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

音声の変化を検知し重複した解読を組み合わせて判断して言語活動を展開する組織の構築

唐澤 信司† 桜庭 弘‡

†‡宮城工業高等専門学校 電気工学科〒981-1239 宮城県名取市愛島塩手字野田山 48
E-mail: †shinji-karasawa@cup.ocn.ne.jp, ‡sakuraba@miyagi-ct.ac.jp

あらまし 音声は神経細胞が活性電位を転送して調音器官などを活動させた結果であり，情報は発声の変化する場所にある．音声の活動単位の列は HMM 対角要素の状態間の遷移に相当し，遅延転送素子の列を介して解読回路の接続点として記憶される．そこで，活動単位を電荷で表し，その電荷をパスゲートとして用いる FAMOSFET の書き込みに用いて接続点群を形成し，書き込んだ活動を再演する IC を実現させる．言語処理で記憶していた認識の要素だけでは確定できない部分には複数の言語候補を選び，広い範囲の活動単位の認識処理と重ね合わせて認識する．

キーワード 音声認識，HMM，スペクトログラム，解読器，文型，文法

Architecture of the Speech Recognition System that Consists of Overlapping of Impulse Driven Look-up Tables

Shinji KARASAWA† and Hiroshi SAKURABA‡

†‡Miyagi National College of Technology 48, Nodayama, Shiote, Medeshima, Natori-shi, Miyagi, 981-1239 Japan
E-mail: †karasawa@cup.ocn.ne.jp, ‡sakuraba@miyagi-ct.ac.jp

Abstract A subset of impulsive activities is decoded through pattern-matching processes. The pattern of activities can be translated by a look-up table (LUT) that possesses the function of a dictionary i.e. an intelligent memory. The pattern of activities is expressed by the pattern of the electric charges those are used to form a circuit of LUT. Here, FA MOS that is used as pass gates operates as a connection of the circuit for the LUT. On the other hand, segmentation of speech voice is detected at the place where voice is changed, because the voice is the result of impulsive activity of brain. The intermittent operations makes possible branch processes for the layered translation of signals those are picked up from speech voice.

Keyword Speech recognition, HMM, Sound spectrogram, Decoder, Sentence pattern, Grammar

1. はじめに

1980 年代より音声認識の分野では隠れマルコフモデル(HMM)の音声処理の研究が成果をあげて実用もされている [1]．しかし、それらの状態遷移の確率を用いた音声処理はコンピュータの計算力に依存しており，システムの小型化や高速化が望まれている．また，音声認識に関する理解も充分とも言えない状態にある．

他方，知能は稼動の前提条件を満たす活動単位を次々と稼動させて実現する．それはアイコンをクリックして活動させるコンピュータの仕組みと同じである．著者らは活動単位を転送して認識を実現するという設計概念で音声処理装置の構築の検討してきた[2][3]．

脳では数ミリ秒の不応期[4]を持つ神経細胞が音声言語の情報を処理している．脳における情報の取り込実の間隔は音声では数十ミリ秒[5],視覚では 100 ミリ秒[6]である．音声で伝えられる情報は音声の状態が数十ミリ秒ごとに変化する場所にある．本研究では脳神経系の活動に基づいて音声処理することを検討した．

まず，音素の区切りを発声の変化が休止する長母音および無声状態の検出により求める．この活動単位(特徴ベクトル)の時系列は隠れマルコフモデル(HMM)の対角要素の状態間の遷移に相当する．

音素の特徴ベクトルの列は上位の桁のデジタル量のパターンマッチングによって認識する．同音異義などの複数の候補を含む場合には，リンクしてオーバーラップする上位の階層の解読器を用いて認識する．

音声信号の処理は非同期の分岐処理で状況に合わせ,また,リンクしてオーバーラップした音声言語の階層構造を利用して認識する．この音声言語処理は，書き込みが容易であり，認識できない文章は直ちに記憶させて認識の機能をダイナミックに適応させる．

2. 脳神経回路が知能を形成するメカニズム

2.1. 言葉の意味と異なる本来の意味

生物は変化する周囲と相互作用しなければ活動を続ける事ができない．そこに知能のルーツがある．生物は生化学反応を連鎖させて生命活動を続けている．活動単位の結果の状態はデジタル的に変化する．

