Vol.88 No.06 510-511

Professional Report

高品位知的音声合成技術の開発

Development of High Quality and Intelligent Speech Synthesis Technology

北原 義典 Yoshinori Kitahara

ヒューマンインタフェースの重要性が叫ばれる昨今，任意の単語や文章を読み上げる「音声規則合成方式」のニーズが高い。今回，入力されてくる漢字かな混じりテキストを，正確に読み分け，明瞭(りょう)性・自然性ともに高く肉声にきわめて近い合成音声に変換する「高品位知的音声合成技術」を開発した。この技術は，係る単語の意味属性をチェックすることで，複数の読み方のある単語であっても前後の文脈から読みを正確に判定，調音コスト関数を導入することにより，接続性のよい音声単位(素片)を効率よく選択，かつ滑らかに接続し，肉声感が非常に高く滑らかな音声を生成，さらに，音声コーパスから学習した韻律モデルにより，自然な抑揚やリズムを表出できるという技術的特徴を持つ。

この技術を用いれば，アナウンスや音声メッセージを，いつでも簡単に低コストで作ることが可能になる。

ここでは，方式の詳細および評価結果について述べる。

1981年日立製作所入社
中央研究所
知能システム研究部 所属
現在，音声合成・認識技術の
研究開発に従事
電子情報通信学会会員
日本音響学会会員
ヒューマンインタフェース学会会員
日本笑い学会会員
工学博士

1 はじめに

人間の発話を機械的に模擬するという試みは1800年頃からすでに行われていた。代表的なものは「フォン・ケンペレンの合成器」で，これは，ふいごのようなもので呼気に相当する風を送って弁を振動させ，革管で作った声道で共鳴させることにより，音声生成系を物理的に実現したものであった。その後，半導体やコンピュータの登場により，1950年以降，音声合成技術は著しい進展を遂げることになる。

現在，駅のホームやデパートのエレベータなどでよく耳にする合成音声は，ほとんどがあらかじめ録音した音声をPCM（Pulse Code Modulation）データあるいはパラメータとして保持しておき，再生する「録音再生方式」である。しかし，ヒューマンインタフェースの重要性が叫ばれる昨今では，任意の単語や文章を読み上げる「規則合成方式」のニーズが高い。現在，音声合成という言葉は，この「規則合成方式」を指す場合も多く，特に，任意文章を合成音声に変換する技術は，テキスト音声変換（TTS：Text-to-Speech）と呼ばれる。このテキスト音声変換技術は，テキスト情報のみから任意の文音声を生成できるため，その適用分野はきわめて広く，各社が競って開発を進めている。

テキスト音声変換技術では，1990年代半ばまでは，いかにはっきり音韻が聞こえるかという，いわゆる「了解性の向上」に研究の力点が置かれてきた。その結果，従来の合成単位をパラメータで持つ方式から，波形で持つ方式が主流となった。さらに，基本周波数を制御するために1周期(ピッチ)波形をずらして加え合わせるピッチ周期波形重畳方式を各社とも採用するようになり，ある水準の了解性は確保されるようになった。

1990年代後半から，各社の研究の焦点は「自然性の向上」に移ってきている。われわれのグループは，音声の抑揚・リズムなどの「韻律」が自然性に最も大きな影響を与えることをつかみ，この韻律の制御に関する幾つかの試みを経て，今回，入力されてくる漢字かな混じりテキストを正確に読み分け，明瞭性，自然性ともに高く，肉声にきわめて近い合成音声に変換する「高品位知的音声合成技術」を開発した。以下，方式の詳細および評価結果について述べる。

2　音声合成技術の課題

音声規則合成方式の構成を**図1**に示す。まず，入力された漢字かな混じりテキストを言語解析処理部によって形態素に分割し，読みとアクセントを決定する。読みは辞書および読み決定規則を用いて決められる。一方，アクセントは辞書およびアクセント結合規則により文節レベルで決定される。この段階で，漢字かな混じりテキストは発音記号列となる。続く韻律計算処理部では，抑揚および音韻継続長といった，いわゆる「韻律」が計算される。抑揚は，韻律モデルにより，近似基本周波数パターンを計算することによって付与される。音韻の時間継続長は，音韻の種類によって異なるが，同じ音韻でも文頭や文末などの置かれる位置によっても異なることから，音韻継続長については，音韻の種類や位置によって規定した音韻継続長モデルを用いて決定する。素片接続処理部においては，発音記号列の読みやアクセントに従って合成の単位である素片を選択して接続する。少ないデータ量で任意の文章を表現できるように,素片として，「a」，「ka」，「sa」などのV（母音）およびCV（子音－母音）を用いる。

