【発明の名称】
離散ウェーブレット変換を用いたテンプレートマッチングによる音声認識
【要約書】
【課題】 データを解像度別に変換する離散ウェーブレット変換と鋭い選択機能を持つテンプレートマッチングを用いて音声を認識する技術を提供する。
【解決手段】 音素の認識は、ピーク値から声帯振動のピッチ程度の時間範囲で切り出した標本および認識対象の音声波形を最大振幅で規格化して、離散ウェーブレットの係数に変換し，その係数を特徴ベクトルとしたテンプレートマッチングで行う。音節の認識は短音節レベルの標本および認識処理対象の音声を同じ時間範囲で採取し、最大振幅で波形を規格化して得た離散ウェーブレット係数のスケール毎の絶対値の加算値を特徴ベクトルとしたテンプレートマッチングで行う。連続音声の音素の区切りは主要な離散ウェーブレット係数の絶対値のスケール毎の加算値の比率が音素遷移領域において 1 に近くなることを利用して行う。
【選択図】
【図1】

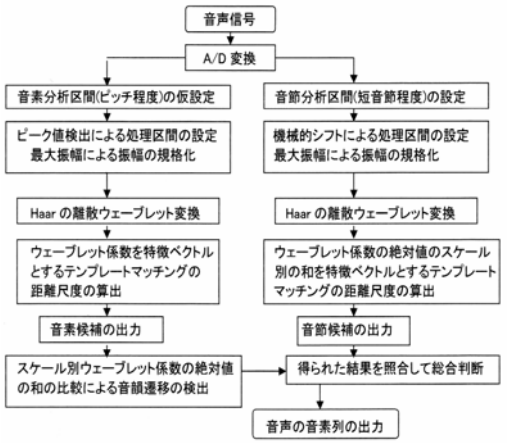


図1 離散ウェーブレット変換とテンプレートマッチングを組み合わせて入力する音声から多重に採取したデータについて標本のそれと比較して判断する音声認識組織の構成を示す。
【特許請求の範囲】
【請求項1】
ピーク値から声帯振動のピッチ程度の時間範囲の音声を切り出し、その領域の波形を最大振幅で規格化してから離散ウェーブレットの係数に変換し，そのウェーブレットの係数を特徴ベクトルとしてテンプレートマッチングにより標本との距離尺度を求めて音素レベルの音声の認識を行うこと。
【請求項2】
短音節単位程度の時間範囲で音声の波形切片を採取し、その領域の波形を最大振幅で規格化してから離散ウェーブレット係数に変換し、スケール毎の離散ウェーブレット係数の絶対値の加算値を特徴ベクトルとしたテンプレートマッチングにより標本との距離尺度を求めて短音節レベルの音声の認識を行うこと。
【請求項3】
音声を構成する主要なスケールの離散ウェーブレット係数の絶対値のスケール毎の加算値のスケール相互間比率が音韻遷移領域において 1 に近くなることを利用して連続音声における音韻遷移のセグメンテーションを行うこと。
【請求項4】
請求項1、請求項 2、および請求項 3 のいずれかを複数含む方法で音声の特徴を現すデータを多重に採取し、多重に採取したデータについて標本のデータとのテンプレートマッチングの距離尺度を求めて音声を認識すること。
【請求項5】
請求項1、請求項 2、請求項 3 および請求項 4に記載の項目のいずれかを特徴として音声を伝送することあるいは記憶することあるいは作動する機能を持つソフトウエアのシステムあるいは電子回路装置。
【明細書】
【技術分野】【0001】
本発明は、解像度別にデータを変換する離散ウェーブレット変換を用いて様々な解像度でテンプレートマッチングを行い音声の認識することに関する技術である。
【背景技術】。
【0002】
現在は、小型化、高性能化、低価格化により普及しているコンピュータ等のヒューマンインターフェースとしてキーボード入力が使われている。しかし、キーボード入力は音声入力と較べれば人間になじみが浅い。また、最近では携帯電話が発達し、個人に特化し時間や場所の制限を受けず利用できる音声認識の技術が求められるようになった。ところが、現在の音声認識の主流は統計処理を駆使した不特定話者の音声を認識する多量生産用であり、個性のある音声から普遍的な特徴を抽出することが課題であった。
【0003】
テンプレートに個人の音声を用いたテンプレートマッチングで個人に特化した音声認識が実

