ロバスト音声認識とその LSI 化について†

早坂 昇*・吉澤 真吾*・和田 直哉*

宮永 喜一*・畑岡 信夫**

A Study of Robust Speech Recognition System and Its LSI Design†

Noboru HAYASAKA*, Shingo YOSHIZAWA*, Naoya WADA*,
Yoshikazu MIYANAGA* and Nobuo HATAOKA**

When a speech recognition system is used in real environment, noises cause a serious problem in the system. In this paper, we propose a method which normalizes feature parameters after filtering the time trajectories of both the power spectra and the log power spectra. As a result of the isolated word recognition experiment where fifteen kinds of noises were added, higher recognition rates were obtained than conventional method by Spectrum Subtraction(SS) and RelAtive SpecTrAl(RASTA). But the proposed method has problems such as processing time and groupe delay of FIR filter. This paper reports on the outline of LSI to solve these problems.

Key Words: filtering, robust, word recognition, LSI

1. ま え が き

近年の音声認識技術の発達に伴い、実用的な音声認識ソフトも現れ、今後さまざまな分野での応用が期待されている。また、携帯電話・カーナビゲーションシステムのような小型機器へ実装する場合、ハードウェア化は処理速度、消費電力、コスト面などにおいて有利であり、その実現への期待が高まっている¹⁾。

音声認識を実環境で行なう場合、背景雑音やマイク特性による歪みの影響により、認識性能が著しく低下してしまうことが報告されている^{1)~6)}。したがって、雑音に対するロバスト技術は不可欠となる。

雑音は音として取り入れられ音声に影響を与える加法的雑音とマイクの周波数特性により伝達特性に影響を与える乗法的歪みに大別される。それぞれに対しさまざまな雑音対策手法が提案されている。

加法的雑音に対する手法として、雑音が付加された音声から雑音のスペクトルを推定しパワースペクトル領域で差し引くスペクトルサブトラクション法(SS法)²⁾をはじめ、雑音

HMMと音声HMMをスペクトル領域で加算し加法的雑音に適應するHMM合成法³⁾などがある。

また、対数スペクトル領域、またはその逆フーリエ変換で与えられるケプストラム領域で乗法的歪みは加法的になる。これを利用しケプストラムの平均値を減算し乗法的歪みを低減するケプストラム平均除去法(CMS法)⁴⁾がある。また、ケプストラムの時間変化にIIR型で設計されたバンドパスフィルタを適用することで音声を強調し乗法的歪みを除去するRelAtive SpecTrAl法(RASTA法)⁵⁾が提案されている。しかし、IIR型フィルタを使用すると位相歪みにより性能が劣化してしまう場合があるため、IIR型ではなくFIR型で設計したフィルタを使用する手法も提案されている⁶⁾。

本論文では、フィルタ特性と認識性能の関係を認識実験を行ない調査する。また、対数スペクトル領域のフィルタリングに加え、パワースペクトル領域でのフィルタリングを新たに提案し、SS法に代わりケプストラムの各次元ごとに振幅最大値で規格化する手法も加えて提案する。しかし、FIR型で急峻な特性を実現するには多大なフィルタ次数が必要になり、演算量の増大に伴う時間遅延やフィルタの群遅延特性は無視できない問題になる。この問題を専用のLSIの実現により解決する⁷⁾。

2. 変調スペクトルと認識性能

2.1 変調スペクトル

音声信号は常に変化している非定常信号である。しかし、100分の1秒程度の短時間区間では比較的定常な信号とみなすことができる。一般的に音声信号は約15~30ms程度に切

† 第3回システムインテグレーション部門講演会で発表(2002・12)

* 北海道大学大学院情報科学研究科 札幌市北区北14条西9丁目

** (株)日立製作所中央研究所 国分寺市東恋ヶ窪1-280

* Graduate school of engineering, Hokkaido University

** Central reserch laboratory, Hitachi, Ltd.

(Received April 28, 2004)

(Revised January 13, 2005)

