九州大学大学院
システム情報科学研究科報告
第3巻 第2号 平成10年9月

Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.3, No.2, September 1998

フレーム間で連続な時間・周波数ワープによる話者正規化の検討

山田 圭* ・ 内田誠一* ・ 迫江博昭**

Speaker Normalization Based on Time-Frequency Warp with Inter-Frame Consistency

Kei YAMADA , Seiichi UCHIDA and Hiroaki SAKOE

(Received June 22, 1998)

Abstract: A new algorithm for speaker-independent spoken word recognition is presented. The algorithm is based on the time-frequency warping technique where frequency axis warping is performed in order to adjust individual spectral difference, in addition to time axis warping. In the conventional algorithm, frequency axis warping is independently determined at each frame (i.e., time). In this case, such warp have a tendency to yield excessive deformations of time-frequency plane, it is feared. In order to suppress such excessive deformations, inter-frame consistency of frequency axis warping is newly taken into account as constraints on the warping. The optimal warping is obtained by using dynamic programming with the constraints. As an implementation technique, beam search based acceleration is also investigated. Experimental results indicates advantageous characteristics of the present algorithm over the conventional algorithm.

Keywords: Spoken word recognition, Dynamic programming, Time-frequency warping, Frame-to-frame continuity, Speaker independent recognition

1. ま え が き

不特定話者を対象とした音声認識において，話者によるスペクトル変動に対応する方法として時間・周波数ワープが試みられた[3),4),5)]．これは特定話者を対象としたDPマッチングアルゴリズムの一つの拡張であり，各フレーム(各時刻の短時間スペクトル)でのスペクトル間距離計算にもDPによる伸縮を適用することで話者毎のスペクトル変動を吸収しようとするもので，比較的簡単な原理で話者変動に対処できるという長所を持つ．しかし，従来の時間・周波数ワープでは周波数軸のワープは隣接フレーム間で独立に決定されており，フレーム間で極端なスペクトル遷移を生じ，結果として誤認識を生起するという欠点があった．

本論文では，最近提案された単調連続2次元ワープ[1),2)]を特殊化し，隣接フレーム間で連続性を保持する時間・周波数ワープアルゴリズムを提案し，不特定話者単語認識に適用する．以下，**2.**でフレーム間で連続な時間・周波数ワープの定式化，およびDPによる解法，**3.**でビームサーチを用いたDPアルゴリズムによる計算量の低減，およびその副作用の改善法について述べ，**4.**で実験結果と考察を示す．

平成10年6月22日受付

* 知能システム学専攻博士後期課程
** 知能システム学専攻

2. フレーム間で連続な時間・周波数ワープ

2.1 時間・周波数ワープの一般的定式化

本論文では音声信号を時間-周波数表現したものを扱う．入力，および標準パターンをそれぞれ$\boldsymbol{A} = \{a(\tau,\phi) \mid \tau = 1,\ldots,I,\ \phi = 1,\ldots,N\}$, $\boldsymbol{B} = \{b(t,f) \mid t = 1,\ldots,J,\ f = 1,\ldots,N\}$ と表す．ここでτ, tは時間方向，ϕ, fは周波数方向のインデックスである．$a(\tau,\phi)$は時刻τ, 周波数ϕでの入力音声のパワーを意味する．この音声の時間-周波数表現において，ある時刻の周波数方向のN次元列ベクトルをフレームと呼び，$\boldsymbol{a}(\tau) = [a(\tau,1),\ldots,a(\tau,N)]$, $\boldsymbol{b}(t) = [b(t,1),\ldots,b(t,N)]$と表す．

時間・周波数ワープは，次の目的関数を最小化するワープ関数$t(\tau)$, $f_\tau(\phi)$として表現される．

$$D(\boldsymbol{A},\boldsymbol{B}) = \frac{1}{I}\min_{\{t(\tau),f_\tau(\phi)\}}\sum_{\tau=1}^{I}\sum_{\phi=1}^{N}|a(\tau,\phi) - b(t(\tau),f_\tau(\phi))| \quad (1)$$

