社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NLC2009-26,SP2009-90(2009-12)

多重ベータ混合モデルを用いた調波時間構造のモデル化による音声合成の検討

中鹿　亘†　立花　隆輝††　西村　雅史††　滝口　哲也†††　有木　康雄†††

† 神戸大学大学院工学研究科
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
†† 日本 IBM 東京基礎研究所
〒 242-8502 神奈川県大和市下鶴間 1623-14
††† 神戸大学自然科学系先端融合研究環
〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: †nakashika@me.cs.scite.kobe-u.ac.jp, †††{takigu,ariki}@kobe-u.ac.jp

あらまし　これまでに数多くの音声合成技術が提案されているが，我々は，音素信号の調波時間スペクトル形状をモデル関数で近似し，音声合成を行うという新たなフレームワークについて検討する．音素スペクトルの調波成分のみを取り出し，各ハーモニクスのエンベロープをスペクトルモデル関数でモデリングする．モデル関数のパラメータから音素信号を復元し，音声を合成する手法について同時に提案する．近似するモデル関数として，ベータ分布をベースにした多重ベータ混合モデルを考案し，評価実験により我々の提案するモデルの有効性について述べる．
キーワード　音声合成，TTS，多重ベータ混合モデル，調波構造，スペクトルモデル関数

A study on speech synthesis by modeling harmonics structure with Multi Beta Mixture Model

Toru NAKASHIKA†, Ryuki TACHIBANA††, Masafumi NISHIMURA††, Tetsuya TAKIGUCHI†††, and Yasuo ARIKI†††

† Graduate School of Engineering, Kobe University
Rokkodaicho 1-1, Nada-ku, Kobe, Hyogo, 657-8501 Japan
†† IBM Japan, Tokyo Research Laboratory
1623-14 Shimotsuruma, Yamato, Kanagawa, 242-8502 Japan
††† Organization of Advanced Science and Technology, Kobe University
Rokkodaicho 1-1, Nada-ku, Kobe, Hyogo, 657-8501 Japan
E-mail: †nakashika@me.cs.scite.kobe-u.ac.jp, †††{takigu,ariki}@kobe-u.ac.jp

Abstract　There are currently some researches related to speech synthesis, but here we present a new framework for speech synthesis, in which we approximate an envelope shape of each harmonic in a phoneme signal by a spectro-modeling function. In this approach only harmonic-parts are extracted from the phoneme spectrum, and the time-varying spectrum corresponding to the harmonics or sinusoidal components is modeled by the modeling function. In addition, we propose a method to synthesize a speech signal using Multi Beta Mixture Model (MBMM) based on Beta distribution. We discuss the effectiveness of our proposed model through the experimental results.
Key words　speech synthesis, text-to-speech, multi beta mixture model, harmonics structure, spectro-modeling function

1. はじめに

1.1 従来の音声合成技術

公共案内システムのコンテンツ自動読み上げや，発話困難な障がい者の代替手段として，人間の音声を人工的に作り出す音声合成技術を用いた，テキスト読み上げシステム (Text-to-speech: TTS) が利用される．この音声合成技術は，これまでに様々な手法が提案されてきた．Concatenative Synthesis (CS) が最も代表的な音声合成技術の一つであり，これは音声の断片波形データを連結して合成する手法である [1] [2] [3]．また徳田らは音声のスペクトル・ピッチ・持続長を HMM によりモデル化する HMM-based Synthesis (HS) を提案した [4] [5]．その他にもフォルマントの生成に必要なパラメータを設定することで音声を合成するフォルマント合成 (Formant Synthesis: FS) がある [6]．

CS は連結的音声合成とも呼ばれ，録音された音声の断片を適当な尺度で連結することで，音声波形を生成する手法である．これは録音された音声データを利用するため，比較的自然な音声合成が可能である反面，断片音声の接続方法が課題となり，適切に素片を結合しなければ音声としての自然性が損なわれる．また，CS では大量の音声データベースが必要となるため，実用上 CPU 時間，記憶容量などの膨大な計算機資源が要求される．

