

HMMに基づく高品質音声合成法と評価法

Development and Evaluation Method of HMM-Based High Quality Speech Synthesis Technology

花園 正也* ・ 渡辺 英樹** ・ 西山 高史** ・ 徳田 恵一***
Masaya Hanazono Hideki Watanabe Takashi Nishiyama Keiichi Tokuda

漢字かな混じりテキストから音声を生成する音声合成法において、HMMに基づく合成音声の聞き取りやすさやイントネーションの自然性を向上できる収録テキスト設計法、技術的な達成度を評価するアプリケーションに共通の評価法、および実際にアプリケーションに特化した評価法を開発した。本技術は、FA機器のプログラマブル表示器の操作方法や状態などを知らせる音声ガイドに、従来の録音音声に代る音声作成機能として活用されている。

In the HMM-based speech synthesis method of generating sound from a text consisting of *kanji* and *kana* characters, a design method of recording text for improving ease of listening and naturalness of intonation, an evaluation method of rating the technical achievement common to applications, and an evaluation method specializing in actual applications have been developed.

Using the developed technology, a speech synthesis service for a programmable FA equipment display unit has been implemented as an audio guide for notifying the user of the equipment operation procedure and conditions, replacing the conventional recorded voice guide.

1. ま え が き

近年、家電や車載情報機器などで、漢字かな混じりのテキストを音声に変換する音声合成が利用されつつある。音声合成法には波形接続法などさまざまな方式があるが¹⁾、近年の計算機によるデータ処理能力の向上を背景に、HMM (Hidden Markov Model: 隠れマルコフモデル) に代表される統計的なアプローチが、音声認識分野だけでなく音声合成分野でも注目されている²⁾。この方式は音声波形を統計パラメータとしてモデル化することで、メモリー容量を小さく抑えることができ、また他の方式に比べてイントネーションが滑らかで自然であるという特徴がある。そこで筆者らもこの点に着目し、HMMに基づく音声合成法を利用した技術開発を進めている。

本稿では、音声合成技術の開発内容と合成音声の評価法について報告する。前者については、合成音声の品質向上法について述べる。合成音声の品質は波形の元となる音声合成データベース (以下、音声合成 D/B と記す) に影響され、またこの音声合成 D/B を構築するうえで、用いる音声データの収録テキストによる影響はとくに大きい。従

来から収録テキストの設計法については提案されているが³⁾、合成音声のイントネーションが不自然であるという問題があったため、筆者らは新しい設計法を提案する。また後者については、録音音声の評価法や JEITA IT-4001「音声合成システム性能評価法」⁴⁾ (以下、JEITA 規格と記す) を参考にし、評価法の詳細を設計した内容と実際に評価を行った結果を報告する。最後に本技術の適用事例として、プログラマブル表示器用の音声合成サービス「しゃべリーナ」について紹介する。

2. HMMに基づく高品質音声合成法

2.1 HMMに基づく音声合成法の概要

HMMに基づく音声合成法の基本構成を図1に示し、その概要を述べる。構成は、音声合成 D/B 構築処理とテキスト音声合成処理の二つに分けられる。音声合成 D/B は収録音声の元に作成しており、まず音声収録するテキスト内容を設計 (収録テキスト設計) し、アナウンサーにそのテキストを読ませ (音声収録)、得られた波形データを用いて HMM 学習を行い、音声合成 D/B を構築する。一方、テキスト音声合成処理では、入力された漢字かな混じりテ

* 新規商品創出技術開発部 New Product Technologies Development Department

** 先行技術開発研究所 Advanced Technologies Development Laboratory

*** 名古屋工業大学大学院 工学研究科 Graduate School of Engineering, Nagoya Institute of Technology

キストから、アクセント位置や読み情報を解析（言語解析）し、中間表記に変換する。そのあとで音声合成 D/B から、HMM を探索・接続し音声波形を生成する（HMM ベース波形生成）。

