話者の声の特徴を直感的な言葉で制御できる音声合成技術

Text-to-Speech Technology to Control Speaker Individuality with Intuitive Expressions

大谷　大和　　森　紘一郎
■OHTANI Yamato　■MORI Koichiro

書籍朗読や，音声広告，エンターテインメントなど，テキスト音声合成の適用分野が拡大するのに伴い，コンテンツに合わせて話者の声の特徴（話者性）を制御できるような音声合成技術へのニーズが高まってきている。

東芝は，このようなニーズに応えるため，“性別”や，“年齢”，“声の明るさ”など，人が聞いたときに感じる声の知覚的特徴を表した言葉により話者性を制御することができ，また“かわいい声”や，“渋い声”といった声の印象を表す言葉からその印象に合った話者性の合成音声を生成できる技術を開発した。これにより，ユーザーがコンテンツに適した話者性を持つ合成音声を手軽に作成できる。

With the expansion of the field of text-to-speech (TTS) applications, including e-book reading, speech advertisements, and entertainment, demand has recently arisen for TTS technologies capable of controlling and generating speaker individuality according to the characteristics of the contents.

In response to these diversifying needs, Toshiba has developed a novel TTS technology that can not only control speaker individuality with perception expressions that represent voice characteristics such as gender, age, brightness of voice, and so on, but also generate synthetic speech that creates a certain effect in accordance with impression expressions such as cuteness, refinement, and so on. This technology allows users to produce synthetic speech with the desired voice characteristics matching the contents.

1　まえがき

音声合成技術とは，文字や文章などのテキスト情報を音声に変換する技術である。これまでの品質の向上により，テキスト情報を音声で伝える点では十分な音質が実現しており，音声案内だけでなく，スマートフォンなどのモバイル機器のアプリケーションや，Webを介したクラウドサービス，音声広告，教育向けコンテンツなど，近年では様々な場面で広く音声合成技術が利用されている。

このような音声合成技術の利用範囲の拡大に伴い，音声の多様性に対するニーズが高まっている。例えば，有名人や近しい人の合成音声，英語や中国語など様々な言語による発声，対話口調や宣伝口調などの発話スタイル，感情豊かな書籍の朗読，コンテンツに適した話者による合成音声といったニーズがある。

東芝は，これらのニーズに応えるため，音声合成の多様性向上を目的に技術開発を進めている。これまでに，特定の人による少量の発話からその人の声色や口調によく似た合成音声を作ることができる日・米・中カスタム音声合成辞書作成システム[(1)]や，合成音声の韻律編集システム[(2)]，通常発声の音声から感情付き合成音声を生成する技術[(3)]を開発してきた。

今回，更なる多様性向上を目指し，声の特徴（話者性）を制御できる音声合成技術を開発した。この技術は，合成音声の声色や口調を“明るい”や“女性っぽい”といった話者性を表す言葉（知覚表現語）により制御できる。また，声の全体的な印象を表す言葉（印象表現語）を指定することで，その印象に合った話者性の合成音声を生成できる。この技術により，コンテンツに適した話者性を持つ合成音声を容易に生成できるため，音声合成の多様性が広がると考えられる。

2　知覚表現語及び印象表現語の選定

知覚表現語や印象表現語による話者性制御では，用いる言葉がシステムの性能や使い勝手に影響すると考えられる。そこで，最初に話者性の制御に用いる知覚表現語や印象表現語を選定した。

各表現語の選定では，一般ユーザーが声の特徴や印象をどのような言葉で表現するのかを調べるために，クラウドソーシングシステム(注1)によるアンケート調査を実施した。

アンケート調査では，**図1**に示すアンケート用のユーザーインタフェース（UI）を用いて，ある話者の音声を提示し，その音声に対してどのような印象が当てはまるかを印象表現語の一覧から一つ選んで投票してもらった。また，提示した話者の音声にどのような話者性が含まれているかを知覚表現語の一覧から全て選んでもらった。その際，図1の各表現語は声質に関する文献[(4), (5)]を参考に事前に選んだものを用いている。

