音声言語情報処理 15－2
（1997. 2. 7）

参照辞書を用いた合成単位辞書自動生成法の検討

斉藤　隆
日本アイ・ビー・エム株式会社　東京基礎研究所

本報告では、ユーザが入力した音声からシステムが全自動的に合成単位辞書を作成し、それを用いた合成が可能となるような合成システムを検討する。合成単位辞書の自動生成方法としては、予めマニュアル法で作成した信頼性の高い合成単位辞書を参照用辞書として利用することによって、比較的小規模で効率的に自動生成を行なう方法を検討する。今回、ベースとする音声合成システムとして筆者らが従来より検討をすすめてきた波形合成システムを適用し、実験システムを作成した。さらに、提案する合成単位自動生成法の基本的な性能について、マニュアルで作成した合成単位辞書と比較検討を行なった。

Automatic Construction of Synthesis Unit Inventories Using a Reference Unit Inventory

Takashi Saito
IBM Research, Tokyo Research Laboratory, IBM Japan Ltd.

In conventional text-to-speech systems, synthesis unit inventories are usually prepared in advance through a laborious process of speech data gathering, analysis, and manual segmentation. This paper propose a new type of text-to-speech synthesizer, which has a function of registering automatically a new speaker's voice to synthesis unit inventories, by providing users with a function of registering their voices. The construction of unit inventories is carried out by making the most of a "reference unit inventory". Experimental results are shown for a waveform-concatenation- based TTS system, which was recently developed by the authors.

1．はじめに

従来より、テキスト音声合成システムで用いられる合成単位辞書は、合成システムに付随した固定データとしてユーザに提供され、合成に使用されるのが一般的である。このような従来型のテキスト音声合成システムでは、出力可能な音声の種類は、当然、提供される合成単位辞書の種類に限定されることになる。合成単位辞書に関するこういった制約は、今後いっそう重要になると考えられる多種多様な音声の合成を実現していく上では、従来型合成システムの大きな課題であるとも言える。

本研究では、合成音声の種類が提供される辞書に限定されるという制約を取り払うことによって、多種多様な音声の合成を可能とするようなシステムを検討する。すなわち、音声合成システム自体に、合成単位辞書を新規作成する機能を持たせることを考える。ユーザが入力した音声から全自動的に単位辞書を生成することが可能な音声合成システムが実現されれば、作成した辞書の切り替えによって、様々な声質を容易に合成することができるようになる。

このような音声合成システムの具体的な応用例としては、たとえば次のようなものが考えられる。

・規則合成音と録音再生音の声質均一化

従来、音声応答システムで録音再生音と規則合成音を併用する場合、単独で使用する場合に比べ、声質の違いが際立って規則合成音の印象を一層悪化させることにもなった。同一話者による合成によって、この問題を大幅に緩和することができよう。

・合成音声にパーソナリティーを付与

教育用ソフト、ゲーム・ソフトなど、対話インターフェースが有効となる応用分野では、コンピュータを擬人化し、そのパーソナリティーに応じた声を与える。これにより、人間にとってより自然で使いやすいインターフェースを提供できる。例えば、電子メールを差出人の声で読み上げるといった応用も可能になる。

合成単位辞書の自動作成手法としては、最近、HMMによる音声認識の強力な音素アラインメント技術を利用したものがいくつか提案されてきている[1,2]。ここでは、あらかじめマニュアル法で作成した信頼性の高い合成単位辞書を参照用辞書として利用することによって、比較的小規模で効率的に自動生成を行なう方法を検討する。

ここでベースとする音声合成システムとしては、筆者らが従来より検討をすすめてきた波形合成システム[3,4]を適用し、実験システムを作成する。さらに、提案する合成単位自動生成法の基本的な性能について、マニュアル法で作成した合成単位辞書と比較検討を行なう。

2．新規話者音声の登録機能をもつ音声合成システム

ここで提案する新規話者音声を用いた合成単位辞書の作成部は前述したように、音声合成システムに組み込まれる一機能として構成される。ベースとするテキスト音声合成システムとしては、以前より我々が開発を進めてきた波形合成システム([3,4])を用いる。その音声生成部の特長としては、

・環境依存型音節を合成単位として採用[4]

・波形切出しと重ね合わせの独立制御[3]

