多様な話者性および発話スタイル・感情表現による音声合成のための韻律生成

Generation of F0 Contours for Speech Synthesis with Various Speakers' Voices and Styles

東京工業大学大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

小林 隆夫
Takao Kobayashi

＜研究協力者＞
東京工業大学大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
* Currently with Tohshiba Corporation

益子 貴史　　山岸 順一
Takashi Masuko*　Junichi Yamagishi

This paper describes several approaches to realizing speaker and style variability including emotional expressivity in text-to-speech synthesis. We first discuss a training method of average voice model for speech synthesis in which arbitrary speaker's voice is generated based on speaker adaptation. To reduce the influence of speaker dependence, we incorporate a context clustering technique called shared decision tree context clustering and speaker adaptive training into the training procedure of average voice model. Then we investigate two methods for modeling speaking styles and emotional expressions, called style dependent modeling and style mixed modeling, based on an HMM-based speech synthesis framework. We also propose a technique for synthesizing speech with an intermediate speaking style or emotional expression from given style models based on a model interpolation technique of HMMs. Finally, we present style adaptation which is a technique for generating speech with a desired speaking style or emotional expression based on a model adaptation technique of style models using a small amount of speech data of the target style. From results of subjective experiments, we show the effectiveness of the proposed approaches.
Key words: HMM-based speech synthesis, average voice, speaking style, emotional expression

1　研究の目的

テキストから音声を合成する技術はヒューマンコンピュータインタラクションを実現するために欠かせない技術の一つである．音声を用いたヒューマンインターフェイスが人間にとって違和感がなく自然であるためには，合成音声の品質が自然であると同時に，自由に合成音声の話者性を変えたり，様々な発話スタイルや感情を表現できることが必要となる．合成音声の品質に関しては，最近の音声合成手法の主流となっている大量の音声コーパスに基づいた素片接続方式により，自然性の高い合成音声が実現されつつある．しかしながら，音声合成における多様な話者性，発話スタイル・感情表現に関しては，いまだに実用的なシステムは実現されておらず，今後の課題として残されたままとなっている．このような背景から，本研究では，多様な話者性および発話スタイル・感情表現による音声合成の実現をめざし，そのための統計的モデル（隠れマルコフモデル）に基づいた韻律生成手法の確立と音声合成システムの開発を研究目的とした．具体的には

1. 多様な話者性による音声合成を実現するために必要となる任意話者の声質・韻律の生成手法
2. 様々な発話スタイルや感情を表現するための韻律生成手法
3. 統計的モデル化における基盤技術となる基本周波数 (F0) 抽出手法

の各項目について検討を行った．

まず，任意話者の声質および韻律特徴による音声の合成手法に対し，かねてより提案してきた平均声

モデルと話者適応に基づく手法について種々の検討を行った．任意話者の音声を合成するには，対象となる話者の少量の発声データを用いて話者適応技術により平均声モデルをモデル適応した後，隠れマルコフモデル (HMM) 音声合成 [1] に基づいて韻律およびスペクトルパラメータの生成を行う．本研究では，モデル適応手法として，最尤線形回帰 (MLLR) に基づいたスペクトルおよび韻律モデルの適応手法 [2] を開発すると同時に，平均声モデルを学習する際，各話者の音声データが大量に存在しない場合にも合成音声の自然性を劣化させないモデル学習法として共有決定木コンテキストクラスタリング (STC) 手法 [3] を，さらに話者適応学習を組込んだ学習法 [4] を提案し，その有効性を示した．

次に，多様な発話スタイルや感情を表現するための韻律生成手法の開発では，まず男女各 1 名が「丁寧」「ぞんざい」「楽しげ」「悲嘆」の 4 種類のスタイルにより読み上げた 503 文章からなるスタイル音声データベースを作成した [5]．そして，これを用いて HMM 音声合成のための二つのスタイルモデルリング手法を提案し，その評価を行った [6]．さらに，多様なスタイルを実現する韻律・スペクトル生成手法として，スタイル補間手法 [7] ならびにスタイル適応手法 [8] を提案した．スタイル補間手法では，あるスタイルから他のスタイルに滑らかに変化する音声を合成できるスタイルモーフィング技術を開発した．またスタイル適応手法では，あるスタイルによる少量の発声データが与えられた際に，読上げ調のモデルからそのスタイルへモデル適応することにより，そのスタイルで任意のテキストに対応する音声が合成できることを示した．

一方，高精度な F0 抽出抽出手法の確立では，瞬時周波数振幅スペクトルの調波構造を利用し，F0 抽出を行う手法 [9] を開発した．信頼性の高い統計的モデルを自動構築するには，大量の F0 データを，種々の音声データベースから自動的に精度良く抽出することが求められる．これに対して提案手法は，F0 を求める際に用いる周波数帯域においてどれだけ明確な調波構造を成しているかを示す尺度となる調波構造指数を定義し，これに基づいて適切な周波数帯域及び分析窓長を自動選択，高精度に F0 を抽出するものであり，有声区間から無声／無声区間への切り替り部分での抽出精度を従来法に比べ向上させることが可能であることを明らかにした [10].

