自然言語処理 125－2
音声言語情報処理 21－2
（1998. 5. 28）

SIG－SLP／SIG－NL合同セッション

ここまでできるぞ音声／言語処理技術

－ 音声編 －

新田 恒雄(東芝) 小林 哲則(早稲田大学) 鹿野 清宏(奈良先端大学) 武田 一哉(名古屋大学) 河原 達也(京都大学) 伊藤 克亘(電総研) 峯松 信昭(豊橋技科大学) 伊藤 彰則(山形大学) 宇津呂 武仁(奈良先端大学) 山本 幹雄(筑波大学) 山田 篤(ASTEM) 西村 雅史(日本 IBM) 甲斐 充彦(豊橋技科大学) 中川 聖一(豊橋技科大学) 服部 浩明(NEC) 阿部 匡伸(NTT) 松浦 博(東芝)

マルチメディア時代が到来し，様々なサービス提供が始まっている。本報告では，今後，ますます重要性を増す音声インタフェース技術に焦点をあて，音声認識および音声合成を中心とした最新技術を紹介している。 内容は，音声認識技術として，日本語ディクテーションソフトウエア，Web 検索ソフトウエア，大語彙音声認識チップを，また音声合成技術として，音声コンテンツ制作支援ツール，テキスト－音声変換ソフトウエアから成る。

SIG-SLP/SIG-NL Joint Session

"Recent Advances in Speech and Language Processing Technologies"

- Speech Processing Technologies -

Tsuneo NITTA (Toshiba), Tetsunori KOBAYASHI (Waseda Univ.), Kiyohiro SHIKANO (NIST), Kazuya TAKEDA (Nagoya Univ.), Tatsuya KAWAHARA (Kyoto Univ.), Katunobu ITOU (ETL), Nobuaki MINEMATSU (Toyohashi Univ. Tech.), Akinori ITO(Yamagata Univ.), Takehito UTSURO (NIST), Mikio YAMAMOTO (Tsukuba Univ.), Atsushi YAMADA (ASTEM), Masafumi NISHIMURA (IBM Japan), Mitsuhiko KAI (Toyohashi Univ. Tech.), Seiichi NAKAGAWA (Toyohashi Univ. Tech.), Hiroaki HATTORI (NEC), Masanobu ABE (NTT), and Hiroshi MATSU'URA (Toshiba)

Computer-human interaction by voice is one of the most important technology in the coming multimedia era. In this report, we introduce recent advances in speech processing technologies through focussing both of speech recognition and speech synthesis. Contents are: Japanese dictation software, Web-page retrieval software, large-vocabulary speech recognition chips, a speech editing tool for designing multimedia applications, and TTS (Text-To-Speech) software for PCs.

1．　音声処理技術の動向

マルチメディア時代が到来し，様々なサービスの提供が始まっている。しかし個々のサービスには，それに見合った利用技術，すなわちヒューマンインタフェース（ＨＩ）技術が必要になる。音声は，文字出現以前から利用された対話手段と考えられる。しかし，コンピュータとの対話に利用されるようになったのは，ごく近年になってからである。以下では，音声処理技術の概要を述べるとともに，２章に紹介する音声認識および音声合成を中心とした最新技術を鳥瞰する。

1．1 音声処理技術の概要

表 1.1 に音声処理技術とそれらの応用をまとめて示した。音声を圧縮・伸張する音声符号化は，古くから通信の中心技術として研究開発が進められてきた。音声合成においても，この符号化技術は様々な音声応答装置に利用されている。また，マルチメディア利用環境が普及するにつれ，テキストからの音声合成（文－音声変換）技術も利用が増えている。概念からの合成は，対話システムなど柔軟性を要求される場面に必要となる技術で，現在は研究段階にある。

音声認識では，単語認識技術が様々な応用システムのコマンド入力や，データ入力装置に利用されている。さらに近年は，ディクテーションの実用化も始められている。本格的な音声対話では自由発話を前提とするため，話者自身から出る非言語音の問題，助詞落ち・倒置など話し言葉のゆらぎの問題，さらに言い淀みなど，解決すべき課題がまだまだ多い。

