高音質で聞きやすい音声合成システム ToSpeak$_{TM}$

ToSpeak$_{TM}$ High-Quality Text-to-Speech System

籠嶋　岳彦
■ KAGOSHIMA Takehiko

東芝は，話者の特徴をリアルに再現した高音質で自然な抑揚の音声を，任意の入力テキストから合成する音声合成システムToSpeak$_{TM}$を開発した。

技術のポイントは，大量の音声データ（音声コーパス）に基づく韻律制御規則の統計的な学習と，複数素片選択融合方式の音声信号生成部（合成器）である。韻律制御では，アクセント句を単位として，音声コーパスの基本周波数（声の高さ）の変化パターンから，誤差が最小となる代表パターンを抽出して利用する。合成器では，合成単位当たり複数の音声素片（合成単位ごとに切り分けられた音声波形）を音声コーパスから選択して融合することで，肉声感と安定感を両立した合成音声を実現している。

当社の音声合成システムは，カーナビゲーションの音声インタフェースをはじめとして，様々な応用分野で実用化されている。

Toshiba has developed ToSpeak$_{TM}$, a new text-to-speech (TTS) system that synthesizes speech in a high-quality, natural manner. ToSpeak$_{TM}$ can generate synthesized speech having the individuality of an original speaker in terms of prosody and voice quality from any input text. This TTS system features corpus-based approaches including (1) statistical training of prosody control rules, and (2) a plural unit selection and fusion method for the speech waveform generation module (synthesizer). In the prosody training, representative fundamental frequency vectors are extracted from the speech corpus so as to minimize errors of the resultant fundamental frequency contours. In the synthesizer, the proposed method achieves stable, humanlike speech quality. Our TTS systems are used in a variety of applications such as the speech interface of car navigation systems.

1　まえがき

テキスト音声合成は，任意の入力テキストを音声に変換する技術で，ヒューマンインタフェースの基盤技術として古くから研究開発が行われてきた。初期の段階では，各音韻のスペクトルの特徴を再現する音声波形を規則によって作り出すアプローチがとられたが，計算機の性能向上に伴って，収録した実際の音声波形を加工して利用する手法が一般的となった。このような技術の進展により，“聞き取れる”という意味で実用的なシステムは，比較的早い段階から実現されていた。

しかし，音質の問題や不自然な抑揚のため，一般に広く普及するには至らなかった。音質の問題は，音声波形の分解や，変形，接続などの音声合成処理により，声の人間らしさ（肉声感）が失われて機械的な声に変質することが原因である。また，不自然な抑揚は，アクセントやイントネーション，リズムなどの自然な変化パターンを十分にモデル化できていないことが原因となっていた。

これらの問題に対して，“コーパスベース音声合成”のアプローチが近年盛んに検討されている。これは，大量の音声データ（音声コーパス）を用い，統計的な手法に基づいて音声合成システムを構築するものである。東芝は，コーパスベース音声合成にいち早く取り組み，コンパクトなメモリサイズで安定した高品質な音声を合成できる音声合成システムを開発し，製品化してきた。この音声合成システムは，カーナビゲーションなど車載機器向けの組込みソフトウェア（ミドルウェア）や，パソコン（PC）のアプリケーションソフトウェアなど，様々な用途で利用されている。

更に，よりリアルな合成音声の実現を目指して，次世代音声合成システムToSpeak$_{TM}$を開発した。特定の人物の声を合成音声で再現するためには，単に人間らしいということにとどまらず，声質の特徴と抑揚（話し方）の特徴を同時に再現する必要がある。ToSpeak$_{TM}$は，抑揚の特徴を学習する韻律制御モデル[1]と，複数素片選択融合方式の合成器[2]により，特定話者のリアルな合成音声を実現している。ここでは，ToSpeak$_{TM}$の原理と特長，及び応用について述べる。

2　テキスト音声合成システムToSpeak$_{TM}$

テキスト音声合成システムは，**図1**に示すように，テキスト解析，韻律生成，及び音声信号生成の三つの処理から構成される。テキスト解析部では，入力されたテキスト（漢字かな交じり文）を言語辞書を参照して解析し，漢字の読みやアクセントの位置，文節（アクセント句）の区切りなど言語情報を出力する。韻律生成部では，言語情報に基づいて，声の高さ（基本周波数）の時間変化パターンと各音韻の長さなどの音韻・韻律情報を出力する。音声信号生成部（合成器）では，音韻の

