次世代音声合成システムToSpeak$_{TM}$ V2を支える多様性向上技術

Enhanced Versatility Technologies Supporting ToSpeak$_{TM}$ V2 Next-Generation Text-to-Speech System

平林　剛　　水谷　伸晃　　籠嶋　岳彦
■ HIRABAYASHI Go　■ MIZUTANI Nobuaki　■ KAGOSHIMA Takehiko

公共施設の情報端末や，コールセンターの自動応答，しゃべる家電など，生活のなかで音声サービスを耳にする機会が増えてきている。音声合成技術は，テキストを入力するだけで任意の音声を生成できるため，これらの音声コンテンツの作成に要していた費用や時間を大幅に削減でき，また，この結果として音声サービスのいっそうの普及も期待できる。

東芝は，高い基本音質に加えて，様々なコンテンツに適したバリエーション豊かな声質や発話スタイルの合成音声の生成を実現するために，音声合成の多様化の研究を進めている。今回，様々な抑揚の特徴を精度よくモデル化できる新しい基本周波数制御手法と，任意の語句を強調できる韻律制御手法を開発した。これらの手法を導入して，幅広い用途に対応できる次世代の音声合成システムToSpeak$_{TM}$ V2を開発した。

Voice guidance is becoming increasingly prevalent in daily life, particularly in contexts such as information terminals in public facilities, automatic responses by call centers, and "talking" home appliances. As text-to-speech (TTS) technology enables users to generate arbitrary voices simply by inputting text, it can not only significantly reduce the cost and time required for creating voice contents, but also assist in the wide dissemination of voice services.

Toshiba has developed ToSpeak$_{TM}$ V2, a new TTS system that can synthesize various voices and speaking styles with high-quality, natural sound for new fields of application.

1　まえがき

近年，ユニバーサルデザインの考え方が普及してきたこともあり，これまでの文字情報に加え，様々な場面で音声によるメッセージやガイダンスが併せて用いられるようになってきた。このような音声サービスの広がりを背景に，音声コンテンツ作成に対する需要が高まってきている。

音声コンテンツを作成するには，プロのナレーターの声を収録して編集する方法が一般的であり，収録した音声の細かな調整や，コンテンツの追加，変更など多くの手間と費用が掛かる。この負担を軽減するために，音声合成技術が利用されるようになってきた。

音声合成技術は，任意の入力テキストを音声に変換する技術であり，文章を入力するだけで必要な音声コンテンツを容易に，しかも低コストで作成することができる。しかし，様々な用途や種類の音声コンテンツを作成するためには，合成音声の基本的な音質や自然性の高さに加えて，声質や話し方などに豊富なバリエーションが求められる。

東芝は，これまでに多言語音声合成システムToSpeak$_{TM}$を開発しており，カーナビゲーションなど組込みシステム用の音声合成ミドルウェアとして製品化し，その合成音声の基本品質は高い評価を得ている[1]。

今回，音声合成技術の用途拡大に向け，新しい音声合成システムToSpeak$_{TM}$ V2を開発した。ここでは，ToSpeak$_{TM}$ V2の概要と，多様な合成音声の生成を実現するために開発した次の手法について述べる。

(1)　決定木(注1)に基づく基本周波数(F_0)制御手法
(2)　強調音声を合成するための韻律制御手法

2　ToSpeak$_{TM}$ V2の概要

ToSpeak$_{TM}$ V2は，従来の書きことばを淡々と読み上げる“読上げ調”に加えて，話しことばに適した“親しげ調”，注意を促す“警告調”などの発話スタイル（しゃべり口調）にも対応しており，様々な状況に適した合成音声を生成できる。このような発話スタイルや話者特徴の多様性を実現するためには，声の高さの変化を表すF_0パターンの制御手法が重要であり，様々な抑揚の特徴を精度よくモデル化できることが不可欠である。そのためにToSpeak$_{TM}$ V2では，代表パターンに基づくF_0制御手法に，決定木に基づいたコンテキスト クラスタリングを導入することで，F_0制御の基本性能を向上させた[2]。

