音声言語情報処理 44－21
(2002. 12. 19)

ウェーブレット変換を用いた音素マッチング処理

吉井 圭吾† 木場 俊暁‡ 金丸 隆‡ 関根 優年‡

†東京農工大学
〒184-8588 東京都小金井市中町 2-24-16
E-mail: †mut@sekine-lab.ei.tuat.ac.jp, ‡sekinem@cc.tuat.ac.jp,
‡{hikaru, t-koba}@sekine-lab.ei.tuat.ac.jp

あらまし 音声や楽音のような信号は局所的に変動し、その周波数が時間とともに変化する非定常信号と思われる。このような信号を解析する方法としては、短時間フーリエ変換、ウェーブレット変換などがあげられる。今回、Haar関数を用いた離散Wavelet変換を使用した。画像認識では、ウェーブレット変換の多重解像度によるテンプレートマッチングの研究が行われその有効性が示されている。これを音声に適応する。本実験では、最も簡単なテンプレートマッチングを使って実験を行った。

キーワード 多重解像度解析, ウェーブレット変換, テンプレートマッチング

Phoneme matching processing using wavelet transform

Keigo YOSHII†, Toshiaki KOBA‡, Takashi Knamaru‡ and Masatoshi SEKINE‡

†Tokyo University of Agriculture and Technology
Nakamachi 2-24-16, Koganei-shi, Tokyo, 184-8588 Japan
E-mail: †k-yoshii@sekine-lab.ei.tuat.ac.jp, ‡sekinem@cc.tuat.ac.jp,
‡{t-koba, kanamaru}@sekine-lab.ei.tuat.ac.jp

Abstract The signal such as audio or musical sound changes locally, and it is considered as the unsteady signal with its frequency varied along the time. To analyze these signals, Short-time Fourier Transform and Wavelet Transform are usually used and we selected The Discrete Haar-Wavelet Transform . On the image recognition, the template matching is widely researched by multiplex resolution of wavelet transform and it revealed so effective. In the present study, we apply this to the audio and make examinations using simplest template matching.

Key words Multi resolution analysis, Wavelet transform, Template matching

1. はじめに

一般的な音声認識のための特徴抽出の方法では、まず入力された音声から数十ms程度の時間長の信号区間を切り出す。切り出された信号をフーリエ変換によるスペクトル解析でケプストラム係数（ＭＦＣＣパラメータ）を求めこれを特徴パラメータとしている。この性能をみるため音声認識ツールHTKによって孤立音声認識システムを構築したが認識率は低いものであった。

その理由の一つとして、音声や楽音のような信号は局所的に変動し、その周波数が時間とともに変化する非定常信号であると思われるからである。このような信号を解析する方法としては,短時間フーリエ変換、ウェーブレット変換などがあげられる。短時間フーリエ変換が窓内では時間-周波数分解能が一定となる解析手法であるのに対して,ウェーブレット変換は時間分解能は高周波では高く低周波では低い。また、周波数分解能は高周波では低く低周波では低い解析法である。本実験では局所的な周波数変化の解析を行うためにウェーブレット変換を用いた解析を行う。

今回の実験では音素二つからなる単語の認識について考えている。全体波形の中から音素と思われる場所に注目して、その場所を分析を行う観測窓（フレーム）の長さ(フレーム長)として、フレーム長によって切り出された信号に対して特徴抽出を行う。次にウェーブレット変換で多重に得たscaling係数とwavelet係数を用いたテンプレートマッチングにより音声認識を試みる。

2. 音声処理手法

2-1 Haarのウェーブレット変換

本研究においては,Haar関数を用いた離散Wavelet変換を利用する。

$$\phi_H(x) = \begin{cases} 1, & 0 \leqq x < b \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

$$\psi_H(x) = \begin{cases} 1, & 0 \leqq x < b/2 \\ -1, & 1/2 \leqq x < b \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Haarのスケーリング関数$\phi_H(x)$とHaarのウェーブレット関数$\psi_H(x)$は式（1）（2）のように定義される。

スケーリング関数$\phi_H(x)$と、ウェーブレット関数$\psi_H(x)$により入力音声（Original audio signal）の変換を行うことで1/2に圧縮されたScaling signalとWavelet signalが得られる(図1)。