現できる。一般にテンプレートマッチングでは鋭い選択性を持たせられるが、その高い選択性では合わせ余裕が損なわれる。そこで、JPEG2000などの画像情報の圧縮に用いられる離散ウェーブレット変換により解像度別に配列したデータに変換すれば、目的とする情報を抽出するために適した解像度でテンプレートマッチングができる。

【0004】

従来はテンプレートマッチングで直接的に行う音声認識では対応が困難であると考えられてきた。これまでウェーブレット変換を音声認識の処理に利用した例があるが、それらは統計的な見地からウェーブレット変換をフーリエ変換の代わりに行うというような見地で試みられており、従来は成果として高く評価されることがなかった。本発明は、音声の離散ウェーブレット変換をテンプレートマッチングの前処理として音声認識処理を行うものである。

【0005】

本発明のテンプレートマッチングによる音声認識のアルゴリズムは脳神経回路網のモデルを基礎にしている。すなわち、脳は多種で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の機能を持つ神経細胞によって組織されている。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インパルスとして神経回路網を転送される時にパターン上のデータが現れる．そのインパルスのパターンが，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．そのインパルスは活動単位であり、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て、活動単位の意味は外界や各細胞の活動自体が担っている．生体ではインパルスという活動単位の転送で情報が処理されており，音声認識もインパルス的な活動単位に量子化され、デジタル的に処理できる。

【発明の開示】

【発明が解決しようとする課題】

【0006】

認識処理では、弁別する機能を高めつつ経済的かつ高速化するために必要な情報を採取し、不必要な情報を除去することが課題である。テンプレートマッチング方式の認識処理ではどのような特徴ベクトルを持つテンプレートで照合するかが機能を決定する。音声認識をテンプレートマッチングで行う際には、照合するパターン状のデータの規格化を含めて、どのような音声の特徴をどのように抽出するかというアルゴリズムを設定する指針が必要である。

【課題を解決するための手段】

【0007】

実際の音声は発声自体にバラツキが多い。これを詳細な特徴を広範囲に行うテンプレートマッチングではテンプレートの数が多くなり処理が困難になる。そこで、【図1】に示すように音素のように時間領域の狭い音声の特徴抽出処理と音節のように中程度の時間領域の音声の特徴抽出処理と音素のセグメンテーションを行う処理を分けて行い、標本の音声から複数の方法で採取したデータと新たに入力する音声から同様な方法で採取したデータとをテンプレートマッチングで照合して認識する。

【0008】

音声の認識処理は音声の発声の特徴に合わせて行う。すなわち、声帯振動のピッチ期間の10ミリ秒程度の音声の特徴抽出にはウェーブレット係数をテンプレートの特徴ベクトルの成分にして音素の分析を行う。発声器官の動作単位として 200 ミリ秒程度の短音節発声期間範囲の音声波形の特徴抽出では周期別変化量に相当するスケール別のウェーブレット係数の成分量を特徴ベクトルとしたテンプレートマッチングで認識する。

【009】

解像度別で位置順に配列されるデータに変換するにはハール（Haar）の離散ウェーブレット変換のスケール別に位置順に配列されるデータを使う。【図2】に示すように高解像度のスケールのウェーブレット係数を段階的に除いて逆変換すれば段階的に低解像度の静止画像のようなデータが得られる。ハールこのウェーブレット変換では、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の符号を変換して加え合わせてウェーブレット係数を求めるので短時間に処理できる。但し、このウェーブレット変換では処理するデータの数を 2 の冪乗とする。