需要と供給の関係を満たす生化学反応を連鎖させる巧妙な組織を再生産して，それを発展させる生物が誕生するには何億年もの試行錯誤を必要とした[7]．

動物は内外の状況に対応して活動を行う．相互作用のある活動を続ける動物では，細胞が活動する時に神経細胞と接続してその活動を記憶する神経細胞の回路が形成される．感覚細胞群の活動単位群のパターンがアクチュエータ群の活動単位群のパターンを生む．

神経回路網では活動した時間で活動という信号を伝えて活動の意味を回路と素子が受け持っている．つまり，感覚細胞が反応したということは外界の状況を示す刺激があったという本来の意味がある．その感覚細胞の反応を伝える神経細胞が活動すれば神経細胞が活動したという意味があり，その神経細胞が筋肉を刺激して体を動かせは，外界に対して活動したという意味がある．途中で中継する神経回路を通過するインパルスを解読する限りではその信号の意味がわからない．神経細胞の活動の効果を理解するには感覚器官や筋肉などの活動を含めて考える必要がある．

2.2. 状態が変化する所に存在する音声の情報

言葉の意味を辞書で説明するのではなく，言葉の本来の意味を理解するしくみには活動の概念が欠かせない．神経細胞が発する信号はインパルス的で，その活動単位はデルタ関数 {δ (t)} で表現できて，その結果の状態は単位階段関数 {u (t)} で表現される．

Fig.1a)に示す単位階段関数 {u (t)} の微分は Fig.1b) に示すデルタ関数 {δ (t)} となる．

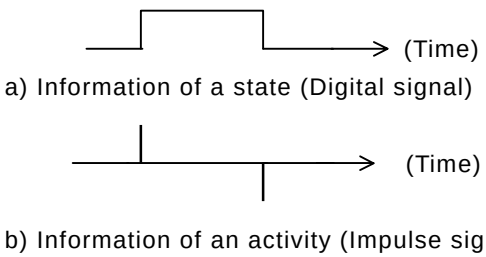

図 1　インパルス的な活動とデジタル的な状態の境界
Fig.1 The relationship between impulse and digital state

神経回路網のインパルス的な指令活動の結果として発声器官の状態が音声に特徴付けるので，神経回路網の活動の情報は音声の状態が変化する所にある．

2.3. 回路網に機能を形成する活動単位の仕組み

電気回路は、静的な状態を伝えることも動的な変化を伝達することもできる．ところが，生体では活動を伝達する動的な仕組みで成り立っている．

一般に単位の知能の活動は If A=B, then X=Y の分岐処理で記述できる．このルールはプロダクションルール[3]とも言い，活動の動機となる活動状況と、影響を及ぼす出力の状況という２組のデータのセットで指定できて，それらのデータは活動の際に採取できる．

一つの神経細胞の論理機能は遅延時間を持つ組み合わせ論理回路としてインパルスで駆動される電子回路により実現できる．遅延する機能は単安定マルチバイブレータを一つ挿入することによって実現し，論理回路の接続点には Floating Gate MOS FET を Pass Transistor として用いる．

入力側に AND 論理回路を挿入して稼動条件を設定すれば，稼動する条件がそろった時だけ出力する．出力側の接続点群はアクチュエータのモニタにより活動状態のパターンを採取して設定し，一つの出力をそれらの接続点群を通してアクチュエータ群に接続する。多数のアクチュエータを同時に動かす場合は OR 論理回路を通してその活動を伝える．

この回路を浮遊ゲート MOS FET を配列した Look-Up Table(LUT)を図２に示す．この回路は Programmable Logic Device (PLD)の機能を持つ．

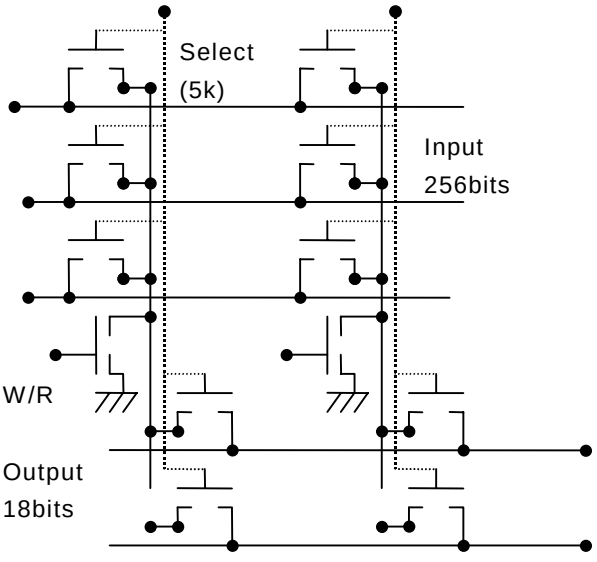

図 2. 浮遊ゲート MOSFET を配列した Look-Up Table
Fig. 2 Programmable Look up table that is composed of floating gate MOS FET

