002

TR-IT-0257

周波数ワーピングに基づく話者正規化手法の検討

Frequency Warping Based Speaker Normalization

杉田 記男 Norio Sugita
内藤 正樹 Masaki Naito

1998年4月9日

話者による音声の音響的特徴の広がりを正規化した後に音響モデルの学習を行い不特定話者音声認識の性能向上を図るため、声道長に着目した周波数ワーピングによる話者正規化手法が提案されその有効性が報告されている。本稿では、周波数ワーピング関数を決定する手段として一般に用いられる (1) ゆう度に基づく周波数ワーピング関数の選択、(2) フォルマントに基づく周波数ワーピング関数の推定の2つの手法について認識実験を通じてその性能の比較・検討を行う。

目次

1 はじめに

音声の音響的な特徴は話者の声道形状の差異などにより話者毎に大きく異なっており、この点が不特定話者音声認識を困難にする大きな要因となる。この話者による音声の音響的特徴の広がりを正規化し音響モデルの学習を行うことで不特定話者音声認識の性能向上を図る話者正規化手法が提案されている。

この話者正規化手法の一つのアプローチとして、声道長に着目した周波数ワーピングによる話者正規化手法が提案されその有効性が報告されている。この手法は、話者毎に音声のスペクトルを周波数方向に伸縮する方法で話者による音声の音響的特徴の広がりを抑える手法であり、周波数ワーピングによる話者正規化を行う際の周波数ワーピング関数を決定する手段として、(1) 予め複数の周波数ワーピング関数を用意し、各周波数ワーピング関数を用いた音響分析結果に対する HMM のゆう度を算出し最もゆう度の高いワーピング関数を選択する方法 [1]、(2) 各話者の音声のフォルマント周波数を基に周波数ワーピング関数を推定する方法 [1]、(3) 各話者の音声から逆推定した声道長等の声道形状の特徴量を基に周波数ワーピング関数を推定する方法等の手法が提案されている。本報告では、この内の (1) ゆう度に基づく周波数ワーピング関数の選択、(2) フォルマントに基づく周波数ワーピング関数の推定の 2 つの手法について認識実験を通じてその性能の比較・検討を行う。

2　周波数ワーピングによる話者正規化

周波数ワーピングによる話者正規化手法は、話者毎に音声のスペクトルを周波数方向に伸縮することで話者による音声の音響的特徴の広がりを抑える手法である。本手法により音声認識の性能向上を計るためには、音響モデルの学習・認識時に、各話者に最適な周波数ワーピング関数を推定する必要がある。本稿では、この周波数ワーピング関数を推定する手段として、(1) 予め複数の周波数ワーピング関数を用意し、各周波数ワーピング関数を用いた音響分析結果に対する HMM のゆう度を算出し最もゆう度の高いワーピング関数を選択する方法、(2) 各話者の音声のフォルマント周波数を基に周波数ワーピング関数を推定する方法、の 2 つの手法について比較検討を行なう。本章では、これら 2 つの手法の詳細について述べる。

2.1　尤度に基づく話者正規化

ゆう度に基づく話者正規化手法は、予め複数の周波数ワーピング関数を用意し、これらの関数を用い周波数ワーピングの後に音響分析を行ない、その結果得られる音響パラメータが初期音響モデルから出力されるゆう度を求め、最もゆう度の高いワーピング関数を選択する方法である。

以下に、尤度に基づく最適な周波数ワーピング関数の選択方法と話者正規化学習の手順について説明する。まず、周波数ワーピング関数の選択方法について以下に示す。ここでは, 予め用意した n 個の周波数ワーピング関数 $F \in \{f_1, f_2 \ldots, f_N\}$ から各話者に最適な周波数ワーピング関数を選択するものとする (図 1)。

1. ある話者 m に対して予め用意した周波数ワーピング関数 $F \in \{f_1, f_2 \ldots, f_N\}$ を用い音響分析を行なう。
2. (1) により得られた音響分析結果それぞれについて、 Viterbi 探索によりそのゆう度を求める。
3. (2) の結果を基に、 $F \in \{f_1, f_2 \ldots, f_N\}$ 中で、最大ゆう度を与える周波数ワーピング関数 f_max を選択する。

次に話者正規化学習の手順について示す。ここで、学習の際には、周波数ワーピング関数選択用音声データ, 学習用音声データの 2 つの異なる音声データセットを用いるものとする。

1. 全学習話者の学習用音声データを用い、音響モデル Λ_0 を学習する $(i = 0)$。
2. 音響モデル Λ_i を基に、各学習話者の周波数ワーピング関数選択用音声データに対して最大尤度を与える周波数ワーピング関数 F_{max} を選択する。
3. 話者毎に選択された周波数ワーピング関数を用い学習用音声データの音響分析を行なう。
4. (3) の結果得られた全話者の音響分析結果を用い話者正規化音響モデル Λ_i の学習を行なう。
5. 指定した回数 (2)-(4) を繰り返す $(i = i + 1)$。

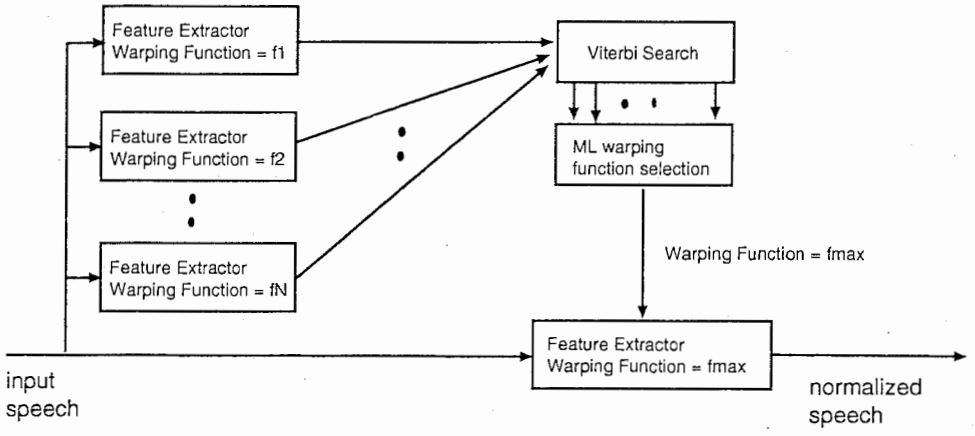


図 1: 尤度に基づく周波数ワーピング関数の選択法

本稿では、ゆう度に基づく周波数ワーピング関数の選択を行なう際に以下に示す3種類の周波数ワーピング関数を使用する。

周波数ワーピング関数1 周波数ワーピング前後の周波数の対応関係を周波数ワーピング係数αによって定まる直線の周波数ワーピング関数で表すもので、周波数ワーピング周波数f'は次式により表される。

$$f' = \alpha \cdot f \tag{1}$$

$$f' = 1.0 \quad (f' > \alpha) \tag{2}$$

ここで、f',fはそれぞれ、周波数ワーピング前後の対応する周波数で、ナイキスト周波数を1として正規化した値である。また、図(2)に周波数ワーピング前後の周波数の対応関係を図示している。

周波数ワーピング関数2 関数2は図3に示すように、基本的に関数1と同様な1次関数であるが、係数ϕを定めることにより、入力音声の正規化周波数fがϕ以下のときは周波数ワーピング関数を

$$f' = \alpha \cdot f \qquad (0 < f \leq \phi) \tag{3}$$

で与え、fがϕから1の区間においては、$(\phi, f \cdot \phi)$と$(1.0, 1.0)$の二点間を結ぶ直線

$$f' = \frac{(\alpha \cdot \phi - 1) \cdot f - (\alpha - 1) \cdot \phi}{\phi - 1} \qquad (\phi < f \leq 1.0) \tag{4}$$