音声処理技術にはこのほか，話者認識と音声加工／復元がある。話者認識はセキュリティが重視されるにつれて，今後様々なシステムに利用されるようになると考えられる。また，話速変換など

表 1.1 音声処理技術とその応用

技　術	用　途
音声符号化	通信，放送，蓄積メディア
音声合成	音声出力（応答），文－音声変換
	音声対話（概念からの合成）
音声認識	コマンド入力，データ入力
	ディクテーション，音声対話（自由発話）
話者認識	個人照合（話者照合）
	個人同定（話者識別（同定））
音声加工／復元	エコーキャンセラ，ノイズ除去
	話速変換，話者変換

も，高齢者のために発話速度を聴きやすい速度に変換するなど，今後重要になる技術といえる。

1．2　音声処理の最新技術紹介

２章では，2.1 – 2.4 で音声認識処理技術を，また 2.5 – 2.6 で音声合成処理技術を紹介する。

近年，音声と言語の大規模データベースを利用して，確率統計的モデルに基づく大語彙連続音声認識が実用に供されるようになった。2.1，2.2 は，この代表的な日本語ディクテーションソフトウエアを紹介する。また，2.3 は，同様のソフトウエアをクライアント・サーバ環境下で，Web 検索に利用した例を述べる。

携帯機器の利用が広がるとともに，音声入力を端末に搭載することへの要望が増している。2.4 は汎用 DSP (Digital Signal Processor) や RISCチップ単体で，音声認識機能を実現した例を紹介する。

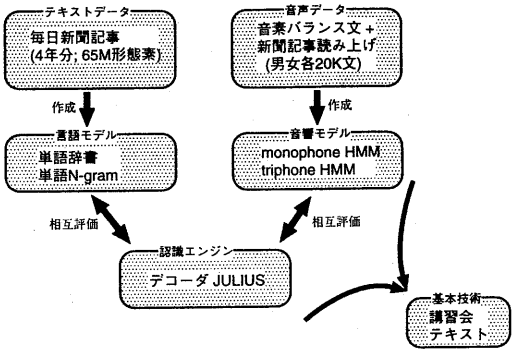
マルチメディア作品やゲームなどに，音声合成を応用できると制作も大変楽になる。2.5 は，このような音声コンテンツ制作支援ツールの例を示している。また，テキストを音声に変換するソフトウエアは，現在，多くのPCに搭載され，画面上のエージェントと同期して動作するものも提供されるようになった。2.6 ではこうした例を紹介する。

2 音声処理技術の紹介

2.1 日本語ディクテーション基本ソフトウェア − 97 年度版 − *

概要

- 大語彙連続音声認識 (LVCSR) 研究・開発の共通プラットフォーム
- 複数の大学・公的研究機関の協力プロジェクト[1]
- 無償で一般に公開[2]



1 音響モデル

- 混合連続分布 HMM(対角共分散)
- HTK のフォーマット

	状態数	混合分布数
monophone	129	4, 8, 16
triphone 1000	1000	4, 8, 16
triphone 2000	2000	4, 8, 16
triphone 3000	3000	4, 8, 16

2 単語辞書

- 5000 語の辞書 (カバレージ 85.8%)
- HTK のフォーマット

3 言語モデル

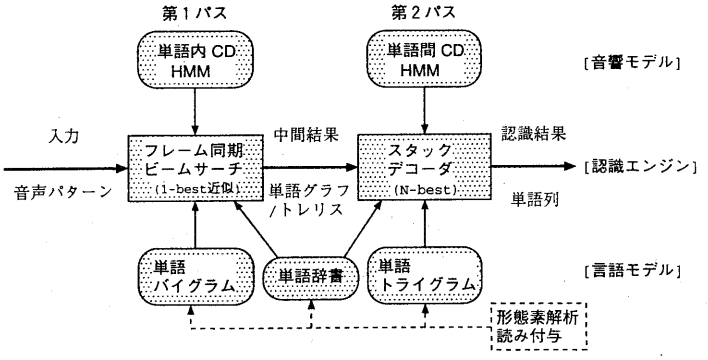
- 単語 2-gram と 3-gram
- CMU-Cambridge SLM-TK のフォーマット

