

論文 / 著書情報
Article / Book Information

論題(和文)	FMニューロンモデルに基づく音声認識のための距離尺度
Title(English)	
著者(和文)	相川 清明, 古井 貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	電子情報通信学会論文誌 A, Vol. J74-A, No. 6, pp. 822-828
Citation(English)	, Vol. J74-A, No. 6, pp. 822-828
発行日 / Pub. date	1991, 6
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 1991 Institute of Electronics, Information and Communication Engineers.

FM ニューロンモデルに基づく音声認識のための距離尺度

正 員 相川 清明[†] 正 員 古井 貞熙[†]

A Distance Measure for Speech Recognition Based on an FM-Neuron Model

Kiyoaki AIKAWA[†] and Sadaoki FURUI[†], *Members*

あらまし 音素知覚上重要とされているホルマント等のスペクトルのローカルピークの時間的な動きを利用した音声認識を行うために、聴覚生理学的に存在が確認されている FM ニューロンの機能モデルに基づいたスペクトルの動きを測る尺度を提案する。まず FM ニューロンを、スペクトルのローカルピークの周波数の推移のみに反応し、かつ移動の方向を識別する機能をもつ、スペクトル時系列に対する時間周波数フィルタとしてモデル化した。次に、FM ニューロンモデルの出力の違いを測る尺度をスペクトルのケプストラム展開に基づき導出した。これをスペクトル移動類似度と呼ぶことにする。更に、聴覚的な非線形周波数軸の上でのスペクトル移動類似度を周波数重み付けにより等価的に実現できることを示した。スペクトル移動類似度を用いた動的時間軸伸縮 (Dynamic Time Warping: DTW) による単語音声認識実験の結果、本尺度と従来のスペクトル距離尺度を併用することにより、従来のスペクトル距離尺度を単独で用いた場合に比べて認識誤りを改善できることがわかった。この改善はスペクトル距離としてケプストラム距離を用いた場合に著しく、認識誤りを 1/4 に減少できた。

1. ま え が き

生理学的知見によると、音声信号は鼓膜から基底膜に伝わった後 4 層の神経回路を経て大脳に伝わり知覚される⁽¹⁾。まず、蝸牛(かぎゅう)内の基底膜が振動して有毛細胞を刺激し、これが第 1 層の神経回路の入力となる。第 1 層と第 2 層の神経回路はほぼ帯域通過フィルタとみなすことができる。層を重ねるに従って神経細胞のもつ通過帯域は次第に狭くなるが、第 3 層になると、単なる帯域通過フィルタの機能をもつニューロンのほかに特殊な機能をもつニューロンが現れる。この中には音のエネルギーの上昇に反応する ON 型ニューロン、下降に反応する OFF 型ニューロン、周波数の一方向への推移に反応する FM ニューロンなどがある。機能的ニューロンの特徴抽出機構を利用すると、より精度の高い音声認識が行える可能性がある。ON 型ニューロンモデルを用いると音素間の境界位置を精度良く検出できることは既に報告した⁽²⁾。

一方、スペクトル距離に基づく音声認識において、スペクトル距離のほかにスペクトルの時間または周波数方向の変化を特徴として併用すると音声認識性能を向上できることが報告されてきた。スペクトルの変化

の特徴を用いた距離尺度としては、スペクトルの時間変化特徴を利用するデルタケプストラム法⁽³⁾、スペクトルの周波数変化を利用するスペクトルスロープ距離⁽⁴⁾、ケプストラムの次数ごとに分散正規化を行った重み付けケプストラム法⁽⁵⁾、これらを組み合わせた重み付けデルタケプストラム法⁽⁶⁾、時間周波数の 2 次元平面上でのスペクトルの最大傾斜方向のベクトルを利用するベクトル場法⁽⁷⁾などがある。スペクトル時系列中の変化成分のうち、ホルマントの動きは音素の知覚上重要な特徴であり、これを用いた音声認識法も報告されている^{(8),(9)}。従って、ホルマントの動きを精度良く抽出できれば良好な音声認識法を提供できる可能性がある。

機能的ニューロンのうち FM ニューロンは周波数の動きをとらえるものであるから、この機能をモデル化できればホルマントの動きを利用した音声認識が可能となると考えられる⁽¹⁰⁾。FM ニューロンの生理学的なモデルに関しては伊藤らの研究があるが音声認識には利用されていない⁽¹¹⁾。本論文ではまずホルマントを含めてスペクトルの動きを抽出する機能をもつ FM ニューロンをモデル化し、これに基づく特徴抽出法を提案する。次に、これを用いたスペクトルの動きを照合する尺度を導出する。

