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

2007－ＳＬＤＭ－128（１）
2007／1／17

聴覚前処理部の回路化

宇佐美 裕也† 荒井 英彦† 曹 越† 高橋 一志† 長野 利隆†
関根 優年†

† 東京農工大学 工学府
〒 184-8588 東京都小金井市中町 2-24-16
E-mail: †{usami,chuchu,caoyue,kazushi,nagano}@sekine-lab.ei.tuat.ac.jp, ††sekinem@cc.tuat.ac.jp

あらまし 人間の聴覚は二つの耳から流入した音に対し，頑健性のある前処理を施すことで，音の特徴量や発生位置などを即座に抽出することができる．そこで本研究ではこの前処理を行う聴覚器官のひとつである蝸牛に注目し，その機能を論理回路によって実現する手法を提案する．具体的には，蝸牛を模擬するモデルを verilog-HDL によって回路化し，本研究室で開発した hwModule 上の FPGA で実装することで音像と呼ばれる音の特徴量を得ることを目指し，聴覚前処理部のディジタル回路化を行う．今回，提案する手法を日本語の母音に対して適用し，得られた音像にテンプレート・マッチングを行った．その結果，音像による母音の判別が可能であり，提案手法の有用性が確認できた．
キーワード 聴覚モデル，多重解像度解析，テンプレート・マッチング，母音認識

Achieve a preprocessing part of auditory sense with circuit

Yuya USAMI†, Hidehiko ARAI†, Etu SOU†, Kazushi TAKAHASHI†, Toshitaka NAGANO†, and Masatoshi SEKINE†

† Faculty of Engineering, Tokyo University of Agriculture and Technology
2-24-16 Naka-chou, Koganei-shi, Tokyo, 184-8588 Japan
E-mail: †{usami,chuchu,caoyue,kazushi,nagano}@sekine-lab.ei.tuat.ac.jp, ††sekinem@cc.tuat.ac.jp

Abstract Auditory sense of human can immediately extract a feature quantity and location that sound happened from sound that flows in ears using robust preprocessing. In this research,we paid attention to cochlea that was a part of the auditory organ, and proposed method that got cochlear function with a logical circuit. Concretely speaking,we compose the model that simulated cochlea of the circuit by verilog-HDL. By implementing this circuit in an FPGA on our hwModule board, we make a digital circuit of a preprocessing part of auditory sense. This aim is to get a feature quantity of sound named "Sound Image". We get "Sound Image" that obtain by applying proposed method to vowel of Japanese, and Template Matching using "Sound Image" is performed for the vowel recognition. As a result,the vowel recognition was possible by "Sound Image", and we could confirm the utility of proposed method.
Key words Auditory Model,Multi-Resolution Analysis,Template・Macthing,Vowel Recognition

1. はじめに

1.1 研究背景

近年，人間の持つ様々な機能を明らかにし，工学的に実現しようとする試みが為されており，人間の聴覚機能に関する研究も行われている．人間の聴覚機能は非常に優れており，雑音下での音声認識やブラインド信号処理，音源の方位定位などを即座に可能とする．これらの優れた処理を可能にしているのは，人間の認識能力もさることながらその聴覚器官が行う前処理にも大きな役割があると考えられており，聴覚器官を工学的に模擬するモデルが考案されている．

しかし，このようなモデルをハードウェアによって実現するという試みはあまり為されていない．人間の聴覚機能を工学的に実現するには，その処理の複雑さ，膨大さを考慮すると，ソフトウェアのような逐次型情報処理とハードウェアのような並列型情報処理を組み合わせることによって分散処理を行うことが

重要であると考えられる.

そこで, 聴覚機能の前処理部を担う器官のひとつである蝸牛と呼ばれる器官のモデル [1] をもとに, ハードウェアとソフトウェアを組み合わせた聴覚の前処理部を構成する.