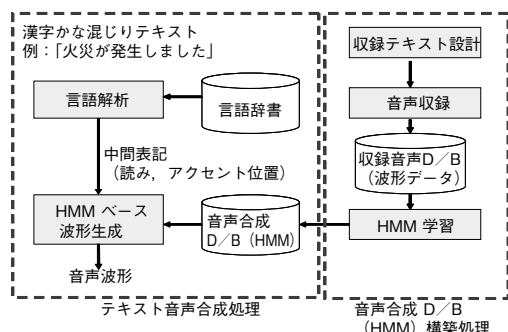


図1 HMMに基づく音声合成法の基本構成

次節では、合成音声の高品質化に大きく寄与する収録テキスト設計法について述べる。

2.2 合成音声の高品質化

一般に大量の収録音声データから音声合成 D/B を構築すれば音声現象を幅広くカバーできるため、合成音声の品質が良くなるとされている。しかし大量の音声収録には多大なコストが掛かるため、収録するテキスト集合はできるだけ小規模に留めることが望ましい。ここでは、HMM に基づく音声合成法において、高品質な文章読上げが可能となる音声合成 D/B を構築することを目的に、収録テキスト集合を最適化する方法について述べる。

収録テキスト集合の音素バランスを考慮した最適化法は、従来から提案されている³⁾。この方法では、音声現象を音素環境（2.2.2 項に詳細記載）で定義し、収録テキスト集合における音素環境のエントロピーを最大化している。これに基づき合成した音声は、各音素の明瞭性は良くなるものの、イントネーションは不自然なものとなっていた。

そこで筆者らは、音声品質において重要とされるイントネーションの自然さ向上を目的に、音素バランスだけでなく、韻律バランスも考慮した設計法を提案している^{5)~8)}。そこでは、半定型文の収録テキストを対象とした設計法⁵⁾、⁶⁾と、さらにそれを拡張した自由文脈でも対応可能な設計法⁷⁾、⁸⁾を述べている。これにより、少ない収録音声データからでも音響的に幅広く自然な音声合成できる。

2.2.1 収録テキスト集合の最適化アルゴリズム

本研究では、収録するテキストセット集合 σ が含むことのできる単語数 M と、集合 σ の母集団 Σ が与えられているという条件のもとで、音素バランスと韻律バランスが取られた集合 σ を求める。以下に集合 σ を求める手順を示す。

(1) 初期化

①収録するテキストセット集合 σ の単語数 M を決定す

る。

② L 個の単語から成る母集団 Σ から、任意に M 個の単語を選択し、集合 σ に入れる。

③ $(L - M)$ 個の単語が残された集合を Σ' とする。

(2) 反復処理

①集合 σ 、集合 Σ' のなかからそれぞれ任意に 1 個の単語を選択して交換したと仮定し、新たに構成される集合 σ' の評価関数 $S(\sigma')$ の値を求める。また元の集合 σ の評価関数 $S(\sigma)$ の値を計算する。

② $S(\sigma') > S(\sigma)$ となる場合は、当該 2 個の単語を交換する。

(3) 終了条件

①反復回数が所定の値 X に達したら終了する。

2.2.2 音声現象の定義法

以下に、例文とともに音声現象の定義法を説明する。

(1) 文章:「ニューヨークを、1 週間ばかり取材した」

(2) 音素表記:「S nyuuyookuo sp iclsyuukaNbakari syuzaishita S」

(3) 韻律表記:「S 6-1 S 9-7 5-0 S」

韻律表記中の「6-1」は、アクセント句「ニューヨークを」におけるアクセントパターン（モーラ数 6 とアクセント型 1 の組合せ）、「S」は無音（ポーズ）を意味する。

音素環境は、「S_ny」、「ny_u」のように 2 連鎖、あるいは「ny_u_u」のように 3 連鎖で定義する。これは、この例の「u」のように同音素であっても前後の音素により音響特性が変わることに対応するためである。韻律環境は、この考え方を応用し、前後関係を考慮して定義する。たとえば「S_6-1」、「6-1_S」のような 2 連鎖や、「S_6-1_S」のような 3 連鎖を定義する。