[†] NTT ヒューマンインタフェース研究所, 武蔵野市
NTT Human Interface Laboratories, Musashino-shi, 180 Japan

2. スペクトル移動の抽出

ホルマント移動は音声知覚上重要であるためホルマントの動きに基づいた音声認識法が提案されてきた^{(8),(9),(12)}。この種の方法では誤りを生じやすいホルマント位置決定が問題となっていた。ホルマント位置を決定することなく音声スペクトルの時間的な動きを記述することができれば、情報の欠落がないので精度の高い音声認識ができると考えられる。そこで、決定過程を経ずにスペクトル時系列からスペクトルの動きを抽出する方法を考察する。なお、スペクトルは周波数の関数、スペクトル時系列は周波数と時間の関数とする。

ここでは音の周波数の一方方向への推移に反応するFMニューロンをモデル化し、これに基づいたスペクトル移動の抽出法を導出する。現在生理学的に見つかっているFMニューロンは単一周波数に反応するものであるが、猫においては、動きに対してではないが、ホルマントのような帯域をもつ音に対して反応する聴覚ニューロンが見つかっている⁽¹³⁾。聴覚心理実験によると、ホルマントの動きの除去は子音の知覚に重大な影響を与えることがわかっており⁽¹²⁾、また、単一周波数の推移のみでなく帯域をもつ音の推移に対してもマスキング効果が観測されている⁽¹⁴⁾。これらの実験的事実に基づき、FMニューロンをホルマントのような帯域をもつスペクトルのローカルピーク（極大）の動きに反応するものとしてモデル化する。このFMニューロンモデルに基づくスペクトルの移動抽出機構の必要条件として

(1) 音声スペクトルの形状が時間的に不変ならば、エネルギー（音量）が時間的に変化しても反応は現れない、

(2) スペクトルのローカルピーク周波数の時間的な動きのみに反応する、

(3) スペクトルローカルピークの動きの方向を識別できる、

の3点を挙げ、これらの条件を満たす機構をスペクトル時系列に対する演算子として導出する。

音声の対数スペクトル包絡は山が鋭く谷が緩い。従ってスペクトルの周波数微分はスペクトルのローカルピーク付近で正から負への大きな値の変化を示す。もしスペクトルローカルピーク周波数が時間的に移動しているならば、スペクトルの周波数微分の時間変化はやはりローカルピーク付近において大きな値をとる。

更に、そのピークが高周波方向に動いているかあるいは低周波方向に動いているかにより極性が異なることがわかる。もし一定のスペクトルが時間的に継続していれば、エネルギーは時間と共に変化していても、スペクトルの周波数微分は時間的に不変となるのでその時間微分は0になる。またスペクトルに起伏がなければやはり0となる。すなわちスペクトルの動きは対数スペクトル包絡時系列 $S(\omega, t)$ (ω は周波数, t は時間) の時間周波数微分

$$F(\omega, t) = \frac{\partial^2}{\partial \omega \partial t} S(\omega, t) \quad (1)$$

により抽出できる。抽出された特徴量 $F(\omega, t)$ をスペクトル移動パターンと呼ぶことにする。図1にスペクトル移動抽出の原理図を示す。

時間と共に移動するスペクトルのモデルとして

$$S(\omega, t) = f(\omega - pt) \quad (2)$$

を考える。 p はスペクトル移動速度 (rad/sec) を表す。式(2)に対するスペクトル移動パターンは

$$F(\omega, t) = -pf''(\omega - pt) \quad (3)$$

となる。関数の2階微分はスペクトルローカルピークの位置においては負となるので、式(3)はスペクトルローカルピークの移動方向すなわち p の符号により正負が決まることがわかる。また、スペクトル移動パターンには一つのスペクトルピークの動きに対して一つのピークが現れ、ピークの両側にはサイドローブが現れる。しかし、音声スペクトルのローカルピークにおける2階微分は通常他の部分に比べて大きいので、サイドローブの大きさはピークに比べて小さい。

一例として単一のスペクトルピークが移動する場合を考えてみる。スペクトルの形状をガウス分布とし、分布の分散を v 、 $t=0$ におけるピーク周波数を ω_0 とするとスペクトルは