ワープ関数に対して，音声信号の時間-周波数平面上での変動，すなわちワープを考える時，その物理的な制約から$\tau_1 > \tau_2$ならば$t(\tau_1) \geq t(\tau_2)$，$\phi_1 > \phi_2$ならば$f_\tau(\phi_1) \geq f_\tau(\phi_2)$ なる単調性条件を考えることは自然である．さらにワープ関数の勾配の上限に対する制約も，極端なワープを除外するため必要である．これらを併せて

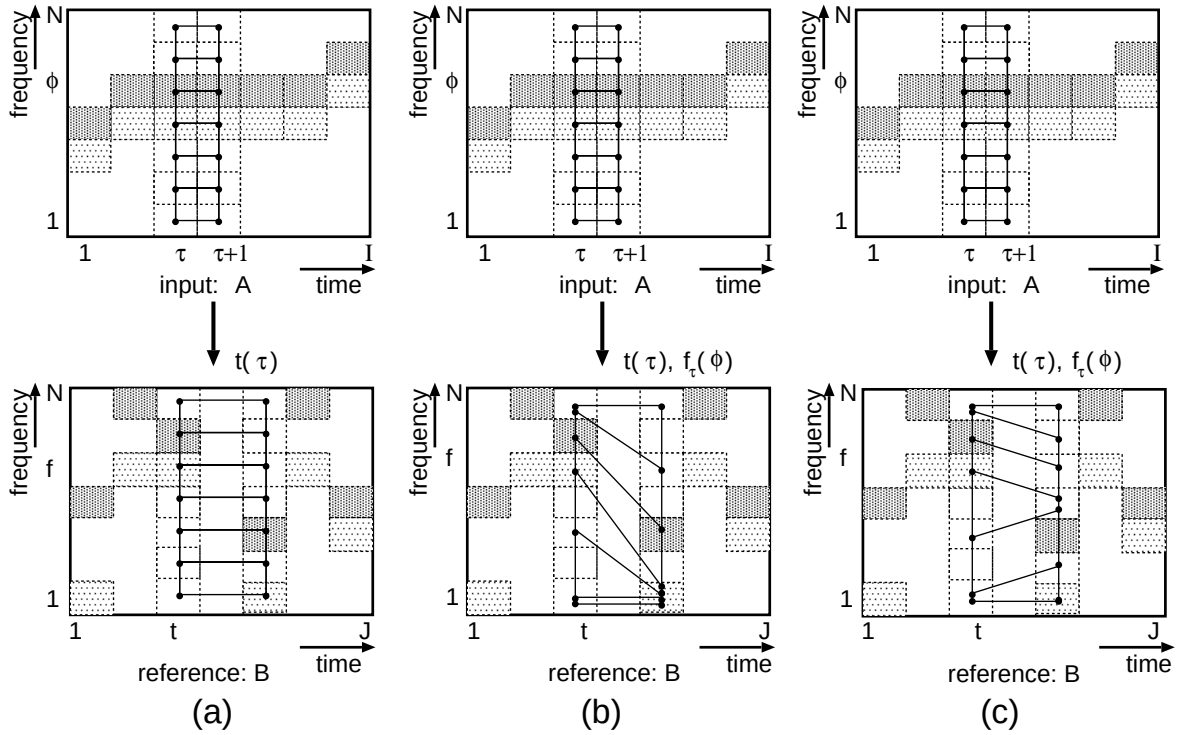

Fig.1 Warp characteristics. (a) conventional time warp, (b) conventional time-frequency warp, having no inter-frame consistency [3),4),5)], (c) the present method, having inter-frame consistency.

考え次式を得る．

$$0 \leq t(\tau) - t(\tau - 1) \leq 2 \tag{2}$$
$$0 \leq f_\tau(\phi) - f_\tau(\phi - 1) \leq 2 \tag{3}$$

(2)(3)式における定数は音声認識で一般的な時間方向の制約を参考にして決定した．必要に応じて整合窓条件(4)(5)式 を考える．

$$\left| t(\tau) - \frac{J}{I}\tau \right| \leq w_t \tag{4}$$
$$|f_\tau(\phi) - \phi| \leq w_f \tag{5}$$