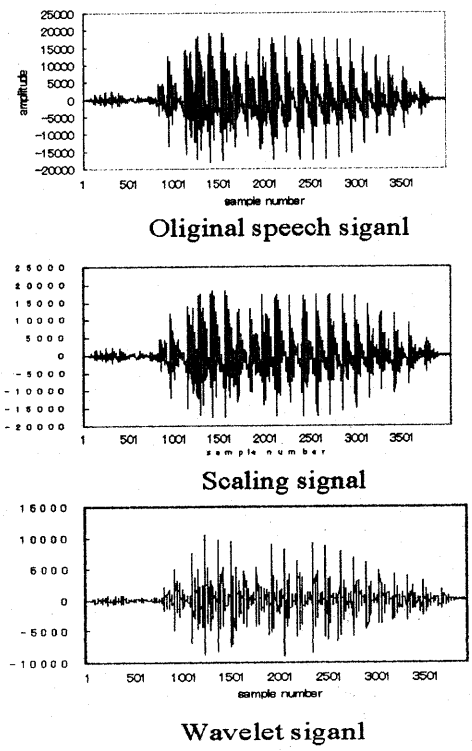


図1 Haar-Wavelet変換による音声波形イメージ

2-2 多重解像度

一回のウェーブレット変換を行うことを1レベル変換という。原音声に、1レベル変換を行うとレベル1のScaling係数とWavelet係数が得られる。データ量は共に1/2となる。レベル1のScalingデータをさらに1レベル変換することで、レベル2のScaling係数とWavelet係数が得られる。データ量は共に1/4となる。このように、ウェーブレット変換してできた低周波成分をさらにウェーブレット変換するという作業を繰り返すことで、様々なレベルのデータを得ることができる（図2)。これを多重解像度表現という。

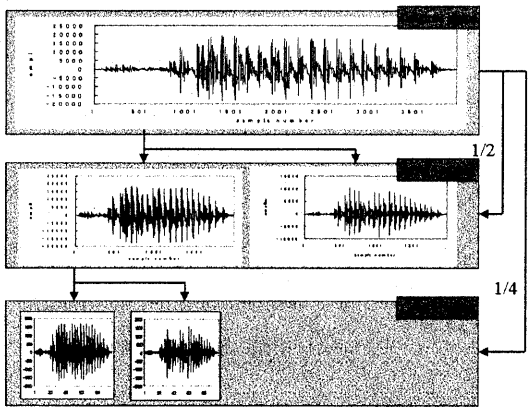


図２　多重解像度

2-3 ピッチ周期

有声音や無声音、または音素境界を知るための方法としてはピッチ周波数を求めることが有効である。

このような音声に含まれる周期性は式（３）で与えられるような自己相関関数を求めることで調べることが出来る。

$$\phi(m)=\frac{1}{N}\sum_{n=0}^{N-1-m} Xn\ Xn+m \qquad (3)$$

式(3)によって得られるΦ(m)はmの増大にともなってその値は徐々に小さくなる。分析対象の音声が有声音であれば音声信号のピッチ周期の整数倍にピークを生じるので、その第一番目のmをピッチ周期とする。Φ(m)にピークが現れなければその分析音は無声音である。

ここで、ピッチ周波数を用いたフレーム位置の決定を考える。例えば、青/ao/という音声波形があるとすると「a」と「o」という音素の境界ではピッチ周波数が変化するはずである。この点に注目してピッチ周波数の値が大きく変動した時刻を「a」と「o」の音素境界時刻とする。また、これにより有声音の始点と終点の時刻も得る。

下の図３(a)は実際の「あお(青)/ao/」の音声波形である。図３(b)がフレーム長を10ｍｓ、観測窓を30ｍｓ、最低ピッチを80Hz最高ピッチを200Hzとした時のピッチ周波数をプロットした図である。

