社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64 (14)
2006／12／21

# ウェーブレット変換を用いた日本語音声の音素分析

唐澤 信司† 桜庭 弘‡

†‡宮城工業高等専門学校 〒981-1239 宮城県名取市愛島塩手字野田山 48
E-mail: †shinji-karasawa@cup.ocn.ne.jp, ‡sakuraba@miyagi-ct.ac.jp

あらまし 特定話者の短音節を標本にして，0.2msec 毎にサンプリングした 1024 個のデータを単位に Haar の離散ウェーブレット変換を行い，ウェーブレット係数（WLC）の絶対値をスケール別に加え合わせた量（SWLC)を成分とした時系列のテンプレートのマッチングによって音節分析をした．音素が変わる連続的な発声の遷移領域では SWLC の比 SWLC(0.8msec 帯)/SWLC(1.6msec 帯)と SWLC(1.6msec 帯)/SWLC(3.2msec 帯)が節となる．ピッチ単位の有声音の波形分析では，ピーク値よりピッチ期間より短い 6.4msec(0.1msec 毎に 64 個)のデータの WLC で，低い解像度(0.8msec 以上)の WLC の成分を 15 個(8,4,2,1)について標本母音と間の Hamming 距離を求めて母音が弁別できる．２通りの方法で同じ話者の発話に含まれる音素の分析ができた．

キーワード Haar のウェーブレット変換，テンプレート・マッチング，短音節認識，音素遷移境界の検出，母音認識

# Making Use of Wavelet Transform in Template Matching for Phoneme Analyses of Japanese Voice

Shinji KARASAWA† Hiroshi SAKURABA‡

†‡ Miyagi National College of Technology 48, Nodayama, Medeshima, shiote, Natori, Miyagi, 981-1239 Japan
E-mail: †shinji-karasawa@cup.ocn.ne.jp, ‡sakuraba@miyagi-ct.ac.jp

**Abstract** Speaker dependent voice recognition performance was achieved with template matching (TM). In order to give a margin to TM, sums of absolute value of wavelet transform coefficients in each scale (SWLC's) are used for vector quantization. Japanese moras are recognized under the condition of 204.8msec (1024 pieces of data those are sampled every 0.2msec) as a unit of processing. As for a segmentation, the ratio of SWLC (the 0.8msec band)/SWLC (the 1.6msec band) and SWLC (the 1.6msec band)/SWLC (the 3.2msec band) become a node in t he transition region of vowel [a,i,u,e,o]. Vowels uttered the same speaker were recognized by TM with 15 piece of WLC's in low resolution (scale is over 0.8msec) where the segmentation of processing is shorter then the pitch in order to make adaptable to the valid speech sound. Here, the data were sampled at each 0,1mces and 64 pieces of data were picked up from each peak of voice and the set of data are transferred to Haar's discrete wavelet coefficients (WLC's).

**Keyword** Haar's discrete wavelet transform, Template matching, Speaker dependent phoneme recognition, Japanese mora

## 1. はじめに

ウェーブレット変換(WLT: Wavelet transform)はパターンのデータを解像度別に変換するので JPEG2000, MPEG-4 など画像データの情報圧縮に応用されている．

音声認識の分野においても，WLT を用い発声活動を抽出すること[1]，会話音声のブロックのクラス分けをすること[2]，音素の区切りを検出すること[3]などの試みが報告されている．WLT を用いて音素，ピッチ，フォルマント，および話者のクラス分けに応用することが試みられた[4]．また，B.T.TAN 等[5]によって離散（Discrete）WLT（DWLT）および Sampled Continuous WLT（SCWT）を用いて不特定話者の HMM 音素認識の特徴ベクトルを検出して認識することも試みられた．しかし，それらの試みの結果は従来の見地からは高く評価されることがなかった．

本研究は脳神経回路網に類似した認識の活動は多種，多重の解像度を持つ非常に鋭い選択性を持つフィルターであるテンプレート・マッチング（TPM）を細胞とする回路組織によって実現すると考えた．

