論文 / 著書情報
Article / Book Information

論題(和文)	話し言葉音声合成の韻律制御に関する検討
Title(English)	
著者(和文)	伊藤 芳幸, 岩野 公司, 古井貞熙
Authors(English)	Yoshiyuki Ito, Koji Iwano, SADAOKI FURUI
出典(和文)	情報処理学会研究報告, Vol. 2009-NL-191, 2009-SLP-76, No. 23, pp. 1-8
Citation(English)	, Vol. 2009-NL-191, 2009-SLP-76, No. 23, pp. 1-8
発行日 / Pub. date	2009, 5
権利情報 / Copyright	

Vol.2009-NL-191 No.23
Vol.2009-SLP-76 No.23
2009/5/22

話し言葉音声合成の韻律制御に関する検討

伊藤芳幸[†1]　岩野公司[†2]　古井貞熙[†1]

本稿では HMM 音声合成方式に基づく話し言葉音声合成において，韻律制御の精度向上を目的とした種々の検討を行う．我々の韻律推定手法では，数量化 I 類によって音素継続時間長（音素長）と基本周波数（F_0）をモデル化している．まず，韻律推定に利用する制御要因の数と種類の違いが，合成音声の「話し言葉としての自然性（話し言葉らしさ）」に及ぼす影響について検討を行った．その結果，音素長の推定に関して，読み上げ音声の合成では有効性が大きくなかった「アクセント句間のポーズ長」が，話し言葉音声の音素長制御に重要であることがわかった．基本周波数（F_0）の推定に関しては，現状の韻律モデル化手法で「話し言葉らしさ」を反映できない原因の分析を行った．その結果，アクセント句ごとの平均 F_0 値の分布が推定 F_0 では小さくなってしまう現象が確認され，それによって合成音声の「話し言葉らしさ」が十分に表現されないことがわかった．特に，本来平均 F_0 値が高いアクセント句に対して，正しい F_0 値が割り当てられれば，合成音声の「話し言葉らしさ」が改善することが確認された．

A study on prosody control for spontaneous speech synthesis

Yoshiyuki Ito ,[†1] Koji Iwano [†2] and Sadaoki Furui[†1]

This paper investigates several topics related to high-quality prosody estimation in HMM-based spontaneous speech synthesis. In our prosody control method, phoneme duration and fundamental frequency (F_0) are modeled by Quantification Theory (Type 1). We first analyzed the effects of the number and kinds of prosody control factors on spontaneity of synthesized speech. The analysis result showed that "pause length between prosodic phrases" was one of the important duration control factors while it was not particularly useful for duration control of reading-type speech synthesis. Next, we investigated reasons why the current prosody control method cannot sufficiently model the F_0 features of spontaneous speech. Through the analysis, it was confirmed that the distribution of estimated phrase-averaged F_0 values was reduced from original/correct distribution, and the reduction caused the low spontaneity of synthesized speech. It was also confirmed that spontaneity could be significantly improved if the correct phrase-averaged F_0 values were assigned to the phrases whose original F_0 values were located in a high-frequency region.

1. はじめに

近年，合成音声の多様化が進む中で，読み上げ調の音声だけではなく，話し言葉調の音声を合成する技術が望まれている[1)]．我々の研究室では，HMM 音声合成方式[2)] を利用した話し言葉音声合成システムの構築を目指して研究を進めている[3)]．我々の先行研究では，日本語話し言葉コーパス（CSJ）を用いて話し言葉音声のケプストラム情報，音素継続時間長（音素長）情報，基本周波数（F_0）情報のモデル化を行い，これらのモデルを用いて合成された音声の「話し言葉音声としての自然性（話し言葉らしさ）」に関する調査を行っている[4)]．各情報のモデル化精度を調査したところ，話し言葉音声のケプストラムは HMM によって十分な精度でモデル化されているが，音素長，F_0 といった韻律情報の数量化 I 類によるモデル化は精度が不十分であり，改善の余地があることがわかった．

そこで本研究では，話し言葉音声合成の韻律制御の改善を目的として，種々の調査・分析を行う．先行研究 4) では，韻律モデル化のための制御要因の種類や数は，読み上げ音声の韻律推定で有効であったものを便宜的に使用しており，十分な検討が行われていなかった．そこで本研究では，利用する制御要因の種類や数を変更した場合に，合成音声の「話し言葉らしさ」がどのように変化するかについて詳しく調査し，それによる韻律制御の改善の可能性を検討する．また，現状の手法では，特に F_0 のモデルが十分に「話し言葉らしさ」を反映できていないことから，このモデルによって推定された F_0 の特徴の分析を行い，問題点とその解決方法について考察する．

以降では，2 章で本研究で用いる HMM 音声合成に基づく TTS システムと，利用するモデルについて説明を行う．3 章では，話し言葉音声の韻律制御に用いる制御要因の数と種類に関する検討を行い，4 章で，話し言葉音声の F_0 推定結果の特徴分析と，推定 F_0 の「話し言葉らしさ」の改善可能性について考察を行う．最後に，5 章で本稿をまとめる．

