■2 群（画像・音・言語）－ 7 編（音声認識と合成）

3 章 音声合成

（執筆者：）

■概要■

【本章の概要】

■2群 － 7編 － 3章

3-3 合成手法

（執筆者：山下洋一）［2009年4月 受領］

人の発声では，声門から送り出される音源波形が，複雑に変化する声道で共振することによって様々な音を生成している．この過程は入力信号と共振回路による電気回路としてモデル化でき，声門における音源と声道での共振を分けて考える音声生成過程のモデルは，線形分離等価回路モデル（あるいはソース・フィルタモデル）と呼ばれ，現在の音声情報処理の基礎を与えている．音声をコンピュータによって合成する一つの方式は，この生成過程モデルに基づいた手法である．すなわち，声道における共振特性を表現するディジタルフィルタを構成し，インパルス系列や白色雑音で近似した音源波形を入力することによって音声生成過程を模擬する手法であり，パラメータ合成方式（あるいはボコーダ方式）と呼ばれる．これとは異なるもう一つの音声合成方式は波形接続合成方式と呼ばれ，収録した音声の時間波形を多数蓄積しておき，適切な波形（あるいはその一部）を選択し，時間領域で接続して再生する方式である．

3-3-1 分析合成と規則合成

一般に，音声分析とは，音声の時間波形を特徴パラメータの時系列に変換する処理をいう．通常，線形分離等価回路モデルに基づいた分析をフレームと呼ばれる20〔msec〕程度の微小区間ごとに行い，声道の特性を表現する特徴パラメータの時系列を得る．分析とは逆の過程によって，すなわち，特徴パラメータに基づいて構成されるディジタルフィルタを音源波形によって駆動することによって，特徴パラメータ時系列を音声波形に変換することができる．このように一度発声された音声に対して，音声分析によって得られたパラメータ時系列からもとの音声波形を復元する処理を分析合成という．

話者が発声していない文（や単語など） の音声を合成するには，合成単位と呼ばれる音素などの短い単位ごとに，音声の特徴を表現した素片をあらかじめ作成しておき，合成すべき文を合成単位に分割し，合成単位系列に対する素片列を決定し音声を合成する．このように比較的短い合成単位を用いて任意の文や単語を合成する方式を一般に規則合成という．合成単位としては，音素のほか，音節やVCV（母音＋子音＋母音の連鎖），CVC（子音＋母音＋子音の連鎖），diphone（2音素連鎖）などが試みられている．

また，音声合成の用途としては，限られた数の単語や句から成る文を合成できれば十分であるような応用も多い．このような応用では，定型文における一部の単語を置き換えることによって音声を合成する録音編集の方式がとられることもあり，この場合には単語や句が合成単位となっていると考えることができる．

3-3-2 パラメータ合成方式（ボコーダ方式）

パラメータ合成方式の音声合成では，音素の特徴を蓄積しておくために，音声波形を特徴パラメータに変換することが必要であり，様々な特徴パラメータを利用した音声合成が行われている．

(1) 声道アナログ合成方式

声道アナログ合成（articulatory synthesis）は，音声の生成過程を模擬するために，声道の形状を近似表現し声道内の音波の伝搬を模擬する方式である．声道を微少な円筒形を接続した音響管とみなし，観測音声波形から各円筒の断面積を推定することによって声道形状を推定する手法や，古くは X 線，近年では MRI など計測技術を用いて，発声時の声道形状を計測し，口唇の開口面積，舌による声道内のせばめ，声門の面積，鼻腔部の広さなどの特徴パラメータによって声道をモデル化する手法が行われている [1]．

(2) フォルマント合成方式（ターミナルアナログ合成方式）

声道アナログ合成が声道内での音波の伝播までを模擬する方式であるのに対し，フォルマント合成（formant synthesis）は，声道における調音の特性を複数の共振回路を接続した電気回路で模擬する方式であり，ターミナルアナログ合成とも呼ばれる [2]．声道における共振をフォルマントといい，その中心周波数（フォルマント周波数）とバンド幅によって表される．i-番目のフォルマント周波数とバンド幅をそれぞれ F_i，B_i，サンプリング周波数を F_S とすると共振の周波数特性は

$$H_i(z)=\frac{1}{1-2e^{-\pi B_i/F_S}\cos(2\pi\frac{F_i}{F_S})z^{-1}+e^{-2\pi B_i/F_i}z^{-2}} \qquad (3\cdot 1)$$

によって表現される．声道における調音は，このような 2 次の IIR フィルタを直接あるいは並列に接続して近似される．直接接続では全極モデルとなり，鼻音などのようにスペクトルに零点を有する音素の合成では並列接続が必要となることから，一般には，両者を組み合わせて合成フィルタが構成される．

(3) その他のパラメータを用いた音声合成

声道形状やフォルマント以外にも，声道断面積へ変換可能な PARCOR 係数（partial autocorrelation coefficient; 偏自己相関係数），補間特性に優れた LSP（line spectrum pair; 線スペクトル対）などを用いたパラメータ合成がこれまでに試みられており，近年では，本章 3-5 節で述べられるようにケプストラムを特徴パラメータとする HMM 音声合成がパラメータ合成の代表的は手法となっている [3]．

3-3-3 波形接続合成方式

パラメータ合成方式が音声の生成過程を模擬して実現されているのに対し，波形接続合成（concatenative synthesis）は，観測された多数の音声波形を合成単位に分割し，それぞれの時間波形を素片としてデータベースに蓄積しておき，合成すべき文に対して最適な素片系列をデータベースから取り出して時間領域で接続して再生することによって音声合成を実現する [4, 5]．この方式では，データベースにおける素片の質と量，更に素片選択の手法によって合成音声の品質が決まる．

(1) 素片の選択

合成すべき対象の文 T が N 個の合成単位 $t_1, t_2, \cdots, t_N$ から構成されているとき，最適な素片系列 $\Theta=\theta_1, \theta_2, \cdots, \theta_N$ は，一般に，

$$d(\Theta,T)=\sum_{i=1}^{N}d_t(\theta_i,T)+\sum_{i=1}^{N-1}d_c(\theta_i,\theta_{i+1}) \qquad (3\cdot 2)$$