で与えるものである。

周波数ワーピング関数3 関数3は次式により表され、図4に見られるように、$\alpha > 0$のときは上向き、$\alpha < 0$のときは下向きの弧を描く曲線である。

$$f' = \frac{f \cdot (\alpha + 1)}{\alpha \cdot f + 1} \tag{5}$$

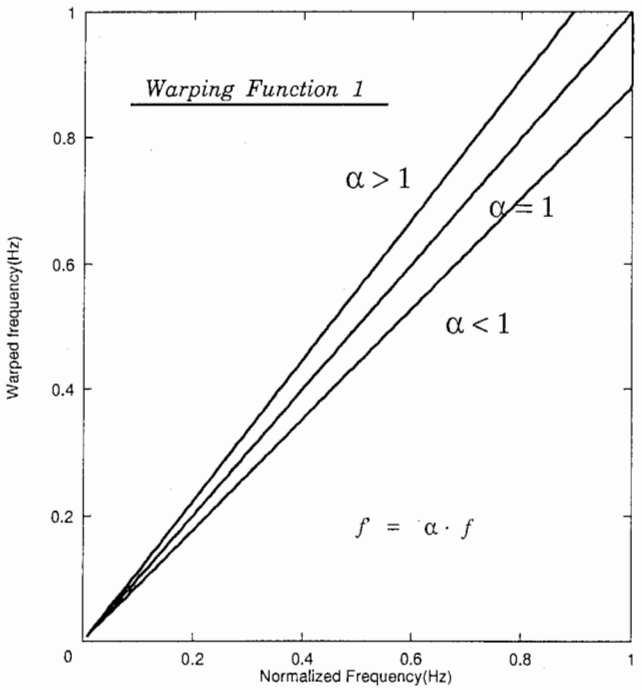


図 2: 周波数ワーピング関数1

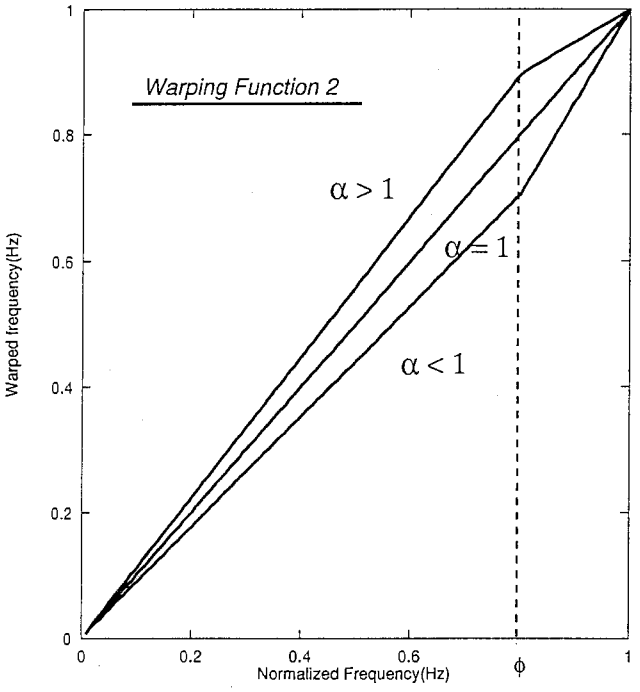


図 3: 周波数ワーピング関数 2

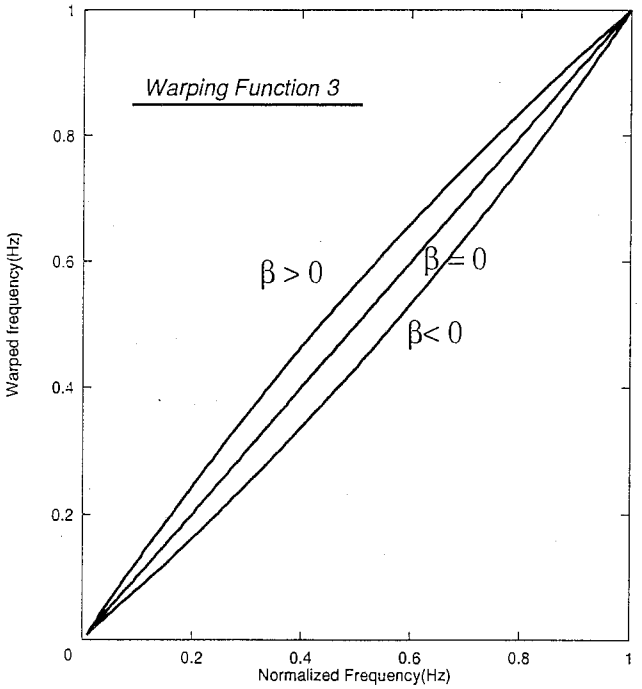


図 4: 周波数ワーピング関数 3

2.2 フォルマントに基づく話者正規化

フォルマントに基づく話者正規化は各話者の音声の母音のフォルマント周波数の平均値を基に以下の手順に従い周波数ワーピング関数を推定するものである (図 1)。

1. 入力音声を音響分析し **Viterbi** 探索により音素境界を求める。
2. フォルマント周波数の抽出を行ない、(1) で得られた音素境界を基に母音のフォルマント周波数 (**F1,F2,F3,F4**) の平均を求める。
3. 学習話者全員のフォルマント周波数の平均値と、話者毎のフォルマント周波数の平均値を基に周波数ワーピング関数を推定する

ここで、(2) でのフォルマントの抽出には **xwaves** の **formant** コマンドを使用し、フレーム毎に求められたフォルマント周波数の単純平均を計算した。なお、フォルマント周波数に基づく周波数ワーピング関数の推定法の詳細については後述する。

次に話者正規化学習の手順について示す。ここで、学習の際には、周波数ワーピング関数推定用音声データ, 学習用音声データの **2** つの異なる音声データセットを用いるものとする。

1. 全学習話者の学習用音声データを用い、音響モデル Λ_0 を学習する。
2. 先に示した方法により求められた、学習用音声データのフォルマント周波数の平均値を基に、各話者の周波数ワーピング関数を推定する。
3. 話者毎に推定された周波数ワーピング関数を用い学習用音声データの音響分析を行なう。
4. (3) の結果得られた全話者の音響分析結果を用い話者正規化音響モデル Λ_1 の学習を行なう。

この場合、前節に示したゆう度を用いた周波数ワーピング関数の推定法と異なり、話者正規化モデルの繰り返し学習は行わない。これは、話者正規化後の音響モデルを用い再度 **viterbi** 探索により音素境界の検出を行った場合でも、検出される音素境界の変化は大きくないと予想されることから、算出されるフォルマント周波数の平均値にも大きな変化は無いと考えるためである。

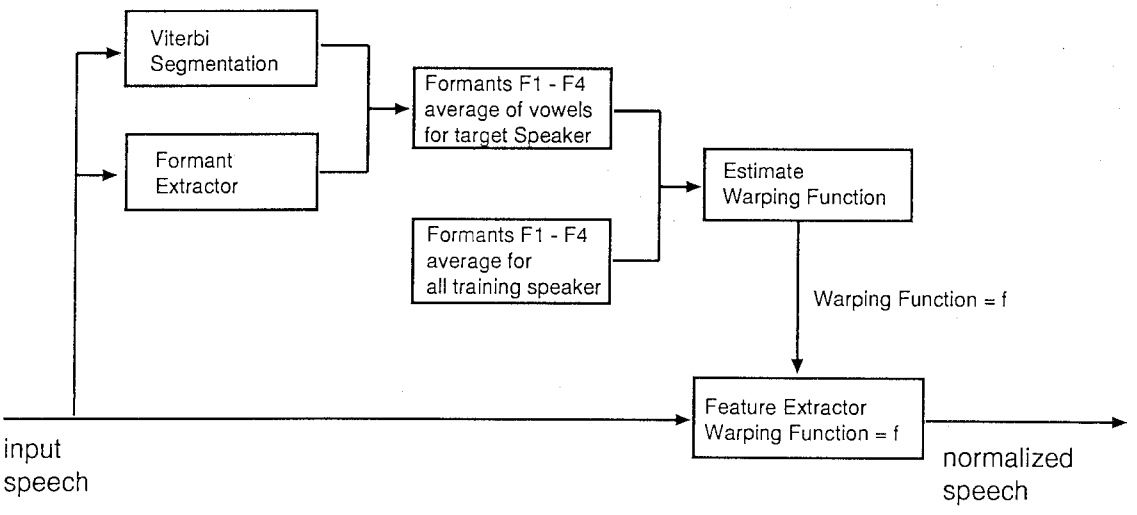


図 5: 母音のフオルマント周波数に基づく周波数ワーピング関数の推定

(2.2.1)　フォルマント周波数を用いた周波数ワーピング関数の推定

本稿では、学習話者全員のフォルマント周波数の平均値と、話者毎のフォルマント周波数の平均値の対応関係を基に以下に示す5種類の方法で周波数ワーピング関数を推定するここでは、学習用の各話者の母音音素のフォルマント周波数の平均を $F_{1s},\ldots,F_{4s}$、学習話者全体の母音音素のフォルマント周波数の平均を $F_1r,\ldots,F_4r$ とする。