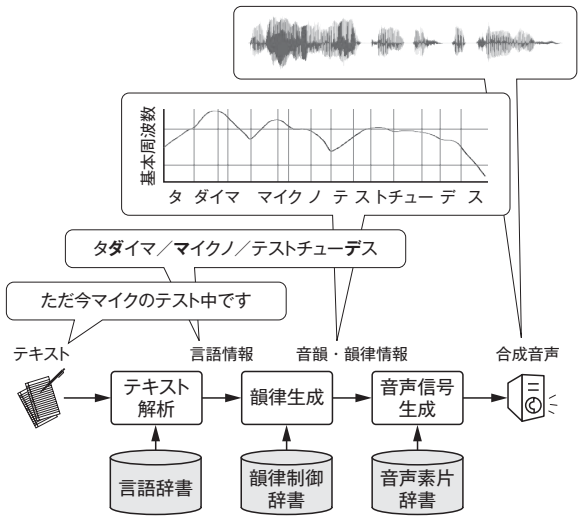

図1. テキスト音声合成 — 任意の入力テキストから，テキスト解析，韻律生成，及び音声信号生成の3段階の処理を経て，合成音声が生成される。

TTS system

系列に従って音声素片を選択し，韻律情報に従って変形して接続することで，合成音声を生成する。

ToSpeak$_{TM}$の特長である韻律生成部と音声信号生成部について，以下に述べる。

2.1 韻律生成

韻律生成処理の中で，合成音声の自然性や個人性にもっとも影響するのが，基本周波数制御である。基本周波数制御は，**図2**に示すように，読みや品詞，アクセント型など質的（定性的）な言語情報から，声の高さという物理量の変化パターンを生成する。そのため，なんらかの基本周波数生成モデルを用い，言語情報から規則によってモデルパラメータを決定することで，基本周波数パターンを生成する。

従来は，基本周波数パターンを関数で近似するモデルなどを用いて，モデルパラメータ生成規則を技術者がチューニングすることで，自然性の向上が図られてきた。しかし，このような方法では十分な自然性を得るのが難しく，また，自然性が技術者のスキルに依存するという問題があった。

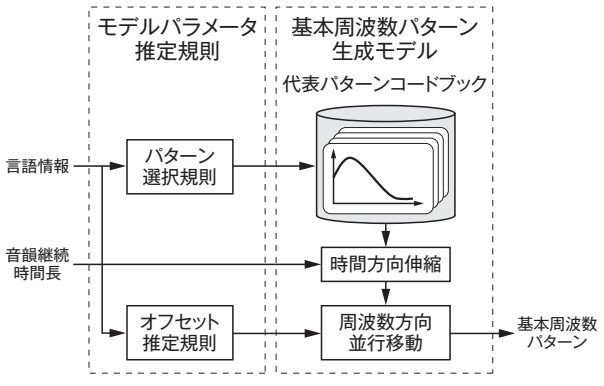

図2. 基本周波数生成モデル — アクセント句ごとに代表パターンコードブックからパターンが選択され，変形して接続することで1文の基本周波数パターンが生成される。

Fundamental frequency generation model

当社は，音声コーパスから抽出された自然音声の基本周波数パターンを教師データとして制御規則を自動的に学習し，話者の特徴をとらえた基本周波数パターンを生成する手法を開発した。自動学習と親和性の良いモデルとして，アクセント句単位の典型的な基本周波数パターンを表す代表パターンに基づくノンパラメトリックなモデルを導入した[1]。このモデルでは，アクセント句ごとに生成された基本周波数を接続して文の基本周波数を生成する。アクセント句ごとのパターンの生成には，図2に示すように，代表パターンの集合である代表パターンコードブックを用いる。コードブックから選択した代表パターンを，音韻継続時間長に従って時間方向で伸縮し，オフセット値に従って周波数方向で並行移動して，アクセント句のパターンを生成する。代表パターンの番号とオフセットの値は，それぞれ言語情報から最適なものを推定して生成する。