生体の認識では感覚細胞が活動を起こし，神経細胞が活動して，筋肉細胞を活動させている．活動の意味は外界や各細胞の活動自体が担っている．音声は周波数成分が時間変化する２次元的データであり，音声に伴う複数の成分の活動が神経回路網を転送される時にパターンが現れる．そのインパルス群が，配線接続をした時のパターンと一致する時に神経細胞が再びインパルスを発生する．認識は活動であり，認識は活動単位に量子化される．

認識の動作の本質は TPM であるとすると，TPM を行う際に，DWLT により得られる解像度別に配列したデータを用いれば，合わせ余裕と同時に処理も簡単化される可能性がある．また，多様な音声には多種のテンプレートを持つことにより対応できると考えた．

本報告で用いた Haar (ハール)の DWLT は，タイムスロット以外は 0 とし、区切られた波形を正負一対の矩形をマザーウェーブレット関数としたものであり，タ

イムスリット内のデータの後半を符号変換して加えて係数を求めるので短時間に処理を遂行できる．

音声の WLT では切り取る処理区間が課題である．音声はピッチという声帯振動の単位で発生されるのでピッチ単位で認識すれば音素を認識することができる．音声の波形切り出しはピーク値を起点にしてその長さをピッチ期間より短くして，6.4msec 単位の TPM による分析のデータ処理を行った．

音素の波形切片との TPM 類似性の評価を Hamming 距離（差の絶対値の和）により求めた．なお，Hamming 距離はユークリッド距離より計算時間が短い．

日本語の短音節（拍:mora）の特徴は発声動作に依存し，音節全体(200msec 程度)の音声の変化となって現れる．そこで，音節全体(204.8msec)の波形のウェーブレット係数(WLC)の絶対値をスケール別に加算した量（周期帯別の成分量に相当）で TPM による分析を行なった．その際に短音節単位の標本は早口で発声し，入力する音声を頻繁にシフトし(6.4msec)，処理区間の最大振幅を 1 とするデータの規格化をして TPM した．以下にそれらの分析方法と得られた結果を報告する．

## 2. テンプレートマッチング(TPM)による認識

### 2.1. 脳神経系おける情報処理が量子化する原理

神経細胞は神経回路網の中で発火条件を満たした時に生化学反応によりインパルスを発生する．その活動を引き起こす条件は同時に発生したインパルス活動単位群のパターンである．この神経細胞の活動が認識活動を担うので認識は活動単位に量子化される[6]．

活動単位を転送するモデルによって入力する刺激のパターンが出力の刺激のパターンに変換される．従来は，活動単位を転送するというモデルで神経回路網の動作を理解することがなかった．むろん，変化しない神経回路に活動の前の状態と活動の後の状態は記憶される．情報の世界と実世界は常に相違する．情報は変化せず，現実は変化を続けてやまない．活動は始まりから盛んになって終わるので，インパルス的である．そこで活動を再演する記憶はルールの単位で記憶され，ルールが変わらず適応性があるので思考の世界は普遍的であり，アナログ的であると考える．

神経細胞の活動には 5 msec 程度の不応期があって，毎秒 200 程度以上の活動単位を発生できない．その神経細胞の発するインパルスのパターンを纏めて上位の活動単位とするので，階層構造の神経回路網が形成される．その神経回路の活動に発話も音声認識の活動も依存するので活動は量子化されて回路は階層構造になる．

### 2.2. 音素の分析処理のデータの切り出し方法

話し言葉の音声は変動が大きいバラツキのある音波であるが，音声の周波数帯の範囲は 200Hz から 2kHz 程度であり，発声器官の動きは 1 秒間に多くても数個の変化で，その遷移は緩やかである．音声の音素の遷移領域では音素の周波数成分が節のように集まる．その節と節の中間の腹の部分で典型的な音素が見出され，音素分析ができる．多種多様な音声から文字に変換する情報を抽出するには音声の特性を考慮した解析が必要である．

