論文 / 著書情報
Article / Book Information

論題(和文)	HMMを用いた話し言葉音声合成に関する検討
Title(English)	
著者(和文)	赤川達也, 岩野公司, 古井貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会2005年秋季講演論文集, Vol. , No. 3-6-17, pp. 361-362
Citation(English)	, Vol. , No. 3-6-17, pp. 361-362
発行日 / Pub. date	2005, 9

HMMを用いた話し言葉音声合成に関する検討*

◎赤川達也, 岩野公司, 古井貞熙 (東工大)

1 はじめに

近年，多様な合成音声を実現するため，読み上げ調の音声だけではなく話し言葉音声を合成する技術が望まれている．そこで本研究では，HMM 音声合成 [1] を用いて話し言葉音声合成の実現を目指す．そのために，日本語話し言葉コーパス (CSJ) の話し言葉音声と読み上げ音声を用い，話し言葉音声合成の実現可能性と合成音声の話し言葉らしさに影響を与える要因についての検討を行う．

2 HMM を用いた音声合成

2.1 HMM を用いたテキスト音声合成システム

HMM 音声合成を用いた TTS システムとして，我々はこれまでに Fig.1 に示すような構成のシステムについて検討を行ってきた [2][3]．このシステムでは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継続時間長 (以下，音素長)，モーラ毎の基本周波数 (F_0) を推定する．推定した音素長と音素 HMM を用いて入力の音素列に対して最尤のケプストラム列を生成し [4]，それを F_0 情報から生成した音源信号とともに MLSA フィルタ [5] に入力することで音声を合成する．音素長と各モーラの F_0 は数量化 I 類を用いてモデル化される [2][3]．本研究においても，同様の構成で話し言葉音声の TTS を実現することを最終的な目標とする．

2.2 HMM 分析合成システム

話し言葉音声の F_0 を推定するためのモデルを作成するためには，高精度なテキスト解析とその情報を用いたモデル化手法が必要となる．しかし現段階ではこれらについて有効な手法が確立されていないため，本研究では F_0 情報には原音声から抽出したものを用い，話し言葉音声合成の実現可能性を検討する．

Fig.2 に本研究で用いる音声合成システムを示す．以下，このシステムを「HMM 分析合成システム」と呼ぶ．入力は音声とその音素ラベルであり，音素長は音素ラベルと数量化 I 類による音素長モデルから推定する．F_0 に関しては，入力音声からモーラごとに特定音素 (母音，撥音，長音) 中心における対数 F_0 値を抽出し，それを推定した音素長を基に決定される新たな音素の中心位置に合わせて配置し直した後，線形補間したものを音源生成に利用する．

3 使用データと音声合成用モデルの作成

話し言葉音声と読み上げ音声を使用して HMM 分析合成を行い，その合成音声の話し言葉らしさ，読み上げ音声らしさを評価する．その際，CSJ の同一話者による話し言葉音声，読み上げ音声を用いることで，話者の違いによる影響を受けずに合成音声の話し言葉らしさを評価することができる．

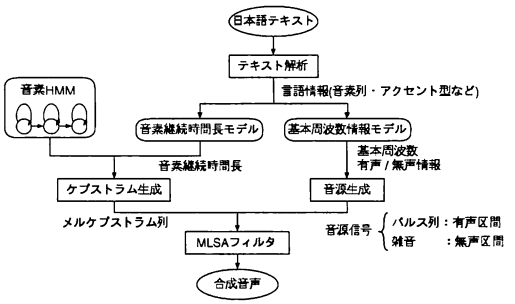

Fig. 1 HMM 音声合成による TTS システムの構成

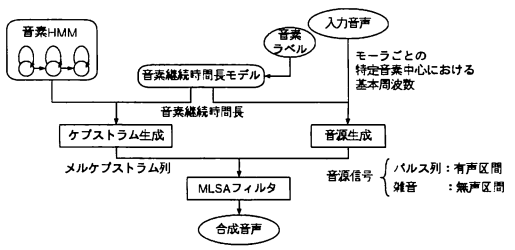

Fig. 2 HMM 分析合成システム

3.1 使用した音声データ

モデルの学習データとして，CSJ のコアに含まれる学会講演音声とその再読み上げ音声を用いた．Table 1 に，使用した音声データの詳細を示す．「講演 ID」が "A" で始まるものが学会講演音声，"R" で始まるものが再読み上げ音声であり，これらを話し言葉音声，読み上げ音声として用いた．表より，話し言葉音声の特徴として，読み上げ音声に比べて発話速度，基本周波数の平均値と標準偏差が増加するといった傾向が見られる．なお，入力音声は学習データの中から 10〜20 秒程度の区間をランダムに取り出して使用した．