これらの研究成果の中から，ここでは平均声モデル学習法，スタイルモデリング，スタイル補間およびスタイル適応について説明する．

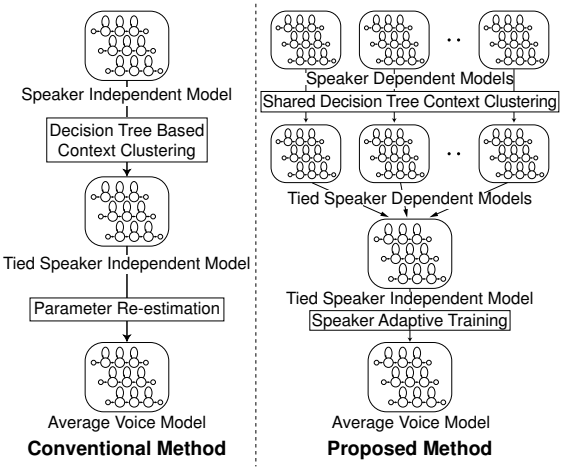


図 1: 平均声モデルの学習法のブロック図.

2 HMM 音声合成のための平均声モデル学習法

2.1 平均声に基づく任意の話者性による音声合成

HMM 音声合成では，合成に用いられる音声単位が隠れマルコフモデル (HMM) によりモデル化されており [11]，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・韻律特徴を変えることができる．本研究では話者適応の初期モデルとして平均声モデルを用いることを考える．平均声モデルとは，HMM 音声合成において，複数話者の音声データベースから学習された音声単位 HMM のことであり，これを用いて合成された音声は複数話者の平均的な声質および韻律特徴を持つと考えられることから，これを平均声と呼んでいる．任意話者の音声を合成するには，対象となる話者の少量の発声データを用いて平均声モデルを話者適応技術によりモデル適応した後，HMM 音声合成に基づいて韻律およびスペクトルパラメータの生成を行う [2].

平均声モデルを話者適応することを考えると，平均声モデルの各分布は全学習話者に対して平均的な分布になっていることが望ましい．しかし，平均声モデルの各分布の学習データ量は各学習話者に対して均一ではなく，分布に話者や性別の偏りが生じることがある．このことは平均声の品質や平均声モデルの適応性能および話者適応後の音声の品質に大きく影響する．そこで，話者による変動の影響を低減し，平均声の品質や平均声モデルの適応性能および話者適応後の音声の品質を向上させるため，共有決定木コンテキストクラスタリングと話者適応学習 (SAT) を併用した平均声モデルの学習法 [4] を導入する．

提案法による平均声モデルの学習法のブロック図を図 1 に示す．まず，複数話者の音声データベース

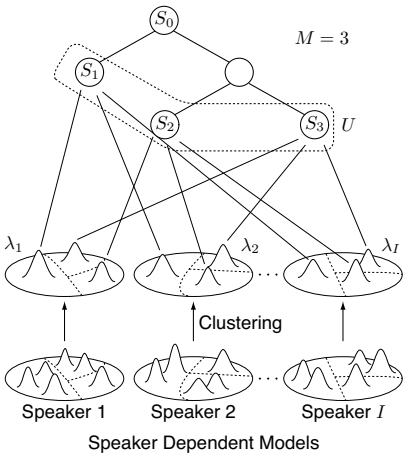


図 2: 共有決定木コンテキストクラスタリング

から話者毎に話者依存モデルを学習し，これらの話者依存モデルの共有決定木に基づいたコンテキストクラスタリングを行う．決定木のノード分割の終了後，平均声モデルのガウス分布は各話者依存モデルのガウス分布から計算する．全学習話者のデータを用いた話者適応学習 (SAT) による パラメータ再推定を行った後，平均声モデルを用いて話者毎に状態継続長分布を求める．最後に，同様のクラスタリング手法で平均声モデルの状態継続長分布を求める．

2.2 共有決定木コンテキストクラスタリング

平均声モデルの決定木のルートノードを S_0 とし，リーフノードの集合 $\{S_1, S_2, \ldots, S_M\}$ により定義されるモデルを $U(S_1, S_2, \ldots, S_M)$ とする (図 2 参照). ノード S_m に対応する話者 i のガウス分布を $\mathcal{N}_{im}$ とし，ノードの集合 $\{S_1, S_2, \cdots, S_M\}$ に対応する話者 i のガウス分布の集合を $\lambda_i(S_1, S_2, \cdots, S_M) = \{\mathcal{N}_{i1}, \mathcal{N}_{i2}, \ldots, \mathcal{N}_{iM}\}$ と定義する．