1.2 研究目的

本研究では, 本研究室で開発した hwModule(後述) を利用する. この hwModule によって, 蝸牛が行う前処理部の一部の演算をハードウェアで行い, 音の特徴量を動的に抽出するディジタル回路の構築を目的とする.

ディジタル回路化によって得られた音の特徴量を音像とし, この音像を用いたテンプレート・マッチングを日本語の母音を対象に行うことにより母音認識が可能か判定し, 聴覚前処理部の回路化についての評価・検討を行う.

2. 聴覚モデル

蝸牛を模擬するモデル, 聴覚モデルが行う音の特徴量を抽出する前処理は以下に示すような処理である.

2.1 周波数成分解析

耳から流入した音に対し, 蝸牛では有毛細胞による周波数解析を行う. この周波数解析は音を聴覚フィルタバンクに通過させることで実現される. 聴覚フィルタには式 (1) で定義されるガンマチャープ関数が用いられる. そして, ガンマチャープ関数を窓関数とし, 式 (2) で定義される畳込み演算を行う. この畳込み演算によって, 音の周波数成分がガンマチャープ関数により抽出され, 抽出された成分が強調される. ガンマチャープ関数を構成しているパラメータ f_r(周波数成分) を変化させることにより, 図 1 に示すように, 複数の周波数帯域を解析することが可能なフィルタバンクを構成できる.

$$g_c(t) = at^{n-1}e^{(-2\pi bERB(f_r)t)}e^{(j2\pi f_r t + jcint + j\phi)} \quad (1)$$

$$S_W(\alpha f_0, t) = \int_0^\infty g_c(\alpha f_0, \tau)s(t-\tau)\,d\tau \quad (2)$$

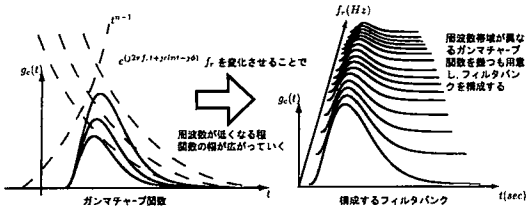

図 1 ガンマチャープ関数によるフィルタバンク

2.2 音の重ね合わせ

耳から流入した音, 特に音声は, ピッチと呼ばれる繰り返して発生する波から構成されており, このピッチは各母音ごとに異なる波形を持ち, 独自の特徴量となる. そこでこのピッチの周期に合わせて音声を重ね合わせて大きな音にする. これにより, ピッチに寄与する成分が強調される.

これを実現するのが式 (3) である. 時間成分 t_p を音声のピッチの周期に合わせる事により, ピッチに寄与する成分を強調することができる. この式は減衰項 $e^{-\zeta\tau}e^{-\eta kt_p}$ があるため, 時間軸である程度離れている音は寄与しないように加算される.

$$A_I(\alpha f_0, \tau) = \sum_{k=0}^{\infty} S_W(\alpha f_0, \tau_1 + Kt_p)\cdot e^{-\zeta\tau}e^{-\eta kt_p} \quad (3)$$

2.3 メリン変換

人間の聴覚は, 男性と女性が発話する母音の"あ"のように, 周期が異なる類似度の高い波形を同一のものとして捉えることができる. これは, 聴覚器官が音の周期を時間軸において正規化しているためと考えられている.

そこで, この時間軸における正規化を行うのにメリン変換を用いる. メリン変換は (4) 式で定義される. この変換は (5) 式で示される性質がある. これにより, 時間軸のスケーリングが係数 α^{-p} として抜き出せるので, この係数を用いれば, 周期を時間軸で正規化することができ, 周期の異なる類似度の高い波形は同一波形として扱うことができる.

これを利用し, 話者の違いによるピッチの周期変動を解決することが可能となる.

$$M[s(t)] = S(p) = \int_0^\infty s(t)t^{p-1}\,dt \quad (4)$$

$$M[s(\alpha t)] = \alpha^{-p}S(p) \quad (5)$$

3. 聴覚モデルによる前処理回路

前章で示した処理をハードウェアで実現するために, 本研究では前処理の本質的な意味を捉え以下に示す処理手法を提案する.