で定義されるコストを最小化する素片系列として決定される．ここで，式(3･2)の右辺第 1 項はターゲットコストと呼ばれ，t_i に対してθ_i がどの程度適切かを評価した値で，音韻やアクセント型などの言語的な素性や F0 周波数などの物理量によって定義される．また，第 2 項は接続コストと呼ばれ，隣接する二つの素片θ_i とθ_{i+1} の接続部での物理的な（音響的な）ひずみの量を評価する．

(2)　PSOLA

波形接続合成では，データベース中に適切な素片がない場合に音質の劣化を引き起こす．特に，音声の韻律変化は多様であるため，基本周波数や時間長がターゲットの合成単位に十分に適合しないことが考えられる．この問題を解決するために，時間領域での処理によって，素片の基本周波数や時間長を変更する手法が提案されている．その代表的な手法として PSOLA（Pitch Synchronous Overlap and Add）がある [6]．

図 3･1 に示すように，PSOLA は音声のピッチ情報が既知としたうえで，ピッチ周期に同期して窓掛けして音声波形をピッチごとに切り出し，ピッチ周期を変更したうえで加算することによって基本周波数の変更を実現する．また，切り出した 1 ピッチ波形を必要に応じて繰り返し加算したり，間引いたりすることによって時間長を制御することも容易に行える．

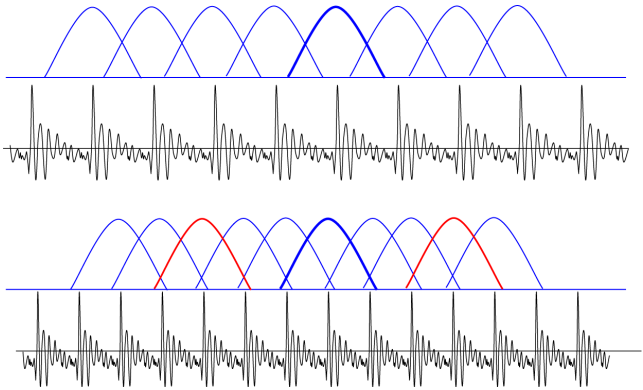

図 3・1　PSOLA による F0 の変更

3-3-4 パラメータ合成と波形接続合成の比較

パラメータ合成では，調音特性を表現したフィルタを音源波形で駆動する処理を必要とするため，やや機械的な音質となる．しかし，素片境界での接続がスムーズであり，基本周波数や時間長の制御を容易に行うことができる．近年の HMM 合成では声質や感情の制御も可能となってきている．

一方で，波形接続合成は，音声の時間波形をそのまま接続して利用するため素片の音質は非常に良く，合成音声の明瞭性も高い．しかし，素片の接続がうまくいかず異音やノイズが発生することがある．また，声質の変更が非常に難しく，パラメータ合成に比べて大量の素片データベースを必要とする．

■参考文献

1) C.A. Bickley, K.N. Stevens, and D.R. Williams, “A Framework for Synthesis of Segments Based on Pseudoarticulatory Parameters,” in Progress in Speech Synthesis, New York, Springer-Verlag, pp.211-220, 1997.
2) J. Allen, M.S. Hunnicutt, and D. Klatt, “From text to speech: The MITalk system,” Cambridge University Press, 1987.
3) 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, 1996.
4) N. Campbell and A. Black, “CHATR: 自然音声波形接続型任意音声合成システム,” 信学技報, vol.SP96-7, pp.45-52, 1996.
5) 籠嶋岳彦，赤嶺政巳，“閉ループ学習に基づく代表素片選択による音声素片の自動生成, ” 信学論，vol.J81-DII, no.9, pp.1949-1954, 1998.
6) E. Moulines and F. Charpentier, “Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones,” Speech Communication, vol.9, no.5, pp.453-467, 1990.

■2群 － 7編 － 3章

3-4 コーパスベース音声合成

（執筆者：河井　恒）［2009年5月 受領］

コーパス*ベース音声合成とは，(a)音声コーパスを用意し，(b)その中から音声素片を抽出し，(c)つなぎ合わせることによって音声を合成する方式である [1]（**図 3･2**）．「単位選択型音声合成」あるいは「波形（素片）接続型音声合成」とも呼ばれる．音声信号生成モデルのパラメータを規則のかたちで実装する「規則音声合成」を計算機による音声合成技術の第1世代とすれば，コーパスベース音声合成は第2世代といえる．

特長として，テキストのジャンルが限定される場合（例えば，株式市況，天気予報など）のように，音素の並び方，アクセント・イントネーションなどの多様性が小さい場合は，自然性の高い実用レベルの音質が得られる．反面，音声コーパスの格納に数十MB以上の記憶容量を必要とするため，携帯機器などへの組込みの用途には適さない．

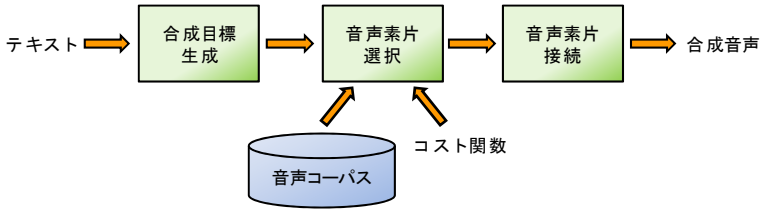

図 3･2　コーパスベース音声合成システムの構造

3-4-1 合成目標の生成

音声コーパスから音声素片を抽出する際に，適合性を判定するために参照する情報は，3種類に分けることができる．まず必須なのが音素列である．

第2のグループは，品詞，係り受け関係などの言語情報とアクセント型，強調など韻律情報である．第3のグループは，音素時間長，F_0†の値と傾き，パワー，スペクトル包絡などの音響パラメータであり，第2グループの情報を入力として，決定木など統計的手法により計算されることが多い．一般的には，第3グループの情報のみが用いられるが，第2グループのみ，あるいは第3グループと併用する方法もある．