120ｍｓ、480ｍｓ、360ｍｓの時にピッチ周波数が大きく変動しているのでここが有声音の始点、音素境界時刻、有声音の終点の３点と推定する。

これをもとにして決めた各音素のフレーム長が図３(c)のようになる。

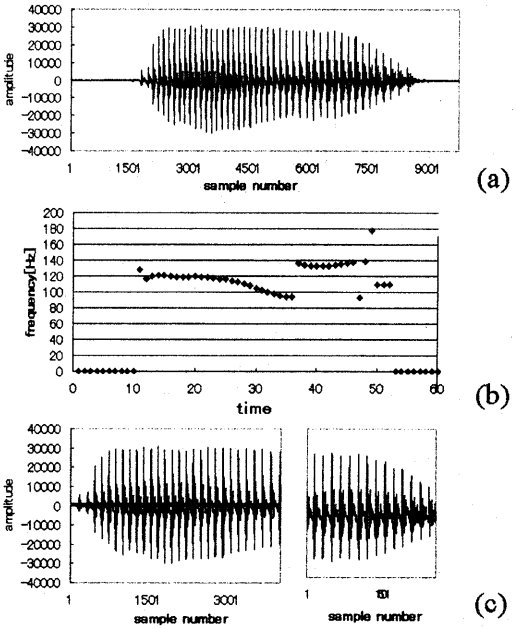


図３　フレーム長の決定方法
(a)「あお(青)」/ao/の音声波形
(b) (a)のピッチ周波数プロット
(c) (b)を元に(a)の音声波形を/a/と/o/と推測されるフレーム長に分離

正確なピッチ周波数を求める研究は種々行われているが、ここでは大まかな音素境界時刻を決めルことを目的としている。ここから、それぞれのフレームにウェーブレット変換を行い詳しい分析をしていく。

2-4 ホルマント

音声波形、特に母音では特定の周波数領域にエネルギーが集中して観測される。その領域の中央値をホルマント周波数とよんでいる。母音はホルマント周波数で特徴付けることが出来る。図４は第１ホルマント周波数と第２ホルマント周波数平面上に母音を配置した母音知覚例である。ホルマント周波数は性別、年齢、話者によってかなり変動する。

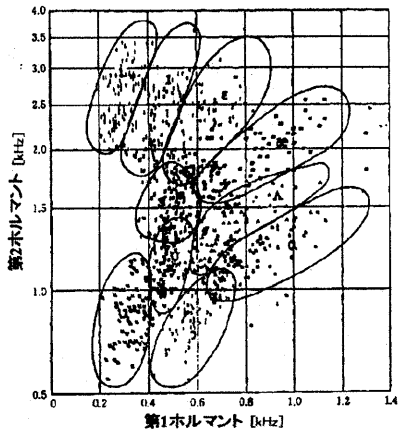


図4　第1ホルマント、第2ホルマントによる母音知覚例[1]

3 音素マッチング処理の実現

3-1 時間周波数タイリング

一般に信号 f(x) は時間軸に沿ってある点 $\tilde{x}$ を中心に幅 Δf の領域を占める。時間軸の幅 Δf と周波数軸の幅 $\Delta \hat{f}$ は次の不等式を満たす。

$$\Delta f \Delta \hat{f} \geqq 1/2$$

Δf は f(x) のほぼ中心 $\tilde{x}$ の片側の幅であるから、信号 f(x) の時間的広がりは $2\Delta f$ となる。同様に周波数的広がりは $2\Delta \hat{f}$ であるから、$2\Delta f \Delta \hat{f} \geqq 2$ が成り立ち面積2が信号の最小単位となる。

ウェーブレット変換 $(W_\psi f)(b, a)$ は式 (4) のようになり

$$(W\psi f)(b,a)=\int \frac{1}{\sqrt{|a|}} \overline{\psi\left(\frac{x-b}{a}\right)} f(x)dx \qquad (4)$$

$(W_\psi f)(b, a)$ を $(b, 1/a) = (2^{-j}k, 2^{-j})$ とおいて離散化すると

$$dk^{(j)} = 2^j \int \psi(2^j x - k) f(x)dx \qquad (5)$$

となる。$(W_\psi f)(2^{-j}k, 2^{-j})$ は $dk^{(j)}$ としている。

離散ウェーブレット変換の信号平面状の表示は図5のようになる。

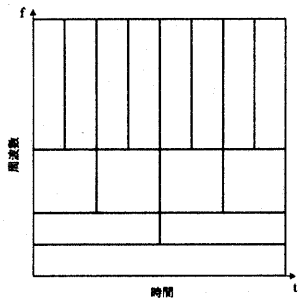


図5　ウェーブレット変換の時間周波数タイリング

3-2 テンプレートマッチング

今回行ったテンプレートマッチングの計算方法を記載する。各レベルでの入力音声 $G(k+j)$ と音素のテンプレート $T(j)$ の一致度を $M(k)$ とする。

