HMM 音声合成における発話スタイルの制御
Speaking styles control in HMM-based speech synthesis

山岸 順一 † Junichi Yamagishi　**大西 浩二** † Koji Onishi　**益子 貴史** † Takashi Masuko　**小林隆夫** † Takao Kobayashi

1. はじめに

音声合成技術は自然なヒューマンインターフェイスを実現するために欠かせない要素技術のひとつである．自然なヒューマンインターフェイスを実現するためには，自由に話者の声質や韻律を変えたり，様々な発話スタイルや感情を表現できる音声合成システムが必要となる．

我々は多様な話者性，発話スタイルを容易に実現することができる音声合成システムの構築を目的として，音声単位を HMM(Hidden Markov Model) によりモデル化した HMM に基づく音声合成システムを提案してきた [1][2]．本システムの特徴は，動的特徴量を考慮して，HMM から直接音声の特徴パラメータを生成 [3] し，音声を合成する点である．また，音声単位として HMM を用いているため，HMM のパラメータを適切に変換することで合成音声の声質や韻律を変えることができる．実際に，話者適応手法を利用し，目標話者の少量の音声データを用いることで目標話者の声質や韻律に近似した合成音声を生成できることを示した [4]．この他にも固有声手法 [5] や話者補間手法 [6] なども声質の多様化手法として検討がなされている．発話スタイルの多様化に関しても，文献 [7] において複数の発話スタイルの音声が合成可能であることが示されている．しかしながら，複数の発話スタイルを自由に制御する手法については検討がなされておらず，声質の多様化と発話スタイルの多様化をともに考慮した手法についても検討がなされていなかった．

そこで本論文では，声質の多様化と発話スタイルの多様化をともに実現するため，話者適応も適用可能である発話スタイルの制御手法について比較，検討を行う．

2. HMM 音声合成システム

HMM 音声合成システム [1][2] のブロック図を図 1 に示す．システムは学習部，合成部から構成される．

学習部では音素単位の HMM を作成する．音声データベースからスペクトルパラメータとしてメルケプストラム，F_0 パラメータとして対数基本周波数を求め，フレーム毎に結合し，静的特徴量とする．得られたパラメータから動的特徴量を求め，静的特徴量と合わせて特徴パラメータとする．スペクトル部には通常の連続分布を，F_0 部には多空間上の確率分布を用いたマルチストリーム MSD-HMM[8] を用いてスペクトルと F_0 とを音素単位で同時にモデル化し，音素 HMM を得る．次に，この音素モデルを初期モデルとして先行・当該・後続音素，アクセント型，モーラ位置などのスペクトル，F_0 パターン，音韻継続長に影響を与える様々な変動要因 (コンテキスト) を考慮したコンテキスト依存モデルを学習する．次に，コンテキスト依存モデルに対し，コンテキストに関する質問を利用して決定木に基づくコンテキストクラスタリング [9][10] をスペクトル部，F_0 部に別々に適用し，出力分布を共有化する．最後に，HMM の各モデルの状態継続長を多次元ガウス分布でモデル化し，スペクトル，F_0 と同様に決定木に基づくコンテキストクラスタリングを適用する．

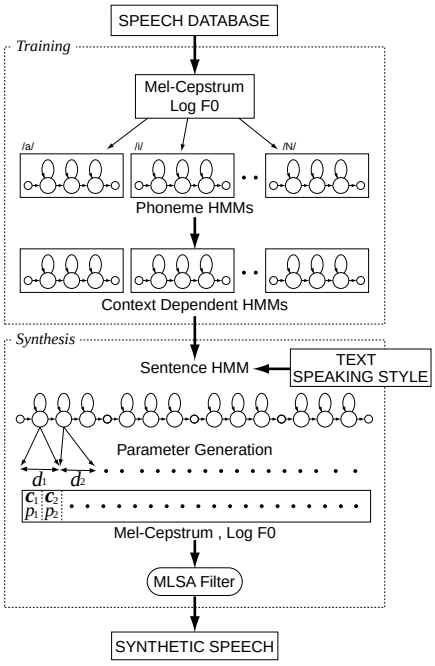


図 1: HMM に基づく音声合成システムのブロック図

合成部では，合成したいテキストと再現したい発話スタイルから韻律情報を含むコンテキスト依存ラベル列を生成する．得られたラベル列に従ってコンテキスト依存 HMM を結合し，一つの文章 HMM を構成する．状態継続長分布に従って各状態の継続長を決定し [1]，尤度最大基準に基づく HMM からのパラメータ生成アルゴリズム [3] により，文章 HMM からメルケプストラム列及び F_0 パターンを生成する．最後に，生成したメルケプストラム列と F_0 パターンを用いて，MLSA フィルタ [11] により音声を合成する．