3-4-2 音声コーパス

(1)　テキスト設計

音声コーパスの作成手順を**図 3･3** に示す．まず，発声内容となるテキストは，音声合成の適用領域（例えば，株式市況の音声化など）から多数の例文を収集して作成する．多様な音素連鎖，韻律的特徴が含まれるような文セットを自動的に作成する手法が提案されているが [2]，音声の総時間数が 2 時間を超える場合は，文をランダムに収集しても文セットを設計し

* 電子化された大規模な言語データ（音声を含む）
† 声帯振動の基本周波数．声の高さを表す物理量

たのと同等の効果が得られる．ただし，いずれの場合でも外来語など低出現頻度の音素については，欠落がないよう配慮が必要である．

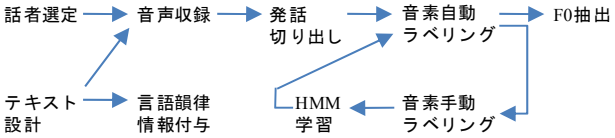

図 3・3　音声コーパスの構築手順

(2)　音声コーパスの規模と合成音品質

音声コーパスの規模が大きくなると韻律が自然で不連続感の少ない合成音声が得られる確率が高くなるが，10 時間前後で音質改善効果が飽和し始める [2]．一方，音声収録期間中の声質変動が原因となって不連続感生ずる確率も高くなるのみならず，収録に要する費用増大も不利な要因である．音質と費用のバランスを考慮すると，2～10 時間が現実的な規模である．

(3)　話者選定

音響パラメータの変動が小さい話者ほど高品質な合成音声を得られるため，試験的に音声合成を行ったうえで話者を決定することが望ましい．また，音声収録は長期間に渡るため，日内及び日間の声質変動が小さい話者を選ぶことも重要である．

(4)　音声収録後の音素ラベリングおよび F_0 抽出

音素ラベルに誤りがあると異聴や不連続感が生ずることがあるため，人手により音素ラベルを付与することが望ましい．しかし，多大な作業時間（音声時間長の 150 倍以上）を要するため，5 分程度の音声に対して人手により音素を付与して特定話者環境非依存 HMM を作成し，発声内容のカタカナ表記を参照して強制音素アライメントを行うこともよく行われる．ただし，その場合は多少の音質劣化が生ずる [3]．

コーパスの F_0 値は，合成音声に直接的に反映されるわけではないため，高精度である必要はなく，むしろ有声／無声誤り，gross error が小さい F_0 抽出手法が適している．

3-4-3 素片選択のためのコスト関数 ― 接続コストとターゲットコスト

個々の音声素片について合成目標からの乖離の大きさを表す数値がターゲットコストであり，音素時間長，F_0 の値と傾き，スペクトル包絡などの音響パラメータごとの差の絶対値（サブコスト）の重み付き和として計算される．サブコストの重みは，MOS 評価実験を行い，ターゲットコストと MOS 値の相関係数最大化するように決定する．

接続コストは，音声素片接続点における音響パラメータの不連続性の大きさと，その接続点における自然な不連続性の大きさの差を定量化した指標である．スペクトル包絡の距離尺度に関しては多くの研究があるが，ケプストラムのユークリッド距離が人間の主観的尺度との対応が最も良好であるとする報告が多い．

一つの音声素片に関してターゲットコストと接続コストの重み付き和を計算し，局所コストとする．次に，局所コストを 1 文内の全音声素片について加算し，最終的なコストとする．

3-4-4 素片選択のための探索処理

(1) 基本アルゴリズム

まず，入力文の音素ごとに音声素片の候補を音声コーパスから抽出する．次に，DP（Dynamic Programming）アルゴリズムによって文全体に渡ってコスト値が最小となる候補の組合せを決定する [4]．音声素片の単位の基本は音素であるが，母音定常部で接続することによって不連続感の発生を低減するために半音素が用いられることもある．

(2) 高速化手法 ― プルーニング，予備選択，枝刈り

10 時間規模の音声コーパスを用いる場合，1 音素の素片候補数は数万に上る．このため，実用上は高速化・省メモリ化のための対策が必要となる．

オフライン，つまり特定の文を音声合成するのに先立って行う対策として，プルーニングがある．これは，多数の文を試験的に音声合成し，音声コーパス中で使用頻度の低い音声素片を削除する手法である．

オンラインの対策としては，予備選択と枝刈りがある．予備選択は，ターゲットコストのいき値によって候補素片の足切りをする手法である．枝刈りは，DP 処理中に，コスト値のいき値によって文頭から現在の音素に至るパス候補の足切りをする手法である．

(3) 短遅延化

素片選択の計算時間は，文に含まれる音素数に比例するため，対話システムなどでは，テキスト入力から音声出力開始までの時間が問題となる．対策として，現時点の音素から 10 音素程度先まで探索した時点で最適パスを仮に決定し，さかのぼって現時点の波形素片を確定する手法がある [5]．

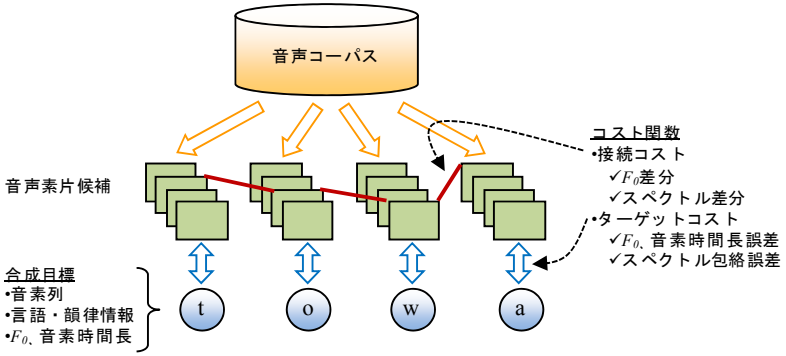

図 3・4 素片選択処理の概要

3-4-5 素片接続と韻律変形

波形素片を接続する際に振幅の不連続性による異音が問題となる場合は，接続予定時点の前後 5 ms 程度の範囲で短時間相互相関係数が最大となる時点を探索し，接続する．