### 2.3. Haar の DWLT の TPM への応用

Haar の DWLC によってテンプレートとの一致度を求める際に，スケール小さな多量のデータを省略することにより合わせ余裕を持つ TPM ができる． 図 1 には，WLC のデータをスケール[1-6]の 64 個で合成した波形と[3-6]の 15 個で合成した波形を示す．

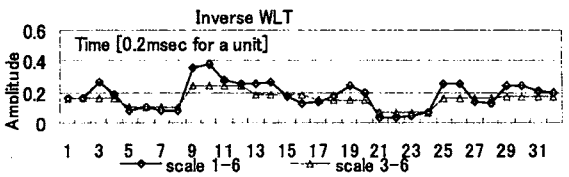


図 1 解像度を下げて TPM の合わせ余裕を得る方法
Fig.1 Wavelet transformation that provides margin on template-matching

### 2.4. 本研究に用いた機器およびソフトウエア

本研究では，音声をマイクロフォンからアンプを通して， Interface 社の A/D カード CSI-320312 を経由してノートパソコンに入力した．

データは Excel のマクロ機能である Microsoft の VBA(Visual Basic for Applications)のプログラムで処理した．A/D 変換カードのプログラムは文献[7]，ウェーブレット変換のプログラムは文献[8]を参考にした．

## 3. 連続的な音声の音素遷移のセグメンテーション

### 3.1. 音声波形のデジタルデータの採取方法

図 1 「おはよう」という男性の音声の波形を示す。データはサンプリング毎秒 10000 個あるいは 5000 個で入力レンジ±５Ｖ，分解能 12bit で採取し，振幅の規格化は WLT 処理の区間内で最大振幅を 1 とした．

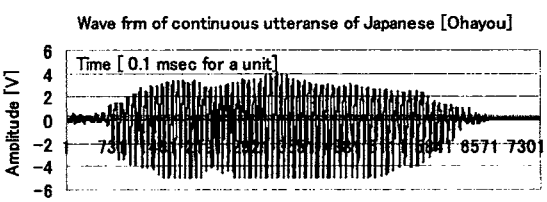


図 2 日本語音声「おはよう」の波形
Fig.2 Wave form of Japanese greeting [Ohayou]

### 3.2. ウェーブレット係数(WLC)に現れる音声の特性

パターン状の音声の波形データを WLC に変換し，それを逆変換すれば元の波形が再現される．その際に得られる WLC の絶対値の和はスケール別の含まれるウェーブレットの量（周期別の変動量）に相当する．

図 2「おはよう」という音声の波形を示す．図 3 はサンプリングを毎秒 10000 個として，256 個(25.6msec)毎に WLT して，各スケール別にタイムスリットをシフトシフトして得た WLC の絶対値の和（SWLC）を求めたものである．スケール 1 のタイムスリットは 0.2msec で Haar のマザーウェーブレット関数（＋と－の矩形波）がシフトして繰り返えされる．

音声のレベルが高くて時間進行とともに音のレベルの変化の多い成分は[0.8msec:1.25kHz]，[1.6msec:625Hz]，[3.2msec:313Hz] の周期帯である．なお，[6.4msec:156Hz]の成分は声帯振動の繰り返しであるピッチに接近している．

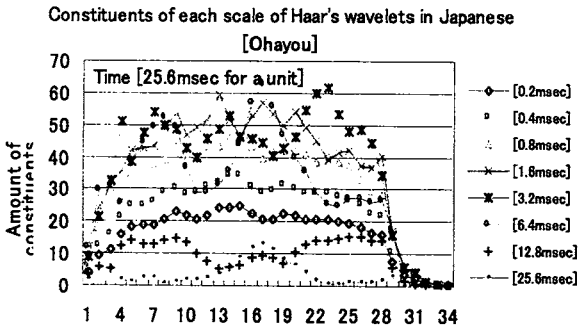


図3 Haar ウエーブレット変換による「おはよう」の周期成分特性
Fig.3 Components in wavelet scale in Japanese greeting [Ohayou] obtained by wavelet transformation

