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

唐澤　信司†　　桜庭 弘‡

†‡宮城工業高等専門学校　電気工学科　〒981-1233 宮城県名取市愛島塩手字野田山 48
E-mail:　†karasawa@miyagi-ct.ac.jp,　‡sakuraba@miyagi-ct.ac.jp

あらまし　デジタル的な状態の変化である活動単位を電荷として転送して知能回路を形成するという設計概念を提案し，音声のスペクトログラムのパターンを二次元的なレジスタに入力と同時にシフトしてパターンを照合して解読するしくみの音声認識装置の構築を検討した．活動単位の発生したパターンを解読する回路を自動的に形成する書き込み可能な接続点に，浮遊ゲートに電子を注入すると選択ゲートに電圧をかけても導通しない浮遊ゲート MOSFET を用いた．この回路は繰り返し書き込むことにより，その共通部分だけを解読する解読回路ができる．また，音声の振幅を規格化する振幅圧縮処理について，それがスペクトログラムにどのように影響するかを調べた．
キーワード　音声認識，サウンドスペクトログラム，書き込み可能解読器，二次元レジスタ

The Speech Recognition System that Decodes Category of Spectrogram

Shinji KARASAWA† and Hiroshi SAKURABA‡

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

Abstract　A speech sound spectrogram recognition system is investigated.　The device decodes serial data on plural frequency components of speech voice. That is, serially inputting sound spectrograms are shifted intermittently by means of 2-dimensional registers. The subsets of impulses on a speech sound are decoded through pattern matching processes. Here, a new semiconductor device is used for programmable look-up table that memorizes a spectrogram as the preconditions of the decoder. The common part of spectrograms on a same word is memorized in the device. Moreover, spectrograms on speech-waves those amplitude are compressed for normalization are measured.
Keyword　Speech recognition, Spectrogram, Programmable look-up table, 2-dimentional register

1. はじめに

従来の音声認識装置は人間が獲得した規則を先行させた演繹的な方法で認識の機能を構築してきた．しかし，元来，動物の認識機能は現実における具体的な事例を重ねて獲得している．新生児や乳幼児,つまり人間の認識は経験が先行する帰納法的な方式である．

そこで，帰納的に認識の機能を獲得できる半導体デバイスが 2004 年に考え出された[1],[2]．このデバイスはデータの存在する発火パターンを接続のパターンとして，その発火パターンの解読器が自動的に形成できて，書き込みを重ねると共通に存在するデータを解読の条件にできる．この素子は特定話者の音声などのデータから言語情報を採取する簡便な装置を構築するのに適している．

他方，音声のスペクトル成分の時間変化をグラフにしたスペクトログラム(spectrogram) [3],[4]のパターン認識で音声認識ができる．今日では FFT(Fast Fourier Transform)機能により，若干の時間遅れだけでスペクトログラムをパソコンで求めることができる．そこで，特定話者の音声から人間が言葉を認識できる範囲で２種の振幅圧縮処理をした音声波形についてスペクトログラフのパターンの変化を調べた．

本報告ではこれらの結果を，特定話者の発話認識システムを実現することを検討するとしてまとめた．

2. 認識システムの帰納法的な形成

2.1. 音声のスペクトログラムのパターン認識

人間は人間が発話した音声はその音声の振幅のレンジを圧縮しても言葉を聞き分けることができる．具体的には，音声信号を対数圧縮した信号や，自動利得調整回路によって音声信号をパルス列にした信号はかなり多くの余分な周波数成分を含むスペクトログラム

となるが言葉を聞き分けることができる．これは，人間は余分な刺激があっても，必要な情報を選択的に採取して認識している証拠である．そこで，振幅条件を規格化した音声を用いて周波数成分のパターンで音声を認識する．

