統計的なパラメータ選択で声の再現性を高める次世代音声合成技術

Next-Generation Speech Synthesis Technology with Enhanced Voice Quality and Speaker Similarity Based on Statistical Parameter Selection

田村　正統
■TAMURA Masatsune

音声合成技術は，特定の人に似た声による音声対話システムなど，声の再現性の高さが求められる応用に広がっており，将来的には，ナレーション音声と遜色ない品質が求められる映像制作など，コンテンツ制作分野への利用拡大が見込まれている。

東芝は，音声合成の利用拡大を図り，トップレベルの音声品質を維持するために，次世代に向けた新方式の音声合成技術を開発した。今回開発した新方式の特長は，パワースペクトルとともに波形形状を表す位相情報も含めて音声を精密に表現する新たな特徴パラメータと，統計的に学習した音響モデルに基づいて特徴パラメータを選択するパラメータ選択とを利用する点にある。評価試験により，新方式では音質の改善と，声の再現性を向上させた高品質な合成音声が得られることを確認した。

The application of text-to-speech (TTS) systems with high voice quality and speaker similarity has been expanded to various fields in which the reproduction of more natural synthesized voices is required, such as speech dialogue systems using synthetic speech similar to that of a specific person. Furthermore, the use of TTS systems is expected to expand in the content production field to realize synthetic speech with high quality comparable to that of a narrator's voice in the future.

With this as a background, Toshiba has developed a next-generation speech synthesis technology incorporating precise speech analysis that uses phase spectrums representing the shapes of waveforms along with power spectrums and a parameter selection method that selects acoustic feature parameters based on statistically trained acoustic models. We have conducted evaluation experiments and confirmed that this next-generation speech synthesis system provides synthetic speech with higher voice quality and speaker similarity compared with conventional methods.

1　まえがき

音声合成技術は，カーナビなどの音声案内に加え，音声対話システムや，ゲームやスマートフォンのアプリケーションを含むコンテンツビジネス，アクセシビリティ対応[1]などへの利用が進んでいる。東芝の音声合成システムToSpeak™[2]は，組込み機器向けのミドルウェアとともに，クラウドサービスのプラットフォームであるRECAIUS™もサポートし，様々な用途の製品に適用されている。このような音声合成技術の広がりに伴って，合成音声に対する要求品質も高まり，より自然な発話で，かつ音声の再現性も高い高品質なシステムが求められている。

当社の現行方式の音声合成技術[3), (4]は，HMM（隠れマルコフモデル）という統計モデルに基づいており，音質の改善と，東芝欧州研究所 ケンブリッジ研究所及び東芝中国社 研究開発センターと共同して進めてきた多言語対応に加え，少量の録音音声で目標話者に似た声（似声）を実現する話者適応や感情制御など，多様性の向上に着目して開発を進めてきた。

一方，次世代に向けた音声合成のコア技術は，高品質化を目指し，音声の再現性を高めることを目標として開発を進めている。今回開発した新方式の主な特長は，位相情報も含めて分析することで音声を精密に表現できる新たな特徴パラメータ[5]と，統計モデルに基づいて特徴パラメータを選択し，これを利用して音声波形を生成する点にある。ここでは，開発した新方式の概要と，その特性評価結果について述べる。

2　次世代音声合成システムの特長

新方式，現行方式，及び他社で採用されている波形選択方式[6]の比較を**表1**に示す。

現行方式では，音響モデルとして学習したHMMから特徴パラメータを生成し，音声波形を生成する。少ないメモリ量で安定した音質が得られるものの，過剰な平滑化が生じるために音声の再現性は低下する。

波形選択方式は，音声波形の選択と接続により波形生成する。選択した区間の再現性は高まるものの，合成する各区間の音韻や韻律属性に対応した適切な音声データが音声コーパス（大規模に音声を集めたデータ）に含まれていない場合には，