ここにw_t, w_fは正の定数とする．$f_\tau(\phi) = \phi$として周波数方向のワープを固定したものが通常DPマッチングと呼ばれる時間ワープである．各フレームで周波数方向の伸縮整合を行なう処理が周波数ワープである．ワープによる入力パターンと標準パターンの対応例を**Fig. 1**に示す．

2.2 従来の時間・周波数ワープ[3),4),5)]

従来の時間・周波数ワープではワープ関数に(2) (3)式の拘束を用いていたが，ワープ関数$f_\tau(\phi)$はフレーム毎に独立に決定される為，フレーム間で極端なスペクトル遷移を許容していた(**Fig. 1**(b))．同じカテゴリに属する単語間の距離は小さくなるが，別のカテゴリに属する単語間の距離も小さくなるという欠点が生じる．すなわちパターン認識の正規化処理によく見られる過剰変形によるクラス間分離性の低下が生じる．

2.3 フレーム間連続性条件の導入

前節で述べた欠点を解決する為，フレーム間に連続性の拘束条件を導入する．

$$|f_\tau(\phi) - f_{\tau-1}(\phi)| \leq 1 \tag{6}$$

これにより，**Fig. 1**(b)のような不連続なスペクトル遷移を伴うワープを排除し，(c)のような連続的なスペクトル遷移だけを許すワープを実現できる．このようにフレーム間で連続なスペクトル遷移を行なう時間・周波数ワープを検討する．

2.4 DP アルゴリズム

フレーム間で連続な時間・周波数ワープでは，各フレームで独立に周波数軸のワープを決定できない．全体として最適なワープを実現するには，各フレームで可能な全ての周波数軸のワープ$f_\tau(\phi)$を評価する必要がある．この最適なワープを実現する問題は以下に示すようにDPを用いて解くことができる[1),2)]．

入力パターン$\boldsymbol{A}$の時刻τでのスペクトル$\boldsymbol{a}(\tau) = [a(\tau,1), \ldots, a(\tau,N)]$を標準パターン$\boldsymbol{B}$の時刻$t$でのスペクトル上にワープ関数$f_\tau(\phi)$を用いて写像する．このワープ関数$f_\tau(\phi)$ と$t(\tau)$の組を局所歪みパターン$\boldsymbol{tf}(\tau)$と呼ぶ．

$$\boldsymbol{tf}(\tau) = (t(\tau), f_\tau(1)), \ldots, (t(\tau), f_\tau(\phi)), \ldots, (t(\tau), f_\tau(N))$$

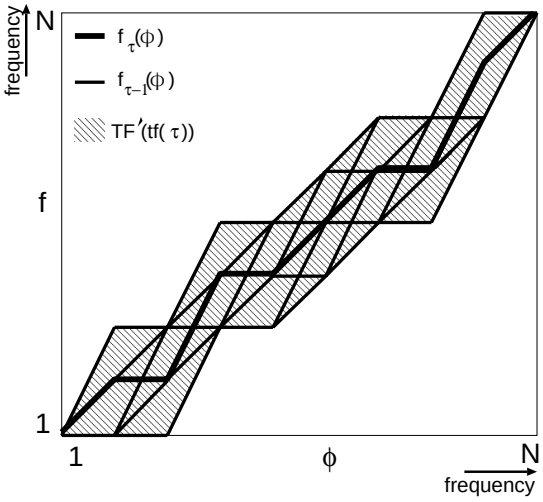

Fig.2 Example of $\boldsymbol{tf}(\tau)$ and $\boldsymbol{TF}'(\boldsymbol{tf}(\tau))$.

(1) Initialization
for all $\boldsymbol{tf} \in \boldsymbol{TF}(1)$
$g(1, \boldsymbol{tf}) = d(1, \boldsymbol{tf})$
(2) DP-recursion
for $\tau = 2$ to I
for all $\boldsymbol{tf} \in \boldsymbol{TF}(\tau)$
$g(\tau, \boldsymbol{tf}) = d(\tau, \boldsymbol{tf}) + \min_{\boldsymbol{tf}' \in \boldsymbol{TF}'(\boldsymbol{tf}(\tau))} g(\tau - 1, \boldsymbol{tf}')$
(3) Termination
$D(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{I} \min_{\boldsymbol{tf} \in \boldsymbol{TF}(I)} g(I, \boldsymbol{tf})$

Fig.3 The present DP algorithm.