周波数ワーピング関数 (関数F1, 関数F2, 関数F3, 関数F4)　これらの関数は、学習用の各話者と学習話者全体の各フォルマント周波数 (F_{ns},F_{nr}) の比を基に次式により周波数ワーピング係数 α_n を定め、

$$\alpha_n = \frac{F_{nr}}{F_{ns}} \tag{6}$$

ゆう度による周波数ワーピング関数の選択の際の関数2と同様に、ワーピング周波数 f' を

$$f' = \alpha_n \cdot f \qquad (0 < f \leq \phi) \tag{7}$$

$$f' = \frac{(\alpha_n \cdot \phi - 1)\cdot f - (\alpha_n - 1)\cdot \phi}{\phi - 1} \qquad (\phi < f \leq 1.0) \tag{8}$$

で与えるものである。

周波数ワーピング関数 (関数F1-4)　これは、周波数ワーピング関数 関数F1-関数4と異なり図6に示すように、学習用の各話者フォルマント周波数と、学習話者全体のフォルマント周波数平均 $F1,F2,F3,F4$ の各点を結んだ周波数ワーピング関数であり、フォルマント周波数 F_n と F_{n+1} を結ぶ関数を f' 以下の式であらわす。

$$f' = \frac{(F_{nr} - F_{(n+1)r})\cdot f + (F_{ns} - F_{(n+1)s}) - F_{nr} + F_{(n+1)r}}{F_{ns} - F_{(n+1)s}} \qquad (n \leq f < n+1) \tag{9}$$

ここで、

$$f'_\phi = \frac{(F_{nr} - F_{(n+1)r})\cdot \phi + (F_{ns} - F_{(n+1)s}) - F_{nr} + F_{(n+1)r}}{F_{ns} - F_{(n+1)s}} \qquad (n < f \leq n+1) \tag{10}$$

と置くと、$\phi \leq f \leq 1$ のときの f' は

$$f' = \frac{(f'_\phi - 1)\cdot f + \phi - f'_\phi}{\phi - 1} \qquad (\phi \leq f \leq 1) \tag{11}$$

2.3　周波数ワーピングのアルゴリズム

本節では、前節までに示した周波数ワーピング関数を用いた周波数ワーピングの実現方法について述べる。周波数ワーピングは、周波数ワーピング関数に従い、周波数ワーピング後の各周波数のパワーの値を、入力音声スペクトル上の対応する周波数のパワーで置き換えることで実現される。本報告では、音響パラメータとしてMFCCを使用するが、この場合、MFCCの計算時にFFTにより入力音声パワースペクトルが計算される。このパワースペクトルはFFTにより周波数方向に離散的に求められており、多くの場合、周波数ワーピング後の各周波数に対応する周波数ワーピング前の周波数のパワーが計算されておらず、周波数ワーピング後のパワースペクトルを直接的に求めることが出来ない。このため、本報告では入力音声の各周波数のパワーを基にした線形補間により周波数ワーピング後の各周波数のパワーを近似することで、周波数ワーピングを実現する (図7 参照)。

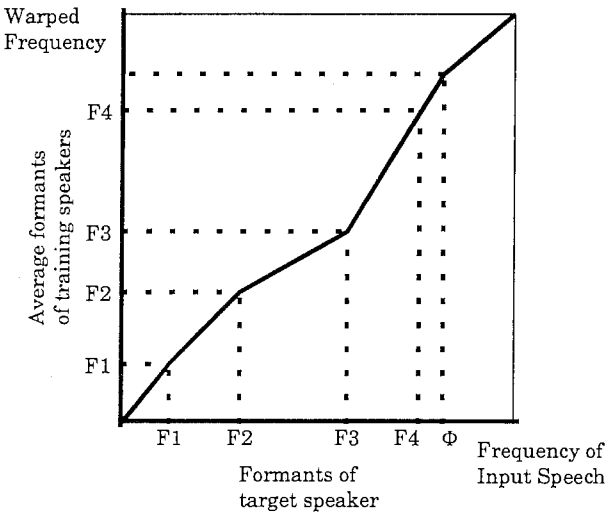


図 6: 母音のフォルマント周波数 $(F1 - F4)$ に基づいた周波数ワーピング関数

周波数ワーピングの手順を以下に示す。

1. **FFT** により入力音声のパワースペクトル $S[f_{in}]$ ($f_{in} = 1, 2, 3, \ldots N$) を求める ($N$ は **FFT** の分解能)。
2. 以下の処理を繰り返し、周波数ワーピング後の **FFT** の各周波数 f_{warp} ($f_{warp} = 1, 2, 3, \ldots N$) のパワースペクトルを求める。
3. 周波数ワーピング関数を基に f_{warp} に対応する入力音声の周波数 f_{rin} (実数) を求める。
4. f_{rin} に隣接する **FFT** によりパワースペクトルが算出された周波数 f_{lin} (低周波数側), f_{uin} (高周波数側) を求める。
5. 式 **12**に従い、入力音声の f_{lin}, f_{uin} におけるパワーの線形補間を行い、入力音声の周波数 f_{rin} におけるパワースペクトルを近似し周波数ワーピング後の **FFT** のポイント f_{warp} のパワースペクトル ($S'[f_{warp}]$) とする。

$$S'[f_{warp}] = S[f_{lin}] + \frac{f_{rin} - f_{lin}}{f_{uin} - f_{lin}}(S[f_{uin}] - S[f_{lin}]) \tag{12}$$

ここで、

$$f_{uin} - f_{lin} = 1 \tag{13}$$

以上の手順によりパワースペクトル上で、周波数ワーピングを行った後音響パラメータ **MFCC** を算出する。

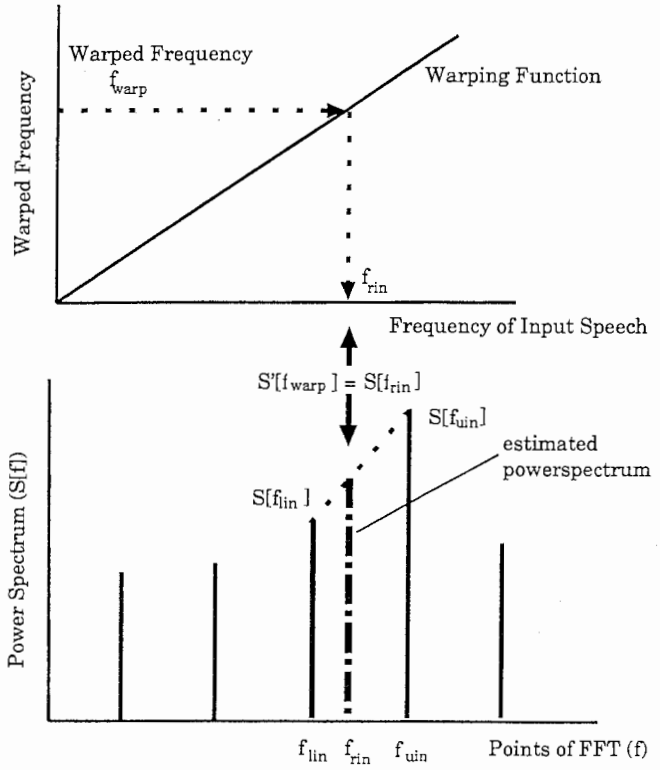


図 7: パワースペクトルの線形補間による周波数ワーピング

表 1: 実験条件

音響分析	
サンプリング周波数 12kHz, preemphasis 0.98	
フレーム周期 10 ms, フレーム長 20 ms (Hamming 窓)	
特徴パラメータ	logpower,Δlogpower, 12 次-MFCC,12 次-ΔMFCC (フィルタバンク次数 24)
音響モデル (HMnet) の構成	
音声：総状態数 800 各状態 1 混合 (ML-SSS により作成)	
無音：3 状態 10 混合	
学習データ	男性 138 名, 音素バランス文 A (50 文)
周波数ワーピング関数推定用データ	
音素バランス文 A N 文	
評価データ	男性 10 名, 音素バランス文 B (50 文)

