解　説

ファジィ理論を用いた音声認識†

藤本　潤一郎††

1. まえがき

古来，人間は，身近な意志伝達手段として声を使ってきた．科学の進歩とともに音声で機械を動かそうとする試みが出てくることは自然である．しかし，現在のところ，SF 映画で見られるような，コンピュータと人間が話し合うところまでは，技術的に進んでいない．文章や会話の音声認識装置が商品として市場を賑わすにはもう少し時間がかかりそうである．

連続音声の認識を難しくしている原因には，調音結合，音の脱落や余分な音の挿入などがあり，音声の特徴パラメータが特定しにくいことがある．調音結合は，注目する音の特徴がその前後に発声された音の影響を受けて変形されるものである．したがって，文章や会話など連続的に発声された音声を認識するには，誤りを含んだ結果の列から，知識や音の連鎖頻度を用いて修復するような技術が必要となる．

一方，単語の音声認識では一つの単語全体を単位として認識するため，単語の内部で調音結合の影響があっても，単語全体としては，常に同じように影響が現れるのであまり問題とはならない．このような理由で単語音声認識が実現しやすく，実用化も進んでいる．

認識のための手法には，統計的手法を用いるもの，知識をルール化して使うもの，最近ではニューラルネットによるものなどがあるが，実用化されているもののほとんどはパターンマッチングを使うものである．**図-1** に音声認識におけるパターンマッチングのブロックを示す．入力された未知の音声の特徴パラメータをあらかじめ登録されている標準パターンと比較し，もっとも類似性の高いものを選びだして認識結果とする．図では破線で示しているが，必要に応じて学習により標準パターンを作成する．

パターンマッチングを行う上での二つの大きな問題点は，(1)同じ単語でも，発声速度などの影響で単語パターン内部の音韻の長さやパターン全体の時間長が異なること（時間変動），(2)発声する人によって特徴パラメータの周波数軸上の変動があること（周波数変動），である．(1)の解決のためには動的計画法を導入した DP (Dynamic Programming) マッチング[1]が提案されている．この方法は二つのパターンの比較に際し，両者の誤差が最小になるように照合するもので，要する計算量は多いが，精度がよく，現在も改良がなされ研究が続けられている．それに対し，(2)の解決策は，決め手を欠いており，一つの単語にいろいろな種類のパターンを登録する方法[2]や，統計的な手法[3]を使うものなどがある．

だれの声でも特別の準備をすることなく認識できる方式を不特定話者音声認識方式，あらかじめ利用に先立ちオペレータの声で装置をトレーニングしてから利用するものを，特定話者音声認識方式という．特定話者方式では，発声者が限定されるので話者間の変動がなく，認識が容易になることはいうまでもない．現在実用化されている音声認識のほとんどが特定話者方式であったり，不特定話者方式ではわずかな単語数しか認識できないのはこのような理由からである．

2. ファジィの導入

ファジィ集合の概念を音声認識に使うことは，“音韻には 5 kHz の成分がある”という代りに，“音韻には高い周波数の成分がある”という表現を可能にする．したがって，音声のような変動要因の多い特徴パラメータを扱ううえでは有効な考え方であろう．

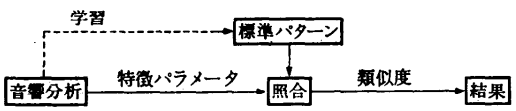


図-1　パターンマッチングのブロック図

† Voice Recognition Using Fuzzy Theory by Jun-ichiroh FUJIMOTO (RICOH Co. Ltd., Research and Development Center).