一方 HMM-based Synthesis [4] は音声を隠れマルコフモデル (HMM) によって統計量から音声を合成する手法である．音素ごとのスペクトル，基本周波数，継続時間を HMM によって同時にモデル化し，尤度最大化基準に基づいて合成音声を出力する．HS による音声合成は CS とは違って滑らかに音声を生成することができる上，連結的合成のような音声データベースを必要としないので，データサイズを抑えることができるという特徴を持つ．しかしながら HMM で出力される音声は CS に比べると肉声感が損なわれ，様々な発話スタイルで音声を合成することが困難である．また，HMM の学習には大量の学習データが必要となる．

Formant Synthesis では基本周波数，音色，雑音レベルなどのフォルマントパラメータを調整することで音声を合成する [6]．HS と同様に音声の原波形データの代わりにパラメータで音声を表現するため，計算機資源を抑えることができるが，CS に比べれば不自然な出力音声が得られる．

本研究では，音素ごとの調波時間スペクトルをある分布関数で近似し，モデル化された関数から音声を出力する，新しい音声合成手法を提案する．このときスペクトルの調波時間構造のモデルとして用いられる関数を，スペクトルモデル関数と呼ぶ．この手法はフォルマント合成と同様に各種パラメータから音声を合成するものであるが，スペクトルの調波構造に着目して，直接的に分布構造を関数でモデル化することが特徴である．

1.2 提案手法の概要

音素スペクトルの調波時間構造をモデル関数で近似し，音声合成を行う手法を本研究で提案する．似たような手法としてフォルマント合成が挙げられるが，これは音素認識に重要な要素であるフォルマント周波数をパラメータとして与えることで音声を人工的に造り出す．本研究では人間の有声音には必ずピッチが存在し，スペクトルの調波パターンが表れることに着目し，音素スペクトルの調波構造だけを取り出してモデル化する．このとき音素間の連結を滑らかにするために，各調波成分に対して時間的に連続な関数を用意することで音素間の不連続性を解消できることが期待される．このように本研究では調波時間パターンに着目して，音素スペクトルの調波時間構造を，スペクトルモデル関数でモデル化する（図 1）．また関数によってモデル化された音素パラメータから，音声を合成する手法についても本稿で提案する．

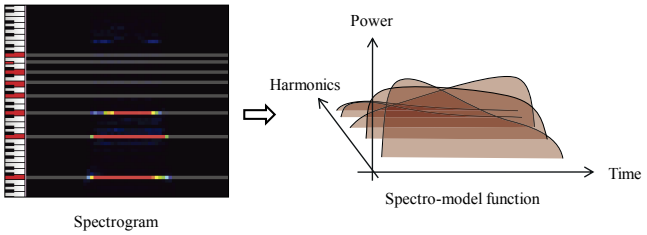

図 1　音素スペクトルの調波時間構造のモデル化
Fig. 1　Modeling of an envelope shape in a phoneme spectrum.

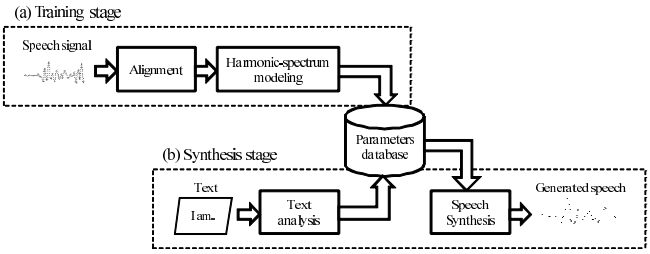

図 2　音声合成システムのフローチャート
Fig. 2　A flowchart of proposed method.

提案手法による音声合成システムの大まかな流れを次に説明する．図 2 のように，音声合成システムは学習部 (Training stage) と合成部 (Synthesis stage) の 2 つに大きく分けられる．まず学習部において音素スペクトルの調波構造を関数によってモデリングする処理について説明する．音素ごとの学習パターンを用意し，学習に用いられる音声信号に対して持続長・ピッチの正規化などのアライメント処理を行う．この信号を用いて，2. で説明するように，スペクトルモデル関数のパラメータを学習させる．こうして得られたパラメータは，スペクトルモデル関数パラメータデータベースに音素レベルで蓄積される．合成部では読み上げたいテキストをテキスト解析し，データベースから必要なパラメータを取得する．最後にそれらのパラメータから音声を合成する．