2.4. 神経細胞の機能を持つ電子回路の形成

同時に発生したインパルスにより接続点を形成し，その接続点の全てにインパルスが発生した時にインパルスを出力する解読器の IC は次のように実現できる．

この解読器は不揮発性半導体メモリ素子である浮遊ゲート MOSFET をパスゲートとして接続点に用い，信号源と負荷に次の条件を満たせて AND 論理を実現して解読機能を実現する．すなわち，信号源にインパルスの存在する状態は抵抗を通して高電位を出力し，信号源にインパルスが無い状態ではその出力端子を接地する．こうした信号源の出力群を接続要素経由で一つの出力側の端子に接続し，それを高抵抗の負荷抵抗に接続する．こうすれば，接続された信号源の中に一つでもインパルスが無ければ出力端子にインパルスが出ない．全ての接続された信号源にインパルスが存在した時だけインパルスが出力される．こうして，インパルスの解読回路が自動的に形成できる．

3. 建て増し方式による知能の高度化

3.1. 活動単位を転送して保持され組織される活動

時間的に変化する複数の成分を持つ活動状況は遅延転送回路網を伝搬する時に，２次元のパターンに展開される．平面上に展開された活動単位群は解読器の接続点として記憶できる。脳神経回路を格子点に活動単位を置く回路というモデルで描けば，脳の活動は格子点にある解読器の稼動状況で示される．ここで，脳神経回路網における活動の保持は遅延要素列を転送させる以外に,遅延要素のループを巡回することがある．ループの巡回活動は外部から活動を点火し，消去を指令することが必要である．さらに，過去の活動を保持すれば，現在の活動と組み合わせて，次の活動を判断することができる．脳神経系では既存の回路の活動によって新たな回路が追加される．この建て増し方式の解読器網は稼動効率が高い上に，写像空間が広い．

3.2. 階層的に活動単位を組織する活動の記憶

コンピュータでは制御バスとデータバスを配置し，アドレスデータを使ってメモリの内容をバスに載せることが信号の変換動作になり，解読の動作となる．

単語の情報圧縮はデータを記憶した単語の番地を単語の識別信号にし，文章のデータを記憶する際にはその文章の識別信号の代わりに登録した番地で表現する．逆に，文章のアドレス番号を指定すれば，単語のアドレス番号列が読み出され，単語のアドレス番号から文字情報が取り出される．このように，LUT の階層構造で情報を記憶すれば接続配線が少なく高速に読み出しができる言語情報のメモリができる．

3.3. 動詞を中心にして構造を持つ文の形成

文章の要は述部である．述部の中心になる要素は動詞であり，実世界の活動を表している。表現される実世界の活動は複数の要素を持っている．その実世界の活動の属性を，言語表現の述部の枠組みに組み込む方法が文法である[8].

文法は言語によってそれぞれに相違するが，同じ実世界の活動を表現するので異なる言語表現の間で翻訳ができる．論理は言語で表現した実世界の属性を整理するものである．論理の根拠は実世界における事物にある．

4. 階層化した活動単位による言語の構造

4.1. 階層的なサブルーチンの組織の動作

神経細胞の活動が数ミリ秒の不応期を持ち，時分割で活動する．構造を持つ言語の音声の処理では音声の状態変化から処理の信号を検出して，上位の区切りの信号によって上位の要素の活動に転送する．

デジタルコンピュータでは階層構造を持つサブルーチンの活動を割り込み動作で行う．そこでは，プログラム実行中に，一旦下位の層の処理を実行し，指示された下位の層の処理を終了した後に元のプログラムに復帰するためにリターンアドレスをスタック（スタックポインタ）に入れている．