このモデルを用いた場合，規則の学習とは，代表パターンコードブック，パターン選択規則，及びオフセット推定規則を音声コーパスから抽出することに相当する。これら規則の学習は，モデルから出力される基本周波数と実音声の基本周波数の誤差を評価尺度として，誤差を最小化するように統一的に行われる。このようにして学習した代表パターンコードブックは，話者ごとに形状が異なっており，話者の特徴を反映した基本周波数の生成が可能である。この手法によって入力テキストから生成した基本周波数パターンの例を，実際に話者が発した音声の基本周波数パターンと比較して**図3**に示す。このように，元になった話者に近い自然な抑揚の音声を生成することができる。

2.2 音声信号生成

音声信号は，基本となる短い合成単位（音素や音節など）の音声データ（音声素片）を，入力された音韻系列に従って選択して接続することによって生成される。このとき，韻律生成部から入力されたとおりの基本周波数パターンの音声を合成するため，選択した音声素片の基本周波数を変形する処理を行う。

コーパスベース手法を導入した従来の音声合成器は，音声コーパスから合成する音声に適した音声素片を探索し，選択された素片を変形して接続するものである。音声素片の選択において，コーパス中の音声素片が合成する音声にどの程度適合しているかを直接に評価することはできないため，間接的な評価尺度が用いられている。すなわち，合成する音声の基本周波数や，継続長，前後の音韻などの属性が，音声素片のそれらとどの程度近いかを表す尺度と，選択された音声素片どうしの接続の滑らかさを表す尺度を用いて，最適な音声素片の系列を探索する。この手法では，適切な音声素片が選択

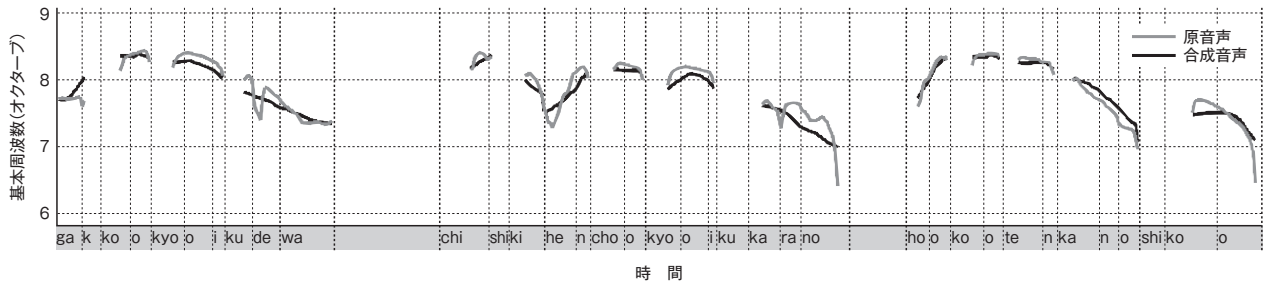

図3. 生成された基本周波数パターンの例 — 入力テキスト「学校教育では，知識偏重教育からの方向転換を志向」から生成した基本周波数パターンと，話者が発声したパターンを比較した。

Examples of fundamental frequency contours

された場合は高い音質が得られるが，音声コーパスに適切な素片がない場合や，あったとしても選択できない場合には，部分的に音質が劣化するという問題があった。特に，コーパスのサイズが小さい場合には，このような問題が顕著である。

これに対して当社は，音声合成器の新しい方式として“複数素片選択融合方式”（**図4**）を開発した[2]。従来法が合成単位当たり1個の音声素片を選択するのに対して，この手法は，音声コーパスから単位当たり複数個の音声素片を選択し（複数素片選択），それらを融合してその合成単位の素片を生成する（素片融合）。そして，融合された素片の韻律を変形して接続することで音声を生成する。

この手法の特長は，肉声感の高い音声を安定して生成できる点にある。高い肉声感を得るためには，声の高さの違いや前後の音韻の影響による音色のバリエーションを表現することが必要で，音声コーパス中の素片を適切に使い分けることによって実現している。また，適切な素片がコーパスに存在しなかった場合でも，目標近傍の複数の素片を融合することにより，目標に近い特徴を持った素片を作り出して音質劣化を抑え，均質で安定した音声を生成する。