3.1 離散ウェーブレット変換と畳込み演算

ガンマチャープ関数による畳込み演算は,hwModule による実装を考慮し, 離散ウェーブレット変換と矩形関数による畳込み演算に変更した. 離散ウェーブレット変換の基底として用いる Haar 関数のスケーリング係数 $\phi_H(x)$ は式 (6) で, ウェーブレット係数 $\psi_H(x)$ は式 (7) でそれぞれ定義される.

$$\phi_H(x) = \begin{cases} 1 & (0 \leqq x < 1) \\ 0 & (otherwise) \end{cases} \quad (6)$$

$$\psi_H(x) = \begin{cases} 1 & (0 \leqq x < \frac{1}{2}) \\ -1 & (\frac{1}{2} \leqq x < 1) \\ 0 & (otherwise) \end{cases} \quad (7)$$

スケーリング係数 $\phi_H(x)$ とウェーブレット係数 $\psi_H(x)$ を用いて入力された信号に対し離散ウェーブレット変換を行う. このとき, 離散ウェーブレット変換を 1 回行うことを 1 レベル変換と呼ぶ. 図 2 のように 1 レベル変換後のスケーリング係数に対して再び離散ウェーブレット変換を行うことを 2 レベル変換という. このように, 離散ウェーブレット変換後の値を用いて再び離散ウェーブレット変換を行うことで各レベルごとの係数を得ることができる. これを多重解像度表現という.

離散ウェーブレット変換後のスケーリング係数に対して, 式 (8) で定義される矩形関数を用いて畳込み演算を行う. 離散信号に対する 1 レベルの離散ウェーブレット変換で得られるスケーリング係数は, 離散信号のサンプリング周波数の 1/2 倍までの

周波数成分を抽出することができる．すなわち,n 回の変換を行うと，サンプリング周波数の $1/2^n$ 倍までの周波数成分を抽出できる．この特性を用いて，各レベルのスケーリング係数に対して，時間パラメータ T により，帯域を変化させた矩形関数による畳込み演算を行うことで，解析可能な周波数帯域を増やし，フィルタバンクを構成する．

$$rect(x) = \begin{cases} 1 & (0 \leq x < T) \\ 0 & (otherwise) \end{cases} \tag{8}$$

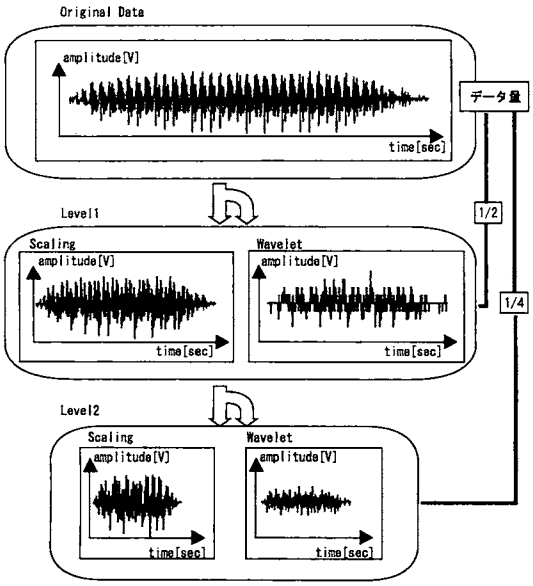

図 2　音の多重解像度解析

3.2　音声の周期検出と重ね合わせ

音の重ね合わせを行うには，まず音声波形を構成するピッチの周期を検出しなければならない．しかし，ピッチの周期は，一つ一つのピッチの波形ごとに微妙に異なるので，音声波形の全体からではなく，ピッチ波形一つ一つから周期を動的に検出しなければならない．

このような処理をソフトウェアで行うのは少々難しいので，ハードウェアによって周期を検出する回路を製作することが望ましい．そこで，本研究では音声波形のピッチの周期を検出する回路を製作した．

