高品質日本語テキスト音声合成システムの開発

5G－5　小林俊一† 片江伸之† 松本達郎† 木村晋太† 加世田光子‡ 大山隆之‡
†（株）富士通研究所　‡富士通株式会社

1 はじめに

音声合成は人間と機械の自然なインタフェースとして注目されており，当社では，NIFTY-Serveでの電子メール読み上げサービスや，視覚に障害のある方を対象とした，ディスプレイ上の文字読み上げシステムFMTALK-IIなどの応用例がある[1]。しかし，これらは音声信号処理用のハードウェアを必要としていた。

ここ数年，パーソナルコンピュータの処理能力が著しく向上し，実時間での音声処理が可能となってきた。また，メモリが非常に安価になったため，システムの蓄積データを圧縮する必要性が小さくなった。このような背景のもとで，全ての処理をソフトウェアで行ない，かつ高品質な音声の合成が可能なテキスト音声合成システムを開発したので，報告する。本システムの技術的な特徴は以下の3点である。

1. ＤＰ照合法と詳細二方向文法による形態素解析
2. 折れ線モデルによる基本周波数パターン生成
3. 波形編集方式による音声波形生成

2 システムの概要

本システムの処理の流れを図1に示す。本システムは，言語処理部，韻律制御部，波形生成部からなっており，その仕様は表1のとおりである。

表1 本システムの仕様

サンプリング周波数	16kHz
量子化ビット数	16bit(または8bit)
プログラムサイズ	C言語約4万行
単語辞書（１２万語）	約5MB
波形辞書（男女声フルセット）	約8MB

2.1 言語処理部

漢字かな混じり文から表音文字列を生成する。形態素解析の方法としてＤＰ照合法を採用している[2]。本方式は，解析単位ごとに，可能性のある全ての単語・形態素の組合せを候補として抽出したのち，種々の規範を評価値として，最適な単語・形態素系列を決定するものである。利用する規範は，(1)文法上の整合性，(2)文節数最小原理，(3)最長一致原理，(4)単語の重要度・頻度などであり，(1)に詳細二方向文法を用いている。これは，ひとつの単語の前後の単語への接続特性を前後別々に持たせたものであり，動詞，名詞などの品詞をさらに細分化して記述している。

さらに言語処理部では，規則による係り受け解析を行ない，アクセント句／フレーズ／呼気段落の境界と文節アクセントを決定する。

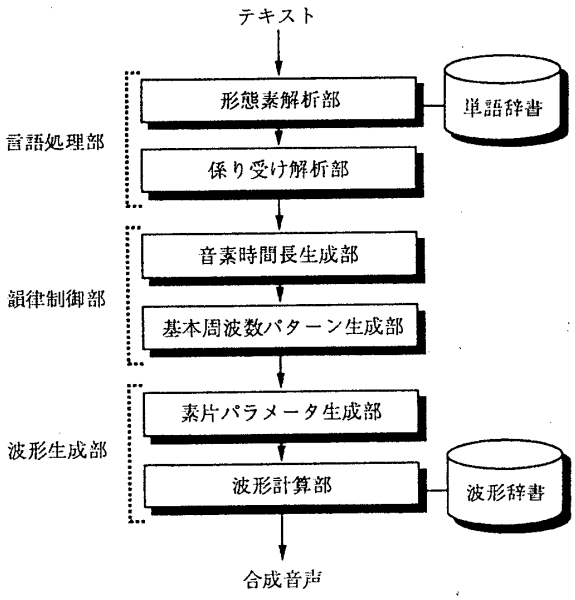

図1 本システムの処理の流れ

2.2 韻律制御部

表音文字列から音響パラメータを生成する。

音素時間長は，自然音声データ１５０文より作成したＣＶ環境音素時間長テーブルを検索し，発声速度と呼気段落長による修正を施すことによって生成する。

基本周波数パターン（対数尺度）は折れ線モデル（図2）により生成する。なだらかに下降するフレーズ成分，台形状のアクセント成分，文末成分を加算することによって基本周波数パターンを生成する。各成分の形状

A Development of High-quality Japanese Text-to-Speech System
Syun-ichi KOBAYASHI†, Nobuyuki KATAE†, Tatsuro MATSUMOTO†, Shinta KIMURA†, Mitsuko KASEDA‡ and Takayuki OHYAMA‡
† Fujitsu Laboratories Ltd., ‡ Fujitsu Ltd.