局所歪みパターン$\boldsymbol{tf}(\tau)$ は単調連続条件(2)(3) (6)式 を満たす範囲で様々な形状をとりうる．その集合を$\boldsymbol{TF}(\tau)$で表す．ここで$\boldsymbol{tf}(\tau)$に対して単調連続条件を満たす$\boldsymbol{tf}(\tau-1)$ を接続可能な局所歪みパターンと呼ぶ．また$\boldsymbol{tf}(\tau)$に接続可能な$\boldsymbol{tf}(\tau-1)$の集合を$\boldsymbol{TF}'(\boldsymbol{tf}(\tau))$で表す(**Fig. 2**)． この局所歪みパターン$\boldsymbol{tf}(\tau)$の導入により式(1)の最適化問題は，最適な$\boldsymbol{tf}(\tau)$の時系列を求める問題，すなわち(7)式に帰着する．

$$
\begin{aligned}
& D(\boldsymbol{A}, \boldsymbol{B}) \\
&= \frac{1}{I} \min_{\substack{\boldsymbol{tf}(1), \ldots, \boldsymbol{tf}(\tau), \ldots, \boldsymbol{tf}(I) \\ \boldsymbol{tf}(\tau-1) \in \boldsymbol{TF}'(\boldsymbol{tf}(\tau))}} \sum_{\tau=1}^{I} \sum_{\phi=1}^{N} |a(\tau, \phi) - b(t(\tau), f_\tau(\phi))| \\
&= \frac{1}{I} \min_{\substack{\boldsymbol{tf}(1), \ldots, \boldsymbol{tf}(\tau), \ldots, \boldsymbol{tf}(I) \\ \boldsymbol{tf}(\tau-1) \in \boldsymbol{TF}'(\boldsymbol{tf}(\tau))}} \sum_{\tau=1}^{I} d(\tau, \boldsymbol{tf}(\tau)) \qquad (7)
\end{aligned}
$$

ここで

$$
d(\tau, \boldsymbol{tf}(\tau)) = \sum_{\phi=1}^{N} |a(\tau, \phi) - b(t(\tau), f_\tau(\phi))| \qquad (8)
$$

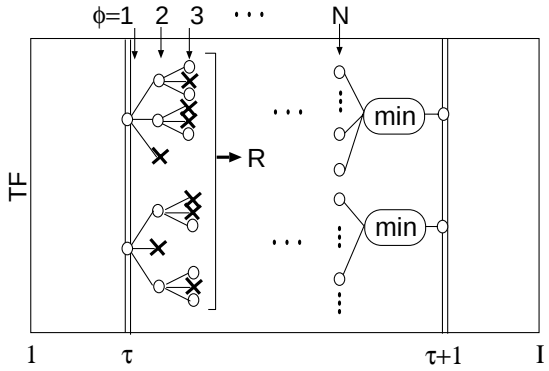

Fig.4 Pruning technique.

Fig. 3にこの最小化問題のDPによる解法を示す．ここで$g(\tau, \boldsymbol{tf})$は初期状態$\boldsymbol{tf} \in \boldsymbol{TF}(1)$ から$\boldsymbol{tf} \in \boldsymbol{TF}(\tau)$までの最適経路の累積コストである．時刻$I$で求まる累積コストの最小値が$D(\boldsymbol{A}, \boldsymbol{B})$となる．必要に応じてバックトラックを行なうことにより最適なワープが求まる．

以上の最適解アルゴリズムにおける計算量は，$|\boldsymbol{TF}(\tau)| = O(J \cdot 3^N)$, $|\boldsymbol{TF}'(\boldsymbol{tf}(\tau))| = O(3^N)$より$O(I \cdot J \cdot 3^{2N})$ となる．

3. 計算量の低減[1)]

Fig. 3の最適解アルゴリズムは指数オーダーの計算量を要する．認識に必要なスペクトルの次元，およびサンプリング周期を考慮すると，現在の計算機の能力では最適解アルゴリズムをそのまま使用することは非現実的である．以下，DPの過程にビームサーチ[1),2),6)] を導入し，多項式時間で準最適解を導出するアルゴリズムについて検討する．