4 認識エンジン JULIUS

- 3-gram / 2 パス デコーダ
- Sun, SGI, DEC, PC Linux で動作

5 5K 日本語ディクテーションシステム

音響モデル	monophone 129x16	triphone 3000x8	triphone 2000x16
デコーディング	candidates reduced	small beam	large beam
認識時間	3x RT	6x RT	12x RT
認識精度 (男性)	85.2	91.3	92.8
認識精度 (女性)	87.3	91.2	93.2



*本ソフトウェアは、情報処理振興事業協会 (IPA) が実施した独創的情報技術育成事業の研究成果である

[1] http://www.itakura.nuee.nagoya-u.ac.jp/~takeda/IPA/

[2] http://www.lang.astem.or.jp/dictation-tk/

仕様と性能の詳細は下記を参照

河原達也, 他. 日本語ディクテーション基本ソフトウェア (97 年度版) の性能評価. 情処研報, 97-SLP-21-8, 1998.

2.2 日本語ディクテーションソフトウェア ViaVoice

2.2.1 ディクテーションに関する研究の現状

1990年以降，欧米では統計的言語モデルを用いたPC用のディクテーションソフトウェアが実用化され，現在では医療所見や判例の入力といった特定分野から，徐々にではあるが個人ユーザーが日常的な文章を入力する手段へと市場が広まっている．そして研究の対象も，読み上げ文からニュース音声の書き起こしなどのより自然な発話へと次第に移行している．

一方，日本語のディクテーションに関する研究は，当初大幅に遅れたものの，最近では日本語と欧米系言語との性能差はなくなりつつある．

2.2.2 日本語版VoiceType/ViaVoice開発の経緯

IBMでは日本語のディクテーションシステムの研究開発にあたり，まず，日本語特有の制約，特に単語の単位が不明確で離散発声に適さないとの制約を外し，欧米言語と同じ土俵の上で検討が出来る枠組みを模索した．その結果得られたのが，形態素解析プログラムの出力である形態素と，日本人が感覚的に捉えている"単語単位"との対応関係を統計的なモデルで表現し，それによって日本語の単語単位（=発声の最小単位）を自動推定する方法[1]である．そして，この単位を使えば，日本語でも離散単語発声によるディクテーションシステムが，欧米語と同様にN-gram言語モデルと音素HMMによって構成される認識システムとして実現できることを示した．それが1996年11月に発表された「VoiceType Dictation日本語版」である．その後，この単位は離散発声が可能なだけではなく，既存の形態素単位などと比較した場合にも，その単語カバレージは十分に高く，一方で単位あたりのモーラ長が長く表記あたりの読みの種類も制限され，音響的に識別しやすい単位となっていることが分かった[2]．

またこの単位は，離散発声可能な発声の最小単位でもあることから，連続音声認識に適用した場合でも自由なポーズの挿入が可能となるなど，連続音声認識の認識単位としても都合がよい．その結果，1997年11月には日本語の連続発声のディクテーション機能を有する「ViaVoice Gold日本語版」が発表された．

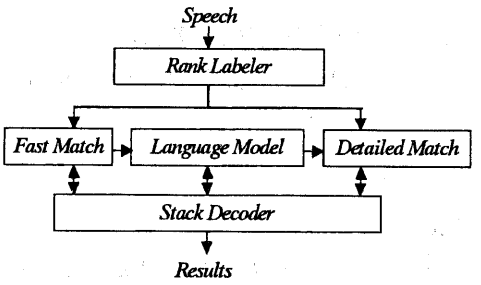


図2.2.1 認識エンジンの構成

2.2.3 ViaVoiceの構成

ViaVoiceは主にアプリケーションの制御などに用いられるナビゲーション機能と，非定型文章の入力を実現するディクテーション機能から構成されている．前者では認識対象語彙をフォーカスのあたっているアプリケーションに応じて動的に決定するか，あらかじめ文法によって記述しておく．一方，後者では一般的な日本語文章の入力用に用意された約44,000語が認識対象となる．なお，この際にはTri-gram言語モデルによる言語制約を用いている．