ピッチの周期検出には式 (9) で定義される自己相関関数をもとに積ではなく和によって周期が検出できるようにした．後は，式 (3) に従い，ピッチの周期値から音声のピッチ波形を切り出し，重ね合わせていけばよい．

$$\phi_{11}(\tau) = \frac{1}{N}\sum_{t=1}^{N} s_1(t)s_1(t+\tau) \tag{9}$$

3.3　周期情報の正規化

メリン変換の意味するところは，周期の異なる類似度の高い波形を同一波形として扱うことができることである．従って，検出したピッチ周期にもとづき，周期波形を全て一定値に正規化すればよい．これにより話者の違いによるピッチの周期変動を抑えることが出来る．

4.　hwModule を用いたアプリケーション

本研究では，このディジタル回路による前処理を利用した母音認識アプリケーションを製作した．図 3 は製作したアプリケーションの処理フローである．まず，外付けのアナログ回路の A/D コンバータ部から音声データを離散値で取得する．取得した値に対し，離散ウェーブレット変換をレベル 4 まで行い，その値を用いて畳込み演算とピッチの周期検出を行う．そして，各レベルのスケーリング係数と畳込み演算結果を検出した周期情報にもとづき重ね合わせる．最後に時間軸での周期の正規化を行い，音の特徴量である音像を得る．本アプリケーションでは，この音像を用いて母音のテンプレートマッチングを行った．

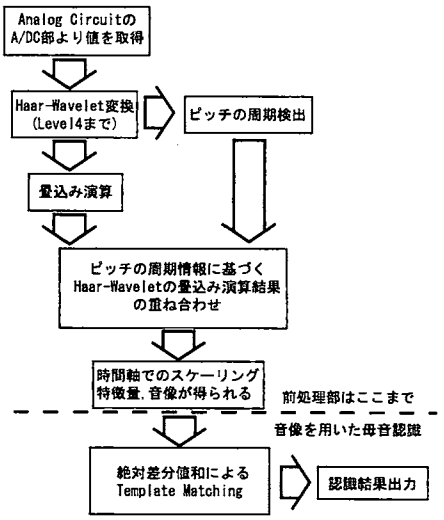

図 3　製作したアプリケーションの処理フロー

4.1　hwModule

hwModule（ハードウェアモジュール，図 4）とは本研究室で開発した再構成可能な LSI である FPGA が搭載された PCI デバイスである．hwModule は hwObject の要素である hwNet(汎用の Verilog-HDL,VHDL 等のハードウェア記述言語である HDL 言語で記述・設計された仮想回路) をダウンロードし，実装するために用いる．

図 5 に hwModule のブロック図を示す．hwModule は FPGA，メモリ，マイクロプロセッサ,GPIF,PCI バスコントローラにより構成される．FPGA は全部で 4 個搭載し,1 個をボード制御の PCI バスコントローラ用に使用し，残り 3 個を hwNet 用に使用する．更に，処理データを格納するためのローカル・メモリ (SRAM) を 3 個搭載している．