（注1）　不特定多数の人を募集し，必要なサービスや，アイデア，コンテンツなどを得るシステム。

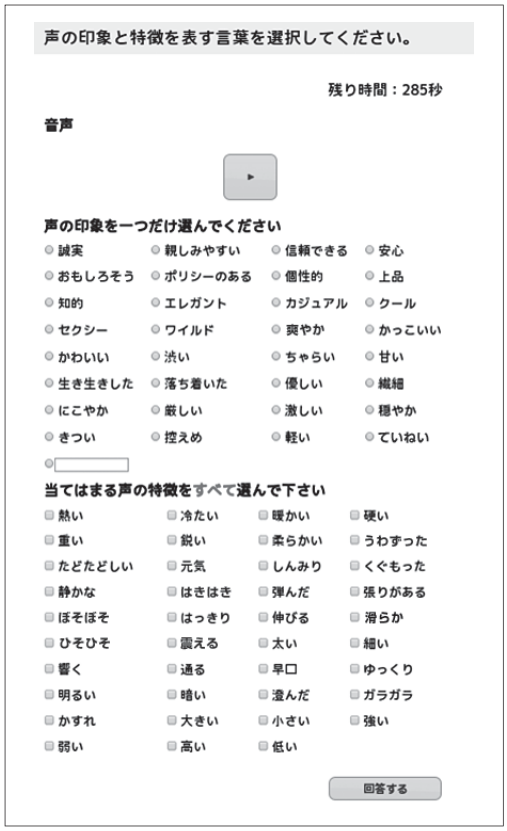


図1．知覚表現語と印象表現語のアンケート用UI — 上段が印象表現語に，下段が知覚表現語に関するアンケート項目である。
Questionnaire items regarding perception and impression expressions

表1．選定した知覚表現語と印象表現語の一覧
List of selected perception and impression expressions

知覚表現語	印象表現語
明るさ （明るい，暗い）	かわいい
硬さ （硬い，柔らかい）	知的
明瞭さ （明瞭な，くぐもった）	ていねい
流ちょうさ（流ちょうな，たどたどしい）	穏やか
透明さ （澄んだ，かすれた）	爽やか
	誠実
	落ち着いた
	クール
	渋い

アンケートでは通常発声や感情発声など30名の話者による合計153種類の音声を用い，各音声サンプルに対して延べ100名の評価者にアンケートを行った。

アンケート集計後，選ばれた知覚表現語及び印象表現語の頻度を求め，それらを相関分析により解析した。その結果，**表1**に示すような知覚表現語の対と印象表現語を選定した。更に話者性制御では，これらの表現語に加え，性別（男性，女性）と年齢（若い，老いた）を知覚表現語として用いている。

3　知覚表現語及び印象表現語による話者性制御法

前章で選定した知覚表現語と印象表現語を用いて所望の話者性を持つ合成音声を生成するには，声色を表すスペクトル包絡や音高を表す基本周波数などの物理特徴量（音響特徴量）が，知覚表現語や印象表現語とどのような関係があるのかをモデル化する必要がある。これらの関係を表すため考案した，話者性に関する3層構造のモデルを**図2**に示す。この3層モデルでは，声の明るさや年齢などの知覚表現語はスペクトル包絡などの音響特徴量の組合せで表現され，また，“かわいい”や“渋い”などの印象表現語は知覚表現語の組合せにより表現されると定義する。この3層モデルに基づき，知覚表現語から音響特徴量を制御する知覚表現語モデルと，指定した印象表現語に基づき知覚表現語の組合せを決定する印象表現語モデルを構築し，それらを用いて話者性制御を実現する。

知覚表現語モデル及び印象表現語モデルの構築では，数十名分の学習用話者の音声データを用いるが，学習用話者の音響特徴量に含まれる知覚表現語が表す声の成分は話者ごとに異なっているため，学習用話者ごとにどの知覚表現語がどの程度含まれているかを表す得点（知覚表現語得点）が必要となる。ここでは，知覚表現語得点の付与方法について述べた後，各モデルの構造と構築方法及び話者性制御方法について述べる。