以降の 2. ではスペクトルモデル関数について解説し，3. で関数パラメータから音声合成を出力する手法について述べる．4. で，評価実験とその結果を報告し，最後に 5. で，結論と今後の課題について述べる．

2. スペクトルモデル関数

スペクトルモデル関数は音素スペクトルの調波時間構造をモデル化するために用いる関数である．これは以下のような条件を満たす時間-周波数の 2 変数関数が望ましい．

- 周波数軸に関して離散的
- 時間軸に関して連続的
- 全領域における積分値は 1
- パラメータを最尤法によって推定可能

これらを満たすため，本研究では式 (1) で表される一般多重関数 (Multi Function: MF) を定義する．これは各ハーモニクスごとに強度時間変化を表す関数を用意し，全体として調波時間スペクトル構造を表現する関数である．

$$q(t, n; \boldsymbol{\Theta}, \boldsymbol{\pi}) = \sum_n \pi_n \; p_n(t; \Theta_n) \tag{1}$$

ここで，t は時刻を表す変数，n はハーモニクスのインデックスを指す．$p(t; \boldsymbol{\Theta})$ は多重関数 $q(x, t)$ の部分関数 (Partial Function: PF) であり，

$$\forall n, \int p_n(t) dt = 1 \tag{2}$$

を満たす関数である．モデル化されるハーモニクスのエンベロープ形状は，この $p(t)$ によって定まる．多重関数のパラメータは $\boldsymbol{\Theta}, \boldsymbol{\pi}$ の 2 つであり，$\boldsymbol{\Theta}$ は部分関数のパラメータ行列を表す．$\boldsymbol{\pi}$ は多重率ベクトルを表し，部分関数間の強度比率を意味する．ただし，多重率 π は

$$\sum_n \pi_n = 1, \quad \forall n, \, \pi_n > 0 \tag{3}$$

を満たし，以下のようにパラメータを求めることができる．

$$\kappa_n = \frac{\int g_n(t) dt}{\int g_1(t) dt} \tag{4}$$

$$\pi_n = \frac{\kappa_n}{\sum_m \kappa_m} \tag{5}$$

ここで κ_n は第 1 ハーモニクスと第 n ハーモニクスの強度比率，$g_n(t)$ は観測値の第 n ハーモニクスの値を示す．

ピアノ音などの楽器音信号のスペクトル形状のモデルには，多重ベータ分布 (Multi Beta Distribution: MBD) を用いていた [8]．これは部分関数にベータ分布を用いた多重関数であり，式 (1)(6)(7) で定義される関数である．

$$p_n(t; \alpha_n, \beta_n) = \frac{1}{B(\alpha_n, \beta_n)} t^{\alpha_n - 1} (1 - t)^{\beta_n - 1} \tag{6}$$

$$B(\alpha_n, \beta_n) = \int_0^1 t^{\alpha_n - 1} (1 - t)^{\beta_n - 1} \, dt \tag{7}$$

ただし，

$$\forall n, \, \alpha_n, \beta_n > 0. \tag{8}$$

楽器音の場合には比較的単純なスペクトル形状を表現できれば良く，多重ベータ分布で十分近似することができた．しかし音声モデルでは楽器音の構造とは異なり，スペクトル形状にピークを複数持つような複雑な構造になるため，単峰的な多重ベータ分布では十分に音素スペクトルを近似することができない (図 3)．そこで本研究では音素スペクトルのスペクトルモデル関数として，多重ガウス混合分布 (Multi Gaussian Mixture Model: MGMM)，多重ベータ混合モデル (Multi Beta Mixture Model: MBMM) の 2 つのモデルを考案し，音素スペクトルのモデリング実験を行った．いずれも混合モデルをベースにしており，複数の強度ピークを持つような構造を近似するのに適していると期待される．次節以降でこれらのモデルについて述べる．

