多様な声や感情を豊かに表現できる音声合成技術

Text-to-Speech Technologies Realizing Various Voices and Expressive Reading

森田　眞弘　■MORITA Masahiro
田村　正統　■TAMURA Masatsune
布目　光生　■FUME Kosei

音声合成が電子書籍の朗読やエンターテインメント（エンタメ）向けの応用などに幅広く使われるようになるにつれ，より多様な話者，発話スタイル，及び感情を表現できる音声合成へのニーズが高まっている。

東芝は，こうしたニーズに応えるため，特定の人物の声質や口調に似た音声を合成できる音声合成辞書を低コストかつ短期間で作成可能な技術や，小説などのせりふを感情豊かに読み分ける技術，意図した抑揚の合成音声を効率よく作り込める韻律編集技術，及びなりすましといった合成音声の悪用を抑止できる電子透かし技術などを開発した。

As text-to-speech (TTS) technologies are now widely used for e-book reading and entertainment applications, improvement of their ability to provide various types of voices, speaking styles, and emotions has become a focus of attention.

In response to this need, Toshiba has developed the following advanced TTS technologies: (1) a custom voice production technology that can build a wide variety of voices closely resembling the voices of specific people at low cost and within a short time; (2) an expressive reading technology that can automatically select emotions from respective dialogues in such works as novels; (3) a prosodic authoring technology that can efficiently create speech contents with the intended intonation; and (4) a digital watermarking technology that prevents the misuse of TTS, such as for identity theft.

1　まえがき

音声合成は，テキスト（文章や文字）を音声に変換する技術である。これまでの性能向上により，単に情報を声で伝える目的には十分な音質が実現されており，様々な機器やサービスで活用されている。音声合成により，プロのナレーターから生の声を録音するよりも低コストかつ気軽に音声コンテンツを作成でき，内容の更新も迅速に行える。また，録音では実現できない，リアルタイムに更新される情報の読上げも可能になる。

近年では更に，合成音声による電子書籍の朗読や，有名人やユーザーといった特定の人に似た合成音声による音声コンテンツの作成など，様々な応用へのニーズが高まっている。これらの応用分野では，ユーザーの好みの話者や，内容に合った話者及び発話スタイルで，感情豊かに読み上げたり，必要に応じて細かく調整したりできることが期待される。また，有名人やユーザーに似た合成音声が悪用されることを極力防ぐ必要がある。

東芝はこれまで，多様な音声の合成が可能な基本技術[1]を開発して，カーナビなどの組込み機器向けミドルウェアや電子書籍の音声読上げ機能，及び音声合成クラウドサービスなどを商品化し，その基本品質や声のバリエーションは高い評価を得ている。更に多様性を高め，前述のニーズに応えるため，特定の人物の声質や口調に似た音声を合成できる音声合成辞書を低コストかつ短期間で作成可能な技術や，小説などのせりふを感情豊かに読み分ける技術，意図した抑揚の合成音声を効率よく作り込める韻律編集技術，及びなりすましといった合成音声の悪用を抑止できる電子透かし技術などを開発した。以下，これらの技術について，概要を述べる。

2　多様性向上技術

当社の音声合成技術は，録音された音声から話者や発話スタイルの特徴を精度よくモデル化できることが一つの大きな特長で，これまでにも多様な音声合成を実現してきた[1]。多様性を更に向上させるために新たに開発した技術について，以下に述べる。

2.1　カスタム音声合成辞書の作成技術

特定の話者に似た声質や口調の音声を合成したいというニーズは以前からあったが，合成音声の品質向上に伴い，そのニーズは高まっている。例えば，エンタメ向けコンテンツを有名人の声でしゃべらせたいというニーズや，喉頭部の摘出で声を失ってしまう人が自分の声を音声合成で残したいというニーズがある。今後，自分や身近な人の音声合成辞書を容易に作れるようになれば，それらを家族や友人と共有したり一般公開したりするといった，新たな使い方が広がる可能性もある。

