社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
SP96-7 (1996-05)

CHATR: 自然音声波形接続型任意音声合成システム

ニック・キャンベル　アラン・ブラック†

ATR 音声翻訳通信研究所
〒619-02 京都府相楽郡精華町光台 2-2
nick@itl.atr.co.jp

あらまし

本報告では、話者性や発話様式の特徴を失わずに任意の音声を合成するための方法として、予め録音された音声データベース中の音素単位の音声波形を、何らの信号処理も行なわずに接続し、連続音声として出力する方法について述べる。本方式は特徴抽出過程、最適重み決定過程、単位選択過程および波形読み出し過程からなり、言語および話者に依存しない。本方式では音声波形自身は生成せず、所望の合成音声の性質に最も近い音素単位の音声波形を、与えられた音声データベース中から取り出すためのインデックスを作成し、音声出力時にはインデックスに従って音声データベースに直接アクセスし、音声波形を取り出す。最適な音素波形の系列を得るために、前処理として、与えられた音声データについて音素記号と音響的および韻律的な特徴の一覧表を作ると共に、それぞれの特徴に対する最適の重み係数を求めておく必要がある。

キーワード • 音声合成　• 信号処理　• 発話様式　• 音声データベース　• 自然な音声

Chatr: a multi-lingual speech re-sequencing synthesis system

Nick Campbell & Alan W. Black†

ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02, Japan
nick@itl.atr.co.jp

Abstract

This paper describes a method for producing speech synthesis without signal processing, using re-sequencing of phone-sized segmets from a pre-recorded speech corpus for the purpose of reproducing the voice characteristics and speaking style of the original speaker to create novel utterances. We describe procedures for indexing and retrieval that make the synthesiser independent of language or speaker. A re-sequencing speech synthesiser doesn't produce speech sounds; it produces an index for a random-access retrieval sequence from the original speech to give the closest approximation to a desired specification from the segments available in a given speech corpus. To find the optimal sequence of segments for concatenation, the synthesiser first creates an inventory of phones and their acoustical characteristics, and then selects from amongst these by a weighted combination of the features to give an index of the segment sequence that best matches the target specification.

Key words • speech synthesis • random access • signal processing • natural speech • multilingual synthesis

†*(currently at Edinburgh University)*

1 はじめに －自然な音声の合成－

CHATR はいろいろな点で従来の音声合成システムと異なっている。これまでの音声合成システムでは、音声は比較的簡単な物理的なモデル化が可能であり、韻律とは切り離して考え、音素の系列から音声波形の合成が可能であるという暗黙の了解に基づいていた。これに対して、本システムは、音声波形は音響的および韻律的な環境によって一意に決まるものであるとの立場を取っており、音声合成システム自身が音声波形を生成することをせず、音響的および韻律的な環境が最も適する音声サンプルの波形をそのまま利用するという手法を取っている [1]。このため、通常の音声合成システムより多くの音声データを必要とするが、出力音声品質は極めて高い自然性を有する。

CHATR のキーワードは自然性である。これは人工的なものと対極をなすものであり、語義から明らかなように自然性は作ることはできず、ただ保存することができるだけである。自然性の高い音声を出力するために、可能な限り自然性を保存し、人間や機械の介在を最小限に抑える必要がある。このような理由から CHATR は音声波形自身を生成するということは行なわず、音声波形間の不連続さを最小化し、目的とする音声に最も近い特徴を持つという規準に基づいて、大規模な音声データベースの中から最適な音素の音声波形を選び出し、順番を並べ変えて出力するという方法を取る。

合成する音声の性質はいろいろな方法で規定することができ、最も高いレベルでは意味論的あるいは語用論的性質を記述する特徴を用いることもでき、やや低いレベルでは個々の音素の音響的および韻律的特徴を用いることもできる。最も低いレベルとしては音響的特徴だけを用いることもできる。

図 1 は従来の音声合成システムと CHATR との比較を示したものである。従来の音声合成システムでは予めシステム内に、単位接続用の音声データと、音声単位接続後の韻律補正を行なうための信号処理部とが含まれていた。これに対して CHATR にはテキスト処理と韻律予測の機能が含まれているだけで、波形に関する情報はすべて外部情報として扱われる。

従来の音声合成システムでは入力テキストから音声波形の生成までが一連の処理として行なわれるのに対して、CHATR では、

- 音声データベースの分析（厳密には音素記号系列の生成、音素アラインメント、特徴抽出を含むが、以下では単に特徴抽出過程という）、
- 最適重み係数の学習（以下では最適重み決定過程という）および
- 音声単位の選択（以下では単位選択過程という）