音声コーパスの規模が小さい場合（おおむね 2 時間以下）は，合成目標と合成される音声の間の韻律パラメータの誤差による音質劣化が，信号処理による音質劣化を上回るため，PSOLA[6] などの手法によって韻律変形を行うことが望ましい．

■参考文献

1) 匂坂芳典，“コーパスベース音声合成技術の動向 [I],”信学誌, vol.87, no.1, pp.64-69, 2004.
2) 河井恒，津崎実，“コーパスベース音声合成技術の動向 [III],”信学誌, vol.87, no.3, pp.227-231, 2004.
3) 河井恒，戸田智基，“波形接続型音声合成のための自動音素セグメンテーションの評価,”信学技報, vol.102, no.619, pp.5-10, 2003.
4) A.J. Hunt, and A.W. Black, “Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database,” Proc. IEEE ICASSP96, pp.369-372, 1996.
5) 西澤信行，河井恒，“波形接続型音声合成における素片選択遅延時間の短縮,”信学論，vol.J90-D, no.1, pp.62-72, 2007.
6) Moulines, E., and Charpantier, F., “Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones,” Speech Communication, vol.9, nos.5/6, pp.453-467, 1990.

■2 群 － 7 編 － 3 章

3-6 声質変換・モーフィング

（執筆者：小林隆夫）［2009 年 8 月 受領］

声質変換（voice conversion）は，ある話者の音声を異なる話者の音声に変換する技術であり，音声に含まれる言語（音韻）情報を変えずに声質や個人性などの非言語情報を制御する技術といえる．既存のテキスト音声合成技術は，合成音声の音韻性や明瞭性といった言語情報の伝達という観点からは既に十分な性能が得られているが，任意の個人性，発話様式，感情などの非言語・パラ言語情報の付加・制御といった観点からはまだ不十分な点が多い．これに対し声質変換は，容易に非言語情報を制御可能にする技術として考えられ，異なる話者間で任意の中間的な仮想話者の音声を生成する音声モーフィング（voice morphing）を始め，対話システム，医療福祉機器，玩具，ゲーム，娯楽など様々な応用が期待できる [1]．

音声に含まれる個人性は，音声のスペクトル概形及び韻律に現れる特徴に依存する．スペクトル特徴は話者個人の調音器官の特性，すなわち声帯や声道形状などの身体的特徴に依存して決まり，主に話者の声質（voice quality）の違いとして現れる，一方，韻律特徴はイントネーションやアクセント，声の高さ，話速，音韻継続長などの違いとなって現れる．このため声質変換を実現するためには，スペクトル特徴及び韻律特徴の両方の変換が必要である．

3-6-1 スペクトル変換

ある話者（元話者）から異なる話者（目標話者）への声質変換を考え，時刻 t における元話者の音声のスペクトル特徴量（例えばメルケプストラム係数ベクトルや線スペクトル周波数ベクトルなど）とそれに対応する目標話者の特徴量をそれぞれ $\boldsymbol{x}_t$, $\boldsymbol{y}_t$ とする．スペクトル特徴の変換に着目した声質変換は，元話者のスペクトル特徴量をもとに目標話者に似せた特徴量を出力する変換関数 $\boldsymbol{y}_t \approx \mathcal{F}_s(\boldsymbol{x}_t)$ を求める問題として定式化でき，音声認識における話者適応と同様の問題となる．変換関数には，$\boldsymbol{A}_i$, $\boldsymbol{b}_i$, w_i をそれぞれ適当な変換行列，バイアス，重み係数，M をある正整数として，次式のかたちが多く用いられている．

$$\mathcal{F}_s(\boldsymbol{x}_t) = \sum_{i=1}^{M} w_i (\boldsymbol{A}_i \boldsymbol{x}_t + \boldsymbol{b}_i) \qquad (3 \cdot 1)$$

ベクトル量子化（VQ）に基づく手法 [2]では，元話者と目標話者が同一内容（音韻系列）を発声した学習用の音声データ（パラレルデータ）を用いてそれぞれの話者の特徴量 VQ コードブック $\{\boldsymbol{a}_j\}_{j=1}^{M}$, $\{\boldsymbol{b}_i\}_{i=1}^{M}$ を作成し，元話者と目標話者の間で VQ 後の特徴量（コードベクトル）の対応関係の度数分布 h_{ji} を求め，$P(i \mid j) = h_{ji} / \sum_{k=1}^{M} h_{jk}$ の値を計算しておく．声質変換を行う際には，元話者の入力音声の特徴量 $\boldsymbol{x}_t$ を VQ したコードベクトルのインデックス j を求め，式(3･1)で $w_i = P(i \mid j)$, $\boldsymbol{A}_i = \boldsymbol{0}$ とおいて変換後の特徴量 $\boldsymbol{y}_t \approx \mathcal{F}_s(\boldsymbol{x}_t)$ とする．

ガウス混合モデル（GMM）に基づく手法 [3]では，パラレルデータをもとに，まず元話者の特徴量を混合数 M の GMM によりモデル化する．ここで，GMM の i 番目の正規分布の平均ベクトル及び共分散行列を $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ として，式(3･1)で

$$w_i = P(i \mid \boldsymbol{x}_t), \quad \boldsymbol{A}_i = \boldsymbol{\Gamma}_i \boldsymbol{\Sigma}_i^{-1}, \quad \boldsymbol{b}_i = \boldsymbol{\nu}_i - \boldsymbol{\Gamma}_i \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i$$