最後に，韻律計算処理部で計算した基本周波数パターンに従って，合成音声の基本周波数を制御し，D/A（Digital to Analog）変換後出力する。

テキスト音声変換における技術課題は次の４点に集約される。

(1) テキストの読み分け性能向上：漢字かな混じりテキストに現れる同表記異読語や同表記異アクセント語をいかに精度よく読み分けられるか。

(2) 高い了解性の実現：各音韻がいかにはっきりと明瞭に聞き取れるか。

(3) 高い肉声感の実現：いかにひずみが少なく，人間の声に近い音声を実現できるか。

(4) 高い自然性の実現：いかに自然な抑揚やリズムを付与できるか。

以下，これらの課題について，これまでに取り組み，開発した手法について述べる。

3　テキスト読み分け性能向上を目指して

漢字かな混じりテキストの読み分けでは，特に同表記異読語,同表記異アクセント語の読み分けが問題となる[1)]。同表記異読語とは，「通った（かよった）」，「通った（とおった）」のように，表記が同じで読みが異なる語であ

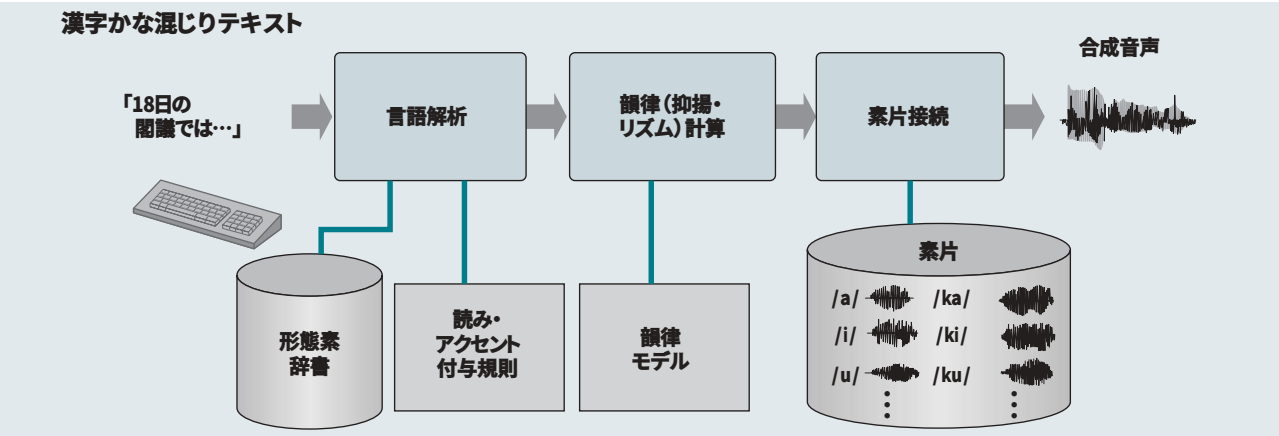

図1　音声規則合成方式のブロック図

入力された漢字かな混じりテキストは，言語解析処理部によって読みとアクセントが付与され，韻律計算処理部にて韻律モデルを用いて抑揚および音韻継続長の計算を行い，発音記号列に変換される。素片接続処理部では，素片と呼ばれる音声の単位片を発音記号列に従い選択，接続する。最後に，韻律計算処理部で計算した基本周波数パターンに対応するように合成音声の基本周波数を制御し，合成音声として出力する。

り，日本語にはこのような同表記異読語が多く存在する。その読み分け方法は，幾つかのレベルに分けられる。例えば「出」は，「出る」の場合「で」，「出す」の場合「だ」と読む。これらは，送りがなによって判断されるため形態素解析レベルで読み分けが可能である。「行った」は，係る先行文節に「に，へ」のような方向を表す格助詞が含まれていれば「い（った）」と読み，同じく対象を表す格助詞「を」が含まれていれば「おこな（った）」と読む。よって，従来は，「行った」に係る格助詞すなわち構文解析レベルで読み分けが可能とされていた。しかし，この方法では「この道を行った」という入力に対しては読み誤りを生ずる。そこで，読み分けすべき語に係る単語の意味属性のチェックまで行い，読み分けを可能にした（図2参照）。