図 3 に一つの発話のシリアルな発声動作をサブルーチンとして表現したモデルを示す．

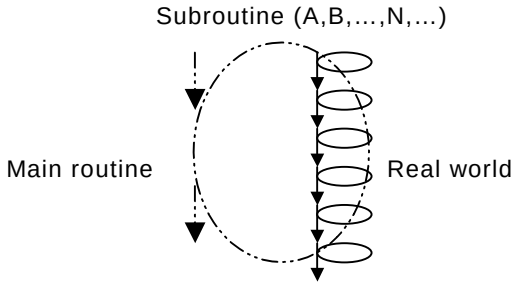

図 3　2 層にリンクしたサブルーチンの動作
Fig.3 Serial activities by double layered subroutines

脳の言語活動では同じ事柄を単語レベル，文章レベル等と複数の情報圧縮のレベルの活動単位で表現している．そこで，発音を意識すれば発音がわかり,単語を意識すれば単語がわかる．その脳神経回路網の階層構造はリンクしたオーバーラップであり，独立しない階層構造である．そこで，意識により顕在化した神経回路の活動は接続された他の層の活動とリンクできる．

図 4 に文章のサブルーチンの下に単語のサブルーチン(1,2,…,m,…),その単語のサブルーチンの下に音節のサブルーチン(ma,mb,..,mn)を割り込むという,リン

クしたオーバーバーラップの様子を示す.

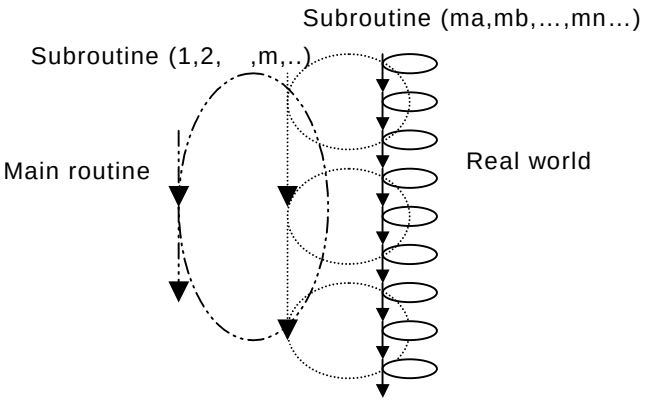

図 4　3 層にリンクしたサブルーチンの動作
Fig.4 Serial activities by triplet layered subroutines

単語のサブルーチンを呼び出すには，音素レベルの活動単位の列をそのアドレスの列で表現した信号を単語レジスタに並べて，無音声の区切り信号の検出により,次の単語の処置を実行するシーケンスを指示する.

息継ぎの区切りで，文章のサブルーチンのレジスタのデータ処理のシーケンスを行う．上位の活動単位にある文章を下位の発声制御で発言の最中に外界から割り込み動作があるとする．その時に発声のどの段階で外部の状況を割り込ませるかを判断しなければならない．そのような時にデジタルコンピュータでは稼動中の状態をフラグによって表し，外部状況の状態と組み合わせて，次の活動を制御している．これと同様な分岐処理により，リンクしてオーバーラップする動作のシーケンスを稼動させる.

4.2. 機械的な活動単位による音声のデータの採取

まず，音声波形を機械的にサンプリングして A/D 変換する．標準的なサンプリングでは毎秒 10k サンプル(周期 0.1 ミリ秒)で 1 サンプルあたり 8 ビット語で，80kbit/sec のデータレートである[9].

音声パラメータの採取の周期は 5～10 ミリ秒とされ区切の期間(分析フレーム長)は 15～30 ミリ秒である．他方，人間は声の動きとしては一回が数十ミリ秒以内で毎秒 15 個程度のレートである.

単語レベルの音声の区切りで無声の状態を差し込み発声の息を途切らせる．通常の日本語の仮名文字単位(拍；モーラ)は 1 秒間に 5～7 拍で，その平均発話時間は 150～200 ミリ秒といわれている[10].

この次々と音素を発声する活動が毎秒 13 個程度のデータレートで進行する．状態遷移確率を示すマルコフモデル(Markov Model)の対角要素(a_{ii})はその状態の継続時間 d と $d=1/(1-a_{ii})$という関係があり[11], 状態の進行は状態間の遷移と見なすことができる.文章レベルの発話の終わりに身体動作を行う息を吸う(0.5msec)程度の無声状態となる.