こうしたカスタム音声合成辞書を作成するには，大きく次の三つのプロセスが必要である。

（1）　話者から音声を収録
（2）　収録した音声を分析してデータベース（DB）化
（3）　音声DBからその話者の音声合成辞書を作成

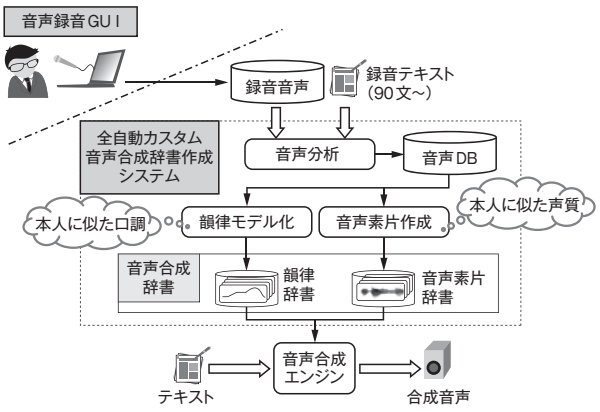

図1．カスタム音声合成辞書作成の枠組み ― 音声を録音する音声録音GUIと，録音した音声から音声合成辞書を全自動で作成する全自動カスタム合成辞書作成システムから成る。

Framework of custom voice production

従来，(1)は音声収録専用のスタジオを利用して数日かけて収録し，(2)は自動分析の後，人手による誤り修正や各種補正を1～2か月かけて行い，(3)は自動で音声合成モデルを学習後に数週間かけて人手で調整していた。このように，ひとりの話者の辞書を作成するのに約2～3か月の期間と多大なコストがかかり，普及の妨げになっていた。

そこで，一般のユーザーでも気軽に音声の収録ができ，それ以降の音声の分析処理から音声合成辞書の作成までを全自動で行える枠組みを開発した[2]。その概要を**図1**に示す。

この枠組みでは，ユーザーはまず，Web上の音声録音GUI（Graphical User Interface）ツールで，最低約90文のテキストを読み上げてその音声を録音する。このツールでは，1文ごとの録音ができ，画面には，読み上げる文とともに読みやアクセントが視覚的に表示される。更にユーザーは，みずからが読み上げた音声を聞いて確認できる。収録用のテキストは，音声合成辞書を作成するのに必要な音韻や韻律のパターンを，少数の文でバランスよくカバーするよう厳選した。

録音された音声は，サーバ上の全自動カスタム音声合成辞書作成システムに送られ，音声中の各音素の開始時刻にラベルを付与したり，抑揚を表す基本周波数を抽出したりする音声分析が全自動でなされる。分析された結果を基に，抑揚やリズムを表す韻律のモデル及び声質の特徴を表す音声素片が全自動で学習される。ユーザーは，30分程度かけて約90文のテキストの読上げ音声を録音すれば，約1時間後に自分の声の音声合成辞書が使えるようになる。

2.2　話者適応技術によるカスタム音声合成辞書の作成

日本語に加え，アメリカ英語（米語）と中国語についても，カスタム音声合成辞書の作成環境を開発した。

これらの言語では，隠れマルコフモデル（HMM）という統計モデルに基づく音声合成方式（以下，HMM方式と呼ぶ）を採用している。HMM方式では，音声信号を分析して得られる，スペクトルや基本周波数などの音響・韻律パラメータの時系列を，HMMと決定木で統計的にモデル化し，これらを音声合成に用いる。声質や韻律を，音声波形ではなく音響・韻律パラメータのレベルで柔軟に操作でき，話者に適応しやすい。言語への依存性が低いことも特長で，近年の音質向上に伴い，音声合成方式の主流となってきている。当社も，話者適応などの活用による多様性の向上や多言語化を効率的に進めるため，当社の海外研究開発拠点と連携してHMM方式を開発した[3]。欧米言語や中国語に適用し，日本語への適用も進めている。

このHMM方式向けのカスタム音声合成辞書の作成では，音声DBから音声合成辞書を作成するステップに，話者適応技術を導入した。具体的には，複数話者の大量の音声から，話者共通の特徴を精緻にモデル化したベースモデルをあらかじめ作っておき，このモデルを音声DBの話者の特徴に合わせ込むことで，その話者のモデルを生成する。その結果，言語的な特徴は精緻に表現しながら，声質や口調は録音音声の話者に似た音声合成辞書が作成できる。

話者適応を用いることで，録音した音声だけから言語的な特徴や話者の特徴の全てを学習する場合よりも，安定した辞書作成が可能になった。現状，100文程度の収録音声から約1時間で新たな辞書が生成できるが，収録音声の量に応じた品質を実現する，より柔軟な辞書作成も可能になると考えている。