### 3.3. SWLC の比率に現れる音素遷移

音声の振幅レベルが高く，その変化が著しい周波数帯の中央の[625Hz 帯]の成分で規格化した[1.25kHz 帯]と [313Hz 帯]の比の対数を求めて図 4 に示す．

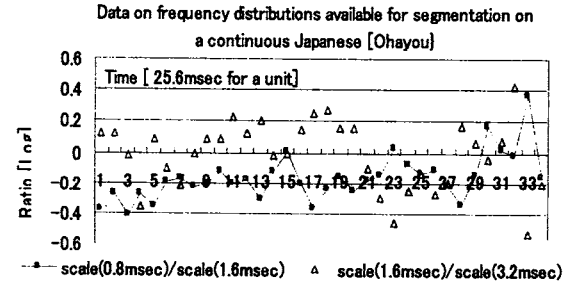


図 4 「おはよう」の音声のスケール主要成分の比の特性
Fig.4 Ratio of principal c frequency components on [Ohayou]

図 4 に示すように，音素が遷移する途中で音素の周波数成分の分布パターンが節のように集まる．この領域は連続音声における音素の遷移「渡り」(Transition) あるいは息つぎである．従って，図 4 に示す比が音素の区切り検出するデータにすることができる．

## 4. ピッチ単位の TPM による音素分析

### 4.1. 音素成分の TPM 分析の方法

音素のわずかな違いがテンプレートとの距離を求めることにより検出できる．解像度の TPM ではその照合の処理を短時間で行うことができる．

本研究では次のような理由でパターンマッチングする際に，テンプレートも照合処理のデータの切り出しを同一に，音声の波形切り出しをピーク値起点にしてその長さをピッチ期間より短くした．

1. 音声は声帯運動により間欠的に発生する空気振動が調音器官の変調を受けるので，ピーク値以後の周波数成分にはピッチの変化の影響が少ない．
2. Haar の WLT では処理するポイント数を 2 べき乗に固定しないと処理のプログラムが非常に煩雑になる（データが 2 の倍数単位で区切られる）．
3. データの切り出しの際に振幅のピーク値などにより位置合わせを行う．

Hamming 距離の差の絶対値の和はユークリッド距離の自乗平均の平方根より計算時間が短いので照合に用いる音素の波形切片との類似性の認識を Hamming 距離により求めた．この距離の値は連続的に類似性を評価できるので音声の変化を連続的に評価できる．

### 4.2. 連続的に発声した母音「あいうえお]の音素分析

図 5 は連続的に「あいうえお」と発声したときの音声を示す．「あ」と「い」発声の間に若干発声が弱くなるが発声にはこのような発声の途切れが混入する．

図 6 は図 5 のデータを HWLC に変換して，HWLC を使って周波数依存性を求めたものである．音素によって変化が著しい周波数領域は前節と同様に[1.25kHz], [625Hz],[313Hz], [157Hz]の帯域である．

図 4 と同様なデータ処理で周波数成分の比を求めると図 7 が得られ，音素が遷移する領域で周波数成分が節のように集まる．

音素レベルの認識をピッチ以内の波形であるテンプレートと同様に区切った入力のデータを HWLC に変換し，両データ間の Hamming 距離を図 8 に示す．TPM による認識では同じ音声でも僅かな処理のずれも見出すことができる．図 8 では一致度を示す距離が同じポイントでは 0 となり，その点から外れる際に Hamming 距離が連続的に変化している．

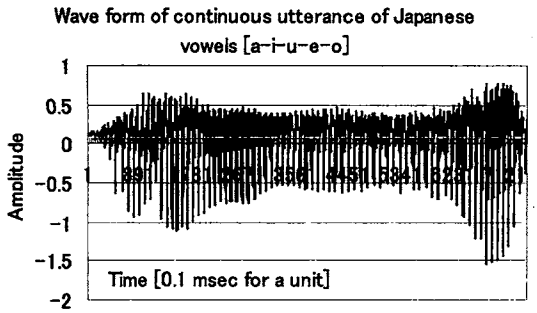


図5連続的に「あいうえお」と発声した音声の波形（男性）
Fig.5 Wave forms of vowels [a-i-u-e-o] uttered by male