2.2.3 評価関数の定義

音素ならびに韻律バランスの度合を評価関数 $S(\sigma)$ で評価し、次式で定義する。

$$S(\sigma) = w_{ph} \cdot S_{ph}(\sigma) + w_{ac} \cdot S_{ac}(\sigma) \quad (1)$$

$$S_{ph}(\sigma) = - \sum_{n=1}^{T_{ph}} p_n \log_2 p_n \quad (2)$$

$$S_{ac}(\sigma) = - \sum_{m=1}^{T_{ac}} q_m \log_2 q_m \quad (3)$$

ここで、 p_n 、 q_m はそれぞれ音素環境 n ($n = 1 \sim T_{ph}$) と韻律環境 m ($m = 1 \sim T_{ac}$) における相対出現頻度である。 w_{ph} 、 w_{ac} は各評価値に対する重みづけであり、 $w_{ph} + w_{ac} = 1$ とする。 $S_{ph}(\sigma)$ 、 $S_{ac}(\sigma)$ は、音素環境、韻律環境それぞれの出現が独立事象であるときには情報エントロピーと解釈することができる。しかし、各単位が互いに重複することがあり、完全な独立事象ではなく厳密には情

報エントロピーとは一致しないが、以下では文献³⁾に基づき評価関数として利用し、単位はbitsを用いる。なお各環境の種類は、あらかじめ母集団を分析することで定義する。

2.2.4 最適化法の評価

(1) 評価方法

提案法を利用することで、収録テキスト集合において音素・韻律バランスを取ることができるかを確認する。ここでは、定義した音素・韻律環境に対する被覆度を示すカバーレッジで効果を評価する。比較対象として、母集団 Σ からランダム抽出した場合(2.2.1項での反復交換処理の初期状態)と、音素バランスのみを考慮した従来法³⁾とでも評価する。

ここでの母集団 Σ は、音素・韻律表記の情報を付与した1056個のテキストで「○○(場所)で、△△(主語)が、□□(述語)」といった2～4個のアクセント句から成る文章とする。定義された音素・韻律環境は母集団に出現するすべての特徴とし、音素環境は429種類、韻律環境は149種類である。また反復回数 X は5万回とする。

(2) 評価結果

最適化過程における評価関数 $S(\sigma)$ の結果を図2に示す。

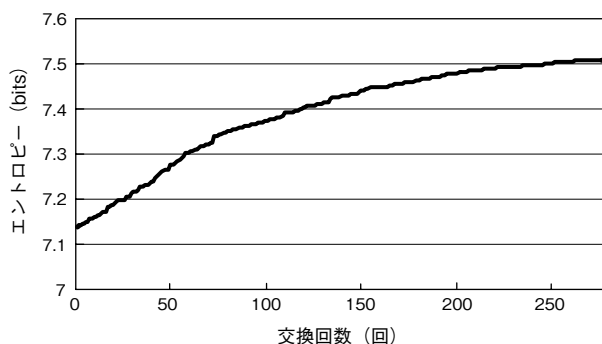


図2 交換回数と情報エントロピーの関係

横軸の交換回数は、集合 σ 、母集団 Σ 間で実際に交換される回数であり、2.2.1項で示した反復回数 X ではない。ここでは単語数 M が100の場合の結果を示す。この図から最適化処理前の情報エントロピーは約7.14 bitsであるが、最適化することで約7.51 bitsに上がり飽和していることがわかる。また、音素・韻律環境に関する評価値についても、その値が飽和することを確認している。

次に収録テキスト内のカバーレッジ（集合 σ に出現する環境の種類数／定義した環境の種類数）を図3に示す。左列から集合 σ の単語数ごとに音素・韻律のカバーレッジを示している。集合 σ の単語数 M にかかわらず、提案法では韻律環境のカバーレッジが他よりも高いことがわかる。たとえば、提案法では単語数 M が100であれば100%に達しているのに対し、提案法以外は母集団 Σ の約半数であ

る単語数 M が500であっても100%に達していない。また従来法では、ランダム抽出に対し韻律バランスが低下することも確認される。

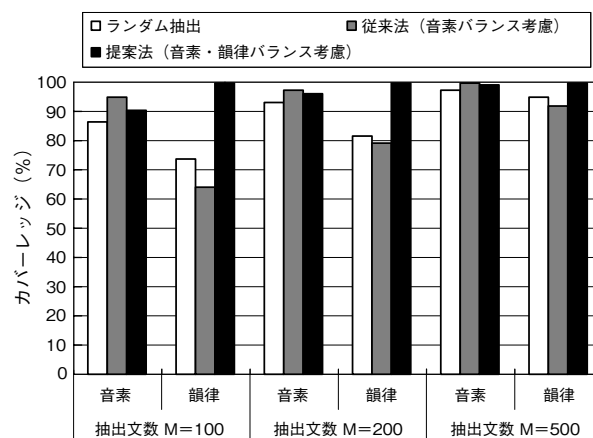


図3 カバーレッジの比較

以上のことから、提案法ではとくに韻律環境においてカバーレッジの向上が確認され、音響的に幅広く自然な音声合成できる。

3. 合成音声の評価法

録音音声に対して、聞き取りやすさや音量の適切さなどを主観評価するためのさまざまな方法が提案されている⁹⁾。しかし合成音声の評価する際は、これらに加えて合成音声特有の評価法が必要となる。