表1．各方式の比較
Comparison of features of speech synthesis methods

項　目	新方式	現行方式	波形選択方式
音質（再現性）	◎	△	◎
音質（安定性）	○	○	△
多様性	○	◎	×
メモリ量，演算量	中～大	小	大

◎：非常に良い　○：良い　△：普通　×：悪い

不連続感や韻律の不一致が生じる。このため，自然な合成音声を得るためには，更に多くの音声データが必要になる。

新方式は音声コーパスを分析した特徴パラメータを選択し，選択された特徴パラメータに基づいて滑らかな特徴パラメータ系列を生成し，音声波形を生成する。現行方式と比較すると，統計的にモデル化した特徴パラメータではなく，選択された特徴パラメータを用いるため，過剰な平滑化が抑制され音声の再現性が向上する。また波形選択方式と比較すると，韻律パラメータとスペクトルパラメータをそれぞれ個別に選択して生成するため，各パラメータを滑らかに生成できる。これにより新方式では，波形接続時の歪み（ひずみ）や韻律の誤りによる音質劣化を抑えた安定した音質を，中規模の音声データ量でも得ることができる。

一方，新方式は音声波形の選択ではなく，特徴パラメータの選択に基づくため，特徴パラメータの分析によって生成する音声波形での音質劣化を抑える必要がある。従来の音声分析処理を用いた場合には，分析元音声と特徴パラメータから合成した音声との間に差異が生じ，音質的な歪みとして知覚される。そこで，この歪みを抑えるために，音声波形の再現性を高める新たな特徴パラメータを開発した。

このように新方式は，再現性の高さと安定性を両立できることが特長であり，クラウドシステムやオンプレミスシステムといったサーバ型の音声合成に適した高品質な合成システムを構築できる。また多様性の観点では，新方式は現行方式と同様にパラメトリックな合成システムであるため，波形選択方式では困難な，統計的なパラメータの変換や重み制御など，将来的に多様性を拡張する技術を導入可能である。

3　音声合成システムの構成

開発した新方式の音声合成システムの構成を**図1**に示す。この音声合成システムは，HMM系列作成部，音声分析部，パラメータ選択部，及びパラメータ生成部から構成される。音声コーパスを分析して特徴パラメータを集め，この特徴パラメータから統計モデルのHMMを学習する。HMM系列作成部では，テキスト解析結果から得られる，音素系列やアクセント情報などを含むコンテキスト情報を用いて，HMM系列を作成する。パラメータ選択部では，HMM系列に基づいて，最長一致基準及び尤度（ゆうど：モデルに対するデータの適合度合い）最大化基準により特徴パラメータを選択する。特徴パラメータ生成部では，選択された特徴パラメータから動的な特徴量の分布も考慮した滑らかな特徴パラメータ系列を生成する。最後に特徴パラメータ系列から波形生成し，合成音声を得る。

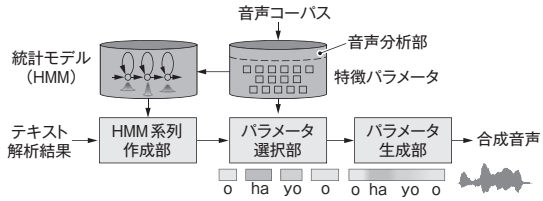

図1．新方式の音声合成システムの構成 ― HMM系列作成部，音声分析部，パラメータ選択部，及びパラメータ生成部から成る。

Configuration of next-generation speech synthesis system

4　ピッチ同期帯域群遅延による音声分析合成

群遅延とは，位相スペクトルの周波数変化であり，帯域ごとにパワースペクトルで重み付け平均したものを，帯域群遅延（BGRD：Band Group Delay）と呼ぶ。これは，各帯域の平均時間を表している。

新方式の分析合成技術[5]の特長は，①固定のフレームレートではなく，波形の周期性に応じて音声分析を行うピッチ同期分析に基づいて精密に音声分析する点と，②波形のパワースペクトルだけでなく形状まで再現するために，BGRDパラメータと位相を補正するパラメータに基づくBGRDC（BGRD with Phase Compensation）パラメータ（位相パラメータ）も特徴パラメータとして導入している点にある。

