北海道大学工学部研究報告
第119号（昭和59年）

Bulletin of the Faculty of Engineering,
Hokkaido University, No.119 (1984)

話者照合における時間整合に関する検討

小田　明*　田中　浩　前田　斉**
斎川　勝男　栃内　香次　永田　邦一

（昭和58年9月30日）

Considerations in Dynamic Time Warping Algorithm for Speaker Verification

Akira ODA, Hiroshi TANAKA, Hitoshi MAEDA
Katsuo SAIKAWA, Koji TOCHINAI and Kuniichi NAGATA

(Received September 30, 1983)

Abstract

The technique of dynamic time warping for time registration of reference and test utterances has found widespread use in the area of word recognition and speaker verification.

Speaker verification experiments have been made using Japanese spoken single digits, and we have considered some modifications to the DP path algorithm.

Spectrum time patterns at transition intervals of an utterance show a relative resemblance among utterances of the same word for each speaker and lengths of vowel intervals are different according to its speaking speed.

Taking these features into account one of the modifications is considered to constrain the DP path to fix its slope unity at transition intervals of a reference utterance.

Another modification is about selections of starting and end points of DP matching path to make better DP matching and to reduce calculation time.

The performance of speaker verification experiments using these modified DP path algorithms, was improved.

1．まえがき

話者照合は通常話者の音声を登録しておき，入力されたテスト音声を登録した参照音声と比較し，それら音声間の類似度を調べ，類似度が大きければ登録話者の音声であると判定し，類似度が低ければ詐称者音声と判定するもので，その結果を以後の対応に反映することになる。

話者音声の個人的特徴を抽出し，それによる判定が行われるのが望ましいが，音声の個人的特徴が何であるかはまだ明らかにされていない。ただ個人的特徴がスペクトル情報に最も多く含まれているということが報告[1]されているが，また同一話者音声スペクトルも発声時期により，大きく変動することも報告されており[2][3]，人間の聴覚系により認識される個人的特徴が何であるかは

電子工学科　電子機器工学講座
* 現在日立製作所勤務
** 現在電電公社勤務

明らかでない。しかしそのような特徴も機械照合に適した特徴であるという保障はない。

現在報告されている話者照合の研究には，単語音声のスペクトル・ピッチ等の時間パターンについて参照音声とテスト音声を比較し，類以度をもとめているものが多い。その場合に，スペクトル・ピッチ等の時間パターンの音声の各時刻についての差異を求め，その差異を音声全区間にわたり加算する方法,発声全区間にわたる平均的なスペクトル,ピッチ周波数等を求め，その差異を求める方法，ならびに両者を併用する方法が考えられる[4)~9)]。そのいずれの方法が優れているかは照合条件とも関係があるが，話者の個人的特徴が十分解明されていない現時点で論ずることは尚早であると思われる。

われわれは話者照合の研究をここ数年行ってきたが，もっぱら第一の方法による照合実験を重ねて来た。これは一般に平均化により失われる情報があるという考えにもとづいている。

二つの単語音声の各時刻におけるスペクトルの類似度を調べるためには，両単語音声の対応する時刻を決定する必要があり，これには通常両音声間の類似度を最大にする時刻対応を求めるDPマッチングの手法が用いられている。

DPマッチングの手法は単語音声識別,話者認識に広く用いられ,各種のアルゴリズムが提案されているが[10)~15)]，一般にDP面上におけるDPパスを決定するに際し，DP時間の節約と照合精度を向上させるためDPパスに各種の制限を加えるのが普通である。

同一話者が同一文章を複数回発声しても，文中の音韻の過渡区間では，スペクトルの動的変化は，比較的一定しているということが報告されている[16)]。本論文はこの様な特徴をDPパス上の制限として持たせたアルゴリズムを単語音声を用いた話者照合実験に利用した場合について検討を行ったものである。

