特集論文

テキスト音声合成技術

大塚貴弘* 山浦 正**
川島啓吾**
古田 訓**

Text-to-Speech Technology

Takahiro Otsuka, Keigo Kawashima, Satoru Furuta, Tadashi Yamaura

要 旨

テキスト音声合成は，入力された任意のテキストから音声を自動生成する技術である。音声による情報提供は，伝達したい内容をタイムリーに分かりやすく伝えられるばかりではなく，注意を喚起できるため，作業中の人にも，さらには同時に多くの人にも，確実に情報を伝えることができる。三菱電機では，テキスト音声合成技術を幅広い分野で活用するため，合成音声の高音質化を進めるとともに，様々な用途での要求に応えるため，合成音声の機能性(明瞭化，多様化)の向上に取り組んでいる。

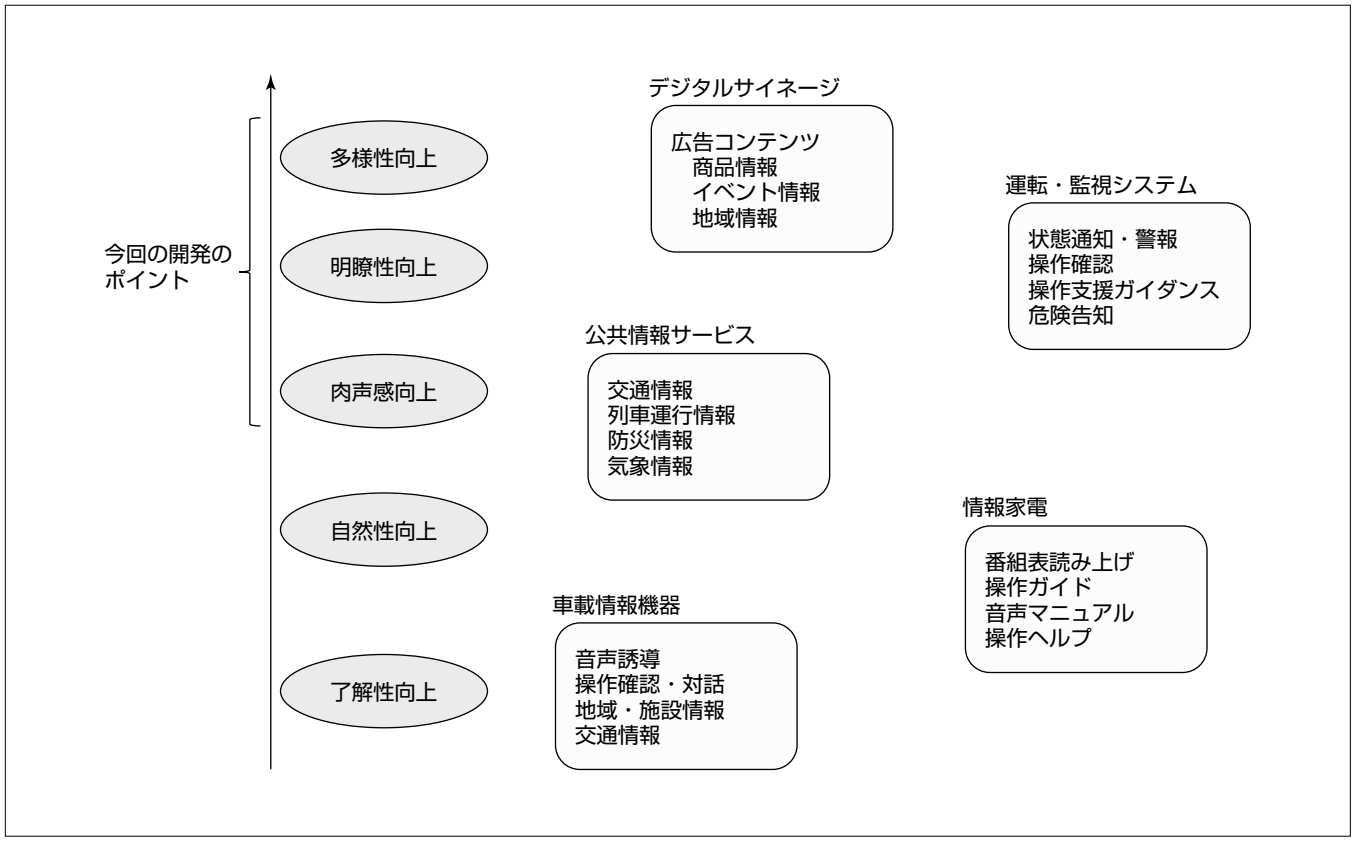
開発したテキスト音声合成技術は，人間が発声した大規模な音声データベース(音声コーパス)を用いたイントネーション・リズム制御と，音片選択とその接続による波形生成によって合成音声を得る方式で，次の特長を備える。