†† (株)リコー中央研究所応用技術開発センター

音声認識の分野にファジィ集合を導入したのは，1970年半ばの De Mori[4] や Pal[5] が最初であると思われる．De Mori はその後もファジィを使った音声認識の研究をしており，それらをまとめた著書[6] もある．その後，筆者らはファジィパターンマッチングによるパターン変動を含んだままでの認識方式[7] を開発して，認識装置を実用化した[8]．ファジィ理論を用いた音声認識装置の実用化はこれが最初であろう．近年では音声認識の分野でもファジィの利用例が目に止まるようになってきている．

3. ファジィパターンマッチング[9]

3.1 ファジィパターンマッチングの原理

標準パターンとして登録された単語名の集合を I，その一つの要素を i とし，x_i を，単語 i を発声したときの特徴パラメータから作られた特徴パターンとする．この特徴パターンはいろいろな要因の変動を受けるため，これをファジィ集合 X_i とする．また，X_i への帰属性を表すメンバシップ関数を m_i と定義する．

さて，未知の単語音声が入力されたとき，この特徴パターン y と x_i の類似度 S_{yi} を求め，S_{yi} $(i \in I)$ を最大にする単語名 i を結果として出力する．ファジィパターンマッチングでは S_{yi} の計算に際し，ファジィ的手法を用いることが特徴である．S_{yi} は

$$S_{yi}=(m_i \wedge y)/(\overline{m_i} \wedge y) \qquad (1)$$

ただし，$\overline{m_i}$ は単語パターン X_i の補集合を定義づけるメンバシップ関数である．分子は入力された未知の特徴パターン y の X_i への帰属性，すなわち y の単語 i らしさの度合いを表し，分母は逆に，単語 i らしくなさの度合を表す．これらの比によって両パターンの類似度を定義する．

3.2 メンバシップ関数の作成

音声認識の特徴パラメータとして，いくつかの種類が提案されているが，ここではスペクトルの時間変化を表すパターン，TSP (Time Spectrum Pattern) を使って特徴パターンを作る．TSP の一例を図-2 に示す．これは男性が発声した「イチ」という単語音声を周波数上で 15 サンプル，時間的には 10 ms に 1 回ずつサンプルしたもので周波数成分の大きさが 16 進で数値化されている．

TSP 上で特に注目すべき共振周波数の近辺を「1」，他を「0」として2値化したものを BTSP (Binary TSP) と呼び，特徴パラメータとする．先の図-2 に示した TSP から作った BTSP の例を図-3 に示す．

```
 0  0  0  0  0  0  0  0  0  0  3  c  a  3  4  0
 8  f  d  4  1  0  1  3  6 11 15 39 44 17  e  5
 f 18  f  7  4  5  7  b 11 22 30 6f 9c 36 22  f
19 20 10  7  6  7  a  f 18 32 3d 8b c3 45 2d 17
18 2b  f  9  8  9  c 13 20 4c 47 93 dd 4e 35 1e
12 29  e  a  9  a  d 12 1f 46 4c aa b9 48 38 19
16 30 10  9  9  a  a  f 1a 3d 43 a4 79 30 28 15
11 37  f  7  8  9  a  e 17 32 40 b0 6f 30 23 14
 f 32  c  4  5  6  7  b 12 25 35 8e 6b 2c 1f 11
 9 25  7  3  2  4  6  9  f 1e 2d 79 5f 2f 20  e
 3 15  2  0  0  0  1  4  8 14 1a 46 3a 23 1a  7
 0  6  1  0  0  0  0  0  0  4  3  e  9  2  1  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  1  2  7  8  b 13 2a 2e 29 2e  3
 0  0  1  1  4  6  9  c 12 1a 2b 65 91 88 8a 15
 0  2  3  2  4  6  8  c 10 17 26 47 84 a5 ae 25
 0  0  0  0  0  2  4  6  a  f 1d 46 56 73 75 2a
 0  0  0  0  0  0  1  3  6  a 17 38 4c 3d 42 24
 0  0  0  0  0  0  0  1  4  9 14 37 3d 2a 24 20
 0  0  1  0  0  0  0  2  6  d 16 3b 46 1e 15 25
 0  1  4  0  0  1  3  5  a 15 25 6f 3a 1f 16 39
 0  0  5  0  0  0  2  4  8 12 1f 62 38 1a 14 3c
 0  0  6  0  0  0  1  4  7 10 1d 5a 38 16  d 3d
 0  0  5  0  0  0  1  3  7  f 1b 57 31 12  a 3b
 0  0  5  0  0  0  0  1  4  d 13 3f 26  c  7 2c
 0  0  6  0  0  0  0  0  3  c  d 24 17  6  4 1c
 1  1  9  0  0  0  0  0  1  a  8 16 17  6  5 14
 0  2  d  1  0  0  0  0  2  9  d 26 32  e  9 19
 0  3  d  0  0  0  0  0  3  b  c 1e 33  e  8 15
 1  7  b  0  0  0  0  0  4  7 14 25  a  6  e
 4 10  5  0  0  0  0  0  0  3  2  a  a  2  3  a
 7  f  1  0  0  0  0  0  0  0  0  0  0  0  0  3
 3  6  0  0  0  0  0  0  0  3  0  1  0  0  0  0
 0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

図-2 「単語（イチ）」の TSP