$$M(k) = \sum_j |G(k+j) - T(j)| \qquad (6)$$

式 (6) を計算すると、テンプレートの形状と一致した位置 k で $M(k)$ が最小となる。

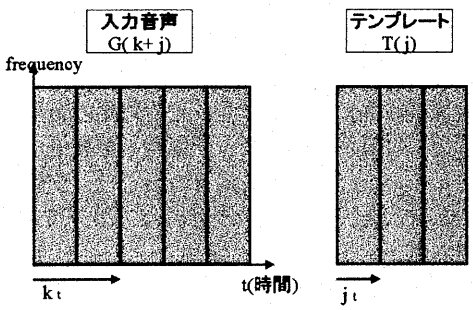
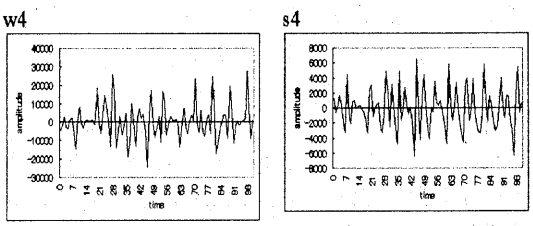


図6　テンプレートマッチング

音素テンプレート a のレベル4とレベル6の scaling 係数と wavelet 係数のテンプレート波形を図7に示す。



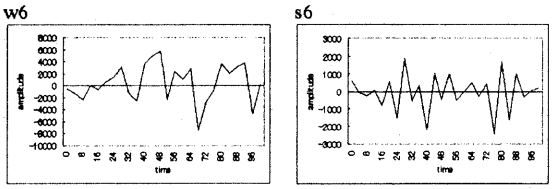


図7　音素aのレベル4とレベル6のテンプレート

図9は、音声マッチングの実現方法である。フレームの位置を決定したあとそのフレーム長で切り出したデータをウェーブレット変換する。各レベルごとにそのレベルに応じた標準のテンプレートが用意してあり切り出されたフレーム長の入力音声がどの音素のテンプレートと似ているかをマッチングさせる。第1ホルマントの周波数帯で選ばれる各音素テンプレートの最小値と、第2ホルマントの周波数帯で選ばれる各音素テンプレートの最小値とを足す。この結果、一番値が最小となる音素テンプレートを音素のマッチング結果とする。

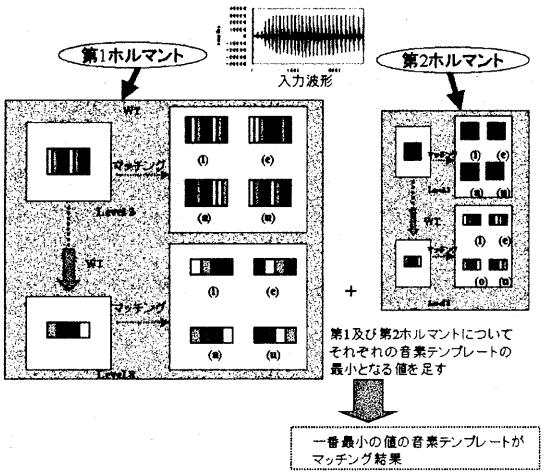


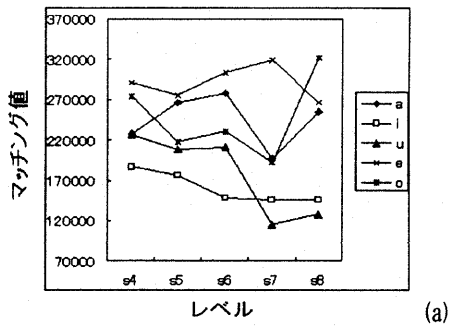
図8　音素マッチング処理の構造

4　実験

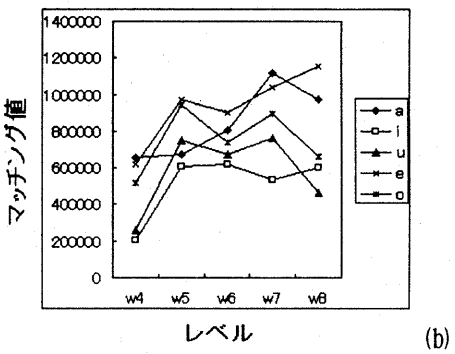
「a」「i」「u」「e」「o」の音素の組み合わせで作られる単語を考えた時、「うえ(上)/ue/」という単語のマッチングについて調べてみる。これらの音声波形はマイクからの入力でそれぞれ16khz 16bitでサンプリングされている。

今回の実験では男性の音声データを使っているので、第1ホルマントは62.5Hzから1000Hz、第2ホルマントは500Hzから2000Hzの範囲で解析する。第1ホルマントでは「レベル5，6，7，8」、第2ホルマントでは「レベル4,5」のscaling係数とwavelet係数に相当する。