【0010】

テンプレートマッチングの決定過程で一致度の評価をユークリッド距離より計算時間が短く距離に差が顕著に現れるハミング距離（差の絶対値の和）の値で評価する。

【0011】

実際に発声される音声は発音記号の種類より多くの発声記号のテンプレートを必要とする。短いテンプレートの方が共通に使えるのでできるだ

け短いテンプレートを用いて認識する。

【0012】

音素の認識として声帯振動のピッチ期間程度の音声波形のテンプレートマッチングを行う際に波形の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以内の処理単位とし、ウェーブレット係数を特徴ベクトル成分としたテンプレートマッチングで音素の認識を行う。

【0012】

音節全体に処理単位を拡大すると発声の都度に伸縮する成分がその中に含まれてテンプレートを非常に多くしなければならなくなる。「いろは・・・」など日本語の発声動作の単位である拍(mora)は200msec程度で発声されており、早口で発声した短音節全体の波形のウェーブレット係数の絶対値をスケール別に加算した周期帯別の成分量に相当する値を特徴ベクトル成分としたテンプレートマッチングで短音節の認識を行う。

【0014】

音節の認識の特徴抽出の前処理として、早口で区切って発声して短音節単位の標本を採取し、採取した区間の最大振幅を1に振幅を規格化し、また入力する音声から照合するデータも標本と同じフレーム長で採取し、同様に規格化して特徴ベクトルを求める。

【0015】

連続音声の中で音素が遷移する領域では音素の認識が難しい。そこで、離散ウェーブレット変換のスケール別成分量を求めてその比率で音素遷移を検出する。【図3】に「い」を短く発声しているが「あいうえお」と連続的に発声した音声の波形を示す。この音声について離散ウェーブレット変換の主要なスケール別成分量の比を【図4】に示す。【図4】では、音素が遷移する領域では主要なスケール別成分量の比率が1に近くなる。この方法を利用して、連続音声の音節のセグメンテーションの検討ができる。

【0016】

音声から認識された音声の情報処理としての音韻記号列を登録番号に変換して、それを解凍する組織を構築すれば音声情報を圧縮して伝送したり、記憶したりすることができる。デジタル信号はAND回路で一つに絞られOR回路により複数の成分に展開できるので、採取されたデジタルデータのパターンをデジタル回路で変換することができる。その際に発明者の特願2004-217828号書き込み可能型双方向論理回路が便利である。そこでは、出力の符号化の書き込みは逆方向から解読器として書き込み、その回路の導通点群を符号器として利用している。

【0017】

現在の高性能パーソナルコンピュータのワークステーションは64ビットを処理するレジスタの構造を持っているので音声言語を処理する技術環境は整っている。ちなみに、8ビットのデジタル信号で256 個の発音記号は区別が可能である。そこで、単語を8個の発音記号で表現すると、単語の発音記号のデータ8×8=64ビットで指定される。64ビットの単語の単位で入力して、8,192個の単語を書き込む場合にはそのマトリックスポイントは524,288個となる。8,192個の単語を登録番号で特定すれば13 ビットで指定できる。ここで、出力の登録番号は登録番号のカウンターで決めることができる。出力を128種7ビットの文字9個 で表現すると文字出力は63ビットとなる。

【0018】

情報処理装置の中では音声情報を文字情報に変換せずに、パターンマッチングを階層化して行えば情報を圧縮して処理できる。すなわち、音韻識別情報を単語レベルの登録番号に変換し、さらにその単語レベルの登録番号列を文章レベルの登録番号に変換する。逆に、文章レベルの登録番号を元の単語レベルの登録番号列に戻し、さらに単語レベルの登録番号を音韻記号列に変換をする。13 ビットのデータで1個の単語を特定して、9 個の単語の組み合わせ13×9 = 117 ビットを1 個の文章として入力し、8,192 種類の文章を登録すれば、その入力側のマトリックスのポイントは958.464 となる。8,192 種に区別された文章は13 ビットの番号で区別できるので、出力側のマトリックス要素は106,496 となる。この程度の規模の回路は半導体集積回路で実現可能である。

