音声言語情報処理　25－4
（1999．2．5）

音声データの自己組織化と音声検索システム

遠藤 隆*，張　建新**，中沢　正幸***，岡 隆一*

*新情報処理開発機構　つくば研究センタ
**メディアドライブ株式会社
***現在，シャープ（株）

〒305-0032 つくば市 竹園1-6-1 つくば三井ビル13F

あらまし　我々は，音声や動画像などの効果的な検索を行うためのに，データを自己組織化する手法として，Incremental Path Method[1]を提案し，自己組織化されたデータベースを高速に検索する連続オートマトン[2]を提案した。またテレビニュースの音声検索に適応し，特定話者音声をクエリーにして高速に検索に検索できることを示した[3]。今回，特定話者の音声をクエリーとした音声検索システムを提案する。

Self-Organizing of Speech data and Searching System

Takashi ENDO*, Zhang Jianxin**, Masayuki NAKAZAWA***, and Ryuichi OKA*

*Real World Computing Partnership Tsukuba Research Center
**Media Drive Corporation, ***Now, Sharp Corporation

Tsukuba Mitsui Building 13F, Takezono 1-6-1, Tsukuba
Ibaraki 305-0032, Japan

Abstract　In a previous work we proposed the Incremental Path Method that can construct a self organized network database from a sequence of label. And we proposed high-speed searching system for speech database with speaker depending speech query using Incremental Path Method. In this paper we propose a speech retrieval system with speaker independent speech query.

1. はじめに

コンピュータ技術の発展や，WWWの普及により，人々が手にすることができるデジタル化された文字，音声，画像などの情報は増加の一途を辿っている。これらの膨大なデータに対して，Altavista[1]，goo[2]などのキーワードベースの全文検索システムを組み合わさることで，有用な情報源として役に立つ場合も多い。

MPEGなどに代表される動画像のコーディック技術や，マイクロプロセッサの性能向上により，画像を含むマルチモーダルデータがデジタルデータとして蓄積可能な環境が整いつつあるが，動画像データはデジタルデータとして蓄積されていても，そのままでは検索することは非常に困難である。

そこで，動画像データは音声を伴っている場合が多いことに着目し，音声を利用した検索とを考える。大量のデータに対して検索を行う場合，すでに特定話者のクエリーを用いた検索は，量子化された音響特徴量系列をIPMネットワークで自己組織化することにより，コンパクトに表現され，高速に検索が可能であることを報告している[3]。今回は，IPMネットワーク不特定話者音声をクエリーとしてサーチするシステムを提案する。今回提案するシステムでは，検索対象となる音声データデータベースに対して，フレーム単位の音素認識を行い音素系列に変換したのち，IPM手法により自己組織化ネットワーク表現に変換する。IPMネットワーク表現されたデータベース中からの検索も，クエリー音声をフレーム単位の音素認識により音素系列に変換し，連続オートマトン手法により，フレーム同期でスポッティング検索を行う。

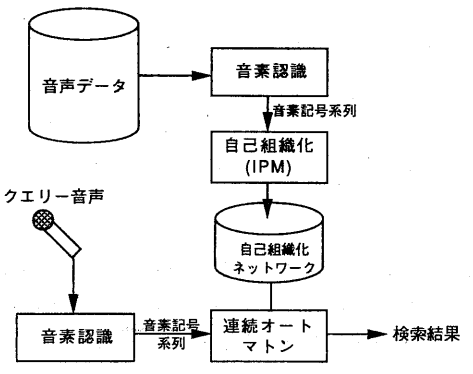

図1　音声検索検索システム構成図

2. IPMネットワーク

入力系列に従ってインクリメンタルにネットワーク構造を構築する手法としてIncremental Path Method（以下IPM）を提案している[9]。IPMでは，ラベル系列中に存在する同一の入力ラベル系列に対応するネットワーク上のトレースが同一になるようにネットワークが構築されるという特徴を持つ。

