションデータに基づく拡張コンテクストセットの使用を提案し，その評価を行った．また，モデルの過学習を避けるため決定木に基づくコンテクストク

ラスタリングにおける分割停止基準の検討や，一部の追加コンテクストを自動的に予測する手法の評価を行った．その結果，追加コンテクストとして有効性が高いと考えられるのは，音素の引き延ばしと X-JToBI のトーン層ラベルに基づくコンテクストであった．また，拡張コンテクストにおいてリーフノードの最小観測サンプル数を制限することで再現性が向上した．更に，トーンラベルの位置を他のコンテクストから決定木により予測しても，ある程度の再現性を得ることができ，提案手法により合成音声の自然性が向上することを主観評価により確認した．

今後は実際の音声対話システムに用いるために，発話意図や語彙の印象表現，話し相手の発話などをもとに，拡張コンテクストを生成する方法を検討していく必要がある．

謝辞 本研究の一部は，日本学術振興会科学研究費補助金 (課題番号 21300063) の助成を得た．

文　献

[1] N. Campbell, “Developments in corpus-based speech synthesis: Approaching natural conversational speech,” IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.376–383, March 2005.

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” Proc. EUROSPEECH, pp.2347–2350, 1999.

[3] 赤川達也，岩野公司，古井貞熙，“HMM を用いた話し言葉音声合成のためのモデルの検討,” 信学技報，SP2007-3, 2007.

[4] S. Andersson, J. Yamagishi, and R. Clark, “Utilising spontaneous conversational speech in HMM-based speech synthesis,” Proc. 7th ISCA Workshop on Speech Synthesis, pp.173–178, 2010.

[5] C. Lee, C. Wu, and J. Guo, “Pronunciation variation generation for spontaneous speech synthesis using state-based voice transformation,” Proc. ICASSP, pp.4826–4829, 2010.

[6] T. Koriyama, T. Nose, and T. Kobayashi, “Conversational spontaneous speech synthesis using average voice model,” Proc. INTERSPEECH, pp.853–856, 2010.

[7] K. Maekawa, “Corpus of Spontaneous Japanese: Its design and evaluation,” ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.

[8] 山田真裕，岩野公司，古井貞熙，“数量化 I 類による F0 パターン生成の制御要因に関する検討,” 情報処理研報，2001-SLP-38-3, pp.15–20, 2001.

[9] 徳田恵一，益子貴史，宮崎　昇，小林隆夫，“多空間上の確率分布に基づいた HMM,” 信学論 (D-II)，vol.J83-D-II, no.7, pp.1579–1589, July 2000.

[10] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, “Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis,” IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.503–509, March 2005.

[11] 森實久美子，中村圭吾，戸田智基，猿渡　洋，鹿野清宏，“HMM に基づく音声合成における強調音声の生成,” 情報処理研報，2009-SLP-75-6, pp.27–32, 2009.

[12] 郡山知樹，能勢　隆，小林隆夫，“HMM に基づく対話音声合成の検討,” 音響秋季講論集，1-2-10, pp.255–256, Sept. 2009.

[13] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” Acoustical Science and Technology, vol.21, no.2, pp.79–86, 2000.

[14] 前川喜久雄，菊池英明，五十嵐陽介，“X-JToBI：自発音声の韻律ラベリングスキーム,” 信学技報，NLC2001-71, SP2001-106, 2001.

[15] 野村大輔，山岸順一，小林隆夫，“HMM 音声合成における決定木の分割停止基準の検討,” 音響春季講論集，I, 3-P-26, pp.291–292, March 2005.

[16] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, “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 Commun., vol.27, no.3-4, pp.187–207, 1999.

[17] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “A hidden semi-Markov model-based speech synthesis,” IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.825–834, May 2007.

[18] Weka 3: Data Mining Software in Java http://www.cs.waikato.ac.nz/ml/weka/

（平成 23 年 6 月 6 日受付，10 月 3 日再受付）

郡山　知樹 （正員）

2009 年 3 月東京工業大学工学部情報工学科卒．2010 年 9 月同大大学院総合理工学研究科物理情報システム専攻修士課程了．現在，同大学院博士後期課程在学中．ISCA，日本音響学会各会員．

能勢　　隆　（正員）

2001 年 3 月京都工芸繊維大学工芸学部電子情報工学科卒業 2009 年 3 月東京工業大学大学院総合理工学研究科博士後期課程了．博士（工学）．同年 4 月より同大学大学院総合理工学研究科助教，現在に至る．音声合成，音声認識の研究に従事．IEEE，ISCA，日本音響学会各会員．

小林　隆夫　（正員：フェロー）

1977 東工大・電気卒．1982 同大大学院総合理工学研究科博士課程了．工博．同年同大精密工学研究所助手．同助教授を経て現在同大大学院総合理工学研究科教授．音声言語情報処理，マルチモーダルインタフェース，ディジタル信号処理の研究に従事．2001 本会論文賞，猪瀬賞各受賞，日本音響学会，情報処理学会，IEEE，ISCA 各会員．