```
...........11.1.
.11.......11...
.1.........11...
11.........11...
.1.......1.11...
.1.......1.11...
.1.........11...
.1.........11...
.1.........11...
.1.........11...
.1.........11...
.1.......1.11...
.1.........11...
................
................
................
................
................
................
................
.......1...11.1.
....111....1111.
..1.........111.
......1....1111.
............111..
...........11...
..1.........11...
..1.........1....
..1.........11...
..1.........11...
..1.........11...
..1.........11...
..1......1111...
..1......1.11...
..1.........11...
..1......1.11...
.11.........11...
.1..........11...
11..............
11.......1.1....
.1..............
```

図-3 図-2 のパターンを2値化して作った BTSP

図-3 ではパターンを見やすくするため，「0」を "." としている．BTSP の採用により，(1)データ量が少なくなる，(2)1フレーム単位のデータがコンピュータの中で取り扱いやすい，といったメリットが生じる．

単語 i の BTSP の集合が X_i であることから，メンバシップ関数は次の手順で作成することにした．

各単語ごとに，多くの種類の BTSP を作り，時間長を一致させて重ね合わせた上で，対応する要素の和をとっている．図-4 は単語「イチ」のメンバシップ関数の例で，コンピュータでの扱いやすさから1要素を4ビットで表している．

3.3 音声認識システム

図-5 に本認識システムのブロック図を示す．特徴抽出部，2値化処理する部分，辞書部，照合部で構成

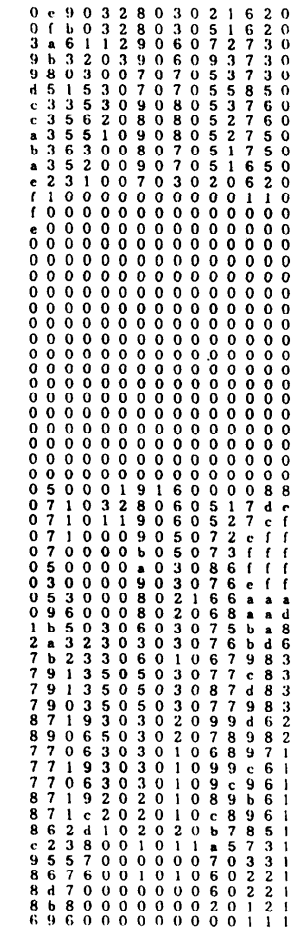


図-4 メンバシップ関数例（単語名「イチ」）

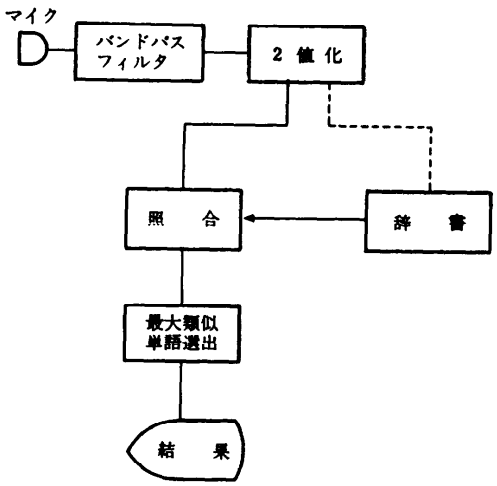


図-5 ファジィパターンマッチングを利用した認識システムのブロック図

されている．図中の破線は辞書へ音声パターンの登録を表している．つまり，このシステムは，利用者が自分の声で装置をトレーニングした場合には特定話者装置として使うことができる．そのためにはメンバシップ関数が，装置の中で組み立てられる必要がある．さらに，この装置では特定話者方式でも不特定話者でもない，その中間的な話者方式であるグループ使用が実現できる．

不特定話者方式，グループ使用，特定話者方式において認識実験をすると，同じ 110 単語を使って，おのおので約 2.5% ずつ認識率が上昇する．ただし，特定話者方式では不特定話者方式に比べて変動も少ないことから，1要素を2ビットで表現している．

この装置の特徴は，不特定話者方式で認識できる単語数も多く，式(1)で定義されるような AND と OR だけの簡単な演算で実現できるため，装置も小型化できる点にある．

