☆1 さらに文全体の中で文節相互間の修飾関係や，対話文などで聞き手に訴えかけたい焦点となる文節が，その文のイントネーションに影響を与えることが知られており[1]，係り受け解析や談話解析で得られる情報を利用してよりきめ細かい韻律情報を生成し，品質の高い合成音声を出力できる可能性がある．しかし解析技術の限界から本格的には利用できていない．

解説

表現力の豊かさを目指した音声合成技術

小原　永
NTT情報通信研究所

合成音声というと，「ロボット音声」とも呼ばれる機械的な音声をイメージする読者は多いかもしれない．しかし，より自然な抑揚や，喜怒哀楽といった感情のこもった表現力を持つ合成音声を作り出す研究が着実に進められ，一部はすでに実用製品に組み込まれる段階まできている．ここでは，一般的な日本語音声合成技術についてまず概観し，これら技術基盤の上で，自然性などの表現力を持った合成音声の発声を狙ったCHATRとSpeed97という2つの音声合成技術について，その技術的ポイントを紹介する．

日本語音声合成技術の概観

音声合成技術の主な適用分野に，漢字仮名混じり文を入力して合成音声で出力するTTS（Text To Speech）がある．TTSは実現技術，具体的なシステムにより差異はあるものの，概念的には**図-1**に示す通り，自然言語処理技術に基づくテキスト解析処理と音響的な処理を中心とした音声合成処理から構成される．ここではまずテキスト解析処理が音声合成処理に渡すインタフェース情報について述べる．次にテキスト解析処理が，入力されたテキストからこのインタフェース情報をどのように生成するかを述べ，最後にこのインタフェース情報を使って合成音声処理がどのように合成音声を生成するかを述べる．

インタフェース情報

テキスト解析処理が音声合成処理に渡すインタフェース情報には，図-1に示す通り，音韻情報と韻律情報がある．音韻情報は読みであるが，一般的な振り仮名ではなく，長音（「先生」を「センセー」），濁音/鼻濁音の区別（「学校」の「ガ」と「漫画」の「ガ」），母音の無声化（「あした」の「シ」）など発音の違いを明示的に区別した情報である．入力されたテキストはアクセント句という単位に分割される．韻律情報は個々のアクセント句のアクセントに関する情報と，アクセント句相互の結合の状態を示す情報（たとえばポーズの挿入など）から構成される☆1．

日本語を自然に話そうとする場合，いくつかの単語がひとまとまりとなって発声される．このひとまとまりをアクセント句と呼んでおり，アクセント句単位で固有のアクセント型を持つことが知られている．一例として日本語東京方言のアクセント型を**表-1**に示す．ここで示される高低の意味は日本語東京方言の場合，ピッチ周波数の違いとして表される．一方モーラという単位はほぼ1つの読みに対応するが，長音「ー」，撥音「ン」，促音「ッ」は独立して1モーラと数える．表-1に示す通りアクセント句はモーラ数に関係なく，アクセント核と呼ぶ，ピッチ周波数が下降し始めるモーラの位置で分類できる．たとえば「水（ミズ）」と「隣村（トナリムラ）」は，いずれも0型（あるいは平板型）のアクセント句である．一方アクセント句相互の結合に関する情報は，アクセント句相互間に設定する無音時間の量と，先行するアクセント句発声後に次のアクセント句を発声する場合に，末尾モーラのピッチの影響を受けて発声するか，あるいは一度リセットして低いピッチから発声するかという，いわゆるピッチの建て直しが起こる／起こらないという2者択一値の，2つのパラメータの組合せとして表現する．無音時間の量は，0，すなわち続けて発声する場合から，文末など十分な間隔を開ける間を1～2段階程度に分けた分解能で示すのが一般的である．また通常ピッチの建て直しの生起は無音時間0の場合にのみ選択するなど縮退した組合せの情報

☆2 例外として「日本／電信電話株式会社」など，名詞が連続する複合語を1単語と認定した場合，1単語中に，「／」で示す境界で分割された複数のアクセント句を形成する場合がある．この場合には，解析辞書にこの情報を掲載するなどの手段で対応しているのが一般的である．