【発明を実施するための最良の形態】

【0019】

本発明はAD変換カードを挿入したノートパソコンを用いて処理前後のデータをマイクロソフトのOffice のExcelで表示する方式で検討したもので、そのソフトウエアはVisual Basic for Application (VBA)でプログラムされており、パソコンやロボットの入力部に組み込みソフトウエアとして使うことができる。

【0020】

多量のテンプレートマッチングを高速で処理するにはテンプレートをハードウエアに書き込めば並列に照合できる。本発明の組織をプログラマブルな半導体集積回路で制作するのが最良

である。半導体集積回路で構成するには活動単位の存在を電荷の有無として転送し、その電荷の転送はCCD やダイナミックMOS IC の回路で行う。活動単位を転送してデータ変換機能を実現する回路としては、発明者の特許第3496065号 インパルス電子装置および発明者の特願2004-217828号の書き込み可能型双方向論理回路がある。

【0021】

離散ウェーブレット変換を用いたテンプレートマッチングによる特定話者の音声認識の実施例を通して、以下に本発明の実施方法を説明する。

【実施例】

【0022】

同様に発声した音声が一致するので、標本も同様な波形切片とすればよい。「あいうえお」と連続的に発声した【図3】の音声波形自身から12.8msec 切り取った5種の波形とのテンプレートマッチングのHamming 距離を【図5】に示す。【図5】では、連続的な音声はテンプレートとの一致度も連続的に変化する様子を示している。

【0023】

波形切片は短いほうが処理単位のデータが少なくて処理時間が短くてすむ。テンプレートの数が少ない場合にはコンピュータで処理する際に処理時間が短くてすむ。そのためには照合に使う標本のテンプレートの区間を短くかつ低い解像度にする。「あいうえお」と連続的に発声した音声の波形を自身の音声から6.4ミリ秒切り取った15個の特徴ベクトルでテンプレートマッチングのHamming 距離を【図6】に示す。【図6】から選択性が低い条件の照合でも識別できることを示す。

【0024】

テンプレートの数が多いと処理時間がかかるので、標本として共通に使える母音の音声を採取したい。声帯振動のピッチは母音の種類や発音の仕方によって相違し、ピッチが変われば波形も変化する。【図7】に「うーう、えーえ、おーお」と発声した時の音声のピッチの変化を示す。

【0025】

認識処理の処理時間を非常に短縮した実施例として、ピッチが9.5ミリ秒の音声波形からピーク値より6.4ミリ秒を切り取った波形切片を共通の標本とした離散ウェーブレット変換を用いたテンプレートマッチングによる特定話者の音素の認識の実施例をしめして、この認識方法の制作の指針を説明する。

【0026】

「あいうえお」と早口で連続的に発声した音声と「あーあ、いーい、うーう、えーえ、おーお」と発声した母音標本とのテンプレートマッチングの距離を【図8】に示す。ここで、5種の母音標本はと発声した音声波形で9.5ミリ秒のピッチの波形から6.4ミリ秒切り取った。なお、入力音声の「い」は短く発声しており、「いーい」と発声した標本の「い」では認識できない。

【0027】

「かきくけこ」と発声した音声でピーク値から6.4ミリ秒切り取った音声と【図8】の処理と同じ母音標本とのテンプレートマッチングの距離を【図9】に示す。「こ」の発声の母音は「う」の標本と認識されている。「こ」の発声を早く切り上げれば認識されるものと考えられる。同じ発音記号でも複数のテンプレートを必要とする。なお、「き」の波形の子音として特徴的な先頭部分を6.4ミリ秒切り取ってKの標本波形とした。Kの特徴とした先頭領域の波形は声帯振動を持たず子音は発声の動作全体にも関係するので、短音節として拡大した音声切片から特徴を抽出して認識すべきである。