3. HMM 音声合成における発話スタイルの制御

HMM 音声合成における発話スタイルの制御手法として，ここでは以下の二つについて考える．まず，単純に各発話スタイル毎に音響モデルを学習し，繋ぎ合わせる手法 [7] であり，この手法をスタイル依存モデル (Style Dependent Model) による制御手法と呼ぶことにする．これに対し，発話スタイルをコンテキストとして扱い，複数の発話スタイルを同時に学習する手法を考え，この

†東京工業大学 大学院総合理工学研究科

手法をスタイル混合モデル (Style Mixed Model) による制御手法と呼ぶことにする．

スタイル依存モデルによる制御手法では，複数の発話スタイルの音響モデルを個別に学習し，発話スタイル毎の決定木のルートへの枝 (パス) を持つ新たな決定木を作成する．新たな決定木のルートにおいて発話スタイル毎の決定木への枝を選択することにより発話スタイルを制御する．この手法では，新たな発話スタイルを加える際にはその発話スタイルの決定木のルートへの枝を追加するだけでよい．しかしながら，全ての発話スタイルに対して適応データが必要となるため，話者適応を行うことは容易ではないと考えられる．

スタイル混合モデルによる制御手法では，発話スタイルをコンテキストとして扱い，複数の発話スタイルを同時に学習する．決定木のノード分割に用いられる質問にも発話スタイルに関する質問が含まれているため，発話スタイルは他のコンテキストと同様に扱われ，決定木が作成される．この 2 分木の決定木により音素と発話スタイルの制御を行う．この手法では，新たな発話スタイルを加える際にはスタイル混合モデルを再学習しなければならないが，一つの発話スタイルの適応データのみでモデル全体の適応ができるため，全ての発話スタイルに対して話者適応が可能であると考えられる．

4. 実験

スタイル依存モデルによる制御手法ついてはおおむね学習した発話スタイルを再現できることが確認されている [7]．しかしながら，スタイル混合モデルについては発話スタイルの再現性の検討はなされていない．そこでスタイル混合モデルを用いた場合の発話スタイルの再現性および合成音声の自然性について評価を行った．

4.1 発話スタイル音声データベース

提案法による発話スタイルの制御法を比較するため，複数の発話スタイルを含む音声データベース [7] を使用した．発話スタイルには「読み上げ (丁寧)」,「ぞんざい」,「楽しげ」,「悲嘆」の 4 つを設定し，男性話者 1 名 (MMI) が ATR 音韻バランス 503 文をそれぞれの発話スタイルで発声した音声データを収録した．

男性話者 MMI の各発話スタイル別に収録した音声データが意図通りの発話スタイルにより発話されているかどうかを調べる予備実験を行った．被験者は 5 名で，評価は各発話スタイル別の収録音声の 503 文章について,「意図した発話スタイルに聞こえる」,「意図した発話スタイルには聞こえない」の二択の形式で行った．表 1 に予備実験の結果を示す．表 1 は過半数の被験者が「意図した発話スタイルに聞こる」と判断した文章数とその割合を示している．この結果より，おおむね意図した発話スタイルで発声されていることがわかる．ただし,「ぞんざい」,「楽しげ」に関しては「意図した発話スタイルには聞こえない」と判断された文章数が若干多くなっていることがわかる．

4.2 実験条件

HMM の学習には各発話スタイルの 450 文章，計 1800 文章を用いた．無音を含む 42 種類の音素を単位とし，コンテキスト情報の含まれるラベルを作成して学習に用いた．考慮しているコンテキストは以下の通りである．

表 1: 意図した発話スタイル通りに聞こえると判断された文章数

読み上げ (丁寧)	ぞんざい	楽しげ	悲嘆
503 (100%)	479 (95%)	491 (98%)	501 (99%)

表 2: クラスタリング後の分布数

	dependent					mixed
	読み上げ	ぞんざい	楽しげ	悲嘆	計	
Spec.	891	752	808	926	3377	2796
F_0	1316	1269	1368	1483	5436	4404
Dur.	1070	1272	1057	950	4349	3182