モデル U のノード S_m が質問 q により二つのノード S_{mqy} と S_{mqn} に分割されることで得られるモデルを U' とする．このとき，分割前後の記述長をそれぞれ $\widehat{D}(U)$, $\widehat{D}(U')$ とすると，その差分は次式で与えられる [3].

$$\begin{aligned}\delta_m(q) &= \widehat{D}(U') - \widehat{D}(U) \\ &= \frac{1}{2}\sum_{i=1}^{I}(\Gamma_{imqy}\log|\boldsymbol{\Sigma}_{imqy}| + \Gamma_{imqn}\log|\boldsymbol{\Sigma}_{imqn}| \\ &\quad - \Gamma_{im}\log|\boldsymbol{\Sigma}_{im}|) + c\sum_{i=1}^{I} K\log W_i \end{aligned} \tag{1}$$

ここで I は話者の総数，$\boldsymbol{\Sigma}_{imqy}$ と $\boldsymbol{\Sigma}_{imqn}$ はそれぞれノード S_{mqy} と S_{mqn} に対応する話者 i のガウス分布の共分散行列，Γ_{imqy} と Γ_{imqn} はそれぞれ話者 i の学習データ中にノード S_{mqy} と S_{mqn} が出現する頻度の期待値，c はモデルサイズを調節するための重み係数，K はデータベクトルの次元，Γ_{im} を話者 i の学習データ中にノード S_m が出現する頻度として $W_i = \sum_{m=1}^{M}\Gamma_{im}$ である．

そこで，決定木を以下の手順で構築する．まず，ルートノード S_0 から構成される集合をモデル U と定義し，$\delta_m(q)$ を最小にするモデル U のノードと質問を選び出し，選び出したノードを $S_{m'}$，質問を q' とおく．そして，$\delta_{m'}(q') > 0$ ならば分割を終了し，そうでなければノード $S_{m'}$ を質問 q' で分割し，その結果得られるノード集合を U と置き換え，上記の操作を繰り返す．ただし，全ての話者依存モデルに対して分割を行える質問のみを採用することにより，決定木の各ノードに対して必ず全ての話者のデータが存在するようにする．

決定木のノード分割の終了後，平均声モデルのガウス分布は各話者依存モデルのガウス分布から計算する．ノード S_m における平均声モデルの平均ベクトル $\boldsymbol{\mu}_m$，共分散行列 $\boldsymbol{\Sigma}_m$ は以下の式で求められる．

$$\boldsymbol{\mu}_m = \frac{\sum_{i=1}^{I}\Gamma_{im}\boldsymbol{\mu}_{im}}{\sum_{i=1}^{I}\Gamma_{im}} \tag{2}$$

$$\boldsymbol{\Sigma}_m = \frac{\sum_{i=1}^{I}\Gamma_{im}\left(\boldsymbol{\Sigma}_{im} + \boldsymbol{\mu}_{im}\boldsymbol{\mu}_{im}^{\top}\right)}{\sum_{i=1}^{I}\Gamma_{im}} - \boldsymbol{\mu}_m\boldsymbol{\mu}_m^{\top} \tag{3}$$

ここで，$\top$ は転置を表し，$\boldsymbol{\mu}_{im}$ はノード S_m，話者 i のガウス分布の平均ベクトルを表す．

2.3 話者適応学習による平均声モデルの学習

話者適応学習 (SAT) では，最尤線形回帰 (MLLR) に基づく話者適応を用いて，各学習話者へ適応したときの尤度が最大となるようにモデルの学習をする．

平均声モデルの状態 m の平均ベクトルを $\bar{\boldsymbol{\mu}}_m$ として，これを話者 i に変換した平均ベクトルを

$$\tilde{\boldsymbol{\mu}}_{im} = \boldsymbol{A}_i\bar{\boldsymbol{\mu}}_m + \boldsymbol{b}_i \tag{4}$$

と表す．ここで，$\boldsymbol{\xi}_m = [1, \boldsymbol{\mu}_m^{\top}]^{\top}$，$\boldsymbol{W}_i = [\boldsymbol{b}_i\boldsymbol{A}_i]$ は平均ベクトルの適応のための回帰行列である．そして，EM アルゴリズムに基づいて，回帰行列の推定と，得られた回帰行列を用いた平均ベクトルおよび共分散行列の推定を収束するまで繰り返す．このとき，話者 i の学習データ $\boldsymbol{O}_i = \{\boldsymbol{o}_{i1}, \boldsymbol{o}_{i2}, \ldots, \boldsymbol{o}_{iT_i}\}$ に対し，SAT による状態 m の平均ベクトルおよび共分散行列の最尤推定値 $\bar{\boldsymbol{\mu}}_m$，$\bar{\boldsymbol{\Sigma}}_m$ は次式で与えられる．

$$\begin{aligned}\bar{\boldsymbol{\mu}}_m = &\Big(\sum_{i=1}^{I}\sum_{t=1}^{T_i}\gamma_{im}(t)\boldsymbol{A}_i^{\top}\boldsymbol{U}_m^{-1}\boldsymbol{A}_i\Big)^{-1} \\ &\times\Big(\sum_{i=1}^{I}\sum_{t=1}^{T_i}\gamma_{im}(t)\boldsymbol{A}_i^{\top}\boldsymbol{U}_m^{-1}(\boldsymbol{o}_{it} - \boldsymbol{b}_i)\Big)\end{aligned} \tag{5}$$