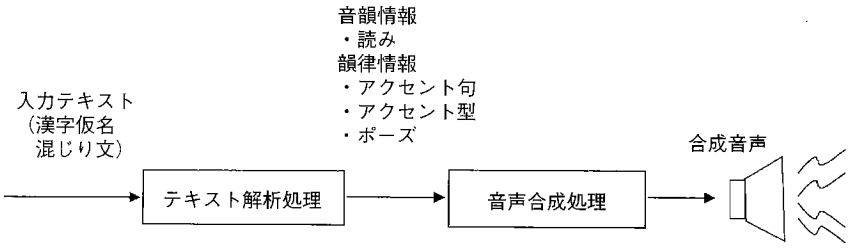


図-1 一般的なTTS（Text To Speech）システムの構成

型＼単語帳	1モーラ	2モーラ	3モーラ	4モーラ	5モーラ
0型（平板型）	（低）葉	水	桜	お花見	隣村
1型	（高）木	春	緑	3月	お月様
2型		山	お菓子	飲み物	おかあさん
3型			休み	湖	山桜
4型				妹	渡し船
5型					桃の花

表-1 日本語東京方言のアクセント型[2]

nモーラの0型とn型は同じ型であるが，後続の助詞などがあるとき，n型では，nモーラでピッチの下降が起こる

で示す．

テキスト解析処理

実用システムでは，上記音韻情報と韻律情報を生成する自然言語解析手法として形態素解析処理が用いられる．形態素解析処理は，分かち書きされていない漢字仮名混じり文を単語単位に分割し，個々の単語に品詞や読みなどの情報を付加する処理であり，一般的な漢字仮名混じり文に対して実用上差し支えない程度の解析精度と処理速度で解析できる技術が確立されている[3]．

テキスト解析処理では，形態素解析処理により得られた単語情報からアクセント句を生成する必要がある．アクセント句は自然言語処理の立場からみると，自立語（1つ以上の付属語以外の単語），あるいは自立語と付属語（1つ以上の助詞，助動詞の単語）から構成される文節とほぼ同等であり，一般的には品詞の接続などの組合せをよりどころとして生成する☆2．

音韻情報は単語辞書より得た情報を元に生成するが，同型異音語（あちらの方→アチラノホウ／アチラノカタ，など）の完全な曖昧性解消は形態素解析処理の範囲では困難である．一方，連濁（目覚まし＋時計→目覚まし時計）や接辞での音韻変化（1＋本，2＋本，など）は，アクセント句形成段階で単語辞書から得られた情報などを元に，ヒューリスティック規則を用意して対応する．