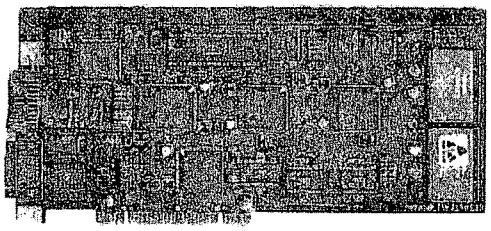

図 4　hwModule

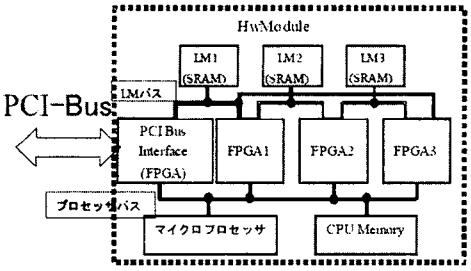

図 5　hwModule の構成ブロック図

4.2　hwObject

hwObject(ハードウェアオブジェクト) とは, ハードウェアの処理をカプセル化したもので, hwModule 上の FPGA にダウンロードし, 実装される hwNet を隠蔽したオブジェクトである. hwObject のホスト PC 上のメンバ関数は,C++言語による記述によって与えられ, ユーザーがハードウェア, ソフトウェアの違いを考慮することなく, ソフトウェアの記述を用いて容易にハードウェアを扱うことが可能となる. hwObject には, ハードウェアが得意とするビット処理, 信号処理などの機能を割り当てることにより, アプリケーション全体のスループットの向上が期待できる. 図 6 に hwObject モデルの概要を示す. hwObject Interface は, ソフトウェア側によって呼び出されるメンバ関数による hwModule へのアクセスを実行し, hwModule は PCI バスを経由したホスト PC とのデータ転送と制御を行う.

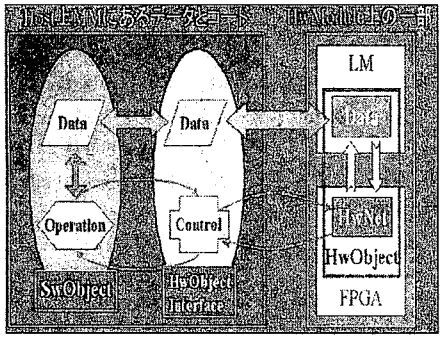

図 6　hwObject モデルの概要

4.3　音声入力回路 (アナログ回路& hwObject)

hwModule に音を離散値データとして取り込ませるために図 7 に示すアナログ回路を 2 つ製作した. この回路はブロック図に示すようにマイク・アンプ回路, ローパス・フィルタ回路,A/D コンバータ回路から構成される. アンプ部は, 表 1 に示すようなゲイン・テーブルを持ち,hwModule 側から動的に利得の変化ができる. また人間の耳を想定し,20Hz から 20KHz 帯を通過させるバンドパス・フィルタとしての機能も有している. ローパス・フィルタ回路は,AD 変換時のエイリアスを抑制するために使用している. ローパス・フィルタは, カットオフ周波数を約 22KHz, 多重帰還型の 5 次のバタワース特性とした. AD コンバータ回路は,8bit, シリアルで, サンプリング周波数は人間の耳の可聴領域を考慮し, 約 43KHz とした. また後段には GPIF に離散値を与える際に使用するフラットケーブルによる減衰やノイズの影響を抑えるためバッファを設置してある.

表 1　マイクアンプのゲイン・テーブル

入力信号 ABC	利得 [dB]
000	46(200 倍)
001	60(1000 倍)
010	66(2000 倍)
011	72(4000 倍)

また, このアナログ回路を hwModule から制御するための hwNet も製作した. この hwNet は,AD コンバータ部の制御信号生成と出力される離散値を取得する回路である.

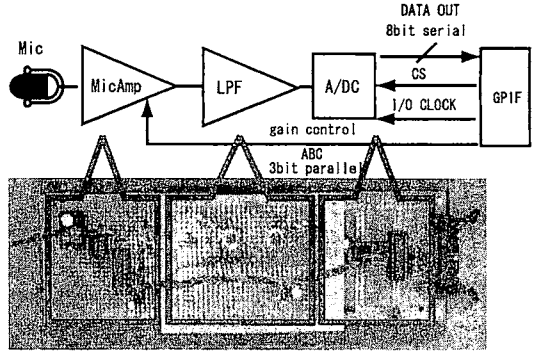

図 7　音取込回路

4.4　離散ウェーブレット変換回路 (hwObject)

アナログ回路から離散値として入力された音に対し,Haar 関数を基底とする離散ウェーブレット変換のスケーリング係数をレベル 1〜4 までを求める回路である.