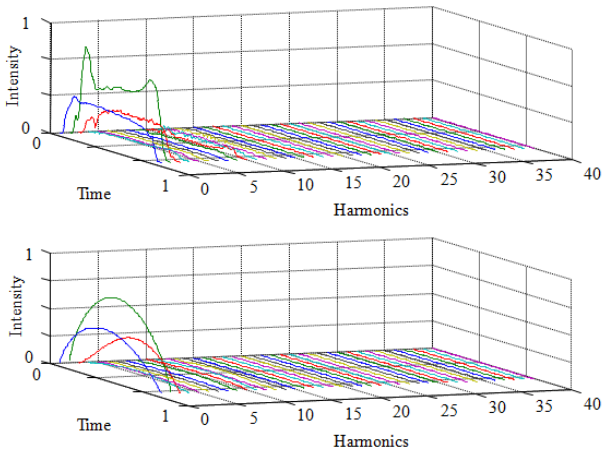

図 3　多重ベータ分布による音声のモデル化の結果．上段が音素 /e:/ のオリジナルのスペクトル構造，下段が多重ベータ分布によるモデル

Fig. 3　A result of modeling speech spectrum using multi-beta-distribution. Observed spectrum envelopes of the phoneme /e:/ (top) and multi-beta-distribution model (bottom).

2.1 多重ガウス混合分布

多重ガウス混合分布は，ガウス混合分布を部分関数とした多重関数である．すなわち，式 (1)(9) のように定義される．

$$p_n(t; \nu_n, \mu_n, \sigma_n) = \sum_l \nu_{n,l} \frac{1}{\sqrt{2\pi}\sigma_{n,l}} \exp\left\{ -\frac{(t - \mu_{n,l})^2}{2\sigma_{n,l}^2} \right\} \tag{9}$$

ただし，$\nu_{n,l}$ は

$$\forall l, \sum_n \nu_{n,l} = 1, \quad \forall n, l, \, \nu_{n,l} > 0 \tag{10}$$

を満たす混合率であり，l は混合コンポーネント表すインデックスである．

ガウス混合分布のパラメータの推定法については，データサンプルから EM アルゴリズムで値を更新させる方法がある [10]．しかし本研究の枠組みの中では，[10] に挙げられるようなデータサンプルは直接観測されず，スペクトル形状そのものをモデル化しようとしているので，このパラメータ更新法を用いようとすれば，観測スペクトルからサンプル値を発生させる処理が必要になる．そこで [11] のように，観測されるスペクトル形状を

直接更新式に取り入れた方法を用いて，パラメータを更新する方が効率が良い．よって本稿ではこの手法のように評価関数を定義し，解析的に更新式を解くことでパラメータの更新を行う．

モデルとなる多重ガウス混合分布を観測スペクトル形状にフィッティングさせるため，これらの擬距離であるカルバックライブラー (KL) 情報量を評価関数として，この評価関数を最小とするようなパラメータを求める．評価関数を式 (11) のようにおく．

$$J = \sum_n J_n = \sum_n \int_{-\infty}^{\infty} g_n(t) \log \frac{g_n(t)}{p_n(t)} dt \tag{11}$$

また，$u_{n,l}, v_{n,l}$ を式 (12)(13) のように定義する．

$$u_{n,l} = \frac{\nu_{n,l}}{\sqrt{2\pi}\sigma_{n,l}} \exp\left\{ -\frac{(t-\mu_{n,l})^2}{2\sigma_{n,l}^2} \right\} \tag{12}$$

$$v_{n,l} = \int_{-\infty}^{\infty} \frac{g_n(t) u_{n,l}}{p_n(t)} dt \tag{13}$$

ラグランジュの未定乗数法を用いて，式 (10) の元で評価関数 J を最小化させるようなパラメータを求めれば，

$$\hat{\nu}_{n,l} = \frac{v_{n,l}}{\sum_m v_{n,m}} \tag{14}$$

$$\hat{\mu}_{n,l} = \frac{\int_{-\infty}^{\infty} \frac{t \cdot g_n(t) u_{n,l}}{p_n(t)} dt}{v_{n,l}} \tag{15}$$

$$\hat{\sigma}_{n,l} = \sqrt{\frac{\int_{-\infty}^{\infty} \frac{(t-\mu_{n,l})^2 g_n(t) u_{n,l}}{p_n(t)} dt}{v_{n,l}}} \tag{16}$$