の 3 つの過程に分割して処理が行なわれる。(1) の特徴抽出過程は新しい音声データベースに対しては必ず一度行なう必要があるが、(2) で求めた最適重み係数は異なる合成条件に対しても再利用が可能である。

本音声合成器は与えられたレベルの入力に基づいて必要とするすべての特徴を予測し、所望の音声の特徴に最も近いサンプルを音声データベースの中から選び出す。最低限、音素ラベルの系列が与えられれば処理は可能であるが、音声基本周波数 (F_0) や音素時間長が与えられていればさらに高品質の合成音声が得られる。なお、入力として単語の情報だけが与えられた場合には辞書や規則に基づいて音素系列を予測する必要がある。また、韻律特徴が与えられなかった場合には音声データベース中のいろいろな環境における音素の既知の特徴を基に標準的な韻律を生成する [5,6,7]。

CHATR では、少なくとも録音内容が正書法で記述されたものがあれば、あらゆる音声データベースが合成用のソースデータとして利用可能であるが、出力音声の品質は録音状態、音声データベース中の音素のバランス等に大きく影響を受け、ソースデータベースが豊富な内容であれば、より多様な音声が合成でき、反対にソースデータベースが貧弱であれば、合成音声は不連続感が強く、ブツブツしたものになる。

2 自然な音声のデータベース

CHATR の前身である ATR ν-Talk [2,3,4] では、約 5,000 語の単語をプロのナレータが読み上げたものをソースデータベースとして用いている。これらは、音素の系列としては十分な量を含んでいるものの、韻律的なバリエーションを最大にする読み方はされていなかった。その結果、これらの単語から選ばれた非均一な音声単位（一つないし複数の音素を含む音声単位）を接続して目的の音声の韻律的特徴を実現する際に大幅な修正を必要とすることになり、音声の自然性に歪みを生じさせ、人工的な音質を生み出す結果となっている。

CHATR では単語音声データベースの代わりに同一話者が読み上げたほぼ同規模の文音声データベースを用いることで豊富なバリエーションを得ることができ、結果として従来の LMA ケプストラム合成方式に代わる方式として自然音声波形をそのままつなぎ合わせる方式が可能になった。すなわち、ソースとなる音声単位の自然なバリエーションを増やすことにより、その後の信号処理の量を減らすことが可能になり、結果として計算量を削減することと自然性を保持することの両方を実現したのである。さらに孤立した文の音声をソースデータベースとする代わりに段落構成を持つ文章を読み上げたものから成る大規模音声データベースを用いれば、まったく信号処理なしにさらに自然性の高い音声が合成できるという見通しを得ている。