音声の品質は大きく分けて、テキストの聞き取りやすさを評価するバーバル情報評価とイントネーションの自然さや感情などを評価するパラ言語情報評価の二つがある。録音音声においては、訓練されたアナウンサの声を録音することでイントネーションなどのパラ言語情報の品質が保障されているが、合成音声の場合はパラ言語情報も含めて評価する必要がある。

また評価の目的から分類すると、以下の二つに分けられる。一つは、合成音声の開発段階での技術的な達成度を評価することを目的とする文脈によらないアプリケーション共通の評価法であり、その内容は本章で述べる。もう一つは、HMMに基づく音声合成の場合は入力するテキストごとに品質がばらつく可能性があることから文脈ごとに評価することを目的とするアプリケーションに特化した評価法であり、4章でアプリケーションの紹介とともに述べる。

合成音声の評価法の開発は、合成音声の評価のガイドラインとして知られているJEITA規格⁴⁾を参考にして行う。この規格には大別すると正読率、了解性、および総合評価の三つの指標がある。

正読率は、入力した漢字かな混じりテキストを言語解析して得られる読み情報の正しさを評価する。本稿では合成音声の評価についてのみ報告することとし、言語解析部の

評価となる正読率は省略する。理解性は、合成音声の明瞭性を評価する。総合評価は、利用される目的や環境などを実使用にできる限り近い状態にし、理解性だけでなく抑揚の自然さなどの評価項目についても総合的に評価する。

なお JEITA 規格によると、合成音声を用いたシステムのどの指標に重点を置いて評価するかに応じて評価結果の解釈や利用法が吟味されるべきであり、また評価したい指標を明確にしたあとでその指標を重点的に評価すべきであると記載されている。したがって筆者らは、アプリケーションや評価目的ごとに評価すべき指標を設定している。

3.1 理解性の評価

アプリケーション共通の評価法の一つとして、理解性の評価を行う。理解性は音節明瞭度と単語理解度の二つに分けられる。筆者らは個々の音節の明瞭度よりも、実使用に近い単語の音声を理解できるかを重視し、単語理解度を評価に用いる。単語理解度は、JEITA 規格にある単語を音声合成させ、正しく聞き取れるかの評価を行う。ここでは表 1 の条件で評価を行う。被験者は音声を受聴し、聞き取ったテキストをキーボードで入力する。なお音声の聞き取りやすさはそのテキストの親密度に影響されるため⁴⁾、ここでは聞き取り評価後に、被験者が各単語に対して、「知っている」、「知らない」の択一で単語を分類したものから、親密度と聞き取りやすさとの関係について確認する。

表 1 単語理解度評価条件

被験者	人数	5名（男性 2 名、女性 3 名）
	平均年齢	46.3 歳
出力機器		密閉型ヘッドホン
音声		女性音声
標準化周波数		8 kHz
評価テキスト		JEITA 規格 単語理解度試験用単語
評価数		1169 語

単語理解度の評価結果を表 2 に示す。親密度に関しては、5 名の被験者のアンケートから、3 人以上が知っていると回答した単語を「知っている」、それ以外を「知らない」単語として分類し、親密度ごとに単語理解度を計算して、被験者の平均値を求めている。「知っている」と回答した単語 948 語に対する単語理解度は 85 % と高いが、「知らない」と回答した単語 221 語に対する単語理解度は 58 % と低く、なじみのない単語は聞き取りにくいことがわかる。つまり、単語理解度は親密度に影響することを考慮して評価結果を考察することが必要である。