4.5　畳込み演算回路 (hwObject)

離散ウェーブレット変換後の各レベルのスケーリング係数に対して, 矩形関数による畳込み演算を行う回路である. 矩形関数の幅を 2,4,8 とし, これを各レベルのスケーリング係数に対してそれぞれ畳込みを行う. 各演算結果を周波数チャンネルとし, 製作した回路では, ウェーブレット変換レベル＋ウェーブレット変換レベル× 3 で合計 16 チャンネルとなる. 畳込み演算に

はシフトレジスタを使用することで並列処理を実現し，各窓幅演算の結果が同時に得られるようにした．

この演算によって得られた合計 16 チャンネルの演算結果をローカル・メモリに格納し，ホスト PC からこの演算結果を取得できるようになっている．

離散ウェーブレット回路と畳込み演算回路のブロック図を図 8 に示す．

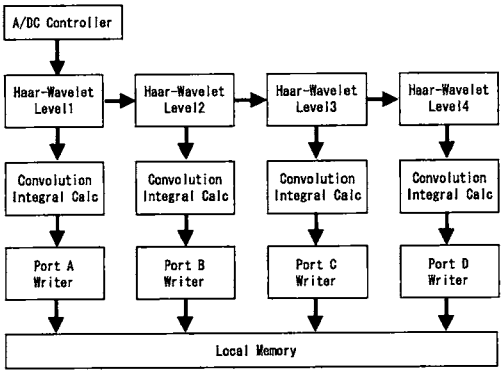

図 8　離散ウェーブレット変換と畳込み回路のブロック図

4.6　ピッチの周期検出回路 (hwObject)

ピッチの周期検出回路は図 9 のような構成である．ピッチの周期検出にはシフトレジスタを用いて，シフトレジスタによるシフトを式 (9) における遅延時間成分 τ に相当させている．そして，各レジスタの値とシフトレジスタへの入力との和を常に求め，ある間隔でシフトしたときの全ての和の値を比較し，最大値を検出することで周期を求めることができる．

しかし，この回路は非常に多量のレジスタ，比較器を使用するため回路規模が非常に大きくなってしまう．

そこで，ピッチの周期は一般に男性のピッチの周期が約 100Hz 付近，女性で約 200Hz 付近であることを考え，検出できる周波数帯を 100Hz から 200Hz 付近に絞り，さらに離散ウェーブレット変換のレベル 4 のスケーリング係数を用いることで回路規模を大幅に削減している．

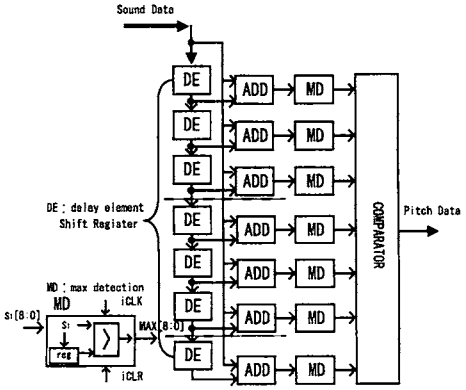

図 9　ピッチ検出回路のブロック図

4.7　hwNet の hwModule 上の構成

今回製作した hwNet の hwModule 上での配置を図 10 示す．

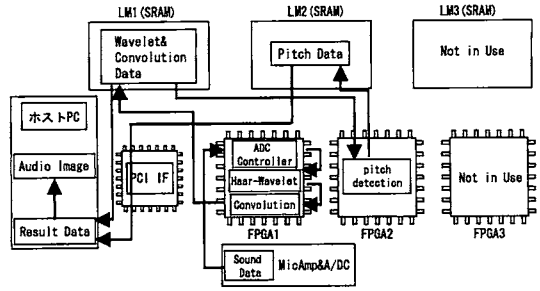

図 10　hwModule 上での hwNet の配置とローカル・メモリの割り当て

FPGA1 に A/D コンバータ部の制御用クロック生成する hwNet と，離散ウェーブレット変換を行う hwNet，さらに離散ウェーブレット変換結果に対して畳込み演算を行う hwNet をダウンロードし，FPGA2 にピッチの周期を検出する hwNet をダウンロードした．今回は,FPGA3 については未使用である．