などが挙げられる。これらの工夫によって、比較的高品質な音声の合成を実現している。また、このシステムは、日本語解析から音声生成までの一連の変換処理をＰＣ上で実時間で行なうことが可能となっている。

(2-1) 新規話者音声登録システム

提案する新規話者音声の登録システムの基本構成を図１に示す。新規話者用の合成単位辞書の自動生成は、あらかじめ用意しておく基準話者の合成単位辞書（参照用辞書）を利用して行なわれる。この参照用辞書は、入力ガイド用合成音声の生成と入力音声のroughな音素アラインメントに利用される。なお、参照用辞書の合成単位情報は信頼性の高いことが必要となるため、視察によって正確に切り出しが行われたものを使用する。

新規話者の音声登録の手順は次のとおりである。まず新規話者は、登録用語彙（ここでは単語を想定）を合成音声によるガイドにしたがって順次入力していく。システムに取り込まれた

音声は、合成単位自動生成部において、単語境界検出後、参照用音声とのＤＰマッチングにより、音素への自動区分化が行なわれる。さらに、有声・無声区間判別の後、波形合成に必要な制御情報の付与が行われ、同時に音素境界の詳細補正も行われる。最後に、合成単位として用いる音節部分が切り出され、合成単位辞書に登録される。

(2-2) ユーザ・インターフェース

ここでは、音節長程度の合成単位の切り出しを目的としているため、認識システムのように全くの自由発声を許す必要はなく、ある程度発声の仕方を規定する方法で入力を安定化させることを考える。

音声登録のセッションでシステムはまず、合成エンジンの波形生成部において、登録すべき単語に対して、基準話者の合成単位辞書とその単語の韻律テンプレート（ピッチ、音韻時間長、パワー）から、ガイド用音声を合成し、ユーザに呈示する。

ユーザは、ガイド音声を復唱する形で、対象音声を発声する。ガイド音声の呈示によって、発声すべき内容を確認しつつ発話登録を行なえるので、基準話者音声との極端な発声様式の違いを未然に防ぐことができる。また、不慣れなユーザに対しては発声補助として働き、比較的安定した発声の登録が期待される。

入力した発声内容については、直ちに出力して、不良な発声になっていないかの確認ができるようになっている。

このように、音声登録のユーザ・インターフェースの工夫によって、自動生成処理に好ましくない入力音声を極力排除し、自動生成部の負荷を軽減するように配慮している。

3．参照辞書を用いた合成単位辞書の自動生成方式

合成単位辞書の自動生成処理は大きく分けて、初期区分化処理、ＧＣＩ検出処理、詳細区分化処理、の３つから成っている。参照辞書は、最初の初期区分化処理において、フレーム単位の音素境界を初期設定するのに使用される。ここで得られるフレーム単位の音素境界値が、それ以降の自動生成処理のベースとなる。次に各部の詳細について説明する。

(3-1) 初期区分化処理

正規化対数パワー閾値に基づいた単語境界検出を経た後、入力音声は、参照辞書の基準話者話者音声テンプレートとのＤＰマッチングを行なうことによって、音素単位に区分化される。ＤＰの結果得られた入力・参照音声両者のフレーム毎の対応関係から、参照話者の音素境界フレームに対応する入力音声フレームを、音素境界位置の初期値として設定する。このフレーム単位の音素区分情報をもとに、波形重ね合わせのための制御情報、サンプル・ポイント単位の音素区分情報の抽出が順次行なわれる。

(3-2) ＧＣＩ検出処理

ベースとしている波形合成システムでは、波形重ね合わせの基準点として、声門閉鎖点（ＧＣＩ：Glottal Closure Instants）を用いており、ここではその自動検出が必要となる。ＧＣＩの検出にわれわれは、ウェーブレット変換を

