

テキスト音声合成技術

大塚貴弘*
川島啓吾**
山浦 正**

Text-to-Speech Technology

Takahiro Otsuka, Keigo Kawashima, Tadashi Yamaura

要 旨

三菱電機は家電製品の高度な機能をらくに使いこなして暮らしを楽しむアシスト機能を“らく楽アシスト”と名づけて展開しており、その中核機能の一つが音声による操作ガイダンスである。また、何らかの作業中の人や、同時に多くの人に情報を提供するには、音声は最も有効な情報伝達手段である。この音声を、あらかじめ録音することなしに、テキストから合成する技術が、テキスト音声合成技術であり、高い期待を受けている。しかし、当初は、機械的で単調な音声しか合成できず、応用分野は限定されていた。そこで、当社では、幅広い分野に適用できるよう、音声合成技術の高品位化と、利用しやすい合成ソフトウェア、ツールの開発に取り組んでいる。

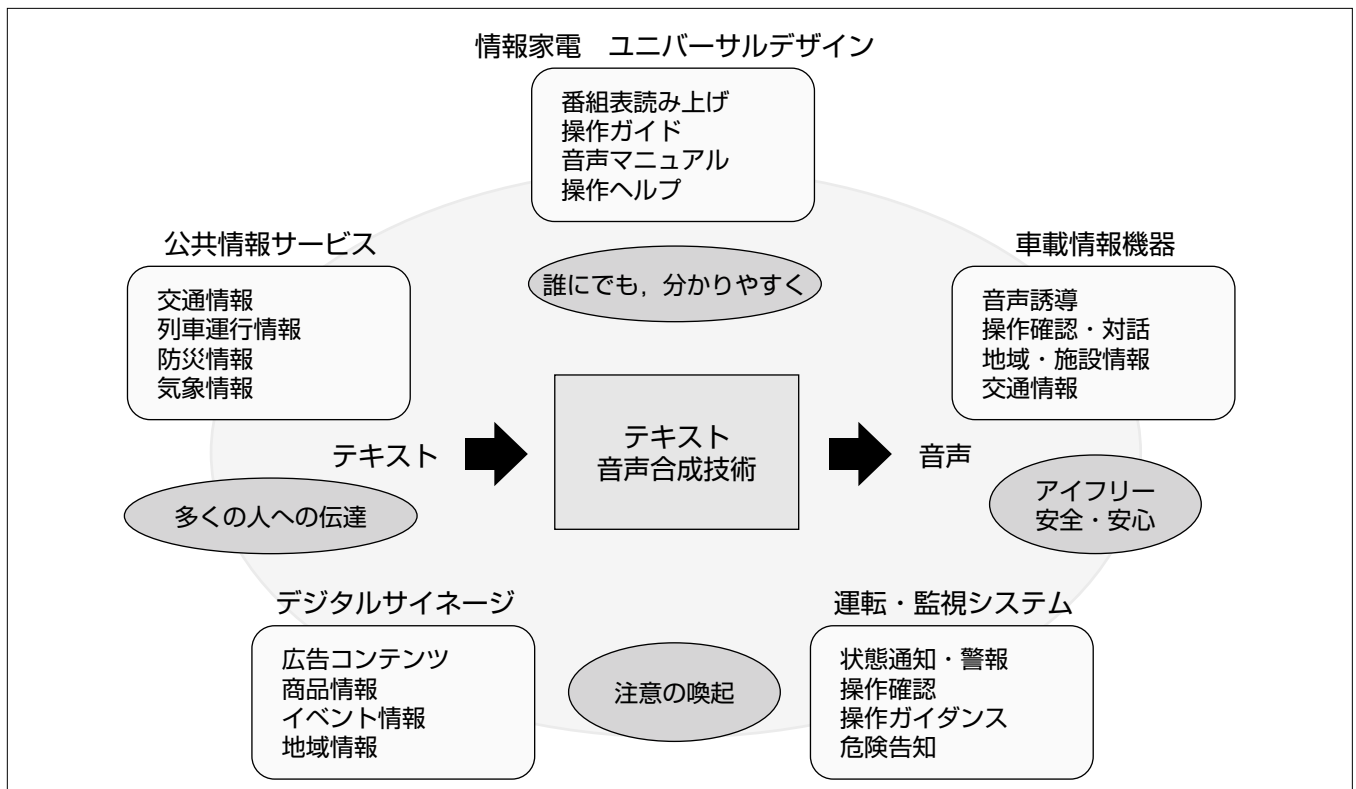
高品位化に対しては、大規模な音声データベースから、イントネーション・リズムの制御規則を学習する方式と、合成に用いる波形辞書をデータ収録時の変動要因を取り除

