手軽に様々な声を使って個性を演出できる音声合成ミドルウェア ToSpeak G2

ToSpeak G2 Text-to-Speech Middleware Offering Individuality to Application Systems through Quick Voice Production Technology

瀬戸　重宣
■SETO Shigenobu

東芝は，ユーザーが新たな声を手軽に追加して利用できる音声合成ミドルウェアToSpeak G2を開発した。ToSpeak G2で用いる声の辞書は数Mバイト規模であり，インターネットを介した授受や，メモリ制約の厳しい組込みシステムへの搭載も容易である。声の辞書の短時間作成技術によるカスタム声の作成サービスを使えば，数十分の収録音声から数時間で試聴用のカスタム声が作成できる。また，既存の声の組合せやパラメータ設定によっても新たな声を作成できる。ToSpeak G2は，情報を声で伝えるという基本機能にとどまらず，製品やサービスのイメージキャラクターである著名人の声を発して個性を演出したり，特徴のある声でメッセージに雰囲気を添えたりすることが可能である。

Toshiba has developed ToSpeak G2, a new text-to-speech (TTS) middleware that can provide users with a wide choice of voices in producing speech contents. ToSpeak G2 features a quick voice production system, and the output voice data of only several megabytes in size can be easily sent via the Internet and installed in an embedded system with a small memory footprint. Using this quick voice production system, a trial version of a new voice can be built within several hours from speech data recorded for several tens of minutes. A new voice can also be created by a combination of two different speakers' voice components, or by changing the voice parameter settings. ToSpeak G2 makes it possible to offer individuality to various application systems with a wide variety of voices.

1　まえがき

近年，音声合成技術は，大幅な音質の向上が図られてきた[1]。種々の製品やサービスに組み込まれて音声コンテンツを生成し，情報を声で伝えるという基本機能が活用されている。その一方で，声の選択肢は僅か数種類しか用意されていないことも多く，また，コンテンツによっては冷静な調子の声がかえってあじけない印象を与える場合もある。意図したようなイメージに合う音声コンテンツを作り込みたいときに，その手軽さにはまだ課題がある。

東芝は，新たな声をユーザーが手軽に追加可能にする基本技術[2]をベースとした音声合成ミドルウェア製品ToSpeak G2を新たに開発した。めりはりのある声や，注意を引くような調子の声，魅力的な雰囲気の声など，伝えたいイメージに合うような声を作成し，それを音声コンテンツ作成に利用できる。また，読み上げるテキストの中に埋込みタグを挿入して，話すスピード（以下，話速と記す），抑揚，及びポーズや読みがなどを詳細な単位で指定し，音声コンテンツを作り込むことができる。

ここでは，今回当社が開発した，新しいカスタム声の短時間作成技術と，音声コンテンツの作込みのための埋込みタグの活用技術について述べる。また，著名人の声を素材にする場合に配慮が望まれる声の使用期間などの制約制御機能についても述べる。

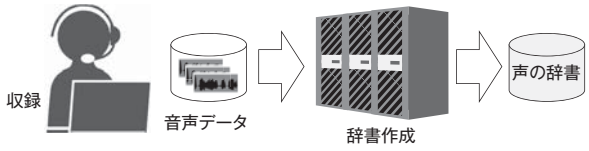

図1．新たな収録から声の辞書作成 ― 数十分の収録音声から数時間程度の計算処理によって新たな声の辞書を作成できる。
Quick voice production from newly recorded voices

2　新しいカスタム声の辞書作成技術

音声コンテンツの制作にどのような声を使用するかは，そのできばえに大きく影響する。ToSpeak G2では，ユーザーが新たな声を追加する際に，以下に述べる二つの手段が使える。

2.1　新たな収録から声の辞書を短時間で作成

一つ目の手段は**図1**に示すように，作成したい話者の発声を収録し，その収録音声から新たな声の辞書を作成する方法である。

音声合成ミドルウェア内部にある音素片や韻律制御の単位の網羅性を考慮した文章セットを，話者が発声した音声データを元にして声の辞書を作成する。収録する音声データの量と作成した声の質はトレードオフになるが，ここでは声の作成が短時間で済む手軽さを優先し，音声データの量を極力減らす文章セットを使うことにした。