2.2. 活動単位を転送する知能回路

一般に，神経細胞は数 msec 以上の不応期を経て元の状態に復元する一過性の活動電位(1msec 程度のインパルス)を転送し，他の細胞に活動を起している．脳神経回路網ではその活動を転送する所を接続して認識回路を形成する．そして，神経回路網は元の状態に復帰しても，その活動により外部の状態は変化する。そして変化した外界に新たな対処をしなければならない．

活動単位を転送して知能活動を行うシステムでは，デジタル的な変化をもたらす活動単位を電荷として表現し，その存在を解読する解読器群により，その活動を記憶する．このシステムでは実時間で同時に発生した活動単位群が一つの活動単位にまとめられる．このようなインパルスで駆動するシステムにより，知能の機能が経験により自動的に形成される装置が実現できる[5],[6].

活動を前後して変化する状態をステップ関数で表すと,変化をもたらす活動はデルタ関数(ステップ関数の微分)で表される．そこで，神経回路網をモデルにした活動単位の部分集合のパターンを選択的に転送する．この活動単位を転送するシステムはデジタルシステムの微分系のシステムといえる．

2.3. パターン状のデータの変換回路の形成

音声の情報処理ではパターンの解読器の出力でパターンの出力動作を指令することが必要になる．

まず，２次元変数を１次元の変数に展開する．ここでは，時間的に変化する活動は活動単位の列を行ごとにシフトして転送する２次元的なレジスタ上に活動蛾パターンとして表わす．

次に，出力のパターンを書き込む方法としてソースとドレインが交換可能である MOSFET の浮遊ゲート MOSFET のソース・ドレイン間を活動単位の通過させる接続点とする．この接続点群は双方に電荷を転送できるので，入力電源群を出力負荷群にすれば、入力の発火パターンであった接点群のパターンが逆に出力のパターンにできる．

2.4. 言語のカテゴリーの形成

人間は共同生活で現実を共有して、言葉を要素にした観念の世界を共有するようになった．その脳神経系に構築される観念の世界は部分的に共通する．言葉は社会を維持字するコミュニケーションの道具であり、構成員がその道具を共有しても、個人はそれぞれ異なる位置を占めるので人によって行動は相違する．

皆が同じことをしたのでは分業ができない．生命の活動は周囲に適応しており，周囲に適応することが組織を作る。神経細胞を含め多細胞生物の細胞も周囲に適応した活動により細胞社会を作っている．

2.5. 音声の周波数成分の時間的変化

音声生成器官の運動によって音道の形状変化することがサウンドスペクトログラフに反映している．

図１にデジタルオッシロスコープで測定した長母音の波形とその周波数成分の例を示す．

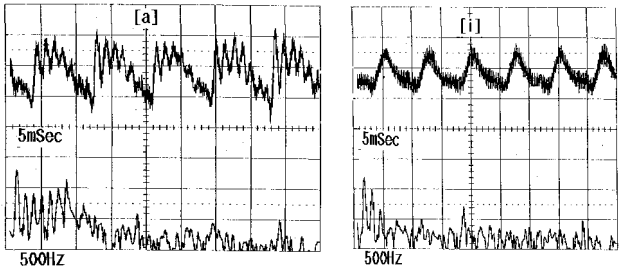

図１ 母音[a][i]の時間領域の波形と周波数領域の特性
Fig.1 Acoustic waveform of typical vowel sounds [a], [i]and the frequency components obtained through FFT analyses

長母音は声帯では発生した音響振動が調音器官で周波数特性に特徴がつけられる。図２に示す母音のサウンドスペクトルは普及版のパーソナルコンピュータに組み込まれた機能で観測したもので，その周波数成分は図１の結果と対応している．