3 認識実験

3.1 実験条件

前章で述べた周波数ワーピングに基づく話者正規化手法の性能比較を行うため、音素タイプライタによる認識実験を行った。実験条件を表 1に示す。なお、実験に用いた HMnet の構造は、旅行対話音声データベース (TRA) を用い ML-SSS アルゴリズム [2] により作成したモデルである [3]。また、ゆう度に基づく周波数ワーピング関数の選択の際の初期音響モデル、フォルマント周波数に基づく周波数ワーピング関数の際の Viterbi 探索による音素境界の推定には、表 1の学習話者 138 名での音声データを用いて作成した男性モデルを使用した。

3.2 周波数ワーピング関数推定用データ量の検討

まず、学習時に各話者の周波数ワーピング関数を推定する際に用いるデータ量を決定するため、推定に用いる文数による、推定された周波数ワーピング関数の変動について調査した。実験は、ゆう度に基づく周波数ワーピング関数 2 を用い、周波数ワーピング係数等の条件については、周波数ワーピング係数を $\alpha = \{0.88, 0.90, 0.88 + 0.02 \times n, \cdots, 1.12\}$ の 13 種類とし、$\Phi = 0.8$ に固定し、これらの α から最もゆう度の高い係数を選択した。実験は評価用話者 10 名を含んだ 20 話者について行ない、５０文で選択した際の α を基準とし、５０文で選択した際の α と N 文で選択したの際の α の差を話者別に図 8に示した。
この結果、周波数ワーピング関数を推定に用いる文数が少ない場合、推定用文数により周波数ワーピング係数が大きく変動しているが、推定に 30 文以上の音声データを使用することで推定される周波数ワーピング係数が安定することが観察された。但し以降の実験では周波数ワーピング関数を推定に要する処理時間を考慮し話者正規化学習を行なう際には、ゆう度に基づく場合、フォルマント周波数に基づく場合各共に、各話者 20 文の音声データを用い周波数ワーピング関数の推定を行なった。

3.3 ゆう度による周波数ワーピング関数選択の評価

ゆう度に基づく周波数ワーピング関数選択による話者正規化の性能評価を行なうため、音素タイプライタによる認識実験を行った。実験には、ゆう度を比較する周波数ワーピング関数として、前章に示した関数 1, 関数 2, 関数 3 を用い、関数 1 を用いる場合, 周波数ワーピング係数 $\alpha = \{0.88, 0.90, 0.88 + 0.02 \times n, \cdots, 1.12\}$ の 13 種類、関数 2 を用いる場合,$\Phi = 0.8$ に固定し周波数ワーピング係数 $\alpha = \{0.88, 0.90, 0.88 + 0.02 \times n, \cdots, 1.12\}$ とした 13 種類、関数 3 を用いる場合, 周波数ワーピング係数 $\beta = \{0.30, 0.25, 0.20, 0.15, 0.10, 0.05, 0, -0.04, -0.08, -0.12, -0.16, -0.20, -0.24\}$ の 13 種類を用意し最もゆう度の高い周波数ワーピン係数を選択するものとした。また、周波数ワーピン関数の選択、話者正規化学習の繰り返しの効果について確認するため、初期モデルも含め周波数ワーピン関数の選択、話者正規化学習が 1 回終了する毎に認識実験を行なっている。その際には話者正規化学習の場

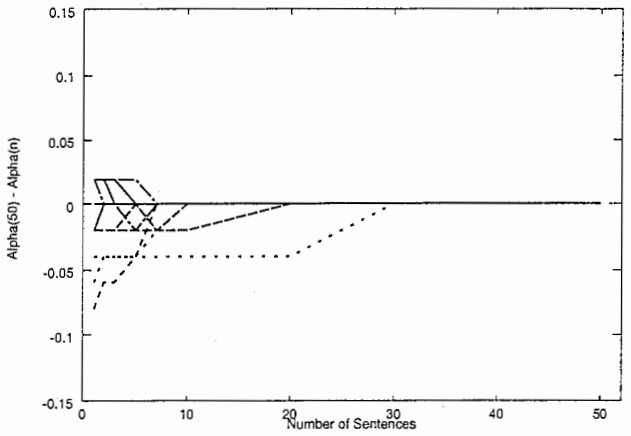


図 8: 周波数ワーピング推定に用いた文数による α の推移 (関数 2)

表 2: 周波数ワーピング関数の形状・推定方法による認識率の比較 (音素認識率 (%))

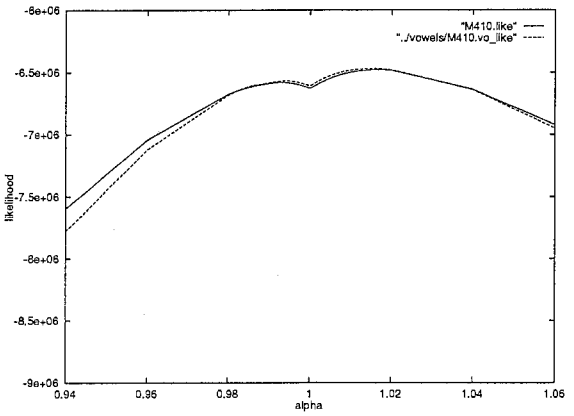
		初期モデル	学習回数 1	学習回数 2
話者正規化無し		80.60	-	-
関数 1	全音素	81.25	81.40	81.46
	母音	81.38	81.55	81.48
関数 2	全音素	81.27	81.37	81.37
	母音	81.44	81.41	81.43
関数 3	全音素	81.49	81.65	81.68
	母音	81.38	81.48	81.84

合と同様各話者 20 文の音声データを用い周波数ワーピン関数の選択を行なった後認識を行なっている。また、音素により話者の声道長の違いが音声の音響的な特徴に与える影響が異なると予想されることから、周波数ワーピング関数の選択において各周波数ワーピング関数別のゆう度を求める際に、全音素のゆう度を用いた場合 (全音素) と、特に声道長の違いが大きく影響すると予想される母音のゆう度のみを用いた場合 (母音) についても比較を行なった。

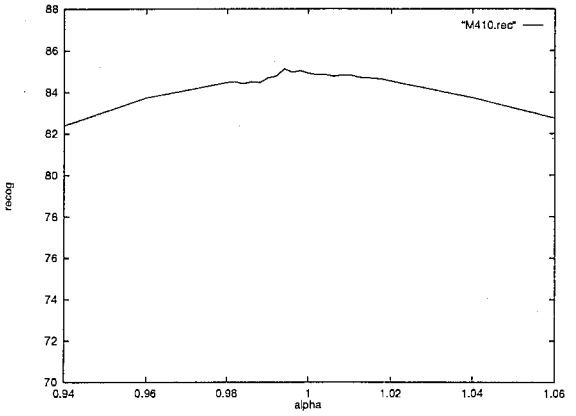
実験結果を表 2に示す。この結果話者正規化を行なわない場合と比較すると、話者正規化学習を行なわない初期モデルを用い周波数ワーピング関数選択のみを行なった場合も含め認識率が改善していることが分かる。周波数ワーピング関数の形状としては関数 3 を用いた場合に認識率が最も高く、声道長の違いが大きく影響すると予想される母音のゆう度のみを用いた場合 (母音) に高い認識性能が得られている。また、認識率の向上はあまり大きくないものの周波数ワーピン関数の選択、話者正規化学習の繰り返しの効果が確認された。総合的に見ると、周波数ワーピング関数の形状としては関数 3 を用い母音のゆう度のみを用い周波数ワーピング関数を選択を行ない話者正規化学習を 2 回繰り返した後に、最も音素認識率 81.84% が得られた。これは話者正規化を行なわない場合の誤認識の約 6.4% 削減に相当している。