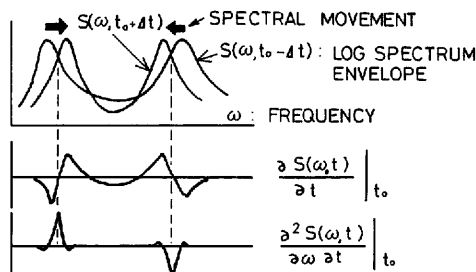


図1 スペクトル移動抽出の原理
Fig. 1 The mechanism of spectral movement extraction.

$$S(\omega, t) = \frac{1}{\sqrt{2\pi v}} e^{-(\omega - \omega_0 - p t)^2 / 2v} \quad (4)$$

により表される。スペクトル移動パターン $F(\omega, t)$ は

$$F(\omega, t) = \frac{p}{\sqrt{2\pi v^3}} \left\{ 1 - \frac{1}{v} x^2 \right\} e^{-(x^2 / 2v)} \quad (5)$$

$$x = \omega - \omega_0 - p t$$

となり、スペクトルのピーク $x=0$ で最大となり、その値は移動速度 p に比例し、スペクトルピークの幅に対応する標準偏差の 3 乗に逆比例することがわかる。

スペクトル移動パターンと整流器 $R[x]$ を組み合わせれば周波数の上昇または下降に反応する 2 種類の FM ニューロンモデル $F^+(\omega, t)$, $F^-(\omega, t)$ が得られる。

$$\begin{aligned} F^+(\omega, t) &= R[F(\omega, t)] \\ F^-(\omega, t) &= R[-F(\omega, t)] \end{aligned} \quad (6)$$

$$R[x] = x \quad x \geq 0$$

$$R[x] = 0 \quad x < 0$$

次にスペクトル移動パターンの性質について述べる。観測される音声パワースペクトルにはさまざまな要因が含まれている。音声パワースペクトルを $P(\omega, t)$ 、伝送路の伝達関数を $H(\omega)$ 、ゲインの変動を $G(t)$ とするとこれらの影響を受けた音声スペクトルに対するスペクトル移動パターンは

$$\begin{aligned} F(\omega, t) &= \frac{\partial^2}{\partial \omega \partial t} S(\omega, t) \\ &= \frac{\partial^2}{\partial \omega \partial t} \ln(P(\omega, t)H(\omega)G(t)) \\ &= \frac{\partial^2}{\partial \omega \partial t} \ln(P(\omega, t)) \end{aligned} \quad (7)$$

となり、スペクトル移動パターンはマイクロホンや伝送路の周波数特性の影響を受けにくいことがわかる。また、長時間平均スペクトルに現れるような時間的に変動しない個人性や発声強度の変動は同様の理由により除去できる。従って、スペクトル移動パターンを特

徴量として用いるとこれらの要因を受けにくい認識が可能である。なお、伝送路特性や平均スペクトルはスペクトル時系列の時間微分によっても取り除ける。

次に、ホルマントのエネルギーがスペクトルの谷の部分に比べて時間的に増大する場合を、基準となる時刻を t_0 として

$$S(\omega, t) = S(\omega, t_0)(\alpha t + \beta) \quad (8)$$

$$\alpha > 0, \beta \geq 1$$

とモデル化すると、スペクトル移動パターンは

$$F(\omega, t) = \alpha \frac{dS(\omega, t_0)}{d\omega} \quad (9)$$

となる。 $S(\omega, t_0)$ がガウス分布の場合には ω_0 に対して点対称な関数、

$$F(\omega, t) = -\alpha \frac{(\omega - \omega_0)}{\sqrt{2\pi v^3}} e^{-((\omega - \omega_0)^2 / 2v)} \quad (10)$$

となる。式(10)は ω_0 を境に高周波側では負、低周波側では正となり、スペクトルピークに向かってスペクトルが移動していることを意味しており、スペクトルピークのエネルギーが他の部分に比べて増大していくことをよく説明している。図2は「あいうえお」と発声した場合のスペクトル移動パターンを示す。

ここで、他のスペクトル距離に関連した特徴量とスペクトル移動パターンの関係について述べる。対数スペクトル時系列の時間微分パターンの、各時点における周波数に関する逆フーリエ展開係数がデルタケプストラムに相当する。また対数スペクトル時系列の周波数微分パターン間の距離はスペクトルスロープ距離と呼ばれるものになるが⁽⁴⁾、これは重み付けケプストラムと類似した関数となる⁽⁵⁾。スペクトルスロープ距離では周波数微分スペクトルの全周波数帯域での周波数積分が常に 0 になるため、パワーを正規化したことに相当する。また、重み付けケプストラムでは各ケプストラム係数の変動の分散を正規化する効果がある。ま

