多重関数を用いた調波時間スペクトル形状のモデル化による音声合成*

◎中鹿亘 (神戸大), 立花隆輝, 西村雅史 (日本 IBM), 滝口哲也, 有木康雄 (神戸大)

1　はじめに

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 (連結的合成) 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]．また，基音とその倍音の正弦波を任意の割合で加算することで音声を合成する Additive synthesis (加算合成) がある [3]．

連結的合成は，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，連結的合成では大量の音声データベースが必要となるため，CPU 時間，記憶容量などの膨大な計算機資源が要求される．

加算合成では音声のフォルマント情報を，調波成分の正弦波を合成することで音声を得る [3]．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できるため，計算機資源を抑えることができるが，出力音声の自然性は連結適合性と比べて劣る傾向がある．

本研究では加算合成法を拡張し，音素ごとの調波時間スペクトルを，本稿で定義する多重関数で近似して，モデル化された関数のパラメータから音声を復元合成する音声合成手法を提案する (Fig. 1)．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．

2　多重関数

調波スペクトル形状のフィッティングを行うには，周波数軸に関して離散的，且つ時間軸に関して連続的，且つ全領域における積分値が 1，且つパラメータを最尤法によって推定可能な時間-周波数の 2 変数

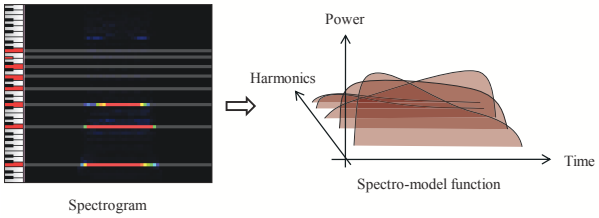

Fig. 1　Modeling of an envelope shape in a phoneme spectrum

関数を用いることが望ましい．そこで本研究では式 (1) で表される多重関数 (Multi Function) 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) 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．

*Speech synthesis by modeling harmonics structure with Multiple Function. by TORU NAKASHIKA (Kobe University), RYUKI TACHIBANA, MASAFUMI NISHIMURA (IBM Japan), TETSUYA TAKIGUCHI, YASUO ARIKI (Kobe University)

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．

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)(6) のように定義される．

$$p_n(t;\nu_n,\mu_n,\sigma_n)=\sum_l \nu_{n,l}\frac{1}{\sqrt{2\pi}\sigma_{n,l}}\exp\left\{-\frac{(t-\mu_{n,l})^2}{2\sigma_{n,l}^2}\right\} \tag{6}$$

ただし，$\nu_{n,l}$ は

$$\forall l,\sum_n \nu_{n,l}=1,\quad \forall n,l,\ \nu_{n,l}>0 \tag{7}$$

を満たす混合率であり，l は混合コンポーネントを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 [7]．しかし本研究の枠組みの中では，[7] 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．本稿では以下に示す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 (8) のようにおく．

$$J=\sum_n J_n=\sum_n\int_{-\infty}^{\infty} g_n(t)\log\frac{g_n(t)}{p_n(t)}dt \tag{8}$$

また，$u_{n,l},v_{n,l}$ を式 (9)(10) のように定義する．

$$u_{n,l}=\frac{\nu_{n,l}}{\sqrt{2\pi}\sigma_{n,l}}\exp\left\{-\frac{(t-\mu_{n,l})^2}{2\sigma_{n,l}^2}\right\} \tag{9}$$

$$v_{n,l}=\int_{-\infty}^{\infty}\frac{g_n(t)u_{n,l}}{p_n(t)}dt \tag{10}$$

ラグランジュの未定乗数法を用いて，式 (7) 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，

$$\hat{\nu}_{n,l}=\frac{v_{n,l}}{\sum_m v_{n,m}} \tag{11}$$

$$\hat{\mu}_{n,l}=\frac{\int_{-\infty}^{\infty}\frac{t\cdot g_n(t)u_{n,l}}{p_n(t)}dt}{v_{n,l}} \tag{12}$$

$$\hat{\sigma}_{n,l}=\sqrt{\frac{\int_{-\infty}^{\infty}\frac{(t-\mu_{n,l})^2 g_n(t)u_{n,l}}{p_n(t)}dt}{v_{n,l}}} \tag{13}$$

のように解析解を得ることができる．したがって式 (9)∼(13) 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)(14) のように表される．ただし $B(\alpha,\beta)$ はベータ関数である．

$$p_n(t;\nu_n,\alpha_n,\beta_n)=\sum_l \nu_{n,l}\frac{1}{B(\alpha_{n,l},\beta_{n,l})}t^{\alpha_{n,l}-1}(1-t)^{\beta_{n,l}-1} \tag{14}$$

式 (14) 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 [6]．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．