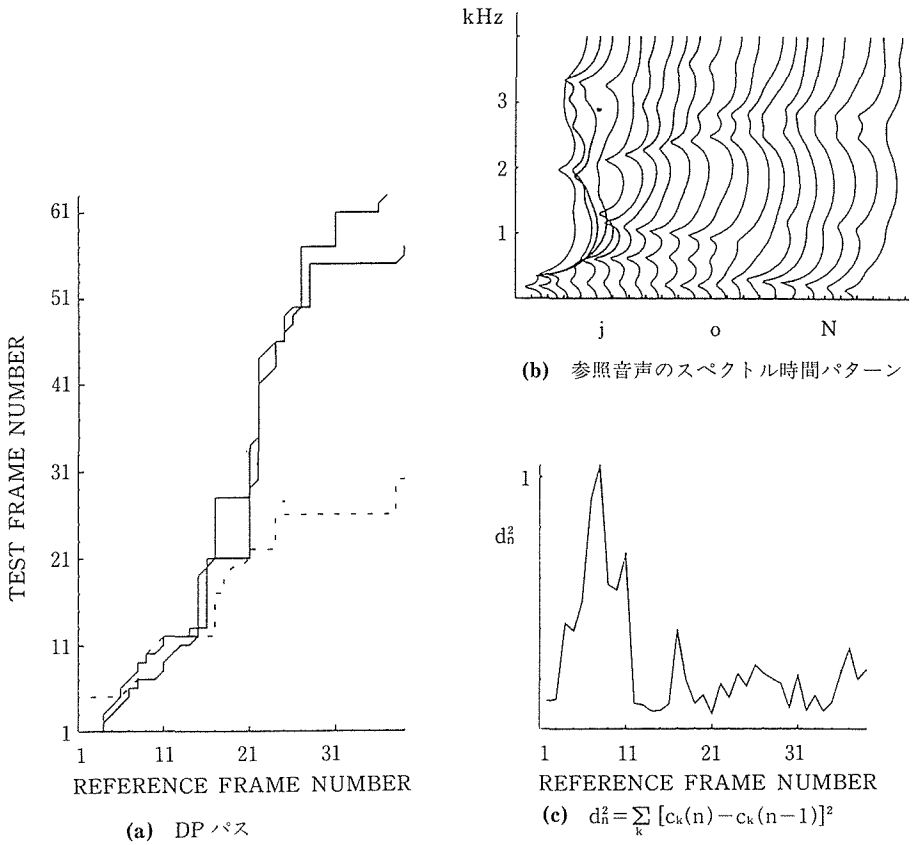
また単語音声の始端，終端の切り出しが不正確であったり，また音声始端，音声終端における音声波形が不安定である場合もしばしば見うけられる。これらは参照・テスト両音声間の類似度に悪影響を及ぼす原因になると考えられる。したがってこのような悪影響を避けるため，本話者照合実験IIでは音声の始端・終端を除外して類似度が求められており，また計算時間の節約も計られている。

2．同一単語音声間のDPパス

DPマッチングによる時間整合を用いて話者照合を行った場合のDPパスの一例を図1(a)に示す。横軸は，話者Aの参照音声“4”/joN/のフレーム番号，縦軸はテスト音声サンプル（話者AおよびBの発声した音声“4”）のフレーム番号を示す。各フレームのスペクトルは4章に述べられる手法で線形予測分析され，14次のケプストラム係数（C_1 ~C_{14}）で表されている。話者Aのテスト音声サンプルは2個で，いずれも参照音声に比しゆっくり発声されており，自己の参照音声とのDPパスは図(a)の実線で示されている。また話者Bのテスト音声と話者Aの参照音声とのDPパスは点線で示されている。図より実線で示されるDPパスは横軸フレーム番号11まではほぼ45°の傾斜に沿って進んでいるが，点線で示されるDPパスは45°の傾斜直線からかなりずれていることがわかる。

図1(b)は参照音声のスペクトル・時間パターンを示す。また図(c)は参照音声スペクトルのフレーム間変動の度合を示すため，各フレームとその直前のフレームとのケプストラム距離を求めたものである。

参照音声はフレーム番号11までの間で，スペクトルの変動が大きく，発声時に声道形状が急速に変化する音韻のわたりの部分に対応するものと考えられる。この様な部分で同一話者のテスト