表 2 単語理解度評価結果

親密度	単語理解度	評価数
知っている	85 %	948 語
知らない	58 %	221 語
合計	80 %	1169 語

3.2 総合評価

総合評価の評価項目として、JEITA 規格のなかから次の三つを選定する。

- (1) 聞き取りやすさ（理解性）
- (2) 抑揚の自然さ
- (3) こもり感（音質）

各項目について 1 ～ 7 点の 7 段階で評価し、平均値を算出する。点数が高いほど品質が良く、こもり感に関しては、それが気にならないほど点数が高い。評価において開発している合成音声に加え、リファレンスとして音声ガイドで一般に用いられる録音音声も評価する。

しかし、想定されるテキストが多く、そのすべてを評価しなければならない場合には、前述の方法を行うことは効率的ではない。その場合は、重要な聞き取りやすさのみを簡易的に評価し、抑揚の自然さやこもり感などは自由意見として聴取することで効率化を図る。

4. アプリケーションへの適用事例

音声合成技術の適用事例として、すでに 2008 年から運用を開始している当社「プログラマブル表示器 GT32T1 シリーズ」用サービス「しゃべリーナ」についての概要と評価結果について述べる。

4.1 アプリケーションの概要

このプログラマブル表示器にはサウンド出力機能が搭載されており、異常検出時に報知音とともに音声を再生することで、視覚や触覚だけでなく聴覚も刺激して操作者への情報伝達をより確実にすることをねらっている。従来、音声収録のコストと手間が課題とされていたが、音声合成技術を利用することで、だれにでも簡単に音声を作成できるようにしている。

音声作成の手順を図 4 に示す。ユーザはウェブサイトで、テキスト入力し「合成する」ボタンを押下することで当テキストデータがセンタサーバに送信される。センタサーバはテキストを受け取りと音声合成し、音声を作成する。この音声はウェブサイト上から試聴、ダウンロードできる。ダウンロード時にはプログラマブル表示器専用のファイル形式に変換される。

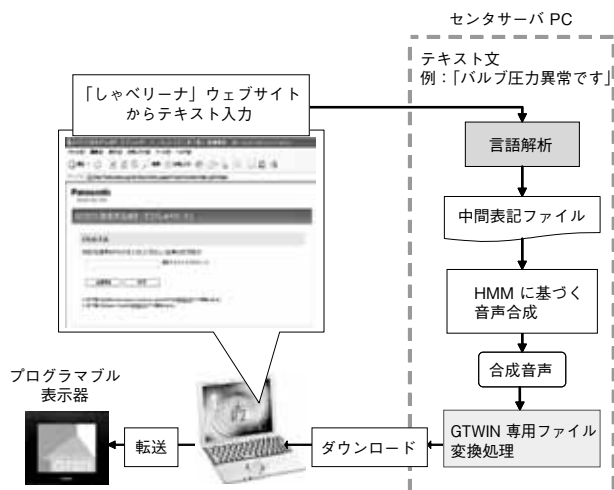


図4 プログラマブル表示器用音声作成の手順

4.2 合成音声の評価

4.2.1 理解性の評価

評価にあたっては、実使用の観点から理解性に重点を置き、アプリケーションで想定されるテキストに基づき合成した音声に対して、表3の条件で簡易的な評価を行う。音声合成D/Bは、3章で単語理解度を確認したものを用いる。出力機器は、さまざまな機器が想定されるため、その違いによる評価結果のばらつきを排除するために密閉型ヘッドホンを使用する。

表3 理解性評価条件

被験者	人数	5名（男性2名、女性3名）
	平均年齢	46.3歳
出力機器		密閉型ヘッドホン
音声		女性音声
標準化周波数		8 kHz
評価テキスト		「しゃべリーナ」想定テキスト 例：「エア圧低下」
評価数		299文

また3章で述べたように、なじみのない言葉は聞き取りにくいことがわかっているため、被験者は音声を受聴して、そのテキストを確認し、そのとおり聞こえたかを評価する。なお、自由意見の聴取も同時に行う。

評価の結果、理解性は被験者平均で96.3%と高く、問題なく聞き取れている。また自由意見の聴取結果から、聞き取れないものの特徴としては名詞を連続させる聞き慣れないテキストなどが挙げられ、その作成において工夫が必要である。