2.3　せりふの感情の自動読み分け技術

喜怒哀楽などの感情は，抑揚やリズムといった韻律に加え，声の質にも強く影響する。情報伝達が主目的の応用では感情は特に必要とされないが，小説のせりふやエンタメ向けコンテンツが感情なく淡々と読み上げられると違和感が生じる。

そこで，典型的な感情である“喜び”，“怒り”，及び“悲しみ”の音声合成辞書を開発した。ナレーターに各感情ごとに数百文ずつのテキストを読ませて音声を収録し，前述のカスタム音声合成辞書の作成技術を用いて，各感情の音声データから辞書を作成した。

更に，これらの感情ごとの音声合成辞書を自動的に使い分ける技術を開発した[4]。**図2**に示すように，小説などのせりふのテキストからもっとも適切な感情を自動で推定し，推定された感情に従って辞書を切り替えることで，せりふの感情を読み分ける電子書籍ビューアを試作した。

テキストからの感情推定では，喜び，怒り，悲しみと感情を含まない“平静”の4種類の感情ラベルを文単位で推定する。入力文に対し，各感情の推定モデルを用いてスコアを算出し，もっともスコアの高い感情を選択する。推定モデルには，メンテナンスの容易さや拡張性を考慮して，ナイーブベイズに基づく統計モデルを採用した。このモデルを用いて，ある文sが感情cになるスコア$Ec(s)$を式(1)で求める。

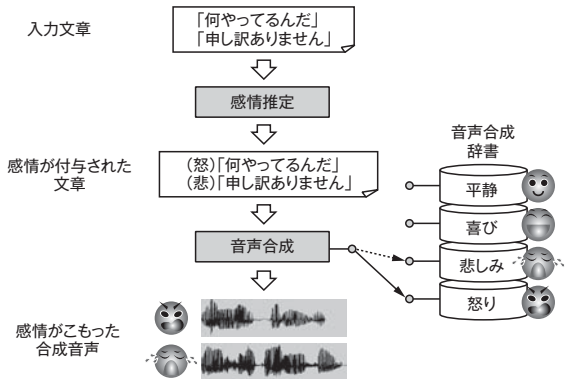

図2．せりふの感情の自動読み分け処理 — 入力されたせりふのテキストに対して感情を推定し，推定された感情に従って音声合成辞書を切り替えることで，感情豊かな合成音声を再現する。

Expressive reading processes using automatic emotion selection from dialogues

$$Ec\,(s) = p\,(c)\,p\,(w_1 \mid c)\,p\,(w_2 \mid c) \ldots p\,(w_n \mid c) \qquad (1)$$

$p\,(c)$ ：感情cの出現確率

$w_1, \ldots, w_n$ ：文sを構成する各単語（nは文s中の単語数）

$p\,(w \mid c)$ ：感情cが与えられたときの単語wの出現確率

ここで，$p\,(c)$及び$p\,(w_k \mid c)$は，文単位の感情ラベル付けを手作業で行ったテキストコーパスから学習する。

この手法により，せりふのテキストから80％程度の精度で感情ラベルが推定できる。

試作した電子書籍ビューアでは，前記の手法でせりふの感情を読み分けるのに加え，地の文とせりふで話者を切り替える。これらの機能により，音声を聞くだけでも内容がわかりやすく，違和感を軽減した読上げが可能である。

2.4　クラスタ適応学習による任意感情の合成技術

東芝欧州研究所では，クラスタ適応学習（CAT：Cluster Adaptive Training）という学習方式をHMM方式の音声合成に導入し，複数の感情を含む音声データからそれらを同時にモデル化して，任意の感情で音声合成できる方法を開発した[5]。

CATでは，モデルを複数クラスタの重み付き和で表し，モデルの学習時には，各クラスタのモデルと重みを，データに合わせて同時に最適化する。HMM方式には，テキスト情報とHMMの統計量を対応付ける決定木のそれぞれを，複数個の決定木の重み付き和に置き換えることで，CATを導入した。

複数感情のモデル化では，複数の感情を含む音声データから決定木と重みを同時に，かつ重みは感情ラベルごとに最適化する（**図3**）。その結果，重みを各感情に対応した値に設定すると各感情が再現でき，重みを感情間で補間することなどにより，中間的な感情など任意の感情が表現できる。