この方法によれば，「最中（もなか）を食べている最中（さいちゅう）」なども読み分けることができる。ただし，「国分寺の方（ほう）」と「国分寺の方（かた）」のような読み分けは文脈レベルで決定されるものであり，この方式ではまだ対応できていない。

他方，「私は悩み（なや ’ み）」と「私の悩み（なやみ）」のように，同表記異アクセント語，すなわち，同じ表記でアクセントの異なる語が存在する。同表記異アクセント語では，同表記異読語の読み分けと同じように，幾つかのレベルの読み分け方法がある。「悩み（なや ’ み）」と「悩み（なやみ）」などは，両者の品詞の違いで区別できるので，形態素解析レベルで読み分けが可能である。しかし，例えば「下記の通り（と ’ おり）」と「石畳の通り（とおり）」では両者ともに同じ名詞であり，読み分けのためには係る単語の意味属性のチェックが必要となる。アクセントの誤りは，読みと同様，合成音聴取時には特に致命傷となるので，このシステムでは，同表記異アクセント語についても意味属性のチェックまで行うことにより，これらの読み分けを可能にし，知的音声合成を実現した。

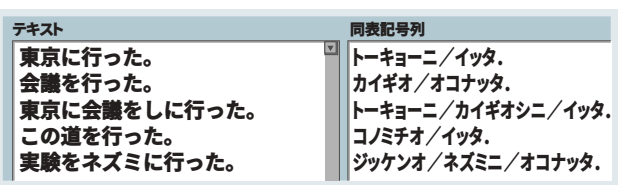

図2　読み分けの例
同表記異読語，同表記異アクセント語について，意味属性のチェックを行うことにより，高い精度で読み分けることが可能になった。

このように，読み分け性能を向上させることは，ニュースや株価，交通情報など，リアルタイムで配信されてくる情報を合成音声で読み上げるアプリケーションにおいては，きわめて重要なことである。

読み分け性能を評価するために，テキスト解析精度を測定した。テキスト解析精度の測定は，JEITA（Japan Electronics and Information Technology Industries Association：電子情報技術産業協会）の制定した「音声合成システム性能評価方法」に準ずる[2), 3)]。ニュース文章273文4,650文節について測定した結果，文節正読率は平均99.8 %となり，98 %台にとどまっている他社に比べて圧倒的に優位であることを確認した。

4　了解性向上を目指して

合成音声は，情報の伝達を主目的にしており，「了解性」の高さは品質のよしあしに影響を与える重要な要因である。了解性の高い合成音声を生成するためには，分節的（Segmental）な特徴がきちんと保存され再現されることが重要である。すなわち，各音韻がその明瞭な音韻性を有することと，これら音韻がスムーズに接続される必要がある。

音声の規則合成方式は，素片をパラメータで保持しておくパラメータ編集方式と，素片を波形のままで持つ波形編集方式とに大きく分けられる。従来は，情報圧縮の観点から，前者のパラメータ編集方式が主として採用されていた。同方式では，音源情報として基本周波数や振幅などのパラメータと，声道情報としてのスペクトル包絡パラメータに分けて保持するため，こ

れらを独立に制御できるという利点がある。また，素片間の接続も比較的容易である。反面，モデル化に伴う音韻のひずみが生ずるため，高い了解性を確保することは困難であった。例えば，線形予測分析系のパラメータでは，極のみの形状のモデル化であるため音韻によっては実際のスペクトルと隔たりがある。

一方の波形編集方式は，素片を波形として保持するため，音韻のひずみが生じにくいという利点はあるものの，データ量が大きいという欠点がある。しかし，昨今のメモリの大容量化および低価格化によって，音質の高さを重要視して同方式が普及してきた。