いて作成する音質安定化手法によって、明瞭(めいりょう)で自然な合成音声を得ている。

さらに、広い応用分野に適用できるよう、次の特長を備えている。

- (1) 漢字仮名交じり文のほか、ITS(Intelligent Transport Systems)車載器用音声合成記号(JEITA((社)電子情報技術産業協会)TT-6004)などの中間言語からも合成可能
- (2) 適用するハードウェアスペックに応じて、適切な処理量・メモリ量で構成することが可能
- (3) 音声作成補助ツールによって、簡単な操作で所望の抑揚、発話速度に調整可能

このテキスト音声合成技術は、AV(Audio Visual)家電やカーナビゲーションシステムをはじめとして、様々な応用分野への適用が進んでいる。



テキスト音声合成技術の応用分野

テキスト音声合成は、入力されたテキストから音声を自動合成する技術であり、音声を収録する必要がないのでコストがかからず、伝達したい内容を即座に分かりやすく、作業中の人や多数の人にも提供ができる。高機能化する家電の使い方のガイダンスや、運転中の人に対する情報提供等に適用されており、今後さらに、運転監視、公共情報サービス、デジタルサイネージ等へ適用が期待されている。

1. ま え が き

音声による情報伝達は、分かりやすいだけではなく、作業中の人や多数の人への情報伝達が可能であり、他に代替しえない情報伝達手段である。任意のテキストから音声を合成するテキスト音声合成技術⁽¹⁾は、録音することなく、自由に音声を合成できる技術として古くから大きな期待を受けてきた。しかし、当初の合成音声は、内容を理解できる了解性はあったものの、肉声感のない機械的音質で、韻律(イントネーション、リズム)も単調なため、応用分野は限定されていた。

当社では、合成音声の高品位化に加え、種々の用途に適用しやすい合成ソフトウェアとツールの開発に取り組んできた。その結果、家電製品をらくに使いこなす当社の“らく楽アシスト”の中核機能の一つである“音声ガイド”に適用されるなど、現在、幅広く応用が進みつつある。本稿では、テキスト音声合成方式の概要を解説し、その特長を述べる。

2. テキスト音声合成方式

2.1 概 要

図1に、このテキスト音声合成方式の構成を示す。音声合成するテキスト(漢字仮名交じり文)が入力されると、言語辞書を参照し、テキストの解析を行い、読み、アクセント位置、品詞等を決定し、中間言語と呼ばれる表現に変換する。

イントネーション・リズム制御では、音声のイントネーションを表すピッチ周期(声の高さ)系列と、リズムを表す各音素の継続時間長を生成する。それらの情報をもとに、波形辞書中の音片(短い合成単位)データを読み出し信号処理によって要求されたピッチ周期、継続時間になるように加工し、接続することで音声波形を生成、出力する。

この方式では、中間言語も直接入力することができる構成となっており、ITS(高度道路交通システム)で規定され

ているJEITA TT-6004 ITS車載器用音声合成記号にも対応している。また、中間言語のみを扱う、テキスト解析部を持たない簡易型の合成ソフトウェアを端末機器で動作させ、テキスト解析をサーバで動作させる分散構成も可能である。さらに、読み、アクセント型、ポーズ位置、ポーズ長等を中間言語上で直接指定することで、所望の合成音声を得ることもできる。

2.2 テキストの解析

テキスト解析部では最初にプリプロセッサを適用する。これは、略語の処理などを行うもので、例えば、“10/9”を“10月9日”又は“9分の10”などへ変換する。種々の規則が用意されており、用途に応じた変換規則が適用できる。

次に、読み・アクセント解析を行う。合成対象であるテキストが入力されると、言語辞書を参照し、候補となる形態素を抽出したのち、接続コスト最小法に基づいて最適形態素列の選択を行う。形態素の接続しやすさを表す接続コストを最小となるように形態素を選択することで、文法の誤りや、通常、解析が困難な話し言葉表現に対しても適切な解析が可能となっている。

次に、選択した形態素列に対し、複合化処理(形態素の結合とアクセントの移動)、前後の音韻環境を考慮した無声化処理を行い、読みやアクセント位置を示す中間言語を出力する。

言語辞書の基本部分は、大量のテキストデータから抽出した分野に依存しない出現頻度の高い単語で構成している。さらに、適用製品で使用するテキストを用いて分野特有の単語を追加することもできる(例えば、交通情報を想定した地名データなど)。言語辞書には、表記・読み・品詞・アクセント位置等が含まれる。

2.3 イントネーション・リズム制御

イントネーション・リズム制御では、統計的予測モデルによって、音声のイントネーション(ピッチ周期系列)とリズム(音素の継続時間長)を決定する。まず、テキスト解析結果又は中間言語から、予測を行うための要因(音節の文