また、図 11- 図 16に学習用話者 138 名に対して、関数 1, 関数 2, 関数 3 を用い全音素のゆう度、母音のゆう度を用いて選択された周波数ワーピング係数の分布を、音響モデルとして学習回数 0(= 初期モデル), 学習回数 1 を用いた場合について示した。またゆう度の算出法による選択される周波数ワーピング係数の比較を行なうため周波数ワーピング関数の選択に音響モデルとして学習回数 1 を用いた場合について、2 つのゆう度算出法により選択された周波数ワーピング係数の対応関係を示した。これらの図からいずれの周波数ワーピング関数を用いた場合にも話者正規化を行なわない周波数ワーピング係数 $\alpha.\beta = 1.0$ がほとんど選択されていないことが分かる。それ以外の周波数ワーピング係数については、正規分布に近い分布の形状をしていることから、実験に用いたソフトウェアに問題があることも考えられる。この点について検証を行なうため、評価用話者の中で、周波数ワーピング係数 $\alpha = 1.0$ に近い周波数ワーピング係数が選択された 2 話者 (M410,M413) について、周波数ワーピング関数と

して関数 2 を用い、各周波数ワーピング係数における、周波数ワーピング係数選択用文章に対するゆう度 (図 9,9) と評価用音声データに対する認識率を示した。ここでゆう度を計算する際の音響モデルとしては話者正規化学習を行なう前の男性モデルを使用している。この結果、ゆう度について見ると周波数ワーピング係数 $\alpha = 1.0$ 付近でゆう度が低下することが分かる。一方認識率で見た場合、周波数ワーピング係数 $\alpha = 1.0$ 付近で若干の認識率の変動が見られるものの認識率の大きな劣化は見られなかった。これらは、周波数ワーピングの際に行なわれるパワースペクトルの線形補間の影響とも考えれれるが詳細な検討は行なっていない。

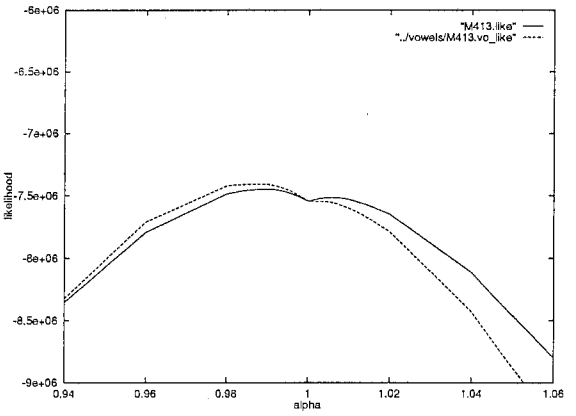


周波数ワーピング係数毎のゆう度

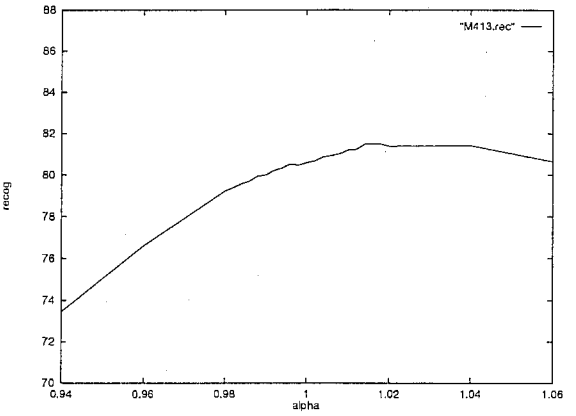


周波数ワーピング係数毎の音素認識率 (%)

図 9: 周波数ワーピング係数別のゆう度と音素認識率 (話者 M410)

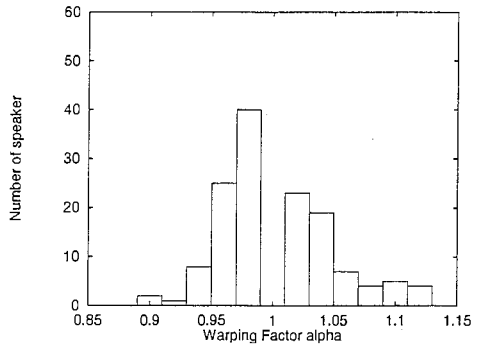


周波数ワーピング係数毎のゆう度

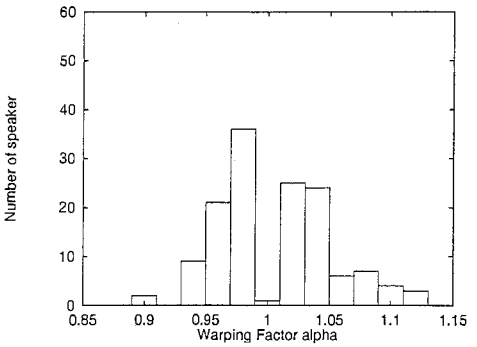


周波数ワーピング係数毎の音素認識率 (%)

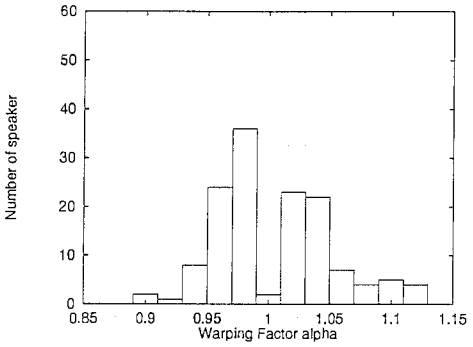
図 10: 周波数ワーピング係数別のゆう度と音素認識率 (話者 M413)



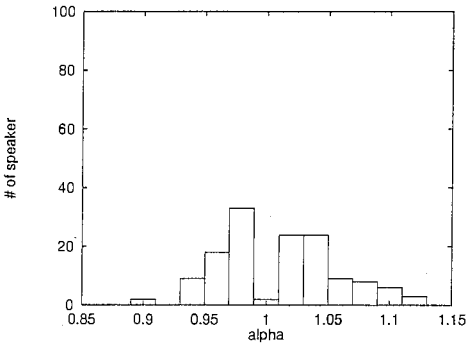
全音素のゆう度により選択 (学習回数 =0)



母音のゆう度により選択 (学習回数 =0)



全音素のゆう度により選択 (学習回数 =1)



母音のゆう度により選択 (学習回数 =1)

図 11: ゆう度により選択された周波数ワーピング係数の分布 (周波数ワーピング関数 1 の場合)

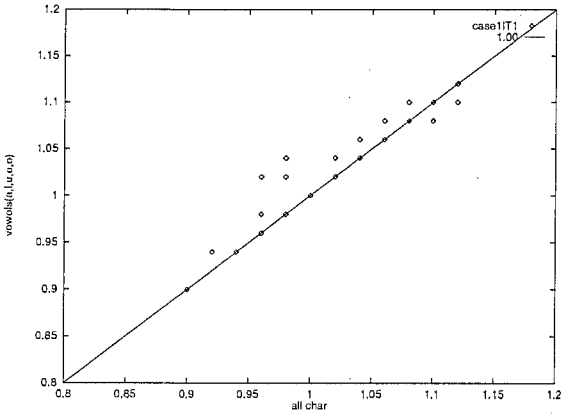
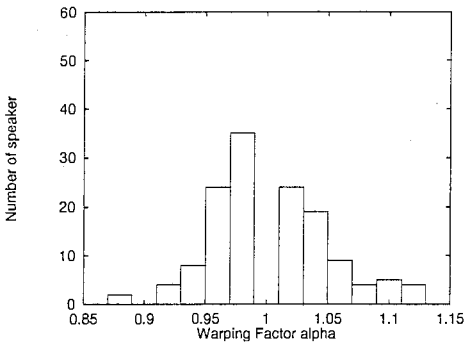
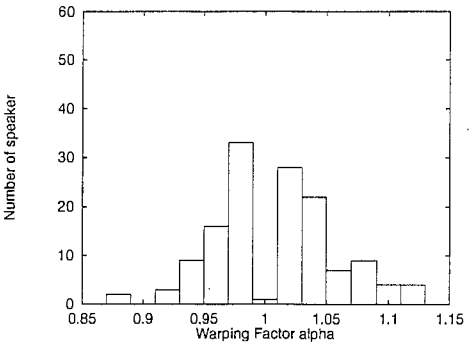


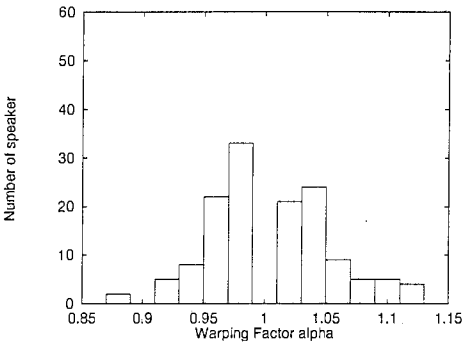
図 12: 全音素と母音のゆう度により選択された周波数ワーピング係数の比較 (周波数ワーピング関数 1, 学習回数 =1)