以上のような理由から，1995年以降，日立製作所中央研究所では波形編集方式を採用しており，その中でも，特に，基本周波数を制御しやすい時間領域ピッチ同期波形重畳法（TD-PSOLA：Time Domain Pitch Synchronous Over-Lapping Addition）を用いている[4)]。同方式は，両端が小さくなるような窓かけをして切り出した2ピッチ分の音声波形を，ずらし幅を変えて加え合わせていき，周波数の制御を行うものである（図3参照）。そのために，新たに，元話者によって動的に探索範囲を限定する高精度ピッチマーキング法を開発し，精度よく素片を切り出すことにより，合成音声の了解性は大きく向上した。

5　肉声感向上を目指して

さらに，肉声感を向上させるために，前後との接続性が高い素片を大量の音声コーパスから選択する方式を開発した（図4参照）。従来は，素片をおのおの1種類ずつ持っておき，これらを組み合わせて音声を合成していた。例えば「閣議で」という言葉を作る場合は，「ka」「ku」，「gi」，「de」という素片を使用するが，「か ’ くぎで」とアクセントは「か」にあるため，素片「ka」の基本周波数を高くし，「ku」，「gi」，「de」は低くすることになる。しかし，この方法では，一つの素片の基本周波数を無理に上下させるため，音がひずみ，肉声感が低下するという問題点があった。そこで，今回，音声コーパスから，人間の調音特性に基づき，接続性が最適な素片を動的に選択する方式を開発した。すなわち，音節S_nに対し，ターゲットコストと接続コストの和として定義した，前後の音節S_{n-1}，S_{n+1}との接続性に関するコスト関数C（S_n｜S_{n-1}，S_{n+1}）を導入し，このコスト関数の和を最小にするような音節S_nを素片として選択する。

例えば，「閣議で」の例では，「かくぎ」の「か」として次に「k」がくる基本周波数の高い「ka」を持ってくる，「ku」は /a/ と /g/ に挟まれた基本周波数の低い「ku」

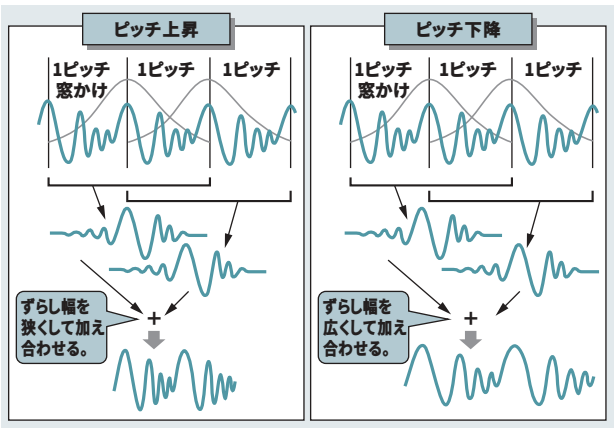

図3　時間領域ピッチ同期波形重畳法によるピッチ制御
窓かけをして切り出した２ピッチ分の音声波形をずらして加え合わせていき，周波数の制御を行う。

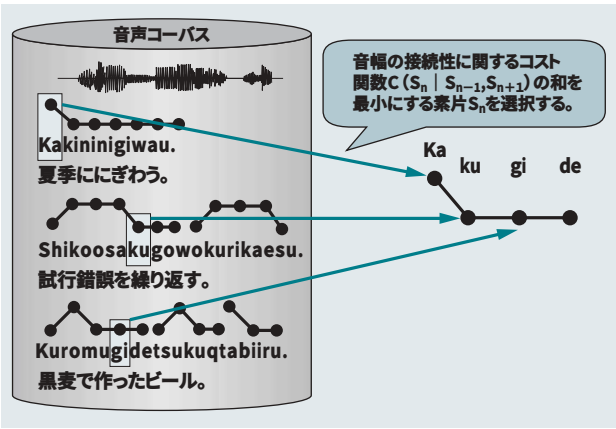

図4　調音コスト最小化素片選択方式
音声コーパスから，人間の調音特性に基づき，接続性が最適な素片を動的に選択する。

を持ってくる，というように，前後の音韻と基本周波数の高さを考慮して，素片を選択する[5), 6)]。このことにより，ピッチ制御によるひずみが減少し，肉声感が格段に向上した。