2.1 音声データベースのサイズ

以上のことから、音声データベースの規模を大きくすれば、より多くのバリエーションを含む音声が得られ、目的

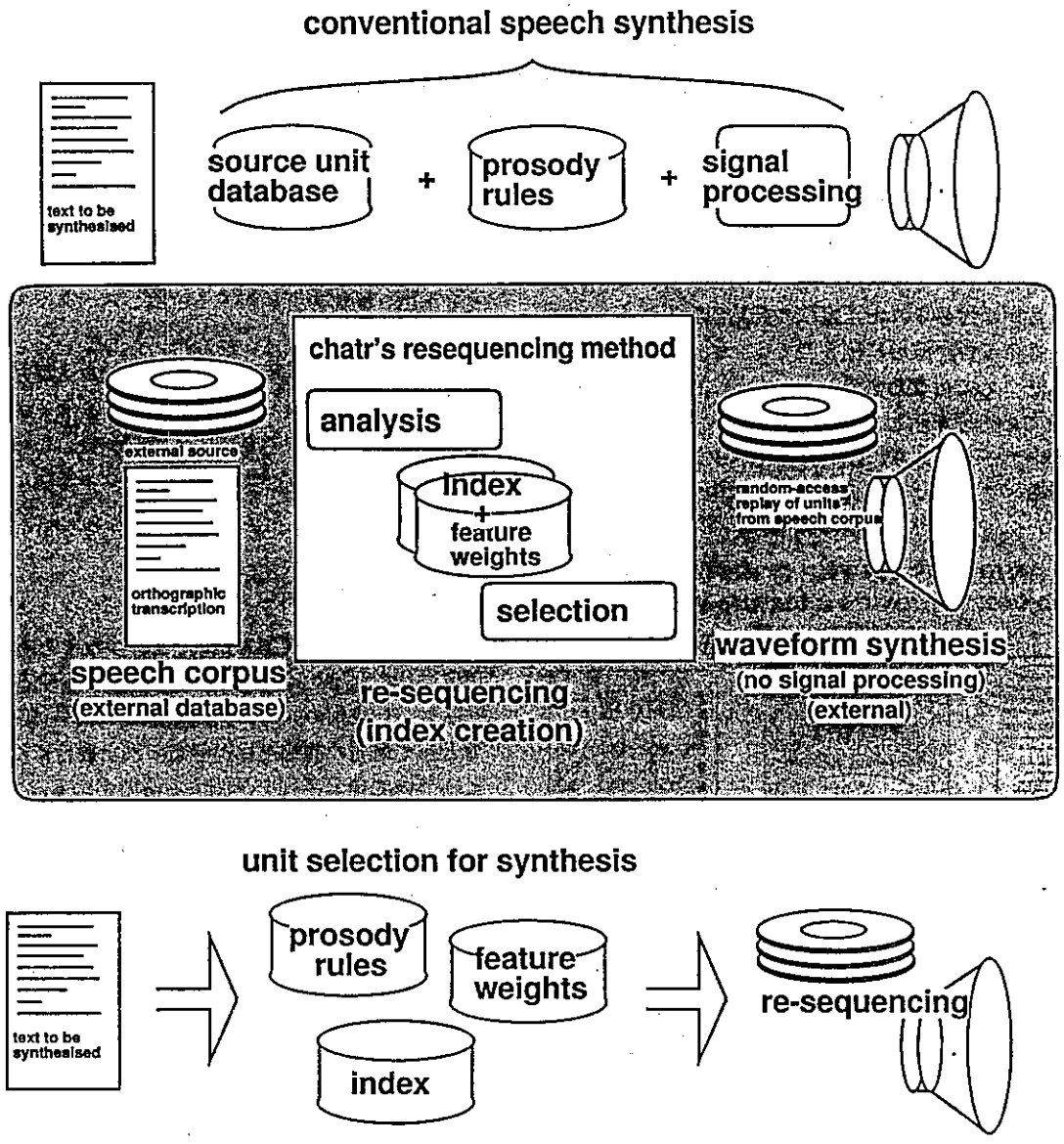

図 1: Overview of the resequencing synthesis system

音声の韻律的特徴に近く、滑らかにつながる音素系列がより簡単に得られることが想像される。日本語については約20分間分の音素バランス文音声が、また他の言語も含めて一般的には約45分間分の音声があれば良いことが経験的に分かっている。最近の例では、約10分間分の音声データでも発話する単語を多少選ぶことにより（もちろん使う単語は音声データベースに含まれるものとは異なっているが）、原話者の話し方に良く似た了解性のある音声が合成できることを確認している。

3 自然な音声に対するインデックシング

音声単位の選択の善し悪しはソースデータベース中の音素のインデックシングと検索の方法に依存する。まず、録音された音声に付与された正書法の発話内容を音素系列に変換し、さらに音声波形に割り当てる。韻律的特徴の抽出はこれに基づいて行なわれる。

図2はこの処理を示したものである。入力は音素表記を伴った音声データであり、出力は特徴ベクトルである。これらの特徴ベクトルは音声データベース中で音声サンプルを表す基本単位となり、最適な音声単位の選択に用いられる。