図－1　DP パスとスペクトル時間パターンの関係

音声の DP パスが 45° の傾斜で進んでいることは，従来指摘されて来たように，発声速度が変っても音韻のわたりの部分では音声器官の形状の時間変化はそれぞれの発声者について安定していることを示しているといえる。

他の話者，他の数字音声についても同様な傾向がえられている。以後スペクトルの変動が比較的大きい音声区間を過渡区間と呼ぶことにする。

3．時間整合 DP アルゴリズム[17]

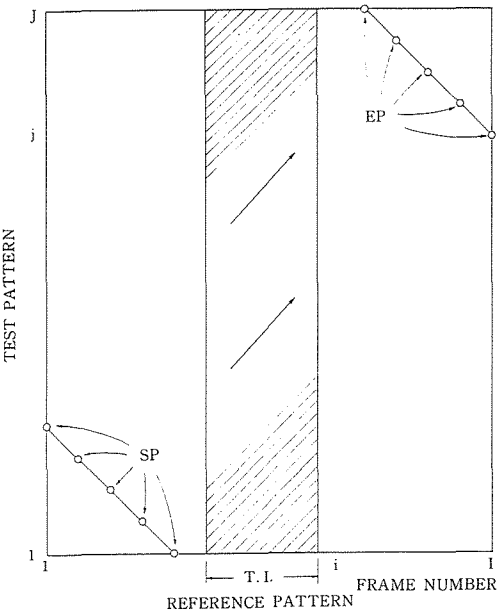
前章で述べられたように，音声の過度区間では音声器官の形状変化が，常にほぼ一定の時間変化をするならば，参照音声のスペクトル変化の大きな過渡区間に対応する DP パスを常に 45° の傾斜に沿って進むように制限を加えても，テスト音声と参照音声が同一話者の場合には最大類似度を与えるパスからのずれは少なく，したがって受ける拘束も小さいと考えられる。

この様な制限を設けることによりパターン間の類似度は低下（パターン間距離は増加）するが，テスト音声が参照音声と同一話者の場合には拘束は小さいため類似度の低下は少なく，他方テスト音声が参照音声と異話者の場合にはパスはより大きい拘束を受け，パターン間類似度の低下も大きく，従って話者照合誤りは減少するものと期待される。

図 2 は本アルゴリズムを用いた DP 面の一例で，図の参照音声は音声区間の中央に 1 ケの過渡区間（T. I.）を有する場合を示している。したがって斜線で示される部分にあるノードから出る

DP パスは 45° の傾斜に固定されることになる。

次章に述べられる話者照合実験に用いられた音声サンプルの中には発声の始端と終端で，声帯励振が欠如したり，スペクトルも不安定なサンプルが見られた。また単語音声の始端・終端の決定は単に信号振幅により行われたためその正確な抽出を期待することは困難である。従って図 2 左下ならびに右上に示されるような 45° の斜線上に並んだ数個の SP 点, EP 点を DP パスの始端, 終端とし, その間で DP パス制限を満足するパスについて最小のパターン間距離を求め両パターン間距離とした。この様にすることにより，音声始端・終端近傍の不安定な音声区間の影響を除き，音声全区間の主要部についての類似度が求められることになる。



図－2　本アルゴリズムによる DP 面の一例

表 1 に話者 7 名の数字音声 0 /rei/（総サンプル数 216），7 /nana/（総サンプル数 217）について SP の数（N_S）を 1 内至 7 に変化して話者照合実験を行った結果を示す。この実験では N_S が，3 内至 4 程度の時最も誤りが少なくなることが示されている。なお本実験には後述の図 3 (a)で示される DP パスアルゴリズムが用いられ，EP の数（N_E）は 3 に選ばれている。N_E を 4 にした場合もほぼ同様な結果が得られている。