- 文の長さ
- 当該呼気段落の位置
- { 先行，当該，後続 } 呼気段落の長さ
- 当該アクセント句の位置，前後のポーズの有無
- { 先行，当該，後続 } アクセント句の長さ，アクセント型
- { 先行，当該，後続 } の品詞 (84 種類)，活用形 (38 種類)，活用型 (74 種類)
- 当該音素のアクセント句内でのモーラ位置
- アクセント位置とモーラ位置の差
- { 先行，当該，後続 } 音素
- 発話スタイル (4 種類)

ただし，ここでの長さ，位置の単位はモーラである．

サンプリングレート 16kHz の音声信号を，フレーム長 25ms，フレーム周期 5ms のブラックマン窓を用いてメルケプストラム分析 [12] し，0 次から 24 次のメルケプストラムを求めた．F_0 パラメータには対数基本周波数を用いた．これらのパラメータに，デルタおよびデルタデルタパラメータを加えた 78 次のベクトルを特徴ベクトルとし，5 状態の left-to-right HMM によりモデル化した．

MDL 基準によるコンテキストクラスタリングにより分布の共有を行った結果，両手法の分布数は表 2 のようになった．表 2 の “dependent” はスタイル依存モデル，“mixed” はスタイル混合モデルの結果を示し，“読み上げ”，“ぞんざい”，“楽しげ”，“悲嘆” はスタイル依存モデルの各発話スタイル毎の結果を示し，“計” は全発話スタイルの和を示す．この表より，スタイル混合モデルはスタイル依存モデルよりも約 3 割の分布数を削減できることがわかる．構築された決定木の一部を図 2，3 に示す．図 2，3 は各々スタイル依存モデル，スタイル混合モデルにより構築された F_0 部の 2 状態目の決定木である．

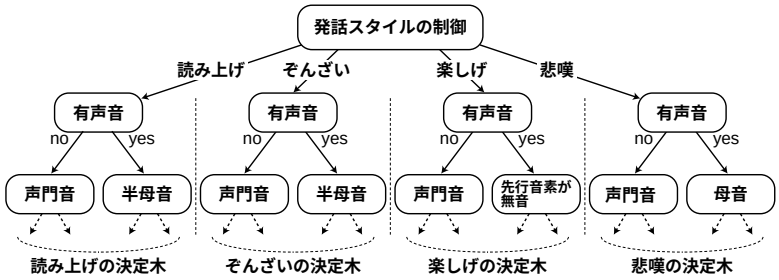


図 2: スタイル依存モデルにより構築された決定木の例

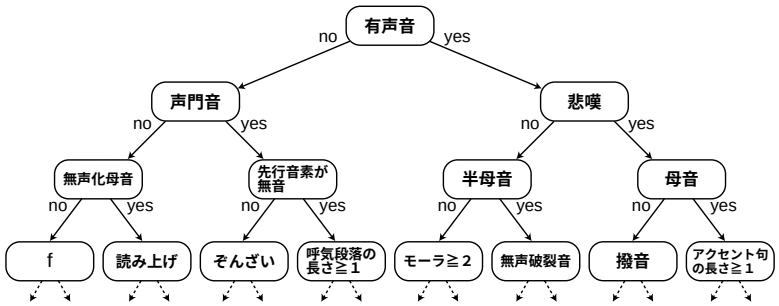


図 3: スタイル混合モデルにより構築された決定木の例

4.3 F_0 の生成例

図 4 に学習データに含まれない「人物の独白があれば空から見下ろすような描写がある」という文章に対して生成された F_0 パターンの例を示す．図 4(a) は読み上げ，(b) はぞんざい，(c) は楽しげ，(d) は悲嘆調に発話スタイルを設定した場合の結果である．また，図中の “natural” は収録音声から抽出された F_0 パターン，“dependent” はスタイル依存モデルから生成された結果，“mixed” はスタイル混合モデルから生成された結果を示す．この図より，スタイル混合モデルによる F_0 パターンはスタイル依存モデルの場合とほぼ同じであることがわかる．

4.4 発話スタイルの再現性

カテゴリーテストによる主観評価試験より，スタイル依存モデルとスタイル混合モデルから合成された音声の発話スタイルの再現性を評価した．男性話者 11 名の被験者に，各音声の発話スタイルが「読み上げ (丁寧)」「ぞんざい」「楽しい」「悲嘆」のどれに認識されたかを判定してもらった．ただし，どれにも当てはまらないと感じられた場合には「その他」と判定してもらった．テストデータは学習データに含まれていない 53 文章とし，被験者毎にランダムに 8 文章を選び，文章毎に全ての発話スタイルの音声を合成し，一文章につき順番をランダムに入れ替えて 2 回繰り返し評価を行った．