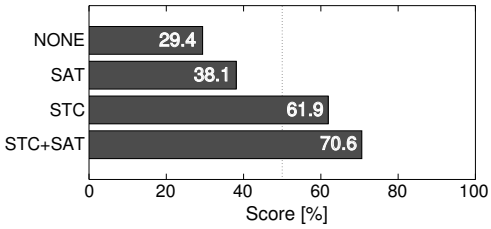


図 3: 平均声の自然性の評価

$$\bar{\boldsymbol{\Sigma}}_m = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T_i}\gamma_{im}(t)(\boldsymbol{o}_{it}-\tilde{\boldsymbol{\mu}}_{im})(\boldsymbol{o}_{it}-\tilde{\boldsymbol{\mu}}_{im})^\top}{\sum_{i=1}^{I}\sum_{t=1}^{T_i}\gamma_{im}(t)} \quad (6)$$

ここで，$\gamma_{im}(t)$ は時刻 t の観測ベクトル $\boldsymbol{o}_{it}$ が状態 m において出力される確率を表す.

2.4 評価実験

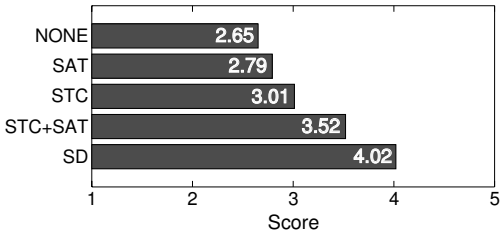
HMM の学習データとして，ATR 日本語音声データベースセット B を用いた．無音を含む 42 種類の音素を単位とし，コンテキスト情報の含まれるラベルを作成して学習に用いた．サンプリングレート 16kHz の音声信号を，フレーム長 25ms，フレーム周期 5ms のブラックマン窓を用いてメルケプストラム分析し，0 次から 24 次のメルケプストラムを求めた．また，対数基本周波数を F0 パラメータとした．これらのパラメータに，デルタおよびデルタデルタパラメータを加えた 78 次元のベクトルを特徴ベクトルとし，5 状態の left-to-right HMM によりモデル化した.

HMM の学習には男性話者 3 名，女性話者 3 名，計 6 名の各話者異なる 150 文章を用いた．SAT に用いる回帰行列は話者毎に一つとし，回帰行列の再推定は行っていない．また，比較のため，共有決定木コンテキストクラスタリングのみを適用したモデルと SAT のみを適用したモデルを作成した.

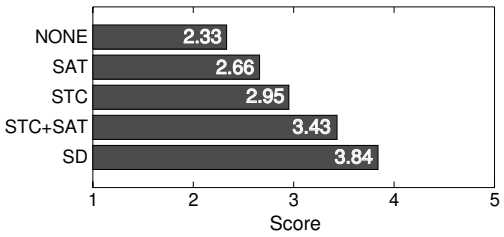
話者適応は学習データに含まれていない女性話者 FTK と男性話者 MMY を目標話者とし，各々10 文章を用いて MLLR による話者適応を行った．なお，継続長分布の話者適応は行わず，適応モデルの継続長分布には平均声モデルの継続長分布をそのまま用いている.

まず，対比較による主観評価試験により，各々のモデルから合成された平均声の自然性を評価した結果を図 3 に示す.．被験者は 9 名で，防音室でのヘッドホンによる両耳受聴により評価を行った．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 5 文章を選び，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.

図より，共有決定木コンテキストクラスタリングを適用したモデル (STC+SAT) は従来法のモデル (NONE) [2] および SAT のみを適用したモデル (SAT) よりも自然性が高いと評価されていることがわかる．更に，共有決定木コンテキストクラスタリングと SAT を併用することで平均声の自然性はより高くなることがわかる．提案法では話者や性別の偏りの影響を低減するため，特に平均声の韻律が改善され，自然性が向上したと考えられる.



(a) 男性話者 MMY



(b) 女性話者 FTK

図 4: 従来法と提案法の話者性の比較

次に，主観評価試験により，話者適応後のモデルから合成された音声の話者性を評価した結果を図 4 に示す．被験者は 7 名で，目標話者の分析合成音を基準に各音声の話者性を「5：非常によく似ている」から「1：似ていない」の 5 段階で評価してもらった．また比較のため，目標話者 MMY および FTK の 450 文章を用いて特定話者モデルを作成した．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 8 文章を選び，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.