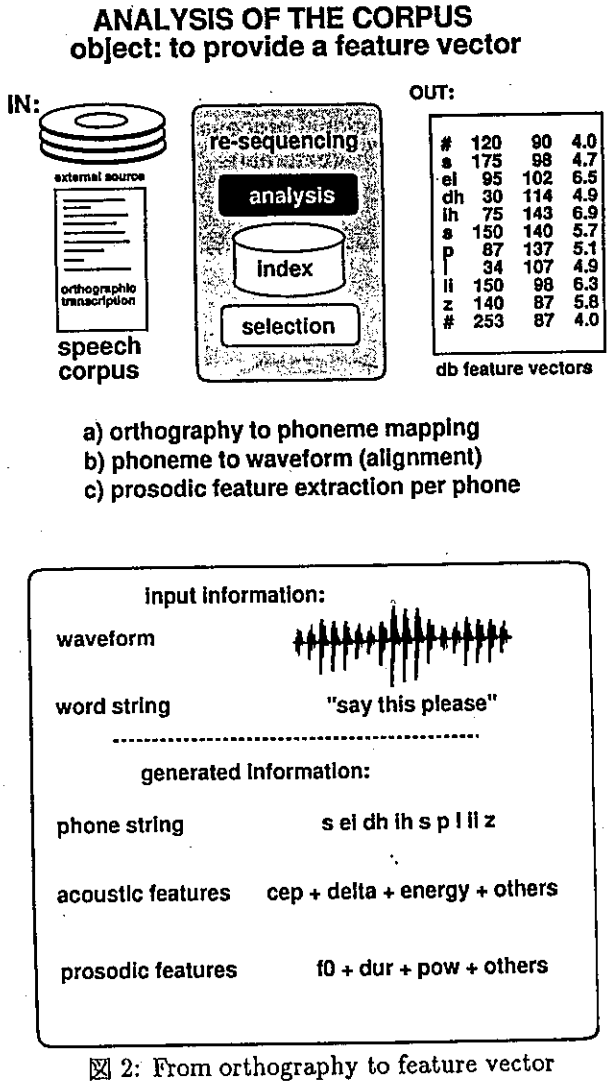

図 2: From orthography to feature vector

第1段階は、正書法で書かれた発話内容が実際の音声データでどのように発音されているかを記述するための正書法テキストから音素記号への変換である。

第2段階は、韻律的および音響的特徴を計測するために各音素の開始および終了時点を決めるべく、各音素記号を音声波形に対応付ける処理である（以下では音素のアラインメントという）。

第3段階は、各音素の特徴ベクトルを生成することである。この特徴ベクトルには、必須項目として音素ラベル、音声データベース中の位置、F_0、音素時間長、パワーの情報が記憶され、さらにオプションとしてストレス、アクセント型、韻律境界に対する位置、スペクトル傾斜等の情報が記憶される。

音声単位を選択する場合に、音響的および韻律的な各特徴がそれぞれの音素でどれだけの寄与をするかを予め調べておくことが必要であり、第4段階では、このために音声データベース中のすべての音声サンプルを用いて各特徴の重み係数を決める。

3.1 音素記号系列の生成

前述した通り、CHATRでは、少なくとも録音内容が正書法で記述されたものがあれば、あらゆる音声データベースが合成用のソースデータとして利用可能である。入力として単語の情報だけが与えられた場合には辞書や規則に基づいて音素系列を予測する必要がある。

3.2 音素のアラインメント

読み上げ音声の場合、各単語がそれぞれの標準の発音に近く発音されることが多く、躊躇したり、言い淀んだりすることもまれである。このような音声データの場合には簡単な辞書検索によって音素ラベリングが正しく行なわれ、音素アラインメント用のHMM（隠れマルコフモデル）の音素モデルの学習が可能となる。

音素アラインメント用の音素モデルの学習では完全な音声認識の場合と異なり、学習用の音声データとテスト用の音声データとを完全に分離する必要はなく、すべての音声データを用いて学習を行なうことができる。まず、別の話者用のモデルを初期モデルとし、すべての単語について標準発音か限られた発音変化のみを許し、適切なセグメンテーションが行なわれるように、全音声データを用いてViterbiアラインメントを行ない、特徴パラメータの再推定を行なう。単語間のポーズは単語間ポーズ生成規則によって処理するが、単語内にポーズがあってアライメントが失敗した場合には人手により修正する必要がある。

どういう音素ラベルを音素表記として用いるかは選択が必要である。もし良く学習されたHMMモデルが利用できるような音素セットが存在するなら、それを用いることが有利である。反対に、音声合成器が完全な辞書を持っ

ているなら、音声データベースのラベルを完全に辞書と照合する方法も有効である。我々はインデックスの作り方に対して選択の余地があるから、後で音声合成器が予測したものと等価なものを音声データベースの中から照合できるかどうかを最も重要な規準とすれば良い。発音の微妙な違いはその発音の韻律的環境によって自動的に把握されるため、特にラベリングを行なう必要はないものと考える。

3.3 特徴抽出

前処理の次の段階として、個々の音素の調音的な特徴を記述するための韻律特徴の抽出を行なう。従来の音声学では、調音位置や調音様式といった素性で言語音を分類した。これに対して（Firth 学派のような）韻律を考慮した音声学では、韻律的文脈の違いから生ずる細かな音質の違いをとらえるために、明瞭に調音されている箇所や強調が置かれている箇所を区別する。これらの違いを記述する方法はいろいろなものがあるが、ここでは 2 つの方法について述べる。

まず低次のレベルでは、1 次元の特徴を求めるために、パワー、音素時間長の伸びおよび F_0 を、ある音素について平均した値を用いる。一方、高次のレベルでは、韻律特徴における上記の違いを考慮した韻律境界や強調箇所をマークする方法を用いる。これらの 2 種類の特徴は相互に密接に関係しているため一方から他方を予測することができるが、両者は共に各音素の特徴に強い影響を与えている。