音声分析部の構成を**図2**に示す。まず，音声波形と周期波形の各周期に対応した時刻を示すピッチマークからスペクトルパラメータとしてメルLSP（MLSP：Mel Line Spectrum Pair）を求め，BGRDCパラメータを求める。次に，各帯域における雑音成分の強度である帯域雑音強度（BAP：Band Aperiodicity）を求める。更に，基本周波数（対数F0）もピッチマークから求めておく。図2は音声波形，ピッチマーク，及びハニング窓により切り出したピッチ波形の例を示しており，ピッチマークの各時刻を中心とするピッチ波形に対して，パワースペクトル分析，BGRD分析，及びBAP分析を行う。

このように，BGRDCパラメータを用いることで，位相スペクトルの再現性を高めている。BGRDパラメータだけから生成した位相（BGRD 位相）とBGRDCパラメータを用いた場合の位相（BGRDC位相）の比較，及び生成したピッチ波形例を，**図3**に

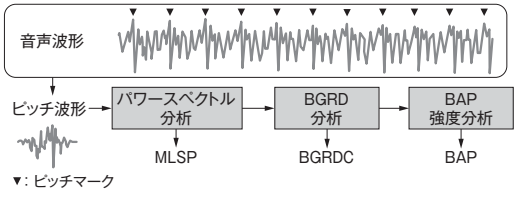

図2．音声分析部での処理フロー ― 音声波形から切り出したピッチ波形に対して，パワースペクトル分析，BGRD分析，及びBAP分析を行う。

Flow of processes in speech analysis part

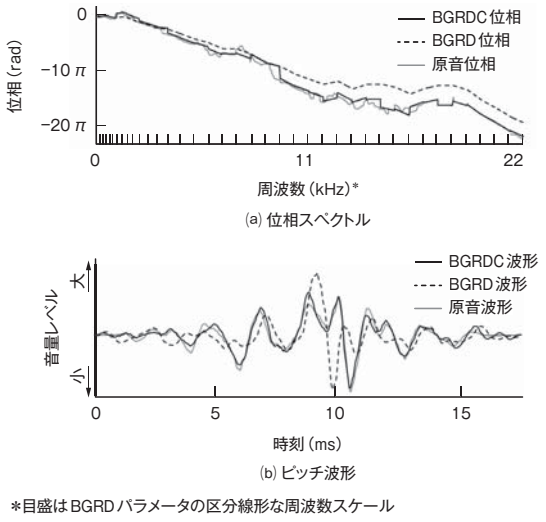

図3. 帯域群遅延の分析例 — BGRDに位相補正も加えた新方式のBGRDCは，原音波形に近い位相スペクトルとピッチ波形が得られることがわかる。

Examples of band group delay analysis

示す。新方式を用いることで，波形の再現性が高まっていることがわかる。また，現行方式や波形選択方式との比較評価を行い，新方式では特徴パラメータにより音声波形の再現性が向上することを確認している[5]。

5　音響モデル学習と統計的パラメータ選択

新方式のパラメータ選択部では，過剰な平滑化を抑えるため，音声コーパスの特徴パラメータを候補とし，HMM系列作成部で作成した分布列に基づいて特徴パラメータを選択する。HMMとしては，各状態の出力分布を決定木クラスタリングしたモデルを用いる。決定木は，音素やアクセントに関する質問など，コンテキストを分割する質問を各ノードに持つ2分木で，リーフノードに出力分布を持つ。固定のフレームレートによる分析ではなくピッチ同期分析を用いているため，フレーム数ではなく時刻をパラメータとする継続長分布を用いる点も，新方式の特長の一つである。このため，合成時には，生成した継続長とピッチから各状態のピッチ波形数を定める。

決定木の各リーフノードに，それぞれの分布の学習データとして用いた特徴パラメータを保持し，それらを候補として特徴パラメータの選択を行う。この概念を**図4**に示す。HMMの各状態の決定木をたどって分布を定め，その学習データを候補（図中の黒丸）として，特徴パラメータの選択を行う。継続長，対数F_0，及びMLSPの各ストリーム（パラメータの系列）に対し，最適特徴パラメータ系列を生成する。各特徴パラメータを選択して生成するため，自然な韻律を生成でき，それぞれのパラメータを編集して用いることもできる。ストリームsに対する特徴パラメータ系列をO_sとすると，最適特徴パラメータ系列$\hat{O}_s$は，式(1)に示すようにモデルλに対する尤度最大系列により求まる。