のように解析解を得ることができる．したがって式 (12)〜(16) を繰り返し更新することで多重ガウス混合分布のパラメータを最適解に近づけることが可能である．

2.2 多重ベータ混合モデル

ベータ分布の混合モデルであるベータ混合モデルを部分関数にした多重分布が，多重ベータ混合モデルである．この多重関数は式 (1)(7)(17) のように表される．

$$p_n(t; \nu_n, \alpha_n, \beta_n) = \sum_l \nu_{n,l} \frac{1}{B(\alpha_{n,l}, \beta_{n,l})} t^{\alpha_{n,l}-1} (1-t)^{\beta_{n,l}-1} \tag{17}$$

式 (17) は部分関数のベータ混合モデルの定義式であり，そのパラメータは EM アルゴリズムによって推定することができる [9]．導出についてはここでは省略するが，M ステップにおける各パラメータの更新式は以下のようになる．

$$\hat{\nu}_{n,l} = \frac{\sum_{i=1}^{K} z^*_{n,l,i}}{K} \tag{18}$$

$$\hat{\alpha}_{n,l} = \Psi^{-1}\left(\frac{1}{K} \sum_{i=1}^{K} \log\left(\frac{X_i}{1-X_i} \right) + \Psi(\beta_{n,l}) \right) \tag{19}$$

$$\hat{\beta}_{n,l} = \Psi^{-1}\left(\frac{1}{K} \sum_{i=1}^{K} \log\left(\frac{1-X_i}{X_i} \right) + \Psi(\alpha_{n,l}) \right) \tag{20}$$

$\Psi(x)$ は digamma 関数を表し，$\Psi^{-1}(x)$ はその逆関数である．X_i はサンプル値であり，観測スペクトルからランダムに発生させることで得られる．K はサンプル X_i の個数である．

ここで $z^*_{n,l,i}$ は，あるサンプル値 X_i が第 n ハーモニクスのベータ混合モデルの第 l コンポーネントから発生する確率を表す潜在変数である．$z^*_{n,l,i}$ は E ステップで以下のように更新される．

$$z^*_{n,l,i} = \frac{\hat{\nu}_{n,l} f_{n,l}(X_i|\hat{\alpha}_{n,l}, \hat{\beta}_{n,l})}{\sum_j \hat{\nu}_{n,j} f_{n,j}(X_i|\hat{\alpha}_{n,j}, \hat{\beta}_{n,j})} \tag{21}$$

$$f_{n,l}(X_i|\hat{\alpha}_{n,l}, \hat{\beta}_{n,l}) = \frac{X_i^{\hat{\alpha}_{n,l}-1}(1-X_i)^{\hat{\beta}_{n,l}-1}}{B(\hat{\alpha}_{n,l}, \hat{\beta}_{n,l})} \tag{22}$$

以上の E ステップと M ステップを十分に繰り返し計算することで，ベータ混合モデルのパラメータ $\Theta = \{\nu_n, \alpha_n, \beta_n\}$ を求めることができる．

3. モデルパラメータからの音声合成

この章ではスペクトルモデル関数のパラメータから音素信号を合成する手法について述べる．音素信号は倍音加算方式によって合成することができる．倍音加算方式では，合成される音素信号 $s(t)$ を (23) 式で表せる [7].

$$s(t) = \sum_n a_n(t) \sin\left(\frac{2\pi f_n t}{T} \right) \tag{23}$$

f_n は第 n ハーモニクスの周波数，T は発音長である．ここで $a_n(t)$ を (24) 式のようにおけば，学習済みのモデル関数パラメータから信号の復元が可能である．