4.2.2 総合評価

ここではアプリケーションとして、家電機器に搭載のスピーカで再生した場合の総合評価結果を述べる。表4の条件で各評価項目に対して7段階で評価する。音声合成D/

Bは、3章で単語理解度を確認したものを用いる。図5に各評価項目に対する評価値の平均値を示す。なお、エラーバーは標準偏差（ $\pm \sigma$ ）である。

評価の結果、合成音声は録音音声と比べてすべての評価項目で平均値の差が1点未満とわずかである。この結果は、出力機器がヘッドホンによる評価結果よりも高いことから、出力機器の性能も考慮した分析が必要であることを示唆している。

表4 総合評価条件

被験者	人数	5名（女性）
	平均年齢	64.0歳
出力機器		家電機器搭載スピーカ（受聴距離：1m）
音声		女性音声
標準化周波数		16 kHz
評価テキスト		案内文 例：「動作が停止します」
評価数		9文

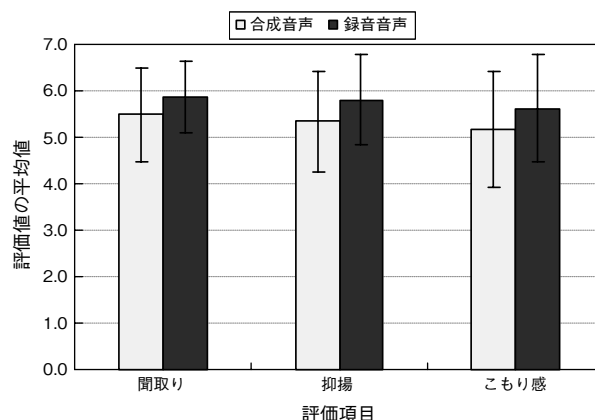


図5 合成音声の評価結果

5. あとがき

漢字かな混じりテキストから音声を生成する音声合成法において、HMMに基づく合成音声の聞き取りやすさやイントネーションの自然性を向上できる収録テキスト設計法、技術的な達成度を評価するアプリケーションに共通の評価法、および実際にアプリケーションに特化した評価法を開発した。本技術は、FA機器のプログラマブル表示器の操作方法や状態などを知らせる音声ガイドに、従来の録音音声に代る音声作成機能として活用されている。

今後、さまざまなアプリケーションへの適用を通じて、合成音声品質の一層の向上に取り組むたい。また評価法に関しては、客観的な評価法も検討中であり、さらに主観評価の効率化や評価の自動化に取り組んでいきたい。

*参考文献

- 1) 古井 貞熙：音声情報処理，森北出版，（1998）
- 2) 徳田 恵一：HMM に基づく音声合成の基礎，電子情報通信学会技術研究報告，Vol. 100, No. 392, SP2000-74, p. 43-50, （2000）
- 3) 磯 健一，渡辺 隆夫，桑原 尚夫：音声データベース用文セットの設計，日本音響学会講演論文集，Vol. 1, 2-2-19, p. 89-90 （1988）
- 4) 音声入出力方式標準化委員会：音声合成システム性能評価法，電子情報技術産業協会，JEITA IT-4001 （2003）
- 5) 花園 正也，西山 高史，徳田 恵一：文脈限定下での高品質 HMM 音声合成のための収録テキストセットの最適化，第 52 回 SCI 学会研究発表講演会，6U4-2, （2008）
- 6) 花園 正也，渡辺 英樹，西山 高史，徳田 恵一：文脈限定下での HMM 音声合成のための収録テキストセットの最適化，日本音響学会講演論文集，2-P-27 （2008）
- 7) 花園 正也，渡辺 英樹，西山 高史，徳田 恵一：HMM に基づく感情音声合成のための収録テキストの最適化設計，第 53 回 SCI 学会研究発表講演会，F36-1 （2009）
- 8) 花園 正也，渡辺 英樹，西山 高史，徳田 恵一：人の感情に訴え人を動かす感情音声合成技術の開発と評価，ヒューマンインターフェイス学会シンポジウム，2323 （2009）
- 9) 渡辺 英樹，有川 泰史：ユニバーサルデザインに配慮した音声ガイドの「聞き取りやすさ」評価法，松下電工技報，Vol. 53, No. 4, p. 28-33 （2005）

◆執筆者紹介



花園 正也

新規商品創出技術開発部



渡辺 英樹

先行技術開発研究所



西山 高史

先行技術開発研究所
博士（工学）



徳田 恵一

名古屋工業大学大学院
博士（工学）