最初に，サンプリング周期を 0.1 ミリ秒,振幅データを 8bit で表す．1 つの分析フレーム長を 20 ミリ秒とし，32 次元のベクトル成分を 5 ミリ秒ごとに機械的に採取して，32 次元 8bit で毎秒 20 回の音素の情報の状態遷移に変換するとその情報圧縮率は 1/10 である.

5. 音声認識システムの構築

5.1. 音声言語処理の信号の流れ

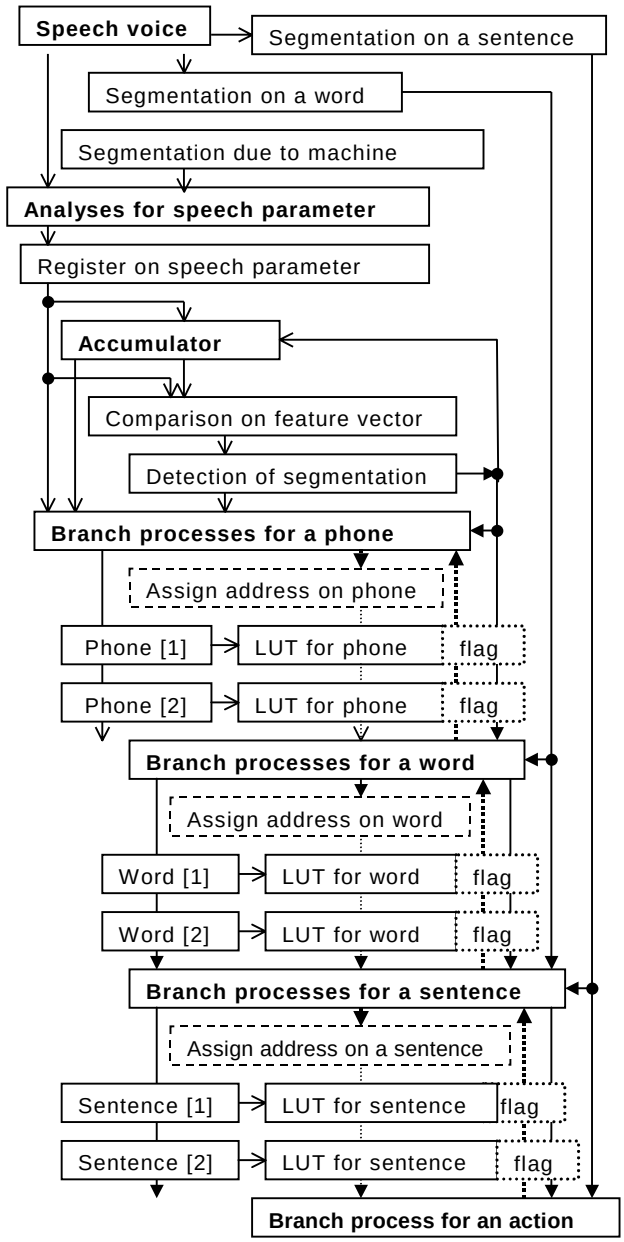

図 5 音声信号処理装置のデータの流れ
Fig.5 Flow of data on a speech signal system

音声言語処理においては，音素，単語および文章などの各要素の継続時間が異なるので，非同期の言語データの流れよってシステムが制御される．書き込み音声信号の活動の流れを図 5 に示す．

5.2. 活動単位群のデータ変換を記憶する LUT

知能動作は状態遷移というモデルで表現できる。その状態の切り替え動作はプログラムでは IF 文であり，組み込み系のフローチャートの分岐動作（branch process）である．

ここでは後述するように最大 256bit のデータのセットを 18bit のデータに情報圧縮する LTP を言語の情報処理の記憶に用いる事にした．

図 5.ではシリアルに入力する音声信号の区切り信号でレジスタに並べたデータを一度に LUP メモリの入力端子群に転送し,出力には登録番号(18bit)を書き込む．この回路では部分集合を指定する出力として書き込まれたアドレスが，入力のデータとその構成成分でを読み出すことになる．図 2．の LUT の接続点群は双方にデータを流ことが可能である．

5.3. 活動単位の流れを仕分ける時分割制御

音声認識では状況に応じて弁別処理を行う．最初の音素認識が同定できない場合には，ベクトルの距離が計算できないので，その領域の発声時間に関する条件だけで認識候補を検索出する．