6　自然性向上を目指して

1章でも述べたように，1990年代後半，各社の研究の焦点は「自然性の向上」に移ってきている。われわれのグループは，音声の超文節的（Supra-segmental）な特徴である「韻律」が自然性に最も大きな影響を与えることを幾つかの実験でつかんできた[7), 8)]。音声の「韻律」とは，基本周波数構造（音の高低），音声の振幅構造（音の強弱），時間構造（リズム）といった音源特性に基づく音声ならではの情報を指す。

今回は，合成音声の時間構造と基本周波数構造に関して，これらを計算する新しいモデルを開発し，自然性を向上させた。

まず，時間構造，すなわち各音韻の長さを決定する音韻継続長モデルの改良である。例えば，「ひたち」という音声を合成する場合には，「ひ」，「た」，「ち」のおのおのの音節の長さを計算し，「hi」，「ta」，「chi」の素片をこれら計算した値に合致するように変形し接続する。音節の長さを計算するためには，音節を構成する「a」，「i」，「u」などの母音，「k」，「s」，「t」などの子音，「N（はつ音）」「Q（促音）」などの特殊音素等，各音素の長さを音韻継続長モデルを用いて求める。従来，この音韻継続長モデルとして，音素別に，文節内における位置を考慮して定めた音韻継続長をテーブルとしたものを用いていた。しかし，このように固定的な音韻継続長では，自然なリズムは得られなかった。そこで，今回，音韻環境を考慮した音韻継続長モデルを肉声データから学習する方式を開発した[9)]（図5参照）。これは，ある音素について，その継続時間長は前後複数個の音韻の種類に依存するとの仮定をおいて，継続時間長を

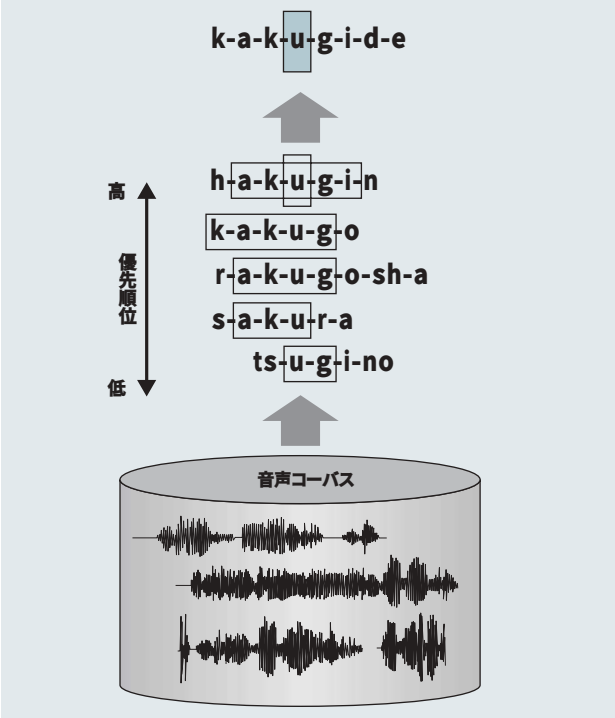

図5　音声コーパスからの学習による音韻継続長モデル
継続時間長を求めるべき音素を前後の音韻環境パターンとともに大量の音声コーパスから抽出し，継続長を決定する。

求めるべき音素を前後の音韻環境パターンとともに大量の音声コーパスから探索し，継続長を決定するものである。すなわち，合成したいフレーズ中における i 番目の音素p_iの音韻継続長をDur(p_i)としたとき，ある整数の定数m，n（m≧1，n≧1）に対して，

$[p_{i-m} \cdots p_{i-2}\ p_{i-1}\ p_i - p_{i+1}\ p_{i+2} \cdots p_{i+n}]$

の音素列パターンを音声コーパス中から見つけ出し，これらについてある基準で優先順位を付け，最も優先順位の高い音素列パターン中の音素piすべての平均継続時間を全体の話速で正規化した値を最終的な音韻継続長Dur(p_i)とする。

一方，ピッチ（抑揚）モデルについては，やはり大量の音声コーパスから学習する方式を開発した。これは，文節内で折れ線近似した基本周波数包絡線を，長さの異なる文節ごとに，上述と同様に前後の音韻環境を考慮し音声コーパスから学習したもので，このピッ