また，ローカル・メモリ 1 に離散ウェーブレット変換結果に対する畳込み演算結果を書き込み，ローカル・メモリ 2 にピッチの周期データを書き込んだ．今回は，ローカル・メモリ 3 については未使用である．

4.8　ソフトウェア部で行う処理

ホスト PC 側には，これらの値を PCI バスを経由してこれらの値を転送する．ホスト PC 側では，転送されたピッチ周期の値をもとに，畳込み演算後の値を重ね合わせを行い，次に，時間軸での正規化を行う．以上の処理によって音像を得る．

得られた音像をもとに，あらかじめ用意しておいた各母音の音像テンプレートとのマッチングを行う．マッチングの評価方法は式 (10) で示す差分絶対値和を用いて，各チャンネルごとに差分絶対値和の値 $M(k)$ を求め，各チャンネルごとに $M(k)$ が最小になったテンプレートによる多数決をとり，最大多数となった母音を入力された母音とする．

$$M(k) = \sum_{j=0}^{pitch} |S(k+j) - T(j)| \qquad (10)$$

$S(k+j)$ は入力された音像,$T(j)$ はテンプレートの音像

5.　動作検証

5.1　製作した GUI

製作したアプリケーションの GUI を図 11 に示す．本アプリケーションは左上のスタートボタンを押すと，すべての hwNet が動作する仕様となっている．hwNet からの値をもとに得られた音像が画面右中段の画像である．この音像に対して図 12 に示す各母音の音像テンプレートとマッチングを行った結果が画面右中段に表示されている"a"である (この場合は母音の"あ"を入力しておりマッチングは成功している)．画面右下にあるのは，各チャンネルごとの判定結果である．

5.2 母音のマッチング結果

本アプリケーションによる母音のテンプレート・マッチングの結果を示す. マッチングの実験は入力する母音の発話者とテンプレートとする母音の発話者が同一話者の場合と, 異なる話者の場合をそれぞれ行う. 発話者はそれぞれ 20 代男性である.

日本語の母音"あ","い","う","え","お"をそれぞれ 10 回ずつ発話した場合の認識率を表 2 に示す. また,"あ"を入力した場合の各チャンネルの差分絶対値和の値を表 3 に示す.

表 2 母音のマッチング結果

母音	同一話者の場合	異なる話者の場合
"a"	60 %	60 %
"i"	80 %	90 %
"u"	60 %	40 %
"e"	40 %	10 %
"o"	30 %	20 %
認識率	54 %	44 %

表 3 "あ"のマッチング時の差分絶対値和 (成功時, 失敗時)

成功時の各チャンネルの差分絶対値和

	a	i	u	e	o
1	4003	6427	6205	7491	8178
2	3951	6382	6212	7450	8169
3	3880	6261	6095	7367	8109
4	3597	6002	5864	7105	7898
5	2105	3227	3227	3672	4125
6	2057	3169	3237	3594	4099
7	1874	3041	3098	3485	4009
8	1442	2653	2632	2999	3576
9	984	1582	1755	1755	2051
10	880	1526	1717	1674	1944
11	715	1301	1481	1456	1769
12	545	855	1009	1051	1236
13	604	732	892	837	885
14	496	631	796	720	820
15	315	423	609	551	663
16	302	214	325	483	623

失敗時の各チャンネルの差分絶対値和

	a	i	u	e	o
1	6931	5597	6053	7993	8804
2	6912	5595	6033	7983	8780
3	6777	5528	5950	7898	8726
4	6332	5271	5697	7628	8539
5	3516	2934	3068	3927	4356
6	3491	2939	3017	3868	4313
7	3276	2824	2895	3738	4244
8	2640	2533	2598	3389	3996
9	1779	1591	1530	1806	2090
10	1713	1581	1472	1741	1995
11	1401	1447	1367	1578	1843
12	939	1253	1163	1339	1622
13	867	807	803	790	914
14	769	716	743	715	877
15	539	617	641	591	805
16	335	401	482	530	764

