音声認識技術を用いた音声文字変換呈示システムの開発

―聴覚障害教育のための情報保障支援機器としての実用化を目指して―

内野權次　教育方法開発センター（聴覚部）

要　旨：聴覚障害教育の情報保障支援機器として，人間の音声を，音声認識装置を用いて文字コードに変換し，ビデオディスプレーに字幕表示するシステムを開発した。このシステムは実用化研究のためのものである。ここにそのシステムの紹介をしたいと思う。

キーワード：音声認識，不特定話者，連続音声認識，聴覚障害，教育工学

1. はじめに

この研究の最終的な目標は，本学の聴覚部の学生に対する講義や，教官の会議などで，音声を文字変換し，情報保障のために使用することである。

この目的に適合した開発システムに要求される特徴と機能は以下のようなものである。

①日本語の変換が可能であること。
②認識変換速度が速いこと。
③不特定話者の音声認識方式が可能であること。
④認識語彙数が多いこと。
⑤認識パラメータや辞書等はユーザー側での変更や組み替えの自由度が充分であること。
⑥連続音声認識方式であること（文節単位の認識が可能なこと）。
⑦話者の映像と字幕を同一画面にスーパーインポーズ表示が可能なこと。この機能は，話者の口形や表情および手話動作等の同時表示のために重要である。

2. システム構成と各部の動作機能

2.1 システム構成

システム構成を図1に示す。大きく分けて，音韻認識装置，ホストコンピュータ，ビデオモニタ，テレビカメラ，ビデオスーパーインポーズボード，音声入力用マイクロホンと文節変換指示入力用スイッチによって構成されている。

2.2 各部の機能とシステム処理

このシステムは，前にも述べたように話者の発声を順次文字変換することが目的であるので，変換速度が速い事が必要である。そのための対策として，本システムでは，ハードウェアー構成や，ソフトウェアー認識変換方式等に各種の高速化のための対策が採用されている。以下にこれらの機能について説明する。

本システム処理の流れを図2．に示す。

2.2.1 音韻認識装置

音声信号のアナログデータをA/D変換部，音響パラメータの変換部，音韻コード変換部に分けられている。

各部の性能と機能は次のようになっている。

(1) 音響処理部

●アナログ回路

サンプリング周波数　16KHz

ゲインコントロール，アナログ・デジタル（A/D）変換

●デジタル信号処理回路

信号処理プロセッサー：DSP5600（Motorola）（20.5MHz）を2個使用したデュアルプロセッサー方式を採用

データフレーム：6.6m sec

フィルター・バンク：20チャンネル

上記プロセッサーで6.6m sec／1フレームのデータを20チャンネルのフィルター・バンクを通して線形予測法（linnear predictive coding）により23種の特徴量に分析する。

(2) 音韻エンコーダでの音韻記号列への変換

処理プロセッサー：Motorola 68020(16MHz)を使用，音韻エンコーダーには，スピーカーモデルを使用してdecision tree（決定木）方式で線形分離の計算をする。この方法を用いると，1023のノードを10回の計算で判定することが可能である。結果の出力は，450種のコード列に変換され，ワークステーションに送られる。

デシジョンツリー方式のベクトル判定原理図は図3．に示してある。この方法は，通常の逐次形計算機上で非常に高速で処理が可能である。図の○印が内部ノードで，□印が終端ノードである。内部ノードは特徴ベクトルX＝（X_1，X_2，…X_N）を用いて$\Sigma_{ai}X_i$の計算をし，判定の

しきい値Tと比較してツリーの下部へと判定を進める。即ち①から③に進むと②以下は評価の対象とならないトップダウン方式となっている。したがって，ノードが1023個あっても，$2^{10}-1$であり10回の計算で評価ができることになる。

この段階で出力される音韻コードは，最終的に決定的な結果を与えるものではなく，曖昧さを残した状態のデータである。後にワークステーション上で言語データの音韻的制約や文法的制約を用いて文字列を決定するようになっている。