ビームサーチとは，最適経路としての可能性の低いものは以後の探索から除外するという枝刈りに基づく方法である．しかし単純に各τで$\boldsymbol{tf}(\tau)$を一定個数残す枝刈りを行なっても計算量は依然指数オーダーとなる．これは$|\boldsymbol{TF}'(\boldsymbol{tf}(\tau))|$が指数オーダーとなっていることによる．そこで局所歪みパターンを各ϕ単位に分解し，(τ, ϕ)毎に累積コストの小さなものからR個だけを残す枝刈り処理を行なう(**Fig. 4**)．

この結果，総計算量は$O(RI(R+N))$となり，多項式時間で準最適解を導出することが可能となる．Rをビーム径と呼ぶ．

4. 解の安定化

4.1 ビーム制御

前節で述べたビームサーチアルゴリズムでは，圧縮された探索空間の中から最適解を探索できるとは限らない．この枝刈りの副作用として生じる時間方向の極端なワープを防ぐために，整合窓の概念を拡張し，ビームを制御

することを考える．各(τ,ϕ)での枝刈り時に，ビームの数が$\exp(-(t-J\tau/I)^2)$に比例するように，各t毎に独立に枝刈り処理を行なう．但し，ビームの総数はRとなるように正規化する．このようにビームを制御することで$t(\tau)\simeq J\tau/I$となり，時間方向の極端なワープを除去できる．周波数方向の極端なワープは，このビーム制御の枠組では除去できないので，次節で述べるペナルティの導入により対処する．

4.2 ペナルティ

周波数方向へのペナルティの導入を検討する．同じカテゴリに属する音声パターン間の周波数方向のずれは少ないと仮定し，この仮定との不一致度をペナルティとして評価することにより最適解からのずれを押える．本論文では**Fig. 3**の局所距離$d(\tau,\boldsymbol{tf}(\tau))$の計算を(9)式に変更することでペナルティの導入を実現する．

$$
\begin{aligned}
&d(\tau,\boldsymbol{tf}(\tau))\\
&=\sum_{\phi=1}^{N}\Big(|a(\tau,\phi)-b(t(\tau),f_\tau(\phi))|+\beta_f\cdot|f_\tau(\phi)-\phi|\\
&\quad+\alpha_f\cdot|f_\tau(\phi)-f_\tau(\phi-1)-1|\Big) \qquad (9)
\end{aligned}
$$

5. 実　　験

5.1 音声試料

東北大・松下単語音声データベース[7] (Vol.1) のうち50単語を用いた．成人男女各12人が発声した各単語をフレーム周期16ms，19次元メルスペクトラム$(N=19)$で分析した．男女各6人分を標準パターン用グループとして，他の男女各6人分をテスト用グループとした．標準パターン用グループの中で音声区間の継続時間が平均的なものを標準パターンとして用いた．なお男女間の変動に対処するために，男性女性のそれぞれに標準パターンを用意した．

5.2 実験条件

テスト用グループの全600データを用いて従来の周波数ワープと本手法の比較実験を行なった．本手法は**3.**で示したビームサーチアルゴリズムを用いている．ここでビーム径は計算時間の制約から$R=1000$とした．さらに**4.**で検討した，ビーム制御，およびペナルティを導入した実験を行なった．ビーム制御は，$t=J\tau/I$のときのビーム径は約200, $t=J\tau/I\pm w_t$のときのビーム径は約40 となるようにした．ペナルティ，整合窓条件に関する定数は予備実験より$\alpha_f=\beta_f=0.01$，$w_t=4$, $w_f=2$とした．

5.3 実験結果および考察

実験結果を**Table–1**に示す．従来の時間・周波数ワープと比較して本手法ではビーム制御，ペナルティを共に用いた場合1.5%の認識率の改善がみられた．ペナルティとビーム制御の効果を比較すると，特にビーム制御により認識率は大幅に向上していることが分かる．このことは平均単語間距離($\sum D(\boldsymbol{A},\boldsymbol{B})$/データ数) を比較することから説明できるように，ビーム制御により時間方向で枝刈りの副作用が押えられ，より良い準最適解が得られているといえる．次にペナルティの効果を検討する．従来法ではペナルティを用いなかった場合と比較してペナルティを用いた場合0.3%の認識率の向上がみられた．これに対し，本手法でペナルティを用いると，ビーム制御を用いないとき3.2%，ビーム制御を用いると0.8%の認識率の向上がみられ，従来法より本手法の認識率の改善効果が高い．ワープによる過剰変形を防ぐ効果と共に，周波数方向で枝刈りの副作用を押えた効果であると考えられる．

