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おけるモデルの構築とその合成音声への影響
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 2007年春季講演論文集, Vol. , No. 1-8-5, pp. 201-202
Citation(English)	, Vol. , No. 1-8-5, pp. 201-202
発行日 / Pub. date	2007, 3

HMMを用いた話し言葉音声合成におけるモデルの構築とその合成音声への影響*

◎赤川達也, 岩野公司, 古井貞熙 (東工大)

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．そこで本研究では，HMM 音声合成 [1] に基づくテキスト音声合成 (Text-to-Speech: TTS) システムを用いて話し言葉音声合成の実現を目指す．我々はこれまで，話し言葉音声のケプストラム情報 (以下，ケプストラム) と音素継続時間長 (音素長) の統計的モデル化を行い，基本周波数情報 (F_0) には原音声から抽出した値をそのまま用いて音声を合成することで，話し言葉音声合成システムの実現可能性について検討した [2]．本研究では F_0 についても話し言葉音声を用いてモデルを構築し，これを用いることで構成した話し言葉音声の TTS システムについてその性能や各モデルの合成音声への影響について調査する．

2 HMM 音声合成に基づく TTS システム

HMM 音声合成に基づく 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 [3, 4]．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のケプストラム列を生成し [5]，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 [6] 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 [3, 4]．本研究で用いる TTS システムも Fig.1 に示したものであるが，本研究ではテキスト解析部については扱わず，日本語話し言葉コーパス (CSJ) のイントネーションラベルから生成した言語情報を TTS システムの入力として用いる．

3 音声合成用モデルの作成

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 6 話者 (男性 3 名，女性 3 名) 分の学会講演音声とその再読み上げ音声をそれぞれ用いた．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．

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，一部を除き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．

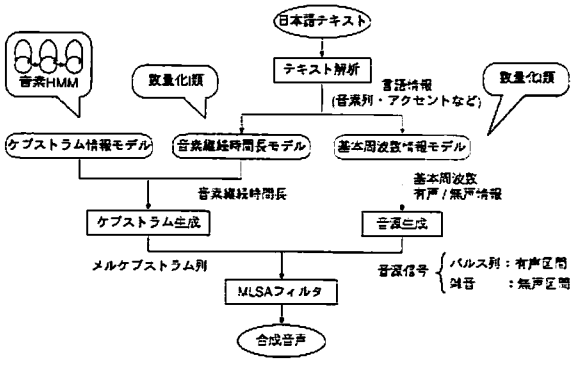

Fig. 1 HMM 音声合成による TTS システムの構成

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 (Hz) 平均値	標準偏差
f01	A06F0128	14,321	7.71	228.8	47.0
	R00F0407	14,664	6.91	212.8	39.9
f02	A05F0043	12,836	8.16	197.5	37.5
	R00F0028	13,111	7.83	191.4	30.2
f03	A01F0122	8,977	7.31	205.7	35.8
	R00F0178	8,998	7.26	191.1	31.8
m01	A11M0846	13,199	8.71	123.0	27.5
	R00M0036	14,082	7.29	115.1	21.7
m02	A01M0056	8,051	8.28	100.1	21.9
	R00M0187	8,242	7.64	89.4	15.8
m03	A11M0369	12,086	8.52	170.5	26.4
	R00M0134	12,452	8.17	179.3	27.2

3.2 モデルの学習

Table 1 に示した 6 話者 × 2 発話スタイル (話し言葉・読み上げ) の 12 音声それぞれについて，テストセット用にランダムに選んだ数文章を取り除き，これを学習データとして 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，数量化 I 類によって音素長モデルを作成した．数量化に用いた制御要因 (アイテム) は，先行/当該/後続音素の種類の 3 つである．F_0 情報も同様に数量化 I 類によってモデル化した．用いた制御要因は音素長モデルのものに加え，アクセント句のアクセント型，アクセント句間の結合の強さなど 18 種類である [3]．

4 合成音声の話し言葉らしさへの影響の調査

Fig.1 に示す TTS システムにより話し言葉らしい音声を合成できるかどうかを調査するため，ケプストラム，音素長，F_0 の 3 つのモデルを全て話し言葉音声または読み上げ音声で学習した場合についてそれぞれ音声を合成し，両者の話し言葉らしさを評価する対比較実験を行った (実験 1)．また，合成音声の話し言葉らしさに影響を与える要因を調査するため，

* Implementation and investigation of the models for HMM-based spontaneous speech synthesis. by Tatsuya Akagawa. Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology)

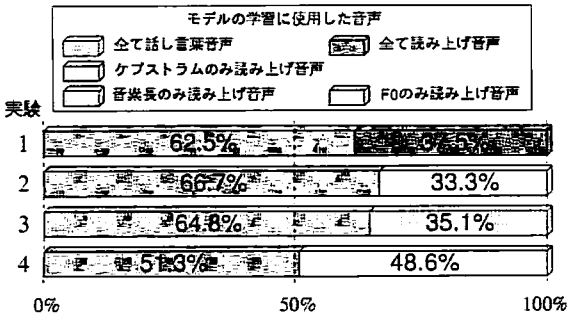

Fig. 2 ケプストラム，音素長，F_0 モデルの学習に読み上げ音声を用いた場合の合成音声の話し言葉らしさのプリファレンススコア

3つのモデル全てに話し言葉音声から学習したものを用いて合成した音声に対し，各モデルのうち1つのみに読み上げ音声から学習したものを用いて音声を合成し，その話し言葉らしさを評価する対比較実験を行った (実験 2〜4)．同一話者，同一文章でモデルの学習に用いた音声の発話スタイルのみが異なる合成音声を対にして 18 名の被験者にランダムな順に提示し，どちらがより話し言葉らしく聞こえるかを主観評価してもらった．