†1 東京工業大学大学院 情報理工学研究科 計算工学専攻
Department of Computer Science, Tokyo Institute of Technology
†2 東京都市大学 環境情報学部 情報メディア学科
Faculty of Environmental and Information Studies, Tokyo City University

Vol.2009-NL-191 No.23
Vol.2009-SLP-76 No.23
2009/5/22

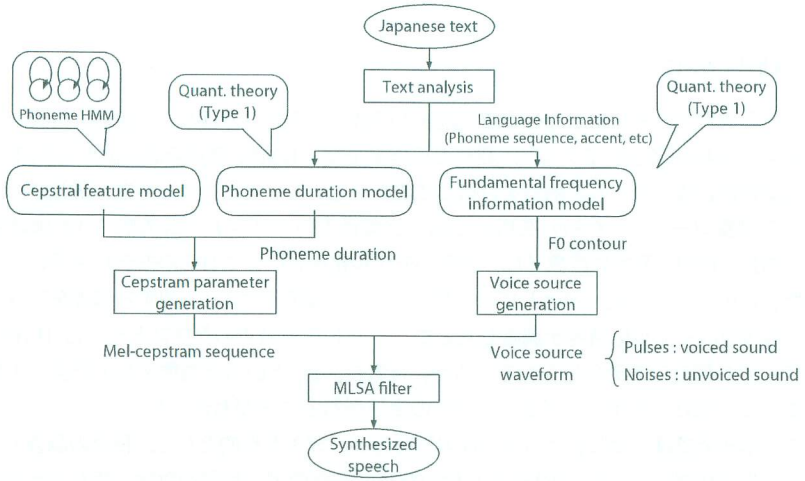

図 1　HMM 音声合成による TTS システムの構成

2. HMM 音声合成に基づく音声合成システム

2.1 HMM 音声合成に基づく TTS システムの構成

本研究で用いた HMM 音声合成に基づく TTS システムの概要を図 1 に示す[5]．このシステムでは，音声を合成するために，ケプストラム，音素長，F_0 の 3 つの音声特徴量を推定する．推定には，予め学習により構築した統計モデルを用いる．本合成システムでは，ケプストラムは音素 HMM で，音素長と F_0 は数量化 I 類によりモデル化を行っている．

音声を合成する流れは以下の通りである．まず入力された日本語テキストを解析して音素列とアクセント句情報を出力し，統計的なモデルを用いて各音素の音素長とモーラ毎の F_0 を推定する．推定した音素長と，ケプストラムをモデル化した音素 HMM を用いて，入力の音素列に対して最尤となるケプストラム系列を生成する[6]．得られたケプストラム系列と，推定された F_0 を元に生成したパルス列と白色雑音からなる音源信号を MLSA フィルタ[7] に入力することで音声を合成する．我々の最終的な目標は，話し言葉音声合成を行う TTS システムの構築であるが，本研究ではテキスト解析の精度の影響を除いて合成音声の分析と評価を行うため，テキスト解析部は用いず，テキスト解析が正しく行われたと仮定して得られる言語情報を利用して音声合成を行う．この場合の言語情報は CSJ のイントネーションラベルから生成する．

表 1　音素クラスの一覧

音素クラス	音素
1. 母音（vowel）	/a/,/i/,/u/,/e/,/o/
2. 撥音（syllabic nasal）	/N/
3. 促音（choked sound）	/Q/
4. 長音（long vowel）	/-/
5. 有声破裂音（voiced stop）	/b/,/d/,/g/
6. 無声破裂音（unvoiced stop）	/p/,/t/,/k/
7. 有声摩擦音（voiced fricative）	/z/,/j/
8. 無声破擦音（affricate）	/ch/,/ts/
9. 無声摩擦音（unvoiced fricative）	/f/,/h/,/s/,/sh/
10. 鼻音（nasal consonant）	/m/,/n/
11. 流音（liquid）	/r/
12. 半母音（semi vowel）	/w/,/y/
13. 拗音（palatalized consonant）	/by/,/dy/,/gy/,/py/, /ky/,/hy/,/ry/,/my/,/ny/

2.2 モデルの構築手法

2.2.1 ケプストラムモデル

ケプストラム情報は音素ごとに triphone HMM でモデル化を行う．音素 HMM にはスキップのない 3 状態 4 混合 left-to-right 型連続分布 HMM を用いた．音声のサンプリング周波数は 16kHz であり，音響特徴量としては 0〜25 次のメルケプストラム係数とその Δ 成分の計 52 次元のベクトルを使用した．フレームシフトは 5ms とした．特徴量抽出では，窓幅 16ms の STRAIGHT 分析[8] で得られたスペクトルをメルケプストラムに変換した．

2.2.2 音素継続時間長モデル

音素長のモデルは，数量化 I 類によって，表 1 に示す 13 の音素クラスごとに作成する．合成時にはこのモデルを用いて，音素ごとに継続長を定める[9]．数量化 I 類の目標値となる音素長は，ケプストラムモデルを用いた学習データの強制切り出しによって得た．

