Windows® PC用音声認識ソフトウェア

Speech Recognition Software for Windows® PCs

太田　治徳　H. Ohta
正井　康之　Y. Masai
鈴木　孝子　T. Suzuki

音声認識（SR：Speech-Recognition）は，人間の話した言葉を処理し，その言葉の意味を認識結果として返す技術である。当社はSRを利用して音声でWindows®（注1）の操作が行えるソフトウェア環境を，Microsoft®（注2）社の日本語Windows® 95上で構築した。直覚的であるという音声の特長を生かして，音声入力をキーボードやマウスとは異なる効果的な入力手段にすることができた。

また，不特定話者の言葉を学習なしに認識するために，声質の違いや話しかたの違いを統計的に表現して認識する方式を採用し，SR処理をすべてソフトウェアで実現した。さらに，各種のアプリケーションからSRが利用できるようにするため，Microsoft® 社が提案しているAPI（Application Programming Interface）に準拠している。

Speech recognition (SR) is a technology that recognizes the spoken word. This paper describes our own SR software environment in Microsoft® 's Windows® 95 operating system, which allows the user to control an application by means of spoken commands. This gives the user more effective control of the application than by using the mouse or keyboard due to the intuitiveness of the spoken word. The SR procedure is speaker-independent, through the use of a method whereby differences in voice and talking characteristics can be expressed statistically. We have also provided our own subset to the Microsoft® speech application programming interface (API).

1 まえがき

音声は，人間にとってもっとも自然な情報伝達手段であり，人間とコンピュータとのインタフェース（ヒューマン インタフェース）に音声を利用する研究は古くから行われてきた。最近は，特にパソコン（PC）のヒューマン インタフェースとして音声が注目されている。その背景には，OSにオーディオ信号を扱えるマルチメディア機能が追加されたことや，サウンドボード，スピーカ，マイクを備えたマルチメディアPCの普及がある。特に，高速のCPUが搭載され，音声処理がソフトウェアだけで安価に実現できるようになってきたことが要因としてあげられる。

さきに，Windows® 95用音声合成ソフトウェアについて報告した(1)。今回，入力と出力を統合した音声ヒューマンインタフェースを実現するため，新たに音声認識ソフトウェアを開発した。SRの基本技術と，SRを利用したヒューマン インタフェースを提供するソフトウェア環境の概要を紹介する。

2 SR技術

SR処理は，図1に示すように，特徴パラメータ抽出部と認識部から成る。特徴パラメータ抽出部は入力音声から音声認識に有効な特徴だけを効果的に抽出する。認識部は，二段階で構成されている。前段の音声セグメント量子化処理では，不特定多数の話者や雑音などの環境による100 ms内の微小な音声変動が表現されている音声セグメント辞書と入力音声を照合し，微小変動を取り除く。後段の単語照合処理では，発声ごとにばらつく100 msを超える音声パターンの変動が表現されている単語辞書と入力音声を照合し，変動の大きい不特定多数の話者の音声に対しても高精度に認識する。

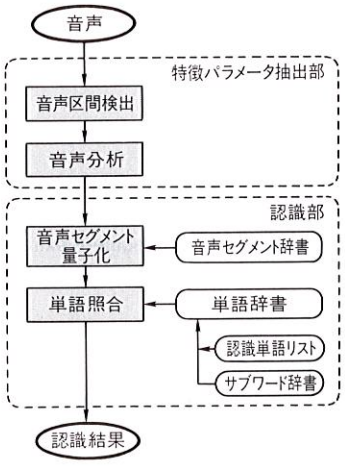

図1．SR処理の概要　人間の話した言葉を処理しその言葉の意味を返す。
SR procedure

（注1），（注2）　Windows，Microsoftは，Microsoft社の商標。

2.1 音声区間検出

音声区間検出処理では，マイクから入力された音声信号から 16 ms ごとに計算したパワー値を使って音声区間の始端と終端を検出する。

2.2 音声分析

検出された音声区間の音声信号を FFT (Fast Fourier Transform) で周波数分析して，音声信号を時間軸と周波数軸から成る二次元パターンに変換する。**図 2** に単語，ピペットの「ピペッ」の部分の音声分析パターンの例を示す。