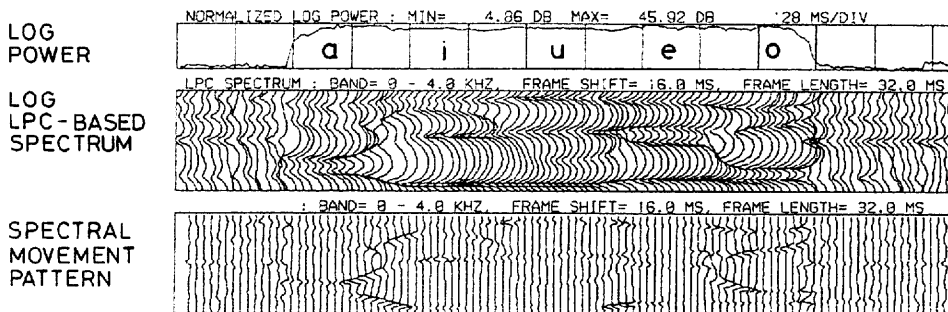


図2 「あいうえお」に対するスペクトル移動パターン
Fig. 2 The spectral movement pattern for the speech "aiueo".

た、スペクトル時系列の時間・周波数変化を特徴量とするもう一つの方法としてベクトル場法があるが⁽⁷⁾、スペクトル移動パターンはスペクトルが時間的に動いたときのみ反応するという点でベクトル場と異なり、ホルマント移動を重視した認識が可能である。

3. スペクトル移動類似度

二つのスペクトル移動パターンの違いを測るための距離尺度を導入する。対数スペクトル包絡 $S(\omega, t)$ は、線形予測分析 (Linear Predictive Coding: LPC) により能率良く高精度で求めることができ、この分析により求まる k 次の LPC ケプストラム係数 $c_k(t)$ を用いて

$$S(\omega, t) = \sum_{k=-K}^K c_k(t) \cos k\omega \quad (11)$$

のように展開できる。 K は平滑化のための打ち切り次数である。従ってスペクトル移動パターンは

$$F(\omega, t) = - \sum_{k=-K}^K \frac{dc_k(t)}{dt} k \sin k\omega \quad (12)$$

により表すことができる。しかし、式(12)のような単なる微分では $S(\omega, t)$ に含まれるスペクトル推定誤差などの不要な揺らぎを強調し過ぎてしまう。そこで、ケプストラム係数の時間微分の代わりに、時間方向ウィンドウ幅 W フレームに入るフレームのケプストラム係数の1次回帰直線の傾き (回帰係数) を用いることにした。時点 t における k 次ケプストラムの回帰係数は

$$c'_k(t) = \frac{\sum_{i=-W/2}^{W/2} c_k(t + \Delta i) i Q(i)}{\sum_{i=-W/2}^{W/2} i^2 Q(i)} \quad (13)$$

により求められる。 $Q(i)$ は三角形の窓関数で、

$$Q(i) = 1 - 2|i|/W \quad (-W/2 \leq i \leq W/2) \quad (14)$$

である。 Δ はフレーム周期を表す。1次回帰係数は平均的な傾きを表し、平滑化された微分係数と言える。

また、 k 次のケプストラム係数にかかる k は周波数微分を行うリフトに相当するが、高次成分を抑えるために単なる周波数微分でなく、周波数軸方向の平滑化のためのリフトを乗じた、

$$\begin{aligned} w(k) &= -L & (k < -L) \\ &= k & (-L \leq k \leq L) \\ &= L & (L < k) \end{aligned} \quad (15)$$

を用いることにした。 L を飽和ケブレンシと呼ぶことにする。従って、スペクトル移動パターン $F(\omega, t)$ は式(12)を改めて、

$$F(\omega, t) = - \sum_{k=-K}^K c'_k(t) w(k) \sin k\omega \quad (16)$$

のように表される。以上によりスペクトル移動パターンの直交関数展開が得られた。

次に二つのスペクトル移動パターンの違いを測る尺度を導入する。従来用いられているケプストラム距離は二つの対数スペクトル $S_1(\omega, t)$, $S_2(\omega, t)$ の2乗距離として以下のように定義される。

$$D_{\text{CEP}}(t) = \frac{1}{\pi} \int_0^\pi (S_1(\omega, t) - S_2(\omega, t))^2 d\omega \quad (17)$$