数量化 I 類の制御要因としては，読み上げ音声の音素長推定に関する我々の先行研究 9) で検討した 21 個の要因を用いた．表 2 に詳細を示す．

2.2.3 基本周波数モデル

F_0 も音素長と同様に，数量化 I 類を用いてモデル化を行う．目的変数は各モーラの母音，撥音，長音の中心時刻における（対数変換された）F_0 値であり，音声合成時には推定した F_0 値を直線補間することで文全体の F_0 パタンを生成する．目標値となるモーラごとの F_0

Vol.2009-NL-191 No.23
Vol.2009-SLP-76 No.23
2009/5/22

表 2　音素長推定に用いる制御要因（括弧内はカテゴリ数）

1	W のモーラ数（9）
2/3	P 内で W に先行/後続するモーラ数（9）
4/5/6	先行/当該/後続アクセント句のアクセント型（7）
7	P 内で W に先行する アクセント核を有する句の数（4）
8/9	W 前/後の音調結合の強さ（4）
10/11	W 前/後の句境界のポーズの長さ（9）
12/13	W の 2 つ前/後の音調結合の強さ（5）
14/15	W の 3 つ前/後の音調結合の強さ（5）
16	O が属するモーラの W 内のモーラ位置（9）
17	音素 O の種類 （1〜9：O の音素クラスによって変化．）
18/19	音素 O の前/後の音素の種類 （18〜29：O の音素種によって変化．）
20/21	音素 O の 2 つ前/後の音素の種類（18〜29）

表 3　F_0 推定に用いる制御要因（括弧内はカテゴリ数）

1	W のモーラ数（8）
2/3	P 内で W に先行/後続するモーラ数（9）
4/5	先行/後続アクセント句のアクセント型（7）
6	P 内で W に先行する アクセント核を有する句の数（4）
7/8	W 前/後の音調結合の強さ（4）
9/10	W 前/後の句境界のポーズの長さ（9）
11/12	W の 2 つ前/後の音調結合の強さ（5）
13/14	W の 3 つ前/後の音調結合の強さ（5）
15	トーンパタン[11]（5〜10：n により異なる．）
16	当該音素（母音，撥音，長音）の種類（8）
17/18	16 の音素の前/後の音素の種類（13）
19/20	16 の音素の 2 つ前/後の音素の種類（6）
21	M の W 内のモーラ位置（$n \geq 5$ の場合）（6）

値は，STRAIGHT 分析によって得られた F_0 系列を元に算出している．

制御要因としては，読み上げ音声の F_0 推定に関する我々の先行研究 10) で検討した 21 個の要因を用いた．表 3 に詳細を示す．ここで，推定対象のモーラを M，モーラ M が属するアクセント句を W，アクセント句 W が属する呼気段落（ポーズで区切られる音声区間）を P とする．モーラ M が，アクセント句 W の第 n モーラであるとすると，数量化 I 類のモデルは n =1, 2, 3, 4, 5 以上，の 5 つの場合に分けて作成する．

3. 話し言葉音声合成の韻律制御に用いる制御要因の検討

読み上げ音声の韻律推定に関する我々の先行研究では，推定に有効な制御要因の種類や数を調べるため，推定誤差に基づいて重要度の順位づけを行った上で，聴取実験を行って上位どこまでの制御要因が自然な韻律の推定に有効であるかを調査している[9),10)]．ここでは，同様の分析を行い，話し言葉音声の韻律推定に有効となる制御要因の種類や数について検討する．学習データには，CSJ のコアに含まれる男性話者 2 名，女性話者 2 名の模擬講演音声（各話者 30〜40 分程度）を用いる．なお，本研究ではフィラーは推定の対象とはせず，モデル学習データや評価データから取り除いている．

3.1 制御要因の重要度による順位付け

まず，韻律推定に用いる制御要因の重要度による順位づけを行う．方法は，以下の通りである．

(1) 対象となるすべて要因を用いて数量化 I 類のモデルを学習し，モデル学習データに対する推定誤差を求める．

(2) 各要因を 1 つだけ取り除き，残りの要因を用いてモデル構築を行った場合について推定誤差を計算する．

(3) 除いたときに推定誤差の増加が最も小さくなる要因を「最も重要度が小さい要因」とみなし，分析対象から外して（1）に戻る．

このように，重要度が小さい要因を順に除きながら韻律モデルの学習・推定誤差の算出を繰り返し行うことで，各制御要因を重要度で順位付けする．話者ごとに制御要因の順位付けを行い，それらの結果から，各制御要因の平均順位を求めて順位を付け直した．

その結果，音素長に関する重要な制御要因の順位は，

19 18 11 21 16 20 1 17 10 5 2 6 3 4 15 13 12 7 14 9 8

となり，F_0 に関しては，

15 17 3 2 1 18 4 9 10 5 6 16 11 12 13 20 19 7 21 8 14

となった．

3.2 聴取実験による重要な制御要因の調査

次に，韻律制御に用いる制御要因数を変化させて音声を合成し，被験者による聴取実験を行って，それぞれの合成音声の「話し言葉音声としての自然性（話し言葉らしさ）」の評価を行い，有効となる制御要因を調べる．