図より，共有決定木コンテキストクラスタリングを適用したモデルから合成された音声は，従来法のモデルおよび SAT のみを適用したモデルよりも目標話者に近いと評価されていることがわかる．また，共有決定木コンテキストクラスタリングと SAT を併用することで合成音声の話者性は更に目標話者に近くなることがわかる．更に，共有決定木コンテキストクラスタリングと SAT を併用したモデルから合成される音声の目標話者の分析合成音に対する話者性は，話者依存モデル (SD) から合成される音声に近いものであることがわかる.

3 発話スタイル・感情表現のモデル化

3.1 スタイル依存モデルとスタイル混合モデル

これまで，平均声モデルから話者適応に基づいて任意の話者性を持った音声の合成が可能になること，話者補間 [12] を用いて様々な声質の音声を合成できることが明らかにされている．一方，発話スタイルや感情表現に関してもこの話者適応や話者補間の「話者」を「発話スタイル」や「感情表現」に置き換えることによって，様々なスタイルや感情を持った音声を合成できることが期待される．そこで，以下では様々な発話スタイル・感情表現を「スタイル」と呼び，音声に現れるスタイルのモデル化と生成について検討した結果を述べる．

まず，HMM 音声合成のためのスタイルモデルリング手法として二つの方法を提案し，その評価を行った．すなわち，単純に各スタイル毎に音響モデルを学習し，繋ぎ合わせる手法であるスタイル依存モデリング [5] と，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スタイル混合モデリング [6] である．

スタイル依存モデリングでは，複数のスタイルの音響モデルを個別に学習し，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スタイル毎の決定木への枝を選択することによりスタイルを制御する．この手法では，新たなスタイルを加える際にはその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という利点がある．

これに対し，スタイル混合モデリングでは，スタイルをコンテキストとして扱い，複数の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スタイルに関する質問が含まれているため，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スタイルの制御を行う．この手法では，新たな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，スタイル間で類似したパラメータの共有が行われるため，より精度の良いコンパクトなモデル化が期待できる．

3.2 音声データベース

実際の音声には様々なスタイルが含まれるが，それらをすべて収録することは容易ではない．ここでは，多様なスタイル音声合成に向けた第一歩として「丁寧」,「ぞんざい」,「楽しげ」,「悲嘆」の 4 種類を設定した．これは「丁寧」と「ぞんざい」,「楽しげ」と「悲嘆」でそれぞれ対比がとれており，比較がしやすいと考えたためである．この他に比較のため，通常の「読上げ」スタイルの音声も用いた．男性話者 1 名 (MMI) と女性話者 1 名 (FTY) に ATR 音韻バランス 503 文をそれぞれのスタイルで発声するように指示し，音声データを収録した．

表 1: 意図したスタイルと判定された文章数

話者	丁寧	ぞんざい	楽しげ	悲嘆
MMI	503	493	499	502
FTY	503	498	502	502

表 2: クラスタリング後のモデルの分布数.

(a) 男性話者 MMI

	スタイル依存					スタイル混合
	読上げ	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

(b) 女性話者 FTY

	スタイル依存					スタイル混合
	読上げ	ぞんざい	楽しげ	悲嘆	計	
Spec.	698	635	735	680	2748	2269
F_0	1464	1545	1343	1249	5601	4598
Dur.	1033	1407	1531	1105	5076	3801

音声データ収録に使用した文章の内容が，指示したスタイルにそぐわないと考えられる場合があることから，各スタイル毎に収録した音声データがそのスタイルにより発話されているかどうかを調べる主観評価実験を行った．被験者は男性 9 名で，各発話スタイル別の収録音声の全ての文章について，そのスタイルに聞こえるかどうかを判定した．表 1 に主観評価実験の結果を示す．表中の数字は，各スタイルの 503 文中，過半数の被験者がそのスタイルに聞こえると判断した文章数を示している．この結果より，男性話者 MMI，女性話者 FTY ともにおおむね指示したスタイルで発声されていることが確認された．

なお，同時に収録した「読上げ」調の収録音声に関して聴取実験を行ったところ,「丁寧」の音声が「読上げ」と認識された割合が，男性話者 MMI では約 38%，女性話者 FTY では約 39%，逆に「読上げ」の音声が「丁寧」と認識された割合が，男性話者 MMI では約 42%，女性話者 FTY では約 43%となった．この結果より，男性話者 MMI，女性話者 FTY ともに「読上げ」の音声と「丁寧」の音声は明確な区別ができないということで，スタイルのモデル化の評価には「読上げ」を用いることにした．