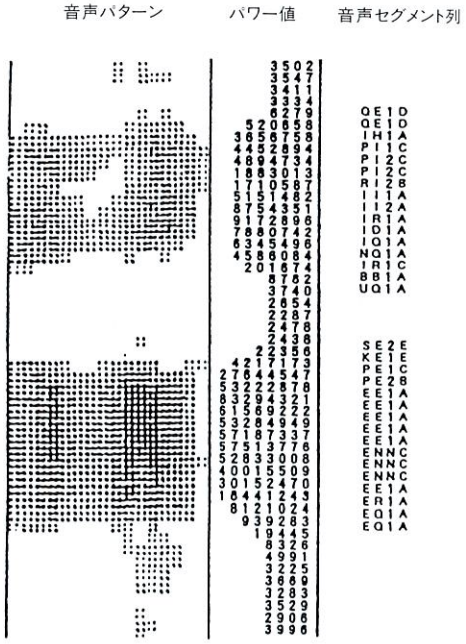

図 2．音声分析パターンと音声セグメント抽出の例　単語ピペットの「ピペッ」の部分である。

Example of speech analysis pattern and phonetic segment extraction

2.3 音声セグメント量子化

連続音声中には，さまざまな形の音声事象が観測され，これを音声セグメントと呼んでいる。

日本語の音声セグメントには，690 種余りの音響的/音声学的構造（音響特徴，音素 (C, V)，Cv，vCv，Vc，C ∪ C) がある。ここで，音声セグメントの種類の表記に用いる C，V はおのおの子音，母音を示し，小文字はそれが過渡的な区間にとどまっていることを示す。なお，C ∪ C は同時調音的発声を示す。今回は，小型化と高速化のために 690 種余りの音声セグメントを 168 種に統合している。たとえば，語頭の Cv，語中の Cv，vCv は同じクラス Cv にまとめた。図 2 に音声セグメント抽出の例を示す。音声セグメント列に記述された文字列 (QE1D，IH1A など) が音声セグメントの名称である。

量子化処理では，音声分析パターンを 16 ms ごとに 96 ms の時間長で切り出し，168 種の音声セグメントの標準パターンとの距離を計算し，切り出した音声パターンの音声セグメントを決定する。

2.4 単語照合

単語照合処理では，**図 3** に示すように指定した認識単語リストに記述された単語の読みに従って，サブワード辞書を接続して単語辞書を作成する。サブワードとは単語を細分化したものであり，例えば音節などをサブワードとすることができる。

単語をサブワードに細分化して表現することによって，単語の読みを与えるだけで任意の単語の認識が可能となる。認識時には，音声セグメント量子化処理で得られた入力音声の音声セグメント列と，各単語辞書との類似度を計算し，類似度がもっとも高い単語を認識結果とする。

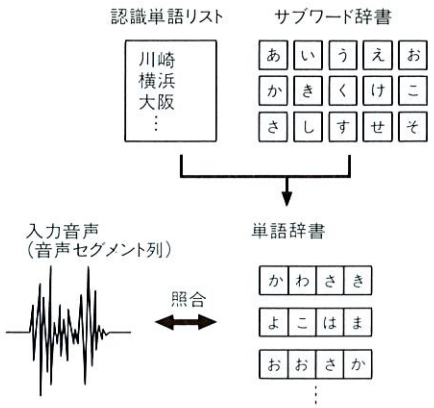

図 3．単語照合処理　入力音声の音声セグメント列と各単語辞書との類似度を計算する。

Word matching procedure

3 日本語 Windows® 95 上での SR システムの実現

SR を利用したヒューマン インタフェースを提供するソフトウェア環境は，**図 4** に示す構造になっている。認識エンジンは，2 章で述べた SR 処理を行うソフトウェア群であり，**表 1** の仕様をもつ[2]。Speech API は，アプリケーションに SR 機能を提供するプログラミング インタフェースである。