チモデルを基に，各文節の抑揚パターンを計算する。

このように，前後の音韻環境を考慮し音声コーパスから学習した音韻継続長モデルおよびピッチモデルを用いることにより，自然性が大きく向上した。

ここまで述べた了解性，肉声感，自然性を含む合成音声の総合品質を評価するために，JEITAの制定した「音声合成システム性能評価方法」に準じた総合性能評価実験を行った。ここでは，組み込み用途向けにデータサイズを削減し，数十Mバイトにまでコンパクト化した素片を用いて作成した合成音声，および，他社2社の合成音声を用いて，22人の被験者に対し，5段階で評定してもらった。音声の提示はランダムで，スピーカ受聴とした。その結果，他社の総合性能の5段階評定平均が2.3から3.1であったのに対し，日立の合成音声は同3.9となり，優位性が確認された。また，被験者の内省報告では，日立の合成音声が最も人間の声に近いというコメントが多く見られた。このようにして，自然で肉声感の高い高品位知的音声合成を実現した。

7　おわりに

ここでは，高品位知的音声合成技術の開発について述べた。

ここ数年，了解性の向上とともに，音声合成技術は実用に供するようになり，カーナビゲーションや各種サービスをはじめとする市場ニーズが急増している。これらの技術は，株式会社日立ケーイーシステムズや株式会社日立超LSIシステムズから，音声合成ソフトウェア開発キット（SDK）や組み込み向けミドルウェアとして製品化され，携帯電話のメール読み上げサービスシステムや，各種アナウンスシステム，ウェブ閲覧支援ソフトウェア，発話訓練ソフトウェアと，立て続けに実用化された。また，日立のパソコン「Priusシリーズ」にも「読みワザ」として標準搭載された。

特に，肉声感の高いことが，多くの顧客に支持され，引き合いが多い。今後も，交通・都市開発・金融・オフィス分野におけるアナウンス市場，通信分野における音声自動応答市場，ロボット・家電分野における音声インタフェースはもちろんのこと，教育・電子書籍・放送分野のコンテンツ市場など，幅広い展開が期待されている。

残る課題を整理すると，（1）さらなる発話のリアル性の向上，（2）文脈まで考慮した読み分け性能向上，（3）表情を持った多様な音声の生成方式の確立，（4）短時間で多種類の音声を生成する方式の確立，となる。特に，（4）については，例えば，簡易な登録だけで，ユーザーが自分の声や好きな有名人の声でさまざまな文音声を作り出したいというニーズが高いが（しかし，犯罪への転用が懸念材料としてある。），まだ，どこの研究機関でも確立していない技術である。

われわれは，常に一歩先んじた技術開発を推進していきたいと考えている。

参考文献など

1）**宮崎，外：日本文音声出力のための言語処理，情報処理学会・自然言語処理技術シンポジウム**（1983.6）

2）**電子情報技術産業協会規格：音声合成システム性能評価方法，**JEITA IT-4001（2003.2）

3）S.Itahashi，et al.：Standard for Japanese Speech Synthesizer Performance Evaluation，Proceedings of INTERNATIONAL SYMPOSIUM ON SPEECH TECHNOLOGY AND PROCESSING SYSTEMS and Oriental COCOSDA-2004，Vol-II（2004.11）

4）F.Charpentier，et al：Pitch-synchronous Waveform Processing Techniques for Text-T-Speech Synthesis Using Diphones，Eurospeech 89 vol.2（1989.9）

5）**額賀，外：組み込み機器向け素片選択型音声合成システムの検討，日本音響学会講演論文集，**2-P-4（2004.9）

6）N.Nukaga，et al：Unit Selection Using Pitch Synchronous Cross Correlation for Japanese Concatenative Speech Synthesis，5th ISCA Speech Synthesis Research Workshop（2004.2）

7）**北原，外：音声言語認知における韻律の役割，電子情報通信学会論文誌，**(D)，Vol.J70-D，No.11（1987.11）

8）**安藤，外：文の多義性解消におけるピッチとポーズの関係について，日本音響学会講演論文集，**2-P-1（1998.3）

9）**永松，外：可変長音素列を用いた継続時間長を用いた継続時間長モデルの検討，日本音響学会講演論文集，**1-6-3（2003.3）