Fig.2 に各対比較実験についての評価結果をプリファレンススコアで示す．いずれの結果も全話者の結果を合計したものであり，それぞれ 108 回の評価により計算されたものである．各スコアについて有意水準 1%で検定を行ったところ，実験 1, 2, 3 においてスコアの間に有意差が確認された．この結果から，HMM 音声合成に基づく TTS システムを用いて話し言葉らしい音声を合成することができること，またその話し言葉らしさがケプストラム，音素長モデルによってもたらされ，今回用いたモデル化手法による F_0 モデルではその効果が小さいことがわかった．

5 各モデルのモデル化精度の調査

F_0 に原音声から抽出した値をそのまま用いた場合には，合成音声の話し言葉らしさに影響を与えていることが先行研究 [2] からわかっている．したがって前節の結果は，話し言葉音声の F_0 を十分にモデル化できていないことが原因であると考えられる．また，ケプストラム，音素長に関しては，より話し言葉らしい音声の合成に向けてそのモデル化の精度にどの程度向上の余地があるかを調査することが必要である．そこで，各モデルが話し言葉音声の特徴をどれだけモデル化できているかを調査するため，Fig.3 に示すように，ケプストラム，音素長，F_0 の全てに原音声から抽出した値を用いて合成した分析合成音声と，3つの特徴のうち1つのみにモデルからの推定値を用いて合成した音声を比較する評価実験を行った (実験 A, C)．ただし音素長に関しては原音声から抽出したケプストラム列に対してその音素長のみを変化させる操作が困難であるため，ケプストラム，音素長ともにモデルからの推定値を用いた合成音声を，ケプストラムのみに推定値を用いた合成音声と比較した (実験 B)．各ペアの合成音声を 18 名の被験者にランダムな順に提示し，どちらが話し言葉らしく聞こえるかを評価してもらった．

Fig.4 に，対比較実験 A〜C の評価結果をプリファレンススコアで示す．いずれの結果も全話者の結果を

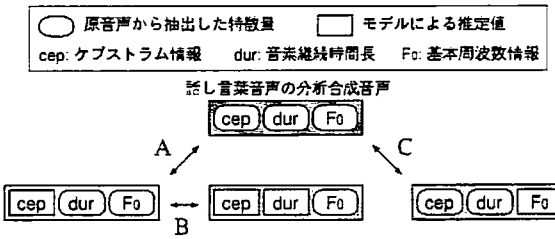

Fig. 3 3つの特徴のうち1つのみに推定値を用いた合成音声と，分析合成音声の比較

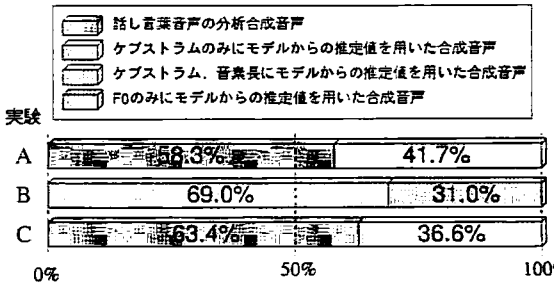

Fig. 4 一部にモデルからの推定値を用いた合成音声の話し言葉らしさのプリファレンススコア

合計したものであり，それぞれ 108 回の評価により計算されたものである．各スコアについて検定を行ったところ，実験 B, C は有意水準 1%でスコアの間に有意差が見られた．この結果は，ケプストラムに関しては話し言葉音声の特徴を十分にモデル化できているが，音素長と F_0 に関してはそのモデル化の精度に改善の余地があることを意味している．また音素長については，モデルの改善の余地があるにも関わらず 4 節の実験結果からは現状のモデルで話し言葉らしさがある程度反映されていることがわかる．これは発話スタイル間の特徴に大きな差があることに起因しているものと考えられる．

6 まとめ

HMM 音声合成に基づく TTS システムにおいて，合成音声の話し言葉らしさに影響を与える要因と各モデルのモデル化の精度について検討した．現状で用いている話し言葉音声のケプストラム，音素長モデルには話し言葉らしさの特徴が反映されており，F_0 モデルには十分に反映されていないことがわかった．また，話し言葉音声のケプストラムは十分な精度でモデル化できている一方で，音素長，F_0 モデルにはまだ改善の余地があることがわかった．今後の課題としては，高精度な音素長，F_0 のモデル化手法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] 赤川 他，音講論，3-6-17, 2005-9.
[3] 山田 他，情処研報，vol.2001. no.100, pp.15-20, 2001.
[4] K.Iwano et al. In S.Narayanan et al.(Eds.), Text to Speech Synthesis. Prentice Hall PTR, pp.155-173, 2004.
[5] 立和 他，音講論，2-3-7, 1999-3.
[6] 今井 他，信学論，vol.J66-A, no.2, pp.122-129, 1983.