同様に二つのスペクトル移動パターン $F_1(\omega, t)$, $F_2(\omega, t)$ の間の距離を

$$D_{\text{SM}}(t) = \frac{1}{\pi} \int_0^\pi (F_1(\omega, t) - F_2(\omega, t))^2 d\omega \quad (18)$$

により定義することができる。これをスペクトル移動類似度と呼ぶことにする。式(16)を式(18)に代入すると式(18)は

$$D_{\text{SM}}(t) = \frac{1}{\pi} \int_0^\pi \left\{ \sum_{k=-K}^K b_k(t) \sin k\omega \right\}^2 d\omega \quad (19)$$

$$b_k(t) = (c'_{1k}(t) - c'_{2k}(t)) w(k) \quad (20)$$

となり、式(19)の右辺を計算すると

$$D_{\text{SM}}(t) = \sum_{k=-K}^K b_k(t)^2 \quad (21)$$

となる。式(20)において、

$$c'_k(t) = c'_{-k}(t)$$

$$\therefore b_k(t) = -b_{-k}(t) \quad (22)$$

また、

$$b_0(t) = 0 \quad (23)$$

であるから、式(21)は

$$D_{\text{SM}}(t) = 2 \sum_{k=1}^K b_k(t)^2 \quad (24)$$

とも書ける。

なお、実際の分散値を代入した重み付けデルタケプストラム距離はスペクトル移動類似度の計算式に近くなることがあると考えられるが⁽⁷⁾、本尺度は、スペクトル移動抽出機構に基づき導出している点が異なる。

4. 周波数重み付け

式(21)は線形の周波数軸上においてスペクトル移動類似度を計算しているが、聴覚系の周波数軸はメルスケールのような低域が拡大された非線形周波数軸であるため、これに近い周波数軸を用いる方が望ましい。低域が拡大された聴覚的な周波数軸においてスペクトル移動類似度を計算するという事は低域に重みを掛けてスペクトル移動パターン差の2乗を周波数積分することに相当する。従って、聴覚的周波数軸におけるスペクトル移動類似度は周波数重み付け関数を $g(\omega)$

とすると

$$D_{FWSM}(t) = \frac{1}{\pi} \int_0^\pi (F_1(\omega, t) - F_2(\omega, t))^2 g(\omega) d\omega \quad (25)$$

により等価的に求めることができる⁽⁵⁾。スペクトル移動パターンは正弦展開により表現されているので $g(\omega)$ を余弦展開して、

$$g(\omega) = 1 + \sum_{m=1}^M a_m \cos m\omega \quad (26)$$

により表せば、周波数重み付けスペクトル移動類似度 D_{FWSM} は

$$\begin{aligned} D_{FWSM}(t) &= \frac{1}{\pi} \int_0^\pi \left\{ \sum_{k=-K}^K b_k(t) \sin k\omega \right\}^2 g(\omega) d\omega \quad (27) \\ &= D_{SM}(t) + \frac{1}{2} \sum_{m=1}^M \left\{ a_m \sum_{k=-K}^K \{ b_k(t)(b_{k+m}(t) \right. \\ &\quad \left. + b_{k-m}(t)) \} \right\} \quad (28) \end{aligned}$$

となる。周波数重み付け関数が1次の展開のみで表される $M=1$ の場合には、

$$D_{FWSM}(t) = D_{SM}(t) + a_1 \sum_{k=1}^K \{ b_k(t)(b_{k-1}(t) + b_{k+1}(t)) \} \quad (29)$$

により求められる。

次に重み付け関数とそれによる座標軸の関係について述べる。 $g(\omega)$ はスペクトルの差の2乗に対する重み付け関数であり、スペクトルに対する重みは $\sqrt{g(\omega)}$ であると考えられる。被積分関数に掛ける重みの大小は座標軸の部分における伸縮に対応するので、重み関数の積分が周波数軸を表している。従って、周波数軸の写像関数 $h(\lambda)$ は以下の式により表される。

$$h(\lambda) = \pi \int_0^\lambda \sqrt{g(\omega)} d\omega / \int_0^\pi \sqrt{g(\omega)} d\omega \quad (30)$$

式(26)の重み付け関数の a_1 が1のとき、すなわち、

$$g(\omega) = 1 + \cos \omega \quad (31)$$

のときには

$$\int_0^\lambda \sqrt{g(\omega)} d\omega = \sqrt{2} \sin(\lambda/2) \quad (32)$$