単語レベルの認識において，内部に音素語の認識候補が複数ある場合には上位の活動単位の入力に複数の候補があることになる．範囲の異なるレベルで検索された候補を照合して認識する言語処理においては，認識の候補を flag で表示して処理する．

認識候補となる個所が幾つもある場合にはそれらの組み合わせが多くなるので，照合する工程を短縮するために，その場所だけは発声の継続時間の系列だけで照合し，上位のレベルで候補を選び出し，下位の候補と一致する候補に絞る．

このシステムでは想定されて組み込まれた文章以外は認識できないので，認識できない文章は新たにそれを装置に書き込む．また，人によってスペクトルが相違する．そこで，それぞれの人に同じ発音記号列について数種類のデータを採取し，同じ発音記号に対する特徴ベクトルと同定される特徴ベクトルについて，その成分と継続時間を採取し参照データとする．

6．音声のから情報の抽出

6.1. デジタル情報の抽出

脳神経系は活動単位を転送して活動しており，脳神経系がデジタル的な状況を発生するにも関わらずアナログ的な世界を認識するのはルールを稼動させる活動により認識していると考えられる.脳神経系で発生し，脳神経系で受けとられる情報はインパルス単位であり，その情報は音声の状態変化の境界にある．そこで，同じ音声の状態を最小の区切りの信号とし，区切り区間の音声の変化でから特徴ベクトルを採取する．

6.2. 音声の雑音の除去

サンプル値からノイズを除去するために前後するデータの平均値を中央値のデータとする(移動平均法;メデアンフィルター）によりノイズを除去する．音声から抽出する言語情報は音の振幅より，周波数情報にあるので，音声の振幅を自動利得調整回路によってほぼ一定の大きさにする．

6.3. 音素の特徴ベクトルの形式

記憶されていた信号と一致することが認識の条件になる．そこでは入力データと記憶していた参照データが必要であり，その認識では参照データも入力データも同様方法で抽出し，同じ形式で表現する．

6.4. 時間変化から空間変化への変換

音声の調音器官の幾何学的形状をフィルターとしているから音声の周波数成分は調音器官の動きを反映している[12]．そこで，波形という時間上の変化をフーリエ変換により空間軸の成分に変換する．その空間的な要素を図面に書き写したサウンドスペクトルから音素の特徴ベクトルが採取できる[2]．なお，機械的にサンプルされる切り口の影響を取り除く除いた自己相関関数は電力スペクトルとの間にフーリエ変換の関係がある．

6.5. 前後の発声に影響される音素の取り扱い

15～30 ミリ秒程度の期間である 1 つの分析フレームの音声パラメータを基本的な特徴ベクトルとする．そこで,スペクトログラムの周波数チャネルを 32 とし，1 チャネル 8bit のデータで記述すれば，1 フレームのデータは 256bit となる．

6.6. 一致度を評価する値としての距離

音声のスペクトログラムの変化を総合的に評価するために，時間を前後する特徴ベクトルの類似性を同じ成分の距離の二乗の和の平方根を求めた値（ユークリッド距離）名殿距離より判定する．

6.7. 音素の継続時間

類似な特徴ベクトルの繰り返し回数を継続時間とする．この特徴ベクトルの状態遷移時間の列は HMM 隠れマルコフモデルの状態間の遷移行列の対角要素列に相当する．マルコフ連鎖モデルでは次の発声状態の遷移確率を等間隔の離散時間で付与した．

6.8. バラツキのある特徴ベクトルの判定方法

数十ミリ秒程度の期間で発生するスペクトログラ

ムの変化を表現するデータとしてその領域の中央と前後の 3 つ組みの周波数成分のデータを採取し，参照データと照合する．データは 8 ビットであるが，その発音記号に相当する音声を何回も発声して，そのバラツキを考慮して同定を評価するデータの桁を決める．ひとつも同定されない場合は一致度のレベルを低くして認識候補を選ぶことができる。

6.9. 音素の出力の記述方法

ひとつの音素を 3 つ組のアルファベットで表現すれば，文字 1 個で 5bit（32 種)その組み合わせは 15bit の 32,768 種となる．ところが，実際に出現するトライフォンは 5000 種程度である[1]pp.37．そこで，その音素の表現を 13bit で 8193 種の登録番号に変換する．

そして，繰り返しの回数を 5bit（32)の値で書き込むと，一つの音素が 18 bit のデータとなる．

6.10. データ変換表による音声データの圧縮

LUT を用いた認識処理では，採取された音声のシンボル系列として音素の特徴ベクトルにその状態の継続時間を加えて，レジスタに載せて入力データとし，出力のデータは登録番号のデジタルデータにする．

図 2 に示す LUT 用の半導体メモリの入力をデータバスに接続し，出力をアドレスバスで平行した 2 つのバスで情報を処理すればそのシステムは簡素化されて処理速度が著しく改善される．その言語処理用の LTP の規模を表 1 に示す．

表 1　言語表現のデータ変換表の規格（図 2 参照）
Table 1. Size of look up table for data compression on elements of language.