音声データベースを記述するための音素セットの規定法に自由度があるのと同様に、韻律的特徴の記述方法についても自由度があるが、これらの選び方は音声合成器の予測能力に依存する。もし音声データベースが予めラベリングされているなら、音声合成器の仕事は内部表現から音声データベース中の実音声をいかに行なうかを適切に学習することである。これに対して、もし音声データベースがラベリングされていないなら、どのような特徴を使えば音声合成器が最も適切な音声単位を予測できるかから検討することが必要となる。

3.4 各特徴に対する重みの学習

与えられた目的音声の音響的および韻律的な環境に最適なサンプルを音声データベースから選択するために、まずどの特徴がどれだけ寄与しているかを音素的および韻律的な環境の違いによって決める必要がある。これは音素の性質によって重要な特徴の種類が変化するためで、例えば、F_0 は有声音の選択には極めて有効であるが、無声音の選択にはほとんど影響がない。また、摩擦音の音響的特徴は前後の音素の種類によって影響が変わる。最適な音素を選択するためにそれぞれの特徴にどれだけの重みを置くかを最適重み決定過程で自動的に決定する。

最適重み決定過程で最初に行なわれることは音声データベース中で該当するすべての発話サンプルの中から最適なサンプルを選ぶときに使われる特徴をリストアップすることである。ここでは調音位置や調音様式等の音素的特徴と先行・当該・後続音素の F_0、音素時間長、パワー等の韻律的特徴等を用いる。

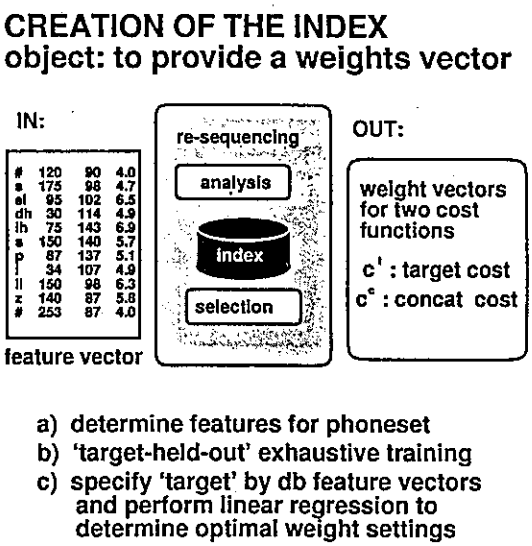

図 3: Training the selection weights

第 2 段階では各音素毎に、最適な候補を選ぶ際にどの特徴がどれだけ重要かを決定するために、一つの音声サンプルに着目し、他のすべての音素サンプルとの音素時間長の差をも含む音響的距離を求め、上位 N 個の類似音声サンプルを選び出す。

第 3 段階では線形回帰分析を行ない、それらの類似音声サンプルを用いて種々の音響的および韻律的環境におけるそれぞれの特徴の重要度を示す重み係数を求める。

図 3 はこれらの処理を模式的に示したものである。なお、これらの処理の詳細については 5 で述べる。

4 自然な音声サンプルの選択

従来の音声合成システムでは目的の発話に対して音素系列を決定し、さらに韻律制御のための F_0 と音素時間長の目標値が計算された。これに対して、CHATR では最適の音声サンプルを適切に選択するために韻律が計算されるだけで、直接韻律を制御することは行なわれない。

図 4 に本システムにおける単位選択過程を示す。この処理の入力は、目的発話の音素系列と、それぞれの音素毎に求めた各特徴に対する重みベクトルおよび音声データベース中の全サンプルを表す特徴ベクトルである。一方、出力は音声データベース中での音素サンプルの位置を表すインデックスで、音声波形を接続するためのそれぞれの音声単位（場合により複数の音素の系列が連続して選択され、一つの音声単位となることがある）の開始位置と長さを示したものである。