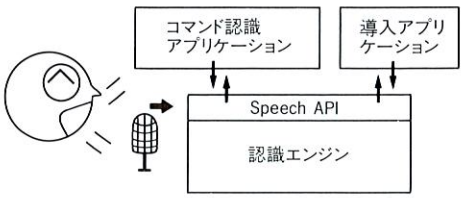

図 4．日本語 Windows® 95 上での SR システム　SR を利用したヒューマン インタフェースを提供する。

SR software environment in Windows® 95

表1．認識エンジンの仕様
SR engine specifications

動作環境	サウンド機能を備えたWindows® 95搭載PC
利用条件	語彙(い)登録：読みをかなで入力 発話形態　：孤立単語 語彙数　　：300語 話者　　：不特定話者 マイク　　：外部または内蔵マイク
応答速度	発声後約0.5秒

さらに，音声認識の特長を生かしたユーザインタフェースを検討し，アプリケーションを構築した。

3.1 プログラミングインタフェース

プログラミングインタフェースは，Microsoft® 社が提案しているSpeech API [3]のサブセットになっている。

Speech API はアプリケーション開発者とエンジン提供者に対して標準インタフェースを提供する。プログラマはどのエンジンを使用しているかを考えずにプログラミングでき，エンジン提供者はすべての音声対応アプリケーションとの互換性を確保できる。

Speech APIは，SRと文音声合成（TTS：Text-To-Speech）を使うWin32 (注3)アプリケーションのためのAPIであり，OLE [4](Object Linking and Embedding) のCOM [4](Component Object Model) オブジェクトの集合として定義されている。そのため，どのような言語で書かれたプログラムからでも音声機能が使用できる。

SR機能を利用するアプリケーションは，Speech APIを通じて認識対象の単語を登録する。単語の登録には以下の項目が必要である。認識に特に重要なのは“認識単語の読み”である。認識エンジンは，この読みデータと入力音声データを照合し認識結果を返す。認識結果には，対応する登録時情報が返る。

項目	説明	例
認識単語ID	：ユニークな値	1
認識単語の読み	：ひらがな入力	へんしゅう
認識単語（表記）	：コマンド名など	編集

Speech APIでは，ルールとグラマーの概念を導入している。認識すべき単語はアプリケーションに依存しており，また各局面でも異なる。認識エンジンに対して一律に認識対象単語を登録すると照合範囲が広くなり認識率にも影響する。したがって，ルールとグラマーの概念を導入し認識対象単語をグループ化する（図5）。認識エンジンはグラマー単位で読み込み，ルールの有効化によって対象単語の照合範囲を絞り込むことができる。

(注3)　Win32は，Microsoft社の商標。

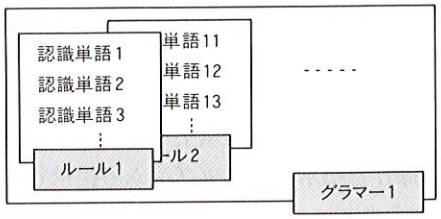

図5．認識単語の管理　ルールとグラマーの概念を導入し認識対象単語をグループ化する。
Management of recognized words

3.2 ユーザインタフェース

音声は，自然な形のコンピュータの入力手段として注目されているが，現状ではマウスやキーボードに取って代わるものではない。しかし，音声のほうが有利な場面はいくつか考えられる。すべてを音声に代える必要はないと考え，この“音声（認識）が有利な場面”を探りながら，PCでの用途の分析を行い，最適なアプリケーションを考えた。また，既存のアプリケーションを変更せずに，音声インタフェースを利用できるくふうも行っている。

3.2.1 用途の分析　音声は直覚的な指示手段であることに注目する。

PCになにかを行わせようとするとき，キーボードやマウスの場合，その指示を各アプリケーションのユーザインタフェースの作法に従った操作にブレークダウンしなければならない。音声であれば，指示をそのまま口に出せばよい。例えば，メールを送信するとき，宛先は「相手の名前」を言い，送るときには「送信」と言う。

PCのアプリケーションでは，キーの複数組合せで操作を行わせたり，リストから該当する項目を選択させるものが多い。このような場合に音声を適用すれば，キーボードやマウスより効果的な入力手段となる。

3.2.2 コマンド認識アプリケーション　音声とコマンドを対応づけ，音声でWindows® の操作を行う手段を提供する。

Windows® 95の標準コマンドはあらかじめ登録しており，ウィンドウの最大・最小化をはじめ，コピー，張付け，切取りなどの標準操作，アプリケーションの起動が音声でできる。たとえば，ウィンドウの最大化は「さいだいか」と言うことにより実行される。また，各アプリケーションのメニューも起動時に自動的に検索して登録している。