となるので、

$$h(\lambda) = \pi \sin(\lambda/2) \quad (33)$$

となる。

$$g(\omega) = (1 + 0.37 \cos \omega)^2 \quad (34)$$

を用いると

$$h(\lambda) = \lambda + 0.37 \sin \lambda$$

となり、メルスケールとほぼ同じ周波数軸を実現できる。式(33)と式(35)の周波数写像関数の違いを図3に示す。なお、式(33)は後に実験の中で述べるが、周波数重み付けスペクトル移動類似度を用いて音声認識実

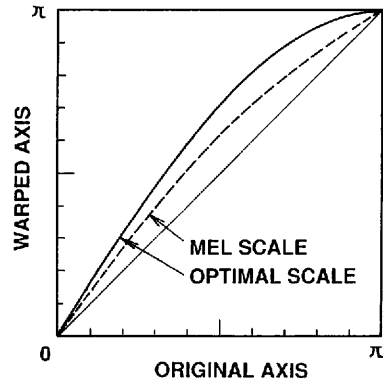


図3 最適周波数軸写像関数とメル周波数軸写像関数の比較

Fig. 3 The Comparison of the optimal frequency axis mapping function and the Mel mapping function.

験を行ったとき最も良い結果が得られる最適な周波数写像関数となっている。

5. スペクトル移動を用いた音声認識のための距離

音素の認識には過渡的な特徴だけでなくスペクトル形状の特徴も必要である。従って、Dynamic Time Warping (DTW) による単語あるいは文章の認識を行う場合には、両特徴を統合した尺度が必要であると考えられる。そこで、従来の線形予測分析に基づくスペクトル距離 D_{SPEC} とスペクトル移動類似度 D_{FWSM} を併用した距離

$$D(t) = qD_{FWSM}(t) + (1-q)D_{SPEC}(t) \quad (36)$$

を用いることにした。 q は両要因のバランスを調整するための結合係数であり、認識実験により最適値を求める。 D_{SPEC} としては D_{CEP} (ケプストラム距離)、 D_{WLR} (Weighted Likelihood Ratio (WLR) 距離)、 D_{PWLR} (Power 込み WLR 距離)、 D_{DCEP} (デルタケプストラム距離)、 D_{WCCEP} (重み付けケプストラム距離) などが考えられる。

デルタケプストラムはスペクトル形状の情報をもたないので D_{DCEP} としては一般的に行われているように D_{CEP} と重み付き加算をしたものを用いる。最適な重みは予備実験により決めた。 D_{WCCEP} では分散の代わりにケプストラム次数の2乗を重みとした。以下に各スペクトル距離の計算法をまとめておく。これらの式において、 r_k は k 次の LPC 自己相関係数、 u はパワーを表す。LPC 分析次数より高い次数の LPC 自己相関係数は低次の自己相関係数と線形予測係数から外挿により

求める。

$$\begin{aligned}
 D_{DCEP}(t) &= 0.05 D_{CEP}(t) \\
 &\quad + 0.95 \frac{1}{\pi} \int_0^\pi \left(\frac{\partial S_1(\omega, t)}{\partial t} - \frac{\partial S_2(\omega, t)}{\partial t} \right)^2 d\omega \\
 &= 0.05 D_{CEP}(t) + 0.95 \frac{1}{\pi} \int_0^\pi \left\{ 2 \sum_{k=1}^K (c_{1k}(t) \right. \\
 &\quad \left. - c_{2k}(t)) \cos k\omega \right\}^2 d\omega \\
 &= 0.05 D_{CEP}(t) + 1.9 \sum_{k=1}^K (c_{1k}(t) - c_{2k}(t))^2
 \end{aligned} \quad (37)$$

$$\begin{aligned}
 D_{WCEP}(t) &= \frac{1}{\pi} \int_0^\pi \left(\frac{\partial S_1(\omega, t)}{\partial \omega} - \frac{\partial S_2(\omega, t)}{\partial \omega} \right)^2 d\omega \\
 &= \frac{1}{\pi} \int_0^\pi \left\{ 2 \sum_{k=1}^K (c_{1k}(t) \right. \\
 &\quad \left. - c_{2k}(t)) w(k) \sin k\omega \right\}^2 d\omega \\
 &= 2 \sum_{k=1}^K \{(c_{1k}(t) - c_{2k}(t)) w(k)\}^2
 \end{aligned} \quad (38)$$

$$D_{WLR}(t) = 2 \sum_{k=1}^K ((r_{1k}(t) - r_{2k}(t))(c_{1k}(t) - c_{2k}(t))) \quad (39)$$