ViaVoiceの認識エンジンは図2.2.1に示すように，ランクラベラー，ファーストマッチとディーテイルマッチを併用したスタックデコーダーおよび言語モデルなどから構成される．音響モデルは混合正規分布で表現された音素環境依存型のHMMである．言語モデルの学習には，主に新聞記事からなる約500万文のテキストデータを用いた．一方，不特定話者用の音響モデルは20代～60代の男女約1,000名の連続発声データから推定した．

VoiceTypeはPentium-133MHz，一方，ViaVoiceはMMX Pentium-200MHzを搭載したPC上でほぼ実時間動作が可能である[3]．なお，現在ViaVoiceは８カ国語（日本語，US英語，UK英語，ドイツ語，フランス語，イタリア語，スペイン語，中国語[4]）版が同時出荷されているが，認識エンジンはすべて共通のものを使用している．

2.2.4 今後の展望

口語体への対応，より自然な発話への対処など解決すべき問題も多いが，放送音声や音声メモの書き起こし，NLU(Natural Language Understanding)システムへの入力など，今後この技術に基づいて実現が期待されるアプリケーションは数多い．

---

[1] 西村，伊東，"単語を認識単位とした日本語ディクテーションシステム，" 電子情報通信学会論文誌，Vol.. J81，D-II，No.1， pp.10-17 (1998.1)

[2] 西村ら，"単語を認識単位とした日本語の大語彙連続音声認識" 情報処理学会，SLP-20-3, pp.17-24 (1998.2)

[3] いずれも日本語版の場合．

[4] 中国語版のみ若干仕様が異なる．

2.3 音声認識公開ソフトウェアとWeb検索への応用

2.3.1 音声認識ソフトウェア*, **

我々は、対話型アプリケーションの構築を容易にするために、クライアント・サーバ構成で利用できる汎用的な連続音声認識サーバ–SPOJUS–を開発、公開している*,**。本システムは、音声認識機能を利用するアプリケーションとネットワークを介して連携させることができる。例えば、対話処理、音声合成処理などの他の処理モジュールとの連携が可能であるが、ここではその一例として、WWWブラウザの音声操作システムの試作例を示す。まず、SPOJUSの概要を説明する。

特徴

- ネットワークを介した利用および単体での利用
- 文法・辞書の記述に基づく認識対象の変更
- 辞書記述に基づく未知語検出機能の設定
- 棄却処理のための信頼性スコアの出力

音声入力・分析サーバ

音声入力・分析サーバは、マイクおよびパソコン内蔵サウンドカード等のA/D変換機能を通して入力された音声を10次元のメルケプストラム係数の特徴ベクトル系列へ変換し、フレーム毎に音声認識サーバへ送信する。

音声認識サーバ

音声認識サーバは、文脈自由文法に基づく言語処理を統合したフレーム同期型のビームサーチアルゴリズムを採用している。音声入力・分析サーバからのフレーム単位の入力に応じて処理が進行する。

不特定話者用の対応のため、音響モデルとしては日本語で主に用いられる113音節分の音節単位のモデルを用意している。これらのモデルは、ATR研究用日本語音声データベースと日本音響学会連続音声データベースの合計36名の話者の音声データ（但し男性話者のみ）により学習されている。

ユーザが入力可能な文の制約は、文脈自由文法に基づいて記述する。文法と単語辞書は、文毎に動的に変更できる。

2.3.2 応用システム例（ブラウザ音声操作）***

一般的なWWWブラウザの操作では、リンク先へのジャンプやページの切替え、スクロール等の操作を全てマウスで行なうことができる。前述の音声認識サーバを応用して、これらの操作を音声入力によって実現するシステムを試作した。