4. ファジィ木探索による音声認識[10]

次に，音韻判別のルールを使って子音を認識する例を紹介する．

4.1 ファジィ木探索による子音認識の原理

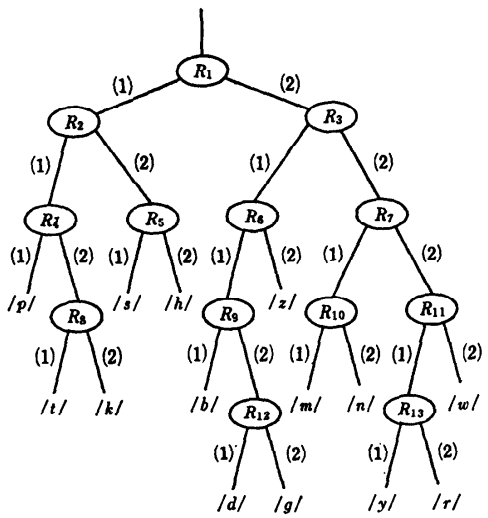
子音の特徴量は調音結合の影響が大きいため，後続母音別に子音判定のルールをもっている．図-6 に後続母音が /a/ のときの子音認識の木を示す．各ノードの R_1～R_{14} は，図の下部に示されているような特徴を判定するルールで，ファジィ木探索[11]を行う．つま

り，各ノードで音素の判別ルールを適用する際に，二つの枝への属性をメンバシップ値として計算する．したがって，どちらの枝がどの程度確からしいかが決められることになる．ルールは根から葉へと順に適用されるから，それぞれの論理の積をとることになり，確信度 Cf は次のような式に従って得られる．

$$Cf_j=\min_k\{Cf_k\} \qquad (2)$$

ただし k は各ノードにおけるカテゴリ，Cf_j は音素 j としての確信度である．このような計算を進めるうちに，ある音素としての確信度が一定値以下となってしまうと，それ以上探索を続ける必要がなく，打ち切ってしまえばよい．

各ノードでの確信度 Cf を表すメンバシップ関数値は $\mu_1, \mu_2, \cdots \mu_p$ と表現し，次のように，$p-1$ 個のしきい値と比較して決定される．



Consonant category for each branch
R_1 (1) Voice-less (2) Voiced
R_2 (1) Explosive (2) Fricative
R_3 (1) Fricative or Explosive
(2) Nasals or Liquids
R_4 (1) Bilabials (2) not Bilabials
R_6 (1) Alveolars (2) Glottals
R_6 (1) Explosive (2) Fricative
R_7 (1) Nasals (2) Liquids
R_8 (1) Alveolars (2) Velars
R_9 (1) Bilabials (2) not Bilabials
R_{10} (1) Bilabials (2) Alveolars
R_{11} (2) Bilabials (2) not Bilabials
R_{12} (1) Alveolars (2) Velars
R_{13} (1) Palatals (2) Alveolars

図-6 後続母音が /a/ のときの子音判定の木構造

$$Cf_j=\begin{cases}\mu_p; & \sum_{i=1}^{n}\omega_i\chi_i\geqq\sigma_{p-1}\\ \mu_{p-1}; & \sigma_{p-1}>\sum_{i=1}^{n}\omega_i\chi_i\geqq\sigma_{p-2}\\ \cdots\cdots & \\ \mu_1; & \sigma_1>\sum_{i=1}^{n}\omega_i\chi_i\end{cases} \qquad (3)$$

ただし，しきい値は $\sigma_1>\sigma_2>\cdots>\sigma_{p-1}$，$\chi_i$ は入力パラメータの値，n はパラメータの数，ω_i は χ_i の重み係数である．つまり，音声から得られたパラメータに，ある特徴が判別しやすくなるような重みをつけて和を求める．その結果をしきい値と比べて，μ の値を決め，確信度とする．

4.2 認識システム

もとの論文[10]では示されていないが，認識の構成をブロック図にすると図-7 のようになるであろう．