(1) 公共交通機関における放送音声で求められる人間らしい肉声感の高い音声(高音質)

(2) 子音特徴を強調することで高齢者でも聞きやすい音声(明瞭性)

(3) 注意喚起の際に聞き逃しを防止する緊張感のある音声など，用途に応じて特徴を制御できる音声(多様性)

このテキスト音声合成技術は，カーナビゲーションシステム，AV(Audio Visual)家電，公共放送を始めとし，様々な分野・製品に適用が進んでおり，マンマシンインタフェースの不可欠な要素技術となっている。



テキスト音声合成技術の進展と当社製品への適用

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生成する技術であり，音声を収録する必要がないのでコストがかからず，情報をタイムリーに分かりやすく，作業中の人や多数の人にも提供できるシステムを実現できる。高機能化する家電の使い方のガイダンスや，運転中の人への情報提供などに適用されており，高音質化，明瞭化，多様化を進めることで，デジタルサイネージ，運転・監視システム等への適用が期待されている。

*情報技術総合研究所(工博) **同研究所

1. まえがき

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生成する技術である。音声による情報提供は，伝達したい内容をタイムリーに分かりやすく，確実に，多数の人にも提供できるため，高機能化する機器・システムの操作方法，点検業務のガイダンスや，作業者への支援，公共施設における放送等で要求が高い。これらの領域では人間らしい肉声感の高い高音質な合成音声が求められるとともに，高齢者でも聞き取りやすい合成音声，注意喚起するなど状況に応じた音声の特徴が制御可能な合成音声も求められる。

本稿では，これらの要求に応えるために開発したテキスト音声合成方式での，高音質化(**2. 2節**)，明瞭化(**2. 3節**)，多様化(**2. 4節**)を実現する技術について述べる。

2. テキスト音声合成方式

2.1 概　要

図1に，テキスト音声合成方式の構成を示す。テキスト音声合成方式は，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，言語辞書を参照して，テキストの解析(読み・アクセント解析)を行い，テキストに対応した読み(音素)，アクセント位置，品詞等を決定する。これらの音声合成のために必要な情報は中間言語と呼ばれ，アプリケーションによっては，この中間言語を直接専門家が記述・編集し，合成音声を得ることも可能である。

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に対応する音声の基本周波数(声の高低)と，発話速度とそのリズムに相当する音素の継続長のパターンを生成する。それらの情報をもとに，音片選択・波形生成では，人間が発声した音声から切り出された子音や母音に対応する短い波形である音片を，波形辞書から選択し，接続することで，音声波形を生成，出力する。

2.2 ナレーター音声に迫る高音質な合成音声

従来のテキスト音声合成方式[(1)(2)]では，小規模な波形辞書から選択した音片を接続し，目的の基本周波数，継続長の音声を得るため，音片に信号処理による変形を行い，基本周波数，継続長を変化させていた。この変形によって，得られる音声波形は品質が劣化するため，高い音質が求められる公共施設における放送音声などへの適用は不可能であった。また，高級車のカーナビゲーションシステムにおける運転者への経路誘導メッセージでも，高級車のイメージを損なわないように，合成音声ではなく，ナレーターによる録音音声が好まれていた。しかし，詳細な情報の提供のためには，多数の音声の録音が必要であり，コストがかかるばかりではなく，最近のクラウドサービスなど，あらかじめ録音ができないようなサービス・コンテンツでは，音声での情報提供ができないという問題があった。

近年，音声コーパス(音声データを多数集めた音声データベース。例えば音声時間10時間以上)を用意し，大量の音片を蓄積しておき，波形の変形を行わないことで，肉声感の高い音声を合成できるテキスト音声合成方式が研究されている[(3)]。

このテキスト音声合成方式は，音声コーパスの中から，入力されたテキストに基づき，適切な音片を選択し，信号処理を行わずに接続し，合成音声を得る。この方式では，大量音片からどのようにして，適切な音片を選択するかが重要な課題となる。一般には，次の尺度で音片を選択する。