Look up table for words

Input 18bits(phone) ×14phoneme<256bits
Output　　18bits=262,144words

Look up table for sentences

Input　18bits(triphone) ×14words < 256bits
Output　　18bits=262,144sentences

表 1.に示す LUT では， 8bit で表した音素 14 個以下の音素列の単語単位をレジスタに載せ，一度に単語の参照データを記憶する。14 個以上の音素の列を持つ単語は分割して照合する．単語の種類を 18bit のバスに載せると 65,536 番まで登録番号ができる．ここで，単語表現は 256bit から 18bit に情報圧縮できる．

表 1.のに示す LUT を文章のメモリとして用いるとすれば，18bit の単語を 14 個の単語を並べることができる．文章の出力を 18 bit で表現すると 262,144 種の文章を識別できるが,図 2 の回路では 2 つの文章を 2 列で記憶するので,必要な文章が 1,000 であれば,回路は 1,000 列となる．

7.　むすび

音素の区切りは音声状態が変化しない場所にある．そこで,音声の変化の状況に焦点を当てて音声を認識する．まず，周波数弁別器群でサンプリングした音声のパターン状のデータを時間進行でシフトして照合する.音素としての情報を抽出する段階でさまざまな処理を行う.音声はリンクして重畳した階層構造を持つので,認識が未確定な領域は上位の階層の照合を利用する．

単語の区切りの信号を受けた時にレジスタに並べたれた音素の活動単位の配列をデータバスに載せて，記憶された単語群に一度に照合し，その結果により，レジスタ上に存在する活動単位群を別の認識候補に切り替えて照合することや，新たな活動単位群としてその活動単位を解読する回路を形成することを行う．

言語情報を情報圧縮して記憶するために，内部に動作の切り替え制御をする 256 本の入力バスと 18 本の出力バスを備えた書き込み可能なデータ変換メモリ(LUT)を提案した.この電荷転送方式の半導体集積回路を設計製作することが今後の課題である．

本報告の音声認識のしくみは自動翻訳電話機や発音をわかりやすい発音に直す装置の改善などに貢献することが期待できる．

文　献

[1]鹿野清宏，伊藤克亘，河原達也，武田一哉,山本幹雄編著「音声認識システム」IT text シリーズ．オーム社，2001.

[2]唐澤信司，桜庭弘“音声のスペクトログラムの共通部分を解読の条件に用いた発話認識”信学技報，HIP2005-95，pp.85-90，2005.

[3]唐澤信司,“活動単位を組織する言語活動に基づいた自然言語の処理方法” 東北大学電気通信研究所第 343 回音響工学研究会，No.343-3，2006.

[4]J. Nicholls，A.R. Martin，and B.G. Wallace，“From neuron to brain，3rd Edition”，pp.109，Sinauer associates，Inc. 1992.

[5]中川俊一，鹿野清宏，東倉洋一「音声・聴覚と神経回路網モデル」,pp.22,オーム社，1990.

[6]乾敏郎「脳と視覚」,pp.12,サイエンス社，1993.

[7]唐澤信司,“活動による生物の進化”宮城工業高等専門学校研究紀要，第 42 号，pp.7-14，2006.

[8]唐澤信司，“神経回路に依る言語表現の構造と実世界の属性に依る言語の普遍文法”信学技報，TL2004-20，pp.5-10，

[9] 今井聖「音声信号処理」,森北出版，1996. pp.5.

[10]中田和男，「音声」，コロナ社， pp.12, 1995.

[11]L.R. Rabiner, B.-H, Juang, “Fundamentals of speech recognition”, pp.325, PTR Prentice Hall,1993.

[12]R-D. Kent, C. Read, “The acoustic analysis of speech”, Singular Publishing Group, Inc., 1992.