$$a_n(t) = \pi_n \cdot p_n(\frac{t}{T}; \Theta_n) \tag{24}$$

ただし，$p_n(t)$ は部分関数である．

4. 評 価 実 験

4.1 実験手順と条件

提案手法の評価を行うために，音素波形をスペクトルモデル関数でモデル化し，音声を合成する実験を行った．実験に用いた学習データは 22.05kHz で録音された女性アナウンスの音声ファイルを使用した．学習させる音素は /a:/, /i:/, /u:/, /e:/, /o:/ の 5 つの長母音である．音声データに対し，発話区間を検出 [12] 後，音素に相当する部分を切り出した．その後 PSOLA [13] を用いてピッチを 440Hz に規定した．2. で述べたスペクトルモデル関数のパラメータ推定法によって音素スペクトルのパラメータを抽出し，データベースに蓄積した．本稿では，スペクトルモデル関数で音素スペクトル構造をモデル化し，音声を合成するという手法自体に重きを置いているので，ここではテキスト解析を用いた入力テキストからの音声合成は行わず，3. で述べた合成手順によって 5 つの長母音の音声を出力する実験を行った．

調波構造のモデルとなるスペクトルモデル関数としては実験条件の異なる 3 種類の MBMM と MGMM を用意した．このときの実験条件は表 1 のようになる．MBMM として B1，B2，B3，MGMM として G1，G2，G3 の 3 種類の実験条件がある．No. of mixtures, No. of iterations, No. of samples はそれぞれ混合数の数，EM アルゴリズムの繰り返し回数，サンプ

表 1 実験条件
Table 1 Experimental conditions.

	MBMM			MGMM		
	B1	B2	B3	G1	G2	G3
No. of mixtures	2	4	4	2	4	8
No. of iterations	20	10	100	200	200	200
No. of samples	2000	1000	5000	-	-	-

リング数を表わす．ただし今回用いた MGMM のパラメータ推定にはサンプリングを用いないので，サンプリング数は不定となる．いずれのモデルにおいてもハーモニクスの数を 20 としている．また，参考としてスペクトルモデル関数に多重ベータ分布を用いたモデルも用意した．

4.2 実験結果と考察

図 4 は提案手法によって音素 /e:/ をモデル化した実験結果を表す．図中の中段，下段はそれぞれ実験条件 G3, B3 の結果である．縦軸がスペクトル強度，奥軸が時間，横軸がハーモニクスを示している．この図を見れば，MGMM, MBMM ともに，入力音素信号の調波時間スペクトル構造のおおよその形状特徴を掴めていることが分かる．例えばいずれのモデルにおいてもハーモニクス間の強度比率や音の立ち上がり，スペクトルピーク（山）の開始時間や持続時間などがうまく推定されている．

またモデルによる比較を，可視性向上のため 2 次元プロットしたものを図 5 に示す．これは音素 /e:/ の第 2 ハーモニクスの強度構造の比較である．横軸が時間，縦軸が強度を示している．図から多重ベータ分布では 2 つ以上山を持つような構造を表現できず，オリジナルのスペクトル形状を再現できていないことが分かる．一方混合モデルである MGMM や MBMM では，多重ベータ分布と比較してオリジナルに近い分布構造になっている．MGMM の中でも混合数の違う G2 と G3 を比べると，混合数の多い G3 の方がよく近似できている．また混合数の等しい G2 と B3 を比較すれば，後者の方がオリジナルに近い形状を表わす．これは次のような事柄から起因していると考えられる．MGMM, MBMM はそれぞれ正規分布，ベータ分布から派生している．ベータ分布の方が正規分布よりも関数形状として曲率の大きな傾向があり，大まかに形状特徴を掴むことを得意としている [8]．よってその混合モデルである MBMM の方が，図 5(a) の時刻 0.3 から 0.7 のように，直線に近い曲線を近似できる．

最後に表 1 の各実験条件でモデル化を行い，パラメータから形状を復元した構造と，オリジナルのスペクトル構造との DP 距離を算出した．この結果を図 6 に示す．図中の B1, B2, B3, G1, G2, G3 がそれぞれ表 1 の実験条件に対応している．縦軸は DP 距離を示しており，この数値が小さいほど，モデルによる近似がよくできていることを示している．このときに用いた DP 距離の算出方法は以下のとおりである．まずハーモニクスごとに時間-強度の 2 変数 DP 距離を求める．次にそれらを全てのハーモニクスについて総和をとったものを図 6 の DP 距離としている．図 4 や図 5 では，おおよその事実として，モデルの混合数が増えるほどオリジナルのスペクトル構造に近づいていることが視覚的に分かった．図 6 ではこのことを数値的に評価することができ，B2 を除いて混合数が多くなるモデルほど，よく近似できていることが読み取れる．B2 と B3 とではサンプル数，EM アルゴリズムの繰り返し回数が異なり，この図から B2 の結果では，まだ十分に学習できていない状態であるため DP 距離が大きくなっていると考えられる．最も DP 距離が小さくなったのは唯一混合数が 8 の G3 である．逆に混合数が等しい場合を比較すると，B1 と G1 とでは B1 が，B3 と G2 とでは B3 の方が値が小さくなっていることが分かる．つまりこのことは，ある程度十分に学習されている状態であれば，多重ガウス混合分布よりも多重ベータ混合モデルの方が，スペクトル構造をより精度よくモデリングできることを意味している．