$$\begin{aligned}
 D_{PWLR}(t) &= D_{WLR}(t) + 0.01 (u_1(t)/u_2(t) \\
 &\quad + u_2(t)/u_1(t) - 2)
 \end{aligned} \quad (40)$$

$$u(t) = \frac{1}{\pi} \int_0^\pi P(\omega, t) d\omega \quad (41)$$

$D_{WCEP}(t)$ では式 (15) のリフト $w(k)$ を用いる。

6. 実 験

男性 1 名の音声データを標準音声、他の男性 10 名を試験音声とみなして DTW を用いた 50 都市名認識実験により本方法の評価を行った。標準音声の話者は他の 10 名の話者の音声データの分布の中心近くに位置している。サンプリング周波数は 8 kHz、フレーム長は 32 ms、フレーム周期は 16 ms とした。スペクトルは 10 次の LPC 分析により求め、ケプストラム 16 次で表現している。式 (36) における D_{SPEC} として D_{CEP} を用いた予備実験により時間平滑化ウィンドウ幅 W は 5 フレームが最適であるという結果を得た。 $W=5$ に固定し、式 (15) の飽和ケフレンシ L をパラメータとして結合係数 q による認識誤り率の変化を調べた結果を図 4 に示す。これにより L が 3 ないし 4 の場合に最も良い結果が得られることがわかる。 $W=5$ 、 $L=3$ に固定し D_{SPEC} として各種の距離尺度を用いた実験結果を表 1 に示す。但し結合係数 q が最適な場合のみ示した。 D_{WCEP} における飽和ケフレンシは予備実験により 3 とした。また D_{DCEP} における回帰係数を求めるウィンド

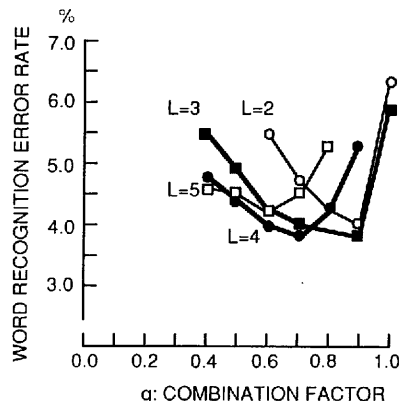


図 4 スペクトル距離とスペクトル移動類似度の結合係数と単語認識誤り率の関係

Fig. 4 Relationship between distance combination factor and word recognition error rate.

表 1 スペクトル距離にスペクトル移動類似度を併用することによる単語認識誤りの改善

スペクトル距離	単 独	スペクトル移動類似度 と併用	
		周波数重み付け	
D_{CEP}	9.5 %	3.8	2.2
D_{DCEP}	5.0	4.2	2.5
D_{WCEP}	5.9	3.5	2.3
D_{WLR}	3.2	2.1	1.6
D_{PWLR}	2.1	1.6	1.5

ウ幅はスペクトル移動抽出の場合と同様 5 フレームとした。

表 1 によるとスペクトル移動類似度 D_{SM} といずれのスペクトル距離を組み合わせても認識誤りを減少させることができる。周波数重み付けにより低域を等価的に引き伸ばした座標軸を用いると更に認識率を改善できる。周波数重み付けは各スペクトル距離に対しても可能であるがここではスペクトル距離としては従来より標準的に用いられている周波数重み付けされていないものを用い、スペクトル移動類似度に対してのみ周波数重み付けを行った。 D_{SPEC} として D_{CEP} を用いた場合には D_{SM} を併用することにより単語認識誤り率を 9.5% から 3.8% に減少でき、更に周波数重み付けをした D_{FWSM} を併用することにより 2.2% にまで減少させることができる。 D_{FWSM} における低域重み付け係数 a_1 は 1 が最適であった。この結果は図 3 に示すようにメルスケールよりも低域を重視した方がよいという結果となっている。この理由としては低域のスペクトルバ

