九州大学大学院
システム情報科学紀要
第6巻　第1号　平成13年3月

Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.6, No.1, March 2001

区分線形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 Piecewise Linear Frequency Warping

Kei YAMADA , Seiichi UCHIDA and Hiroaki SAKOE

(Received December 15, 2000)

Abstract: An efficient algorithm for speaker-independent spoken word recognition is presented. This algorithm is based on the time-frequency warping with inter-frame consistency, where each frame of an input pattern is mapped to a reference pattern by controlling the mapping of several points(pivots) on the frame. The mapping of non-pivot points is given by linear interpolation between mapping of two consecutive pivots. The optimal mapping is obtained by using a dynamic programming based algorithm. The computational complexity of the algorithm is reduced to less than that of the previous time-frequency warping algorithm with inter-frame consistency. Experimental results show advantageous characteristics of the present algorithm.

Keywords: Spoken word recognition, Dynamic programming, Frequency warping, Deformable template, Speaker independent recognition

1. ま　え　が　き

不特定話者を対象とした単語音声認識を目的として，筆者らはフレーム間で連続な時間・周波数ワープ[2] の検討を行ってきた．これは非線形なスペクトルマッチングを2つのスペクトル間の非線形写像(ワープ)の最適化問題として定式化し，最適なワープを動的計画法(DP)を用いて探索する手法である．この手法は従来の時間・周波数ワープ[3)–6)]と異なり，隣接するフレームのスペクトル間に連続性の拘束を導入し，極端なワープを防ぐという特徴を持っている．

しかし実際には，DP実行の計算量はスペクトルの次元数に対して指数オーダーとなり，現実的な時間では実行不可能であった．高速化のためビームサーチを導入した近似解法を検討してきたが，認識精度に対する副作用が大きく，問題になっていた．

本論文では区分線形２次元ワープ[1] をスペクトルマッチング用に特殊化し，計算量を現実的なものとしたワープ手法を提案する．本手法では周波数方向のワープを区分線形近似し，その節点(ピボット)を制御し，ピボット間を線形内挿する．ワープの最適化はこのピボットに対して行う．このような区分線形関数によっても，個人性に由来するスペクトル変動の大部分は吸収できるものと考えられる．

最適なワープを求める時間計算量は，ピボットの個数に関しては指数オーダーであるが，スペクトルの次元数に関しては多項式オーダーとなる．よってピボット数を妥当な個数に設定できれば，従来法[2)]では事実上不可能であった最適なワープの探索が可能となる．

以下，**2.**で区分線形周波数ワープの定式化，およびDPによる解法，**3.**で実験結果および考察を示す．

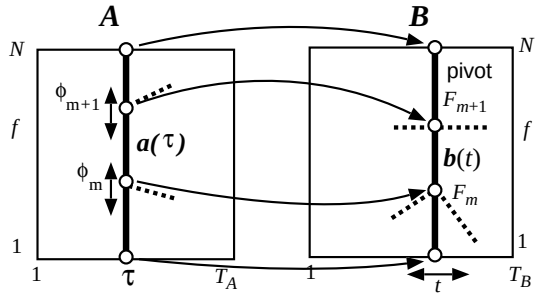

Fig.1　Piecewise linear frequency warping.

2. 区分線形周波数ワープ

本論文では，時間・周波数表現された入力パターン $\boldsymbol{A} = \{a(\tau,\phi) \mid \tau = 1,\ldots,T_A,\ \phi = 1,\ldots,N\}$ と標準パターン $\boldsymbol{B} = \{b(t,f) \mid t = 1,\ldots,T_B,\ f = 1,\ldots,N\}$ との間のマッチング問題を考える．ここに 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．パターン $\boldsymbol{A}$, $\boldsymbol{B}$ のある時刻における N 次元周波数スペクトル特徴ベクトルをフレームと呼び，それぞれ $\boldsymbol{a}(\tau) = \{a(\tau,\phi)|\phi = 1,\ldots,N\}$, $\boldsymbol{b}(t) = \{b(t,f)|f = 1,\ldots,N\}$と表す．

話者正規化を目的として各フレームで独立にスペクト

平成12年12月15日受付
* 知能システム学専攻博士後期課程（現在　沖電気工業（株））
** 知能システム学部門

ルマッチングを行う従来の時間・周波数ワープ[3)–6)]では，時間的に見て不連続にスペクトルを変形してしまう場合があり，結果として誤認識を生起する悪影響が懸念される．これに対し，フレーム間で連続な時間・周波数ワープ[2)]では極端なスペクトル遷移を排除できる．しかし計算量がスペクトルの次元数Nに対して指数オーダーになるという問題が新たに生じる．