(1) 接続する前後の音片との連続性が高い音片(自然なつながり)

(2) 音片の継続長が，接続時に目的の継続時間のパターンに近くなる音片(自然なリズム)

(3) 音片の基本周波数が，接続時に目的の周波数パターンに近くなる音片(自然なイントネーション)

従来，これらの尺度に対して，コスト関数，すなわち目標値からの差異を評価する関数を経験的に定義し，コスト最小となる音片を選択していた。このコスト関数の重要な部分である目標値の算出では，リズムやイントネーションに影響を与えると経験上考えられる限られた言語要因(入力されたテキストから得られる情報の種類・組合せ)から目標値を推定する学習が行われていたが，十分な要因を考慮することができず品質劣化の原因となっていた。

そこで，我々は，与えられたテキストの様々な言語要因を条件として，音片系列全体の音響的特徴量の確率分布を音声コーパスから学習する方法を考案した(**図2**)[(4)]。この方法は，従来の経験に基づく，限定された要因による局所的な評価尺度の最適化ではなく，種々の言語要因の種類・組合せを同時に考慮し，音声の種々の特徴を確率的に学習し，最も高い確率で観測される音声特徴を再現する音片を

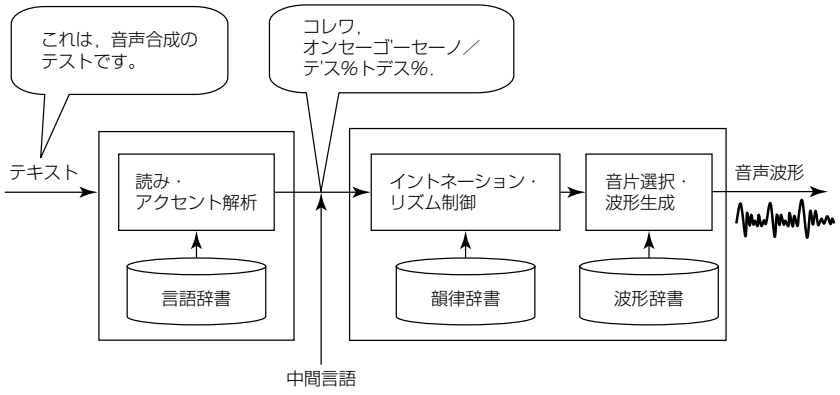


図1．テキスト音声合成方式

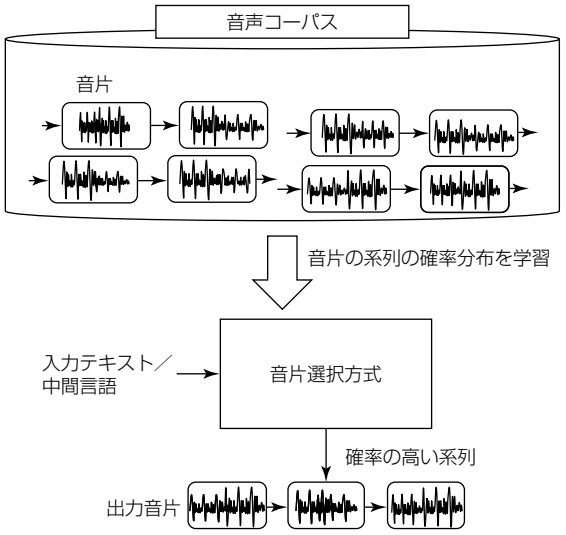


図 2．学習型の音片選択方式

選択する。この方法は，人間の発声で確率高く現れる現象が，品質の高い人間らしい肉声感のある声だという考えに基づく方法と言える。

この音片選択方式では，音片を選択するときに考慮する言語要因は，音片の系列の確率分布を特徴付ける素性関数によって表す。素性関数は，共起条件を満たす場合に反応する関数（共起条件を満たす場合に1，満たさない場合に0を与える）であり，この共起条件を用いることで，音片の音声特徴と，それらに関連する言語要因とを考慮した確率分布を学習することが容易となる。ここで，音声特徴は，音声分析によって得られる情報で，例えば，韻律（基本周波数・継続長），スペクトル振幅等である。また，言語要因は，入力テキスト又は中間言語から得られる言語情報で，例えば，音素の種類や，音高（声の高い，低い），音片の文内・単語内の位置等である。