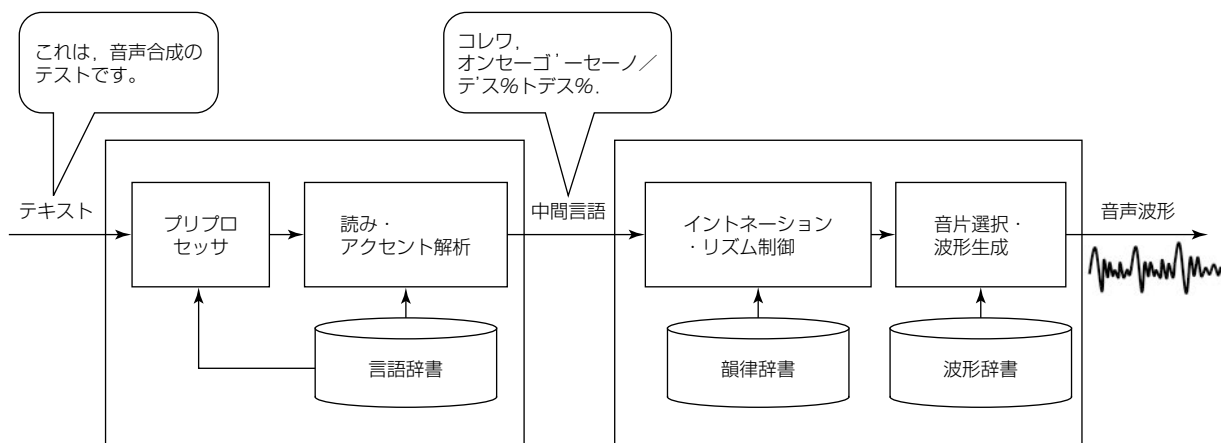


図1. テキスト音声合成方式の全体構成

内位置、フレーズの音節数、アクセント型、音素名等)を抽出する。これらの要因を入力とし、予測モデルを用いてイントネーションとリズムを決定する。

予測モデルで用いる要因に対するパラメータセットをここでは韻律辞書と呼び、学習用音声コーパス(音声と言語情報等を体系化したデータベース)から機械学習によって求める(図2)。一般には予測値とコーパスの観測値の平均誤差を最小とるように学習を行うが、このとき、全観測値に対する平均誤差最小を基準とすると、考慮していない要因による変動を平均化してしまい、合成音声の自然性が損なわれるという問題があった。そこで、学習過程で、要因に対する固有の変動を持つ観測値を選択し、それに対する誤差を最小化する学習方式を開発した。この方法によって大量コーパスの学習でも自然性が維持できる。

2.4 音片選択・波形生成

波形辞書には、音声コーパスから抽出した種々の音声素片が記憶されており、音片選択では、合成する音素の継続時間長やピッチ周期系列に近く、前後の音片データとのつながりの良い音片データを選択する。

波形生成では、合成する音素の継続時間長やピッチ周期系列に合わせるよう、選択された音片データを信号処理によって変形・接続して音声波形を作成する。

波形辞書は、ナレータが発声した大量の音声コーパスから、学習によって最適な区間を選択し、音素波形を切り出して構築する。一般にデータが大量であれば、前後の音の種類、声の高さ、継続時間による変化のパターンが増え、有効な学習によって品質が向上する。しかし、収録ごとの声質が変化してしまうことも多く、必要のない変動データを学習してしまうなどの問題が生じ合成音声が不安定なものになるという問題があった。そこで、同じ特徴を持つと

考えられる音から収録ごとの変動成分を予測し、変動成分を引き去ることで、音質の均一化を行う音質安定化技術を開発した。

2.5 動作ハードウェアのスケーラビリティ

処理速度及びメモリサイズは、使用するハードウェアに応じて、スケーラブルな構成をとることができる。処理能力の低いハードウェアで使用する場合は、波形辞書からの音片データの選択時に、予備選択方法を導入することで音質劣化を最小限に抑えながら演算量の低減が可能である。(200MHzのCPU(Central Processing Unit)で使用率25%未満)

組み込み向け機器で省メモリサイズが求められる場合には、音片データ数を絞り込むことで波形辞書のサイズを削減でき、最小構成では約7MBのメモリで実装可能である。さらに、入力を中間言語に限定し、かつ波形辞書を外部ディスクに格納することができれば約3MBまで低減できる。

一方、サーバ向け機器でメモリや処理能力の制約が緩い場合には、音片データのバリエーションを増やすとともに、波形辞書からの音片データ選択を高精度に行うことで、最高の音質を得ることができる。