韻律情報では，アクセント型とアクセント句相互の結合状態に関する

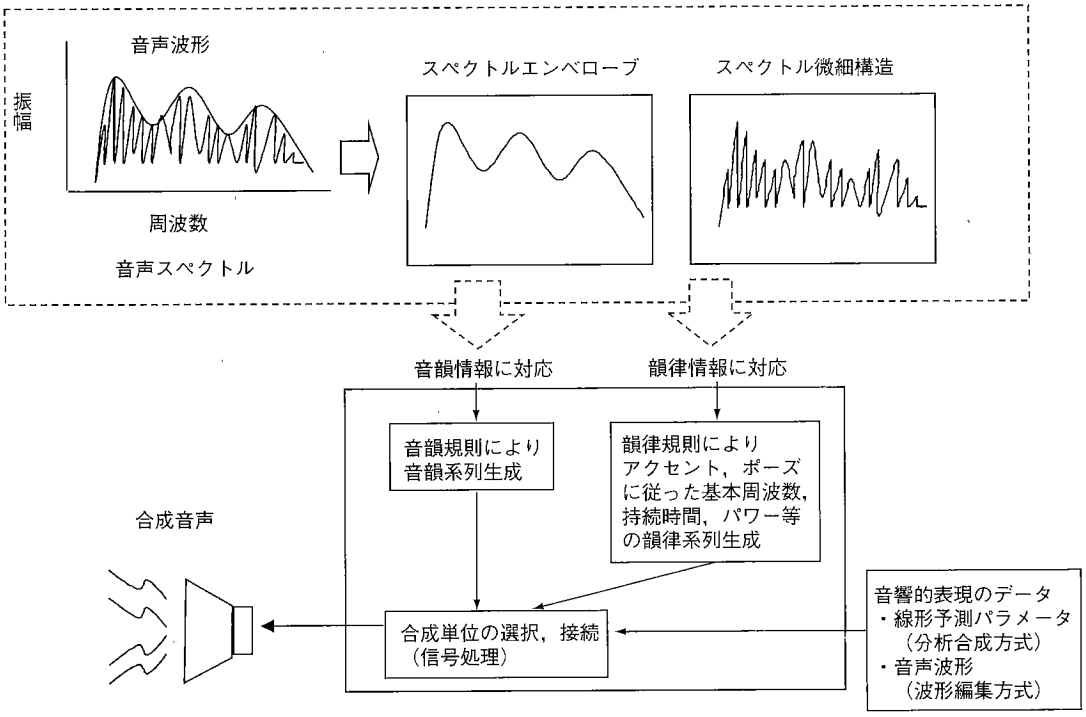


図-2　音声合成処理の構成

情報を生成する必要がある．アクセント型は，あらかじめ単語辞書に格納しておき，この情報を元にして生成するが，アクセント句の形成により，単語が本来持っていたアクセント核の消失，移動（百点＋満点→百点満点，など）が起こるため，この変形に対応できるヒューリスティック規則が用意される．アクセント句相互の結合状態については，形態素解析処理の範囲に閉じることから文の埋め込みなどの構造を認識するにはおのずと限界があり，句読点の存在，モーラ数などをよりどころとしてヒューリスティックルールで生成するのが一般的である．

音声合成処理

音声波を音声スペクトル分析し，音声波形の持つ周波数スペクトル構造を明らかにすると，比較的ゆっくり変化するスペクトルエンベロープ（包絡：spectral envelope）と，短時間で変化するスペクトル微細構造に分離できる．スペクトルエンベロープは声道の共振特性に対応しており，音素などの単語や文を構成する個々の音の特徴を示す．一方，スペクトル微細構造は音源特性に対応しており，アクセントやイントネーションを表す音響的な特徴である基本周波数F0を示す．

音声合成処理の基本的な処理は，実現技術，システムにより差異はあるものの，この分析に基づいて，**図-2**に示す通り，音韻情報から音声の音韻系列を選び出す処理と，韻律情報からアクセント，リズム，イントネーションを表現した音声波形を生成するための韻律系列を求める処理の，2つの処理の組合せにより実現する．前者を狭義の意味で音声合成方式と呼ぶことがあり，このうちTTSを目標とした実用技術としては，自然音声の分析結果を元に作成した音響的表現をあらかじめデータベースに蓄積しておき，合成の段階で適切なものを取り出し信号処理により接続する方法として，分析合成方式，波形編集方式が順次実用化されている[4]．

音韻情報に基づく音声の合成単位の選定処理

蓄積する音響的表現の選択と，その音響的表現をどういう単位で蓄積するかという合成単位の決定が重要になる．ピッチ制御が容易などの理由から音響的表現として線形予測パラメータ（LPC: Linear Predictive Coding）を利用した分析合成方式がすでに実用化されている．しかし，音質が不自然，合成音の明瞭性が低く了解性に欠けるなどの問題があった．この問題を解決する手段として，直接音声波形を接続していく手法が考えられるが，この接続のための波形変換処理で，スペクトルの影響から韻律などの側面で音声品質が劣化するという問題があった．この問題に対して，波形データのピッチや時間長を制御する方式が開発され[5]，さほど合成音声品質を落とすことなく波形を接続して合成音を生成することが可能となり，収集した音声波形を直接音響的表現として蓄積，利用することで，自然音声に近く，し

☆3 連続する3つ組音素のことをいう．日本語の場合約15,000種類存在する．その中でCVC(Consonant-Vowel-Consonant)の組合せが一番多く，約5,800種類存在する．この他にVCV，CVV，VVVなどがある．