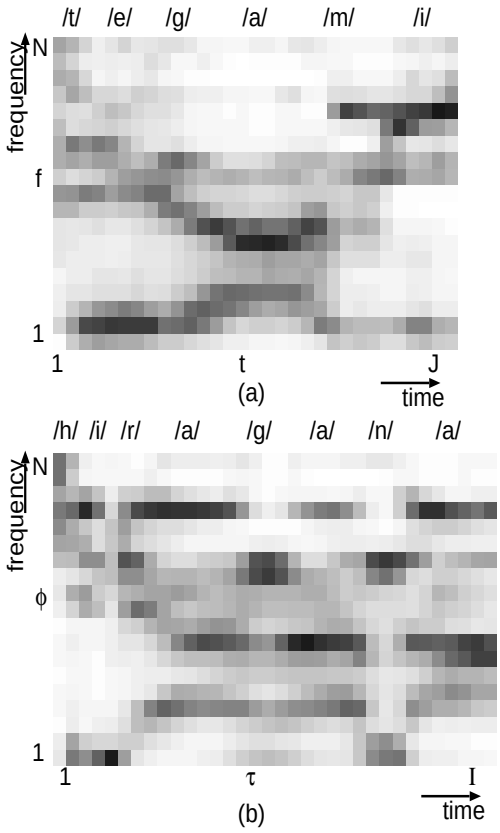

Fig.5 Example of (a) reference pattern(/tegami/) and (b) input pattern(/hiragana/).

Fig. 5に従来法で誤認識となった単語が本手法(ビーム制御，ペナルティあり)を用いることで正しく認識されるようになった例を示す．**Fig. 5**(a) に入力パターン(/hiragana/)，(b)に入力パターンと別のカテゴリの標準パターン(/tegami/)を示す．標準パターン(/tegami/)を入力パターン(/hiragana/)に合わせてワープを行なった

Table 1 Recognition result($w_t = 4, w_f = 2$).

		Without penalty ($\alpha_f = \beta_f = 0$)		With penalty ($\alpha_f = \beta_f = 0.01$)	
		Error rate(%)	Average distance	Error rate(%)	Average distance
Conventional algorithm		6.5	–	6.2	–
Present algorithm	Without beam control	22.2	1.53	19.0	1.55
	With beam control	5.5	1.38	4.7	1.40