3.2 音声合成用モデルの作成

Table 1 に示した 6 話者× 2 発話スタイルの 12 音声それぞれについて，4 混合の triphone HMM を学習した．学習の際には 16kHz の音声信号をフレーム長 32ms, フレーム周期 5ms のハミング窓を用いてメルケプストラム分析し，求めた 0〜24 次のメルケプストラムとその Δ 係数を音響パラメータとした．次に，強制切り出しにより音素長を求め，文献 [3] の手法に従い数量化 I 類によって音素長モデルを作成した．

4 主観評価実験

4.1 合成音声の話し言葉らしさに関する評価実験

HMM 分析合成により話し言葉らしい音声が合成できるかどうかを調査するため，モデル学習及び入力音声全てに話し言葉音声を用いて合成した音声と，全てに読み上げ音声を用いて合成した音声を被験者に提示し，話し言葉らしさを 5 段階 (1. 読み上げ調で

---

* Toward realization of HMM-based spontaneous speech synthesis
By Tatsuya Akagawa, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology)

Table 1　使用した音声データ

話者 ID	講演 ID	総音素数	発話速度 [モーラ/秒]	基本周波数情報 平均値 [Hz]	標準偏差
f01	A06F0128	14,321	7.79	228.8	47.0
	R00F0407	14,664	6.93	212.8	39.9
f02	A05F0043	12,836	8.19	197.5	37.5
	R00F0028	13,111	7.89	191.4	30.2
f03	A01F0122	8,977	7.44	205.7	35.8
	R00F0178	8,998	7.11	191.1	31.8
m01	A11M0846	13,199	9.07	123.0	27.5
	R00M0036	14,082	7.43	115.1	21.7
m02	A01M0056	8,051	8.62	100.1	21.9
	R00M0187	8,242	7.76	89.4	15.8
m03	A11M0369	12,086	8.80	170.5	26.4
	R00M0134	12,452	8.20	179.3	27.2

Table 2　HMM 分析合成音声の話し言葉らしさの評価

使用データ	話者						全体
	f01	f02	f03	m01	m02	m03	
話し言葉	3.7	3.0	2.5	4.1	4.2	3.6	3.51
読み上げ	2.6	3.1	2.3	2.2	2.5	2.5	2.53

ある，2. やや読み上げ調である，3. どちらとも言えない，4. やや話し言葉調である，5. 話し言葉調である) で評価してもらった．被験者は 10 名であり，音声は被験者ごとにランダムな順序に並べ変えて提示した．

Table 2 に平均評価値を示す．全話者での結果を見ると，話し言葉音声と読み上げ音声を用いた場合の平均評価値には危険率 1%で有意差が確認された．このことから，HMM 分析合成による合成音声で話し言葉らしさが表現されていることが確認された．話者ごとに見ると，話者 f02, f03 では評価値に有意差が見られなかった．これらは原音声の話し言葉らしさの調査において，話し言葉音声と読み上げ音声の話し言葉らしさの差が小さい話者であることがわかっている [6].