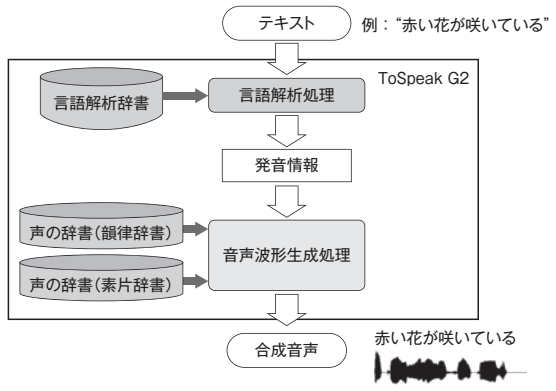

図2．音声合成ミドルウェアToSpeak G2の構成 — 声の辞書は，韻律辞書と素片辞書から構成される。

Configuration of ToSpeak G2 TTS middleware

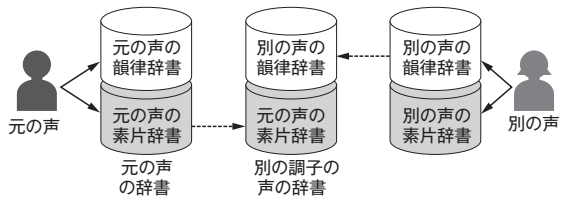

図3．韻律辞書と素片辞書の組替え — 素片辞書を別の声の韻律辞書と組み合わせた例である。

New voice creation by partial exchange of other speaker's component

例えば数十分の収録音声からは，数時間程度の計算処理によって新たな声の辞書が作成可能である。声の辞書のサイズは数Mバイト規模であり，インターネットを介したダウンロードや，アプリケーションへのセットアップも短時間で行うことができる。また，アプリケーション内で声の辞書をメモリにロードする処理も一瞬で済むため，ユーザーは様々な声を軽快に切り替えて利用できる。

当社では，辞書の動作検証を経た製品を10日程度で提供可能な製品化プロセスを構築し，辞書作成サービスに向けて運用を開始している。

2.2 既存の声を活用した新たな声の作成

新たな声を作成するもう一つの手段は，既作成の合成辞書の組合せやパラメータ設定により新たな声を作成する方法である。

声の辞書は**図2**に示すように，抑揚や話速，リズムに関するその話者のしゃべる調子の特徴を反映する韻律辞書と，声の音色や滑舌を反映する素片辞書から構成される。これらの内容を，**図3**に示すように元の声と別の声とで組み替えることで，元の声とは異なる調子の声を新たに作ることができる。これに併せて，全体的な声の高さや話速などのレベルを設定する音声合成ミドルウェアのパラメータも変更すれば，元の声とは大きく異なる印象の声にすることも可能である。

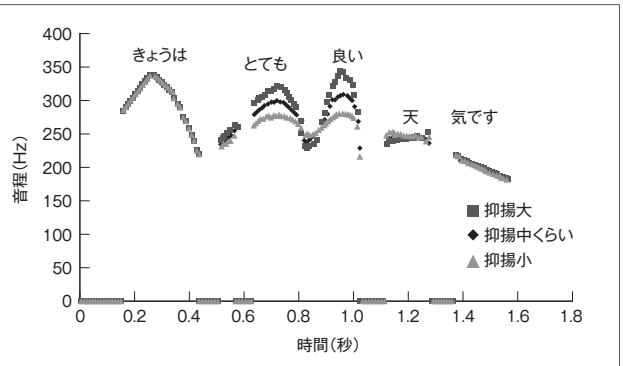

図4．抑揚の大きさを指定した例 — “きょうはとても良い天気です。”というコンテンツの“とても良い”の部分の抑揚の大きさを3段階に変えた例である。

Examples of SSML tag usage setting different prosody range tags

3　イメージに合う音声コンテンツの作込み

同じテキストを読み上げる場合でも，読む人により，あるいはその場面により，読み上げる調子は様々である。音声合成の音質が大幅に向上したといえども，テキスト情報だけから自動生成する合成音声が音声コンテンツの制作者のイメージに最初から合致するとは限らない。既存の音声コンテンツが存在していて期待する読み方のイメージがある場合には，その読み方に合成音声の読上げを近づける仕組みが有効である。

ToSpeak G2では，音声コンテンツの制作者がテキスト中に埋込みタグを挿入することによって細部の調整を可能にする機能を持っている。例えば，ミドルウェアが自動付与する，語の読み方，及びポーズの挿入位置や長さが制作者の意図と異なる場合，所望の読み方やポーズの内容を埋込みタグの書式で直接指定することによって調整できる。また，ミドルウェアが生成した，合成音声の抑揚や話速，リズムについても，所望のテキスト範囲に対して抑揚，話速，及びリズムの内容を同様に埋込みタグの書式で指定して調整できる。

テキスト中の一部に抑揚の設定値を指定した例を**図4**に示す。埋込みタグは，音声合成用の記述言語であるSSML（Speech Synthesis Markup Language）で定義されたフォーマットに従い，**表1**に示すような指定が可能である。