表 2 モデルパラメータの数
Table 2 Number of model parameters

	MBD	MBMM			MGMM		
		B1	B2	B3	G1	G2	G3
No. of parameters	60	140	260	260	140	260	500

一方各モデルにおけるパラメータの数は，表 2 のようになる．このときに用いたパラメータ数に関する式を以下に示す．MBD, MGMM, MBMM のそれぞれのパラメータ数を $\gamma_{mbd}, \gamma_{mgmm}, \gamma_{mbmm}$，ハーモニクスの数を ρ，MGMM や MBMM の混合数を λ とすると，

$$\gamma_{mbd} = 3 \cdot \rho \tag{25}$$

$$\gamma_{mgmm} = \rho \cdot (3 \cdot \lambda + 1) \tag{26}$$

$$\gamma_{mbmm} = \rho \cdot (3 \cdot \lambda + 1) \tag{27}$$

と表すことができる．式 (26)(27) より，混合数，ハーモニクス数が共に等しい場合，MGMM と MBMM のパラメータ数が等しいことが分かる．ある程度の精度を保ちつつ，パラメータをなるべく抑えるとすれば，B3 が妥当なモデルだと考えられる．

5. おわりに

本稿では，スペクトルモデル関数を用いて音素信号の調波時間スペクトル形状をモデル化し，音声合成を行う手法について提案した．我々は音素スペクトル形状のモデリングに相応しいモデル関数として多重ガウス混合分布，多重ベータ混合モデルの 2 つのモデルを考案し，学習の繰り返し回数や混合数などを変えた，複数の実験条件を用意して評価実験を行った．実験結果によって提案手法の妥当性が示され，パラメータの数と近似精度のバランスを考慮すれば，多重ベータ混合モデルが最適なモデルであると考えられる．今後は，さらに表現力の高く音声のモデルに適したスペクトルモデル関数の考案，MBMM のパラメータ推定時の収束速度の改善，プリファレンススコアによる HS や FS との比較について検討していきたい．

文　献

[1] Andrew J. Hunt and Alan W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” IEEE ICASSP, pp.373-376, 1996.
[2] Romain Prudon and Christophe d’ Alessandro, “A selec-