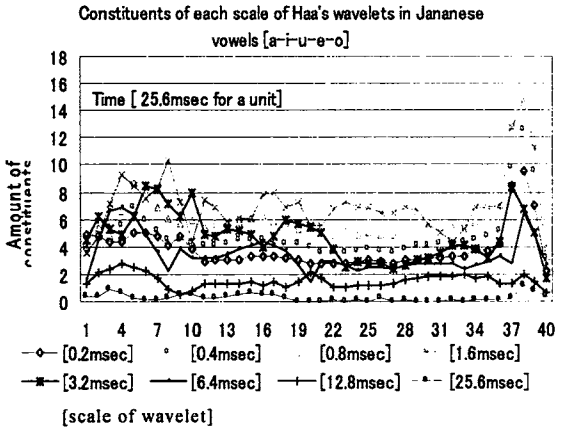


図6連続して発声した母音「あ-い-う-え-お」についてSWLCより求めた各スケール成分の時間変化]
Fig.6 Sums of coefficients (SWLC) those are obtained from Haar's wavelet transformation on vowel [a-i-u-e-o].

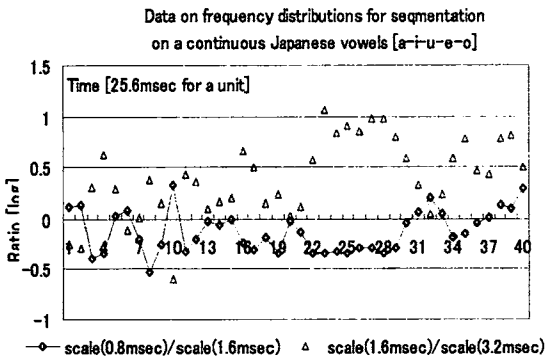


図7　音素の遷移の区切り検出方法（[1.25kHz]と[313Hz]の周波数成分に対する[625Hz] 成分の比の対数）
Fig. 7　Data on frequency distribution available for segmentation on continuously uttered vowels [a-i-u-e-o]

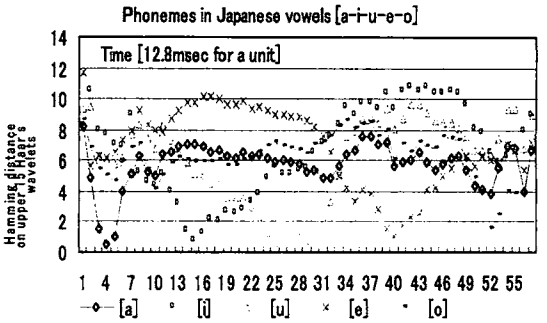


図8 連続発声母音[あ-い-う-え-お]の音素分析
(テンプレートを自身から切り取り，15個のWL係数の差を取ってその絶対値の総和でTPMを評価)
Fig.8 Phonemes in Japanese vowels [a-i-u-e-o] where the templates for recognition are picked up from itself

### 4.3. 音声のピッチの変化

音声の高音，低音の変化にも対応できるテンプレートのデータを採取するために，「うーう」「えーえ」「おーお」と母音のトーンを変えて発声した男性の音声のピッチの時間を求めたデータを図4に示す．

図9から，ピッチは6.4msec以上であり，母音の種類によってもピッチが相違している．

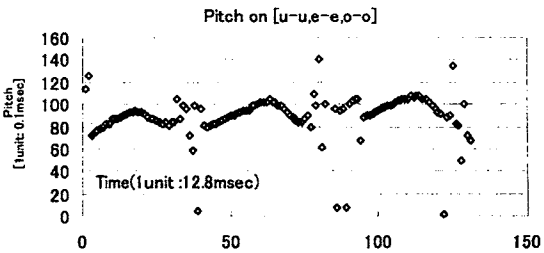


図9「う～う,え～え,お～お」と発声した男性の音声のピッチの変化
Fig.9 Variations of pitch on [u-u, e-e, o-o] uttered by a male