3.1　知覚表現語得点の付与方法

各学習用話者の知覚表現語得点は，クラウドソーシングシステムを用いて人手で得点化する。得点化では，10名の評価者に対して**図3**に示すUIを利用し，基準音声と比較対象の音声を順番に聴いた後，基準音声から比較対象の音声がどのくらい異なるのかを得点付けしてもらう。これらの平均値を知覚表現語得点として用いる。

このとき用いる基準音声は，平均声モデル[6]と呼ばれる平均

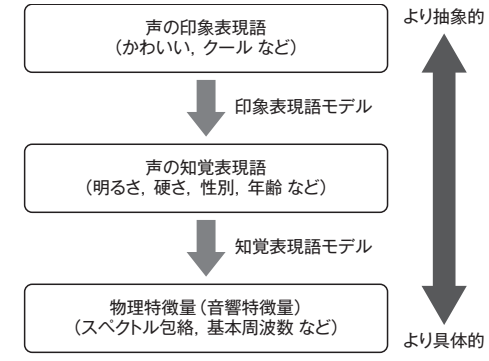


図2．話者性制御のための3層モデル — 印象表現語で知覚表現語を制御し，知覚表現語で声の物理特徴量を制御する。
Three-layered model for controlling speaker individuality

一般論文

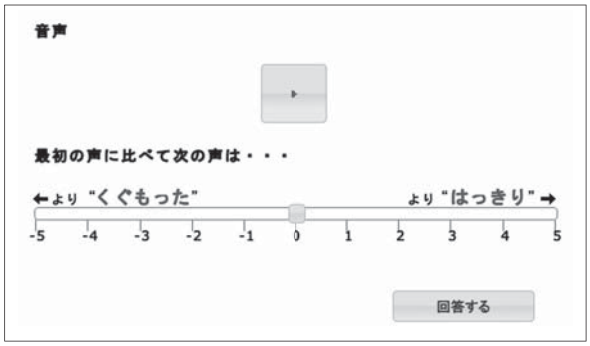


図3. 知覚表現語の得点化用UIの例 ― 基準音声と比較対象の音声を聴いてもらった後，対象の知覚表現語対に対して-5から+5の間で得点を付けてもらう。

Display for scoring of perception expressions

的な話者性を持つモデルから合成した音声を用いた。また，知覚表現語得点は音響特徴量ごとに付与している。

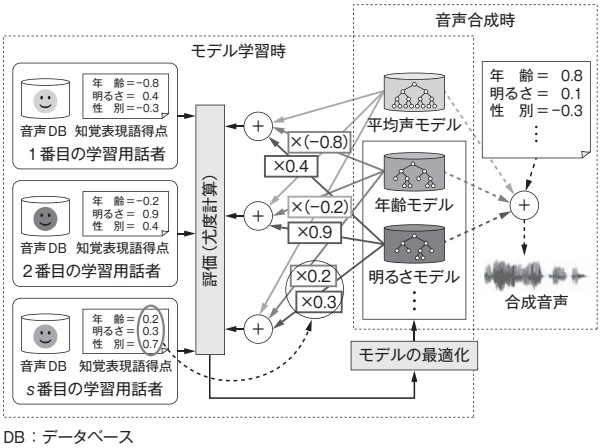
3.2 知覚表現語モデルの構造と構築方法

知覚表現語モデルの概要を**図4**に示す。知覚表現語モデルは“年齢モデル”や“明るさモデル”といった知覚表現語それぞれに対応した話者性を表すモデルである。ある話者性を示すモデル$\tilde{\theta}$は知覚表現語モデルと平均声モデルθ_{AVM}を用いて式(1)のように表せる。

$$\tilde{\theta}=\sum_{c=1}^{C} w_c\,\theta_c+\theta_{\mathrm{AVM}} \tag{1}$$

ここで，θ_c及びw_cはc番目の知覚表現語モデルと知覚表現語得点である。

モデル学習では，まず全学習用話者の尤度（ゆうど：モデルに対するデータの適合度合い）$L(\theta)$を式(2)により計算する。



DB：データベース

図4. 知覚表現語モデルの概要 ― 学習により各知覚表現語に対応したモデルを構築し，合成時はそれらのモデルを各知覚表現語得点に基づいて重み付けし，音声を合成する。