【0028】

「さしすせそ」と発声した音声で6.4ミリ秒切り取った音声切片と【図7】と同じ母音標本とのテンプレートマッチングの距離を【図10】に示す。「し」の発声の母音はいくつもの母音に認識されている。「し」の発声は「si」ではなく「shi」であると考えられる。なお、Sの標本は「し」の波形の子音として特徴的な先頭部分を6.4ミリ秒切り取って波形である。Sの特徴とした先頭領域の波形はKと同様に声帯振動を持たず子音は発声の動作全体にも関係する。

【0029】

スケール別成分量を特徴ベクトルにしたテンプレートマッチングでは周期別の変化量成分に相当するので、振幅レベルの小さな量は影響が少ない。そこで、「あ、か、さ、た、な」と発声して、各音節を全てカバーする区域[409.6ミリ秒]の音節波形標本についてスケール別成分量を特徴ベクトルにしたテンプレートマッチングの照合を試みたところ【図11】に示すように同じ音声から波形の切り出し位置を0.8ミリ秒シフトしただけで影響が現れる。

【0030】

短く発声した204.8ミリ秒の短音節単位を標本としてその範囲のスケール別成分量を特徴ベクトルにして、普通に発声した音節とのテンプレートマッチングの照合をした様子を【図12】に示す。中央部でその音節を認識し、終端部で母音を認

識している。短音節期間の音声標本でテンプレートマッチングする時にはフレームのシフトを図12に示すほど頻繁にする必要はない。

【0031】

以上、本発明の要旨を説明してきたが、本発明は図面で示す実施例の条件に限られるものではなく、本発明の主旨に逸脱しない範囲における変更や追加があっても本発明に含まれる。多様な音声と多様な音声認識の用途があるので、本発明を実際に用いる際には本明細書で説明した事柄を指針として具体的に構築する。

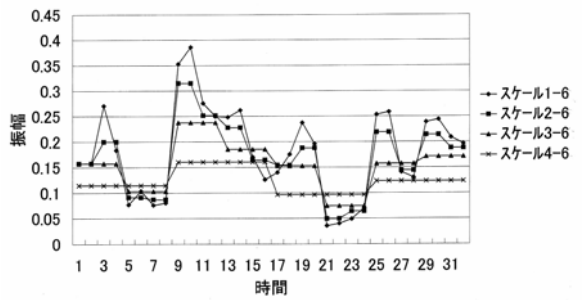
【産業上の利用可能性】

【0032】

本発明の音声を入力とする組織の活用例として音声の情報圧縮あるいは、レストラン等の注文書の音声入力、音声タイプライター、音声矯正装置、自動翻訳電話、産業ロボット、介護ロボットなどがある。

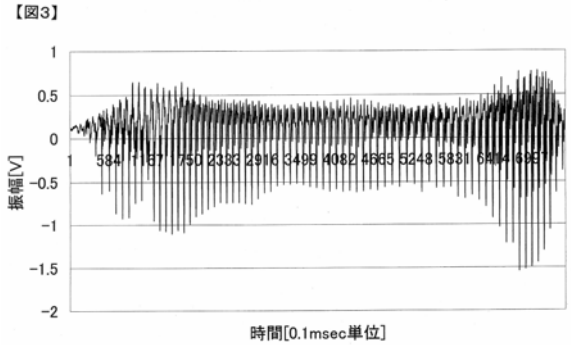
【図面及び簡単な説明】

【図2】



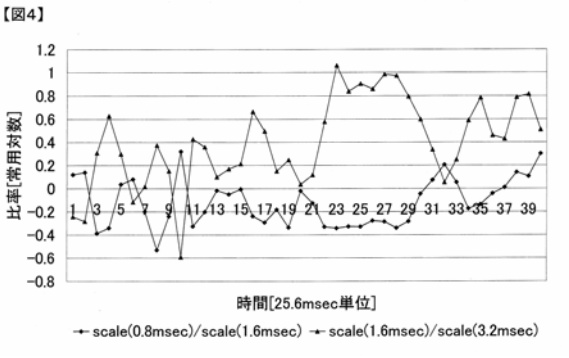
【図2】高解像度のスケールのウェーブレット係数を段階的に除いて逆変換によって求めた波形により離散ウェーブレット変換の多重解像度を示す。