Fig.1 Modulation spectra of clean speech data for power spectra

Fig.2 Modulation spectra of noisy speech data for power spectra

り出され、約 5~15ms づつシフトしながらスペクトル分析され特徴量に変換される。このシフト幅ごとに求められる特徴量やスペクトルの時系列データを周波数領域で表現したものが変調スペクトルである。

$$C_n(f) = \int_{-\infty}^{\infty} c_n(t) e^{-j2\pi ft} dt \quad (1)$$

ここで $C_n(f)$ は特徴量の変調スペクトル、 $c_n(t)$ は特徴量を表わし、 n, t は n 次元目の特徴量の t フレーム目を表わす。

2.2 変調スペクトルと雑音

加法性雑音を $a(k)$ 、乗法性歪みを $h(k)$ とし、発声された音声信号を $s(k)$ とすると認識システムに入力される信号 $x(k)$ は次式ようになる。

$$x(k) = \{s(k) + a(k)\} \otimes h(k) \quad (2)$$

これがあるフレーム幅、シフト幅でスペクトル分析が行なわれ周波数領域では、

$$X_i(t) = \{S_i(t) + A_i(t)\} H_i(t) \quad (3)$$

となる。ただし、 $X_i(t)$ 、 $S_i(t)$ はそれぞれ入力信号、音声信号のスペクトル、 $A_i(t)$ 、 $H_i(t)$ はそれぞれ加法性雑音、乗法

Fig.3 Comparison between modulation spectra without noise (solid line) and with white noise (dotted line) at about 1 kHz

性歪みのスペクトルである。ここでの i, t は t フレーム目のスペクトルの i 番目のポイントを表わしている。これらをパワースペクトル上で表現すると、

$$\begin{aligned} |X_i(t)|^2 &= |\{S_i(t) + A_i(t)\} H_i(t)|^2 \\ &= |S_i(t) H_i(t)|^2 + |A_i(t) H_i(t)|^2 \\ &\quad + 2|H_i(t)|^2 |S_i(t)| |A_i(t)| \cos(\theta_i(t)) \end{aligned} \quad (4)$$

となる。(4) 式における $\theta_i(t)$ は音声信号 $S_i(t)$ と加法性雑音 $A_i(t)$ との位相差である。

雑音を付加していない状態におけるパワースペクトルの変調スペクトルを Fig. 1 に示す。横軸は変調周波数、縦軸が周波数、色の濃淡がエネルギーの強さを表わしている。Fig. 1 に示すように、変調周波数の約 10Hz 以下の部分は特に強くなっている。この部分は音節の変化の周期などに対応していると言われており、特に約 4Hz 付近は最も重要であるとされている^{6), 8)}。これに白色雑音 5dB を付加すると Fig. 2 のようになる。雑音により Fig. 1 で見られたエネルギーの強弱が目立たなくなっているのがわかる。ここで Fig. 1, Fig. 2 の約 1kHz の部分の変調スペクトルを取り出し比較する (Fig. 3)。雑音を付加することにより変調周波数 0Hz 付近の成分が大幅に増大している。これは、加法性雑音 $|A_i(t) H_i(t)|^2$ は音声に比べ定常的でありその影響が集中したためであると考えられる。また、加法性雑音の影響は (4) 式の第 3 項 $2|H_i(t)|^2 |S_i(t)| |A_i(t)| \cos(\theta_i(t))$ にも含まれている。観測時間が無限であれば音声と雑音には相関が無いので平均的にこの項は 0 と見なすことができるが、実際には観測時間は有限であるため、この項の影響は無視できないものになり、0Hz 付近以外に現れている雑音の影響はこの項が原因であると考えられる。

つぎに、ケプストラムに変換するため (4) 式のと対数を求めると次式になる。

$$\log |X_i(t)| = \log |S_i(t) + A_i(t)| + \log |H_i(t)| \quad (5)$$

対数変換後の無雑音、白色雑音付加それぞれの状態におけ

Fig. 4 Modulation spectra of clean speech data for log-power spectra

Fig. 5 Modulation spectra of noisy speech data for log-power spectra

る変調スペクトルをそれぞれ Fig. 4, Fig. 5 に示す。雑音を加えてもエネルギーの分布はほとんど変化がなく、どちらの図においても変調周波数約 10Hz 以下にエネルギーの大部分が存在している。また、0Hz 付近にエネルギーが著しく強く現れているが、これは、対数変換により加算の形に変換されたマイクの伝達特性等による時間変化がほとんどない定常的な乗法性歪み $H_i(t)$ であると考えられる。

3. 頑健な特徴抽出法

変調スペクトルの特徴をまとめると以下の通りとなる。

- 認識に有効な音声情報の多くはどちらの変調スペクトルにおいても変調周波数約 10Hz 以下にあり、約 4Hz 付近は音節の変化の周期に対応しているため特に重要である。
- 加法的雑音は音声に比べ定常的であるためパワースペクトルに対する変調スペクトルの低域 (0Hz 付近) に多く存在する。
- 相互相関項 $2|H_i(t)|^2|S_i(t)||A_i(t)|\cos(\theta_i(t))$ により、0Hz 付近だけではなく帯域全体にその影響がみられ、雑音が無い状態における変調スペクトルに比べ強弱が目立たなくなってしまう。