音素長に関する調査では，上位 1, 4, 7, 21 の制御要因を用いた場合の合成音声を作成し，

Vol.2009-NL-191 No.23
Vol.2009-SLP-76 No.23
2009/5/22

それぞれの合成音声の「話し言葉らしさ」に関して対比較を行うことで，上位どこまでの制御要因が有効であるかを調べる．この際，F_0 は抽出値（正解値）を用いた．なお，ポーズ長については有効な推定手法が確立されていないため，正解値を用いることとした．F_0 に関する調査では，上位 1, 4, 7, 10, 21 個の制御要因を用いた場合の合成音声を作成し，同様の対比較実験を行う．その際の音素長は正解値を用いた．

話者 4 名それぞれについて，学習に用いた文のうちランダムに 5 文を選び聴取実験に用いた．同じ文を異なる制御要因数の韻律モデルで合成した音声のペアを 11 名の被験者にヘッドホンで提示し，どちらが「より話し言葉音声として自然か」を評価してもらった．音素長，F_0 のそれぞれの評価に対して，一人の被験者あたり，各制御要因数間（例えば制御要因数 1 と 7）の比較評価を 2 回行っている．合成音声の話者や文の種類，制御要因数の組み合わせやペアの提示順などはランダムで提示した．

音素長の各制御要因数におけるプリファレンススコアと，聴取実験に用いた合成音声に対する音素あたりの平均推定誤差を図 2 に，F_0 の各制御要因数におけるプリファレンススコアと，モーラあたりの平均推定誤差を図 3 に示す．平均推定誤差は，二乗平均平方根（RMS）で計算されており，音素長の単位には ms，F_0 の単位には semitone（$= 12\log_2(F_0[\mathrm{Hz}]/55)$）を用いた．

3.3 考　　察

音素長に関する結果について，各モデル間のプリファレンススコアの差を二項分布に基づいて有意水準 5% で検定したところ，制御要因数 1 と他の制御要因数との間には有意差がみられたが，制御要因数 4, 7, 21 の間には有意差がみられなかった．この結果から，音素長のモデル化に対しては，上位 4 つ程度の制御要因（周辺音素の種類や当該／後続アクセント句間のポーズ長）が合成音声の「話し言葉らしさ」の知覚上，重要であることがわかる．読み上げ音声の音素長制御に有効な制御要因の分析結果[9] では，周辺音素種は同様に上位の要因であったが，ポーズ長は中位（8 番目程度）であったため，これまでの話し言葉音声の音素長推定[4] ではポーズ長の情報は制御要因として利用していなかった．したがって，この「当該／後続アクセント句間のポーズ長」の情報を用いることで，話し言葉音声合成の音素長推定の精度が向上する可能性がある．

F_0 推定に関する結果についても同様にスコアについて検定を行ったところ，制御要因数 1 と他の制御要因数との間には有意差がみられたが，制御要因数 4, 7, 10, 21 間には有意差はみられなかった．したがって，上位 4 つ程度の制御要因（トーンパタン，周辺音素の種類，周辺アクセント句のモーラ数）が合成音声の「話し言葉らしさ」の知覚上，特に重要であることがわかる．これらの上位の制御要因については，読み上げ調音声の F_0 制御要因の分析結果[9] との間に，大きな違いは見られなかった．先行研究 4) において，これらの制御要因を用いて話し言葉音声の F_0 推定を行った場合に，再合成音声と比較して十分な話し言葉らしさが得られないことが確認されている．今回の分析により，制御要因の種類や数を変化させた場合でも，推定される F_0 の話し言葉らしさに改善がみられないことが確認されたことから，現状の制御要因・制御方法では F_0 のモデル化精度が不十分であることが改めて

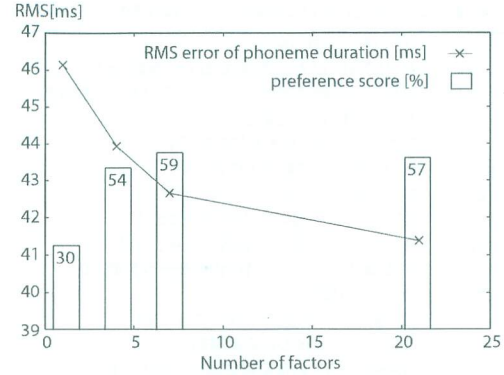

図 2　各制御要因数における音素長の推定誤差とプリファレンススコア

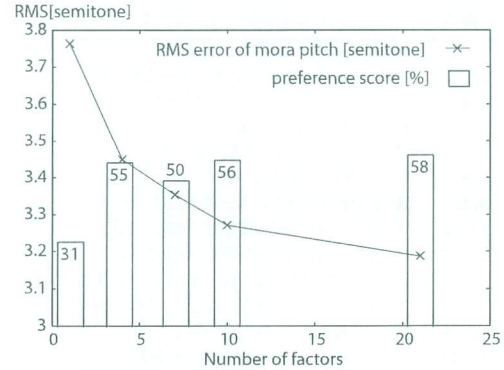

図 3　各制御要因数における F_0 の推定誤差とプリファレンススコア

Vol.2009-NL-191 No.23
Vol.2009-SLP-76 No.23
2009/5/22

示された．したがって，この原因を分析し，別の制御要因の導入や，数量化 I 類以外のモデル化手法の検討などを行う必要がある．

4. 話し言葉音声の F_0 推定における問題点の分析

話し言葉音声の F_0 は，現状の数量化 I 類による制御手法で十分にモデル化できないことが確認された．そこで，本章ではこの方法によって推定された F_0 が「話し言葉らしさ」を反映できない原因について検討する．