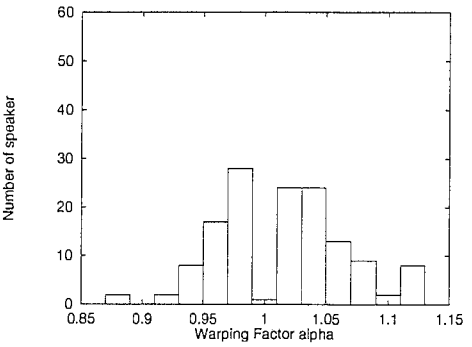
全音素のゆう度により選択 (学習回数 =0)



母音のゆう度により選択 (学習回数 =0)



全音素のゆう度により選択 (学習回数 =1)



母音のゆう度により選択 (学習回数 =1)

図 13: ゆう度により選択された周波数ワーピング係数の分布 (周波数ワーピング関数 2 の場合)

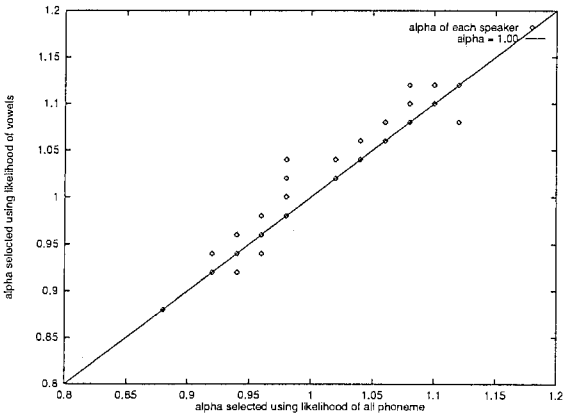
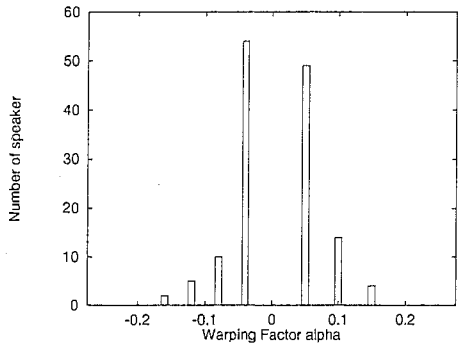
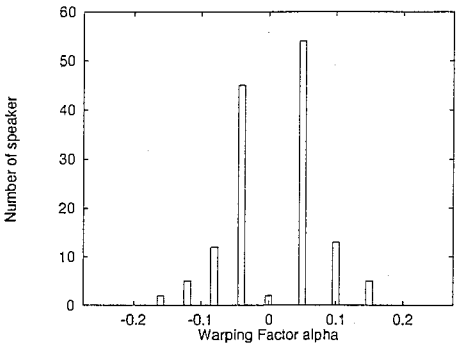


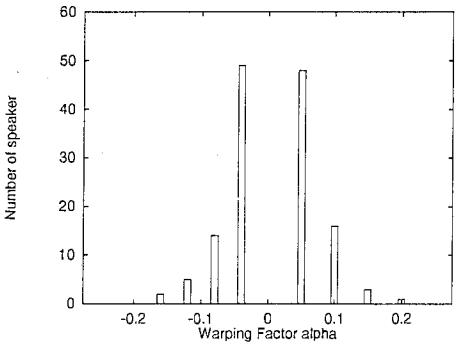
図 14: 全音素と母音のゆう度により選択された周波数ワーピング係数の比較 (周波数ワーピング関数 2, 学習回数 =1)



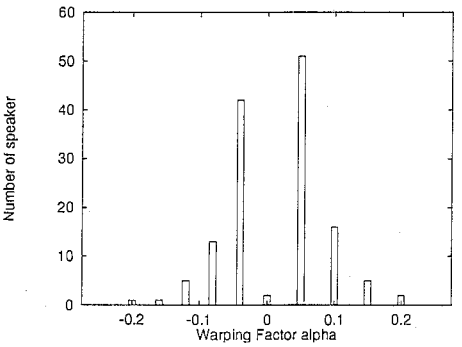
全音素のゆう度により選択 (学習回数 =0)



母音のゆう度により選択 (学習回数 =0)



全音素のゆう度により選択 (学習回数 =1)



母音のゆう度により選択 (学習回数 =1)

図 15: ゆう度により選択された周波数ワーピング係数の分布 (周波数ワーピング関数 3 の場合)

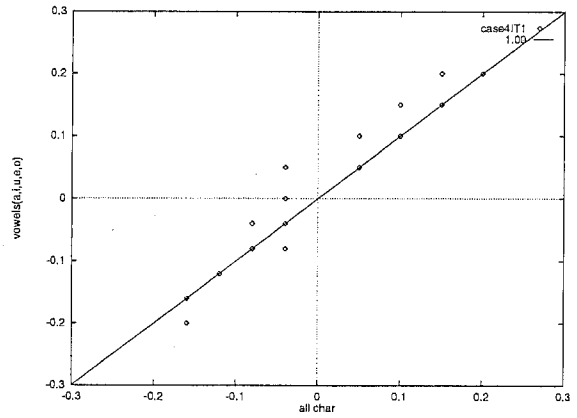


図 16: 全音素と母音のゆう度により選択された周波数ワーピング係数の比較 (周波数ワーピング関数 3, 学習回数 =1)

3.4 フォルマントによる周波数ワーピング関数推定の評価

フォルマント周波数に基づく周波数ワーピング関数推定による話者正規化の性能評価を行なうため、音素タイプライタによる認識実験を行った。実験においては、ゆう度に基づく周波数ワーピング関数選択の場合と同様に、話者正規化学習、認識時共に各話者 20 文の音声データを用い周波数ワーピン関数の推定を行なった。実験は、周波数ワーピング関数推定法として前章に示した関数 **F1**, 関数 **F2**, 関数 **F3**, 関数 **F4**, 関数 **F1-4** を用いて行なった。図 17に学習話者 138 名のフォルマント周波数 F1-F4 の分布を示す。各フォルマント周波数の平均値は、$\overline{F1} = 451.06$, $\overline{F3} = 2836.76$, $\overline{F2} = 1597.50$, $\overline{F4} = 3880.75$ である。

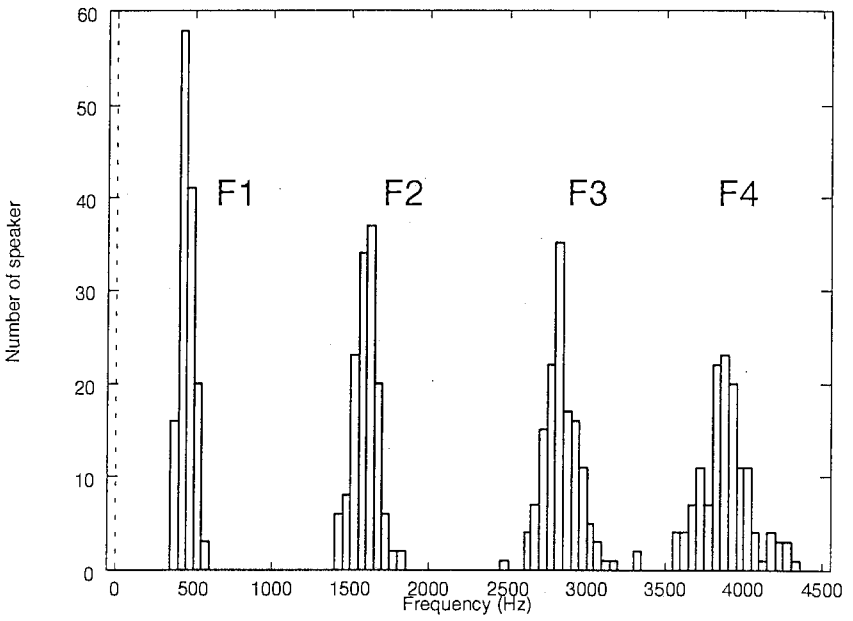
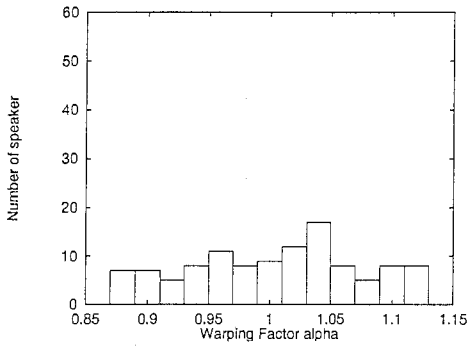