特徴パラメータはケプストラム係数，線形予測係数，など合計 82 個が用いられる．認識は，音声の特徴パラメータを取り出したあと，その中を音素の境界で区切ってセグメンテーションする．しかし，正確にセグメンテーションすることが困難であるため，複数の候補を作っておく．さらに，セグメンテーションされた中から比較的容易な 5 母音と撥音を認識して，この結果から単語辞書でつづりを調べ，子音が決定されないまま候補単語をしぼる．候補単語から判定すべき子音をファジィ木探索法で判定する．最終的な単語の確信度は，セグメンテーションの確かさ，母音認識の確かさ，子音認識の確かさから判定される．

学習は 72 単語を発声して行う．この 72 単語は，さまざまな調音結合を含み，すべての後続母音ごとの子音を網羅するように作られた，単語セットである．学習で，母音の標準パターンや，子音判別のための重み係数などを求めておく．認識は JR の駅名 100～320 単語を使って実験しており，おのおの 91～97% の結果を得ている．

従来の二進木探索法では，木の根に近い部分の判定結果が最後の結果に与える影響が大きかった．しかし，ファジィ木探索法では，二つの枝のそれぞれへの

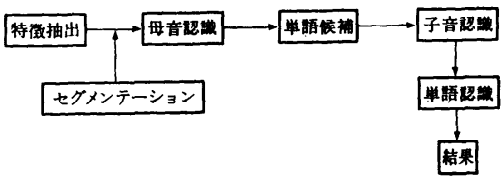


図-7 ファジィ木探索を利用した認識システムのブロック図

確信度を求めて進むため，途中で正解が失われることが少ない．さらに，ファジィ木探索によると最終結果の第1，第2候補を作ることができるので，誤認識をしても，人間がカバーしやすい．

5. ファジィ関係を利用した音声認識[4]

ここでは日本語の VCV（母音-子音-母音）と類似したシラブル（pseudo-syllable）を認識単位として使っている．VCV は，子音の特徴が前後の母音によって変形されることを考慮して，前後の母音を含めて認識単位としたものである．したがって，単音節（CV）に比べると，その前に付く母音の組合せ分だけ種類が多い．

5.1 認識の原理

発声された音声をセグメンテーションし，そのスペクトルパターン ω_i から仮説をたてて シラブル $S_{i1}\cdots S_{wi}$ に分類するとき，その仮説があいまい性をもつと考え，ファジィ集合 H_i とする．

$$H_i=\mu_{i1}/S_{i1}+\cdots+\mu_{ji}/S_{ji}+\cdots+\mu_{wi}/S_{wi} \quad (4)$$

論文中[4]ではメンバシップ関数値 μ の決め方は具体的に述べられていないが，シラブル S_{ji} の音響的特徴を知識として与えておいて決めるようである．

認識は，まずセグメントされたスペクトルパターン ω_i から ディスクリプタによって解釈結果 D が出力される．このときの ω_i と D の関係をファジィ関係で R_1 (ω_i, D) と表し，その関係の強さを μ_{R1} (ω_i, D) で表す．この結果はトランスジューサで G に変換され，さらに，G からシラブル S を出力する．これらの関係を R_1 同様に R_2, R_3, その関係の強さを μ_{R2} μ_{R3} とすると，最終的な認識結果は

$$\mu_R=\underset{G}{V}\left(\underset{D}{V}(\mu_{R1}(\omega_i, D)\wedge\mu_{R2}(D, G))\wedge\mu_{R3}(G, S)\right) \quad (5)$$

の確信度をもって得られることになる．

5.2 単語辞書の構成

この方式では特定のシラブルが認識できたところで単語名が分かるようにして，演算を少なくできるように単語辞書を構成している．そのために，おおまかな特徴を表すクラスを決め，各シラブルがどのクラスに属すかを表しておく．クラスとはたとえば，