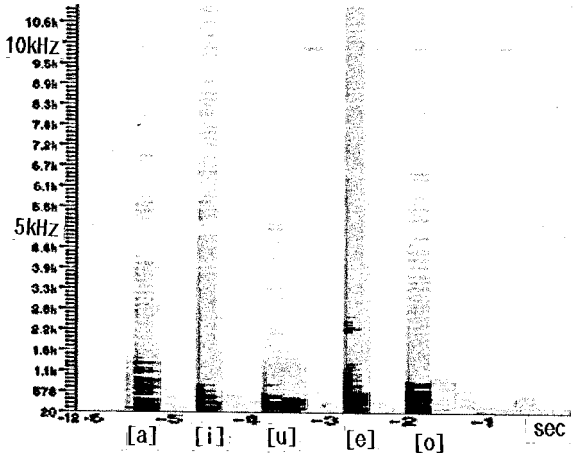

図２ 長母音[a][i][u][e][o]のサウンドスペクトログラム
Fig2. Spectrograms of the vowel sounds [a], [i], [u], [e], [o].

2.6. 情報採取する過程における処理の区切り

音声の波形によれば，単語の区切りを一般的に音響データから見出すのは簡単ではない．日本語音声の音節「モーラ(mora)」[7]を採取するのは比較的少ないデータの照合で識別できる。実際には解読する発話も，音節を短くし切って，頻繁に成分を組み合わせて認識する方が，利用効率が高く、経済的である．

会話では，多く場合にポイントとなる事柄が認識で

きれば良いので，キーワードに助詞や助動詞をつけた程度の発話が多い．書く文章と相違して，会話では必ずしも完全な文章を必要としない．

音素と単語をラベリングして，単語に助詞や助動詞を接続する程度で，会話を補助する発話認識装置で足りる。文節の解読ならばデータマッチングの率も多く長い文脈効果を処理しないで済むが，長い話しを一度に照合したのでは一致をみることがまれになる．

通常の会話での発話は 1sec から 2sec 程度で，音素は 10msec 程度である．10msec 間隔で 1.5 秒間，16 チャンネルのデータを載せると，2,400 個となる．

3．振幅圧縮処理した音声のスペクログラム

自動利得調整回路によって音声波形をパルス列にして 5kHz 以上の周波数成分を除去した音声のスペクトログラムを音声認識に使うことを検討した．

3.1．対数変換処理音声のスペクトログラム

図３に示すダイオードを用いた対数変換回路により音声信号の振幅を圧縮して得た図４に示す波形でも言葉を認識することができる．そこで，振幅の対数圧縮の処理がスペクトログラムにどのように現れるかを調べた．

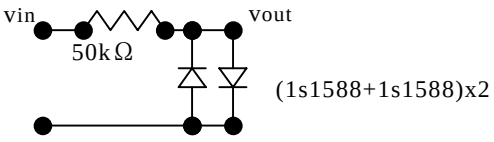

図 3 振幅圧縮処理に用いた対数変換回路
Fig.3 Logarithm transformation circuit for amplitude compression.

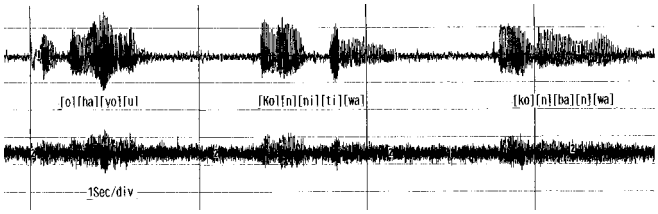

図４「おはよう,こんにちは,こんばんは」という音声の波形（上）と振幅圧縮の対数変換処理語の波形（下）
Fig.4 The lower waveforms are logarithmic transferred from the upper waveform on [ohayou], [konnitiwa], [konbanwa].