3.3 スタイルモデリングの評価

各スタイル 503 文中，450 文を用いてスタイルのモデル化を行った．音響分析および HMM の構成は，

表 3: スタイルの再現性の評価

(a) 男性話者 MMI

スタイル依存	判定結果 (%)				
	読上げ	ぞんざい	楽しげ	悲嘆	その他
読上げ	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

スタイル混合	判定結果 (%)				
	読上げ	ぞんざい	楽しげ	悲嘆	その他
読上げ	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

(b) 女性話者 FTY

スタイル依存	判定結果 (%)				
	読上げ	ぞんざい	楽しげ	悲嘆	その他
読上げ	92.5	1.9	5.0	0.0	0.6
そんざい	3.1	85.6	1.3	9.4	0.6
楽しげ	8.8	0.0	90.6	0.0	0.6
悲嘆	3.8	6.9	0.0	88.7	0.6

スタイル混合	判定結果 (%)				
	読上げ	ぞんざい	楽しげ	悲嘆	その他
読上げ	90.0	1.9	7.5	0.6	0.0
ぞんざい	0.6	90.0	0.0	8.1	1.3
楽しげ	3.1	1.9	92.5	0.0	2.5
悲嘆	1.3	5.6	0.0	91.8	1.3

前節の平均声の学習の場合と同じである．

MDL 基準を用いた決定木に基づく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後の各スタイルのモデルの分布数を表 2 に示す．この表より，スタイル混合モデルの方がスタイル依存モデルより分布の総数が少なくてすむことがわかる．

各スタイルモデルから HMM 音声合成により生成された合成音声のスタイル再現性について評価した結果を．表 3 に示す．被験者は 11 名で，評価は各評価用音声について,「読上げ」,「ぞんざい」,「楽しげ」,「悲嘆」,「その他 (どれにも当てはまらない)」のスタイルから一つを選択してもらった．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 8 文章を選び，文章毎にスタイルの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．

いずれのスタイルモデリング手法においても，男性話者，女性話者ともにおおむね意図した発話スタイルどおりに認識される結果となった．なお,「ぞんざい」は「その他」と認識されることが多かったが，これは収録音声について同じ実験を行った場合，同様の傾向が見られたことから，合成音声には収録音声と同様のスタイルが反映されていると言える．また，スタイル依存モデルとスタイル混合モデルの合成音声に対し対比較試験を行ったところ，いずれのスタイルに対してもほぼ同等のスコアが得られた．

4 スタイル補間

4.1 モデル補間に基づくスタイルの補間

話者補間手法 [12] をスタイルモデルに応用したスタイル補間の検討を行った結果を述べる．補間手法としては，分布の補間に基づく手法，カルバック情報量に基づく手法なども考えられるが，ここでは出力ベクトルの補間に基づく手法を用いる [7].

N 種類のスタイル $S_1, S_2, \ldots, S_N$ のモデル $\lambda_1, \lambda_2, \ldots, \lambda_N$ から，補間により新たなスタイル $\tilde{S}$ のモデル $\tilde{\lambda}$ を求めることを考える．スタイルの補間比率を $a_1, a_2, \ldots, a_N$ (ただし，$\sum_{k=1}^{N} a_k = 1$) とすると，補間後のスタイルの特徴ベクトル $\tilde{\boldsymbol{o}}$ は各スタイルの特徴ベクトル $\boldsymbol{o}_k$ を線形的に補間することによって求められる．このとき，スタイル S_k の特徴ベクトル $\boldsymbol{o}_k$ の分布としてガウス分布を仮定すると，新たなスタイル S の特徴ベクトル $\boldsymbol{o}$ の分布の平均ベクトル $\tilde{\boldsymbol{\mu}}$ および共分散行列 $\tilde{\boldsymbol{U}}$ は

$$\tilde{\boldsymbol{\mu}} = \sum_{k=1}^{N} a_k \boldsymbol{\mu}_k, \qquad \tilde{\boldsymbol{U}} = \sum_{k=1}^{N} {a_k}^2 \boldsymbol{U}_k \tag{7}$$

で与えられる．

各スタイルのモデルの構造が等しければ直接モデル間で補間することも容易であるが，一般にはスタイル毎にコンテキストクラスタリング後の分布の共有構造が異なるため，共有構造も含めて補間することは困難であると考えられる．そこで，音声合成時に動的に分布列の補間を行うことによりモデル補間を実現する．具体的には，N 個のモデルを補間するとき，まず，与えられたコンテキストラベル列に従って各スタイルモデルからそれぞれ文 HMM を作成する．次に，得られた N 個の文 HMM のメルケプストラム，F0，状態継続長の各分布を式 (7) に従って補間し，新たな文 HMM を作成する．そして，この文 HMM からメルケプストラムおよび F0 パラメータを生成することにより，補間して得られた新たなスタイルの音声を合成する．