- 乗法性の歪みは対数領域では加法的になり、時間変化が少ないため、対数パワースペクトルに対する変調スペクトルの低域に集中する。

これらの特徴を考慮し、変調スペクトルを効果的に操作することで認識性能が向上すると考えられる。SS 法はパワースペクトルに対する変調スペクトルの直流成分を低減し加法的雑音を抑える手法であり、CMS 法は対数スペクトル (ケプストラム) に対する変調スペクトルの直流成分を取り除き乗法性歪みを抑える手法である。また、RASTA 法は 0Hz 近傍のみではなく 10Hz 以上の成分も抑制し音声を強調する手法である。このように、広く用いられている耐雑音手法の多くは変調スペクトルに対し効果的な処理を施している。

3.1 対数スペクトル領域フィルタリングに用いるフィルタの特性と認識性能

対数スペクトル領域フィルタリングについては RASTA 法⁵⁾が提案されているが、位相歪みの影響を受け、波形が歪んでしまい認識率が低下する。そのため、FIR 型フィルタを用いる手法が提案されている⁶⁾。フィルタ次数が高くなると演算量や群遅延などにより処理が遅くなってしまうため、従来はより少ない次数で実現することが望まれていた。しかし、本論文後半で紹介する専用ハードウェアの実現により高い次数のフィルタを用いても実時間処理が可能であるため、フィルタの特性を急峻にすることで認識性能がどう変化するかを認識実験により調べた。学習条件、分析条件、認識に用いた音声データは 4. での条件と同じで、SNR 比ごとにすべての認識率の平均値で比較した。結果を Table 1 に示す。BP 以下の数字はフィルタの次数を表わし大きいほど急峻な特性である。RASTA 法や CMS 法に比べ FIR 型フィルタを用いるほうが良い結果が得られた。RASTA 法は位相歪みの影響を受けているため性能が劣化していると考えられ、CMS 法は直流成分は除去しているが、RASTA 法のように高域部分を低減していないため十分な認識性能が得られていないと考えられる。また、フィルタ次数と認識率の関係については 164 次に比べ 200 次のほうが性能がよく、それより高い次数になるとほとんど性能が変わらないが、200 次以上の高い次数が必要であることがわかった。

3.2 パワースペクトル領域フィルタリング

対数スペクトル領域のフィルタリングでは乗法性歪みの影響は抑えることが出来るが、加法的雑音の影響を低減することはほとんど出来ない。そこで、加法的雑音の影響が著しいパワースペクトル領域フィルタリングを提案する。パワースペクトル領域の変調スペクトルと雑音の関係から、雑音が多く集まっている直流成分の抑制と相互相関項の影響により

Table 1 Recognition results to variety log-domain filtering methods

SNR	BP_300	BP_240	BP_200	BP_164	RASTA	CMS	BASELINE
20dB	97.3	97.3	97.3	97.2	96.8	97.3	95.5
10dB	76.2	76.4	76.2	75.5	69.4	71.8	65.1
0dB	25.9	25.8	25.8	24.6	21.7	19.6	19.8

Fig. 6 Comparison between the 4th MFCC without noise (solid line) and with white noise (dotted line) before normalization

Fig. 7 Comparison between the 4th MFCC without noise (solid line) and with white noise (dotted line) after normalization

埋没した音声成分を強調するため、バンドパスフィルタの適用が最も妥当であると考えられる。しかし、直流成分を阻止域にするようなフィルタのフィルタ係数には必ず負の値が含まれ、フィルタリング後のスペクトルが負になってしまう場合がある。広く用いられているSS法も推定した雑音を減算するときにスペクトルが負になるため、1以下の係数を乗じ負になるのを防いでいる。後述する規格化により直流部分の雑音の影響を補正するため、ここでは直流成分は保持し、高域側のみ抑制する正の係数のみで構成されるローパスフィルタを適用し、音声の強調を行なう。なお、このフィルタはMATLAB FDATool⁹⁾を用い、フィルタ係数が負にならないことを確認しながら特性を変化させ設計した。

3.3 ケプストラム振幅最大値規格化

2回のフィルタリングで認識に重要な部分の強調並びに乗法性歪みの影響を取り除いた。しかし、パワースペクトルに対する変調スペクトルの直流成分は低減していないため、加法的雑音の影響のほとんどは残ったままになっている。無雑音と白色雑音に加わったフィルタリング後のケプストラム4次の波形をFig. 6に示す。雑音が付加されることにより、ダイ