ターンの方が声道長などの個人差の影響を受けにくい
ためであると考えられる。

7. む す び

音素の認識にとって重要なスペクトルピークの動きを抽出する FM 聴覚ニューロンをモデル化し、これに基づくスペクトル移動の抽出法を提案した。本方法によりスペクトルピーク周波数の動きを伝送路特性や定常的な個人性、ゲイン変動の影響を受けずに抽出できる。次に、FM ニューロンモデルにより抽出したスペクトル移動パターンを比較するためのスペクトル移動類似度を導入した。本尺度はホルマント位置を決定することなくホルマント移動を直接比較できる特徴をもっている。本類似度を併用することにより従来のスペクトル距離を単独で用いた場合に比べて音声認識誤りを減少できることがわかった。この改善は特にスペクトル距離としてケプストラム距離を用いた場合に著しかった。更に聴覚的な非線形周波数軸上でのスペクトル移動類似度を周波数重み付けにより等価的に求める方法を示した。低域を引き伸ばした聴覚的な周波数軸を用いることにより更に認識誤りを減少できる。スペクトル距離としてケプストラム距離を用いた場合、周波数重み付けスペクトル移動類似度を併用することにより単語認識誤り率を 1/4 に減少できた。

今後は、FM ニューロンモデルの精密化について考察を進める必要がある。

謝辞 討論して頂いた嵯峨山茂樹主幹研究員（現在 ATR 出向中）をはじめとする音声部の諸氏に感謝する。

文 献

- (1) 大沢文夫、大井龍夫、岡崎令治、石浜 明、朝倉 昌、江口吾朗、塚原伸晃、森 主一、巖佐耕三編集：“神経回路と生体制御”，生物科学講座 6，朝倉書店（1979-04）。
- (2) 相川清明：“ON 型聴覚ニューロンモデルに基づく音素セグメンテーションパラメータ”，信学論(A)，J71-A，3，pp. 592-600（1988-03）。
- (3) Furui S.: "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Trans., Acoust Speech & Signal Process., ASSP-34, 1, pp. 52-59 (1986-02)。
- (4) Klatt D.H.: "Prediction of Perceived Phonetic Distance from Critical-band Spectra: A First Step" Proc. ICASSP82, pp. 1278-1281 (May 1982)。
- (5) Tohkura Y.: "A Weighted Cepstral Distance Measure for Speech Recognition", IEEE Trans. Acoust Speech & Signal Process., ASSP-35, 10, pp. 1414-1422 (1987-10)。

- (6) Soong F.K. and Rosenberg A.E.: "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", Proc. ICASSP86, pp. 877-880 (1986-04)。
- (7) 岡 隆一：“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(D)，J69-D，11，pp. 1704-1713 (1986-11)。
- (8) 辻野克彦、櫻井 徹、野村康雄、千種俊輔、溝口理一郎、角所 収：“知識ベース構築支援環境を備えた連続音声認識エキスパートシステム：SPREX II”，信学論(D)，J71-D，3，pp. 531-542 (1988-03)。
- (9) Cole R.A.: "Perception and Production of Fluent Speech", Lawrence Erlbaum Associates Publishers (1980)。
- (10) Aikawa K. and Furui S.: "Spectral Movement Function and Its Application to Speech Recognition", ICASSP88, S5. 11, pp. 223-236 (1988-04)。
- (11) 伊藤崇之、福島邦彦：“聴覚系の特徴抽出機構の神経回路モデル”，信学論(D)，J70-D，2，pp. 451-462 (1987-02)。
- (12) Edelman G.M., et al.: "Auditory Function", Chapter 23, John Wiley & sons (1988)。
- (13) 工藤雅治：“ネコ聴覚領のホルマント識別ニューロンについて”，新潟医学会雑誌，96，5，pp. 311-326 (1982-05)。
- (14) 平原達也：“フォルマント移動音に対する聴覚系の応答”，日本音響学会講演論文集，1，pp. 249-250 (1985-09)。
- (15) 相川清明：“周波数重み付けスペクトル移動マッチング尺度”，音響学会講演論文集，pp. 297-298 (1988-10)。

(平成 2 年 8 月 20 日受付，3 年 1 月 23 日再受付)

相川 清明



昭 50 東大・工・電子卒，昭 55 同大学院博士課程了，同年 NTT 基礎研究所入社。以来，連続音声認識，不特定話者音声認識の研究に従事。現在，HI 研究所音声情報研究部所属。音素認識，聴覚モデル，ニューラルネットに興味をもつ。日本音響学会，日本 ME 学会，IEEE 各会員，工博。

古井 貞熙



昭 43 東大・工・計数卒，同 45 同大学院修士課程了，同年 NTT 電気通信研究所入社。以来，音声認識，話者認識等の研究に従事。現在同社 HI 研究所音声情報研究部長，工博。本会米沢賞，論文賞，著述賞，日本音響学会論文賞，IEEE 論文賞，科学技術庁長官賞受賞，著書「ディジタル音声処理」など。