今回はデータベースを表現するラベル系列として，フレーム毎に33種類の音素識別認識を行い，その結果得られる音素系列を用いる。

2.1 音声のラベル系列化

フレーム毎に音響特徴量を計算し，ベクトル量子化し，33種類の音素に対する事後確率を求め，最も高い事後確率の音素を当該フレームの音素ラベルとした。このように得られたラベル系列を用いて，自己組織化を行う。音響特徴量が得られたときの音素ラベルの事後確率の学習には，ATR連続音声データベースの話者9人分のデータ[11]を用いた。音響特徴量は，メルケプストラムおよびスペクトルベクトル場[6]の方向性パターン[7][8]を認識実験により比較し，スペクトルベクトル場の方向性パターンを用いることとした。図2に，量子点数とフレーム単位の不特定話者データに対する音素認識率を示す。

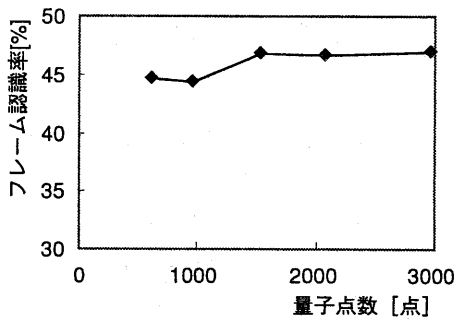

図2　量子点数とフレーム認識率

【セントロイド作成アルゴリズム】

t:時刻

C_n:セントロイドベクトル

N:セントロイドベクトルの数

I_t:入力ベクトル

d:許容量子化誤差

(i) $N = 0, t = 0$

(ii) if $\exists n$ minimum of $n < N$ so that

$$|I_t - C_n| < d$$

then goto (iv)

else goto (iii)

(iii) $C_N = I_t, N := N + 1$

goto (iv)

(iv) $t := t + 1$

if $t \leq T$ then goto (ii)

else stop

2.2 IPMネットワークの構築

ＩＰＭで用いるネットワークは，ノードと有向アークから構成され，アークには1つのシンボルラベルが対応付けられている。1つのノードペア間には1つのアークが対応付けられている。