【図3】



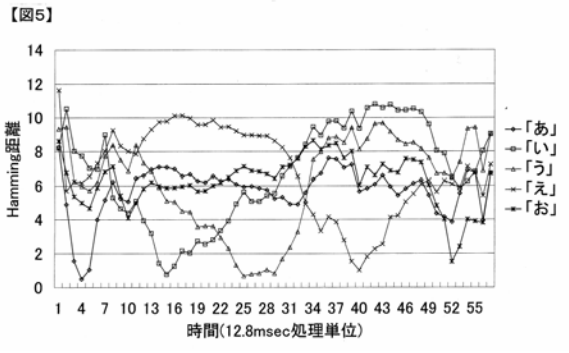
【図3】離散ウェーブレット変換を用いた音声の分析および認識処理の実施例で用いた「い」を短く「あいうえお」と連続的に発声した音声の波形を示す。

【図4】



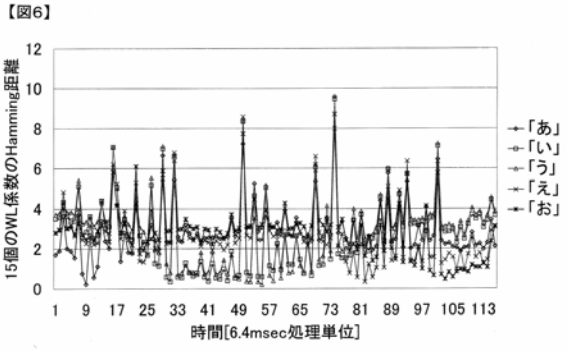
【図4】図3に示す音声の離散ウェーブレット変換のスケール別成分量の比率で音素遷移が検出できることを示す。

【図5】



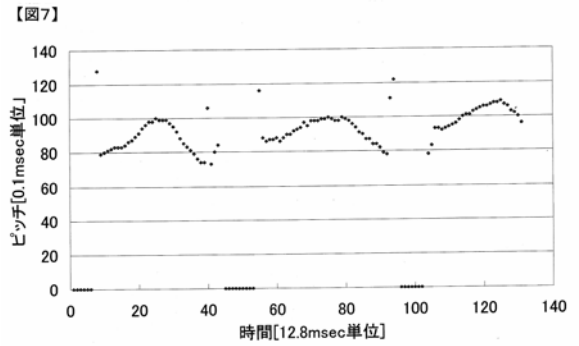
【図5】図3に示す音声の波形から 12.8msec 切り取った 5 種の標本波形とのテンプレートマッチングの Hamming 距離を示す。

【図6】



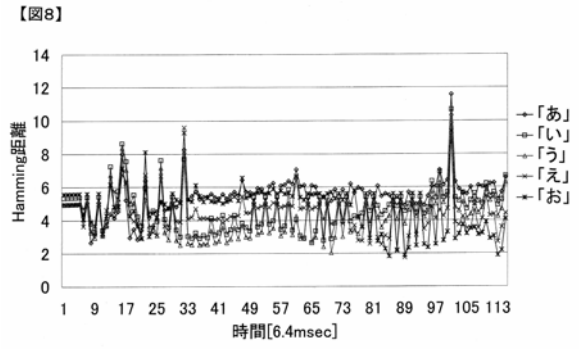
【図6】図3に示す音声の波形を自身の音声から 6.4msec 切り取った 5 種の低解像度の波形でテンプレートマッチングの距離を示す。