ある話し言葉音声について，モーラごとの F_0 の正解値と推定結果を比較した例を図 4 に示す．実線が正解値の F_0 系列を表しており，破線が推定値の F_0 系列を表している．縦線は句境界を示しており，この例では 10 個のアクセント句がある．図を見ると，推定がほぼ正しく行われているアクセント句と，推定結果が正解から大きく離れているアクセント句があることが分かる．具体的には，5, 6, 7 番目のアクセント句については F_0 推定結果は正解とほぼ同じであるが，他の句では，推定結果が正解のものと大きく異なっている．そこで，以降ではアクセント句を単位として推定された F_0 の特徴分析を行う．

4.1 句を単位とした話し言葉音声の推定 F_0 の特徴分析

まず，F_0 推定結果と正解（原音声）について，アクセント句ごとにモーラ F_0 の平均値と標準偏差を算出し，その分布を調べた．結果を図 5 に示す．使用したデータは日本語話し言葉コーパス（CSJ）中の男性話者 1 名の講演音声（約 18 分）であり，合計 1,271 アクセント句が分析対象となっている．これを見ると，推定結果の分布が正解の分布よりもかなり縮小していることがわかる．

次に，アクセント句ごとの F_0 のばらつき具合と音声の「話し言葉らしさ」の関係を調べるため，実際の「読み上げ音声」と「話し言葉音声」との間で，句ごとの F_0 の平均値と標準偏差の分布がどのように異なるかについて分析を行った．この実験には，先ほどの分析で用いたデータと同じ男性話者 1 名の講演音声（約 18 分）を話し言葉音声として利用し，その再朗読音声（約 23 分）を読み上げ音声として利用した．なお，先行研究 4) より，この話者の講演音声と再読み上げ音声について聴取実験を行ったところ，両者の「話し言葉らしさ」には大きな差が知覚できることが分かっている．読み上げ音声と話し言葉音声における，アクセント句ごとの F_0 の平均値と標準偏差の分布を，それぞれ図 6 に示す．これを見ると，標準偏差の分布については両者の間で大きな違いは見られないが，平均値については話し言葉音声よりも読み上げ音声の方が，分布が小さくなっていることが分かる．特に，話し言葉音声については 17〜20 semitone（約 145〜175 Hz）周辺の高い F_0 値が多く観測

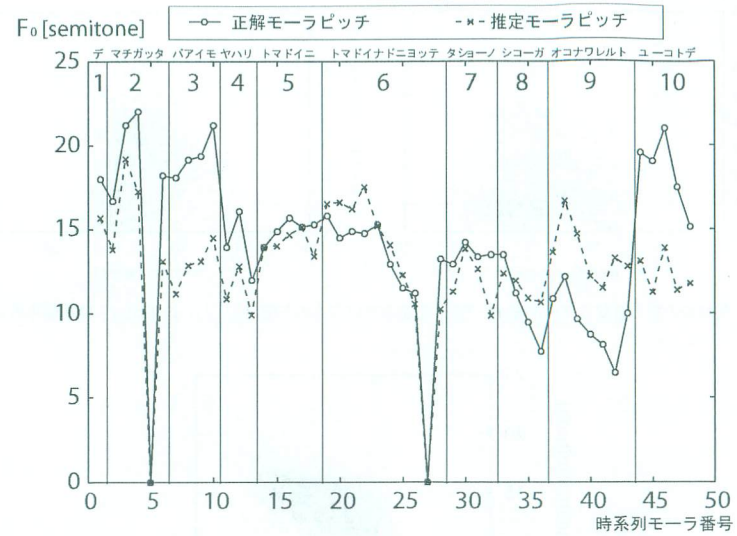

図 4 ある話し言葉音声における正解 F_0 と推定 F_0 のモーラごとの値の比較

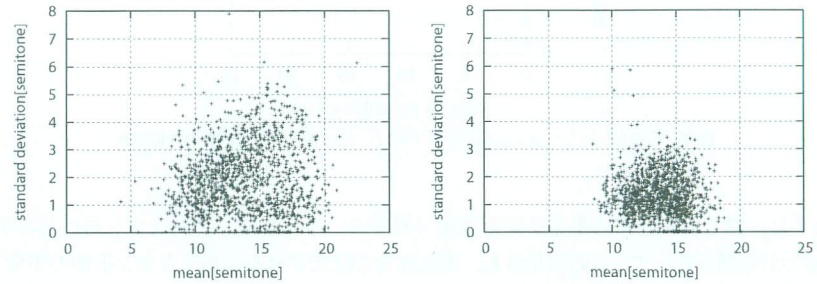

図 5 句単位の話し言葉音声の正解 F_0（左）と推定 F_0（右）の平均値・標準偏差の分布

されているのに対し，読み上げ音声ではこの領域の F_0 が少ないことがわかる．図 5 の推定された話し言葉音声の F_0 値の分布では，標準偏差も平均値も共にばらつきが小さくなっているが，特に平均 F_0 値の分布が小さくなってしまうことが，「話し言葉らしさ」が十分に反映されない原因の一つとして考えられる．