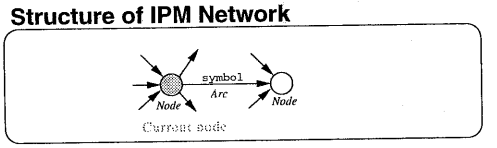

図3　IPMネットワーク基本構造

音声の定常部などを効率良く表現し同時に伸縮を許すための，自己ループ，既存のアークを飛び越すナル遷移（ϕ）を生成する。

したがって，今回用いるIPMでのネットワーク構築の基本戦略は，

(1). 現在の状態から現在の入力に対応するアークがあれば，それを使って遷移する

(2). 同一のシンボルが連続する場合は，自己ループを形成する

(3). 現在の状態から現在の入力に対応するアークが無い場合，次の入力シンボルに対応するアークが存在するノードに遷移する。

(4). 既存のアークをひとつ挟んだ先に現在の入力に対応するアークが存在した場合，ナル遷移アークを生成する。

の4点にまとめられる。

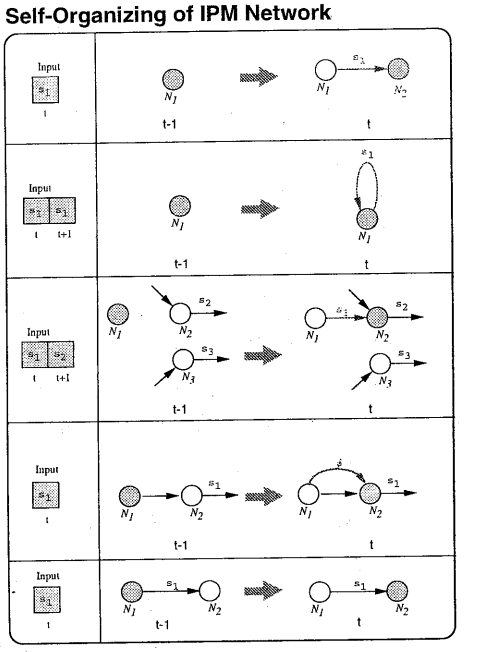

図4 IPMネットワークの自己組織化

【IPMによる自己組織化ネットワークの構成アルゴリズム】

N:ノードセット

M:ノードの数。$M = |N|$

L:1つのノードが持つことができるアークの最大数

A_{ij}:ノードiからノードjへのアークに対応したラベル

P：新たにアークを作成する際に参照するラベルの先読み数の最大値

K:ラベルセット。$A_{ij} \in K, \varnothing \in K$

$\varnothing$:null ラベル

$u(t)$:時刻tにおける入力ラベル

$1[x]$:$1[x] = 1$ if $x \neq \varnothing$, $= 0$ otherwise

(i) $t := 0$ *and* $\forall i, j \in N, A_{ij} := \varnothing$

(ii) $i_0 := 1$

(iii) $t := t+1$

(iv) $J \equiv \{ j | A_{i_{t-1}, j} = u(t) \}$,

if $J = \varnothing$ then goto (vii)

else goto (v)

(v) if $\exists j_t$minimum of $j \in J$ and $\exists k$ so that

$$A_{j_t,k} = u(t+1) \text{ and } \sum_{m=1}^{M} 1[A_{k,m}] \le L-1$$

then $i_t := j_t$ and goto (iii)

otherwise goto (vi)

(vi) if $\exists j_t$minimum of $j \in J$ so that

$$\sum_{m=1}^{M} 1[A_{j,m}] \le L-1$$

then $i_t := j_t$ and goto (iii)

otherwise goto (viii)

(vii) if $\exists j_{new}$minimum of $j \in N$ and

$\exists p$ maximum of $1<p<P$ and

$\exists k_{1\cdots p}$ so that

$$A_{i_{t-1},j_{new}} = \varnothing \text{ and } \sum_{m=1}^{M} 1[A_{i_{t-1},m}] \le L-1$$

$$k_0 := j_{new}$$

$$A_{k_0,k_1} = u(t+1) \text{ and} \cdots$$

$$\cdots \text{and } A_{k_{p-1},k_p} = u(t+p)$$

$$\text{and } \sum_{m=1}^{M} 1[A_{k_p,m}] \le L-1$$

then $A_{i_{t-1},j_{new}} = u(t)$, $i_t := j_{new}$ and goto (iii)

otherwise goto (viii)

(viii) if $\exists j_{new}$minimum of $j \in N$ so that

$$A_{i_{t-1},j_{new}} \neq \varnothing \text{ and } A_{j_{new},k} = u(t)$$

$$\text{and } \sum_{m=1}^{M} 1[A_{k,m}] \le L-1$$

then $i_t := j_{new}$ and goto (iii)

otherwise goto (ix)

(ix) if $\exists j_{new}$minimum of $j \in N$ so that

$$A_{i_{t-1},j_{new}} \neq \varnothing \text{ and } \sum_{m=1}^{M} 1[A_{j_{new},m}] = 0$$

then $A_{i_{t-1},j_{new}} := u(t)$ and $i_t := j_{new}$

and goto (iii)

otherwise stop

3．スポッティング検索

IPMネットワークで表現された音声データから中から，連続オートマトン手法を用いて特定のデータをスポッティングする。このスポッティング機能により，検索queryである音声は切れ目のないものとして良いこととなる。連続オートマトンでは，ネットワーク中のすべてのノードに累積コストを定義し，1フレーム毎の入力ごとに，アークのラベルと入力ラベルの距離によって決まるコストの累積値として再計算される。図5に累積コストを算出する際のDPパス[9]とIPMネットワークの対応関係を示す。

累積コストが一定値以下となるノードが，活性化される。活性化されたノードのうちから，最も低い累積コストを持つノードをとりだし，遷移してきたアークの履歴を参照し，遷移アークに対応したデータベース中の部分をスポッティング結果とする。データベース中の部分を表現するために，IPM ネットワークにデータを登録する際に，同一のカテゴリーに属するデータに単語番号を付与して管理する。これは，便宜上単語と呼んでいるが，さらに大きな文等の単位に対して同一の単語番号で管理することができる。