4.2　ケプストラム，音素長，F_0 の影響に関する対比較実験

HMM 分析合成では，音素 HMM の学習データ，音素長モデルの学習データ，F_0 を抽出する入力音声の 3 つにそれぞれ話し言葉音声か読み上げ音声のどちらかを使用することにより，8 通りの音声を合成することができる．ケプストラム，音素長，F_0 が合成音声の話し言葉らしさに与える影響の強さを調査するため，各話者についてこの 8 通りの音声を合成し，ケプストラム，音素長，F_0 の 3 つのうち 1 つの使用データのみが異なる合成音声の組合せについて，どちらが話し言葉らしく聞こえるかを評価する対比較実験を行った．被験者には 8 通りの合成音声の中から，ケプストラムのみ使用データが異なる組合せを 4 ペア，音素長のみが異なる組合せを 4 ペア，F_0 のみが異なる組合せを 4 ペアの計 12 ペアについて，各 4 話者分ずつ，計 48 ペアの音声を提示し，それぞれどちらの音声がより話し言葉らしく聞こえるかを評価してもらった．被験者は 18 名である．

Fig.3 にケプストラム，音素長，F_0 それぞれについて，異なる発話スタイル (話し言葉，読み上げ) の音声を用いた場合の評価結果をプリファレンススコアで示す．各話者の結果は 48 回の評価によって計算されており，ALL は全話者に対する結果を表す．各 ALL のスコアについて有意水準 1%で検定を行ったところ，話し言葉音声を用いた場合の方が読み上げ音声を用いた場合よりもスコアが大きいことが確認された．この結果から，合成音声の話し言葉らしさにはケプストラム，音素長，F_0 のいずれも影響を与えていることが確認された．話者ごとに見た場合にはスコアにばらつきがあり，一部スコアの逆転や有意差が確認されないことも見受けられる．この点については今後検討していく必要がある．

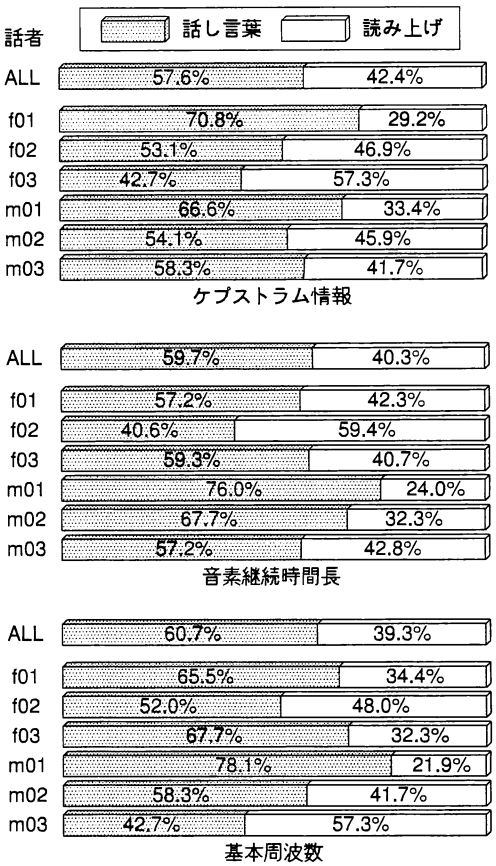

Fig. 3　ケプストラム，音素長，F_0 それぞれについて，異なる発話スタイルの音声を用いた合成音声に対する話し言葉らしさについての対比較実験の結果

5　まとめ

HMM 音声合成を用いた話し言葉音声合成の実現可能性について検討した．HMM 分析合成により話し言葉らしい音声を合成することは可能であり，また合成音声の話し言葉らしさにはケプストラム，音素長，F_0 のいずれも影響を与えていることが確認された．

今後の課題としては，話し言葉音声の F_0 情報のモデル化や HMM 音声合成法の改良による話し言葉音声の合成音の高品質化，話し言葉特有の発話であるフィラーの扱いについての検討などが挙げられる．

参考文献

[1] 益子 他，信学論，vol.J79-D-II, no.12, pp.2184-2190, 1996.
[2] 山田 他，情処研報，vol.2001, no.100, pp.15-20, 2001.
[3] K.Iwano et al. In S.Narayanan et al.(Eds.), Text to Speech Synthesis, Prentice Hall PTR, pp.155-173, 2004.
[4] 立和 他，音講論，2-3-7, 1999-3.
[5] 今井 他，信学論，vol.J66-A, no.2, pp.122-129, 1983.
[6] 赤川 他，信学技報，vol.105, no.98, pp.25-30, 2005.