Fig. 8 The filter characteristic for power spectra

Fig. 9 The filter characteristic for log-power spectra

ナミックレンジの差が大きくなっているのがわかる。われわれの実験調査の結果このダイナミックレンジの差が加法的雑音混入時の認識率の低下を招いていることがわかった^{10), 11)}。これを補正するためケプストラムの各次元ごとに振幅最大値で規格化を行なう。

$$c'_n(t) = c_n(t) / \max |c_n(t)| \quad (6)$$

$c'_n(t)$ は規格化後のケプストラム、 $c_n(t)$ は規格化前のケプストラムである。Fig. 7は規格化後のケプストラム4次の波形であるが、Fig. 6に見られたダイナミックレンジの差が補正され類似度が高くなっている。

以上、2度のフィルタリングと規格化により、加法的雑音、乗法性歪みそれぞれの影響を抑え、かつ音声成分を強調することで高い認識率が得られると考えられる。それぞれの領域で用いたフィルタ特性をFig. 8, Fig. 9に示す。これらのフィルタ特性は認識実験を行なった結果最も認識率の高かったものを採用している。

4. 認識実験

4.1 実験条件

提案法を評価するため孤立単語音声認識実験を行なった。音声データは電子協データベース日本の地名100単語を用い、

Table 2 Analysis conditions

Window length	23.2ms (256point)
Frame shift	11.6ms (128point)
Windowing	Hanning window
Preemphasis	$1 - 0.97z^{-1}$
Feature vectors	MFCC 12, Δ MFCC 12, $\Delta\Delta$ MFCC 12 Δ energy 1, $\Delta\Delta$ energy 1

サンプリング周波数は 11.025kHz にダウンサンプリングした。学習には男性話者 40 名 1 人 3 回発声分を用い、認識には学習に用いなかった男性話者 10 名 1 人 1 回発声分を用いた。雑音は NOISEX-92 にある 15 種類を SNR0, 10, 20dB の 3 種類の大きさで付加し、結果は SNR, 雑音ごとに平均値で比較した。また、無雑音環境における認識性能についても実験を行なった。

本実験の分析条件と用いた音響特徴量を Table 2 に、提案手法を組み込んだシステムの流れを Fig. 10 に示す。認識には HMM 法を用い、32 状態 1 混合の単語モデルを採用した。

実験は雑音対策を行わないもの (BASELINE), 現在広く用いられている SS 法と RASTA 法を組み合わせさせた手法 (SS+RASTA) と提案手法 (PROPOSED) で行なった。比較対象である SS 法は相互相関項 $2|H_i(t)|^2|S_i(t)||A_i(t)|\cos(\theta_i(t))$ が 0 であると仮定し、つぎの式のような処理を行なう。

$$|X'_i(t)|^2 = \begin{cases} |X_i(t)|^2 - \alpha|\hat{N}_i(t)|^2 & |X_i(t)|^2 > \alpha|\hat{N}_i(t)|^2 \\ \beta|X_i(t)| & \text{else} \end{cases} \quad (7)$$

$\hat{N}_i(t)^2$ は事前に推定した雑音, $|X'_i(t)|$ は SS 法適用後の観測信号であり, α は過剰推定係数, β はフロアリング係数である。SNR や雑音種によらず, $\alpha = 1.5$, $\beta = 0.1$ とした。

Fig. 10 The overview of the analysis process

Table 3 Recognition rates for clean speech

SS+RASTA	PROPOSED	BASELINE
98.5	98.9	99.4

4.2 実験結果

Table 3, Table 4 に認識結果を示す。無雑音環境下における認識性能は, SS+RASTA, PROPOSED 共に BASELINE に比べ若干の性能の低下が見られた。しかし, PROPOSED は劣化が少なく SS+RASTA に比べ良好な結果を示した。また, 雑音を付加した実験においては雑音対策を行なった結果, 低 SNR においては著しく認識率が向上しているが, PROPOSED は SS+RASTA よりもさらに高い認識率を得ている。また, 20dB 程度の高 SNR では SS+RASTA を適用すると性能が大きく劣化する雑音種が存在するのに対し, PROPOSED はほとんど劣化することがなく高い認識率を示している。SS+RASTA が性能を落としてしまう原因のひとつとして, SS 法にある α や β が適切に与えられていないことが考えられる。これらの値は SNR により最適値が変わることが知られているが, 一般に SNR を知ることは困難であるため, 適応的にその値を変えることができない。一方, 提案手法は α や β のようなパラメータに依存したモデルではないため, 安定した認識が可能であり, 高 SNR においても高い認識率を保っている。しかし, 提案法は雑音がある程度定常的であることを仮定しているため, 突発的な雑音には効果が少ない。今後の課題として検討していきたい。