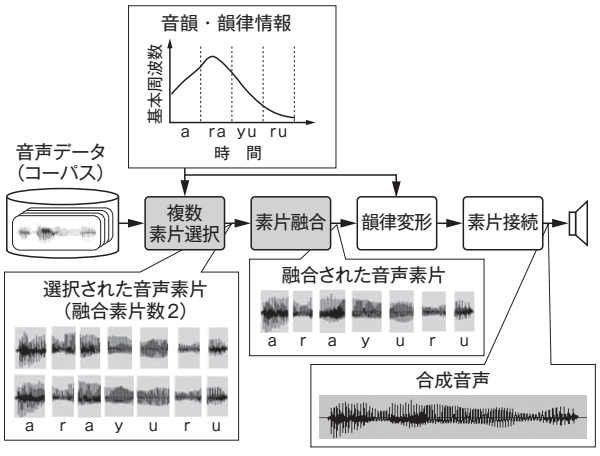

図4. 複数素片選択融合方式 — 合成単位当たり複数の音声素片をコーパスから選択し，それらを融合して接続することにより，肉声感と安定感を兼ね備えた合成音声を生成することができる。

Plural unit selection and fusion method

従来方式と提案する複数素片選択融合方式で合成した音声の音質を，主観評価実験により比較した結果を**図5**に示す。“従来法2”は，前述した従来手法で，合成単位当たり一つの音声素片を選択し接続する。また，“従来法1”は，合成単位ごとにその合成単位を代表するスペクトルを持つ音声素片をあらかじめ用意し，その合成単位には常にその代表素片を用いる手法で，バラつきの小さい安定した音声を生成できるという特長がある。

図5に示すように，実験に用いた女声3種類，男声1種類のすべての音声について，提案法は，従来法と比較してスコアが高く，音質が良いことがわかる。これは，肉声感重視で安定感に欠ける従来法1及び安定感重視で肉声感に欠ける従来法2と比較して，肉声感と安定感をバランス良く備えた複数素片選択融合方式の音質が高く評価されたものと考えられる。

この手法は，融合する素片の数（融合数）を大きくすると多くの素片が平均化されて，音色のバリエーションが減少し肉声感が低下していくが，逆に安定感は増していく傾向がある。そこで，コーパスサイズが小さく音質が不安定になりがちな場合は融合数を増やして安定感を高め，逆にコーパスサイズが大きい

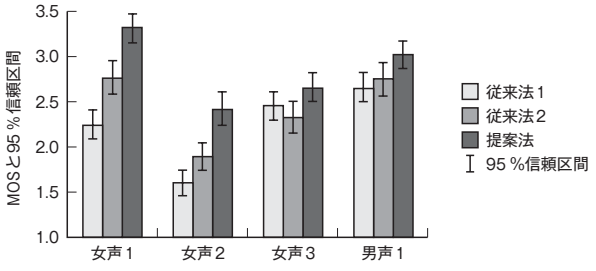

図5. 合成音声の音質評価 — 縦軸は，合成音声の音質を，1点（悪い）から5点（良い）までの5段階で主観評価した点数の平均値（MOS：Mean Opinion Score）を表している。

Subjective evaluation of synthesized speech

場合は融合数を減らして肉声感を高めるなど，安定感と肉声感のバランスを制御することができる。ToSpeak$_{TM}$は，このようにコーパスのサイズに応じて音質を最適化できることから，利用可能なメモリやストレージの容量が異なる様々なプラットフォームに適用できる，スケーラビリティの高い音声合成システムであると言える。

複数素片選択融合方式は，従来法と比較して，素片融合のため計算量が増加する。計算量削減のために，素片融合処理をあらかじめ行うことで，ランタイムの処理から素片融合を取り除くこともできる。この場合は，あらかじめ大量のテキストを合成し，選択された素片の組合せごとに出現頻度を記録して，高頻度の組合せだけ事前に融合素片を作成しておく。音声コーパスの代わりに融合素片を利用することで，大幅に計算量を削減してほぼ同等の音質を実現することができる。