### 4.4. 共通のテンプレートを標本に用いた音素分析

4.4節では，母音のテンプレートとして「あ～あ」，「い～い」，「う～う」，「え～え」，「お～お」と男性が発声した母音でピッチが9.5msecの波形をピーク値から6.4msecの波形を切り出した．

図10に，図5に示した音声波形について，共通テンプレートによるTPM音素分析の結果,母音はほぼ識別できる結果を得た．ここで，Hamming距離(値が小さいほど一致度が高い)はマザーウェーブレットのスケール[0.8msec], [1.6msec], [3.2msec], [6.4msec]それぞれに8個，4個，2個，1個として合計15個のウェーブレット係数の差の絶対値の和として求めた．

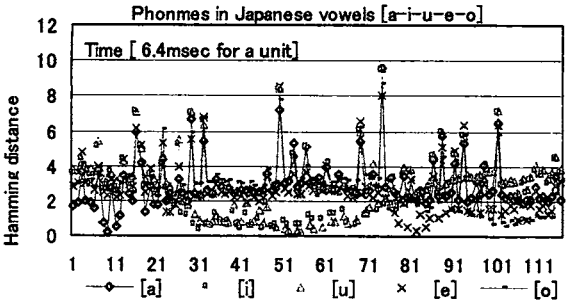


図 10 連続発声母音「あ-い-う-え-お」の母音分析
Fig.10 Vowels in [a-i-u-e-o] uttered continuously

図 11 に示す「か,き,く,け,こ」の音声波形について，母音の共通のテンプレートと「き」の先頭部をテンプレートとした音素分析の結果を図 12 に示す．

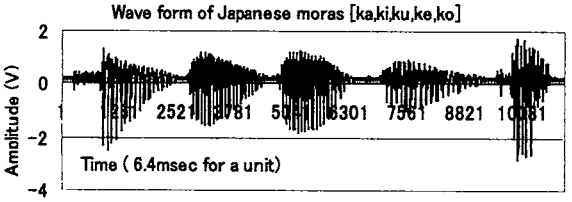


図 11 日本語音節「か-き-く-け-こ」の波形
Fig.11 Wave form of Japanese mora [ka-ki-ku-ke-ko]

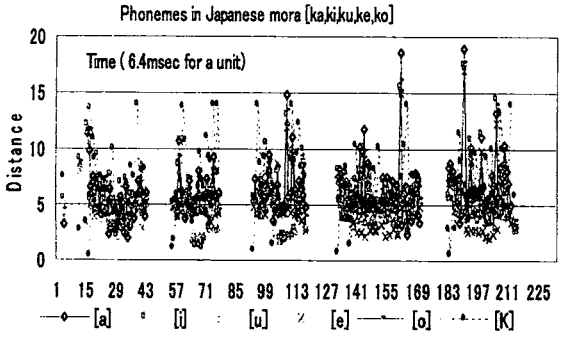


図 12 日本語音節「か-き-く-け-こ」の音素分析
Fig.12 TPM of phonemes in Japanese mora [ka-ki-ku-ke-ko]

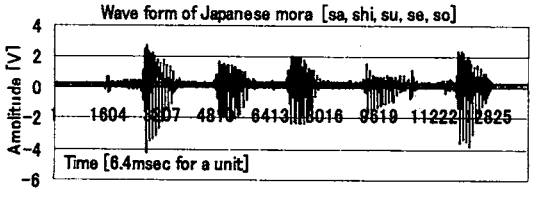


図 13 日本語音節「さ-し-す-せ-そ」の波形
Fig.13 Wave form of Japanese mora [sa-shi-su-se-so]