図５は振幅圧縮の処理をしない「おはよう,こんにちは,こんばんは」という音声のスペクトログラムである．他方，図６の対数変換処理をした音声のスペクトログラムである．定電流特性を得るために直列に挿入した抵抗を大きくしたので出力信号のレベルが小さくなり，図６のスペクトログラムの強度が小さく現れている．

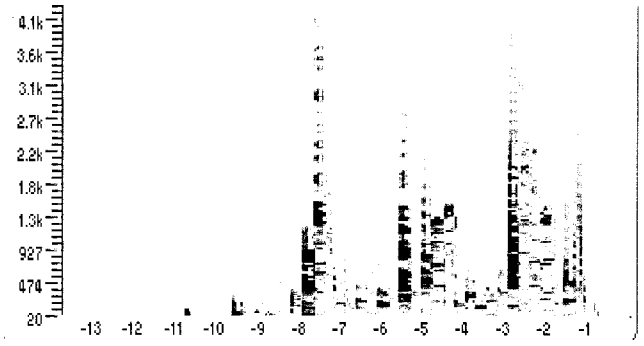

図５ 振幅圧縮処理をしない「おはよう,こんにちは,こんばんは」という音声のスペクトログラム
Fig.5 Spectrograms on original speech sound of [ohayou], [konnitiwa], [konbanwa].

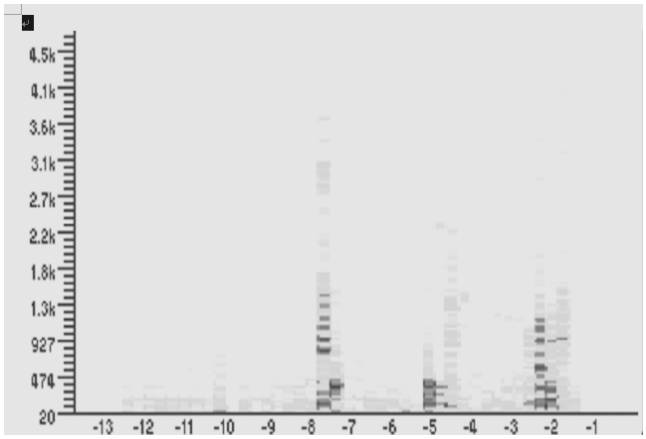

図6「おはよう,こんにちは,こんばんは」という音声の対数圧縮した波形のスペクトログラム
Fig.6 Spectrogram of legalistic transferred waveforms on [ohayou],[konnitiwa],[konbanwa].

3.2．自動利得調整回路による振幅レベル圧縮した音声のスペクトログラム

図７に示す自動利得調整回路[8]で振幅を圧縮して得たパルス状の音声波形を図８に示す．

図８に示すパルス列状の音声を聞いても人は何を話しているかを聞き取ることができる．

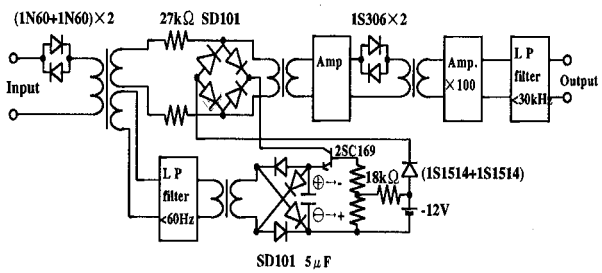

図８ 音声の振幅レベルを圧縮する自動利得調整回路
Fig.8 Circuit diagram for an amplitude compression.