また，通常のテキスト入力に加え，音声合成向けの記述言語であるSSML（Speech Synthesis Markup Language），及

(注1)　データの中にあるパターンや構造を抽出する手法で，データの分類や，パターン認識，予測に使われる。

びSAPI（Microsoft®(注2) Speech Application Programming Interface）で定義されたXML（Extensible Markup Language）タグフォーマットによる入力が可能である。これにより，入力テキストの中で，任意の語句の強調度合いや話速，声の高さなどを指定できるようになり，言語や話者などを文章の途中で切り替えることもできる。特に，強調タグは，発話の中で重点が置かれている語句を明示するために使われ，対話音声の生成などで有用である。例えば，“これは<emphasis>東芝の</emphasis>音声合成システムです。”のようにタグを用いれば，この発話の強調点が“東芝の”にあることを，リズム（継続時間長）や抑揚（F_0パターン）(注3)といった韻律の変化によって表現することができる。ToSpeak$_{TM}$ V2では，このような任意の語句を強調するために，強調韻律の生成を可能とするF_0制御手法と時間長制御手法を新たに開発して導入した[3]。

3　決定木に基づくF_0制御手法

3.1　F_0制御モデル

文節（アクセント句）単位の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たF_0制御モデルを，**図1**に示す。

文章全体のF_0パターンは，アクセント句ごとのF_0パターンを接続することによって生成される。アクセント句ごとのF_0パターンは，代表パターンコードブックから選択された代表パターンを，音節単位の継続時間長に従って時間軸上で伸縮し，対数F_0軸上で平行移動することによって生成される。代表パターンの選択と対数F_0軸上の移動量（オフセット）の決定は，テキスト解析によって得られるアクセント音節の位置（アクセント型），品詞，係り受けなどの言語属性情報（コンテキスト）に従って行われる。

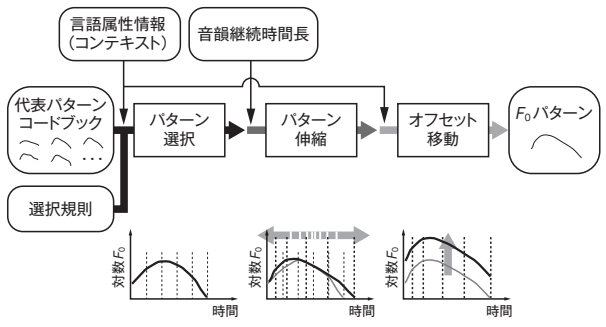

図1. F_0制御モデル — コンテキストに基づいてパターンが選択され，時間軸上，対数F_0軸上でそれぞれ変形されてアクセント句ごとのF_0パターンが生成される。

Fundamental frequency (F_0) control model

（注2）　Microsoftは，米国Microsoft Corporationの米国及びその他の国における商標又は登録商標。

（注3）　イントネーションのことで，基本周波数の変化パターン。

3.2　決定木に基づくクラスタリング

従来のF_0制御手法では，代表パターンコードブックの学習と，代表パターンコードブックに格納されている代表パターンを選択するための規則の学習を個別に行った。このため，代表パターンを生成する際の基準と，コンテキストによって表現される選択規則が必ずしも一致せず，代表パターンの選択誤りによる自然性の劣化が問題であった。そこで，コードブックに格納すべき代表パターンと，代表パターンの選択規則を同時に構築するため，決定木に基づいたコンテキスト クラスタリング手法を導入した。決定木を構築する過程で行うクラスタの分割では，各クラスタに属する学習データに関する誤差が最小となる分割を求めることになる。また，一般的な決定木と同様に，このモデルでも，一つのクラスタを二つに分割する処理は，ほかのクラスタには影響を与えない。そのため，着目している一つのクラスタを分割するときには，局所的に誤差を最小化すればよい。