図 13 に示す「さ,しす,せ,そ」の音声波形について，母音の共通のテンプレートと「し」の先頭部をテンプレートとした音素分析の結果を図 14 に示す．

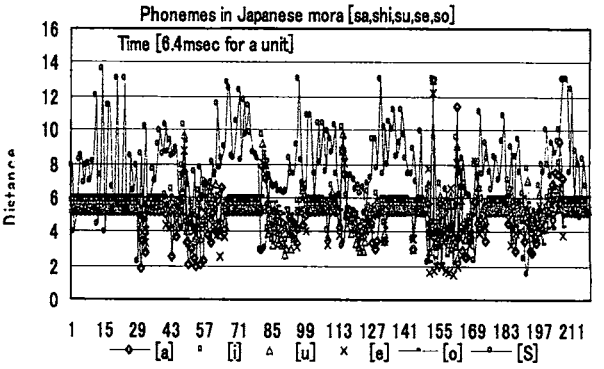


図 14 日本語音節「さ-し-す-せ-そ」の音素分析
Fig.14 TPM of phonemes in Japanese mora [sa-shi-su-se-so]

## 5. Mora 単位の WLT による音声認識

### 5.1. 切り取り位置のスケール別成分量への影響

別の音節の識別方法として 3.2 節および 4.2 節に示したスケール別の成分量 SWLC の特徴が利用できる．

そこで，音声が途切れる期間中にシフトして Mora の全期間をカバーする 2048 個（409.6msec）単位の SWLC で TPM することを試みた．振幅 1Vp-p 以下の部分を除外して区間内の最大振幅を 1 として TPM をした結果，図 15 に示すように同じ音声を 0.8msec ずらすと TPM には一致度に距離が検出される．SWLC は位相差が反映される点で周波数成分と相違する．

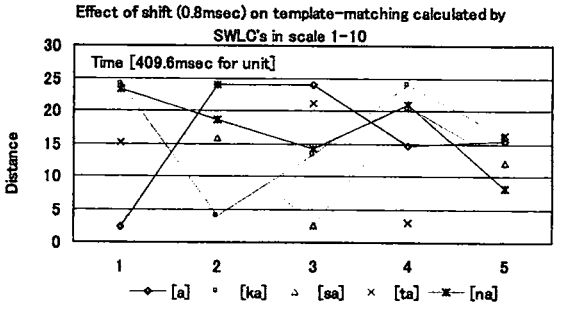


図 15 スケール成分量 TPM に及ぼす切り出し位置の影響
Fig.15 Effects of phase shift on template-matching of SWLC

### 5.2. Mora 単位のスケール成分による TPM

特定話者が話すスピードを変えた発声について時間区間が相違すると加算量が相違するので Mora 毎に SWLC の総和を 1 とする規格化を SWLC に行い，スケール別の成分比率を TPM した結果を図 16 に示す．この方法では，「あ」の発声で「あ」の他に「な」が認識

され，「た」の発声で「た」ではなく「さ」と「な」が認識され，「な」の発声で「な」の他に「さ」と「た」が認識される．Mora 全体の SWLC で TPM による弁別では音節の特徴が発声のバラツキに隠れてしまう．

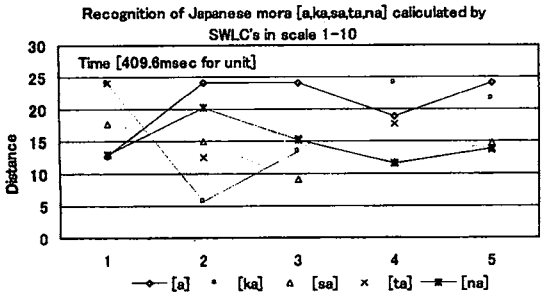


図 17　異なる速度で発声した音声の Mora 単位の TPM
Fig.17　Recognition of Japanese mora[a, ka, sa, ta, na]

### 5.3. 短音節単位フレームの頻繁なシフトによる音節分析

Mora をやや早口で発声した短音節を標本にし，シフトは機械的に 32 個(6.4msec) 間隔で頻繁に，短音節のと同じ切り取り期間 (204.8msec；5000 個/sec,1024 個) の SWLC の Hamming 距離を求めた特性を図 18 に示す．ここで，規格化は処理区間内の最大振幅を 1 とし，スケールが 1 から 8 まで(2.5kHz, 1.25kHz, 625Hz, 313Hz, 156Hz. 78Hz, 39Hz, 20Hz)帯の SWLC を照合の成分に使った．