2.2.2 ワークステーション上での処理

音韻コード列はRS−232C経由で，UNIXワークステーションに入力される。機種はUNISYSのUSモデル70Eを使用し，X−Windowで稼動している。ここでは各音韻コードに対して，複数の音素記号を確率付きで割り当てた音韻コードブック，各単語を音素記号列で記述した音韻辞書，および有限状態法を用いて単語間の接続を記述した文法（Syntax）が用意されて居り，これらの情報の検索参照は，ビタービ・ビーム・サーチ（Viterbi Beam Serch）法，又は，ビーム・サーチ法と呼ばれる方法を用いた音韻デコーダを通して行われる。結果は確率的に高い，確からしい文字列を見つけて出力される。

ここで使用される音韻デコーダと前項で説明した音韻エンコーダでは，スピーカー・モデルが用いられ不特定話者の音声認識を可能にしている。スピーカモデルには，現在は，1000文／人×10人［男女それぞれ別］のデータが使用されている。また，男女の区別は，前もって設定する方式となっている。

2.2.3 知的かな漢字変換（AI）辞書

これまでの出力段階で，かな漢字混じりの文章表現が可能であるが，本システムでは，新たな試みを実施した。音韻辞書からの出力は，かな文字扱いとし，つぎにAI辞書を検索してかな漢字混じりのASCIIコードに変換する方式とした。このようにすると，同音異義語などを，前後の文脈によって判定させることで，Syntaxに登録する記述文章のデータ量を大幅に節減することがが可能である。また，この段階でも入力の曖昧さのデータを文脈判定することで，最終結果の正当率を向上することが可能である。

2.2.4 ビデオ出力制御

かな漢字コードの文章データは，ビデオキャラクター変換ソフトを駆動してビデオ・スーパーインポーズボードでビデオカメラからの話者の画像と重ねられ，ビデオモニターに表示される。

2.3 日本語シンタックスの記述例

このシステムの音声認識用辞書に相当するシンタックスと呼ばれる文章データの記述方法の簡単な，例を次に示す。

```
FILE_NAME--test.jas
S-〉 ｛ここ｜この大学｝ は ｛つくばぎじゅつたんき
       ｜つくば｝ だいがく  です
       ｜ ｛くうきが｜さいばん  を｝ ぼうちょうする
       ｜ ｛かれ  は［わ］｛あし｜やさい｝ を  いためる
```

この例は，一番簡単な文章例である。まず｛｝はその中に2個以上のORとして使用する単語を書くことが可能である。括弧内の単語の区切りは｜を使用する。行のはじめの｜は上の文とORとしてあつかうこを意味する。［ ］の中は読みの音を記入する。

シンタックス文の中に変数が使える。

```
S-〉 ｛にっぽん｜にほん｝ では  どのようにして
       きっぷ_  を  かうのですか
きっぷ_→［きっぷ］
         ｜［じょうしゃけん］
         ｜［とっきゅうけん］
         ｜［ぐりーんけん］
         ｜［しんだいけん］
```

ここでは，ごく簡単な例を紹介したが，これらを複数組み合わせることによって，多くの組み合わせを少ない文章データで取り扱うことが可能である。

3. 実験結果の評価

これらのシステムの構築が完了し，全体の動作確認が終了した。現在Syntaxファイルの構築作業中である。最初のテストでは，50単語で500文章の組み合わせでテストした結果では，単語の認識率が97%，文章で88%であった。判定のしきい値やマイクのセット位置の調整を念入りに行えば，もう少し認識精度が上げられると思う。ただし，複数の話者（話者が何人もになったら）の場合では，認識率は低下することになる。また，Syntaxの文章が増加した場合でも同様に，認識率は低下する。なお本格的テストはこれからである。またこのシステムは開発用なので，認識テストの結果を認識確率データとして確認することが可能となっている。