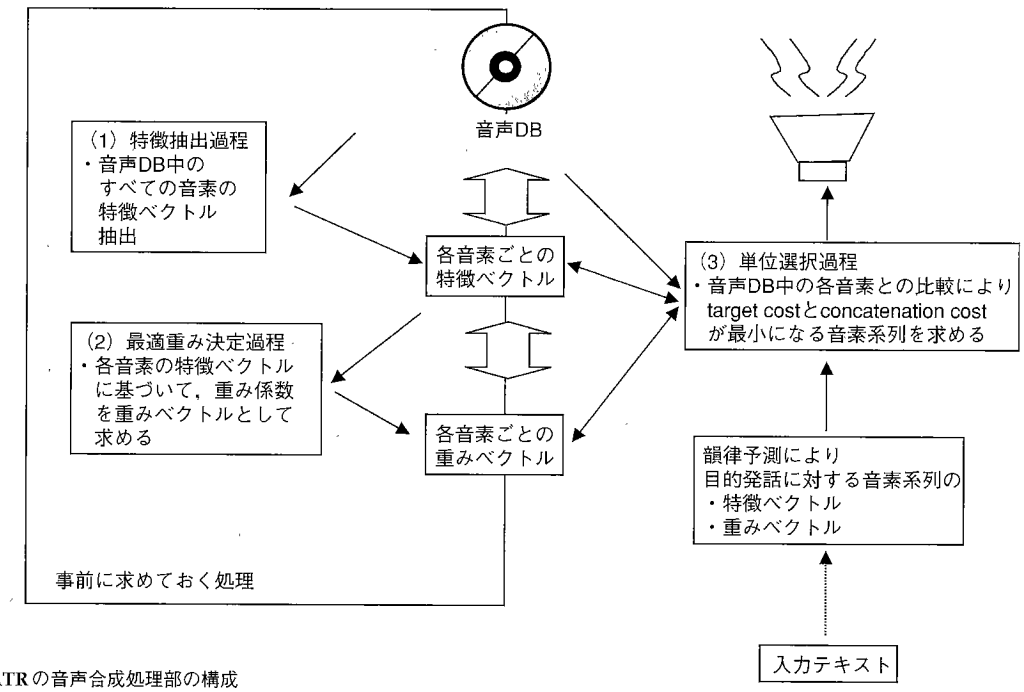


図-3 CHATRの音声合成処理部の構成

かも明瞭性に優れた波形編集方式が実用化された．

一方，合成単位はできるだけ長いほうがより自然な合成音声を生成できるものの，TTSのような読み上げる文が特定できない用途では，用意すべき数が爆発的に増大し現実的でない．このためその合成単位は音素レベル程度とする必要がある．連続音声中の音素の特徴が，隣接音素の特徴によって変形することが調音結合現象として知られており，音素レベルで調音結合するためには，最低限その音素に対して前後の音素を考慮したTri-phone☆3をすべて用意する必要がある．合成単位の作成は，このTri-phoneを含む無意味単語セットをナレータに発声してもらい，その録音音声から，中心音素ごとにセグメントを行った音素波形をピッチマーク情報と共にDBに蓄積することで実現する．

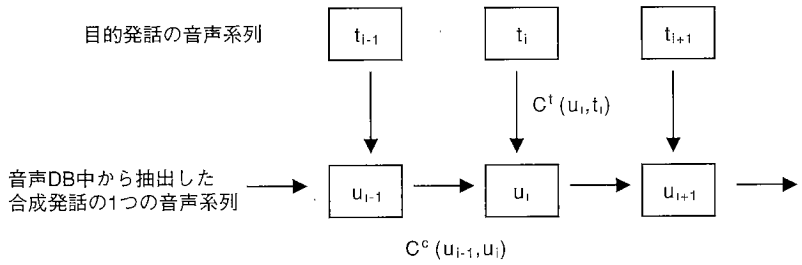
韻律情報に基づく音声波形の生成処理

合成音声波形の生成では，アクセント，イントネーションの変化を示す基本周波数F0パターンの生成が重要である．アクセント型によって決定したアクセント句内に閉じた局所的アクセント句成分と，アクセント句相互の関係に基づいて決定するイントネーションに対応する文のフレーズ成分を独立に求め，これを重畳することによって基本周波数パターンを生成する手法が提案されている[6]．フレーズ成分は，文の係り受け関係などの利用でより自然性の高いパターンを生成できる可能性を持つが，テキスト解析処理でも述べた通り，実用システムでは句読点やアクセント句のモーラ長に基づく選択程度にとどまっている．実現技術，システムによる差異はあるものの，概念的には選定された音素単位（波形編集では音素波形）の系列をこの重畳モデルに当てはめながら個々の波形，および波形相互の接続を信号処理により波形変形することで，目的とするアクセント，イントネーションを持った音声合成波形を生成する．