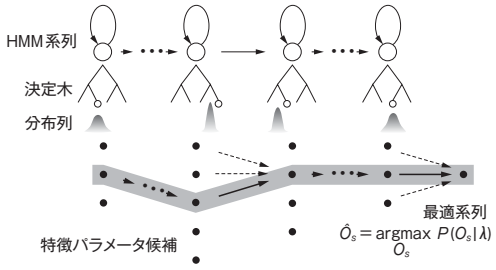

図4. 統計的パラメータ選択 — HMM系列の各状態の決定木をたどって分布を定め，最適特徴パラメータ系列を選択する。

Example of statistical parameter selection

$$\hat{O}_s = \underset{O_s}{\mathrm{argmax}}\, P(O_s \mid \lambda) \tag{1}$$

ここで，尤度関数$P(O_s|\lambda)$は式(2)によって表され，モデルq，状態j，ストリームsにおける各状態の尤度と，先行状態$pre(jq)$に$O_{s(pre(jq))}$が観測されたときの接続に関する条件付き尤度により求められる。

$$P(O_s \mid \lambda) = \prod_{all\ j,\ q} P(O_{sjq} \mid s, j, q, \lambda) P(O_{sjq} \mid O_{s(pre(jq))}, \lambda) \tag{2}$$

式(1)から$\hat{O}_s$を求めるために，動的計画法を用いて，最尤（さいゆう）特徴パラメータ系列を選択する。選択の際には，最長一致優先探索を行い，音素列や韻律属性列が一致する場合は連続する音声データから選択される系列を優先する。これにより，音声コーパス内に入力文と同じ文が含まれる場合には，その音声によるパラメータ系列が選択され再現性が向上する。また，最適系列だけでなく，各状態に対して複数の特徴パラメータを選択し平均化する処理も行うことで，不自然な抑揚や不連続感が抑えられる。最適特徴パラメータと，複数を選択して平均化した特徴パラメータを切り替えて用いることで，再現性の高さと安定性を両立している。

パラメータ生成部では，選択された特徴パラメータを用いて，分布系列の平均ベクトルを置き換える。このとき更新した分布列から特徴パラメータ系列生成を行い，波形を生成することで，合成音声を得る。

6　評価試験

新方式の音声合成システムを用いて，評価試験を行った。試験では，女性2名と男性1名の話者に対して，それぞれ約3,500文の音声コーパスを用いて音声合成システムを構築した。音声

コーパスの文章は，現代日本語書き言葉均衡コーパス[7]から，音韻バランス文や韻律バランス文を含む収録原稿を作成して用いた。また特徴パラメータは，39次のMLSP，38次のBAP，75次のBGRD，及び対数F_0をパラメータとして用い，動的特徴量としてはΔパラメータ（1次動的特徴量）及びΔ^2パラメータ（2次動的特徴量）を使用した。評価音声は，原音，新方式による合成音声，及び現行方式のHMM音声合成による合成音声を用いた。評価文章として，音声コーパスに含まれない学習外文章（13文），及び音声コーパスに含まれる学習内文章（10文）と学習外文章（10文）を使用した。被験者は約20名で，クラウドソーシングによる一般被験者の評価を用いた。

音質評価結果を**図5**に示す。それぞれの合成音を5段階（5：非常に良い，4：良い，3：普通，2：悪い，1：非常に悪い）の平均オピニオン評点（MOS）で評価した結果である。図5より，どの話者においても，現行方式（平均3.21），新方式の学習外文章（平均3.48），新方式の学習内文章（平均3.91）の順にMOSは高くなり，原音（平均4.19）に近づくことがわかる。新方式の学習内文章は原音に近いMOSが得られており，新方式の学習外文章も現行方式に比べて約0.3ポイント改善した。これらの結果から，新方式による音質改善の効果を確認した。