Vol.2009-NL-191 No.23
Vol.2009-SLP-76 No.23
2009/5/22

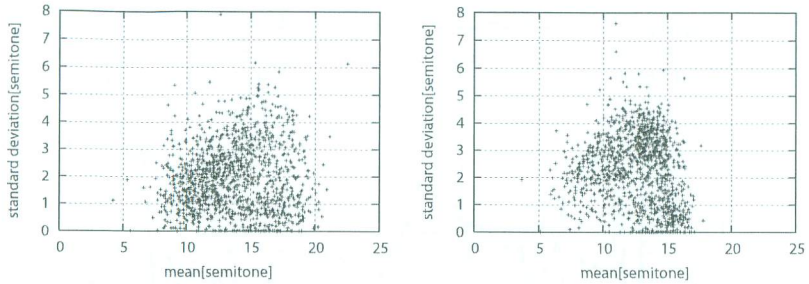

図 6 句単位の話し言葉音声の正解 F_0（左）と読み上げ音声の正解 F_0（右）の平均値・標準偏差の分布

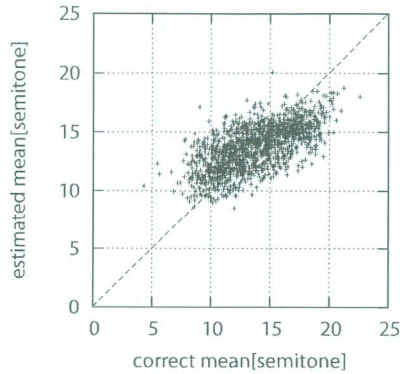

図 7 句を単位とした話し言葉音声の平均 F_0 値の推定値と正解値の対応関係

図 7 に，話し言葉音声の推定結果と正解（原音声）との間の，アクセント句ごとの平均 F_0 値の対応関係を示す．この図からも，本来大きく推定されるべきアクセント句の平均 F_0 値（横軸の 17〜20 semitone 周辺）が，それよりも小さく推定され，本来小さく推定されるべきアクセント句の平均 F_0 値 (横軸の 7〜10 semitone 周辺) が，それよりも大きく推定されている様子がわかる．

4.2 平均 F_0 値が正しく推定された場合の「話し言葉らしさ」の調査

アクセント句単位の平均 F_0 が精度よく推定された場合に，合成音声の「話し言葉らしさ」がどれだけ改善するかについて検討する．推定が理想的に行われた場合を想定するため，推定 F_0 値に対して，以下のような変換操作を行った．

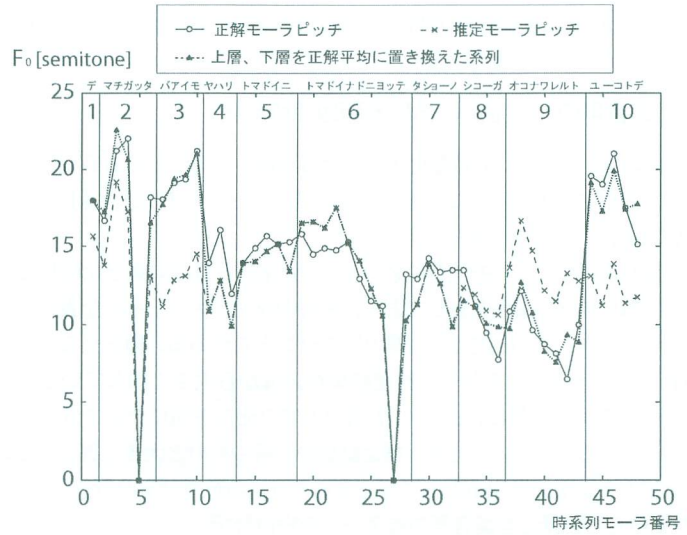

図 8 推定 F_0 に対して，上下層句の平均 F_0 値を正解値になるように変換した例

(1) 推定された話し言葉音声のアクセント句ごとの平均 F_0 値に対して，平均値（μ）と標準偏差（σ）を求める．

(2) 対応する正解のアクセント句の平均 F_0 値が $\mu+2\sigma$ を越えるものを「上層句」，$\mu-2\sigma$ を下回るものを「下層句」とする．これらの句については，平均 F_0 値を正解のものに置き換え，該当する句のモーラ単位の F_0 値を計算しなおす．その際，句内でのモーラ F_0 の相対関係は変化させないため，この F_0 変換操作は，アクセント句ごとの F_0 パタンの（上下方向の）平行移動操作となる．

図 8 に，図 4 で挙げた話し言葉音声の例に対して，上記の変換を施した場合の結果を示す．変換後の F_0 パタンが正解に近づいている様子がわかる．節 4.1 で用いたデータに対して，上述の変換操作を行った場合の句ごとの F_0 平均と標準偏差の分布を図 9 に示す．変換後の句単位の平均 F_0 値の分布が拡張されていることががわかる．そこで，この F_0 変換後の発話が，どの程度の「話し言葉らしさ」を持っているかについて，聴取実験により調べる．