より自然で表現力豊かな合成音声生成の試み

なお一層の自然な韻律の再現が重要な課題であり，音声言語現象を観察し，韻律を再現するために必要な規則を導出する必要がある．このためには (1) より自然な韻律を再現する要因となる特徴パラメータの選定，(2) 韻律記述指標の開発，(3) 韻律記述指標に従った大量の分析データの作成，さらに (4) 大量分析データからの韻律規則の導出，と検討すべき課題は多い．たとえば音素の継続時間は，文中での位置，モーラ数，品詞など種々の要因が関係しているため発見的手法では定式化が困難であるなど，特に (3)，(4) において統計的手法に基づいた研究が進められている．

自然性という点で特徴的な音声合成システムもすでに実現されている．ここでは喜怒哀楽などの感情表現生成までも狙ったまったく異なる2つ



$C^t(u_i,t_i)$：音声DB中の音声単位u_iと，合成音声として実現したい音声単位t_iとの間の差の予測値
$C^c(u_{i-1},u_i)$：接続単位u_{i-1}とu_iとの間の接続で起こる不連続の予測値
この両者のコストが最小になるような音声系列を求める

図-4　単位選択過程の処理

のアプローチについて述べる．

音韻自然音声波形接続型任意音声合成システムCHATR[7)]

CHATRの狙い

CHATRは広義には図-1に示したTTS（および音声対話システム）として具備すべき言語処理，音響処理を含んだ音声合成技術研究用のワークベンチの総称である．ここでは特徴的である音声合成処理で実現している技術について述べる．

CHATRの音声言語処理方式は，音声波形は音響的，韻律的環境によって一意に決まるものという考え方に基づいて実現されている．音韻情報から求めた音素系列を，韻律情報に基づいて波形処理するという従来の考え方を否定した技術である．CHATRは音声合成システム自体が音声波形を生成することは基本的には行わず，音響的および韻律的環境が最も適する波形をそのまま利用するという手法をとっている．

CHATRの音声DB

CHATRの大きな特徴の1つに音声データベースがある．先に述べた通り従来の波形編集方式は，音韻的にみて十分な量の単語をプロのナレータが読み上げることで作成していたが，韻律的なバリエーションについては，むしろニュートラルになるよう無意味単語を読ませるということが行われてきた．韻律的な特徴は信号処理による波形処理を行うことを前提にしているからで，この結果，波形が本来持っていた自然性は損なわれ人工的な音質になっていた．CHATRはこの代わりに同一話者が読み上げた文音声DBを用いる．文音声DBから音韻だけでなく韻律のバリエーションも持った豊富な音声単位を選択し，できる限り信号処理を行わないで波形を接続していく．音声単位選択基準として韻律情報も利用することで，従来の音声合成システムより多くの音声単位を必要とはするものの，信号処理の負荷を減らすことができ，計算量削減と自然性保持の両方を実現できるようになっている．文音声DBの量については，開発者の経験によれば，日本語の場合，20分間分の音素バランス文音声があればよいとのことである．

CHATRの処理概要

図-3に示す通り，CHATRの音声合成処理部は韻律予測の機能が含まれているだけで，波形に関する情報はすべて外部情報として扱う．処理は以下の3つのプロセスに分割して行われる．

(1) 特徴抽出過程：音声DBの分析
(2) 最適重み決定過程：最適重み係数の学習
(3) 単位選択過程：音声単位の選択

以下に各処理の概要を示す．

(1) 特徴抽出過程

厳密には，音素記号系列の生成，音素のアライメント，特徴抽出を含み，音声DB中のすべての音素について，それらの性質を与える特徴ベクトルを与える過程であり，新しい音声DBを作成するときに1度行う必要がある．音素記号系列の生成とは，音声DB中で読み上げられる文の読みを求める，いわゆる書き起こしのことであるが，その表記は先に述べた音韻の区別ができる音素記号を用いる．音素のアライメントは，この音素記号を実際の音声波形に対応づける処理である．特徴抽出は，各音素の特徴ベクトルを生成する処理である．特徴ベクトルには音素表記や音響的特徴を示す音素ラベル，音声DB中の位置に加え，F0，音素時間長，パワーなどの韻律的特徴が含まれる．