この方式は，感情だけでなく話者や言語にも適用可能で，多様性を向上させる様々な展開が期待できる。

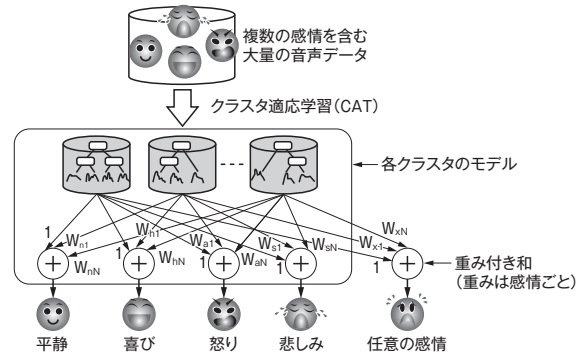

図3．クラスタ適応学習による任意感情の合成 — 複数のクラスタのモデル（決定木）を感情ごとの重みを付けて加算することで，複数感情を表現する。

Synthesis of arbitrary emotions applying cluster adaptive training (CAT)

3　効率的な韻律編集技術

自動で生成した合成音声では，部分的に不自然な韻律になったり，韻律が多様なあいさつや語尾表現などで満足できない韻律になったりする場合がある。

このような場合にユーザーが韻律を編集できるツールがいくつか実用化されているが，韻律パターンをマウスで直接編集するなど音声に対する高度な知識が必要なものが多く，いずれも思いどおりの音声を作るのは難しい。

これを解決するため，ナレーターの実際の発声から取り出し

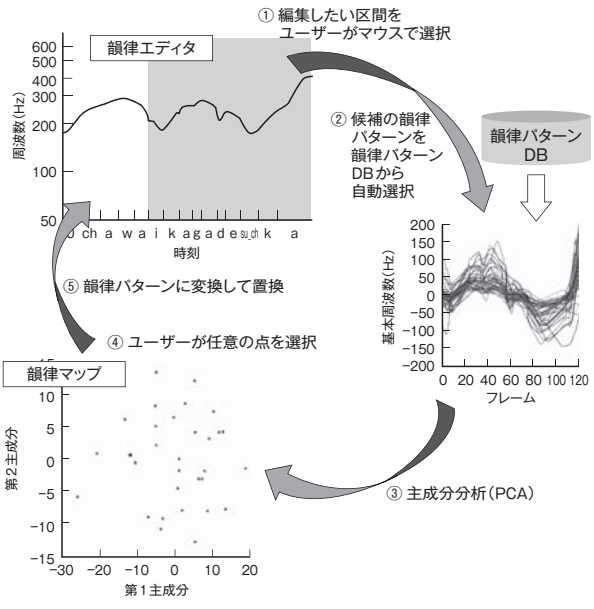

図4．PCAを用いた韻律パターンの選択・置換方法 — 韻律エディタで編集したい箇所を指定すると，候補の韻律パターンが韻律マップ上の点として表示される。韻律マップ上の任意の点を選択し，対応する韻律パターンで置き換える。

Selection and replacement of prosodic patterns using principal component analysis

た韻律パターンをあらかじめ大量に用意し，その中から提示される候補のパターンからユーザーが適切な韻律パターンを選択することで，思いどおりの韻律に調整できる手法を開発した[6]。その概念を**図4**に示す。

この手法では，提示される韻律パターン候補の中からユーザーが適切なパターンをいかに簡単に選べるかが鍵であるが，大量の韻律パターンがそのまま提示されても選択は困難である。そこで，韻律パターンの候補を主成分分析（PCA）し，パターンを，第1主成分と第2主成分を座標とする二次元平面上の点として可視化することで，その平面上を走査して思いどおりの韻律パターンが容易に探索できるインタフェースを開発した。

この二次元平面では，異なる韻律パターンは離れて分布する一方，似たパターンは近い位置に分布するため，まずは大ざっぱに探索して当たりをつけ，次に周辺を細かく探索することで，思いどおりの韻律パターンを効率的に探索できる。