ここで，ある一つのクラスタ$G(=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots\})$を分割する状況を**図2**に示す。ベクトル$\boldsymbol{p}_n(n=1,2,\cdots)$は学習データ（$F_0$パターン）である。$q_k$は，コンテキストによる分割条件（質問）を表し，クラスタGを二つの子クラスタ$G_{q_k^-(\mathrm{No})}(=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots\})$と$G_{q_k^+(\mathrm{Yes})}(=\{\boldsymbol{p}_3,\cdots\})$に分割する。質問$q_k$の候補としては，“着目するアクセント句のアクセント型が2型(注4)以上か？”，“着目するアクセント句の音節数が10音節以下か？”など，コンテキストにより表現される様々なものが挙げられる。クラスタGを分割する最適な質問q_{best}としては，式(1)で表されるように分割前後での誤差の変化量ΔEが最大となる質問を選択することになる。ただし，分割前のクラスタGでの誤差は，質問によらず一定であるため，実際の基準は式(2)となり，二つの子クラスタそれぞれに関する誤差を計算すればよい。

$$q_{\mathrm{best}}=\arg\max_{q_k}\Delta E(q_k,G) \tag{1}$$

$$q_{\mathrm{best}}=\arg\min_{q_k}\left(\varepsilon(G_{q_k^-},\boldsymbol{c}_-)+\varepsilon(G_{q_k^+},\boldsymbol{c}_+)\right) \tag{2}$$

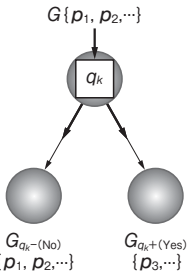

図2. 決定木に基づくクラスタリング — 学習データを，ある質問q_kに対して真（Yes）となる集合（クラスタ）と偽（No）となるクラスタに分割する。

Decision tree-based clustering

（注4）　アクセント句中でアクセントのある音節の位置を型数で表す。

$\boldsymbol{c}_-$及び$\boldsymbol{c}_+$は，それぞれ子クラスタG_{q_k-}，G_{q_k+}に対応する代表パターンを表す。誤差εは，クラスタGとそれに対応する代表パターンベクトル$\boldsymbol{c}$から(3)式で定義される。

$$\varepsilon(G,\boldsymbol{c})=\sum_{\boldsymbol{p}_j\in G}(\boldsymbol{p}_j-\hat{\boldsymbol{p}}_j)^{\mathrm{T}}(\boldsymbol{p}_j-\hat{\boldsymbol{p}}_j) \quad (3)$$

ここで，ベクトル$\boldsymbol{p}_j$は，前記のコンテキスト クラスタリングによって分割されてクラスタGに属する学習データで，ベクトル$\hat{\boldsymbol{p}}_j$は，$\boldsymbol{p}_j$を目標として代表パターンベクトル$\boldsymbol{c}$からこのモデルによって生成されるアクセント句単位のF_0パターンである。代表パターンベクトル$\boldsymbol{c}$は，式(3)の誤差を最小化するように立てた連立方程式を解くことによって求めることができる。

以上によって，代表パターンと決定木による選択規則を同時に生成できる。

3.3 有効性の確認

決定木に基づく代表パターン選択手法の有効性を確認するため，数量化I類(注5)モデルによる代表パターン選択規則に基づく従来手法との比較評価を行った。

学習データとして，1名の成人話者が発声した，音韻と韻律のバランスを考慮した625文章分の収録音声を用いた。代表パターンコードブックには，従来手法と提案手法ともに，0型，1型，…，及び5型以上のアクセント型それぞれに8個ずつ，合計48個の代表パターンを格納した。対数F_0軸上で平行移動するオフセットは，従来手法及び提案手法ともに，数量化I類による推定値を用いた。評価者は8名で，学習データに含まれていない韻律テスト用の102文章から15文章を選んで提示し，どちらの手法がより自然な抑揚かを選択形式で評価してもらった。

この主観による比較評価試験の結果を**図3**に示す。従来手法と提案手法では，95％信頼区間の幅を考慮しても，選択率に20％以上の差異があることがわかり，F_0制御の基本性能が向上していることが確認できた。