Vol.2009-NL-191 No.23
Vol.2009-SLP-76 No.23
2009/5/22

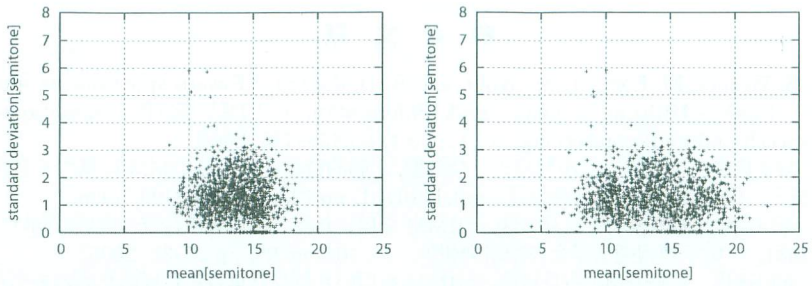

図 9 句単位の話し言葉音声の推定 F_0（左）と上層・下層句変換後の F_0（右）の平均値・標準偏差の分布

4.2.1 聴取実験の条件

聴取実験には，CSJ の学会講演音声（153 発話）を用いた．合成対象とする文には，推定 F_0 を変換する際に，上層句・下層句がバランスよく含まれるように，「上層句を全体句数の 20〜40%，下層句を全体の 20〜40%」含むような 21 発話を選択した．その中の 10 発話を学習用データから取り除き，143 発話を用いて F_0 モデルの学習を行った．その際，F_0 モデルの制御要因数は 21 とした．open 条件での聴取実験には学習に用いていない 10 文から無作為で選択した 5 文を，closed 条件の実験には，合成対象発話の残りの 11 発話から無作為に選んだ 5 文を使用した．なお，ケプストラム情報は HMM から推定された値を用い，音素長には強制切り出しによって得られた値（正解値）を用いた．被験者は 18 名である．

4.2.2 聴取実験の結果

「推定 F_0 の上・下層句に対して変換を行ったもの」，「正解 F_0」，「(変換前の) 推定 F_0」それぞれを用いた場合の合成音声について，それぞれの間で対比較実験を行い，どちらがより話し言葉らしく聞こえたかを主観評価してもらった．音声はヘッドホンを用いてランダムな順序で提示し，音声は何度でも聞けるように配慮した．それぞれの実験に対して，一話者あたり，1〜2 回の対比較を行っており，一つの実験あたり 30 回の比較実験で構成されている．プリファレンススコアの結果を図 10 に示す．*印は，二項分布に基づく有意検定を行ったときに，有意水準 10%でスコア間に有意差が認められたことを表している．図より，open，closed 実験ともに，変換によって「話し言葉らしさ」が改善していることがわかる．しかし，open 条件では正解 F_0 と変換 F_0 間で有意差がみられることから，変換操作による改善が不十分であり，句ごとの平均 F_0 以外にも「話し言葉らしさ」に関連のある要因が存在することが考えられる．

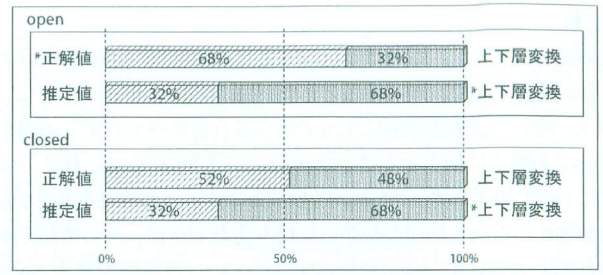

図 10 正解 F_0，推定 F_0，変換 F_0 を用いた合成音声の「話し言葉らしさ」を比較した場合のプリファレンススコア

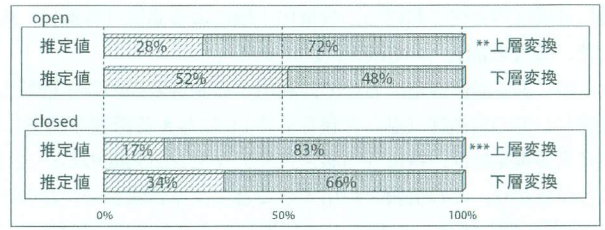

図 11 推定 F_0 と，上層・下層の一方を変換した F_0 を用いた合成音声の「話し言葉らしさ」を比較した場合のプリファレンススコア

4.2.3 変換箇所の違いによる「話し言葉らしさ」の変化の調査

前節の実験では，上下層句全てを対象に変換を行って，「話し言葉らしさ」に与える影響を調査した．ここでは，それぞれの層（上層，下層）を個別に変換した場合に，どの程度「話し言葉らしさ」に変化があるかを調査した．対比較項目は，「下層句のみ変換した F_0 と推定 F_0」「上層句のみ変換した F_0 と推定 F_0」の 2 つとした．それぞれの F_0 値から合成された音声のプリファレンススコアを図 11 に示す．**,***印は，二項分布に基づく有意検定を行ったときに，それぞれ有意水準 5%，1%でスコア間に有意差が認められたことを表している．open，closed 条件ともに，上層句のみ変換した F_0 を用いた場合に，「話し言葉らしさ」が改善していることがわかる．下層句のみを変換した場合には，open，closed 条件ともに有意差は見られなかった．このことから，「本来大きな F_0 平均値となるアクセント句を，より精度よく推定する」ことが重要であり，これにより「話し言葉らしさ」の改善をはかることが可能であることが示された．