本論文では，ワープの区分線形化によりフレーム間で連続な時間・周波数ワープの計算量を現実的なものとした手法(区分線形周波数ワープ)の検討を行う．

2.1　区分線形周波数ワープの定式化

標準パターン$\boldsymbol{B}$の第tフレーム，すなわち$\boldsymbol{b}(t)$上にM個の節点(t, f_m)，$(1 \leq m \leq M, 1 < f_m < f_{m+1} < N)$をあらかじめ設定しておく．これらの節点をピボットと呼ぶ．ピボットの位置(t, f_m)を時間tに関する関数として見たものをピボット軌跡と呼び，次式で表す．

$$f_m = F_m(t) \tag{1}$$

本論文ではピボット軌跡に対して以下の連続性条件を仮定しておく．

$$-1 \leq F_m(t+1) - F_m(t) \leq 1 \tag{2}$$

なおピボットの設定基準については，**3.1**で詳細を述べる．

区分線形周波数ワープ(**Fig.1**)では，パターン$\boldsymbol{A}$の第τフレーム$\boldsymbol{a}(\tau)$をパターン$\boldsymbol{B}$の第tフレーム$\boldsymbol{b}(t)$に対応付けることで時間変動を吸収しながら，さらにフレーム$\boldsymbol{a}(\tau)$上の各点(τ, ϕ)をフレーム$\boldsymbol{b}(t)$上に対応付けることで周波数変動を吸収する．この周波数変動の吸収に際し，ピボット(t, f_m)に対応する点のみ動的に決定し，それ以外の点の対応付けは線形補間により定める．具体的には，$\boldsymbol{a}(\tau)$上の点$\phi_m(\tau) = (\tau, \phi_m)$が$\boldsymbol{b}(t)$上のピボット$(t, f_m)$に対応するとすれば，$\boldsymbol{a}(\tau)$上のそれら以外の点$(\tau, \phi)$ $(\phi_m(\tau) < \phi < \phi_{m+1}(\tau))$の$\boldsymbol{b}(t)$上での対応点$(t, f)$を次式により与える．

$$(t, f) = \left(t(\tau),\ f_m + \frac{(f_{m+1} - f_m)(\phi - \phi_m(\tau))}{\phi_{m+1}(\tau) - \phi_m(\tau)} \right) \tag{3}$$

なお，$f_0 = 1$，$\phi_0(\tau) = 1$，$f_{M+1} = N$，$\phi_{M+1}(\tau) = N$とする．

以上の定義より，フレーム$\boldsymbol{a}(\tau)$のワープは，$(t(\tau), \phi_1(\tau), \ldots, \phi_m(\tau), \ldots, \phi_M(\tau))$という$M+1$個の変数により制御されることになる．ここで，$t(\tau)$が時間変動を吸収するための変数であり，それ以外が周波数変動を吸収するための変数である．よって区分線形周波数ワープでは，変数$(t(\tau), \phi_m(\tau) | \tau = 1, \ldots, T_A, m = 1, \ldots, M)$を制御し，パターン$\boldsymbol{A}, \boldsymbol{B}$の最大一致を図ることになる．

フレーム$\boldsymbol{a}(\tau)$のワープに関する評価量(コスト)を

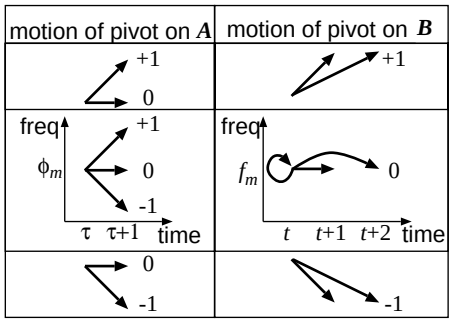

Fig.2　Example of pivot control.

$$d(t(\tau), \phi_1(\tau), \ldots, \phi_m(\tau), \ldots, \phi_M(\tau) | \tau) = \sum_{m=0}^{M} \sum_{\phi=\phi_m(\tau)}^{\phi_{m+1}(\tau)} \left| a(\tau, \phi) - b\left(t(\tau), f_m + \frac{(f_{m+1} - f_m)(\phi - \phi_m(\tau))}{(\phi_{m+1}(\tau) - \phi_m(\tau))} \right) \right| \tag{4}$$