4.2 スタイル補間の評価

前節で述べた「読上げ」,「楽しげ」,「悲嘆」の各スタイルモデルを用いて実験を行った．スタイル補間により各スタイルの中間的なスタイルを生成することが可能であることを示すため，以下の 3 種類の組合わせ「(読上げ，楽しげ)」,「(読上げ，悲嘆)」,「(楽しげ，悲嘆)」に対して，式 (7) によりスタイル補間

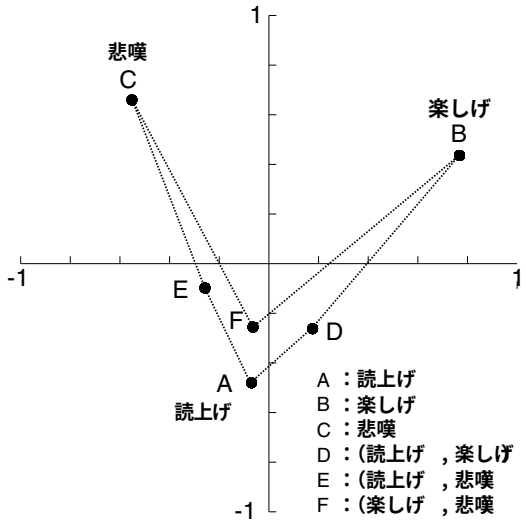


図 5: スタイルの類似度の評価 (男性話者 MMI).

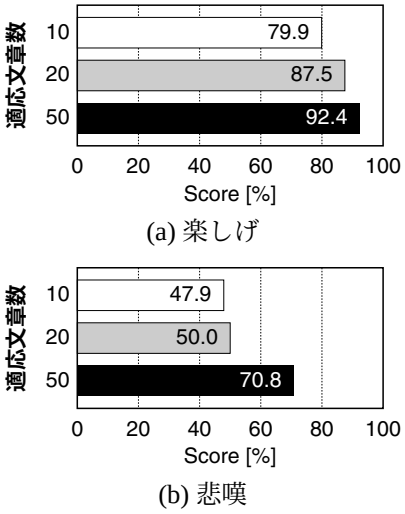


図 6: ABX 法による主観評価結果.

をし，得られたモデルから音声を合成した．ここで，スタイル A とスタイル B を補間することによって得られた新たなスタイルを，(A，B) で表すことにする．なお，補間比率はすべて $1:1$ $((a_A, a_B) = (0.5, 0.5))$ とした．

主観評価実験により，A:「読上げ」，B:「楽しげ」，C:「悲嘆」，D:「(読上げ，楽しげ)」，E:「(読上げ，悲嘆)」，F:「(楽しげ，悲嘆)」，のスタイルの合成音声の類似度を評価した．被験者は 8 名で，上記の 6 種類のスタイルの合成音声から二つを聞かせ，それらのスタイルの類似度を「5：非常によく似ている」から「1：全く似ていない」の 5 段階で評価した．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，文章毎にスタイルの組合せの順番をランダムに入れ替えて評価を行った．

主観評価実験で得られたスコアを数量化 4 類により分析し，各スタイルを類似度を表す 2 次元平面上に配置したものを図 5 に示す．図は男性話者 MMI の結果である．

この図より,「(読上げ，楽しげ)」と「(読上げ，悲嘆)」は，それぞれ補間元の 2 つの発話スタイルの間に配置されていることがわかる．よって，補間元のスタイルの組合せによっては，モデル補間の手法をスタイルに用いることにより，中間的なスタイルの音声を合成することができると考えられる．また,「読上げ」と「(楽しげ，悲嘆)」が近くに配置されていることから,「楽しげ」と「悲嘆」の中間のスタイルが「読上げ」であると考えることができる．なお，女性話者 FTY に対しても同様の結果が得られている．

5 スタイル適応

5.1 MLLR に基づくスタイル適応

話者適応に用いたモデル適応手法をスタイルモデルに適用し，スタイル適応の検討を行った結果を述べる．

一般に適応データ量は少量であるため，幾つかの状態の分布間で回帰行列を共有することで，適応データに存在しない状態に対する適応も可能としている．文献 [2] では，各分布の平均ベクトルのユークリッド距離をもとにリーフノードが分布となる二分木の回帰木を構築し，適応データ量の期待値がある閾値より大きくなる最下位のノードにおいて分布の適応を行う方法が用いられている．しかしこの手法では，分布間の時間軸上での接続関係が考慮されていないため，本質的にフレーム単位の情報しか適応できないと考えられる．そこでここでは，セグメント単位の特徴も反映させるため，学習時に構築されたコンテキストクラスタリング決定木を回帰行列の共有に利用する手法 [8] を用いる．

5.2 スタイル適応の評価

男性話者 MMI の「読上げ」スタイルモデルを用い，適応データには「楽しげ」,「悲嘆」,「ぞんざい」の各スタイルで発声した学習文章に含まれている 10，20，50 文章の音声データを用いた．スペクトルパラメータの回帰行列には，ブロック対角行列を用いた．また，無音とポーズに対しては通常の音素とは別に各々決定木を構築した．