a=AN+EN+O
b=BA+CA+CO+DO+GE+PA+PO
　+QUA+TO
c=ANO+ENE+ENO+IMO+ENE+OMA
　+OMO+ONA+ONO
……
f=FI+SA+SSO+VE+ZA+ZE+ZIA

である．$a, b, c\cdots$ がクラスで，＝ の右側がそのクラスに属するシンボルである．このように，音響的特徴が類似しているものを同じクラスに入れておいて，単語の辞書はシラブルの連結だけでなく，クラスとシラブルの両方で表す．しかも辞書に登録するすべての単語が決まると，クラスの認識できた段階で単語の仮説がたてられるものや，あるいは，クラスも知る必要がないものが分かる．たとえば，論文中では単語 VENEZIA は次式のように登録している．

$$\begin{aligned}H(\text{VENEZIA})=&f(\text{VE})+c(\text{ENE})+f(\text{ZIA})\\&+*(*)c(\text{ENE})+*(*)f(\text{ZIA})\\&+c(\text{ENE})*(*)+f(\text{VE})*(*)\\&+f(*)c(*)*(*)\end{aligned} \quad (6)$$

式(6)で $*(*)$ はセグメントさえできればクラスやシラブル名は知る必要のないもの，$x(*)$ はクラス x が分かればよいものを表している．式(6)からセグメントがうまくできなくても VE, ENE, ZIA のどれか一つのシラブルが分かれば，単語 VENEZIA は認識できるし，三つのシラブルのセグメントができるならば，先頭から f と c のクラスに属していることが分かった時点で認識できることになる．

5.3 認識システム

認識システムは図-8 に示すように2段階になっており，第1段階のほうがクラス認識，第2段階がシラブル認識である．まず，セグメントされたスペクトルパターン ω からおおまかな分析によってクラス ξ が仮定され，次に，クラスと ω から式(5)によってシラブルが認識される．これらと単語辞書によって最終結果を得る．

この方法では，シラブルの認識で生ずる不確かな情報は不確かなままで結果に導く．また，単語として認識するための必要最低限の情報が何であるかをもっているため，効率的に，確信度に応じた検索ができる．

もとの報告には認識実験の結果がなく，どの程度の

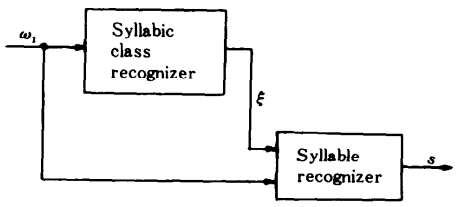


図-8　ファジィ関係を用いた音声認識のブロックダイヤグラム

ものかは分からないが，シラブルの記述に冗長性が多いとき，つまり登録単語数の少ないときには相当の効果が期待できる．ただし，単語数が多くなったり，文章や連続発声された音声の認識をする場合，シラブルの記述に冗長性が少なくなると式(6)から(*)で表される表現が少なくなって，効果が乏しくなる．

6. あとがき

本解説では，単語認識を中心にしてファジィ理論が導入されている例を述べた．ここに紹介した例は，すでに実用化されているもの，初期に研究されていたものなどそれなりに意義のあるものだと思われる．

現在，音声の認識課程の中の二つの段階でファジィ理論が使われている．第1に，ファジィパターンマッチングや音韻の特徴を定性的な表現で表すもののように，変動を含む特徴パラメータを分類する段階で利用するものである．第2に，分類された結果の確からしさを考慮に入れて扱うようにしたものである．前者は変動をともなうパターンの取扱い，後者は不確さをともなったものの解釈を目的としている．さらに第3として，音声認識の技術がさらに進み言語処理を含む段階で，言語のもつあいまいさの処理にも使われるようになるであろう．ファジィと音声認識にはこのような三つの接点ができると考えられる．