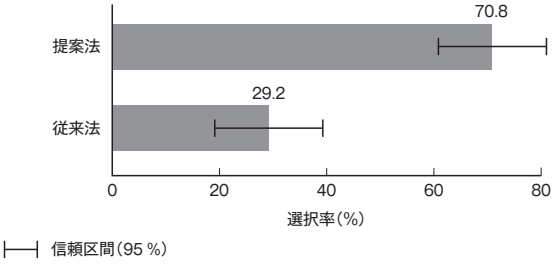

図3. F_0制御手法の対比較評価結果 — 二つの手法によって生成された合成音声を聞いて，どちらの抑揚がより自然かを比較した結果（選択率）である。

Preference scores for F_0 control methods

(注5) 数値ではない質的なデータを説明変数に用いて分析する手法の一つで，ダミー変数を用いた重回帰分析と等価。

4 強調音声を合成するための韻律制御手法

音声合成技術を用いて様々な用途の音声コンテンツを作成するためには，人間の発声に含まれるような局所的な強弱の変化を合成音声に付与できることが重要である。

そこで，入力文章の一部を強調する合成音声を生成するために，新たに強調発声を含んだ音声データベースを作成し，強調情報に基づいて継続時間長やF_0パターンなどの韻律情報を変化させる手法を開発した。

4.1 強調音声データベース

同一文章の指定した箇所（アクセント句）を，“強調なし”と“強調あり”で発声した音声データを収集するために，強調するアクセント句の音節数やアクセント型，品詞などのバリエーションを考慮して約300文章の収録テキストを作成した。これを基に，1名の成人話者による強調部分の異なる約900発声を収録し，強調音声データベースを作成した。

4.2 ベースモデル

既存の韻律制御モデルでは，継続時間長については，積和数量化モデルによって音素ごとの時間長を推定する。また，F_0については，前章で述べた決定木に基づく代表パターン選択と数量化I類によるオフセット推定のモデルによって，アクセント句単位でF_0パターンを推定する。ここでは，これをベースモデルとして用い，以下の改良を行った。

4.3 強調韻律制御モデル

図4(a)に示すように，ベースモデルに強調の有無を表現するための属性を追加し，強調音声データベースを用いて学習することで，強調韻律制御モデルを作成した。追加した属性は，“強調アクセント句との距離”に関する属性で，着目するアクセント句と強調アクセント句の位置関係を表したものである。

時間長制御については，“強調アクセント句との距離×音韻種別×着目する音素のアクセント句中の位置”という項を積和

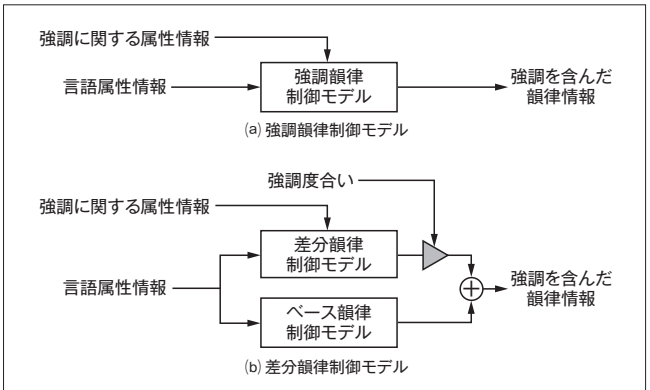

図4. 強調韻律制御モデルと差分韻律制御モデル — 差分韻律制御モデルでは，ベースモデルから生成される韻律情報に，強調の有無によって生じる韻律情報の差分を加算することで強調韻律を生成する。

Emphasized prosody control model and differential prosody control model

一般論文

数量化モデルに加えることで，強調時間長制御モデルとした。

F_0制御については，代表パターン選択モデルとオフセット推定モデルのそれぞれに“強調アクセント句との距離”に関する属性を追加することで，強調F_0制御モデルとした。

4.4　差分韻律制御モデル

4.3節で述べた強調韻律制御モデルは，強調音声データベースを用いてモデル全体を学習することにより，強調に伴う韻律変化を精度よく生成できることが期待される。これに対して，差分韻律制御モデルは，図4(b)のようにベースモデルから生成される韻律情報に，強調の有無によって生じる韻律情報の差分を加算する構成である。このため，強調韻律制御モデルと同等の性能が得られれば，強調発声を収録していない話者や発話スタイルのベースモデルに対しても強調成分を付与することができ，また，強調度合いの制御が容易にできるというメリットがある。

時間長制御については，強調属性を加えて学習した4.3節の強調時間長制御モデルにおいて，強調の有無によって変化する項だけを利用することによって，差分時間長制御モデルとした。つまり，**図5**に示すように，強調音声データベースから学習した積和数量化モデルの係数の一部を用いることで，強調の有無によって生じる差分時間長を推定する。