図 17: 学習話者 138 名のフォルマント周波数 F1-F4 の分布

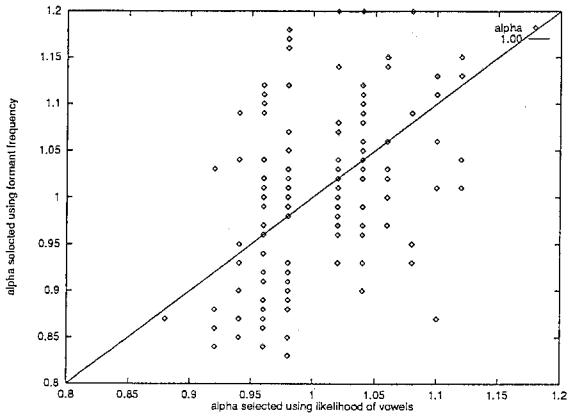
実験結果を表 2に示す。表には比較のため話者正規化を行なわない男性モデルの認識率、フォルマント周波数により推定される周波数ワーピング関数と関数の形状が類似している母音のゆう度に基づく周波数ワーピング関数選択 (関数 2) についての認識率を共に示した。ここで、表中の初期モデルは音響モデルとして男性モデルを用い認識時にのみ話者正規化を行なった場合の認識率、話者正規化モデルは話者正規化学習を行なったモデルの認識率である。この結果、周波数ワーピング関数の推定法 (関数 **F2**) を用いた場合に最も高い認識性能 (音素認識率 **81.66%**) が得られた。周波数ワーピング関数の推定法 (関数 **F1-4**) を用いた場合にもゆう度に基づく周波数ワーピング関数選択 (関数 2) を上回る認識性能が得られた。関数 **F1-4** は関数 **F2** と比較してより詳細な周波数ワーピング関数を記述できるものの認識性能で見る場合関数 **F2** を上回る認識性能を得ることはできなかった。フォルマントに基づくその他の周波数ワーピング関数の推定法 (関数 **F3**)(関数 **F4**) については話者正規化を行なわない場合と比較すると認識性能は向上するものの、ゆう度を用いた周波数ワーピング関数選択を上回る認識性能は得られなかった。さらに (関数 **F1**) については、話者正規化を行なわない場合より逆に認識性能が低下している。この原因について確認するため関数 **F1**, 関数 **F2**, 関数 **F3**, 関数 **F4** それぞれについて、フォルマントに基づき推定された周波数ワーピング係数の分布および母音のゆう度を用い選択された周波数ワーピング係数との対応関係について図 18-21に示した。図から明らかなように、関数 **F1** の場合、周波数ワーピング係数が広く分散しており、ゆう度を用い選択された周波数ワーピング係数との差も大きいことから、推定精度の点で問題があると考える。関数 **F2**, 関数 **F3**, 関数 **F4** については、ゆう度を用い選択された周波数ワーピング係数との間に幾らかの差はあるものの関数 **F1** の場合と比較すると推定された周波数ワーピング係数の分布も正規分布に近い形状をしている。

表 3: フォルマント周波数に基づく話者正規化モデルの認識性能 (音素認識率 (%))

	初期モデル	話者正規化モデル
話者正規化なし	80.60	-
ゆう度 (関数 2)	81.44	81.43
関数 F1	78.19	78.87
関数 F2	81.46	81.66
関数 F3	80.67	80.92
関数 F4	80.96	80.94
関数 F1-4	81.47	81.58