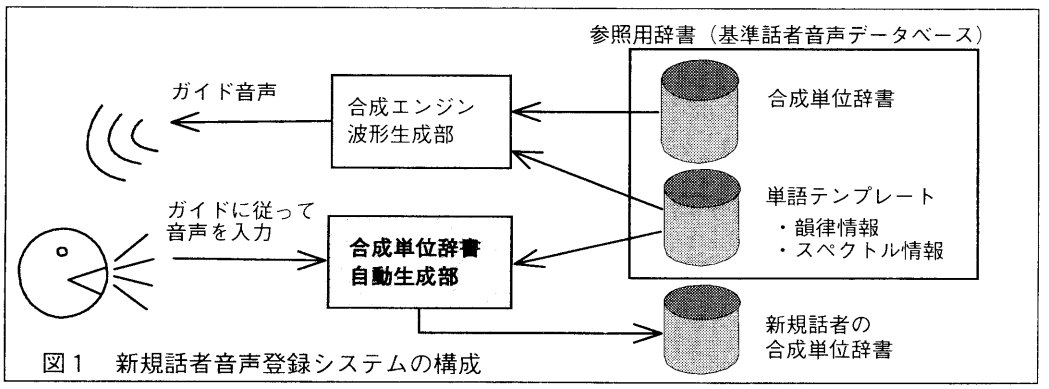


図１　新規話者音声登録システムの構成

用いる方法を採用している。ウェーブレット変換は信号の急峻な変化点を抽出するのに適しており、ＧＣＩを有声音声波形の急峻な変化点とみなして求める。この方法は元々Kadambeら[6]によって提案されたが、われわれはこの方法を波形合成システムに応用し[3]、今回完全自動化のため、さらに抽出精度の向上を図った。

ウェーブレット変換は一般的にかなりの計算量を要する。ここでは、ＧＣＩ検出の処理量を削減するために、検出の前処理としてフレーム単位での有声・無声区間の判別を行なう。その結果から有声区間についてのみ、ＧＣＩ検出処理を適用する。この有声・無声判別は現在、対数パワーと零交差数に基づいた判別[5]を参考に行なっている。判別閾値の設定は合成処理への影響の度合いを考慮して、有声から無声への誤りを押さえるように調整がなされている。

ＧＣＩ検出を行った後、ピッチ波形切り出し窓の中心点として用いるためのローカル・ピーク点を設定する。このローカル・ピーク点は、２つの隣り合うＧＣＩ点間を対象区間として探索され、決定される。

(3-3) 詳細区分化処理

ここでは、(3-1)で得られたフレーム単位の音素区分情報を、波形合成の合成単位を切り出すため、サンプル・ポイント単位の音素区分情報に変換するとともに、音素境界の種別に応じた区分初期値の修正を行なって、最終的な区分化精度の改善を図る。

まず(3-1)のフレーム単位の音素区分位置をサンプル・ポイント単位に変換し、境界初期値とする。

次に、(3-2)のフレーム毎の有声・無声値に基づいて、各音素区分毎に有声・無声判別を行なう。この際、母音（/i/,/u/）や有声摩擦音（/z/,/j/）等揺らぐ音素については、両方の可能性を考慮する。

最後に、得られた各音素の有声・無声値データから、有声・無声区間の境界となる音素境界では、音素境界の強制的な修正を行なう。すなわち、無声から有声への遷移境界ではピッチ・マークの始点に、また、有声から無声への遷移境界ではピッチ・マークの終点に、それぞれ、音素境界位置を一致にさせるように修正する。

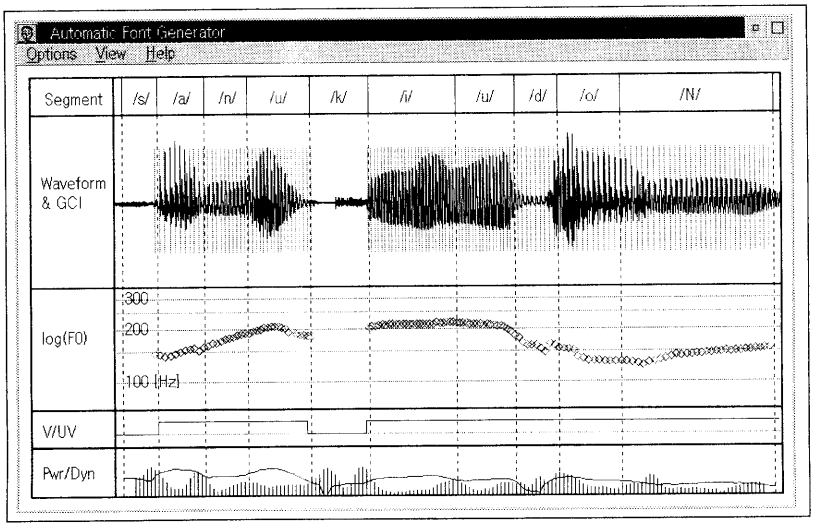
以上、３つの処理を経て、ＣＶ音節の波形データが合成単位として切り出され、合成単位辞書に制御情報とともに登録される。