とおいた変換関数を用いて変換誤差 $\varepsilon = \sum_t \| \boldsymbol{y}_t - \mathcal{F}_s(\boldsymbol{x}_t) \|^2$ が最小となるようにパラメータベクトル及び行列 $\boldsymbol{\nu}_i, \boldsymbol{\Gamma}_i$ を決定する．GMM と最小二乗誤差に基づく点は同じであるが，パラレルデータから得られた元話者と目標話者の対応する特徴量を結合した $\boldsymbol{z}_t = [\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{y}_t^{\mathrm{T}}]^{\mathrm{T}}$ を用いてモデル化した GMM に基づく手法[4]では，式(3･1)の変換関数において

$$w_i = P(i \mid \boldsymbol{x}_t, \{\boldsymbol{\mu}_j^{(xx)}\}, \{\boldsymbol{\Sigma}_j^{(xx)}\}), \quad \boldsymbol{A}_i = \boldsymbol{\Sigma}_i^{(yx)} \boldsymbol{\Sigma}_i^{(xx)-1}, \quad \boldsymbol{b}_i = \boldsymbol{\mu}_i^{(y)} - \boldsymbol{\Sigma}_i^{(yx)} \boldsymbol{\Sigma}_i^{(xx)-1} \boldsymbol{\mu}_i^{(x)}$$

としている．ただし，結合特徴量 GMM の i 番目の正規分布の平均ベクトルと共分散行列を

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^{(x)} \\ \boldsymbol{\mu}_i^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(xx)} & \boldsymbol{\Sigma}_i^{(xy)} \\ \boldsymbol{\Sigma}_i^{(yx)} & \boldsymbol{\Sigma}_i^{(yy)} \end{bmatrix}$$

とする．

これらの手法は，基本的に音声の分析フレーム単位で変換が行われるため，変換後の音声に聴感上の不連続が生じやすいほか，モデル学習のデータ量が十分でない場合に変換音声の明瞭性が低下したり，モデル学習にパラレルデータが必要となるといった問題がある．これらの問題を解決するために，複数フレームにわたる特徴量を導入した手法[5]やパラレルデータが不要な手法など，様々な改良手法が提案されている．

3-6-2 韻律変換

韻律特徴は声質とともに話者の個人性を左右する重要な要因である．しかし，スペクトル特徴の変換に比べると，韻律特徴の変換に関する手法は限られており，スペクトル変換と同様な VQ に基づく手法，基本周波数（F0）や話速の平均値を単純に目標話者の学習データの平均値に合わせる手法，分散を考慮した F0 の線形変換手法，HMM 音声合成を用いた韻律生成に基づく手法などがある．分散を考慮した F0 の線形変換では，元話者の対数 F0 値 x_t の平均及び標準偏差を $\mu^{(x)}$, $\sigma^{(x)}$，目標話者の対数 F0 値 y_t の平均及び標準偏差を $\mu^{(y)}$, $\sigma^{(y)}$ として，元話者から目標話者への変換関数 $y_t \approx \mathcal{F}_p(x_t)$ に次式を用いる．

$$\mathcal{F}_p(x_t) = \left(\sigma^{(y)} / \sigma^{(x)}\right)\left(x_t - \mu^{(x)}\right) + \mu^{(y)} \qquad (3 \cdot 2)$$

3-6-3 スペクトル・韻律変換

声質変換において言語情報があらかじめ与えられる場合には，HMM 音声合成と話者適応に基づいてスペクトル特徴と韻律特徴双方の変換が可能である．HMM 音声合成では，スペクトル，F0，音韻継続長を同時にモデル化した音声単位 HMM を用いており，音声認識で用いられる話者適応と同様な手法に基づいて音声単位 HMM をモデル適応することにより，スペクトル特徴と韻律特徴を含む任意の話者性の音声を生成できる．元話者のモデルとして複数話者の音声データから学習した平均声モデルを用いることで，任意の目標話者が発声した少量の適応データがあれば，当該目標話者に近い話者性を持った音声が合成できる[6]ほか，発話様式や感情表現を含むパラ言語・非言語情報の制御への応用も容易である．

■参考文献

1) Yannis Stylianou, “Voice transformation: A survey,” 2009 ICASSP, pp.3585-3588, 2009.

2) M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, “Voice conversion through vector quantization,” J. Accoust. Soc. Jpn. (E), vol.11, no.2, pp.71-76, 1990.
3) Y. Stylianou, O. Cappè, and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech and Audio Process., vol.6, no.2, pp.131-142, 1998.
4) A. Kain and M.W. Macon, “Spectral voice conversion for text-to-speech synthesis,” 1998 ICASSP, pp.285-288, 1998.
5) T. Toda, A. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” IEEE Trans. Audio, Speech, Lang. Process., vol.15, no.8, pp.2222-2235, 2007.
6) J. Yamagishi and T. Kobayashi, “Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training,” IEICE Trans. Inf. & Syst., vol.E90-D, no.2, pp.533-543, 2007.

■2群 － 7編 － 3章

3-7 韻律の生成

（執筆者：匂坂芳典）［2009年8月 受領］

声の高さ，速さ，強さを表す韻律は，各々，主に声帯の基本周波数，音素に対応するセグメントの持続時間長(以下,音韻長と呼ぶ),音声振幅といった音響特徴量によって担われる．合成音声に適切な韻律を与えるために，韻律制御の仕組みについての分析，数理モデル化がなされている．以下に示すように，各特徴量値は，入力テキストなどから得られる言語情報をもとに，実際の音声データに基づく統計的なモデルを用いて算出される．

3-7-1 声帯の基本周波数の制御

声の高さは，アクセントやイントネーションを表出するうえで重要であり，声帯の基本周波数（F0）の制御によって実現される．特に，日本語のアクセントはいわゆる高低アクセントであり，**図 3･5** の実際の F0 の時間変化例（図中＋で示す変化）に見られるように，「スーツケースと」と「航空カバン」といったアクセント核直後に急な下降を伴う局所的な F0 の起伏として実現される．また，より広い範囲にわたる全体的な下降特性で代表される，イントネーションとしての F0 の制御がなされる．この結果，例えば図 3･5 中の「航空カバン」は直前の「スーツケースと」に引き続いて全体的な下降特性を示す．これらは図 3･5 が示すように，アクセントに起因する局所的な起伏（アクセント成分，実線）とそれらをまとめた大局的な下降特性を示す話調（フレーズ成分，破線）との重畳により，実際の F0 の時間変化特性がよく近似できることが知られている [1-3]．特に，指令応答モデル [2-3]では，F0 生成機構の機械的なモデルに基づき，アクセント成分とフレーズ成分の生起指令時刻とその大きさといった少ない生成パラメータにより，F0 の動特性を規定でき，日本語のみならず多くの言語でその有用性が確認されている．