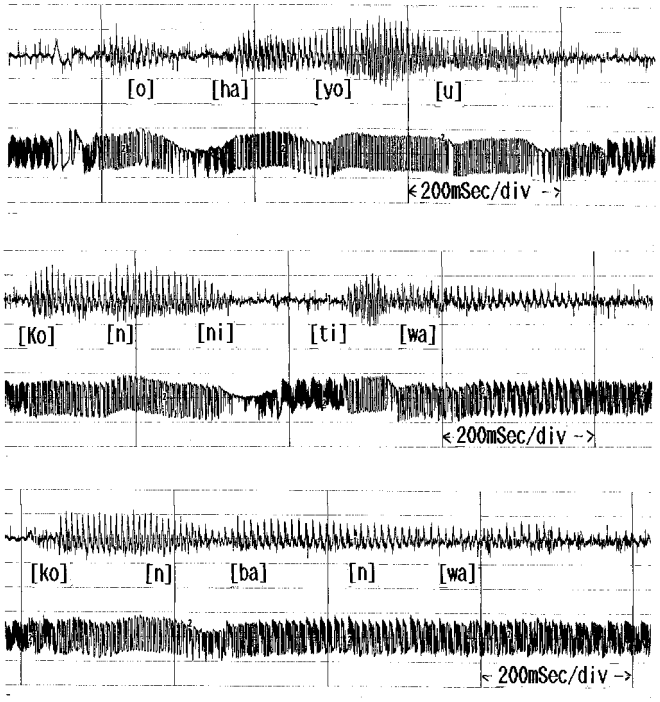

図7「おはよう,こんにちは,こんばんは」という音声の波形（上）と図８に示す自動利得調整回路によって得られたパルス化波形（下）
Fig.８ Lower waveforms are obtained from the upper waveforms on [ohayou], [konnitiwa], [konbanwa] by using an automatic gain control circuit shown in Fig.8.

図９ は，図７に示した「おはよう，こんにちは，こんばんは」という音声のスペクトログラムである．パルス化した波形では多くの周波数成分を含む．

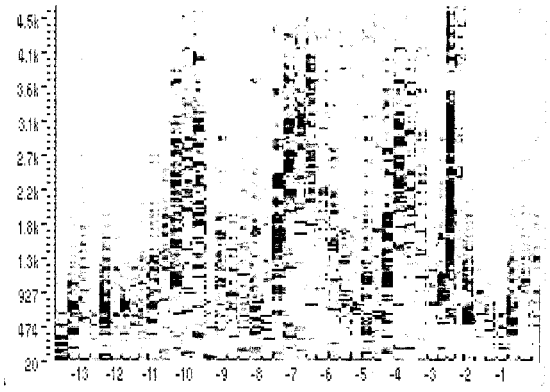

図9 「おはよう,こんにちは,こんばんは」という音声のパルス化音声のサウンドスペクトログラム
Fig.5 Spectrograms on speech of [ohayou],[konnitiwa], [konbanwa].

図 10. は，図７に示したパルス化した音声について 5kHz 以上の高周波域を除去し，その信号のサウンドスペクトログラムを求めた．まだ，この処理の適切な条件を出していないが，この方法でスペクトルグラムの認識に用いる２値データを採取することにした，

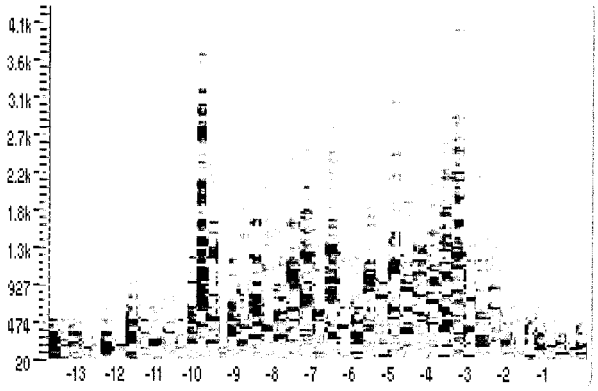

図10 5kHz 以上の高周波を除去したパルス化音声「おはよう,こんにちは,こんばんは」のサウンドスペクトログラム
Fig.10 The spectrogram on the waveforms on [ohayou], [konnitiwa], [konbanwa] those are obtained by 5kHz low pass filter.

4. 経験で認識機能を形成する半導体デバイス

4.1. 帰納法的な学習をする回路

符号器付き解読器(データ変換回路)の書き込み動作を図 11 に示す,書き込みには反転した信号によってホットエレクトロンを浮遊ゲートに注入して切断し，図 12 に示す如く，非反転信号で読み出し動作を行う．