4　声の辞書の制約制御機能

声の主が著名人である場合，声を広く活用するだけでなく，一定の使用条件を設定して，その条件下での活用を促進する仕組みを持つことが望まれる。このため**図5**に示すように，ToSpeak G2では，対象とする声の使用条件に関する制約を設定あるいは変更できる機能を持たせた。声の辞書の使用を所定の期限までに限定したい場合は，その制約情報を声の辞書に設定すると同時に，使用条件の範囲内であれば制限を解

表1. サポートするSSMLタグの例

Examples of supported SSML (Speech Synthesis Markup Language) embedded tags

項　目	タグの利用例
強調	きょうは<emphasis level="moderate">とても良い</emphasis>天気です。
話速	<prosody rate="fast">時間がないので，少し早口でしゃべります。</prosody>
声の高さ	<prosody pitch="high">少し声を高くします.</prosody>
抑揚の大きさ	<prosody range="x-low">抑揚のないロボット声です.</prosody>
音量	ちょっと<prosody volume="loud">大きな声で</prosody>しゃべります。
ポーズ	ここで<break time="2000ms"/>ちょっと長めのポーズを入れてみました。ここで<break strength ="x-strong"/>ちょっと長めのポーズを入れてみました。
読み	文脈なしで，<phoneme ph="ｼｼﾞｮｰﾄ">市場と</phoneme><phoneme ph="ｲ'ﾁﾊﾞ">市場を</phoneme>読み分けることは，人間でもできません。
文境界	<s>私の好きなグループは○○○○。でした。</s>

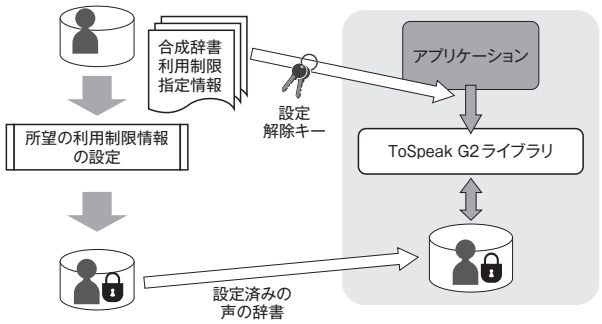

図5. 声の使用条件に関する制約の制御 — 声の使用条件に関する制約を，声の辞書と設定解除キーに埋め込んだ設定情報の照合により制御している。

Availability control of voices

除する設定解除キーを，ミドルウェアを呼び出すアプリケーションに埋め込む。アプリケーションがその声の辞書を使用する際に，設定解除キーと声の辞書に設定された使用条件を確認し，両者を満足すれば声の辞書を使用できる。

製品プロモーションや商業イベントなどのように使用期限を限定したい場合や，使用可能な機器を限定したい場合，声の辞書の使用権を購入した人だけに限定したい場合などにこの機能を使う。設定済みの声の辞書と設定解除キーとに分けているので，これらをそれぞれ異なるルートでユーザーに配布し，ユーザーの手元でこれらの設定内容が合致すれば制約が解除されて使用できるようになる。

前述のような声の辞書ごとの制約制御だけでなく，ミドルウェアの音声合成機能全体に対しても同様に制約制御する機能を持たせている。

(注1)，(注2)　ARM，Cortexは，英国ARM社の商標。
(注3)　Androidは，Google Inc.の商標又は登録商標。

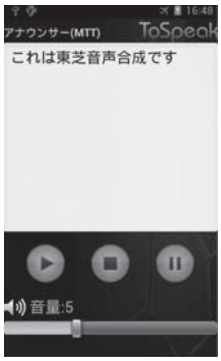

図6. ToSpeak G2 のデモンストレーション アプリケーションの画面例 — Android環境で動作するToSpeak G2のデモンストレーション アプリケーションの画面例である。

Example of demonstration display of ToSpeak G2

5　動作環境

ToSpeak G2は，ARM(注1) Cortex™ (注2)-A9 (動作周波数：1 GHz)クラスの動作環境において実時間動作する。また，同等の処理能力のあるプラットフォームであれば，移植も可能である。一例として，Android™ (注3)環境で動作しているアプリケーションのデモンストレーション画面を**図6**に示す。

6　あとがき

生身の人間の声と聞きまちがえるほどのクオリティの実現を目指すとともに，声の選択肢を増やして表現力を増すことが，音声合成の応用拡大にとって重要であると考える。ToSpeak G2はそのような目指す方向に向けた第一歩であり，今後，より手軽に種々の声を活用でき，実用的な応用を促進するための技術開発を更に進めていく。

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)　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.

瀬戸　重宣　SETO Shigenobu
セミコンダクター＆ストレージ社 システム・ソフトウェア推進センター 企画・管理担当主幹。音声ミドルウェアの開発に従事。電子情報通信学会，日本音響学会会員。
System & Software Solution Center