４．合成単位自動生成実験

(4-1) 実験システム

・登録語彙セット

図２　音素自動区分化の結果例（男声,「さぬきうどん」）



合成単位を構成するために登録する語彙セットの選定には、まず、５～１０分程度の発声登録で最小構成可能であること、さらに、語彙追加によって品質向上できること、などを念頭においている。今回の検討は最小構成の性能について確認することに重点をおき、301単語からなる比較的小さめの語彙セットを用いた。この語彙セットは、文献[4]の環境依存性の検討で使用した語彙のサブセットで、合成単位として必要な音節をカバーする最小セットとなっている。

・音声分析条件

音声はサンプリング周波数22kHz、16ビットでA/D変換したものを、ＧＣＩ検出・合成単位抽出に使用する。他の自動化処理のための音声分析については、1/2の11kHzにダウン・サンプリングしてから処理を行なっている。

また、ＤＰマッチングの特徴ベクトルとしては、音声スペクトル（臨界帯域フィルタ出力＋パワー）の静的特徴、および、動的特徴を使用し、そのユークリッド距離をマッチンクの距離尺度とした。FFT分析の窓幅、フレーム周期は、それぞれ23.2ms、5.4msである。

・自動区分化結果

図２に音素自動区分化およびＧＣＩ検出結果の一例を示す。概ね、自動化処理は良好に動作している。特に、ＧＣＩ検出の精度は、以前の我々の合成システムのものよりもかなりの向上がみられる。しかしながら、パワーの小さい有声子音(/r/,/b/,/d/)や、mixed voicingの子音(/z/,/j/)などで、検出誤りがまだ見受けられ改善の余地がある。

(4-2) 音素自動区分化の性能

音素自動区分化の基本性能を調べるため、マニュアル・セグメンテーションとの比較を行なった。参照用辞書データとしてマニュアルで合成単位の切り出しを行った男声データを用意し、新規話者として女性話者１名が発声したデータについて、本方式による自動区分化と、視察による区分化とを行なった。

区分化の対象となった1964音素境界について、自動区分化と視察による区分化の結果を比較した。平均設定誤差は、ＤＰマッチングによる初期区分化の結果では、3.8フレーム（約20ms)であった。視察での作業に本質的に内在する誤差

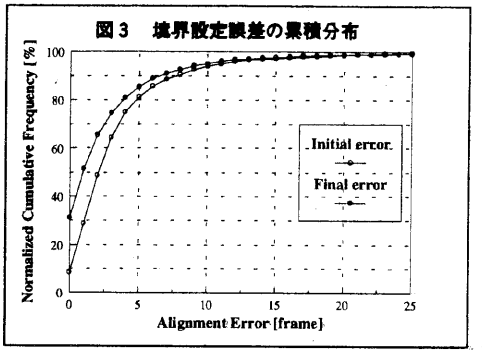


図３　境界設定誤差の累積分布

を考慮すれば、平均的には良好な区分化が行われていると考えられる。さらに、詳細区分化の結果、平均誤差は 2.8フレーム（約15ms）まで減少し、初期区分化から約1.0フレームの改善がみられた。

境界設定誤差の累積分布を図３に示す。このグラフは、許容誤差値を与えたとき、その値以内で区分化されている音素境界の割合を示す。なお、図中のInitial error、Final errorはそれぞれ、初期区分化、詳細区分化の設定誤差を表わす。誤差が10フレーム以内の境界は、詳細区分化によって、かなり改善されていることが分かる。また、4フレーム(21.8ms)以内の誤差で区分化されているものの割合は全体で80%程度であるが、表１に示すように、音素境界のタイプによって、かなり異なった傾向をみせる。これはスペクトル変化の緩やかな境界ほど誤差が大きくなることを示唆している。半母音や母音の連鎖等では、視察でも境界設定の大きな揺れは避けられず、当然の結果ともいえるが、一方で、音韻時間長制御・タイミング制御を安定に行なうことを考えると、やはり、一貫した境界設定基準が必要と考えられる。