2.6 評価

開発技術(統計的予測モデルによるイントネーション・リズム制御、及び独自の音質安定化技術)の効果を確認するために、5段階MOS(Mean Opinion Score:複数の評定者による5段階主観評価値の平均)による主観評価試験を行った。評定者8名で、提示された合成音声に対して、情報案内音声として5段階(5:非常に良い、4:良い、3:普通、2:悪い、1:非常に悪い)で評価した。合成音声は、単語及び文の24サンプルを用いた。

組み込み向け機器を想定し、使用メモリ7MBという条件で方式を構成した。開発した技術を適用前の合成音声(旧版)と適用後の合成音声(新版)とを比較した。図3に示す結果から、旧版と比較して1点以上の大きな改善効果が見える。

3. 音声作成補助ツール

高い肉声感と了解性が求められる公共放送、カーナビゲーション音声、情報家電の操作ヘルプ音声等では、編集音

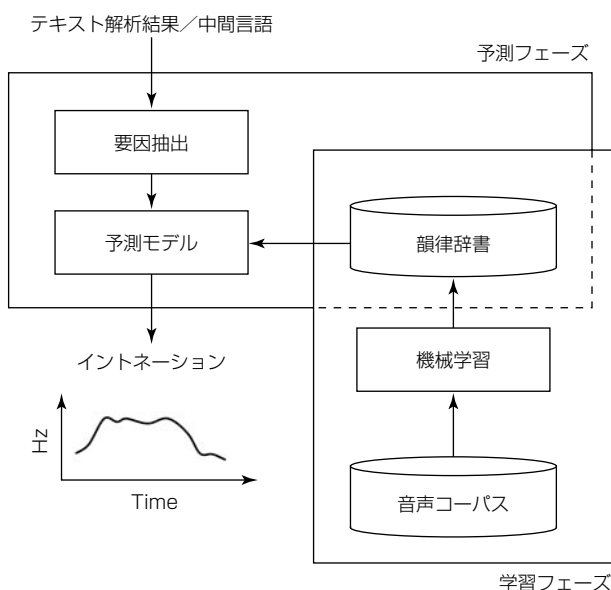


図2. イントネーション制御

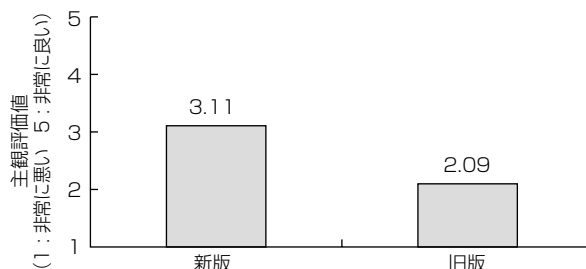


図3. 主観評価結果



図 4. 音声作成補助ツールのメイン画面



図 5. 音声作成補助ツールのサブ画面

声合成方式や録音再生方式が用いられることが多い。しかし、これらの方法は、音声収録・整備にコストがかかること、一貫した音色で多数の音声を収録することが困難であること、後日の音声データの追加の際に同一ナレータを確保できないリスク等の課題があり、任意のテキストから音声を即時作成できるテキスト音声合成方式への期待が高い。

そこで、製品開発者が、ナレータ音声の代わりとなる、より高品位な合成音声を作成するために、次の機能を持つ音声作成補助ツール(図 4, 図 5)を試作した。

- ①任意のテキストから合成音声を生成できる。
- ②合成音声のアクセント位置などを中間言語で調整できる。
- ③合成音声全体の、声の高さ、話速、声の大きさ、抑揚の大きさを調整できる。
- ④合成音声の部分的な声の高さや話速を調整できる。

このツールによって、操作に慣れれば、録音音声に匹敵する韻律特徴を持った合成音声を実現できることを確認した。今後、さらなる操作性と音質の改良を行い、より高い合成音声品質が要求される分野に適用を図る予定である。

4. む す び

当社で開発したテキスト音声合成方式と音声作成補助ツールについて述べた。テキスト音声合成方式は、様々な製品への適用に高い期待を受けている。車載機器・情報機器の応答音声⁽²⁾、プラント運転・監視システムにおける機器の誤操作を防ぐためのガイダンス音声や状態通知音声、視覚障がい者のサポート⁽³⁾、複雑になる家電製品の“らく楽アシスト”のさらなる展開等である。今後も、音質の向上に加え、音声作成補助ツールによる音声コンテンツ作成など、利用用途の拡大に努める。

参 考 文 献

- (1) 藤井洋一, ほか: テキスト音声合成技術, 三菱電機技報, **76**, No. 8, 507~510 (2002)
- (2) 成田知宏, ほか: カーナビゲーション向け音声認識の新たな展開, 三菱電機技報, **84**, No. 9, 495~498 (2010)
- (3) 川勝かがり: デジタルテレビのユニバーサルデザイン, 三菱電機技報, **85**, No. 3, 167~170 (2011)