最適な音声単位は目的発話との差を表す target cost と、隣接音声単位間での不連続性を表す concatenation (con-

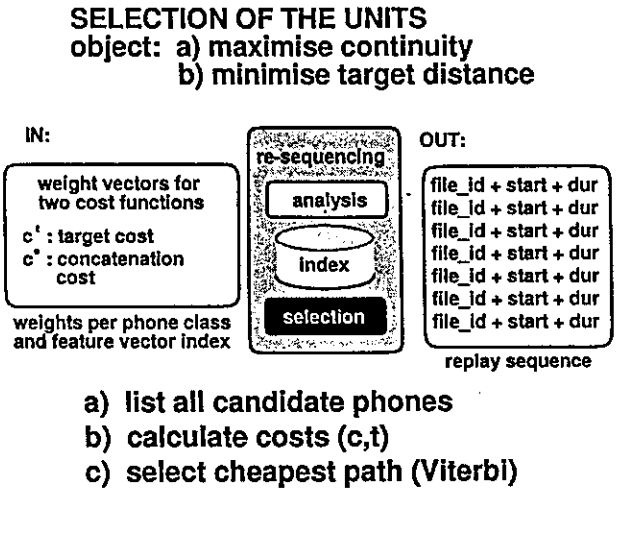

図 4: Selecting the units sequence

cat) cost の和を最小化するパスとして求められる。経路探索には Viterbi アルゴリズムが利用される。目的とする音声 $t_1^n = (t_1, ..., t_n)$ に対しては、target cost と concat cost の和を最小化することで、各特徴が目的音声に近く、しかも音声単位間の不連続性が少ない音声データベース中の音声単位の組合せ $u_1^n = (u_1, ..., u_n)$ を選ぶことができ、これらの音声単位の音声データベース内での位置を示すことにより、任意の発話内容の音声合成が可能になる。

5 音声単位選択コスト

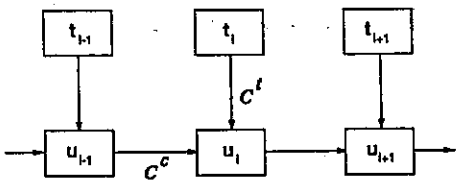

図 5: Minimising two distance measures

Target cost $C^t(u_i, t_i)$ は、音声データベース中の音声単位 u_i と、合成音声として実現したい音声単位 t_i の間の差の予測値であり、concat cost $C^c(u_{i-1}, u_i)$ は接続単位 u_{i-1} と u_i との間の接続で起こる不連続の予測値である [7,8]。

5.1 コストの計算

Target cost は実現したい音声単位の特徴ベクトルと音声データベース中から選ばれた候補の音声単位の特徴ベクトルの各要素の差の重み付き合計であり、各 target sub-cost $C_j^t(t_i, u_i)$ の重み w_j^t が与えられた場合、target cost $C^t(t_i, u_i)$ は次式で計算できる。ここで、特徴ベクトルの各要素の差は p 個の target sub-cost $C_j^t(t_i, u_i)$ (ただし、j は 1 から p まで) で表され、特徴ベクトルの次元数 p は現在のところ 20 から 30 の範囲で可変としている。

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) \qquad [1]$$

一方、concat cost $C^c(u_{i-1}, u_i)$ も同様に q 個の concat sub-cost $C_j^c(u_{i-1}, u_i)$ (ただし、j は 1 から q まで) の重み付き合計で表される。concat sub-cost は接続する音声単位 u_{i-1} と u_i の音響的特徴から決定することができる。現在のところ concat sub-cost としては、

(1) 接続点におけるケプストラム距離、
(2) 対数パワーの差の絶対値、
(3) F_0 の差の絶対値

の 3 種類を用いている。各 concat sub-cost $C_j^c(u_{i-1}, u_i)$ の重み w_j^c が与えられた場合、concat cost $C^c(u_{i-1}, u_i)$ は次式で計算できる。

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i) \qquad [2]$$

もし、u_{i-1} と u_i が音声データベース中の連続する音声単位であった場合には、接続は自然であり、concat cost は 0 になる。

N 個の音声単位の接続コストはそれぞれの音声単位の target cost と concat cost の和となり、次式で表される。

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S) \qquad [3]$$

このとき、S はポーズ を表しており、$C^c(S, u_1)$ および $C^c(u_n, S)$ はポーズから最初の音声単位へおよび最後の音声単位からポーズへの接続における concat cost を表している。この表現からも明らかなように、本システムではポーズも音声データベース中の他の音素とまったく同じ扱い方をしている。さらに上の式を sub-cost で直接表現すると次式のようになる。

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n}\sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) + \sum_{i=2}^{n}\sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S) \qquad [4]$$

単位選択過程は上式で決まる total cost を最小にするような音声単位の組合せ $\overline{u}_1^n$ を決定するためのものである。

$$\overline{u}_1^n = \min_{u_1, ..., u_n} C(t_1^n, u_1^n) \qquad [5]$$

5.2 重み係数の学習

Target sub-cost の重みは音響的距離に基づく線形回帰分析を用いて決める。重み係数の学習過程ではすべての音素毎に異なる重み係数を決めることもできるし、音素カテゴリ (例えば、すべての鼻音) 毎に重み係数を決めることもできる。また、すべての音素について共通の重み係数を