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

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

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

$\Psi(x)$ 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．

$$z^*_{n,l,i}=\frac{\hat{\nu}_{n,l}f_{n,l}(X_i|\hat{\alpha}_{n,l},\hat{\beta}_{n,l})}{\sum_j \hat{\nu}_{n,j}f_{n,j}(X_i|\hat{\alpha}_{n,j},\hat{\beta}_{n,j})} \tag{15}$$

$$f_{n,l}(X_i|\hat{\alpha}_{n,l},\hat{\beta}_{n,l})=\frac{X_i^{\hat{\alpha}_{n,l}-1}(1-X_i)^{\hat{\beta}_{n,l}-1}}{B(\hat{\alpha}_{n,l},\hat{\beta}_{n,l})} \tag{16}$$

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 (17) 式で表せる [4].

$$s(t) = \sum_n a_n(t) \sin\left(\frac{2\pi f_n t}{T}\right) \tag{17}$$

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 (18) 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．

$$a_n(t) = \pi_n \cdot p_n(\frac{t}{T}; \Theta_n) \tag{18}$$

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 [8] 後，音素に相当する部分を切り出した．その後 PSOLA[9] 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 2 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．B1，B2 は MBMM モデル，G1，G2 は MGMM のモデルを表し，それぞれ混合数や繰り返し回数などの実験条件が異なる．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，サンプリング数を表わす．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 (A1) も用意した．

Table 1　Experimental conditions

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

4.2　実験結果と考察

Fig. 2 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 G2, B2 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 Fig. 3 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．混合数の等しい G2 と B2 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 [5]．よってその混合モデルである MBMM の方が，Fig. 3 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．

最後にそれぞれの実験条件で，パラメータから復元した形状とオリジナルのものとの DP 距離を算出した．この結果を Fig. 4 に示す．図中の B1, B2, G1, G2 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 Fig. 4 の DP 距離

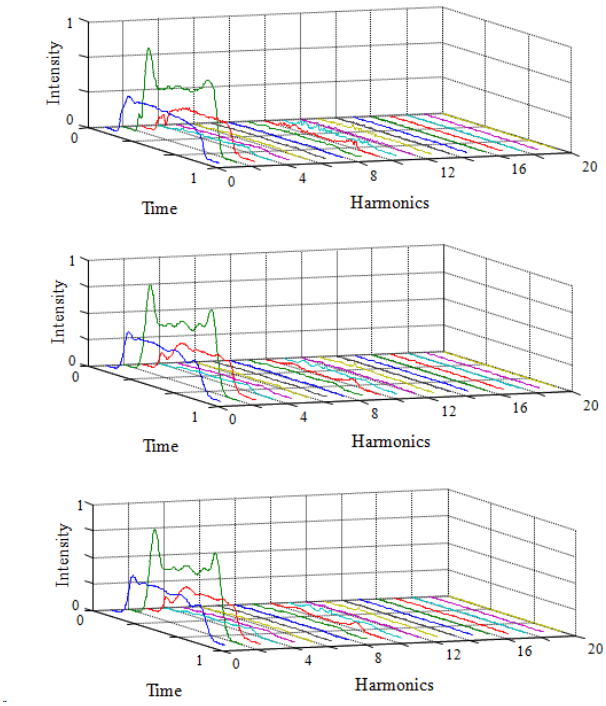

Fig. 2 Experimental result. Observed spectrum envelopes of the phoneme /e:/ (top), modeling result of multi-gaussian-mixture-model (middle) and result using multi-beta-mixture-model (bottom)

としている．Fig. 4 から，混合モデルの場合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分かる．また混合数を等しくしてモデル間で比較した場合，多重ベータ混合モデル，多重ガウス混合分布，多重ベータ分布の順にスペクトル構造をより精度よくモデリングできることが読み取れる．

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．

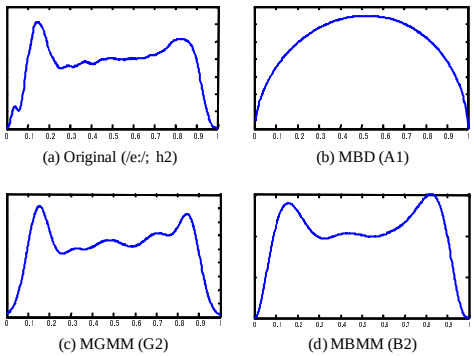

Fig. 3 Comparison of spectrum-modeling function shapes (two-dimensional view)

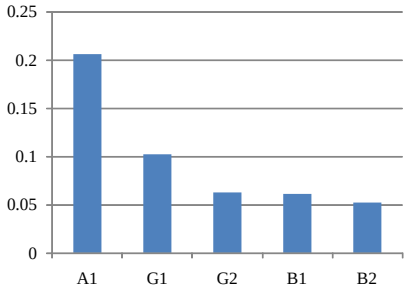

Fig. 4 Comparison of DP distances

参考文献

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

[3] Remez, R. E. et al., “Talker identification based on phonetic information, ”Journal of Experimental Psychology: Human Perception and Performance, vol23, pp.651-666, 1997.

[4] Xavier Rodet, “Musical Sound Signal Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models,” TFTS’97, 1998.

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

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

[7] Christopher M. Bishop, “Pattern Recognition and Machine Learning,” Springer. 2006.

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

[9] X. Huang et al., “Spoken Language Processing: A Guide to Theory, Algorithm and System Development,” 2001.