F_0制御については，強調の有無による抑揚の差分である差分F_0データを抽出し，この差分F_0データから差分F_0パターンを推定するモデルを作成した。差分F_0データは，強調の有無以外の発声内容（音素系列やポーズ位置）がまったく同じ収録音声データ対について，“強調なし”のF_0パターンを，“強調あり”の時間長に合わせて時間軸上で線形伸縮し，F_0の差分を計算したものである。一例として，“10時に”を強調した文章での差分F_0データを**図6**に示す。

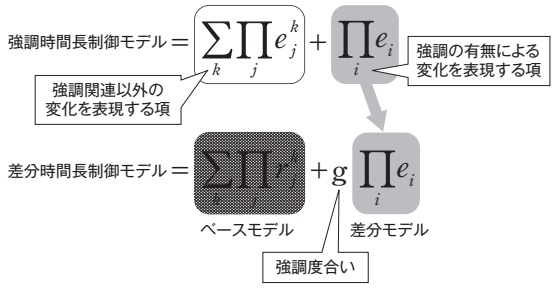

図5．継続時間長の差分制御モデル — 強調時間長制御モデルから強調の有無によって変化する項を取り出し，差分モデルとして利用する。
Differential duration control model

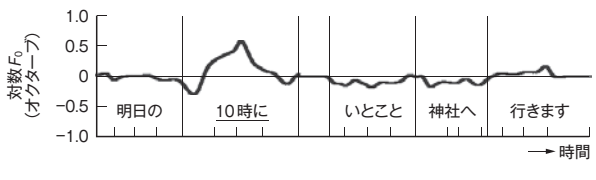

図6．差分F_0データの例 — 収録音声データ対において，強調の有無が異なる部分（“10時に”）に大きな変化（差分）が生じている。
Example of differential F_0 data

4.5　有効性の確認

提案手法の有効性を確認するために，強調の了解性（強調箇所の適切さ）についての主観評価試験を行った。評価は，提示された質問文に対して，回答として再生される合成音声の強調箇所が適切か否かを3段階（3：適切，2：強調が知覚できない，1：不適切）で行った。ここで，各合成音声には，強調箇所が適切となる質問文（Q_c）と強調箇所が不適切となる質問文（Q_w）の2種類が必ず提示される。例えば，“目的地まで，およそ20キロで，30分かかります。”という“20キロで”の部分を強調制御した合成音声に対しては，次の2種類の質問文との組合せで評価が行われる。

- Q_c：“目的地までの距離はどのくらいですか？”
- Q_w：“ここから目的地まで何分くらいかかりますか？”

もし，合成音声に強調による変化が感じられなければ，Q_cが提示された場合とQ_wが提示された場合で評価点に差がつかず，強調が感じられれば，Q_cとQ_wで評価に差が出るはずである。1名の成人話者のデータから学習した次の4種類のモデルを用いて，評価対象の合成音声を生成した。

- A：既存の読上げ調韻律制御モデル
- B：強調韻律制御モデル（強調箇所指定なし）
- C：強調韻律制御モデル（強調箇所指定あり）
- D：差分韻律制御モデル（ベースモデルはAの読上げ調韻律制御モデル）

4.4節で述べたように，Dのモデルでは強調度合いを制御できるが，ここでは特別な調整などは行わず，重み1で差分を加算した。

評価文章は，強調アクセント句のバリエーションを考慮した10文章で，各文章においてC及びDのモデルに対しては強調箇所を1アクセント句だけ指定し，強調するアクセント句を変えて2セットの評価試験を行った。

強調箇所の了解性に関する評価結果を**図7**に示す。A及びBから生成した合成音声については，強調が知覚できないため，質問文の違いによる評価点の間に有意差がなかった。一方，C及びDから生成した合成音声では，強調箇所が適切となる質問との組合せでは強調箇所が適切（3点）と評価されたものが多く，不適切となる質問との組合せでは強調箇所が不適切（1点）と評価されたものが多くなっており，強調に関して十分な了解性が得られていることがわかる。また，同時に実施した韻律の自然性の評価によって，強調効果を付与することで自然性が大きく劣化しないことも確認できた。