音声インタフェースとしてユーザが可能な発話様式は、リンク先指定の発話と表示切替えのコマンド

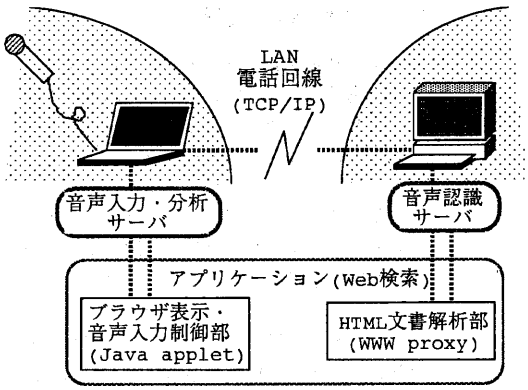


図2.3.1: 応用システム例（ブラウザ音声操作）

発話である。ユーザが、閲覧するページを開くかリンク先へジャンプすると、そのページの表示とは別にリンク先一覧をブラウザ内の別のフレームに表示する。ここで、リンク先指定の発話の場合、ユーザはリンク一覧の番号、またはリンク（下線部分）に対応するキーワードと、「です」、「のホームページ」などを付加した発話を入力できる。入力内容に応じて、発話に近いキーワードを表示するか、またはURLへ変換して新たなページの表示へ自動的に切替える。

リンク先に対応するキーワードは一般に単語系列からなる。そこで、形態素解析システム****により妥当なキーワード断片の単位を自動的に抽出する。また、閲覧するページが切替わる毎に、得られたキーワード単位と読みの情報に基づいて、文法および辞書を再生成する。

試作システムのアプリケーション部は、ブラウザ表示・音声入力制御部と、HTML文書解析部のコンポーネントで構成される。前者は、Java appletとCGI(Common Gateway Interface)に基づいて実装されており、Javaが動作する一般的なブラウザを利用できる。

2.3.3 今後の課題など

音声認識ソフトウェアと応用例について述べた。音声認識ソフトウェアは、対話型システムに応用可能で、マルチモーダル音声対話システムも開発されている*****。言語モデルとして文脈自由文法を利用しているため、大規模な対話型システムの構築には人手が必要であるが、応用例のような小規模なタスクでは少数の定型的な文型があらかじめ定まるため、容易に実現できた。

現在の試作システムは、対話的な機能を持っていないため、複数の候補からの選択、フォームの入力などの機能が不十分である。また、連想辞書や過去の履歴の利用等によって関連のあるWebページを効率的に検索する方法を検討することも今後の課題である。

*http://www.slp.tutics.tut.ac.jp/SPOJUS/.
**[甲斐, 伊藤, 山本, 中川] 音講論集, 2-Q-30 (1997.9).
***[甲斐, 中野, 中川] SLP20-14 (1998.2).

****JUMAN ver3.2 を使用.
(http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/)
*****[中川, 傳田, 伊藤] 人工知能学会誌, Vol.13, No.2 (1998)

2.4 大語彙音声認識チップ

近年、携帯電話、カーナビゲーションシステム等の携帯情報端末において音声認識機能を搭載するものが市場に登場して来ている。これらの機器においてはキーボードやマウスといった従来のマンマシンインターフェイスを用いることが困難であり、音声という人間にとって自然で容易なコミュニケーション手段をインターフェイス手段として利用できるようになった意義は大きい。

このような音声インターフェイスが実現した背景には音声認識技術と半導体製造技術の両技術領域での進歩が上げられる。ここでは、NECの大語彙音声認識技術について述べるとともに、音声認識チップを紹介する。

2.4.1 音声認識方式

大語彙音声認識を実現する場合、単語を認識単位とすると学習用語彙の発声、収集が困難となるため、単語よりも小さな単位を用意し、これを連結することで単語、文を作成する方法が一般に用いられる。

NECでは「半音節」と呼ぶ独自の単位を用いている[1]。半音節は単音節をその母音中心で分割した単位となっており、知覚にとって重要な子音から母音へのわたりを含んでいるため、比較的少数の認識単位により高い認識性能が得られるのが特徴である。

この半音節の音響的特徴を多数の話者の音声から抽出、用意しておき、平仮名等で表記された発音に従って連結することで、不特定話者の任意単語の認識が可能となる。

人名、地名等の入力を考えると１０万単語以上の大語彙認識が必要となる。この際、入力と単語パターンとのマッチング処理が増加しリアルタイム動作が困難となる。そこで、木構造辞書とビームサーチによる効率のよいマッチング手法を開発[2]し、１０万単語以上の大語彙であってもリアルタイムの認識動作を可能とした。