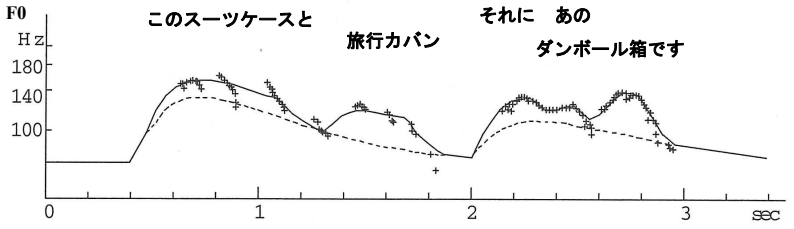

図 3・5 指令応答モデルによる基本周波数（F0）パターンの表現

音声合成では入力文が持つ言語情報，主に文節を構成する単語の性質によって文節のアクセントが決定され，それに基づいた局所的な F0 の制御がなされる．この制御には，アクセント句の長さ（拍数），アクセント型などが関与することが知られている．また，文節の依存関係や句の構造に基づいて形成される話調境界とその大きさが決定される．実際の合成システムでは，統計的なモデルを用い，入力言語情報に対応して観測される F0 時間変化特性が実現されている．統計的モデルとしては，F0 生成機構を陽に取り入れたものや，内在的に取

り入れているもの，F0 時間変化特性そのものをモデル化したものが用いられている．統計的な手法としては，HMM（隠れマルコフモデル）に代表されるような最適化アルゴリズムを備えた数理モデルにより，入力である句のアクセント型，句長，隣接句間の関係などの言語情報と出力としての F0 時間変化との対応を大量の音声コーパスでモデル学習する手法がとられている [4]．

3-7-2 音韻長，ポーズ長の制御

音韻セグメントの持続時間長の設定には種々の要因が関与し，日本語では**表 3･1** に示されるような制御要因の存在が知られている [5]．日本語では，拍（モーラ）を単位とした時間制御特性（リズム）が特徴的であり，英語などに見られるストレス単位リズムとの対比をなす．前述した基本周波数の場合と同様に，大規模な音声データベースを用い音韻セグメント長モデルの最適化が試みられている．計算モデルとしては，最適化のための統計的手法としてよく知られた線形回帰，回帰木，及びこれらを発展させた方法が提案されている [6-10]．線形回帰のモデルとしては，連続値をとる制御要因をカテゴリ値に適用できるようにした，数量化第 I 類が用いられている [6, 7]．このモデルでは次式で示すように，各制御要因についてカテゴリごとの特性関数δfc の一次式と全音韻のセグメント長平均値μの和によって当該サンプルの音韻セグメント長の dur を表す．

$$\mathrm{dur} = \mu + \sum_{f}\sum_{c} X fc\,\delta fc$$

ここで，f は表 3･1 に示したような，音韻種類，音韻環境，発話区分長，発話区分内位置などの制御要素を表し，c は各制御要素のカテゴリを表す．特性関数δfc は当該サンプルが一致する制御要素カテゴリのみ 1 の値をとり，そのほかでは 0 となる．Xfc は制御要素 f カテゴリ c の寄与度を示す係数であり，この値は前記の式によって予測された音韻時間長と実測された時間長の平均 2 乗誤差を最小化するように求める．

表 3･1　音韻セグメント長に影響を与える要因

影響範囲	観測される音響的特徴	影響要因
当該音韻	固有平均長，伸縮傾向の相違	調音上の制約
近傍音韻モーラ	隣接音韻間の時間長補償，長短リズム	モーラ・タイミング
単　語	内容語伸張・機能語短縮	単語の重要度
発話区分頭・末尾	句・呼気段落末伸長，区分頭短縮	発話区分境界の明示
発話区分全体	句・呼気段落内モーラ数増加に伴う音韻長短縮	発話区分内テンポ
文全体	発話区間全体の伸縮	発話テンポ

英語のように音素数が多く複雑な音節構造をもつ音声言語では，大量のデータを必要とし，多くの要因を一度に考慮した線形回帰モデルをそのまま適用することは難しいため，種々の工夫がなされている．なかでも回帰木は広く用いられている．回帰木では，データ中に見られる要因による分布の偏りをもとに，統計的に有為な範囲でモデルが得られるため，データ量に応じたモデル化が可能となる [8]．このほかにも，要因間の依存関係を統計的に調べ，要

因変数の双一次形式で音韻セグメント長を表現した積和型のモデル [9]，線形回帰を拘束条件付の回帰木としてみなし，その拘束条件を緩めた部分的拘束条件付回帰木 [10]といった統計的モデルの工夫もなされている．

文章構造の影響は F0 同様，音声生成の時間制御にもみられる．特に，句境界にみられる休止（ポーズ）区間の挿入傾向・時間長に顕著に現れる．ポーズの挿入には自由度があり，読み方や発話者によって挿入箇所・個数は変化する．右枝分かれ境界では多くの話者が比較的長いポーズを挿入し，左枝分かれ境界でのポーズ挿入は話者によるバラツキが見られ，比較的短いポーズが多い．先に述べた F0 のフレーズ成分とアクセント成分の違いに対応するような二種類のポーズ区間長の存在が確認されており，それらが構文構造に深く関係することが知られている．

3-7-3 振幅の制御

基本周波数，音韻長の制御に比較して，振幅の制御は，文章の読み上げなどの目的のためには大きな品質低下要因とはなっていなかった. このため, あまり調べられてきてはいない．音韻長の制御同様に，数量化第 I 類を用いた線形回帰のモデルにより，基本周波数，隣接音韻，呼気段落内位置，当該音韻種類などによる影響が報告されている [10]．合成システムへの組み込みには F0，音韻長と共に，HMM を用いた最適化が用いられている [4].