5.3 考　　察

提案手法によって得られた音の特徴量, 音像はピッチの 1 周期をほぼ正確に切り出せており, 母音の特徴量であるピッチの周期を回路によって動的に捉えることができている. さらに, 母音マッチングにおいて, 表 3 に示すように, 誤判定時でも, 畳込み演算のチャンネルは正しい判定をしている (チャンネル 12,15,16). これは, 音の特徴量の抽出に畳込み演算が寄与しており, 離散ウェーブレット変換と矩形関数を用いた畳込み演算の組み合わせによる周波数解析が有効であることを意味している.

この提案手法による母音のマッチング精度は表 2 に示すように話者が異なる場合においても認識率が極端に低下しないことが確認できた. マッチングが失敗する原因としては話者の違いよりも, 話者の滑舌やイントネーションの変化によるものが大きいことも確認できた.

これは, あらかじめ選択したテンプレートがこれらの音の変化に対応できていないことが考えられる. 人間が母音だと認識する音の幅はかなり広く, テンプレートもその点も考慮されなければならない.

6. むすび

本研究では, 聴覚機能を担う器官の一つである蝸牛を既存のモデルをもとに, 聴覚機能の前処理部の回路化を行う手法について提案した. 提案手法を hwModule を用いることで実現し, ハードウェアに前処理部の演算を行うことで, 前処理部のみでの CPU 使用率 (CPU:Pentium4 2.6GHz) を 4% としハードウェアによる分散処理を実現した. また, 母音の音像を用いたテンプレートマッチングを行った結果, ある程度の母音の認識が可能であり, 本手法によって音声の特徴量の抽出が可能であることが分かった.

現状のアプリケーションは, 選択する母音の音像テンプレートによって, マッチング精度が変動してしまう問題がある. これは, 多くの音声サンプルを用意し, より頑健性のある音像テンプレートを得ることで, 認識率の変動がある程度抑制されることが期待できる.

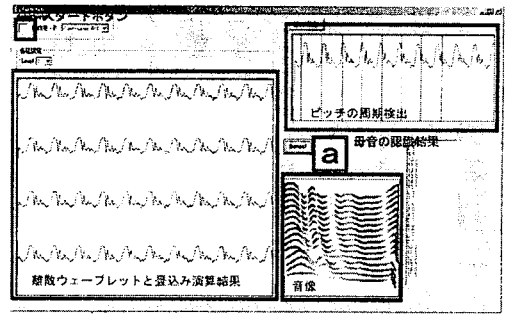

図 11 製作したアプリケーションの GUI

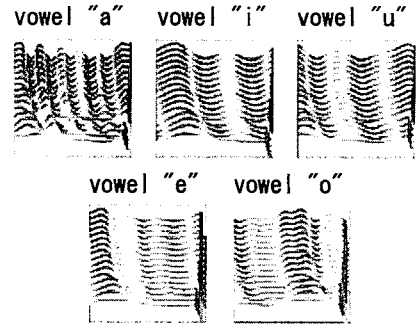

図 12 テンプレートとして利用した各母音の音像

7. 付　　録

本研究で製作した各 hwNet の回路規模と動作速度を表 4 に示す.

表 4 hwNet の回路規模

hwNet 名	周辺回路有無	スライス数	最大動作速度
ウェーブレット	周辺回路有	1404/2352	38.900MHz
畳込み回路	周辺回路無	734/2352	86.415MHz
ピッチ検出	周辺回路有	1622/2352	37.354MHz
回路	周辺回路無	934/2352	37.354MHz

文　　献

[1] ToshioIrino,Roy D.Patterson. 『Segregating information about thesize and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet Mellin transform』.Speech Communication 36, 2002.

[2] 吉井圭吾. 『ウェーブレット変換を用いた音素マッチング処理に関する研究』. 東京農工大学, 関根研究室 H15 年度修士論文.