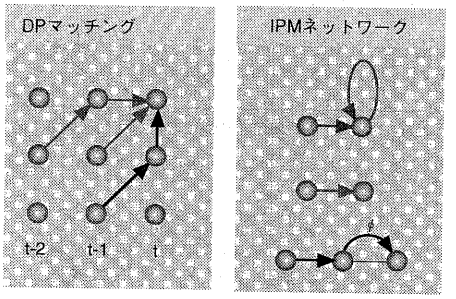

図5　DPパスとIPMネットワークの対応

【連続オートマトンによる検索アルゴリズム】

$S_j(t)$:時刻tにおける，ノードjの累積コスト

D_{ij}:{0,1} ノードiからノードjへのアークの有無

A_{ij}:ノードiからノードjへのアークに付与されたラベル

$u(t)$:時刻tにおける入力ラベル

h_3:ノード活性判別のためのスレシホールド。

$$0 \le h_3 \le 1$$

$$d_{ij}(t) = \frac{|u(t) - A_{ij}|}{|u(t)| + |A_{ij}|}$$

α：忘却係数$(0<\alpha<1)$

$MinPathNode_i(t)$：時刻tに，ノードiへ至る最適パス上の時刻t-1にお

けるノード

$MinPathArcNode_i(t)$:時刻tに，ノードiへ至る最適パス上の，時刻tにおいて遷移にしようされたアークの遷移元ノード。

$WordTags(i,j)$:ノードiからノードjに向かうアークを通過する単語の集合

W:単語セット

Step 0 初期化

$t := 0$

$S_j(t) := 1.0$ for all $j \in N$

Step 1　ナル遷移を含まないパスの計算

for all j do

if $S_j(t) < \min_i\{(1-\alpha)S_i(t-1) + \alpha d_{ij}(t)\}$ then

$S_j(t) := \min_i\{(1-\alpha)S_i(t-1) + \alpha d_{ij}(t)\}$

$i_{\min} := \operatorname{argmin}_i\{(1-\alpha)S_i(t-1) + \alpha d_{ij}(t)\}$

$MinPathNode_j(t) := i_{\min}$

$MinPathArcNode_j(t) := i_{\min}$

endif

Step 2　ナル遷移を含むパスの計算

for all j do

if $S_j(t) < \min_k\{(1-\alpha)S_k(t-1) + \alpha d_{ij}(t)\}$ and

$D_{ki} = 1$

then

$k_{\min} := \operatorname{argmin}_k\{(1-\alpha)S_k(t-1) + \alpha d_{ij}(t)\} | D_{ki} = 1$

$i_{\min} := \operatorname{argmin}_i\{(1-\alpha)S_{k_{\min}}(t-1) + \alpha d_{ij}(t)\}$

$MinPathNode_j(t) := k_{\min}$

$MinPathArcNode_j(t) := i_{\min}$

endif

Step 3 活性化ノードの履歴保存

if $S_j(t) < h_3$ then

$Active_j(t) := 1$

else

$Active_j(t) := 0$

endif

Step 4 最適パス上の登録単位の識別

$j_{\min} := \operatorname{argmin}_j S_j(t)$

if $S_{j_{\min}}(t) < h_3$ then

$Hist_w := 0$ for all $w \in W$

$p := 0$

$i_0 := j_{\min}$

while $p < P$

$i_{p+1} := HistNode_{i_p}(t-p)$

if $Active_{i_p}(t-p) = 1$ then

$w \in WordTags(HistArcNode_{i_p}(t-p), i_p)$

endif

$p := p + 1$

end while

$w_{spotted} := \operatorname{argmax}_w Hist_w$

endif

goto step 1

ネットワークを生成する際に，データベースのデータを入力したときの遷移アークに，入力データが属する単語ラベルを通過履歴として各アークに記録しておく。単語のスポッティングの際には，連続オートマトンで活性化されたノードの中から最も累積コストの小さいノードを取り出し，そのノードに至る最適パス上のアークの単語の通過履歴を全て加算し，最も通過回数の高い単語をスポッティング結果としている。