また、携帯情報端末は様々な場所、雑音下で用いられるため、周囲環境によらずに高い認識性能が得られなければならない。あらかじめ、周囲環境が既知である場合にはその環境での音声を収集することで、対処することができるが、モバイル環境では時々刻々と周囲環境が変わるため、このような手法では対処困難である。そこで、認識単語そのものから周囲環境の騒音等を推定、瞬時に適応化する手法[3]を開発し、種々の雑音下での高い認識性能を実現した。

2.4.2 大語彙音声認識チップ

現在、入手可能なNECの大語彙音声認識チップを紹介する。これらはどちらも半音節認識単位による音声認識アルゴリズムを搭載している。目的、用途にあわせてDSPと汎用RISCチップが用意されている。

DSPベースの認識チップは消費電力が小さい、外付け回路が小さい等のメリットがあり、携帯電話等のコンパクトかつ低消費電力を求められる用途に適している。一方、汎用RISCチップベースの方はリアルタイムOSの採用により高いプログラマビリティを有し、多機能を求められる情報端末に適している。

- μ PD77524

  DSPベースの音声認識チップ[4]。単語登録は平仮名表記を与えるだけでよく、音声登録は不要。一度に不特定話者の１０００単語(５音節／単語換算)の認識が可能である。２マイク入力による雑音抑圧処理を搭載しており、雑音環境下でも高い認識性能が得られる。また、音声応答のためのADPCM音声再生機能を内蔵しており、１チップで音声対話処理が実現できる。

- ULTALKER-V

  汎用RISCチップV830[5]ファミリの音声認識ミドルウェア。木構造辞書とビームサーチアルゴリズムの採用により、１０万単語以上の大語彙をリアルタイムで認識可能。μ PD77524と同様の耐環境アルゴリズムを搭載、種々の環境で利用可能。リアルタイムOS RX830上で動作し、用意されているエコーキャンセラ、CODEC、音声合成等のミドルウェアライブラリにより、１チップで快適なマルチメディア環境を実現することができる。

2.4.3 今後の課題

以上、携帯情報端末における自然なマンマシンインターフェースを実現するための大語彙音声認識チップについて述べた。今後、より自然な発話認識、より頑健な耐環境技術の開発を行い、音声認識、自然言語処理、音声合成の融合を図り真に人に優しいインターフェイスの実現を目指して行く。

---

[1] 磯谷他、音講論集 1-8-19 (1990.9)

[2] 服部他、音講論集 2-6-58、(1997.3)

[3] 高木他、音講論集 2-P-8、(1994.3)

[4] NEC 資料 S10876JJ4V0DS00)

[5] NEC 資料 SUD-T-2183

2．5 音声合成による音声コンテンツ制作

コンピュータの発展に伴って、コンピュータが人間の創造活動を支援することが増えてきた。例えば、デスクトップミュージック（DTM）と呼ばる電子楽器による音楽制作や、コンピュータグラフィックス（CG）による映像制作等がそれである。同様なことを音声で行うことが音声合成による音声コンテンツ制作であり、その１つの試みとして *Speed97*（Speech editor）[1]を開発した。

Speed97 は、音声メッセージの作成を支援するツールである。テキストを入力として、GUI（グラフィカルユーザインタフェース）によってパラメータをエディットしながら、音声を作成していく。システム構成を図１に示す。図１のテキスト解析部、韻律パラメータ生成部、音声合成部は、入力されたテキストを処理して、規則を適用することによって音声を自動的に生成する。これは、テキストからの音声合成と呼ばれる技術である。*Speed97* のユーザは、自動合成された音声を出発点として、韻律パラメータのエディット、音声の合成、合成音の視聴の３つのステップを繰り返しながら音声を作成する。

これまでの実験によれば、*Speed97* によって、音声ガイダンスや情報アナウンスのための音声メッセージばかりでなく、方言や感情を込めたせりふ[2]等を制作できることが明らかとなっている。音声品質の点では声優らが発声した音声には及ばないものの、用途によってはマルチメディアコンテンツやゲームで利用できると考えられる。さらに、*Speed97* で作成された音声は、音声信号と音素記号等との対応が明確になっているため、自然音声にはない次のようなメリットがある。