4. 今後の課題

これからの作業として，実際に使われる音声会話，又は講義での話し言葉などを，効率のよい組み合わせで，

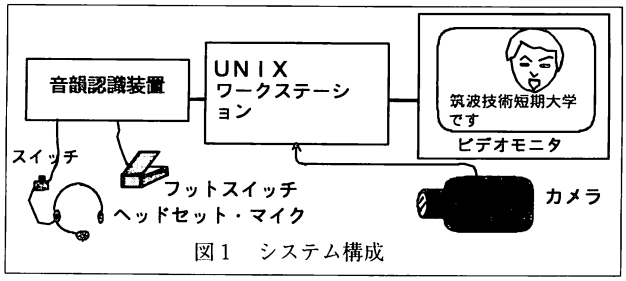

図1　システム構成

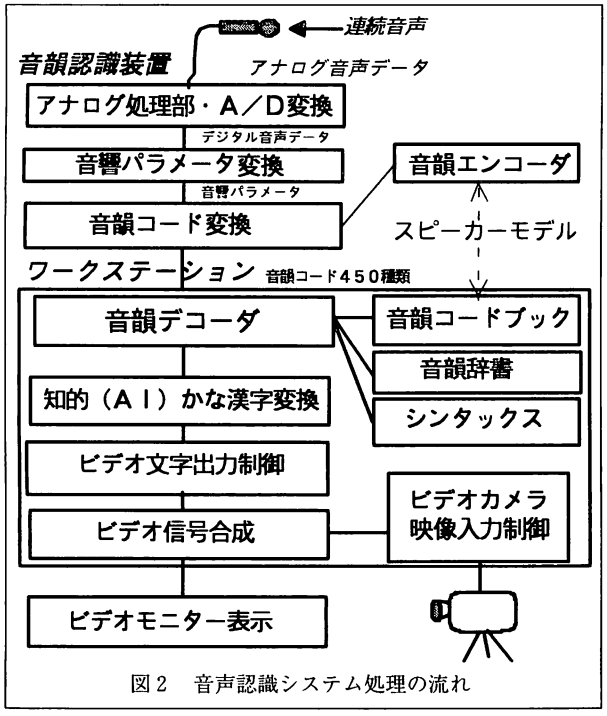

図2　音声認識システム処理の流れ

Syntax を作成して実用化のためのテストをくり返し，問題点を抽出して行くことであると思う。また，実用化システムでは，開発されたソフトを，小型のラップトップ形ワークステーション上で稼動するシステムにして置き換えて行くことも必要である。将来は，このようなシステムが，低価格のパソコンで稼動できるようにすることが望ましいと思う。

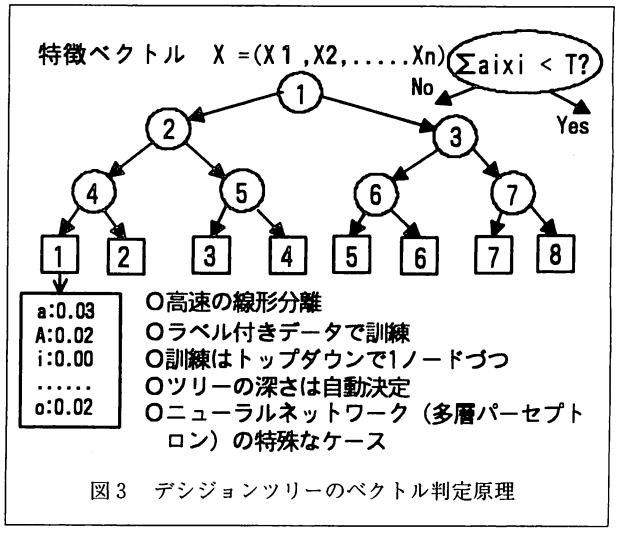

図3　デシジョンツリーのベクトル判定原理

5. 参考文献

1）内野權次：音声認識システムの聴覚障害教育への活用，第27回全日本聾教育研究大会 研究収録，石川大会，1993, pp.184-185

2）平山　輝，平島充雄：不特定話者，連続音声認識システムの開発とその応用 “Computer Woerld '91” 論文集，pp.189-196, Sep, 91