ファジィ理論を使う上で常に問題になるのがメンバシップ関数の決定のしかたである．質的な事象では多くの人からアンケートを取り，その結果からメンバシップ関数を決めていくやり方[12]がある．しかし，たとえば，音声認識における“高い周波数”とは，どのような関数で表せば良いかが一定ではなく，この決め方が性能に影響することがある．このあたりが，一部の人から非難を受けたりもする．3. で述べたものは，認識装置の中でメンバシップ関数を形成するようになっているが，音声認識の場合，話者に応じてメンバシップ関数が修正されるほうが，有効な結果が得られることが多い．最近ではニューラルネットを使ったメンバシップ関数のトレーニングの報告もあり[13]，今後もこのような研究が必要であろう．

参考文献

1) 迫江，千葉：動的計画法を用いた音声の時間正規化に基づく連続単語音声認識，日本音響学会誌 27，9，p. 483 (1970).
2) 中津，長島，小島，石井：電話音声の認識方法，電子通信学会論文誌，J66-D，4，p. 377 (1983).
3) 新田，村田，松浦，斎藤：複合類似度法を用いた不特定話者の単語音声認識，電子通信学会論文誌，J67-A，11，p. 1076 (1984).
4) De Mori, R. and Trasso, P.: Lexical Classification in a Speech Understanding System Using Fuzzy Relation, ICASSP p. 565 (1976).
5) Pal, S.K. and Majumder, D.D.: Fuzzy Sets and Decision Making Approaches in Vowel and Speaker Recognition, IEEE Trans. Syst. Man. Cyb. SMC-7, 8, p. 625 (1977).
6) De Mori, R.: Computer Models of Speech Using Fuzzy Algorithm, Plenum Press (1983).
7) 藤本，中谷，米山：2値の TSP を用いた単語音声認識方式，日本音響学会講演論文集，3-1-8, p. 195 (1983).
8) Fujimoto, J. et al.: Speaker-independent Word Recognition Using Fuzzy Theory, IFSA Congress '87, p. 819 (1987).
9) Fujimoto, J., Nakatani, T. and Yoneyama, M.: Speaker-independent Word Recognition Using Fuzzy Pattern Matching, Fuzzy Sets and Systems, 32, 1 (1989).
10) 森島，原島：音響処理と記号処理を融合した単語音声認識システムの構成，電子情報通信学会論文誌，J70-D，10，p. 1890 (1987).
11) 森島，原島：統計的手法に基づくプロダクションルールの自動抽出法とファジィ木探索，電子通信学会論文誌，J69-D，11，p. 1754 (1986).
12) 中島：メンバーシップ関数の構成，和歌山県立医大紀要，13，p. 63 (1983).
13) 林，高木：神経回路網モデルによるファジィ推論の定式化，4th Fuzzy System Symposium, p. 55 (1988).

(平成元年4月12日受付)