(2) 最適重み決定過程

音声DBの特徴ベクトルと音声DBの原波形を用いて，目的の音声を合成する場合に最も適切な音声単位を選び出すための，各特徴の最適重み係数を重みベクトルとして求める過程である．目的音声の音響的，韻律的環境に最適な音素を音声DBから選択するためには，どの特徴が音素的，韻律的環境の違いによりドミナントとなるかを決める必要がある．

☆4 http://www.itl.atr.co.jp/chatr/

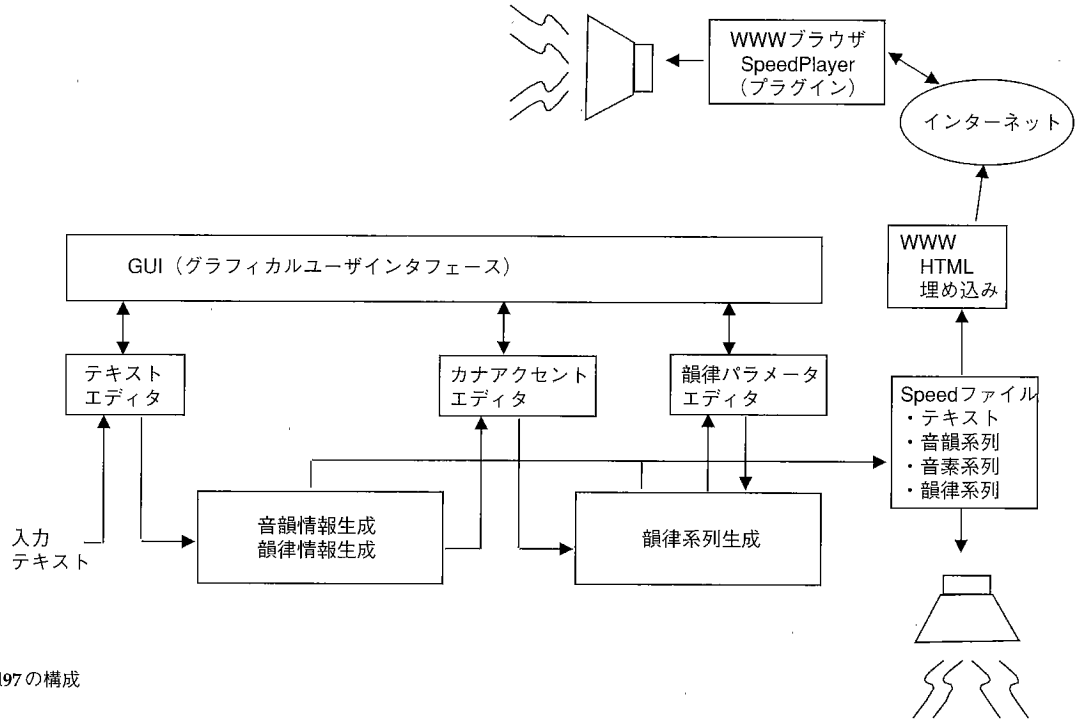


図-5 Speed97の構成

これは音素の性質によって重要である特徴の種類が変化するためで，たとえばF0は有声音の選択にはきわめて有効であるが無声音の選択には影響がない．また摩擦音の音響的特徴は前後の音素の種類によって影響が変わるなどといわれている．調音位置や調音様式などの音素的特徴とTri-phoneのF0，音素時間長，パワーなどの韻律的特徴など音声DBの特徴ベクトルを構成するパラメータがすべて利用される．各音素ごとに，最適な候補を選ぶ際にどの特徴がどれだけ重要かを決定するための重み係数の学習には，音声DB中のすべての音素サンプルに対して，それぞれの音素サンプルに着目し，他のすべての音素サンプルとの音響的距離を求め，上位N個の類似音素サンプルを選び出し，特徴ベクトルの差分（target sub cost）を求める．この差分から線形回帰分析を行い，当該音素に対して線形重み係数を求める．