4．評価実験

4.1 1時間のTV番組の音声の自己組織化

実験試料として，1時間のＴＶニュース番組の音

声を無音区間で分割し，無音区間に挟まれた一続きの音声を1つの文単位として IPM ネットワークを構築する実験を行った。I P M ネットワークを構築には，フレーム長16[ms]フレームシフト8[ms]で分析し，フレーム毎に音素識別を行った結果得られる音素系列を用いた。無音部分をカットした結果，397,045フレーム，486文，53分間のデータになった。

4.2 実験結果

図6 に，I P M ネットワーク表現した際のノード数とアーク数を示す。グラフの横軸は，1つのノードが持つことのできるアークの制限値を示す。アークの制限数を1にした場合は，ネットワークは一本の線のような形状となり，ノードの数はデータベースに含まれるデータの数と等しくなる。逆にアークの制限本数を大きくすると，ネットワークの構造はバイグラムに近づき，ノードの数は減少し，最終的には量子化されたラベルの種類とノード数が等しくなる。アークの制限数は量子化されたラベルの種類までふやすことができるが，通常その手前で飽和する。

計算時間は，連続オートマトンのアルゴリズムではノード数とアーク数に比例するため，アークの制限数の増加と共に計算量が減少する。

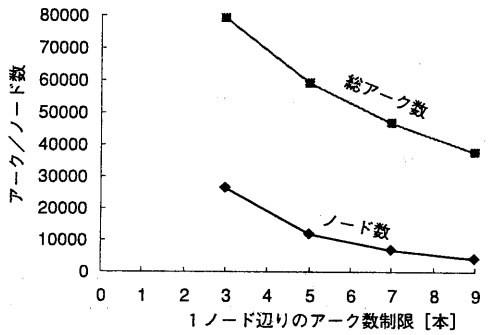

図6 IPMネットワークのノード数とアーク数

5. まとめ

本報告では，時系列情報からネットワークを自己組織化するIncremental Path Methodを用いて，音声データを自己組織化し，不特定話者音声による検索システムを提案した。1時間のT V ニュース番組の音声データを自己組織化ネットワークで表現すると，397,045フレームのデータが，26,335ノード，78,985アークで表現できた。今後は実際に検索性能と計算量で自己組織化されたデータの評価を行いたい。

謝辞 本研究に日頃支援いただく，新情報処理開発機構島田潤一所長に感謝いたします。また，熱心に議論いただく情報ベース機能研究室・マルチモーダル機能研究室の皆様に深く感謝いたします。

参考文献

[1] 豊浦，岡："テキストの知識ベース化のための自己組織化ネットワークの提案"，信学技報，NLC96-59(1997-3).

[2] 岡，伊藤，木山，張："概念スポッティングのための画像オートマトン"，音講論集，3-4-12(1995-3).

[3] Endo, Zhang, Nakazawa, Oka:"Mutual Spotting Retrieval between Speech and Video Image using Self-organized Network Databases", Proc. of the INternational Conference on Advanced Multimedia Content Processing, pp.106-121(1998-11).

[4] Alta Vista:http://www.altavista.digital.com/

[5] goo:http://www.goo.ne.jp/

[6] 岡："スペクトルベクトル場とスペクトルの音声認識における有効性比較について"，信学論誌，D, J69-D, No. 11, pp. 1704(1986,11).

[7] 松村，岡，木暮，小島："スペクトルベクトル場の方向性パターンを用いた不特定話者の単語音声認識"，信学論(D-II), Vol.J72-D-II, No.4(1989-4).

[8] 古川，中沢，岡："Reference Interval-free 連続DPを用いた音声検索方式における特徴量削減の方法"，音講論集，3-8-1(1996-3).

[9] Oka:"Spotting Method Approach Towards Information Integration", Proc. of 1997 Real World Computing Symposium, pp. 175-182(1997-1).

[10] 遠藤，中沢，長屋，高橋，岡："Incremental Path Mehtodによる音声・動画像の自己組織化データベース構築"，人工知能学会 情報統合研究会資料，SIG-CII-9710, pp. 28-32(1997,10).

[11]阿部，匂坂，梅田，桑原："研究用日本語音声データベース利用解説書"，ATR，ATR Technical Report, TR-I-0166(1990-9).