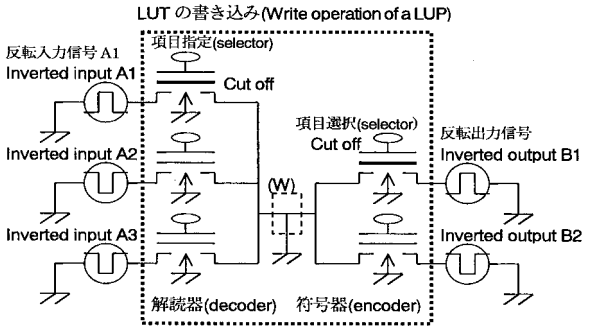

図11 データパターン変換回路の書き込み動作
Fig.11 Write operation of a programmable look-up table

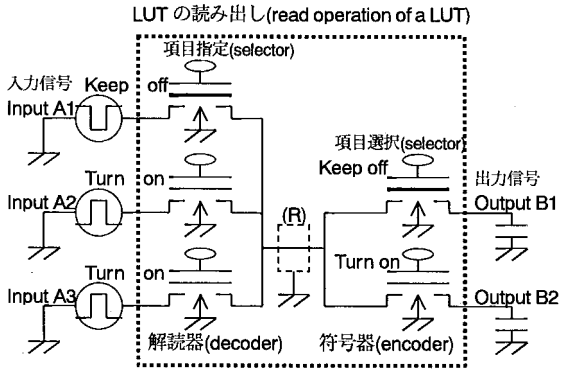

図 12 データパターン変換回路の読み出し動作
Fig.12 Read operation of a programmable look-up table.

図 11，図 12 の回路構成で同じ言葉のスペクトログラムのデータを書き込み，その書込みを重ねることにより，共通する成分だけを解読条件にする．

4.2. 認識の判定の閾値の制御

中央の解読器から符号器に接続する中央接続点は書き込み際して接地する回路を構成する．読み出しの過程において，信号源がＨレベルのとき内部抵抗に比較して，Ｌレベルの時の内部抵抗を小さくすれば中央接続点の電位はＬレベルの存在の影響を強く受ける．

中央接点の電位によって接続点に接続された入力群に含まれるＨレベルとＬレベルの比率がわかる．この点の電位をセンスアンプで検出して，一致度を判定して，出力を出す．

4.3. 並列する解読器群とのデータマッチング

最初の画像の空間位置が行列要素の位置で示される。二次元レジスタでは列に並べたデータをクロック信号で一行ずつシフトする際にはデータが同じ列で隣の行に転送される．

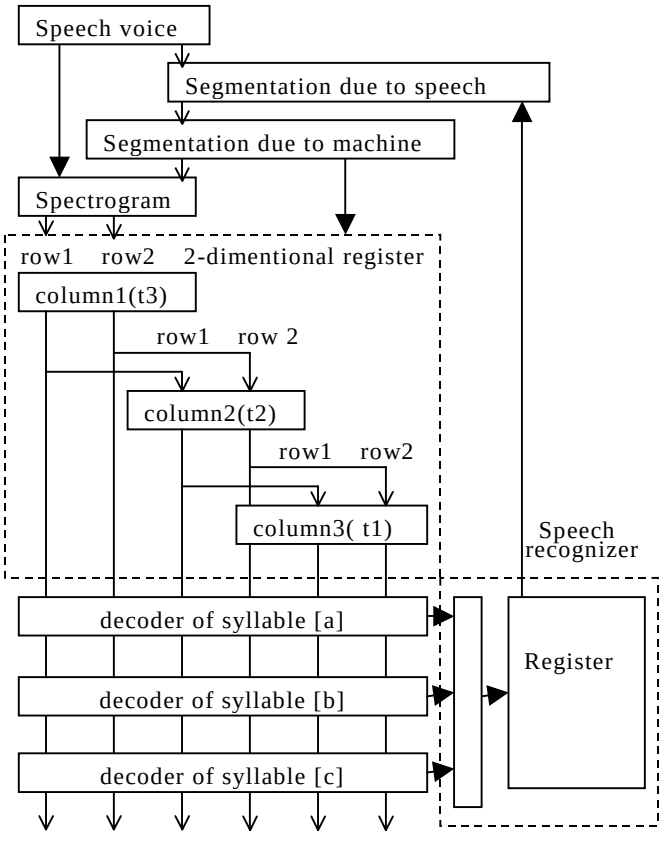

図９ 音節のサウンドスペクトルグラフをシフトするレジスタ-と発話を解読するためのパターン解読器
Fig.9 2-dimensional register for the pattern recognition of syllable on sound spectrogram