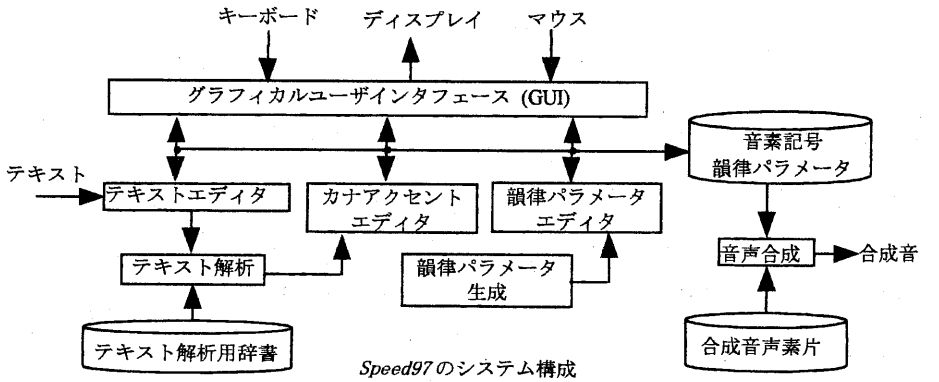
(1) 他のメディアとの同期が容易に可。 マルチメディアタイトルの制作では、画面切り替えのタイミングなど、音声と画像の同期が重要となる。*Speed97* で作成された音声は、音素記号レベル、分析フレーム長レベルなど、詳細な時間分解能で他のメディアとの同期が容易にとれる。

(2) 音声の検索が可。 対応づけられた文字情報をたより音声信号を検索でき、必要な部分のみを聞くことなどに利用できる。

(3) 再利用可。 一度作成した音声を、文字情報および韻律情報と共に、文単位、フレーズ単位などで辞書に蓄積しておけば、これを再利用して、新たな文を容易に作成できる。言い回しや語彙の限られたアナウンス文等の作成に有効である。

また、*Speed97* で作成された音声は、1kbit/s 以下の高能率符合化音声である。音声信号を 80 分の 1 以上に圧縮することができるため、回線の伝送レートが低い場合には有効な音声出力手段となる。例えば、ホームページでの利用が有望である。一方、*Speed97* で韻律パラメータをエディットする際に、自然音声から抽出した韻律パラメータを参照するようにすれば[3]、高能率音声符合化方式としても利用できる。

「音声を制作」するツールとしては、様々なレベルで、様々な機能が必要であろう。このうち、*Speed97* は、最も低レベルに位置し、音声合成のパラメータを直接操作する機能を有する。また、より上位の音声作成法として、我々は MSCL（Multi-Layered Speech/Sound Synthesis Control Language）と呼ばれる記述方式を提案している[4]。これは、HTML のようなマークアップランゲージに類似しており、抽象化された制御コマンドで音声現象を表現することによって音声作成を行う。詳細は、文献[4]を参照されたい。



Speed97 のシステム構成

[1] 阿部他，音声言語情報処理研究会資料，17-12，pp. 67-72（1997. 7）
[2] 篠崎他，音声言語情報処理研究会資料，17-14，pp. 81-88（1997. 7）
[3] 阿部他，音響学会講演論文集，2-2-1，pp. 225-226（1997. 9）
[4] 水野他，音声言語情報処理研究会資料，17-13，pp. 73-79（1997. 7）

2.6 テキスト音声合成ソフトウエア（東芝音声システム）

東芝は 95 年 11 月、Windows® 95 の日本での発売に合わせてパソコン用のテキスト音声合成(TTS:Text-To-Speech)ソフト「東芝音声システム Ver1.0[1] 」をプリインストールの形で発売した。96 年 6 月には音質を改良した「同 Ver1.5」を、96 年 11 月には音声認識機能を加えて「同 Ver2.0」を、97 年 6 月には、合成音質と認識機能を大幅に改善した「同 Ver2.5」を発売した。さらに 98 年 2 月に発売した「同 Ver3.0」は Microsoft® Agent に対応し、アニメの口の動きと同期させて音声を出力する。