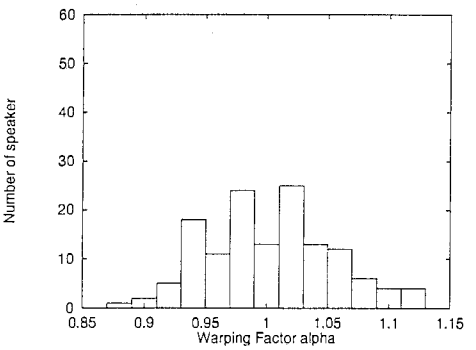


選択された周波数ワーピング係数の分布

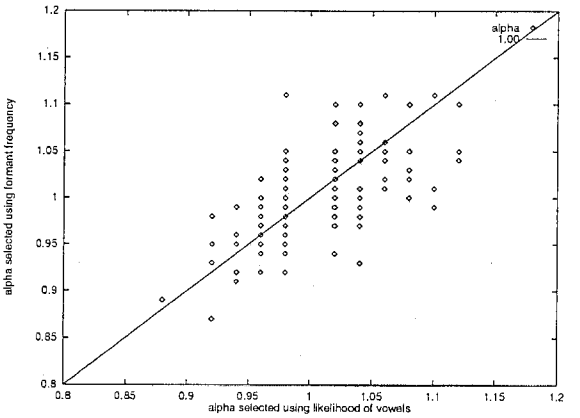


母音のゆう度により選択された周波数ワーピング係数との比較

図 18: フォルマント周波数 F1 を基に決定した周波数ワーピング係数

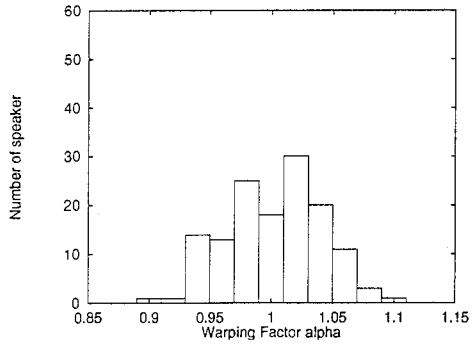


選択された周波数ワーピング係数の分布

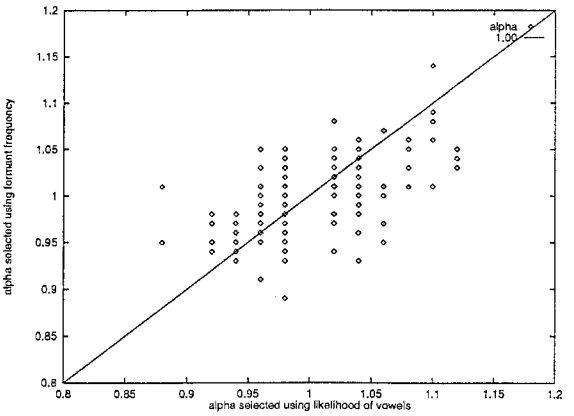


母音のゆう度により選択された周波数ワーピング係数との比較

図 19: フォルマント周波数 F2 を基に決定した周波数ワーピング係数

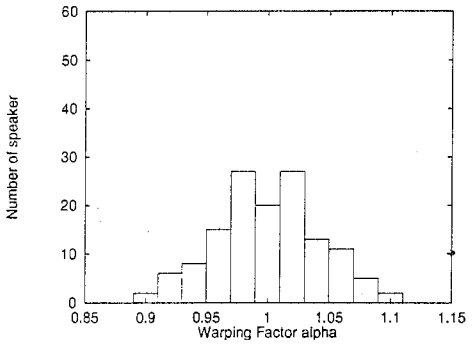


選択された周波数ワーピング係数の分布

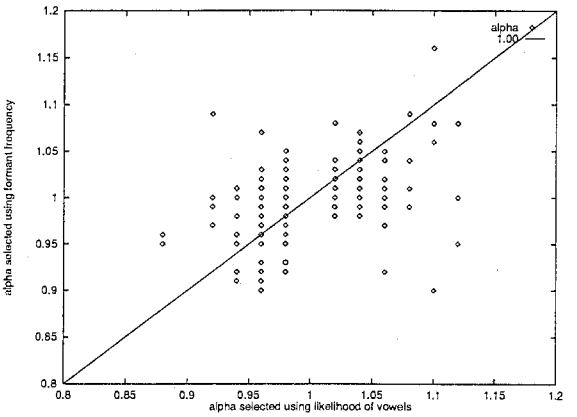


母音のゆう度により選択された周波数ワーピング係数との比較

図 20: フォルマント周波数 F3 を基に決定した周波数ワーピング係数



選択された周波数ワーピング係数の分布



母音のゆう度により選択された周波数ワーピング係数との比較

図 21: フォルマント周波数 F4 を基に決定した周波数ワーピング係数

3.5 認識時の周波数ワーピング関数選択文数による性能比較

認識時に必要となる、周波数ワーピング関数の推定に必要となる音声データ量について検討を行なうため、周波数ワーピング関数推定に用いる文数を 1,2,3,4,10,20 文とし、得られた認識性能について調査を行なった。実験は、(1) 全音素のゆう度 (関数 2) を用い周波数ワーピング関数を推定した場合、(2) 母音のゆう度 (関数 2) を用い周波数ワーピング関数を推定した場合、(3) フォルマント (関数 F2) を用い周波数ワーピング関数を推定した場合、(4) フォルマント (関数 F1-4) を用い周波数ワーピング関数を推定した場合について行なった。実験結果を表 6, 図 22に示す。この結果、フォルマント周波数を基に周波数ワーピング関数を推定する場合、推定用文数が増加しフォルマント周波数の推定精度が増すにつれ認識性能が向上し、フォルマント (関数 F2) の場合 10 文程度でフォルマント (関数 F1-4) の場合 4 文程度で認識率が安定している。その反面、ゆう度を用いた周波数ワーピング関数推定においては、推定に用いる文数により認識性能が安定していないことが分かる。これは特に全音素のゆう度を用い周波数ワーピング関数を推定行なった場合に顕著である。

表 4: 認識時の周波数ワーピング関数推定に用いる文数別の音素認識率 (%)

	1	2	3	6	10	20
全音素のゆう度 (関数 2)	81.43	81.40	81.41	81.36	81.51	81.37
母音のゆう度 (関数 2)	81.33	81.27	81.22	81.19	81.32	81.43
フォルマント (関数 F2)	81.28	81.34	81.47	81.53	81.62	81.66
フォルマント (関数 F1-4)	81.34	81.52	81.61	81.59	81.59	81.58

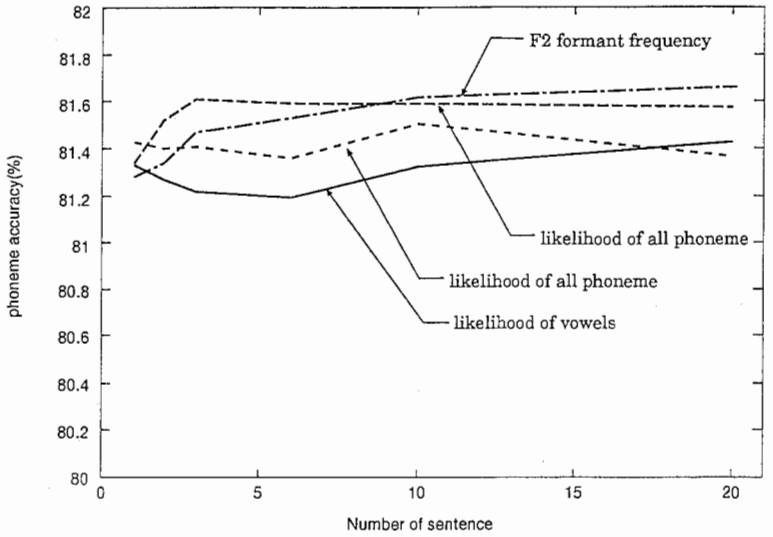


図 22: 認識時の周波数ワーピング関数選択文数別の音素認識率 (%)

4　まとめ

声道長に着目した周波数ワーピングによる話者正規化手法として提案されている周波数ワーピング関数の推定法、(1) 予め複数の周波数ワーピング関数を用意し、各周波数ワーピング関数を用いた音響分析結果に対する HMM のゆう度を算出し最もゆう度の高いワーピング関数を選択する方法、(2) 各話者の音声のフォルマント周波数を基に周波数ワーピング関数を推定する方法の 2 つの手法について認識実験を通じてその性能の評価を行なった。この結果、ゆう度に基づく周波数ワーピング関数の推定においては、周波数ワーピング関数関数 3 を用い母音のゆう度を用い選択した周波数ワーピング関数を基に、話者正規化学習を行なうことで音素認識率 81.84% が得られ、話者正規化を行なわない場合の誤認識の約 6.4% が削減された。またフォルマント周波数に基づく周波数ワーピング関数の推定においては、フォルマント周波数に F2 を用い推定した周波数ワーピング関数において音素認識率 81.66%) が得られた。また、認識時の認識時に必要となる、周波数ワーピング関数推定に必要となる音声データ量について検討を行ない、フォルマント周波数を基に周波数ワーピング関数を推定においては、推定用文数が増加しフォルマント周波数の推定精度が増すにつれ認識性能が向上し認識性能が安定するのに対し、ゆう度を用いた周波数ワーピング関数推定においては、推定に用いる文数が少ない場合認識性能が安定しない傾向が見られた。

5　謝辞

本研究を進めるにあたり、研究全般にわたり、親切丁寧に御指導頂いた匂坂 芳典 室長を始めとする ATR 音声翻訳通信研究所第一研究室の皆様に深く感謝致します．さらに実務訓練の機会を与えて下さった豊橋技術科学大学 知識情報工学系の阿部 英次 教授，及び ATR 音声翻訳通信研究所の山本 誠一 社長に心から感謝致します．

参考文献

[1] Micheiel Bacchiani: "Practical Vocal Tract Length Normalization for Automatic Speech Recognition", *ATR Technical Report*, TR-IT-0248 (1997-12).

[2] M. Ostendorf and H. Singer: "HMM topology design using maximum likelihood successive state splitting," *Computer Speech and Language*, 11, pp. 17–41, 1997.

[3] 深田 俊明, 柘植 覚, シンガー・ハラルド, 内藤 正樹: "連続音声認識用音響モデル (Version 2.0)", *ATR Technical Report*, TR-IT-0241 (1997-10).

付録 A　話者別認識率

以下に話者正規化モデルを用いた認識実験により得られた話者別の認識率を示す。ここで、認識に用いた音響モデルは、ゆう度を用いた周波数ワーピング関数による話者正規化においては、話者正規化学習を 2 回繰り返した後のモデルを、フォルマントに基づく周波数ワーピング関数による話者正規化においては、話者正規化学習を 1 回繰り返した後のモデルを使用している。

表 5: 話者別音素認識率 (ゆう度を用いた周波数ワーピング関数)

		関数 1		関数 2		関数 3	
speaker	GD	全音素	母音	全音素	母音	全音素	母音
M014	74.74	77.76	77.83	77.21	76.96	77.89	77.70
M109	79.93	78.54	80.30	80.33	80.24	80.15	80.21
M130	78.91	81.57	81.32	81.23	81.26	81.35	81.41
M303	78.20	80.92	80.86	80.86	80.80	80.64	80.43
M305	79.86	80.98	80.91	81.01	80.64	80.98	80.95
M311	79.96	81.14	81.23	79.87	81.07	79.31	79.25
M320	82.80	82.59	82.53	82.59	82.93	82.90	83.02
M410	84.87	84.75	84.96	84.87	84.87	84.01	84.16
M413	80.55	79.01	78.73	78.82	78.48	82.43	80.77
M414	86.17	86.72	86.85	86.94	86.82	86.85	86.88
total	80.60	81.40	81.55	81.37	81.41	81.65	81.48

表 6: 話者別音素認識率 (フォルマント周波数に基づく周波数ワーピング関数)

	F1	F2	F3	F4	F1-4
M014	72.67	77.76	76.74	77.33	77.55
M109	79.44	80.06	80.06	80.64	80.33
M130	80.18	80.98	81.04	80.80	80.98
M303	79.81	80.30	80.80	79.72	80.33
M305	77.76	79.74	75.29	76.93	79.56
M311	76.44	80.80	80.70	80.40	80.95
M320	81.29	82.59	82.65	82.06	82.87
M410	77.49	85.18	84.35	84.78	85.21
M413	77.18	82.40	81.63	80.18	81.54
M414	86.42	86.79	85.95	86.54	86.48
total	78.87	81.66	80.92	80.94	81.58