Outline of perception expression model

$$L(\theta)=\prod_{s=1}^{S}\prod_{t=1}^{Ts} P(O_s(t)\,|\,\tilde{\theta}_s) \tag{2}$$

ここで，$O_s(t)$は時刻tにおけるs番目の学習用話者の音響特徴量，$\tilde{\theta}_s$は式(1)で計算されるs番目の学習用話者の話者性を表すモデルである。$P(O_s(t)\ |\tilde{\theta}_s)$は$\tilde{\theta}_s$に対する$O_s(t)$の条件付き確率である。

次に，計算した尤度が最大となるように全ての知覚表現語モデルを同時に更新する。この処理を繰り返し，知覚表現語モデルを最適化する。

3.3 印象表現語モデルの構造と構築方法

印象表現語モデル$\lambda(i)$は，式(3)に示す混合正規分布モデル（GMM）によりモデル化する。

$$P(\boldsymbol{w}|\lambda(i))=\sum_{m=1}^{M} a_m(i)\,N(\boldsymbol{w};\boldsymbol{\mu}_m(i),\boldsymbol{\Sigma}_m(i)) \tag{3}$$

ここで，$N(\cdot)$は正規分布を表し，$a_m(i)$，$\boldsymbol{\mu}_m(i)$及び$\boldsymbol{\Sigma}_m(i)$はそれぞれm番目の分布の重み，平均ベクトル及び全共分散行列で，iは印象表現語モデルのインデックスである。また，$\boldsymbol{w}$は知覚表現語得点のベクトルで，$P(\boldsymbol{w}\,|\,\lambda(i))$は印象表現語モデル$\lambda(i)$に対する$\boldsymbol{w}$の条件付き確率である。

各印象表現語モデルの学習はまず，2章で述べた印象表現語に関するアンケートで対象とした印象表現語への得票数が，10以上の話者を学習用話者として用意する。次に，学習用話者の$\boldsymbol{w}$を用いてGMMを学習する。

3.4 話者性制御方法

この手法の話者性制御では，ユーザーによる手動制御と印象表現語を用いた話者性制御が行える。

手動制御の場合は，ユーザーが設定した知覚表現語得点を式(1)に代入することで話者性を制御し，所望の話者性を持つモデルを構築できる。

印象表現語モデルを用いる場合は，所望の印象表現語モデル$\lambda(i)$からランダムサンプリングを行い，目標の印象を持ったある話者性を示す$\boldsymbol{w}$を生成する。これを式(1)に代入し，話者性制御を行う。

このような話者性制御により得られるモデルを用いて音声合成を行うことで，所望の話者性の合成音声が生成できる。

4 評価実験と応用例

知覚表現語モデル及び印象表現語モデルによる合成音声の品質評価について述べる。また，これらのモデルを用いた話者生成のインタフェースについても述べる。

4.1 知覚表現語モデルの評価

2章のアンケートで評価した中の25名の話者を用いて知覚表現語モデルを構築した。学習に含まれない12文のテキスト

を用いてランダムに生成した10種類の合成音声の5段階平均オピニオンスコア（1：悪い～5：良い）による音質の評価と，知覚表現語モデルを一つずつ制御した際の変化した話者性を識別するテストを行った。

評価の結果，音質の平均値は2.87，最低で2.01，最高で3.39となった。評価のばらつきが大きいのは，主に声色に依存するところが大きく，人が出さないような声色の場合に低い点数となっていた。しかし，音声に不快なノイズなどが発生しなかったことから，実用に耐えられる音声が得られた。また，話者性の識別テストでは平均で70％の識別率となり，適切に話者性制御ができることがわかった。

4.2　印象表現語モデルの評価

印象表現語モデルの評価では，表1の九つの印象表現語モデルからランダムサンプリングにより生成した音声を用いて，それらの音質と印象の表出度合いの平均オピニオンスコア（1：印象が出ていない～5：印象がよく出ている）で評価した。