2.3　多言語対応

前述したToSpeak$_{TM}$の韻律生成と音声信号生成は，言語に依存しない技術である。音声コーパスを言語ごとに開発し，基本周波数パターンの単位や制御に用いる言語情報を各言語に適したものにカスタマイズすれば，日本語と同様にほかの言語の音声合成システムを開発することができる。これまでに，アメリカ英語やイギリス英語，ドイツ語，フランス語，スペイン語，中国語（北京語）などで効果が確認されている。

3　テキスト音声合成システムの応用

音声は，人間にとってもっとも自然なコミュニケーションの手段である。ヒューマンインタフェースでも，録音しておいた音声の再生や，単語単位の録音音声をつなぎ合わせて再生するなどの手段で，従来から音声が利用されてきた。近年，これらの録音音声に加えて，テキスト音声合成による合成音声の利用が広まっている。テキスト音声合成をヒューマンインタフェースに用いると，以下に述べるような様々な利点があり，音声合成ならではの応用はもとより，録音音声に代わって合成音声が用いられることも増えている。

音声合成を用いることで，あらかじめ録音することができない変化する情報を音声で伝えることができる。例えば，PCのディスプレイ上の文字の読上げやWebページの読上げなどは，視覚障害者向けのインタフェースとして広く利用されている。また，近年の自動車向けテレマティクスサービスでは，電子メールや交通情報，ニュースなどがテキスト情報として車に送信され，音声合成により運転中でも安全に情報が受け取れるようになっている。そのほかにも，ビデオゲームで，プレーヤーが入力した名前などを合成し，音声で呼びかけるなどの応用もある。

あらかじめ録音可能な内容でも，音声合成を利用することで，データ容量の削減や音声収録コストの低減などの効果がある。例えば，カーナビゲーションが発声する交差点名称や施設名称などは，既定の内容ではあるが件数が非常に多く，頻繁に更新されることから音声合成が用いられる場合が多い。この理由の一つは，音声合成を利用することで，音声の収録に要する多大な費用と労力が削減されることにある。また，もう一つの理由として，その音声を合成するためのテキストのデータ量のほうが，音声のデータ量よりもサイズが小さいため，記録媒体に記憶させるデータのサイズを節約できることが挙げられる。これらと同様の理由で，電子辞書の見出し語や用例文の読上げにも，音声合成が利用されている。

当社の音声合成システムは，前述したようなPCのテキスト読上げソフトやカーナビゲーション，ビデオゲームソフト，電子辞書をはじめ様々な機器に搭載されて利用されており，音質及び自然性の更なる向上と多言語化に伴って，応用の拡大が期待されている。

今後は，合成音声の韻律や声質のバリエーションの向上が課題である。人間の音声は，話す内容（言語情報）だけでなく，その場の状況や話し手と聞き手の関係，話し手の意図・感情などによって話し方が変化し，ニュアンスの違いを伝えることができる。このような非言語的な情報も含めて合成音声で表現することができれば，よりユーザーフレンドリーなインタフェースが構築できると考えられる。

4　あとがき

より柔軟なユーザーインタフェースを低コストで実現するため，機器の音声出力へのテキスト音声合成の利用が広まっている。音声合成システムToSpeak$_{TM}$は，話し方や声質が，より人間の声に近づいた高品質な音声の合成が可能で，聞き取りやすく，自然な音声のインタフェースを提供することができる。また，ハードウェアリソースに対するスケーラビリティが高いことから，種々のプラットフォームの様々な応用が期待されている。今後は，合成音声の韻律や声質のバリエーションを広げるための研究開発を進めていく。

文　献

(1) 籠嶋岳彦，ほか．代表パターンコードブックを用いた基本周波数制御法．電子情報通信学会論文誌 D-II．**J85-D-II**，6，2002，p.976－986.

(2) Mizutani, T., et al. Concatenative Speech Synthesis Based on the Plural Unit Selection and Fusion Method. IEICE Trans. **E88-D**, 11, 2005, p.2565 － 2572.

籠嶋　岳彦　KAGOSHIMA Takehiko, D.Eng.
研究開発センター　マルチメディアラボラトリー主任研究員，工博。音声合成の研究に従事。
電子情報通信学会，日本音響学会会員。
Multimedia Lab.