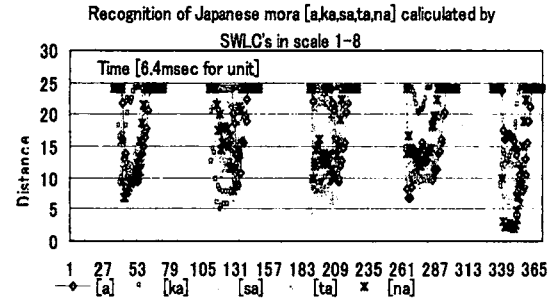


図 18 早口で発声した Mora の音節による音節の認識
Fig.18　Recognition of Japanese mora[a, ka, sa, ta, na]

図 18 はデータを採取する範囲を 200msec 程度の短音節単位で集積した SWLC をベクトル成分にした TPM により音節の認識ができることを示している．

## 6. まとめ

解像度別のデータに変換する DWLT を駆使して，シャープな選択機能を持つ TPM の処理によって多様な音声を弁別できるが，DWLT を用いた TPM では照合するデータの範囲(切り出し区間)の強い影響を受ける．

母音分析はピーク値を基準に 6.4msec の 64 個のデータを WLT して得た WLC の低解像度成分の 4 スケール 15 個の TPM によって単母音の識別ができた．

他方．日本語の「あ，か，さ，た，な」の各 Mora を単位とする認識は，Mora の発声単位(409.6msec)で SWLC を求め TPM の距離を求めた．この TPM 方法では，Mora 毎に SWLC の総和で規格化し時間区間の相違を補償したが，音節単位内の SWLC の比率には発声のバラツキが混合して音節を正しく認識できない．そこで音素の遷移の検出と組み合わせた分析として，標本のデータの処理区間を短音節程度 (204.8msec)とし，区間の最大振幅を 1 とする規格化を行い，細かくシフトして SWLC の TPM により音節の分析ができた．

この音声の音素と単音節を別々に認識する方法を組み合わせて多様で変化に富んでいる音声を認識し，そのデータから脳神経系の情報処理の量子化の原理により階層的に言語活動を展開する組織が制作できる．

本報告の多重解像度の DWLT を用いた TPM はプログラムが簡単であるので特定用途の音声認識装置に組み込むような応用に適している．

今後，Concatenative Synthesis (連結的音声合成)技術と結びつけた音声認識の開発や多重解像度を利用したパターンマッチング認識技術の応用などの分野の展開に期待が持てる．

[謝辞]　本研究は，宮城工業高等専門学校電気工学科卒業生菊池進氏の励ましにより促進されました．ここに，謝意を表します．

## 文　献

[1] Y. C. LEE, S. S. AHN, Statistical Model- Based VAD Algorithm with Wavelet Transform, Proc. IEICE Transaction on Fundamentals of Electronics, Communications and Computer Sciences, 1594-1600, E89-A(6) 2006

[2] J.O. Kim, et al. On the Extraction of the Valid Speech-Sound by the Merging Algorithm with the Discrete Wavelet Transform, Inter. Conference on Computational Science, 619-628, 2003.

[3] B.Thipakom, B. Kaewkamnerdpong, Thai Phoneme Segmentation using Discrete Wavelet Transform, International Journal of Smart Engineering System Design, 389-399, Vol 5, No.4 2003.

[4] C.J.Long, S.Datta, Wavelet Based Feature Extraction for Phoneme Recognition, International Conference on Spoken Language Processing, 1996.

[5] B.T.Tan, M.Fu, A.Spray, F.Dermody, The Use of Wavelet Transforms in Phoneme Recognition, International Conference on Spoken Language Processing, 1996.

[6] S.Karasawa, Attributes of Language Use Explained by Activities of Neuron, IEICE Technical Report, pp.31-36, TL2006-11, 2006

[7] 大川善邦，“波形の特徴抽出のための数学的処理” pp.66-67, CQ 出版社, 2005.

[8] 大川善邦“Excel 実験データ処理” pp.181-183,工学社, 2005.