ABX 法による主観評価試験により，適応した音声が適応元の「読上げ」と各目標スタイルのどちらに近いかを評価した．ABX 法では，A を「読上げ」

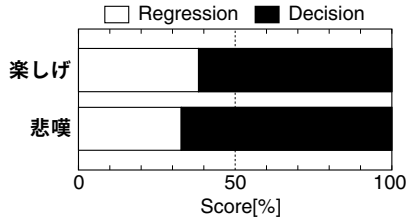


図 7: スタイル適応音声の対比較評価結果.

のモデルから合成された音声，B を目標スタイルのモデルから合成された音声，X を適応モデルから合成された音声とし，A，B，X または B，A，X の順に被験者に提示し，X が 1 番目と 2 番目のどちらに近いかを判定させる．被験者は 9 名で，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，スタイル毎にランダムに 3 文章を選び，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．

図 6 に ABX 法による主観評価結果を示す．スコアは適応モデルから合成された音声が目標スタイルモデルから合成された音声に近いと判定された割合を表す．この図より，適応文章数が増えるに従って目標スタイルに近付く傾向があることがわかる．

さらに，適応の際に回帰木を用いる従来手法 (Regression) と決定木を用いる提案法 (Decision) について，スタイル適応を行った後のモデルから生成された合成音声の対比較評価を行った結果を図 7 に示す．図のプレファレンススコアより，提案法のスタイル適応手法がより良いスタイル適応音声を生成できることがわかる．

6 まとめ

多様な話者性および発話スタイル・感情表現によるテキスト音声合成の実現をめざし，平均声のモデル化手法ならびに平均声からの任意話者音声の合成手法，様々な発話スタイルや感情を表現するスタイルモデリング手法，スタイル補間手法ならびにスタイル適応手法の検討・開発を行った結果について述べた．まず，平均声のモデリングでは，共有コンテキストクラスタリングと話者適応学習を導入した手法を提案し，平均声と話者適応後の合成音声の自然性が向上することを示した．次に，二つのスタイルモデリング手法を提案し，HMM 音声合成の枠組で発話スタイルや感情表現のモデル化が可能なことを示した．また，スタイル補間やスタイル適応のアプローチが話者補間や話者適応と同様に可能であることを示し，その有効性を主観評価試験により示した．

今後は，韻律のモデル化，特に継続長モデルの精密化を行うとともに，話者やスタイルの種類を増やして，より人間に近い多様な音声合成システムの実現をめざす予定である，さらに，スタイル制御，HMM 音声合成の品質改善等も行う予定である．

参考文献

[1] 小林 隆夫, 徳田恵一, “コーパスベース音声合成技術の動向 [IV] —HMM 音声合成方式—,” 電子情報通信学会誌, Vol.87, No.4, pp.322–327 (2004-4).

[2] 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,” 電子情報通信学会論文誌, Vol.J85-D-II, No.4, pp.545–553 (2002-4).

[3] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “A context clustering technique for average voice models,” *IEICE Trans. Information and Systems*, Vol.E86-D, No.3, pp.534–542 (2003-3).

[4] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “A training method of average voice model for HMM-based speech synthesis,” *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, Vol.E86-A, No.8, pp.1956–1963 (2003-8).

[5] 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生成の検討,” 電子情報通信学会技術研究報告, SP2002-172, pp.17–22 (2003.1).

[6] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, “Modeling of various speaking styles and emotions for HMM-based speech synthesis,” *Proc. 8th European Conference on Speech Communication and Technology*, EUROSPEECH ’03, Vol.III, pp.2461–2464, Geneva, Switzerland (2003.9).

[7] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, “HMM-based speech synthesis with various speaking styles using model interpolation” *Proc. 2nd International Conference on Speech Prosody*, SP2004, pp.413–416, Nara, Japan (2004.03).

[8] J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, “Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis,” *Proc. 2004 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2004, Vol.I, pp.5–8, Montreal, Canada (2004.05).

[9] 田中 智宏, 益子 貴史, 小林 隆夫, “瞬時周波数振幅スペクトルに基づくピッチ抽出法の検討,” 電子情報通信学会技術研究報告, SP2000-160, pp.1–8 (2001.3).

[10] D. Arifianto and T. Kobayashi, “Performance evaluation of IFAS-based fundamental frequency estimator in noisy environments,” *Proc. 8th European Conference on Speech Communication and Technology*, EUROSPEECH ’03, Vol.IV, pp.2877–2880, Geneva, Switzerland (2003.9).

[11] 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,” 電子情報通信学会論文誌, Vol.J83-D-II, No.11, pp.2099–2107 (2000-11).

[12] 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 (2000-4).