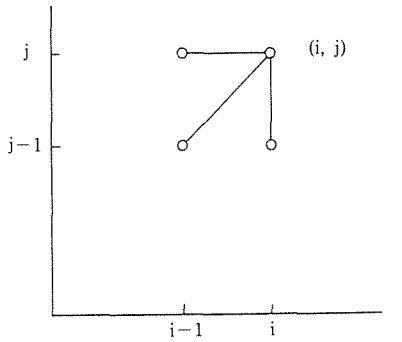
図 2 で DP パスがいずれの SP から出発し，いずれの EP に終ろうともパターン間の集積距離を求めるのに参加する縦軸，横軸の合計フレーム数は $I+J-N_S-N_E+2$ となるため，1 回の DP パス整合演算を行うだけで最適パスを求めることができる。

図 3 は過渡区間以外の音声区間に対応する DP 面上でとりうるパスのアルゴリズムを示したものである。図 3 (a), (b)中の g(i, j) は参照パターンの i 番目フレーム，テストパターンの j 番目フレームを対応させるノード (i, j) における集積距離を示し，d(i, j) は参照パターンの i フレームとテストパターンの j フレーム間の距離を示している。g(i, j) は図に示される 3 種類の集積距離の内最小のものが選ばれ，それに対応するパスが決定される。

同一発声者の数字音声を約 2 年にわたって収録した音声サンプルについて，サンプル長の変動を調査したところ，発声者は普通の早さで発声したにも拘らず，音声長の比は 2.6 に及ぶ場合のあることがわかった。したがって DP パスアルゴリズムとしては上記音声長比に追従できるようにする必要がある。図 3 (a)は傾斜制限の無いパスアルゴリズムで，図 3 (b)は図 2 の始端群 SP，終端群 EP の設定に適合し，かつ上記条件を満足するように考慮されたパスアルゴリズムである。

表 1　DP開始点数（Ns）を変化した場合の話者照合誤り率　NE；3 の場合

音声数字 ＼ Ns	1	2	3	4	5	6	7
0	5.9	5.5	5.1	5.1	5.5	5.8	6.2
7	6.0	6.1	6.1	6.1	6.2	6.3	6.4
平　均	6.0	5.8	5.6	5.6	5.9	6.1	6.3

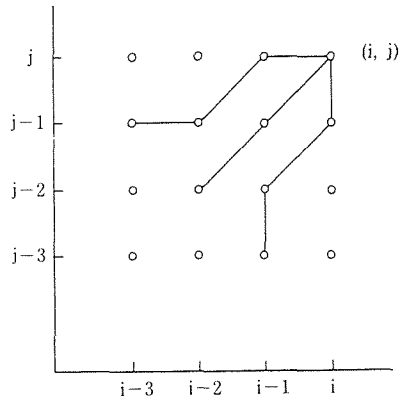


$$g(i,j)=\min\begin{cases} g(i,j-1)+d(i,j) \\ g(i-1,j)+d(i,j) \\ g(i-1,j-1)+2d(i,j) \end{cases}$$

INITIAL CONDITION

$2d(i,j)$

(a)



$$g(i,j)=\min\begin{cases} g(i-1,j-3)+d(i-1,j-2)+2d(i,j-1)+d(i,j) \\ g(i-2,j-2)+2d(i-1,j-1)+2d((i,j) \\ g(i-3,j-1)+d(i-2,j-1)+2d((i-1,j)+d(i,j) \end{cases}$$

INITIAL CONDITION

$2d(i,j)$

(b)

図－3 許容 DP パスと評価関数

4．話者照合実験 I

前章に述べた過度区間拘束を設けることにより，どの程度の話者照合誤りの改善が計られるかを実験により確かめた。

本章の話者照合実験 I に用いられた数字音声は“4”/joN/と“7”/nana/で，いずれも 7 名の男性話者が区切って発声したものである。それぞれの数字音声は最初に各話者について 7 サンプルずつ採取し，その後 1 週間後，1 ケ月後，4 ケ月後，6 ケ月後，9 ケ月後，1 年後の各時期に各話者 4 サンプルずつ，各話者計 31 サンプル，総計 217 サンプルが用いられた。