と定義すれば，

$$\sum_{\tau=1}^{T_A} d(t(\tau), \phi_1(\tau), \ldots, \phi_m(\tau), \ldots, \phi_M(\tau) | \tau)$$

を最小化するワープが最大一致を与えることになる．上式の最小値を$D(\boldsymbol{A}, \boldsymbol{B})$とする．これは，時間周波数変動を吸収した後の2パターン$\boldsymbol{A}$,$\boldsymbol{B}$間の距離である．本論文の認識実験においては，この$D(\boldsymbol{A}, \boldsymbol{B})$をパターン間距離として用い，最短距離法により識別を行う．

2.2　制約条件

2.2.1　単調連続性制約

周波数変動は時間的に滑らかであり，またその範囲は比較的小さいと考えられる．そこで周波数方向のワープに連続性制約と整合窓制約を課する．

$$-1 \leq \phi_m(\tau+1) - \phi_m(\tau) \leq 1 \tag{5}$$

$$|\phi_m(\tau) - f_m| \leq w_f \tag{6}$$

一方，時間方向のワープ$\tau \mapsto t(\tau)$についても，その物理的な特性から，単調連続性制約および整合窓制約を課するのが自然である．

$$0 \leq t(\tau+1) - t(\tau) \leq 2 \tag{7}$$

$$|t(\tau) - \frac{T_B}{T_A}\tau| \leq w_t \tag{8}$$

2.2.2　ピボットの設定に応じた制約

異カテゴリに属する単語の過度の整合を排除するため，$\boldsymbol{B}$上のピボットの時間・周波数平面での動きと，その$\boldsymbol{A}$上での対応先$(\tau, \phi_m(\tau))$の動きを同期させることを考える．これはピボットの設定に応じた$\phi_m(\tau)$の制約に相当

●*Initialization*

for all $[\phi_1,\ldots,\phi_m,\ldots,\phi_M]$

$g(1,\phi_1,\ldots,\phi_M|1) := d(1,\phi_1,\ldots,\phi_M|1)$

●*Recursion*

for $\tau := 2$ **to** T_A

for all $[\phi_1,\ldots,\phi_m,\ldots,\phi_M]$

$g(t,\phi_1,\ldots,\phi_M|\tau) := d(t,\phi_1,\ldots,\phi_M|\tau)$

$$+ \min_{\substack{s\in\{0,1,2\}\\ p_m\in\{-1,0,+1\}}} g(t-s,\phi_1-p_1,\ldots,\phi_M-p_M|\tau-1)$$

●*Termination*

$$D(\boldsymbol{A},\boldsymbol{B}) = \min_{\phi_1,\ldots,\phi_m,\ldots,\phi_M} g(T_B,\phi_1,\ldots,\phi_M|T_A)$$

Fig.3 The present DP algorithm.

する．**Fig.2**にその様子を示す．

2.3 DPアルゴリズム

最適なワープを探索するDPアルゴリズムを**Fig.3**に示す．ここでは$(t(\tau),\phi_1(\tau),\ldots,\phi_m(\tau),\ldots,\phi_M(\tau)|\tau)$を$(t,\phi_1,\ldots,\phi_m,\ldots,\phi_M|\tau)$と略記している．ワークエリア$g(t,\phi_1,\ldots,\phi_m,\ldots,\phi_M|\tau)$は時刻1から$\tau$までの最小累積コストを格納している．なお，認識実験では距離$D(\boldsymbol{A},\boldsymbol{B})$のみを利用するため，各変数の最適値を求めるバックトラック処理は不要である．

以上のDPアルゴリズムの計算量は，$W_t = 2w_t + 1$, $W_f = 2w_f + 1$と表すと，各時刻tでの$(f_1,\ldots,f_m,\ldots,f_M)$の総数$W_tW_f^M$，DP漸化式の計算量$O(3^M+N)$より，$O(T_AW_tW_f^M(3^M+N))$となる．すなわち$M$に関しては依然指数オーダーであるが$N$に関しては多項式オーダーとなる．よってピボット数$M$を妥当な個数に設定できれば，現実的な時間で最適なワープの探索が可能となる．