例えば，接続する音片の基本周波数の差に着目した場合，それに関連する言語要因である先行音片と後続音片の音素の種類，音片の位置の3種を考慮する素性関数ϕi（i=1,..., N）は次のように設定する。

$$\phi i(\text{音声特徴，言語情報}) = \begin{cases} 1 & \text{if 条件}i \\ 0 & \text{上記以外} \end{cases} \quad \cdots\cdots\cdots\cdots(1)$$

条件i：先行音片と後続音片の基本周波数の差＝I
＆先行音片の音素＝P1
＆後続音片の音素＝P2
＆音高＝T

ここでI，P1，P2，Tは，接続される可能性のある音片の組合せから得られる値である。すなわちNは，I，P1，P2，Tの組合せの数である。このように定義した素性関数によって，種々の言語要因の種類・組合せを考慮した学習を行う。

開発方式による効果を確認するため，主観評価試験を行った。評定者8名で，提示された14音声に対して，非常に悪い（−3）～非常に良い（3）の7段階で評価した。**図 3** に

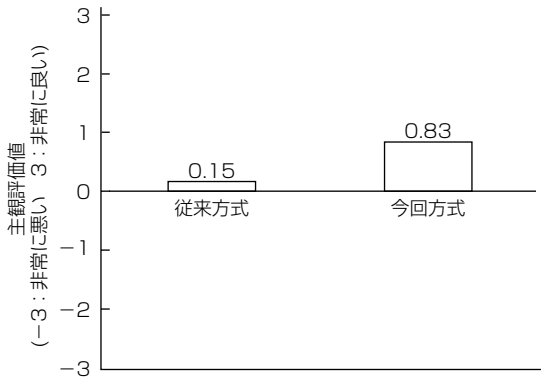


図 3．主観評価の結果

示すように，今回開発した方式は，従来方式[2]と比べ，大幅な改善（0.68の向上）が得られ，高い肉声感が得られることが確認できた。

2.3　音声強調による合成音声の聞きやすさの改善

一般に高齢者は加齢による聴覚器官の衰えによって聴力が低下し，特に高域（高い周波数帯域の音声）が聞き取り難（にく）くなる傾向がある。高域の明瞭性を改善する方法として，イコライザによって高域を強調し，明瞭性を高めることが考えられるが，DA（デジタル・アナログ）変換器などの再生機器の制限で音声再生周波数帯域を狭くせざるを得ないシステムでは，既に合成音声を生成する音片から高域の情報が欠落しているため，イコライザによる方法は明瞭性の改善にはつながらない。そこで，音声の再生周波数帯域より高域の信号を，あらかじめ音片辞書の再生周波数帯域内に重畳しておくことで，音声合成のためのメモリ量や処理量に負担をかけず，疑似的な高域特徴を再現し，合成音声の明瞭性を改善する音声強調方式を開発した[5]。

図 4 に開発方式の処理の概要を示す。

①入力音声をスペクトルに変換し，音声再生周波数帯域外の高域スペクトルを切り出す。
②入力音声の特徴を分析し，音声再生周波数帯域内で，高域を重畳するための適切な周波数境界（重畳帯域周波数Fc）を決定する。
③重畳帯域周波数Fcに基づいて，再生周波数帯域内の上限周波数帯域内に収まるように切り出した高域スペクトルを周波数方向に線形圧縮する。
④得られた圧縮スペクトルを，再生周波数帯域内に重畳する。

これは，本来の高域特徴に応じて，帯域内の高域を適応的に強調していることになり，制限された帯域内だけで疑似的な高域感が得られる。この方式によって，子音の明瞭性が向上することを確認した。