決めることもできるが、ここでは各音素で別々の重み係数を用いることとする。以下に線形回帰分析における処理の流れを示す。

1. 現在学習を行なっている音素（または音素カテゴリ）に属する音声データベース中のすべてのサンプルについて繰り返し以下の4つの処理(a-d)を行なう。
   (a) 取り上げた音声サンプルを目的の発話内容と見なす。
   (b) 音声データベース中の同一音素（カテゴリ）に属する他のすべてのサンプルと当該音声サンプルとの音響的距離を計算する。
   (c) 目的音素に近いもの上位N（例えば、上位20個）を選び出す。
   (d) 目的音素自身と(c)で選んだ上位N個のサンプルについてtarget sub-cost $C_j^t(t,u_i)$を求める。
2. すべての目的音素と上位N個の最適サンプルについて音響的距離とtarget sub-cost $C_j^t(t,u_i)$を求める。
3. 線形回帰分析を行ない、当該音素（カテゴリ）に対してt個のtarget sub-costの線形重み係数を求める。この重み係数を用いて前節で述べたコストを計算する。
4. 1.から3.の過程をすべての音素（カテゴリ）について繰り返す。

もし仮に目的音声単位の音響的距離が直接求められた場合に最も近い音声サンプルを選び出すためにはそれぞれのtarget sub-costにどのような重み係数をかければ良いのかを決定するのが、本学習過程の目的である。本方式の利点は音声データベース中の音声波形が直接利用できることである。

6 システムの実装

CHATRはこれまで4カ国語を含む各種の音声データベースで評価を行なっている。周知の通り、従来は女性話者の音声を用いて高品質の音声を合成することは技術的に大変難しかったが、本手法では性別・年齢などによる差がなくなった。現在のところ、日本語では若年の女性話者が短い物語を読み上げた音声データベースを用いた場合に最も高品質の合成音声が得られている。

一方、ドイツ語については韻律ラベルと詳細な音素ラベルが付与された読み上げ文のCD-ROMデータを用いた合成音声を出力している。これは特別に音声合成用に録音した音声データではなく、本方式が既存の各種の音声データを技術的には自由に利用できることを示している。また、英語ではボストン大学ニュース・コーパスのラジオ・アナウンサの45分間の音声データで最も良い音質が得られている。

なお、韓国語については短い物語の読み上げ音声を用いている。

7 問題点と今後の課題

現在、ほとんどの処理は自動的に行なわれており、完全に新しい話者に対しても録音から合成音声出力までを1日以下の短い時間で行なうことができるが、分析および最適重み決定過程では人手による処理が必要となっている。また、自動音素アラインメントについては現在盛んに改良を進めているところであるが、人手による修正が必要なケースも多く、より多くの手間をかけて修正を行なうと合成音声品質はそれに応じて向上する。しかしながら、人手が介在することにより音素ラベルに誤りが生ずることがあり、これが最適重み決定過程での処理を複雑化させることもある。今後の研究はすべての処理を可能な限り自動化することに焦点を置いて進めることになるが、この際音声認識技術をより有効に活用することができよう。

次の段階として、入力テキストから音声合成の各段階までへの処理における韻律の予測の問題がある。強調箇所や品詞、韻律境界位置等の特徴は音声データベースに記憶されているが、これらの特徴からF_0や音素時間長を予測し、これを使って音声単位選択を行なっていて、冗長な構成となっている。実際にはF_0や音素時間長等も同様にラベリングを行なっているが、この予測過程が誤りの原因となっている。予測機能は完全ではないため、音声データベース中に記憶されている特徴を直接用いて韻律的な環境を特徴付け、中間的な数値の予測を行なわない方法がより賢明な方法であると考えられる。適切な韻律的環境から直接に選ばれた音声単位は我々が予測しようとしているF_0や音素時間長を持っている可能性が高いものと考えられる。これによって、一部のコンポーネントを削除することが可能になり、さらにスリムで高速で簡単なインデックシングと選択の過程が実現できるものと考えられる。

また、CHATRでは音声データベース中の音素タイプのバリエーションがさらに豊富であることが望まれるので、ソースデータベースのサイズが重要な検討項目となる。今後の課題の第3のテーマは全音声データをそのまま記憶するのではなく、その中から豊富なバリエーションを持つサブセットを残して、音声データベースの規模を削減することである。しかしながら、コンピュータは、マルチメディア化に伴ってさらに大容量・高速化する傾向にあり、現時点では優先度は低いと考えられ、音声処理用のデータが画層処理のデータの量を上回るようなら、さらに検討する必要があろう。