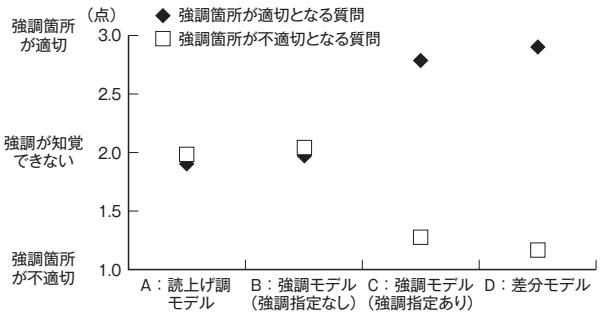

図7. 強調の了解性に関する主観評価結果 — 強調制御を行ったC及びDのモデルについては，質問内容に応じて評価点に差が付いており，強調が十分に知覚されていることがわかる。

Results of subjective evaluation

5　音声合成体験Webサイト“Studio ToSpeak”

当社は，多種多様な顧客の用途にいち早く低コストで対応するための，サービス事業の展開に向けた検討を始めており，Web対応型の音声合成サーバ・クライアントシステムの開発を進めている。

その第一弾として，当社の音声合成技術のアピールとユーザーからの意見や要望の収集を目的とした音声合成体験WebサイトStudio ToSpeak（http://tospeak.toshiba.co.jp/）を一般公開した（**図8**）。このサイトでは，簡単なユーザー登録を行うだけで，ToSpeak$_{TM}$ V2を利用したWebアプリケーションにより，任意の文章から様々なキャラクターの声の合成音声を無料で生成することができる。また，制御タグによる合成音声の調整や，作成した合成音声のダウンロードなどもできるようにしており，ユーザーに音声合成技術の可能性を体感してもらえる仕組みとなっている。

図8. 音声合成体験WebサイトStudio ToSpeak — 当社の音声合成技術のPRとユーザーからの意見や要望の収集を目的としたWebサイトで，簡単なユーザー登録だけで無料で利用できる（http://tospeak.toshiba.co.jp/）。

Website "Studio ToSpeak"

6　あとがき

音声合成システムToSpeak$_{TM}$ V2は，精度向上を達成した新しいF_0制御手法や強調音声のための韻律制御手法などにより，多様な音声コンテンツに応じた合成音声を生成できる。当社では更に，これまでのミドルウェア製品に加え，サービス事業の展開に向けて，ToSpeak$_{TM}$ V2を利用したWeb対応型の音声合成サーバ・クライアントシステムの開発も行っており，音声合成体験WebサイトStudio ToSpeakを公開した。

広がりつつある音声合成技術の新しい用途に対応するため，今後も，基本音質の向上に関する研究に加え，任意の話者の合成音声を少量のデータから生成する話者適応技術や，感情表現などの新たな発話スタイルについて研究開発を進め，合成音声のバリエーションの強化を図っていく。

文　献

(1) 籠嶋岳彦．高音質で聞きやすい音声合成システムToSpeak$_{TM}$．東芝レビュー．62，12，2007，p.34－37.
(2) 水谷伸晃，ほか．“決定木に基づいた代表パターン選択手法の検討”．日本音響学会講演論文集．仙台，2005-09，日本音響学会．2005，p.340－341.
(3) 平林　剛，ほか．“強調音声のための韻律制御手法の検討”．日本音響学会講演論文集．福岡，2008-09，日本音響学会．2008，p.343－344.


平林　　剛　HIRABAYASHI Go
研究開発センター 知識メディアラボラトリー研究主務。
音声合成技術の研究・開発に従事。日本音響学会会員。
Knowledge Media Lab.


水谷　伸晃　MIZUTANI Nobuaki
研究開発センター 知識メディアラボラトリー。
音声合成技術の研究・開発に従事。日本音響学会会員。
Knowledge Media Lab.


籠嶋　岳彦　KAGOSHIMA Takehiko, D.Eng.
研究開発センター 知識メディアラボラトリー主任研究員，工博。
音声合成技術の研究・開発に従事。電子情報通信学会，日本音響学会会員。
Knowledge Media Lab.