表 3 に発話スタイルの再現性の評価結果を示す．表 3 の数値は，各テスト音声がどの発話スタイルに認識されたかを割合で示し，(a) はスタイル依存モデルによる合成音声の結果，(b) はスタイル混合モデルによる合成音声の結果を示している．この表より，両手法とも同様の傾向を示しており，おおむね意図通りの発話スタイルに認識されていることがわかる．ただし，両手法とも「ぞんざい」の再現性がやや劣化していることがわかる．これは，収録音声において,「読み上げ (丁寧)」,「楽しげ」,「悲嘆」では音素境界が明確であるのに対し,「ぞんざい」ではやや音素境界が不明確となっており，そのことが合成音声にも影響していると考えられる．また，**4.1** の予備実験より,「ぞんざい」の収録音声のなかに若干ぞんざいと聞こえない音声データがあることがわかっており，そのことも合成音声に影響していると考えられる．

表 3: 発話スタイルの再現性の評価

(a) スタイル依存モデルによる合成音声の評価

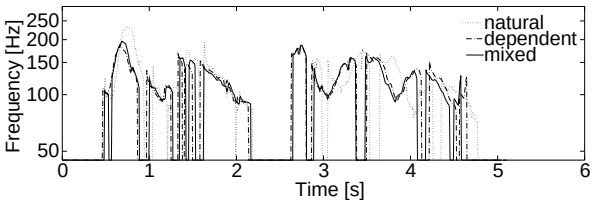
		判定結果 (%)				
		読み上げ	ぞんざい	楽しげ	悲嘆	その他
音声	読み上げ	98.3	0.6	0.0	0.0	1.1
	ぞんざい	6.9	82.3	0.0	0.0	10.8
	楽しげ	1.1	0.0	94.9	0.0	4.0
	悲嘆	0.6	1.1	0.0	94.9	3.4

(b) スタイル混合モデルによる合成音声の評価

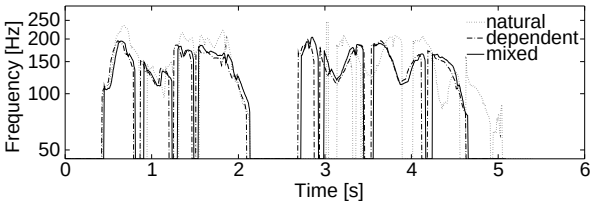
		判定結果 (%)				
		読み上げ	ぞんざい	楽しげ	悲嘆	その他
音声	読み上げ	98.9	0.0	0.0	0.0	1.1
	ぞんざい	2.8	89.8	0.0	1.1	6.3
	楽しげ	0.6	0.0	96.0	0.0	3.4
	悲嘆	0.0	0.6	0.0	96.0	3.4

4.5 発話スタイルの制御法の比較

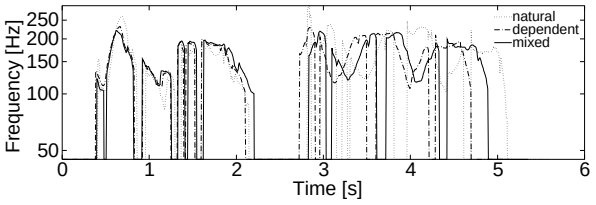
対比較による主観評価試験により，スタイル依存モデルとスタイル混合モデルから合成された同じ発話スタイルの音声の自然性について評価した．被験者は 16 名である．テストデータは学習データに含まれない 53 文章とし，被験者毎にランダムに 4 文章を選び，文章毎に全ての発話スタイルの音声を合成し，一文章につき順番を



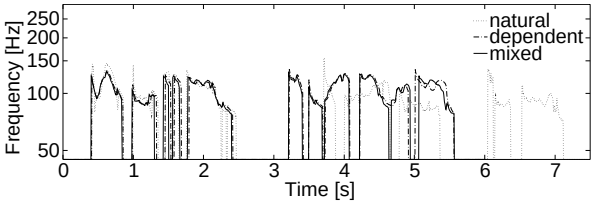
(a) 読み上げ



(b) ぞんざい



(c) 楽しげ



(d) 悲嘆

図 4: F0 パターンの生成例

ランダムに入れ替えて 2 回繰り返し評価を行った．

図 5 に比較結果を示す．図の横軸はプリファレンススコアを表し，“dependent”，“mixed” はそれぞれスタイル依存モデル，スタイル混合モデルのスコアを示す．棒グラフは上から「読み上げ (丁寧)」「ぞんざい」「楽しい」「悲嘆」の比較結果を示す．この図より，両手法により合成された同じ発話スタイルの音声の自然性はほぼ同等であることが分かる．