フレームで切り取られた入力音声/ue/の前半の音声波形と/ue/の後半の音声波形とを/a//i//u//e//o/の音素のscaling係数とwavelet係数のそれぞれのテンプレートでマッチングさせた時のレベル4から8までの結果をプロットしたのが図9、図10である。

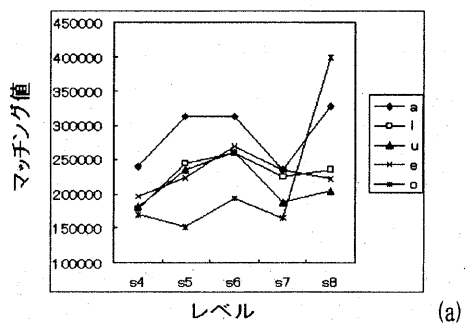


(a)



(b)

図9　ueのuと推測される音声波形のマッチング値
(a) scaling係数　(b) wavelet係数



(a)

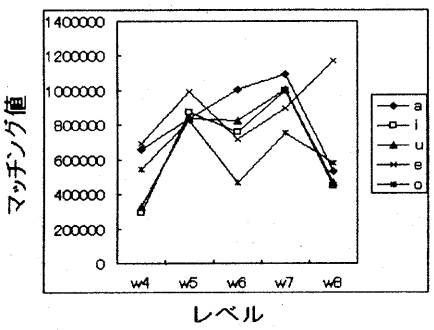


図 10　ue の e と推測される音声波形のマッチング値
(a) scaling 係数　(b) wavelet 係数

scaling 係数のマッチングは基本周波数のマッチングであると考えられる。フレーム長を決めるために求めた基本周波数によってどの音素も 100Hz－200Hz の範囲に持っていることが分かった。この周波数帯に当てはまるレベルはレベル 7 に相当する。この結果 ue の前半の音声波形では u の音素テンプレートが 114439 で第 1 候補に選ばれた。同様に ue の後半の音声波形では o の音素テンプレートが 165000 で第 1 候補に選ばれた。e の音素テンプレートは第 5 候補となっている。

次に、第 1 および第 2 ホルマントによる音素分類を行う。第 1 ホルマントと第 2 ホルマントで、それぞれの音素テンプレートが最小となるマッチング値を求め、第 1 および第 2 ホルマントのマッチング値を足した時の値が表 1、表 2 である。

表 1　/ue/の前半の音声波形と各音素テンプレートとのマッチング値

a	i	u	e	o
1325976	737656	721413	1523800	1173184

表 2　/ue/の後半の音声波形と各テンプレートとのマッチング値

a	i	u	e	o
1191144	745872	800917	1412240	1010880

表の結果から ue の前半の音声波形とのマッチング値を見てみると、u の音素テンプレートが 721413 で第 1 候補に選ばれた。同様に ue の後半の音声波形では i の音素テンプレートが 745872 で第 1 候補に選ばれた。e の音素テンプレートは第 5 候補という結果になった。

これは、基準となるマザーウェーブレットの選択が不十分であると考えられ、式 (1) 式 (2) の b を変化させた時の細かな解析をすることがさらに必要である。

5. むすび

本報告では、離散ウェーブレット変換を用いた音素の簡単なテンプレートマッチングを行い、その有効性を示した。

参考文献

[1]G.E.Peterson & H.L.Harney, “The Journal of the acoustical Society of America”, 1952.

[2] 斎藤秀昭, 森晃徳, “視覚認知と聴覚認知”,オーム社, 1999.

[3] 中野宏毅, 山本鎮雄, 吉田靖男, “ウェーブレットによる信号処理と画像処理”, 共立出版 ,1999.

[4] 新島耕一, “ウェーブレット画像解析,”科学技術出版, 2000.

[5] 榊原進, “ウェーブレットビギナーズガイド,”東京電機大学出版局, 1995.

[6] 高橋進一, 池原雅章, “ディジタルフィルタ”, 培風館, 1999.

[7] 鹿野清広, 伊藤克亘, 河原達也, 武田一哉, 山本幹夫, “音声認識システム”,オーム社, 2001.

[8] 甘利俊一, “音声・聴覚と神経回路モデル”, オーム社, 1990.

[9] 半谷精一郎, “ディジタル信号処理-基礎から応用-”, コロナ社, 2000.

[10] 石井直樹, “音声工房を用いた音声処理入門”, コロナ社, 2002.