3. 実　　験

3.1 音声試料および実験条件

東北大・松下単語音声データベース[7]の男性22人の50単語，計1100データを用いて従来法[2]と本手法の比較実験を行った．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．2名を標準パターン用グループ(話者ID: sp104，sp106)，残り20名分1000単語をテスト用の入力パターンとした．

ピボット数Mは1,2,3のいずれかを用いた．整合窓条件に関する定数は予備実験より$w_t = 4$，$w_f = 2$とした．

実験においては，$\boldsymbol{B}$上のピボットの配置法として2つの方式を採用した．すなわち時間軸に平行な直線で等間隔に設定したものと，スペクトルピークに設定したものを用いた．また定常母音部での過変形を排除するため，後者においては特に$|F_m(t+1) - F_m(t)|$=0となる区間で

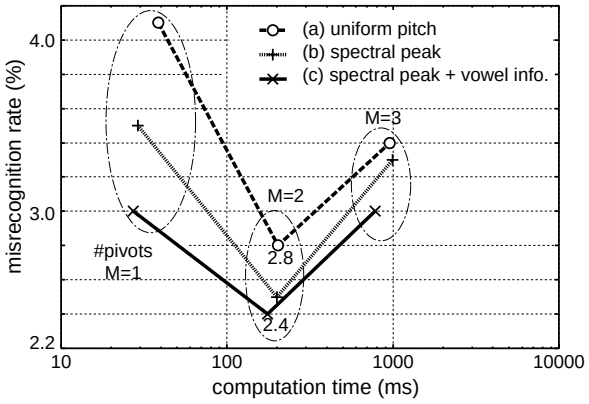

Fig.4 Misrecognition rate of the present algorithm as a function of computation time.

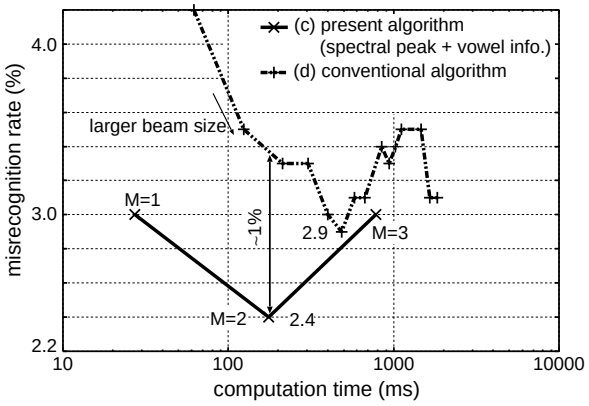

Fig.5 Misrecognition rate of the present algorithm and the conventional algorithm as a function of computation time.

$|\phi_m(\tau+1) - \phi_m(\tau)|$=0となるよう，母音情報を用いて制約条件を強化した場合についても実験を行なった．

3.2 実験結果および考察

ピボットを等間隔に設定した場合とスペクトルピークに設定した場合それぞれについて，ピボット数Mを変化させながら測定をした誤認識率と計算時間の関係を**Fig.4**のカーブ(a)，(b)に示す．ここで計算時間とは1組のパターン$\boldsymbol{A},\boldsymbol{B}$間にワープを求めるためにワークステーション(SPECint_95=12.3, SPECfp_95=20.2)が要した時間である．同じ計算時間で両者を比較すると，スペクトルピークに設定した場合(b)の方が高い認識率を得た．この理由は2つ考えられる．第1は(b)の場合**Fig.2**に示したピボット制御が効果的に働いたためと考えられる．第2はパワーの大きなスペクトルピークをピボットとすることで，線形補間による歪みを抑制できたためと考えられる．ピボットをスペクトルピークに設定した場合については**3.1**で述べたように，母音情報を用いて制約条件を強化した場合について実験を行なった．結果を**Fig.4**のカーブ(c)