この実験結果と **4.4** の実験結果よりスタイル混合モデルによる発話スタイルの制御手法はスタイル依存モデルによる制御手法と同等の性能であることがわかる．しかしながら，スタイル混合モデルの分布数はスタイル依存モデルより少ないことを考慮すると，スタイル混合モデルによる発話スタイルの制御手法の方がより良いといえる．

4.6 発声途中での発話スタイルの切替

発声途中において発話スタイルのラベルを変えることにより，複数の発話スタイルを切り替えることができると考えられる．例えば,「その夫人は眼をこっちに向けているが見てはいない」と言いう文章の「眼をこっちに向けてはいる」の部分を楽しげ調で,「見てはいない」を悲嘆調で，残りの部分を読み上げ調の音声で合成すること

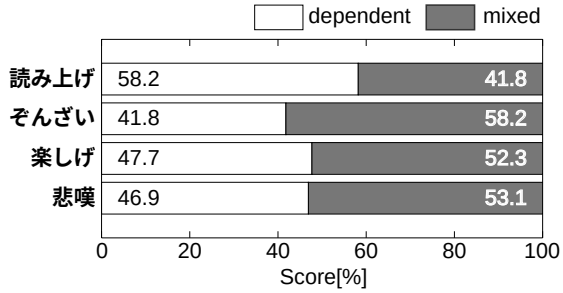


図 5: 発話スタイルの制御法の比較

が可能であると考えられる．非公式の受聴の結果，スタイル依存モデル，スタイル混合モデルのどちらの制御手法を用いても，意図通りの発話スタイルを制御できることが確認できた‡.

5. むすび

本論文では，話者適応も適用可能である発話スタイルの制御手法ついて，スタイル混合モデルとスタイル依存モデルの比較，検討を行った．主観評価試験により，スタイル混合モデルによる制御手法はスタイル依存モデルによる制御手法と同等の発話スタイルの再現性をより少ない分布数で実現できることを示された．今後の課題は，話者適応を用いて多様な話者の声質で様々な発話スタイルを表現する検討や，更に多くの発話スタイルでの音声合成である．

参考文献

[1] 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,” 信学論 (D-II), vol.J79-D-II, no.12, pp.2184–2190, Dec. 1996.

[2] 吉村 貴克, 徳田 恵一, 益子 貴史, 小林 隆夫, 北村 正, “HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化,” 信学論 (D-II), vol.J83-D-II, no.11, pp.2099–2107, Nov. 2000.

[3] 徳田 恵一, 益子 貴史, 小林 隆夫, 今井 聖, “動的特徴を用いた HMM からの音声パラメータ生成アルゴリズム,” 日本音響学会誌, vol.53, no.3, pp.192–200, Mar. 1997.

[4] 田村 正統, 益子 貴史, 徳田 恵一, 小林 隆夫, “HMM に基づく音声合成におけるピッチ・スペクトルの話者適応,” 信学論 (D-II), vol.J85-D-II, No.4, pp.545–553, Apr. 2002.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, “Speaker Interpolation for HMM-based Speech Synthesis System,” J. Acoust. Soc. Jap. (E), vol.21, pp.199–206 Apr. 2000.

[6] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, “Eigenvoices for HMM-based speech synthesis,” Proc. ICSLP-2002, p.1269–1272, Sep. 2002.

[7] 大西 浩二 , 益子 貴史, 小林 隆夫, “HMM音声合成における異なる発話スタイルの生成の検討,” 信学技報, vol.102, Jan. 2003.

[8] 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く HMM,” 信学論 (D-II), vol.J79-D-II, no.7, pp.1579–1589, July 2000.

[9] S. J. Young, J. Odell and P. Woodland, “Tree-based state tying for high accuracy acoustic modeling,” Proc. ARPA Human Language Technology Workshop, pp.307–312, Mar. 1994.

[10] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” J. Acoust. Soc. Jap. (E), vol.21, pp.79–86, Mar. 2000.

[11] 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,” 信学論 (A), vol.J66-A, no.2, pp.122–129, Feb. 1983.

[12] 徳田 恵一, 小林 隆夫, 深田 俊明, 斎藤 博徳, 今井 聖, “メルケプストラムをパラメータとする音声のスペクトル推定,” 信学論 (A), vol.J74-A, no.8, pp.1240–1248, Aug. 1991.

‡http://sp-www.ip.titech.ac.jp/research/demo/ に合成音声のサンプルがある．