これまで，音声合成では，書き言葉を対象として，文章の読み上げを中心とした応用に向けて技術開発が進められてきた．合成音声の高品質化に伴い，人間とのインタフェースとしての音声出力として，読み上げ音声に加え，話し言葉としての音声の要求が高まってきている．しかしながら，人間を模倣した話しかけ音声を実現するためには，文自動生成も含む自律的な対話機能が不可欠であり，完全な実現は難しい．話し言葉としての韻律は，発話者の意図や感情を如実に表す重要な情報であり，音声コミュニケーションとしての豊かな情報伝達を実現するため，対話音声の韻律生成をはじめとする韻律の多様性に関する制御実現への探索が続けられている．

■参考文献

1) 箱田，佐藤，“文音声合成における音調規則，”電子通信学会論文誌, vol.J63-D, no.9, pp.715-722, 1980.
2) 広瀬，藤崎，河井，山口，“基本周波数パターン生成過程モデルに基づく文章音声の合成，”電子情報通信学会論文誌，vol.J72-A, no.1, pp.32-40, 1989.
3) 河井，広瀬，藤崎，“日本語文章音声の合成のための韻律規則，”日本音響学会会誌，vol.50，no.6, pp.433-442, 1994.
4) 徳田，益子，小林，今井，“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，”日本音響学会論文誌，vol.53, no.3, pp.192-200, 1997.
5) 匂坂，東倉，“規則による音声合成のための音韻時間長制御，”電子通信学会論文誌，vol.J67-A，no.7, pp.629-636, 1984.
6) K. Takeda, Y. Sagisaka and H. Kuwabara, “On sentence-level factors governing segmental duration in Japanese,” JASA, vol.86, no.6, pp.2081-2087, 1989.
7) 海木，武田，匂坂，“言語情報を利用した母音継続時間長の制御，”電子情報通信学会論文誌，vol.J75-A, no.3, pp.467-473, 1992.
8) Riley M.D., “Tree-based modeling of segmental durations,” in ‘Talking Machines’ , Elsevier Science Publishers B.V., pp.265-273, 1992.
9) van Santen J.P.H., “Contextual effects on vowel duration,” Speech Communication, vol.11, pp.513-546, 1992.

10) N. Iwahashi and Y. Sagisaka, “Statistical modeling of speech segment duration by constrained tree regression,” IEICE Trans., vol.E83-D, no.7, pp.1550-1559, 2000.
11) 三村，海木，匂坂，“統計的手法を用いた音声パワーの分析と制御，”日本音響学会論文誌，vol.49, no.4, pp.253-259, 1993.

■2 群 － 7 編 － 3 章

3-8 韻律コーパス

（執筆者：前川喜久雄）［2009 年 8 月 受領］

音声認識ないし合成に利用される音声コーパスは，主に音声の分節的特徴（子音や母音とその結合に関する特徴）に着目して設計される．これに対し，韻律コーパスは音声の韻律的特徴に着目して設計されたコーパスである．韻律的特徴には，アクセント，ストレス，声調，プロミネンス（卓立），イントネーションなどが含まれる．言語学的には持続時間長の情報も韻律的特徴の一部であるが，これは通常の音声コーパスでも提供される情報である．

3-8-1 韻律的特徴の階層構造

音声の韻律的構造の記述にあたっては，通常，階層構造を想定する．日本語（東京方言）の場合，最下位の階層は通常モーラ（M）であり，そこから，音節（S），語（W），アクセント句（AP），中間句（IP）などを経て，最上位階層である発話（Utt）に至る構造を想定する [1]．

東京方言には「飴」と「雨」のようなアクセントの対立があり，アクセントをもつ語ではアクセントの位置でピッチが急激に下降する．また，名詞と名詞が結合して複合名詞を構成したり，名詞と助詞が結合して文節を構成する際に構成要素である語のアクセントが変化することから分かるように，語と語の交互作用を考える必要がある．アクセント句は多くの場合に文節に該当する大きさの単位であり，アクセントの交互作用の領域を示すと同時に，句末イントネーションが生ずる領域でもある‡．

アクセント句頭ではピッチの上昇が生ずる．アクセント句内にアクセントが存在すれば，ピッチはそこで下降するが，アクセントが存在しなければ句末まで高いピッチレベルが続く．そしてアクセント句末には，句中におけるアクセントの有無とは無関係に，様々な句末イントネーション（上昇調とその様々な変種，上昇下降調，上昇下降上昇調など）が生ずる．

東京方言ではアクセントによってピッチが下降するため，それ以降の発話のピッチレンジがせばめられるが，1 発話内に複数のアクセントが存在する場合，ピッチレンジが再帰的にせばめられることがある．この現象はダウンステップと呼ばれ，文の修飾関係や語の間の意味的限定の有無などの伝達に利用されている．中間句（intermediate phrase）はアクセント句と発話の中間にあって，ダウンステップが持続する領域を示す [1]．

図 3･6 は東京方言の疑問詞疑問文「何が見える？」と単純疑問文「何か見える？」の韻律階層構造とそこから生成される典型的な音声基本周波数パターンである．両者を構成するトーン（次節参照）は同一であり，いずれも 2 個のアクセント句から構成されているが，前者では発話全体が 1 個の中間句から構成されているため，疑問詞「何」のアクセントが引き起こしたダウンステップによって述語句「見える」のピッチレンジがせばめられている．一方，後者は 2 個の中間句から構成されているので，「何か」のアクセントによって引き起こされたダウンステップが中間句境界でリセットされる．そのため，音声基本周波数パターンには顕著なピークが 2 個観察される．

‡ 参考文献 1)では句末イントネーションは「発話」階層を領域として生ずると記述されているが，これは正しくない．図 3･6 でも上昇イントネーションを担う H%トーンはアクセント句に所属させてある．