並列する解読器群に同時に照合する配線を平面上で行う際に、入力データが２次元のパターンであっても、そのデータは一次元上に載せなければならない．入力データを並列するｎ列に一度に載せて、そのデータを並列するｍ種の解読器に照合する間合いには，一つパターンを解読する解読器にはデータのバスと同数ｎの並列する行を入力に持つ必要があり、それをｍ種類、並列に並べるのでその接点数はｎ×ｍとなる．

4.4. 電荷を転送して知能回路を実現する技術

デジタル回路では演算のタイミングを合わせる機能のために状態を保持するといってもよい．組み合わせ論理は状態の変化が伝えられると同時期に演算を実行する．

今日,デジタル技術で使われる CCD をはじめダイナミック MOS　LSI では各ゲート電極の下（バックゲート）に電荷を引き込むことによって動作させている．そこで，活動単位を電荷で表現して知能を実現する回路システムに従来の LSI,VLSI,ULSI の技術が駆使される．また，図 11, 12 に示すパターン解読器は半導体不揮発メモリ（フラッシュメモリ）の設計製作技術によって実現可能である．

4.5. データパターン変換用 MOS LSI の規模

半導体不揮発メモリ（フラッシュメモリ）をデータパターン変換回路の接続点として集積回路の規模を簡単に推定みると，16 チャンネルの周波数帯で 5msec ごとに 100 回分（500msec）を音素の解読の処理をすると，レジスタに載せられるデータは 1600 個となり，2 値データとした解読器の入力条件との接続点総数は，最大で 2.56 万個となる．

500sec まで音素を参照すると，解読候補のデータは１行 16 個単位に横にシフトするレジスタで 1600 個のデータを 5 msec 毎（毎秒 200 回）照合する．

5.　音声言語の認識理解の帰納法的な理解

5.1. 言語構造の形成

活動単位を転送する神経回路の仕組みから同時に発生する活動だけが解読されるので，現在の活動単位のパターンが次ぎの活動を決める．神経回路網では活動単位が遅延されて転送されるので活動単位の部分集合とそれをまとめた概念を表す活動単位が回路網に中で同時に存在する．

活動単位の組織が神経回路網の中にできるとそれらの活動組織全体を解読する回路もできる．こうした活動単位の組織化によって，神経回路の情報処理の機能は高度化する．

神経回路網により構築される言語機能にはさまざまな活動単位が組織される．その活動の全てが言語活動に寄与している．単語には発音を担う活動単位もあり，文章には単語などを組み合わせた要素もある．そ

の言語表現の構造が神経回路網の形成に依存する[9].

言語活動を担う神経細胞が増し方式で回路網を形成して言語の機能が高度化する．他方，言語は現実の表現対象に依存しており，言語の普遍文法は言語表現対象の属性から説明できる[10].

会話では言語表現による刺激により脳神経系を活動させ，その活動と現在の活動状態により活動状態を更新し，次の入力でその活動状態が更に更新される．言語認識あるいは画像認識でも組織的な脳神経系の活動状態が遷移するように更新される.

5.2. 現実の理解と情報の創出

人間は言葉という共通の道具をアドリブで使って会話し，思考している[11]．その言葉や知識などは人間が存在しなければ存在しない人間が作ったものである．言葉を理解することを言葉で説明するとなると言葉の世界のなかに留まる.ところが,「ある感覚細胞が刺激に反応したということに意味がある．その感覚細胞の反応を伝える神経細胞が活動したことに意味がある．その神経細胞が筋肉を刺激して体を動かしたことに意味がある.」このように考えると，神経細胞の伝えるインパルスから情報を採取する必要がない.