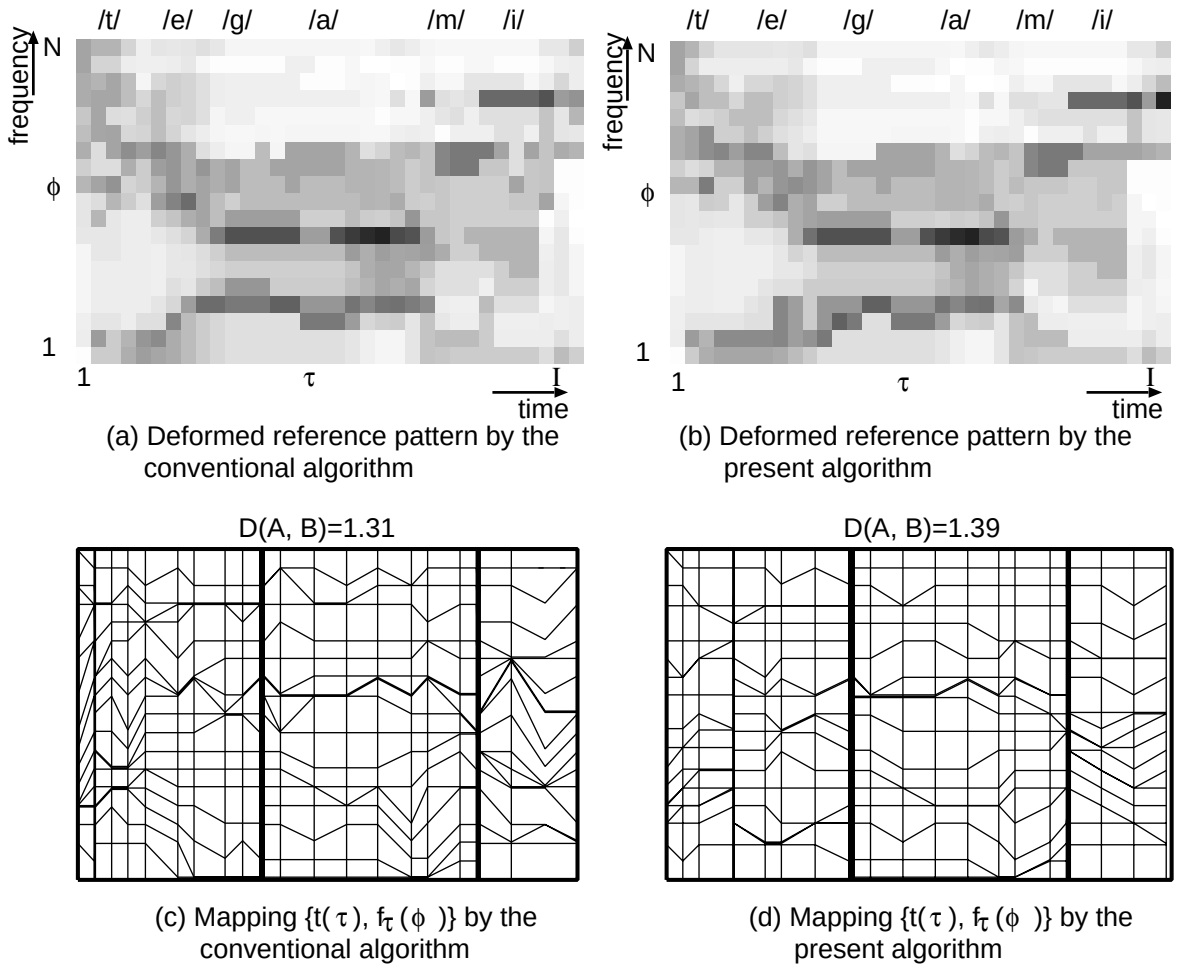

Fig.6 Deformed reference pattern and mapping (/tegami/ ⇒/hiragana/).

結果を**Fig. 6**(a)(c)に示し，その時のワープによる対応関係を**Fig. 6**(b)(d)に示す．従来法では入力パターンの調音結合部/ra/と，連続的に遷移すべき母音部/e/と極端な対応付けが行われていることがわかる(**Fig. 6**(c))．一方，本手法では連続性の拘束により極端な対応付けが排除されている(**Fig. 6**(d))．この対応付けを行なうワープの局所距離の関係を**Fig. 7**に示す．従来法と比較して本手法では不連続なスペクトル遷移を排除した時に，連続性の拘束により局所距離が大きくなる．これはカテゴリ間の分離能力の高さを示すものである考えられる．次に**Fig. 5**(a)の入力パターン(/hiragana/)と同じカテゴリの標準パターン(/hiragana/)とのワープを行ない，このワープの局所距離の関係を**Fig. 8**に示す．従来法に対して，本手法の局所距離が大きくなるフレーム数は，**Fig. 7**と比較すると少ない．同じカテゴリに属する単語間では，連続性の拘束を用いても従来法と同様の正規化能力が得られるためだと考えられる．

Table–2に従来法，本手法(ビーム制御，ペナルティあり) のどちらか一方で誤認識となった例を示す．従来法で誤認識となった単語について本手法では9単語が正しく認

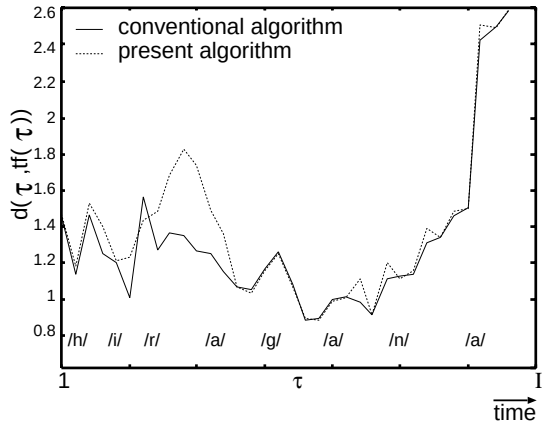

Fig.7 Local distance between /hiragana/ and /tegami/.