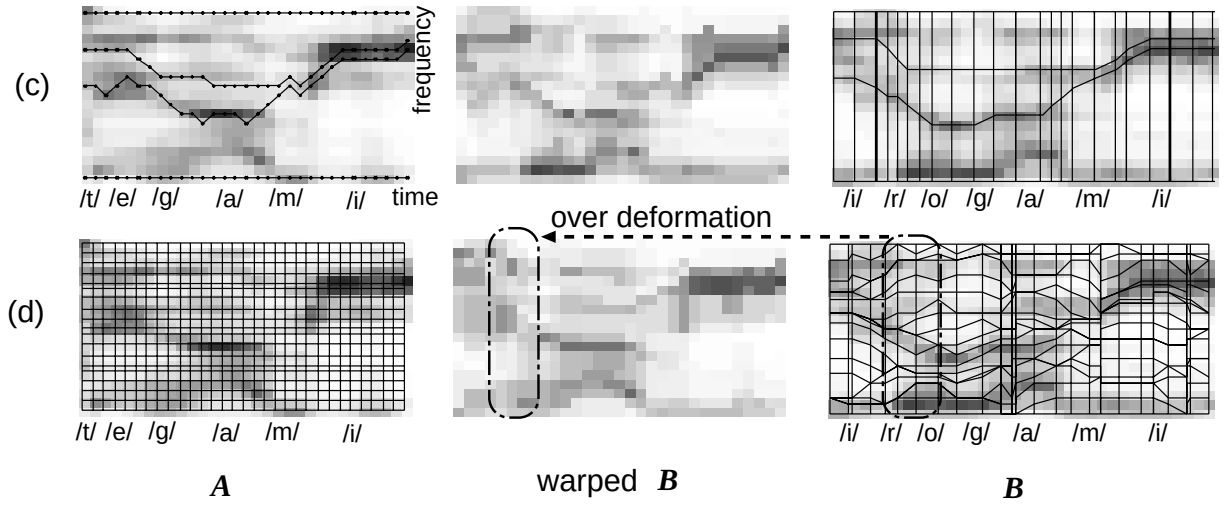

Fig.6 Warping between input $\boldsymbol{A}$(/tegami/) and reference $\boldsymbol{B}$(/irogami/) and warped $\boldsymbol{B}$. (c) Warping by the present algorithm with pivots on spectral peak and vowel information. (d) Warping by the conventional algorithm[2].

に示す．母音情報を用いない場合(b)と比較して母音情報を用いた場合(c)では，少ない計算時間で高い認識率が得られた．これは母音情報を用いてワープの自由度を押え探索空間を圧縮することで高速化が実現でき，母音部定常性がワープの特性に反映できたためと考えられる．

次に本手法及と従来法[2]の比較を行った．従来法ではビームサーチ(枝刈)を組み込み高速化を図ったアルゴリズムを用いた．それぞれの結果を**Fig.5**のカーブ(c)，(d)に示す．(c)においてはピボット数，(d)においてはビームサイズを変えることで条件を変化させている．同じ計算時間でみると，従来法と比較して本手法では常に高い認識率が得られた．理由としては2点考えられる．第1は，計算効率化による副作用の有無である．従来法ではビームサーチの利用により探索幅が圧縮されているため，その副作用として認識精度が劣化していると考えられる．一方，元々計算量の少ない本手法ではビームサーチを利用する必要がないため，そのような副作用は存在しない．第2に，本手法のワープの自由度がスペクトル変動を吸収するために必要十分であったことが考えられる．すなわち，区分線形化を図ってもスペクトル変動の吸収能力は十分保っており，むしろ従来法で問題視されていた過変形を回避する効果があったと思われる．

従来法で誤認識となった単語が本手法を用いることで正しく認識されるようになった例を**Fig.6**に示す．従来法(d)では/e/と/ro/が過剰にマッチングするという過変形が生じており，結果的に誤認識された．これに対し，本手法(c)ではこのような極端な整合が排除されており，パターン$\boldsymbol{A}$は正しく/tegami/と認識された．

4. む　す　び

フレーム間で連続な時間・周波数ワープの一方式として，区分線形周波数ワープの検討を行った．ワープの最適化処理をスペクトル上の節点(ピボット)に対して行うアルゴリズムの検討を行った．従来のフレーム間で連続な時間・周波数ワープに比べ，計算量は大幅に低減され，ピボット数Mを妥当な個数に設定すれば現実的な時間で最適なワープが求まることを示した．男性20人の50単語を対象とした認識実験では，従来法と同じ計算時間で約1%の認識率の改善がみられ，本手法の効果が示された．

参　考　文　献

1) 内田誠一, 迫江博昭. “区分線形２次元ワープ法の検討,” 信学論, Vol.J83-D-II, No.12, pp.2622–2629, 2000.
2) 山田 圭, 内田誠一, 迫江博昭. “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,” 九州大学大学院システム情報科学研究科報告, Vol.3, No.2, pp.197–202, 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, 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) 岡 隆一. “時空間DPによるパターン・マッチング・アルゴリズムとそのセル空間表現,” 信学技報, PRL77-6, pp.1–9, 1977.
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.