本 TTS の構成を図 2.6.1 に示し、主な特長を以下に述べる。

・肉声に近い高品質な合成音[2]
・音声素片の容量は男女声についてそれぞれ 150kB ずつで済み、かなり小さい。
・調音器官の動きの制約を考慮し音韻継続時間長を適切に制御しているので、話し方が自然[3]
・男女の音声素片を基に一部の合成パラメータを変更することによって、様々な声を作成可能
・精度の高い言語処理により読み誤りが少ない。 [例]漢字が同じでありながら違う読み方をする場合についても正しく読上げる：「公園を通って学校へ通った。」「彼女は神奈川の山北町（やまきたまち）から新潟の山北町（さんぽくまち）へ嫁いだ」

「東芝音声システム Ver3.0」に含まれる数々のアプリケーションのうち「おしゃべりテキスト」など「おしゃべり」が付くアプリケーションには TTS 技術が使用されている。「おしゃべりテキスト」は、テキストファイルの読み上げやクリップボードの自動読み上げ、OLE2.0 による埋め込み等の機能を持つ。また「音声のプロパティ」を開いて、読み上げる声（キャラクタボイス）の選択、読み上げ方の選択、ユーザ辞書登録等ができる。キャラクタの声質や声の高さ、抑揚等を変更するツール「おしゃべりキャラクタ」を使えば、新しいキャラクタの声を作成し登録することが可能である。

「東芝音声システム」の TTS 機能を使ったアプリケーションを作成するには、「東芝音声合成コントロール」(OLE コントロール)が利用できる。読み上げるキャラクタや読み上げ方の選択をするプロパティ、読み上げや WAVE ファイル作成をするメソッド、読み上げ位置や読み上げ終了を知らせるイベント等を利用してプログラムを作成する。

また Microsoft® Agent を利用して Web のページやアプリケーションを作成して、その中で TTS 機能を利用することもできる。「東芝音声システム Ver3.0」では Microsoft® Agent に対応したうさぎのキャラクタ「スーパーミミ」を提供している。TTS による読み上げ音声に同期して口を動かしたり、プログラムによりさまざまな動作をさせることができる。

なお「東芝音声システム」の動作する環境は、
Microsoft® Windows® 95 日本語版
Pentium®以上の CPU
16MB 以上の RAM(24MB 以上を推奨)
18MB 以上のディスク空き容量
Sound Blaster®互換のサウンド機能
である。

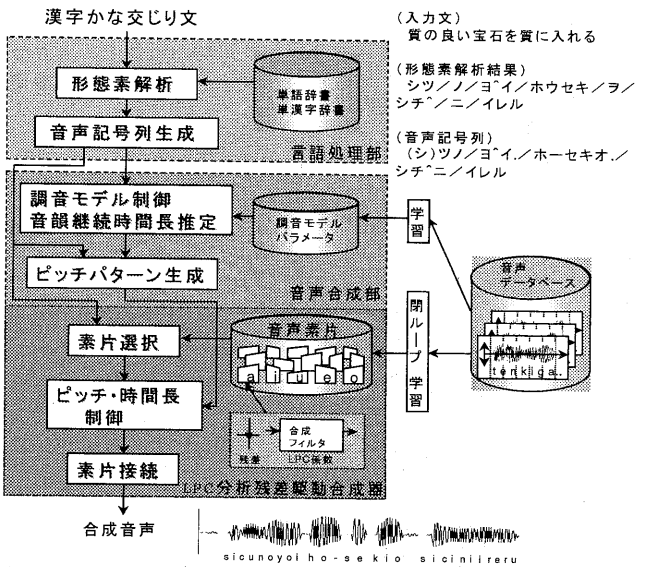


図 2.6.1 東芝音声システム Ver2.5,Ver3 の TTS の概略構成

1.桃崎ほか，音学講論 2-4-10(96.9) 2.赤嶺ほか,97-SLP17-16(97.7) 3.志賀ほか，音学講論 1-7-7(97.3)

(注)Microsoft®は Microsoft 社の登録商標です。