音声サンプルは3.4kHzの低域フィルタ通過後，8 kHz，12ビットでA－D変換された。音声始端・終端の検出は信号振幅によってなされた。差分処理による高域強調後 10 mS 毎に 30 mS のハニング窓を乗じてフレーム列を構成し，各フレームに対し自己相関法により 14 次の LPC 分析を行い，14 次の LPC ケプストラム係数を求めた。

各話者の参照音声は初時期発声サンプル中より無作為に抽出され，残りの本人音声サンプル，詐称者音声 186（6 ×31）が各話者の照合実験に用いられた。

参照音声 i フレームとテスト音声 j フレームのフレーム間距離としては次式で示される LPC

ケプストラム距離が用いられた。

$$d(i,j)=\sum_{k=1}^{14}[C_k^{(R)}(i)-C_k^{(T)}(j)]^2$$

ここで$C^{(R)}_k$ (i) は参照音声iフレームのk次のLPCケプストラム係数，$C^{(T)}_k$ (j) はテスト音声jフレームのk次のLPCケプストラム係数である。

図3(a)のDPパスアルゴリズムを用いた話者照合実験の結果を表2，表3に示す。

表2は数字音声"4"/joN/についての照合結果で，第1列は話者名，第2列は過渡区間拘束を設けた場合の誤り率A(%)，第3列は過渡区間拘束を設けない場合の誤り率B(%)を示している。

AとBの比を第4列に，過渡区間は/jo/の区間に設けられ，そのフレーム番号を最終列に示す。表より過渡区間拘束を設けることにより平均誤り率が12%改善されることが示されている。しかし話者HHでは逆に誤りの増加が見られる。

また表3は数字音声"7"/nana/についての照合結果で，表中の各列は表2と同様な項目を示しているが，過渡区間は2区間で，いずれも鼻音/n/から母音/a/へのわたりの区間に設けられている。この場合も過渡区間拘束を設けることにより平均10%の誤り率改善が得られている。しかし話者HO，MSについては誤り率の改善が見られない。

表2　話者照合実験I(数字音声/joN/の場合)の誤り率

A：過渡区間拘束を設けた場合の誤り率
B：過渡区間拘束を設けない場合の誤り率

	A(%)	B(%)	A／B	Nos. of T. I.
HH	14.1	13.3	1.06	3 － 6
HO	3.8	6.7	0.57	3 － 17
KN	3.4	3.4	1.00	3 － 11
KS	20.0	20.0	1.00	3 － 14
MS	6.5	9.7	0.67	3 － 12
NM	0.5	0.5	1.00	3 － 8
TN	0.5	1.6	0.31	3 － 17
MEAN	7.0	7.9	0.88	

表3　話者照合実験I(数字音声/nana/の場合)の誤り率

A：過渡区間拘束を設けた場合の誤り率
B：過渡区間拘束を設けない場合の誤り率

	A(%)	B(%)	A／B	Nos. of T. I.
HH	11.3	12.9	0.88	1－6，24－27
HO	3.3	3.3	1.00	1－6，24－35
KN	12.8	13.4	0.96	1－4，17－30
KS	3.3	3.8	0.87	3－5，20－25
MS	3.3	3.3	1.00	1－5，21－30
NM	16.7	17.7	0.94	1－5，26－38
TN	1.1	3.3	0.33	1－5，30－40
MEAN	7.4	8.2	0.90	

表4　話者照合実験II(数字音声/joN)の誤り率

A：過渡区間拘束を設けた場合の誤り率
B：過渡区間拘束を設けない場合の誤り率

	A(%)	B(%)	A／B	Nos. of T. I.
AR	8.7%	10.3	0.84	1 － 21
FR	0.0	0.0		1 － 21
HK	17.4	22.1	0.78	1 － 17
HS	0.0	0.0		1 － 15
IT	4.8	4.8	1.00	1 － 14
KA	13.0	13.2	0.98	1 － 18
KD	4.4	4.4	1.00	1 － 12
KO	8.8	8.7	1.01	1 － 17
MD	4.4	4.4	1.00	1 － 11
NG	4.3	4.3	1.00	1 － 17
NK	7.4	4.4	1.68	1 － 14
SA	2.9	2.9	1.00	1 － 17
SG	7.4	10.3	0.72	1 － 19
SI	8.8	8.8	1.00	1 － 20
TK	4.4	4.4	1.00	1 － 19
UM	5.9	8.7	0.56	1 － 10
YM	13.2	14.7	0.89	1 － 17
YS	5.9	8.7	0.68	1 － 17
MEAN	6.8	7.5	0.90	