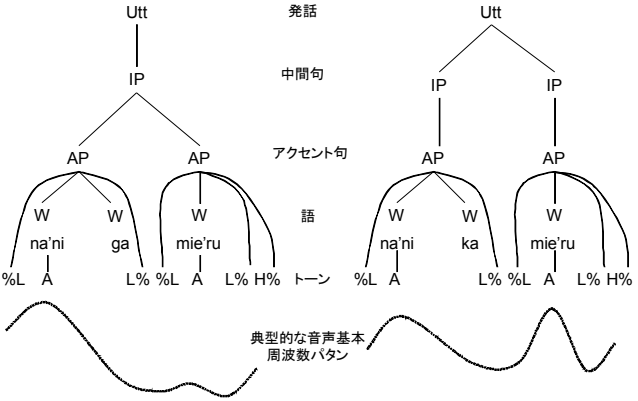

図 3・6　発話の韻律階層構造と音声基本周波数パターンの関係

3-8-2 ToBI による韻律的特徴のアノテーション

コーパスへの韻律的特徴のアノテーション方式として，現在最も普及しているのが ToBI 方式である．ToBI は Tones and Break Indices の略称であるが，ここでトーンとは，発話に伴う音声基本周波数パターンを記述するために利用される抽象的なピッチレベルの指定を意味しており，中国語などの声調言語におけるトーン（語に対するピッチレベルの指定）とは異なる概念である．東京方言の場合，アクセントのほかに，アクセント句境界を表現する境界トーン（L%, H%など）とアクセント句頭におけるピッチ上昇のピークを表現するトーン（H-）が用いられる．日本語の場合，いずれのトーンも抽象化された高低二段階のトーン（H, L）の組合せで表現される（これは英語にあっても同様である）．

Break Indices（BI）とは発話の韻律構造の階層構造から導かれる構成素境界の強度（階層の深さ）を整数で近似的に表現したもので，日本語であれば，語境界が 1，アクセント句境界が 2，中間句境界（及び得発話境界）が 3 に指定される．

日本語の ToBI としては，まず朗読音声を主な対象とした J-ToBI[2)]が提案され，音声合成のための韻律特徴の研究に利用された．次いで『日本語話し言葉コーパス』[3)]の韻律アノテーションのために，J-ToBI を自発音声用に拡張した X-JToBI[4)]が提案された．現在では X-JToBI をただ Japanese ToBI と呼ぶこともある．**図 3･7** に X-JToBI による『日本語話し言葉コーパス』のアノテーション例を示す．最上段は音声波形，その下がサウンドスペクトログラムと音声基本周波数曲線であり，その下に X-JToBI ラベルが 3 層に分けて表示されている．上から単語層（語境界と語の音素構成を示す），BI 層，Tone 層である（X-JToBI にはほかに分節音層，卓立層などがあるが本例では省略）．アクセント句境界を縦線で，中間句境界を二重縦線で示し，アクセントをアポストロフで，アクセント句末の上昇イントネーションを上向き矢印で示すと，この発話は「ユージントフタリ'デ‖オ'ーランドノ‖ディズニーワ'ールドニ↑|イッテキマ'シタ‖」である．BI 層中の「2+bp」は X-JToBI の拡張仕様で，この境界にはポーズと句末イントネーションが伴っているため，通常よりも強いアクセント句境界であることを示している．

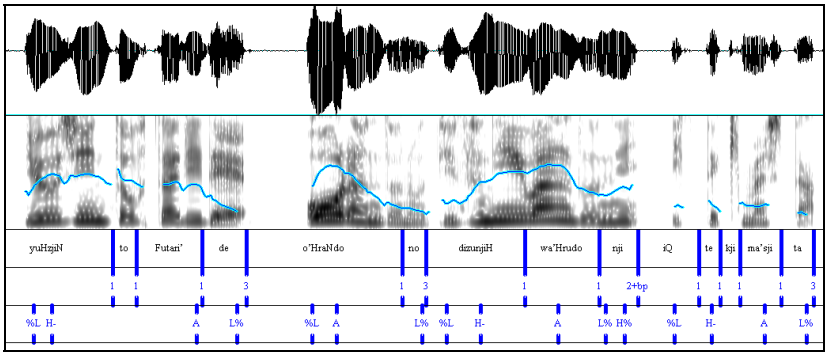

図 3・7　X-JToBI による韻律アノテーションの例

（※画像をクリックすると音声が流れます）

ToBI は対象言語の言語学的特徴に立脚したアノテーション方式であるから，対象言語（方言）が変われば，新たなシステムを考案する必要がある．これまでに ToBI が考案された言語には，英語，ドイツ語，朝鮮語，ギリシア語，ポルトガル語，カタロニア語などがあるが，韻律的特徴のアノテーションが施された音声コーパスは世界的にみても僅少である．日本語には先述の『日本語話し言葉コーパス』のほかに，朗読音声，感情音声を対象とした『MULTEXT 日本語版』があり [5]，J-ToBI によるラベリングが施されている．

3-8-3 韻律コーパスの応用

語の意味にかかわる情報（言語情報）の伝達という観点から評価すれば，韻律的特徴の重要度は分節的特徴よりも低い．しかし，韻律的特徴は，強調などの談話的情報を豊富に伝達しており，更に話者の身体性（性別や年齢など），感情などの非言語情報，話者の心的態度にかかわるパラ言語情報を豊富に伝達している．談話情報とパラ言語情報は対話処理ないし意味理解にとっては言語情報と並んで重要な情報であるから，韻律特徴の適切な処理は今後の音声研究の課題として第一級の重要性をもっている．韻律コーパスを構築する目的も多くの場合そこにある [6]．

■参考文献

1) J. Pierrehumbert and M. Beckman, “Japanese Tone Structure,” MIT Press, 1988.
2) J. Venditti, “Japanese ToBI Labelling Guidelines,” Ohio State University Working Papers in Linguistics, vol. 50, pp.127-162, 1997.
3) 前川喜久雄, “『日本語話し言葉コーパス』の概要,” 日本語科学, vol.15, pp.111-133, 2004.
4) 前川喜久雄, “話し言葉コーパスの韻律ラベリング,” 広瀬編『韻律と音声言語情報処理』, 丸善, pp.85-93, 2006.
5) http://minny.cs.inf.shizuoka.ac.jp/~multext/index-j.html
6) J. Venditti, K. Maekawa and M. Beckman, “Prominence marking in the Japanese intonation system,”, S. Miyagawa and M. Saito (eds.), The Oxford Handbook of Japanese Linguistics, Oxford University Press, pp.456-512, 2008.