4　なりすまし防止のための電子透かし技術

前述のカスタム音声合成辞書の作成や韻律編集の技術が進み，本人と区別がつかない音声が容易に作れるようになれば，合成音声がなりすましなどに悪用される危険性が高まる。

そこで，音質にほとんど影響しない電子透かしを合成音声に埋め込むとともに，合成音声からこの電子透かしを検出できる技術を開発した[7]。人間の聴覚が位相の変化に鈍感であることを利用し，合成音声の位相を緩やかに変化させることで透かしを埋め込む。具体的には，有声音の音源信号（ピッチ周期間隔のパルス）の各パルスの位相を式(2)により変調する。

$$ph_f(t) = 2\pi at \mod 2\pi \qquad (2)$$

$ph_f(t)$　：時刻tに中心があるパルスの，周波数fの成分の位相
a　：位相の変調周波数
$x \bmod y$：xをyで割った余り

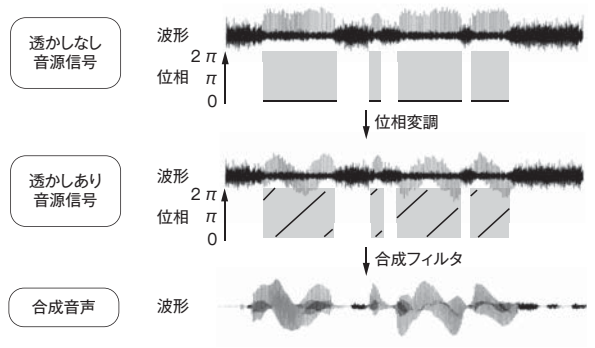

図5．位相変調による電子透かしの挿入 ― 有声音の音源信号を位相変調して電子透かしを埋め込み，合成音声を生成する。
Embedding of digital watermark by means of phase modulation

これを合成フィルタに通すことで，位相が変調された音声波形が生成される（**図5**）。一方，電子透かしの検出では，合成音声の波形を分析して位相の時系列を求め，その傾きがaに近い値かどうかで判定する。

ノイズや残響，音声符号化といった音声の劣化要因がまったくない条件では，透かしを100％検出でき，いずれか一つの劣化要因を与えた条件では，90％以上の精度で検出できる。しかし，様々なひずみが複合する実環境下での検出精度は条件によって差が大きく，その改善が今後の課題である。

5　あとがき

音声合成の多様性を向上させるために開発した様々な技術について述べた。これらの技術により，利用シーンや内容に応じた多様な話者や発話スタイル及び感情を表現できる合成音声が生成できる。今後も，多様性と音質の両面を更に向上させながら，音声合成の適用範囲を広めていく。

文　献

(1) 平林 剛 他. 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. 東芝レビュー. **65**, 4, 2010, p.43－47.
(2) 橘 健太郎 他. “個人声の合成音作成フレームワークの開発”. 日本音響学会2011年春季研究発表会講演論文集. 東京, 2011-03, 日本音響学会. 1-Q-34C.
(3) 田村正統 他. “HMM音声合成による英語音声合成システムの開発”. 日本音響学会2011年春季研究発表会講演論文集. 東京, 2011-03, 日本音響学会. 3-7-7.
(4) 布目光生 他. 自然で聞きやすい電子書籍読上げのための文書構造解析技術. 東芝レビュー. **66**, 9, 2011, p.32－35.
(5) Latorre, J. et al. "Speech Factorization for HMM-TTS Based on Cluster Adaptive Training". Proc. INTERSPEECH2012. Portland, OR, USA, 2012-09, ISCA. p.971－974.
(6) 森 紘一郎 他. “主成分分析を用いた韻律編集インタフェース”. 日本音響学会2013年春季研究発表会講演論文集. 八王子, 2013-03, 日本音響学会. 3-P-30B.
(7) 橘 健太郎 他. “位相変調に基づくHMM音声合成向け電子透かし方式の提案”. 日本音響学会2013年春季研究発表会講演論文集. 八王子, 2013-03, 日本音響学会. 1-9-2B.


森田　眞弘　MORITA Masahiro
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。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.


田村　正統　TAMURA Masatsune, 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，IEEE会員。
Knowledge Media Lab.


布目　光生　FUME Kosei
研究開発センター 知識メディアラボラトリー研究主務。文書解析及び情報抽出技術の研究・開発に従事。ACM会員。
Knowledge Media Lab.