さらに，利用者が，ある指示を音声コマンドに対応づけするための登録ツールを提供している。3.1節で認識対象の単語登録に必要な項目を述べたが，そのなかの“認識単語”の項目に注目する。この項目を“認識単語の読み”に合った音声データが認識されたときの動作の記述として使用する。

たとえば，“CTRL＋F5”のキーボード操作がスペルチェ

ック機能になっている文書作成アプリケーションで，その操作を音声で実行させる場合は，認識単語の読みを「すぺるちぇっく」，認識単語を“CTRL＋F5”と登録する。

登録ツールに加えて，認識結果を受け取り該当アプリケーションに操作を指示するプログラムも提供している。つまり，「すぺるちぇっく」が認識されると該当アプリケーションに“CTRL＋F5”のコードを送ることでスペルチェックを実行する。この方式により，アプリケーションの変更なしに音声インタフェースを使えることになる。

3.2.3 音声導入アプリケーション　アニメーションで作成したうさぎのミミと会話するアプリケーションは，話しかけると音声と動作で応答する（図6）。応答音声はTTS機能で合成している。5種類の動作に対応する話かけることばと返事の内容は，利用者が自由に設定できる。このアプリケーションによって，一般利用者にSR，TTSに親しんでもらうことができる。同時にコンピュータとの対話という新しい体験を提供する。

上述のアプリケーションは一例であり，次のようなことも容易にできる。「今何時」と言うと，TTS機能をもつ時計アプリケーションが現在時刻を読み上げる。「売上報告書」と言うと，表計算ソフトウェアを起動し報告書シートを開く。さらに「表の読上げ」と言うと，表の各項目の値を読み上げる。

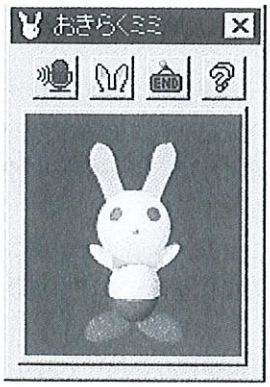

図6．SRとTTSを使用したアプリケーション　話しかけるとうさぎが音声と動作で応答する。

Example of application using SR and text-to-speech (TTS)

4 あとがき

SRとTTSを利用したソフトウェア環境は“東芝音声システム”として当社のPCに標準搭載されている。このシステムは，長年にわたり研究所レベルで培われてきた要素技術をPCに適用した成果である。

これは，音声の特長を生かしたインタフェースとなっているものの，通常のPCの使いかたにおいては，表示装置やキーボード，マウスに取って代わるものではない。

真に音声が重要となるには，音声以外にインタフェースのない状況においてである。例えば，外出先から自宅や会社のPCに電話を掛けて操作する。電話でPCと対話し，到着しているメールやファクスの読上げを指示し，内容を電話口で聞く。

このような状況では音声の真価を発揮する。こうした時代の到来もそれほど先のことではない。認識率の向上，最適なユーザインタフェースの構築など克服すべき課題も多いが，今後とも音声技術をPCに積極的に取り入れていく所存である。

文　献

(1) 太田治徳，他：パソコンにおける文音声合成を利用したヒューマンインタフェース，東芝レビュー，51，1，pp.14-17 (1996)
(2) 正井康之，他：パソコン用音声認識ソフトウェアの応用，電子情報通信学会総合大会論文集，A-15-22 (1997)
(3) Speech API Developer's Guide Version2.0, Microsoft Corporation (1995)
(4) Kraig Brackschmidt : INSIDE OLE2, Microsoft Press (1994)

太田　治徳　Harunori Ohta

青梅工場 パソコンソフトウェア設計部グループ長。
パソコン基本ソフトウェアの開発設計に従事。
Ome Works

正井　康之　Yasuyuki Masai

マルチメディア技術研究所 開発第六部開発主務。
音声認識システムの研究開発に従事。
Multimedia Engineering Lab.

鈴木　孝子　Takako Suzuki

青梅工場 パソコンソフトウェア設計部。
Windows ソフトウェアの開発設計に従事。
Ome Works