Vol.2009-NL-191 No.23
Vol.2009-SLP-76 No.23
2009/5/22

4.3 考　　察

現状の F_0 制御手法において，このような本来大きな F_0 平均値を有するアクセント句を上手く制御できない原因として，句を単位とした「強調発声」を扱っていないことが考えられる．強調発声は，話者の意図や感情によって生じると考えられるが，意図や感情といった情報をテキストから自動的に抽出することは困難である[12]ため，これまではこのような情報を F_0 制御に用いていなかった．今後，話し言葉音声の F_0 推定精度の向上のためには，例えばこれらの情報をシステム使用者が明示的に与えた場合に，その情報を制御要因として利用して，自然な話し言葉音声の F_0 を推定する，といった手段を考える必要がある．

5. ま と め

本稿では，話し言葉音声合成における，数量化 I 類による韻律（音素長・F_0）推定の精度向上を目的とした，種々の調査・分析を行った．

まず，数量化 I 類による韻律推定において，利用する制御要因の数と種類の違いが，合成音声の「話し言葉としての自然性（話し言葉らしさ）」に与える影響を調査した．その結果，話し言葉音声の音素長推定には，特に「当該／後続アクセント句間のポーズ長」が重要な制御要因であることがわかり，この情報を利用することによる音素長モデルの改善可能性を示した．また，F_0 制御に関しては，これまでに利用してきた制御要因のみでは「話し言葉らしさ」を十分にモデル化できないことがわかった．

次に，F_0 が現状の手法で高精度にモデル化できない原因について分析を行った．その結果，推定された F_0 では，アクセント句ごとの平均値のばらつきが実際の話し言葉音声に比べて小さくなっていることがわかり，この影響を取り除くことができれば，合成音声の「話し言葉らしさ」が改善する可能性があることを確認した．特に，本来大きな F_0 平均値を有するアクセント句を正しく推定することが重要であることがわかった．

今後の課題としては，1) ポーズ長の言語情報からの推定，2) F_0 推定における「強調発声」の扱いの検討，3) 音源強度の推定手法の確立，などが挙げられる．

謝辞　STRAIGHT 分析のためのコードをご提供頂いた和歌山大学の河原英紀教授に深く感謝致します。また，実験に用いた音声合成システムの構築に多大な貢献をして下さった，当研究室の神山歩相名君に感謝致します．

参 考 文 献

1) S. Werner, M. Eichner, M. Wolff, and R. Hoffmann, “Toward spontaneous speech synthesis – Utilizing language model information in TTS,” IEEE Transactions on speech and audio prosecing, vol.12, no.4, pp.436–445, 2004.
2) 益子貴史，徳田恵一，小林隆夫，今井聖，“動的特徴を用いた HMM に基づく音声合成，” 電子情報通信学会論文誌，vol.J79-D-II, no.12, pp.2184–2190, 1996.
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実現に向けての検討，” 電子情報通信学会技術研究報告，vol.105, no.98, pp.25–30, 2005.
4) 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，” 電子情報通信学会技術研究報告，vol.107, no.77, pp.13–18, 2007.
5) K. Iwano, M. Yamada, T. Togawa, and S. Furui, “Prosody control for HMM-based Japanese TTS,” In S. Narayanan and A. Alwan (Eds.), Text to Speech Synthesis – New Paradigms and Advances –, Prentice Hall PTR, New Jersey, Ch.8, pp.155–173, 2004.
6) 立和　航，古井貞熙，“HMM による規則音声合成の検討，” 日本音響学会講演論文集，2-3-7, pp.239–240, 1999.
7) 今井　聖，住田一男，古市千枝子，“音声合成のためのメル対数スペクトル近似（MLSA）フィルタ，” 電子情報通信学会論文誌，vol.J66-A, no.2, pp.122–129, 1983.
8) H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a reptitive structure in sounds,” Speech Communication, vol.27, pp.187–207, 1999.
9) 岩野公司，山田真裕，外川太郎，古井貞熙，“HMM に基づく音声合成における様々な発話速度の実現，” 電子情報通信学会技術研究報告，vol.102, no.292, pp.11–16, 2002.
10) 山田真裕，岩野公司，古井貞熙，“数量化 I 類による F_0 パターン生成の制御要因に関する検討，” 情報処理学会研究報告，vol.2001, no.100, pp.15–20, 2001.
11) 箱田和雄，塚田元，吉田由紀，広川智久，水野秀之，“波形合成法を用いたテキスト音声合成ソフトウェア（FLUET），” 電子情報通信学会ソサイエティ大会講演論文集，D-462, p.465, 1996.
12) 阿部匡伸，佐藤大和，“音節区分化モデルに基づく基本周波数の 2 階層制御方式，” 日本音響学会誌，vol.49，no.10，pp.682-690，1993.