(3) 単位選択過程

従来の音声合成システムでは目的の発話に対して音素系列を決定し，さらに韻律制御のためのF0と音素時間長の目標値が計算された．CHATRでは最適な音声サンプルを選択するために韻律が計算されるだけで，直接韻律を制御することは行わない．単位選択過程の処理では目的発話の音素系列と，それぞれの音素ごとに求めた各特徴に対する重みベクトルおよび音声DB中の全サンプルを表す特徴ベクトルが入力され，音声DB中での音素サンプルの位置を表すインデックスが出力される．このインデックスには，音声波形を接続するためのそれぞれの音声単位の開始位置と長さが示される．

最適な音声単位の集合は，**図-4**に示す通り，目的発話との差を表すtarget costと，隣接音声単位間での不連続を表すconcatenation costの和を最小化することで求める．経路探索にはViterbiアルゴリズムが用いられている．target costを最小にすることで，各特徴が目的音声に近く，concatenation costを最小にすることで音素単位間の不連続性が少ない音声DB中からの音声単位の組合せを選び出すことができ，これら音声単位の音声DB中での位置を示すことで，任意の発話内容の音声合成が可能になる．

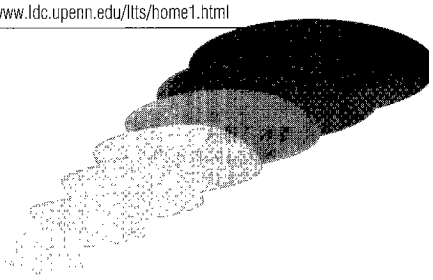
CHATRの効果

CHATRでは，話者の個性や発話様式の特徴を失うことなく任意の音声を合成できる．これは音声合成エンジンとはまったく独立して外部情報化した音声DB中の音素をそのまま利用することによる．この結果単にCD-ROMなどに記憶した音声データを取り替えることで，任意の言語の任意の話者で合成音声を発声させることができるようになる．現在すでに日，英，独，韓，中5カ国の男女，子供を含めて50種類以上のバリエーションの音声が用意されており，該当ホームページをアクセスすることで視聴が可能である☆4．

音声メッセージデザインツールSpeed97[8)]

Speed97は，合成音声による豊かな表現力を持った音声メッセージを作成するためのデザインツールを意図して開発された．マルチメディア

☆5 この値は，電話音声の約1/80，携帯電話音声の約1/3程度に相当する．
☆6 http://www.hil-unet.ocn.ne.jp　このホームページでSpeed97で作成したデモ音声も視聴可能である．
☆7 http://www.itl.atr.co.jp/cocosda/synthesis/3rd_ws.html
☆8 COCOSDAでは，世界中の合成音声を視聴できるホームページを用意している．アドレスは次の通りである．http://www.ldc.upenn.edu/ltts/home1.html



サービスの実現において，コンテンツ作成が重要な位置を占めるようになってきており，DTM（Desk Top Music）による音楽制作や，CG（Computer Graphics）による映像制作と同様，音声のためのコンテンツ作成ツールがあってもよいのではないかという発想から開発されたものである．

Speed97の実現機能

図-5はSpeed97のシステム構成である．TTSシステムで最初に入力されるテキスト，テキスト処理から音声合成処理に渡されるインタフェース情報，ならびに音声合成処理部分で韻律的特徴を示すパラメータの3つの情報をGUI環境で自由に変更できるエディタが用意されている．ユーザはテキストを入力して自動合成された音声を元に，韻律的特徴を変更しては，その変更に基づく合成音声を視聴するというステップを繰り返すことで所望の音声を制作できる．韻律的特徴の設定は2段階で行われる．まずはスタイル設定と呼ぶ段階で話者の区別（男／女），速度，音量，平均的ピッチの高さなどの標準値を事前に決めておく．次に実際にこのスタイルに基づいて作成した合成音声を，GUI環境で波形を表示しながら，各音素ごとに基本周波数，パワー，継続時間長を自由に変更できるようになっている．すでに音声ガイダンスや情報アナウンスのための音声メッセージばかりでなく，方言や感情を込めたせりふなどもこのツールで試作されている．