評価の結果，音質の平均が3.54，印象の表出度合いの平均が3.12となった。音質は，当社製品の音声合成システムと同等の品質が確認された。また印象の表出も，実際に聴いて，印象表現語の持つ印象を感じられる程度の品質が得られた。

4.3　話者生成インタフェースの例

知覚表現語と印象表現語を用いた話者性の制御及び音声合成を一般ユーザーにも手軽に行えるようにするためには，GUI（グラフィカルUI）による操作が不可欠である。ここでは，どのようなプラットフォーム上でも利用できるように，**図5**に示すようなWebベースの話者性制御・音声合成インタフェースのデモシステムを構築した。

テキストボックスに任意のテキストを入力した後，印象表現語の一覧から選択した印象表現語を用いて，ランダムサンプリングにより所望の印象を示す知覚表現語得点を生成する。そして，レーダチャート形式の知覚表現語操作Webインタフェースを使って話者性制御を行い，所望の話者性を持つ合成音声を生成する。

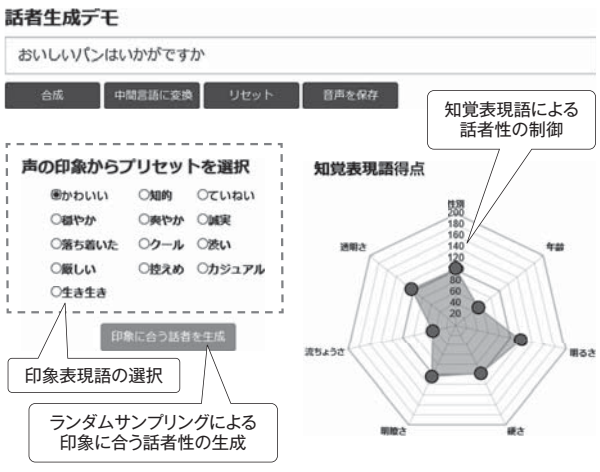


図5．Webベースの話者性制御・音声合成インタフェースの例 — 声の印象表現語を選択し，知覚表現語により制御することで所望の話者性を持つ合成音声が得られる。

Example of web-based interface display for speaker generation demonstration system

5　あとがき

当社は，様々な話者性を自由に制御できる音声合成技術を開発した。話者性制御では声の知覚的特徴を表す言葉により，ユーザーが直感的に操作することができる。また，声の印象を表す言葉から最適な知覚的特徴の組合せを算出し，印象に合った話者性の音声合成を容易に生成できる。

今後は，このシステムを更に発展させて，話者性だけでなく，感情や様々な口調を持つ合成音声を手軽に生成できるようにし，様々な音声合成技術を用いたコンテンツが容易に制作できる環境をユーザーに提供できるように技術開発を進めていく。

文　献

(1) 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.
(2) 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.
(3) 大谷大和 他. “平静音声から予測した感情付与モデルに基づく統計的感情音声合成”. 日本音響学会2015年秋季研究発表会講演論文集. 会津若松, 2015-09, 日本音響学会. 2-1-12.
(4) 木戸 博 他. 通常発話の声質に関連した日常表現語：聴取評価による抽出. 日本音響学会誌. **57**, 5, 2001, p.337－344.
(5) 高椋琴美 他. “声の印象を表現する単語による認知構造モデルの検討”. 日本音響学会2014年春季研究発表会講演論文集. 東京, 2014-03, 日本音響学会. 2-Q5-2.
(6) Yamagishi, J.; Kobayashi, T. Average-Voice-Based Speech Synthesis Using HSMM-Based Speaker Adaptation and Adaptive Training. IEICE Trans. Inf. & Syst. **E90-D**, 2, 2007. p.533－543.

大谷　大和　OHTANI Yamato, D.Eng.
技術統括部 研究開発センター 知識メディアラボラトリー研究主務，博士（工学）。音声合成に関する研究・開発に従事。日本音響学会，電子情報通信学会，情報処理学会，ISCA，IEEE会員。
Knowledge Media Lab.

森　紘一郎　MORI Koichiro
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主務。音声合成に関する研究・開発に従事。日本音響学会，人工知能学会会員。
Product and Service Marketing Div.