【図7】



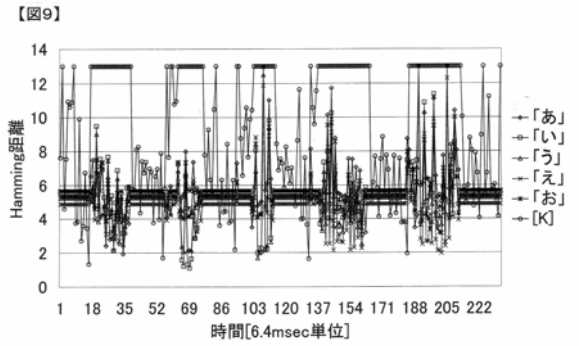
【図7】共通のピッチとして 9.5msec を採用した際に参照した「うーう、えーえ、おーお」と発声した時の音声のピッチの変化を示す。

【図8】



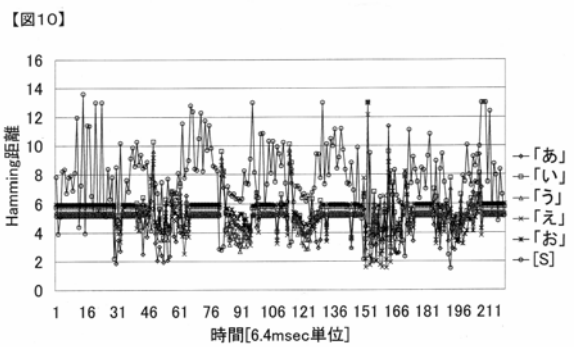
【図8】一部は図 6 に示す 9.5msec のピッチの波形から 6.4msec 切り取った 5 種の低解像度の波形で図3に示す音声のテンプレートマッチングの距離を示す。

【図9】



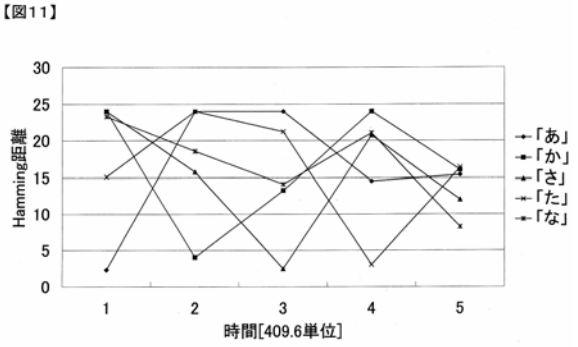
【図9】 9.5msec のピッチの波形から 6.4msec 切り取った 5 種の低解像度の波形で「かきくけこ」と発声した音声のテンプレートマッチングの距離を示す。

【図10】



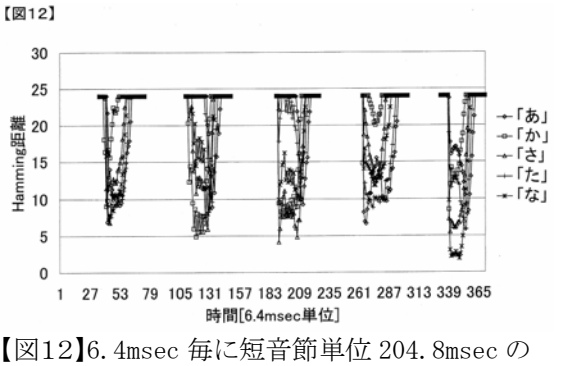
【図10】 9.5msec のピッチの波形から 6.4msec 切り取った 5 種の低解像度の波形で「さしすせそ」と発声した音声のテンプレートマッチングの距離を示す図で ある。

【図11】



【図11】同じ音節[409.6msec]のスケール別成分量を特徴ベクトルにしたテンプレートマッチングの距離において、時間を 0.8msec シフトした影響を示す。

【図12】



【図12】6.4msec 毎に短音節単位 204.8msec のフレームをシフトして、短音節単位のスケール別成分量を特徴ベクトルにしたテンプレートマッチングの Hamming 距離で音節が認識できることを示す。