2.4　韻律変形による注意喚起可能な合成音声

公共施設における放送音声などは，聞き手が必要な時にだけ耳を傾けるため，災害時の避難誘導案内などの緊急放

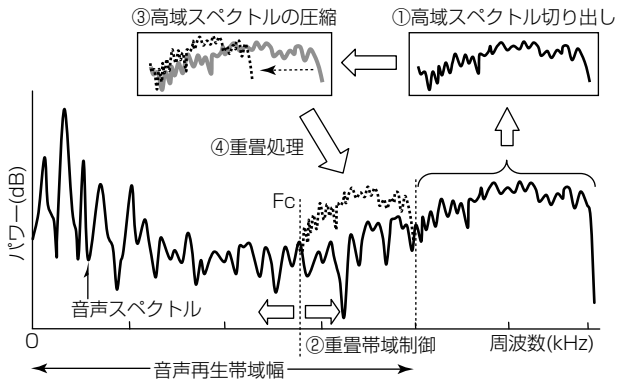


図4．音声強調方式の処理

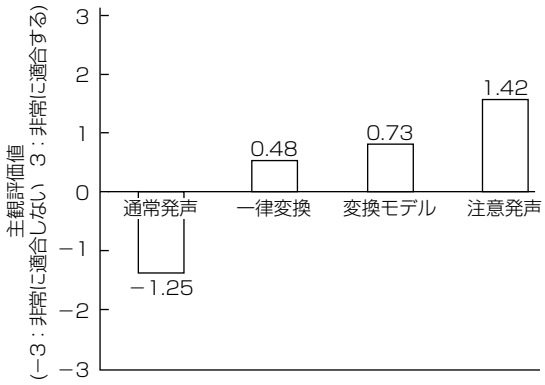


図5．主観評価の結果

送を通常のアナウンスと同様の音声で放送すると，聞き逃すおそれがある。これを防止するには，発話の特徴を変化させ，緊張感のある注意喚起可能な音声を合成する必要がある。

3名のナレーターの通常のアナウンス(通常発声)と注意喚起を行うアナウンス(注意発声)を収録・分析したところ，注意発声は，通常発声と比べて継続長が短く，基本周波数が高いこと，またその特徴を再現すると，注意喚起の効果が知覚できることを確認した。しかしこの特徴を持つ合成音声を生成するために，注意発声の音声コーパスを作成することは，大量の音声データの録音のコストからも，さらには合成音声が収録した声の特徴再現にしかならないため，多様性の制御の観点からも現実的ではない。

そこで，通常発声の韻律と注意発声の韻律の間の相対的な変化を予測し，通常発声(従来方式)の韻律を変換することで，自然性を保ちつつ，注意喚起可能な韻律を再現する方式を開発した。特に句頭と句末の部分で変化が大きいことに着目し2段階の変換モデルを構築した。

(1) 発話全体の継続長・基本周波数の平均値を変換
(2) 句頭と句末の継続長・基本周波数を局所的に変換

評定者6名で，提示された音声に対して，緊急時の注意喚起の放送音声としてのふさわしいかどうかを“適合度合い”の7段階で判定する主観評価試験を行った。評価音声は，緊急放送文を含む10文を用いた。適合度は，3：非常に適合する，2：適合する，1：やや適合する，0：どちらでもない，−1：やや適合しない，−2：適合しない，−3：非常に適合しない，の7段階である。

次の4方式の音声を比較した。

(1) 韻律変換を行わない合成音声(通常発声)
(2) 継続長・基本周波数の平均値だけを変換(一律変換)
(3) 平均値に加え句頭と句末を変換(変換モデル)
(4) 韻律の変換モデルの学習に使用したナレーターの発声(注意発声)

図5に示す結果から，通常発声では適合しないという判断であったのに対し，一律変換だけでも1点以上大きな改善効果があった。さらに変換モデルによって0.2点以上の改善があり，適合性を向上させることができた。開発した変換モデルによって，合成音声によってユーザーに注意を促すアナウンスが可能となった。

3. むすび

当社で開発したテキスト音声合成方式の最近の取り組みとその成果について述べた。開発したテキスト音声合成方式は，様々な製品への展開が可能である。

合成音声による情報伝達は，ここで述べた“注意喚起”にとどまらず，発話スタイルの制御や，強調等のより詳細な制御ができるようになれば，極めて豊かな情報を提供できる。ユーザーも，なんら学習する必要なく，人と人とのコミュニケーションで得た経験から，情報の意図や，種類，重要性を容易に理解できる。様々な分野で，効果的，かつ豊かなユーザーインタフェースが実現できるよう，今後も品質の更なる改良に加え，多様性の向上を図る予定である。

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) 大塚貴弘，ほか：テキスト音声合成技術，三菱電機技報，**85**，No.11，641〜644（2011）
(3) N. Campbell，ほか：CHATR: 自然音声波形接続型任意音声合成システム，電子情報通信学会技術研究報告.SP，**96**，No.39，45〜52（1996）
(4) 大塚貴弘，ほか：条件付き確率場に基づく波形接続型音声合成，2014春季-日本音響学会講演論文集，1-R5-10（2014）
(5) 古田　訓，ほか：入力信号の帯域外の高域成分を低域重畳する音声強調，電子情報通信学会ソサイエティ大会講演論文集，A-4-19（2013）