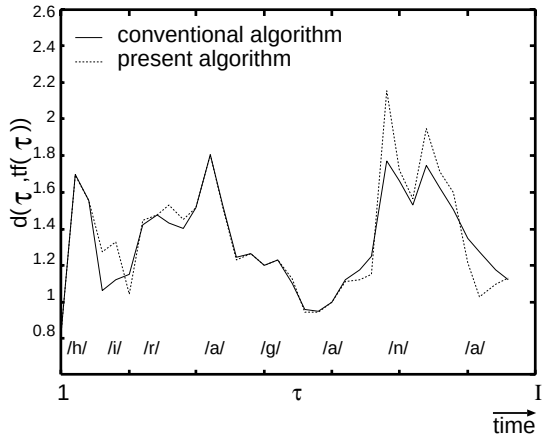

Fig.8 Comparison of local distance between /hiragana/ and /hiragana/.

Table 2 Misrecognized words.

Input (speaker ID)		Misrecognition	
		Conventional algorithm	Present algorithm
/bizyucu/	(208)	/nukiuci/	
/hiragana/	(208)	/tegami/	
/guusuu/	(226)	/diizeru/	
/hiragana/	(504)	/curara/	
/tosiue/	(504)	/buqkyoo/	
/ginkoo/	(606)	/hendoo/	
/hiragana/	(606)	/curara/	
/yomici/	(609)	/dageki/	
/boosoo/	(609)	/oosama/	
/boosoo/	(301)		/roodoo/

識され，新たに1単語が誤認識となった．連続性の拘束を用いることで，カテゴリ間の分離性は高くなるが，カテゴリ内の正規化能力は保たれるためであると考えられる．これはフレーム間の不連続なスペクトル遷移を排除する，連続性の拘束を用いた本手法の有効性を示すものである．

6. むすび

不特定話者単語認識を目的としてフレーム間のスペクトル遷移に連続性の拘束を持たせるフレーム間で連続な時間・周波数ワープアルゴリズムについて検討した．

成人男女各6人が発声した50単語について認識実験を行なった．結果として，従来法に対し本手法では1.5%の認識率の改善がみられた．これは極端に不連続なスペクトル遷移をともなう対応づけを排除した本手法の有効性を示している．

参考文献

1) 内田誠一, 迫江博昭. “単調連続２次元ワープの効率化と拡張,” 信学技報, PRMU97-18, 1997.
2) 内田誠一, 迫江博昭. “動的計画法に基づく単調連続２次元ワープ法の検討,” 信学論, Vol.J81-D-II, No6, pp.1251−1258, 1998.
3) 三輪譲二, 小野政彦, 牧野正三, 城戸健一. “非線形スペクトルマッチングによる単語音声認識の一方式,” 信学論, Vol.J64-D, No.1, pp.46−53, 1981.
4) 松本　弘, 脇田　壽. “Frequency Warping による話者の正規化,” 音講論, 3-2-6, pp.587−588, Jun. 1984.
5) 中川聖一, 神谷伸, 坂井利之. “音声スペクトルの時間軸・周波数軸・強度軸の同時非線形伸縮に基づく不特定話者の単語音声の認識,” 信学論, Vol.J64-D, No.2, pp.116−123, 1981.
6) 迫江博昭, 藤井浩美, 吉田和永, 亘理誠夫. “フレーム同期化, ビームサーチ, ベクトル量子化の統合による DP マッチングの高速化,” 信学論, Vol.J71-D, No.9, pp.1650−1659, 1988.
7) 牧野正三, 二矢田勝行, 真舟裕雄, 城戸健一. “東北大-松下単語音声データベース,” 音響誌, Vol.48, No.12, pp.889−905, 1992.