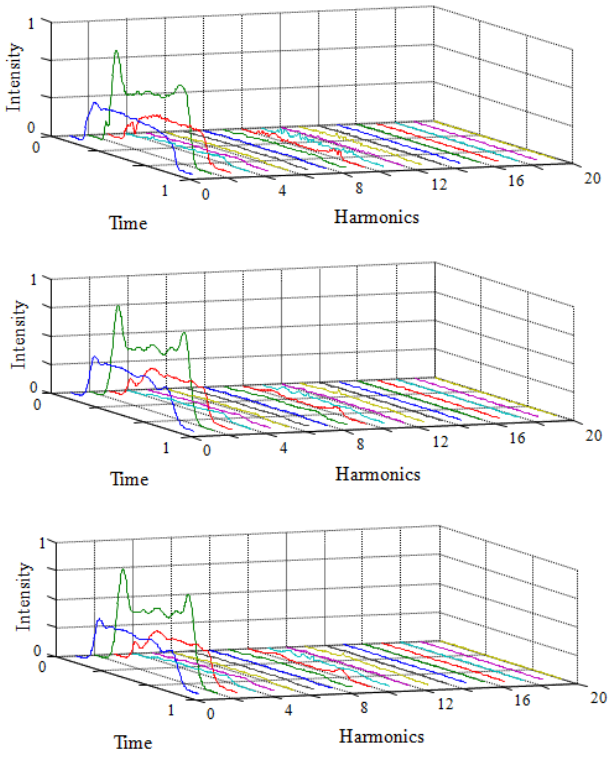

図 4　実験結果．上段から順に音素 /e:/ のオリジナルのスペクトル構造，多重ガウス混合分布によるモデル，多重ベータ混合モデルの結果を表す

Fig. 4　Experimental result. Observed spectrum envelopes of the phoneme /e:/ (top), modeling result of multi-gaussian-mixture-model (middle) and result using multi-beta-mixture-model (bottom).

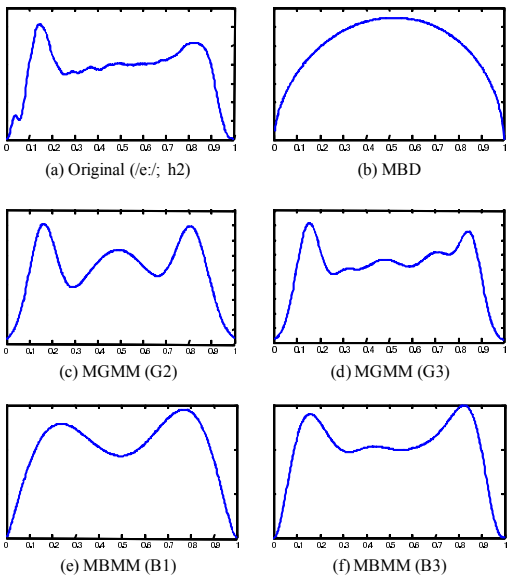

図 5　スペクトルモデル関数のスペクトル構造比較

Fig. 5　Comparison of spectrum-modeling function shapes (two-dimensional view).

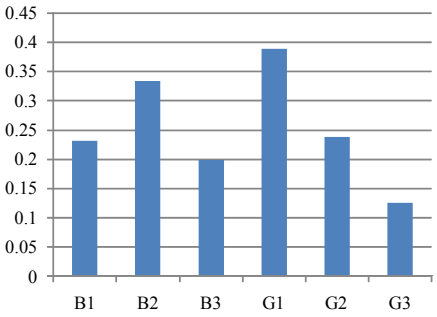

図 6　DP 距離の比較

Fig. 6　Comparison of DP distances.

tion/concatenation TTS synthesis system: Databases development, system design, comparative evaluation," Speech Synthesis Workshop, 2001.

[3] Gregory Beller *et al.*, "A hybrid concatenative synthesis system on the intersection of music and speech," JIM, 2005.

[4] M. Tamura *et al.*, "Speaker adaptation for HMM-based speech synthesis system using MLLR," ESCA/COCOSDA workshop on Speech Synthesis, pp.273-276, 1998.

[5] M. Plumpe *et al.*, "HMM-based smoothing for concatenative speech synthesis," ICSLP, vol.6, pp.2751-2754, 1998.

[6] T. Styger and E. Keller, "Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges," pp.109-128, 1994.

[7] Xavier Rodet, "Musical Sound Signal Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models," TFTS'97, 1998.

[8] T Nakashika *et al.*, "Mathematical Modeling of Harmonic-Timbre Structure with Multi-Beta-Distribution," IEEE Workshop on Statistical Signal Processing, pp.769-772, 2009.

[9] Yuan Ji *et al.*, "Applications of Beta-Mixture Models in Bioinformatics," Bioinformatics, vol.21, no.9, pp.2118-2122, 2005.

[10] Christopher M. Bishop, "Pattern Recognition and Machine Learning," Springer. 2006.

[11] H Kameoka *et al.*, "Harmonic-temporal structured clustering via deterministic annealing EM algorithm for audio feature extraction," ISMIR2005, pp.115-122, 2005.

[12] J. Ramirez *et al.*, "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness," Robust Speech Recognition and Understanding. pp. 1-22, 2007.

[13] X. Huang *et al.*, "Spoken Language Processing: A Guide to Theory, Algorithm and System Development," 2001.