5．話者照合実験II

本章の話者照合実験IIに用いられた数字音声は“4”/joN/で，18名の男性話者が2年間にわたる4時期に発声した音声サンプルが用いられた。各時期の音声サンプルは6個で総計432サンプルが利用された。

各話者の参照音声は初時期発声サンプル中より各話者サンプル平均の音声長に近い音声長を有するサンプルが選ばれた。各話者の照合実験には本人音声として残りの音声サンプル全数(23)，詐称者音声として各他話者の各時期音声サンプルより1サンプルずつ選ばれ，総計68(1×4×17)個の詐称者音声が用いられた。信号の処理・話者照合実験の方法は図3(b)のDPパスアルゴリズム(N_S；8～10，N_E；5～6）を用いた以外前章の場合とほぼ同様である。

表4に照合結果を示す。この場合も前述の照合実験Iの場合と同じく，過渡区間拘束を設けることにより誤り率が平均10%改善されている。しかし誤り率の改善の見られない話者も多く，今後検討する必要がある。

6．むすび

数字音声/joN/, /nana/について，DPマッチングによる時間整合を用いた話者照合実験を行い以下の結果を得た。

1）音韻のわたりに対応する過渡区間を設け，参照音声の過渡区間に対応するDPパスを45°に固定することにより照合誤り率を平均約10%改善することができた。

2）過渡区間拘束を設けることにより誤り率の改善されない話者も存在し，その原因について今後検討する必要がある。

3）単語音声の始端，終端近傍の情報を棄て，図2に示されるような複数のDPパス始端SP，終端EPを用いることにより，一回のDP整合で良好な照合結果が得られた。

なお前章の照合実験IIに用いられた数字音声サンプルは電電公社武蔵野通信研究所より提供されたもので，当時の斎藤収三特別研究室長(現東大教授)，古井貞熙調査役ならびに関係各位に深謝致します。

参考文献

1）伊藤，斎藤：電子通信学会論文誌，Vol. **J65-A**（昭57），No. 1，P. 101
2）古井：電子通信学会論文誌，Vol. **57-A**（昭49），No.12，P.880
3）小倉，小田，栃内，永田：日本音響学会音声研究会資料S 80－98（1981－3）
4）古井，板倉，斎藤：電子通信学会論文誌，Vol. **55-A**（昭47），No. 10，P. 549
5）古井，板倉：電子通信学会論文誌，Vol. **56-A**（昭48），No. 11，P. 717
6）R.C. Lummis：*IEEE Trans.* **AU-21**，No. 2，1973，P. 80
7）小倉，広瀬，栃内，永田：北大工学部研究報告，**第101号**（昭55），P. 61
8）古井，斎藤：研究実用化報告第**29**巻第7号（昭55），P. 71
9）古井：電子通信学会論文誌，Vol. **J65**（昭57），No. 2，P. 183
10）迫江，千葉：電子通信学会連合大会論文集，136，（昭45）
11）迫江，千葉：日本音響学会誌，Vol. **27**（昭46），No. 9，P. 483
12）H.Sakoe, S.Chiba：*IEEE Trans.* **ASSP-26**，No. 1，1978，P.43
13）H.Sakoe：*IEEE Trans.* **ASSP-27**，No. 6，1979，P. 588
14）C.S.Myers, L.R.Rabiner：*IEEE Trans.* **ASSP-29**，No. 3，1981
15）鹿野，相川：日本音響学会音声研究会資料S 82－15（1982－6）
16）伊藤，古井，斎藤：日本音響学会研究発表会講演論文集II，Oct. 1978，2－2－8，P. 367
17）小田，栃内，永田：電気四学会北海道支部連合大会講演論文集，（昭56）P. 221