類似度評価結果を**図6**に示す。原音とともに合成音を提示し，同じ人の声に聞こえるかどうかを5段階（5：よく似ている，4：似ている，3：やや似ている，2：あまり似ていない，1：似ていない）のMOSで評価した。音質評価と同様に，現行方式（平均2.88），新方式の学習外文章（平均3.52），新方式の学習内文章（平均4.57）の順にMOSが高くなっている。特に，新方式の学習内文章は高いMOSが得られており，新方式の学習外文章も現行方式に比べて約0.6ポイント向上した。このように，類似性評価結果には明確な差が見られ，音質改善ともに，声の再現性が大きく向上していることを確認した。特に，音声コーパス内の文章の再現性が高まるため，用途に応じた文章をあらかじめ収録することで，定型的な文章は録音音声に近い合成音声を生成できる。

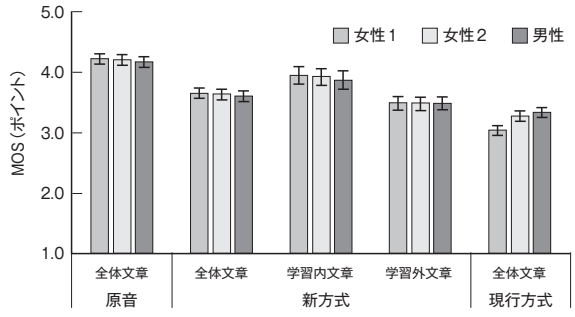

図5．音質評価結果 — どの話者についても，現行方式よりも新方式のほうが，音質が改善した。
Results of speech quality evaluation using mean opinion score (MOS)

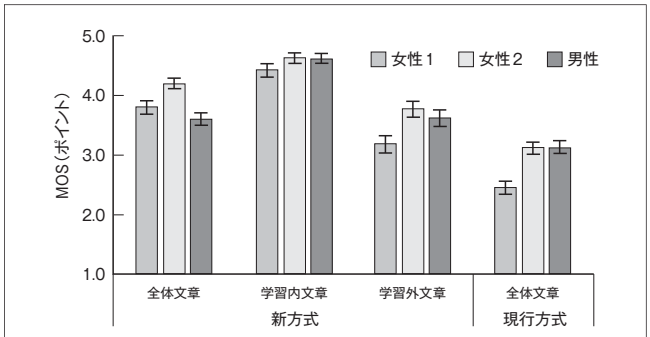

図6．類似性評価結果 — どの話者についても，現行方式に比べて新方式のほうが，声の再現性が向上した。
Results of speaker similarity evaluation using MOS

7　あとがき

次世代音声合成技術として検討した，統計的パラメータ選択に基づく新しい音声合成方式について述べた。評価試験により，現行方式からの音質改善とともに，声の再現性が大きく向上することを確認した。

今後の課題としては，実用性の更なる向上が挙げられる。計算量やメモリ量を削減し，サーバ上で運用できることを目標に開発を進めている。製品化とともに，学習内外文章の音質差の軽減，多言語対応，更には多様性への拡張も進めていく。

文　献

(1) 布目光生 他．DAISYコンテンツ作成のための音訳支援システム：Daisy-Rings™の実装と予備評価．信学技報．**113**，366，2013，p.135－140.
(2) 森田眞弘 他．多様な声や感情を豊かに表現できる音声合成技術．東芝レビュー．**68**，9，2013，p.10－13.
(3) Tokuda, K. et al. Speech Synthesis Based on Hidden Markov Models. Proceedings of the IEEE. **101**, 5, 2013, p.1234－1252.
(4) 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．2011，3-7-7.
(5) 田村正統 他．高品質音声合成のためのピッチ同期帯域群遅延ボコーダ．信学技報．**115**，392，2016，p.33－38.
(6) Ling, Z-H. et al. "The USTC System for Blizzard Challenge 2012", The Blizzard Challenge 2012 Workshop. Portland, OR, 2012-09, ISCA. 2012. <http://www.festvox.org/blizzard/bc2012/USTC_Blizzard2012.pdf>, (accessed 2016-08-08).
(7) 国立国語研究所．“概要 現代日本語書き言葉均衡コーパス（BCCWJ）”．<http://pj.ninjal.ac.jp/corpus_center/bccwj/>，（参照2016-08-08）.

田村　正統　TAMURA Masatsune, D.Eng.
技術統括部 研究開発センターを経て，インダストリアルICTソリューション社 商品統括部 メディアインテリジェンス商品推進部参事，博士（工学）。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。
Product and Service Marketing Div.