表１　音素境界のタイプによる区分化誤差の違い

Boundary type	Less than 4-frame error (%)
All	80.8
C-to-V	87.5
V-to-C	77.8
V-to-V	67.1

(4-3) 合成音声の品質

4名の話者（男2名、女2名）について合成単位辞書の自動作成し、合成音声を試聴した。現時点では自動区分化の性能はけっして十分でなく、しかも、登録語彙セットが小さいこともあって、全般的に滑かさの欠けた感じは否めない。ただし、発声者の韻律を使って試聴したところでは、話者性についてはある程度保存されており、自動区分化の性能向上・語彙セットの拡充によって改善できる感触を得ている。

5．考察

(5-1) グロス・エラー対策

大きな設定誤差、例えば、10フレーム以上の誤差が生じているものも、図3より、5%程度存在することがわかる。このようなグロス・エラーは合成音の品質に重大な影響を与える恐れがあり、さらに低く抑えていく必要がある。対策としては、図4に示されるような境界設定誤差とDPスコア（累積距離）との相関を利用して、スコアに基づいたリジェクション機能も有効な方法の一つと考えられる。この種のエラーには発声の取り込み時の誤りが多く含まれており、再入力が必要であることからも妥当な方法といえる。

(5-2) 詳細区分化基準

今回は主として、GCI検出結果を有効に使うことによって、詳細区分化を行なった。ところが、これは有声音どうしの音素境界に対しては、当然ながら、あまり効果がない。そこで、スペクトル変化率や波形包絡変化率を用い、他の音素境界クラスに対しても有効な詳細区分化法について、現在検討を行なっている。

6．おわりに

マニュアル法で作成した信頼性の高い合成単位辞書を参照辞書として利用して、合成単位辞書を全自動的に作成するシステムについて報告し、自動生成法の基本性能をみるためマニュアル法との比較検討を行なった。合成音声の品質では、マニュアル法に比べると、現時点ではまだまだ改善の余地がある。今後、さらに、音素区分化、GCI検出の改良を進めていく必要がある。また、今回、音声登録に用いた語彙セットは音素環境数としてはかなり少ないため、明瞭度は確保できても、滑らかさについては明らかに課題があった。語彙セットのサイズと品質との関係についても、今後調べていく予定である。

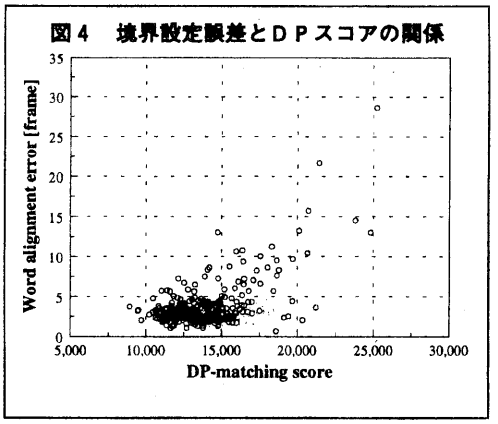


図4　境界設定誤差とDPスコアの関係

謝辞

本研究の初期段階においてご協力頂いた橋本泰秀氏（現在、日本アイ・ビー・エム（株）音声情報システム事業推進部）に深謝いたします。

参考文献

[1] R. E. Donovan et al., "Automatic Speech Synthesizer Parameter Estimation Using HMMs," Proc. of ICASSP '95, pp.640-643, 1995.

[2] ニック・キャンベル他、「CHATR:自然音声波形接続型任意音声合成システム」、信学技報SP96-7、1996.

[3] 阪本他、「波形重畳法を用いた日本語テキスト音声合成システムについて」、信学技報SP95-6、1995.

[4] T. Saito et al., "High-Quality Speech Synthesis Using Context-Dependent Syllabic Units," Proc. of ICASSP '96, pp.381-384, 1996.

[5] 都木他、「ピッチ同期音声処理のためのピッチ区間自動区分化の一手法」、信学技報SP93-6、1993.

[6] S. Kadambe et al., "A Comparison of Wavelet Functions for Pitch Detection of Speech Signals," Proc. of ICASSP' 91, pp.449-452, 1991.