8 まとめ

出力音声の自然性を最大にするために、大規模な自然音声のデータベースを用いて処理を最小に抑える方法について述べた。CHATRは3つの基本要素から構成される。

特徴抽出過程 正書法の書き起こしテキストを伴った任意の音声データを入力とし、この音声データベース中のすべての音素について、それらの性質を記述する特徴ベクトルを与える過程。

最適重み決定過程 音声データベースの特徴ベクトルと音声データベースの原波形を用いて、目的の音声を合成する場合に最も適するように音声単位を選ぶための、各特徴の最適重み係数を重みベクトルとして求める過程。

単位選択過程 音声データベースの全音素の特徴ベクトルと重みベクトルと目的音声の発話内容の記述からインデックスを作成する過程。このインデックスに従って、音声データベース中の音声波形に飛び飛びにアクセスし、目的音声波形を出力する。

本方式においては、音声波形の圧縮や F_0・音素時間長の修正は不要になったが、代わって音声サンプルを注意深くインデックシングし、大規模なソースデータベースの中から最適なものを選択することが必要となる。本合成方式の基本単位は音素であり、これは辞書やテキスト-音素変換プログラムで生成されるが、同一の音素であってもソースデータベース中に音素の十分なバリエーションを含んでいることが要求される。音声データベースからの単位選択の過程では目的の韻律的環境に適合し、しかも接続したときに隣接音声単位間での不連続性が最も低い音素サンプルの組合せが選ばれる。このために、音素毎に各特徴の最適重み係数が決められる。

本システムの特徴として、

単位選択規準としての韻律的情報の利用 スペクトル的特徴は韻律的特徴と不可分であるとの立場から、音声単位の選択規準に韻律的な特徴を導入した。

音響的・韻律的特徴の重み係数の自動学習 音素環境や音響的特徴、韻律的特徴等の各種の特徴量が音声単位の選択にどれだけの寄与があるかを音声データベース中の全音声サンプルを利用することで自動的に決定し、コーパスベーストな音声合成システムを構築した。

音声波形の直接接続 上記の自動学習により、大規模音声データベース から最適な音声サンプルを選び出すことにより、何らの信号処理も利用しない任意音声合成システムを構築した。

音声データベースの外部情報化 音声データベースを完全に外部情報として扱うことにより、単にCD-ROM等に記憶した音声データを取り替えることで任意の言語、任意の話者に利用できる音声合成システムを構築した。

等が挙げられる。

CHATRを用いた場合、極めて高い自然性を保持した合成音声が得られることは既に確認されているが、本方式をある程度限定したドメインに利用した場合にはさらに有効性を発揮するものと考えられる。今後さらに処理の自動化・高速化を行なうと共に、より多くのバリエーションを含む音声の収録方法およびその利用方法に検討を進めていく予定である。

謝辞

The authors would like to take this opportunity to express their sincere thanks to Dr. Norio Higuchi (Department Head, Dept 2, ITL) for helpful discussions (and especially for his assistance and advice on the use of Japanese, without which this paper could not have been written in its present form), and to Dr. Yoshinori Sagisaka (Department Head, Dept 1, ITL) for his continuing advice and inspiration.

参考文献

[1] W. N. Campbell. Synthesis units for natural English speech. Technical Report SP 91-129, IEICE, 1992.

[2] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. ATR ν-talk speech synthesis system. In *Proc. 1992 Intl. Conf. on Spoken Language Processing*, pages 483–486, Banff, Canada, 1992.

[3] Iwahashi, N. Kaiki, and Y. Sagisaka. Concatenative speech synthesis by minimum distortion criteria. In *ICASSP '92*, pages II–65–68, 1992.

[4] Y. Sagisaka and N. Iwahashi. Objective optimisation in algorithms for text-to-speech synthesis. In *Spech Coding and Synthesis*, W. B. Klein & K. K. Paliwal, Eds., Elsevier Science B. V. 1995.

[5] W. N. Campbell. Prosody and the selection of source units for concatenative synthesis. In Proc 2nd ESCA Workshop on Speech Synthesis, Mohonk, N.Y., 1994.

[6] A. W. Black and W. N. Campbell. Optimising selection of units from speech databases for concatenative synthesis. In *EUROSPEECH '95*, Madrid, Spain.

[7] W. N. Campbell and A. W. Black. Prosody and the selection of source units for concatenative synthesis. In J. van Santen, R. Sproat, J. Olive, and J. Hirshberg, editors, *Progress in Speech Synthesis*. Springer Verlag, 1996. (forthcoming)

[8] A. J. Hunt and A. W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In Proc. Int'l. Conf. on Acoust., Speech and Signal Proc., 1996. (forthcoming)