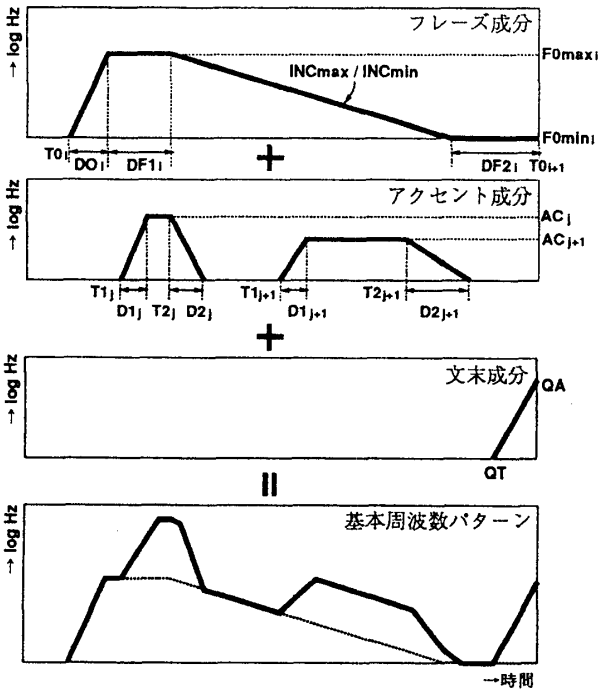

図 2 基本周波数パターンの生成モデル

は図中に示すパラメータにより制御する。

2.3 波形生成部

音響パラメータから音声波形を生成する。合成方式には波形編集方式を採用した。本方式は，あらかじめ自然音声から切り出した素片波形（1ピッチ周期程度の短い波形）を波形辞書に蓄積しておき，音響パラメータに応じて逐次なめらかに接続するものである。

波形辞書に蓄積した素片波形データは3音素連鎖環境を網羅した無意味単語リストを，男女1名ずつのアナウンサが発声した音声データから，音声波形の視察により抽出した。3音素連鎖環境は150音節，39音素について，7737環境からなる。また，素片波形データは，音素の開始部，定常部，終了部の3箇所より抽出した。

周期性波形の処理を図3に示す。合成窓を用い，ピッチ周期ごとに重畳し接続することによって，接続時の不連続を軽減し，なめらかな合成音声を生成している。合成窓は中央部分が平坦で，両端が固定長(=1msec)のハニング窓の関数とし，ピッチ周期に応じて平坦部分を伸縮する。二つの蓄積波形間の補間には，これらの二つの蓄積波形に重み付けしたのち，足し合わせることによって補間波形を生成する。

本方式は線形予測モデルによる方式に比べ，自然かつ明瞭な合成音声が得られる。

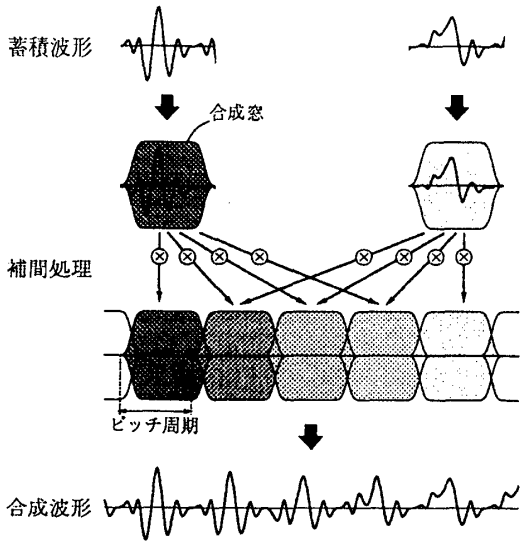

図 3 波形生成部の処理

3 評価

以下の文章を用いて性能評価を行なった。

- 小説「カッコーはコンピュータに卵を生む」
- 当社ニュースリリース
- 技術書「はじめてのC」
- 評論（大学入試問題より）
- 雑誌「日経コンピュータ」

（計 4464 文字。2759 単語。92 文。）

別途行なった，単音節明瞭度試験の結果とあわせて表2に示す。

表 2 本システムの性能（カッコ内は当社従来製品）

読み正解率		99.8%	(99.0%)
アクセント正解率		95.9%	(84.7%)
単音節明瞭度	男声	88.9%	(83.2%)
	女声	73.4%	(66.9%)

4 まとめ

全ての処理をソフトウェアで行ない，かつ高品質な音声の生成が可能なテキスト音声合成システムを開発した。本システムは当社の TownsOS V2.1 L31 に標準搭載されている。今後は，声質の多様化，また合成方式の改良による品質の向上を計るとともに，音声を用いた新しいアプリケーションへの応用について検討したい。

参考文献

[1] 「富士通の音声合成技術の紹介」音響学会誌 49(12)

[2] 神山 他「日本語音声合成における文章解析部の検討」音響学会講演論文集 3-6-15 (1987,3)