Speed97の効果

Speed97で作成した音声コンテンツはSpeedファイルと呼ばれる形式でファイルに格納される．この形式では音声の情報量を約800bit/sに圧縮しており☆5，同時に開発されたWWWブラウザのプラグインソフトであるSpeedPlayerによって再生できる仕組みを用意している．Speedファイルを HTML文章中に埋め込んでホームページのデザインに利用することで，インターネット上での音声も利用した情報発信が可能となる．またSpeedファイル上では，音声信号とテキストとの対応がつけられているため，文単位，フレーズ単位で蓄積，管理することで，文字情報を用いて必要な音声の検索を自由に行うことができる．この結果，駅のアナウンスやナビゲーションなどのメッセージ作成のための再利用や，画像などの他メディアとの同期も容易に実現が可能である．なお，Speed97，SpeedPlayer共に，試行版をインターネット経由で入手可能である☆6．

今後の課題

音声対話処理なども視野に入れ，研究レベルのものも含めて広く合成音声技術をサーベイしたい方々には当会誌で特集が組まれており[9]，大変参考になる．

表現力豊かな合成音声を生成できるようになったことで，従来校閲支援など限られた領域でのみ使用されていた音声合成技術も一般ユーザに広く利用されていく段階にきている．最近当学会のSLP研，NL研合同で「ここまでできるぞ音声／言語処理技術」[10]という研究会が開催されたことや，ICSLP'98のサテライトアクティビティとして開催されるCOCOSDA 3rd International Workshop on Speech Synthesis☆7においては，世界中から音声合成技術を持ち寄って，相互に比較評価するという特別セッションが計画されているなど，音声合成技術に関する研究成果の積極的なアピールの現れと見てとれ，このような企画からも，合成音声技術者の自信を伺うことができる☆8．

一方，合成音声技術が広く利用されることで，広範なユーザ層を対象としてホームページや電子メールの内容を読み上げるなどのサービスがこれから本格化してくることが予想される．このような場合，そもそも読んで，あるいは見て分かりやすく書かれている文章を，聞いて分かりやすくする必要がある．このような問題に答えるため文書書き換え技術の研究が進められており[11]，合成音声システムの広い普及において重要な技術となると考える．

参考文献
1) 杉藤美代子編: 講座日本語と日本語教育, 第2巻日本語の音声・音韻(上),明治書院(1989).
2) 小池恒彦, 筧　一彦, 古井貞熙, 北脇信彦, 東倉洋一: 音声情報工学, NTTアドバンステクノロジ(株)(1990).
3) Fuchi, T. and Takagi, S.: Japanese Morphological Analyzer using Word Co-occurrence, Proc. COLING-ACL'98, pp. 409-413 (1998).
4) 北脇信彦: テキスト音声合成とその多様化技術 , NTTR&D, Vol.45, No.10, pp.47-54 (1996).
5) Charpentier, F. and Moulines, E.: Pitch Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones, Proc. Eurospeech'89 (1989).
6) Fujisaki, H. and Hirose, K.: Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese, J. Acourst. Soc. Japan, Vol.5, pp.233-242 (1984).
7) ニック・キャンベル,アラン・ブラック: CHATR: 自然音声波形接続型任意音声合成システム, 電子情報通信学会技術研究報告, SP96-7, pp.45-53 (1996).
8) 阿部匡伸, 水野秀之, 中鳶信弥: 様々な音声表現を実現できる音声作成ツールSpeed97, 音声言語情報処理, 17-12, pp.67-72 (1997).
9) 特集「音声処理技術とその応用」, 情報処理, Vol.38, No.11, pp.970-1018 (Nov. 1997).
10) SLP/NL合同セッション「ここまでできるぞ音声／言語処理技術」: 情報処理学会研究報告, 98-NL-125, pp.1-16 (1998).
11) Matsuoka, K., Takeishi, E. and Asano, H.: Natural Language Processing in a Japanese Text-to Speech System for Written-style Texts, Proc. 3rdIVTTA, pp.33-36 (1996).

(平成10年11月9日受付)