人間社会で通用する情報の世界の中では情報を送り出す側には意図がないとは言えないし，情報を受ける側にも意図がないとは言えない．人間の持つ意思や意図には普遍性や一般性を持つとは限らない．それぞれ異なる場所を占める人間が変化しつつある状況にどのように対応行動をとるかを決めることが意思を持つことにつながる.

第３者的な立場で居られる事象には科学的に対処できる．科学は一般性と普遍性を備えているものとしている．また，言葉の世界においてもその要素となる単語などの理解は一般性と普遍性を求められている．また，現実に迫る音響や映像が提供されている.

しかし，情報や思考は実時間で変化を進行させている具体的な現実世界とは異なる．生きている個人はそれぞれの位置で活動し，その活動によって周囲の状況を変える．実際の人間は相違するので認識機能の個性がある．言葉を使わない神経回路網の領域における知能のしくみを科学することよって，人間の知能は帰納法的に形成され個性があることを知ることができる.

6. むすび

本報告で論議した認識システムは個別の個体が独自に認識の機能を形成する方式の知能装置である.

音声は時間と周波数成分を変数とする２次元の平面に表現できて，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．簡単な認識システムではデータをデジタルの値にしてその２次元のデータの認識する.

脳神経系の活動では神経細胞によってほぼ同時に発生したインパルスというしきい値論理の活動単位のパターンを一つのインパルスに情報圧縮できる．照合のためにデータを並列する線群（バス）に載せる１次元のレジスタでは２次元データの移動が行単位になる.

インパルスが同時に存在する並列線群を接続点群として，活動単位群を解読する回路ができる．この電子回路網によってパターンを認識するシステムが自動的に形成されるしくみを理解することができる．こうした認識回路により，機械に個性を持つ知能を付与することができる.

現段階では電荷転送方式の半導体集積回路で認識機能を書き込む LSI システムを実現するための準備段階である．本報告の音声認識装置のしくみは将来，自動翻訳電話機や発音を判りやすい発音に直す装置として発展させることができる.

文　献

[1] S. Karasawa, Dynamic MOS circuits for neuromorphic hardware implementation based on the paradigm of activity, Inter. Conf. on Computing, Communications and control technologies, Vol.5. Austin, Texas, pp.194-199, Aug. 2004.

[2] 唐澤信司, “活動単位を転送して３次元空間に活動を組織する神経回路網モデル”, 信学技報, MBE2004-52, pp.1-4, 2004.

[3] Ray D. Kent, Charles Read, “The acoustic analysis of speech”, Singular Publishing Group, Inc., 1992.

[4] 三浦種敏監修「新版聴覚と音声」電子情報通信学会、昭和 55 年

[5] 唐澤信司, “人間の視覚及び聴覚神経系に類似したインパルス転送回路網モデル”, 信学技報, HIP2003-77, pp.1-6, 2003.

[6] S.Karasawa, “The dialectical architecture of visual intelligence where every activity is available as a tool for the next activity”, Abstracts, pp.238, ECVP, A Coruna, Spain Aug.22-26, Website presentation on ECVP2005.

[7] 風間喜代三、上の善道、松村一登、町田健「言語学」東京大学出版会 pp.220, 1993.

[8] 唐澤信司、具龍会、閔泰泓「レベル圧縮した音声信号による音声認識」東北大学電気通信研究所、第 281 回 No.2 音響工研、第 11 回 No.2ME 研究会,1996.

[9] 唐澤信司, “活動単位により活動単位群を統合して言語表現する神経回路モデル”, 信学技報, TL2003-14, pp.79-84, 2003.

[10] 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,

[11] S. Karasawa, Model of linguistic activities as Ad hoc interactive activities in an impulse driven multi-agent system, The 7th world conf. on Systemics, Cybernetics and Informatics, Vol.14. Orland, pp.365-370, Jul. 2003.