5. ロバスト認識システムの LSI 実現

5.1 ロバスト処理回路

3. ではパワースペクトルおよび対数スペクトルの時系列に対してフィルタリングを行なう手法を述べたが, 実際の認識システムの動作において, 計算機シミュレーションでは現れなかったフィルタの群遅延特性に関する問題が発生する。パワースペクトルに対するフィルタリングを定式化して説明する。伝達関数 $H(z)$ の FIR フィルタに対して, ある周波数でのスペクトル時系列の入力を $X(k)$, 出力を $Y(k)$ としたときの数式は以下の通りである。

$$Y(k) = \sum_{i=0}^Q h(i)X(k-i) \quad (8)$$

ここで, k はフレーム時間, $h(i)$ は FIR フィルタ係数, Q は FIR フィルタ次数である。完全直線位相特性を有する FIR フィルタはフィルタ係数が対称形であるためフィルタ次数の半分の群遅延特性をもつ。Fig. 11 は FIR フィルタの入力と出力の関係を示している。サンプリング間隔を T_s とした時, フィルタの入力と出力の間には $(Q/2)T_s$ の時間遅延が発生する。この時間遅延は図の AB 間の間隔に該当する。音声信号に対するフィルタリングではサンプリング時間は $1\mu\text{s}$ 以下 (サンプリング周波数 11.025kHz に対して $T_s = 9.0 \times 10^{-5}\text{s}$) と群遅延特性による影響は少ないが, スペクトルの時系列

Table 4 Recognition results

Noise Name	SNR	BASELINE			SS+RASTA			PROPOSED		
		0dB	10dB	20dB	0dB	10dB	20dB	0dB	10dB	20dB
White noise		0.8	33.2	92.4	41.4	90.3	98.2	47.9	85.6	97.5
Pink noise		0.7	53.5	96.9	47.3	94.7	98.9	53.3	91.5	97.6
HF channel noise		3.3	26.6	82.1	33.2	78.5	87.3	49.3	85.1	95.5
Speech babble		7.8	58.2	91.2	5.9	30.6	61.8	20.2	77.0	95.7
Factory floor noise 1		2.0	62.9	97.4	34.0	89.5	97.2	33.4	88.5	98.4
Factory floor noise 2		15.0	80.0	98.1	70.9	94.3	97.6	75.2	96.3	98.1
Jet cockpit noise 1		0.9	49.4	95.5	31.8	88.8	96.7	45.5	88.5	97.9
Jet cockpit noise 2		0.2	41.4	95.5	44.4	91.7	98.8	51.7	90.0	97.7
Destroyer engine room noise		3.8	55.6	94.7	50.5	92.0	96.8	61.0	90.7	97.0
Destroyer operation room noise		4.1	76.1	98.5	48.2	87.6	96.2	58.9	94.7	98.5
F-16 cockpit noise		1.7	69.4	97.3	50.2	94.1	98.2	56.2	92.1	97.8
Military vehicle noise		52.9	87.5	98.4	78.9	89.4	96.4	88.4	94.4	97.5
Tank noise		29.7	91.2	98.5	81.9	97.4	98.9	80.7	97.7	98.7
Machine gun noise		80.1	92.6	97.5	72.0	80.5	87.9	84.6	92.5	96.8
Car interior noise		93.9	99.0	99.2	95.0	98.2	99.1	97.0	98.5	98.9
Average		19.8	65.1	95.5	52.4	86.5	94.0	60.2	90.9	97.6

に対するフィルタリングではサンプリング間隔はフレーム間隔と同じ 10ms 以上であるため大幅な時間遅延が生じる。たとえばフレーム間隔 11.6ms、フィルタ次数 240 の場合には $(Q/2) \times T_s = 1.392s$ となる。A 地点は音声認識システムに入力する発話が終了する時刻であり、その時刻からフィルタリング処理が完了するのは B 地点の時刻であるので、発話終了後からのシステム応答時間が 1 秒以上と長時間になってしまう問題点をもつ。

スペクトル時系列に対して直接フィルタ計算を行なうと前述の問題が発生する。そこで発話区間内のスペクトル時系列データをバッファリングしておいて、不足するデータをバッファリングしておいた一部のデータから補完するフレーム拡張方式を適用する。Fig. 12 はその方式の手順を示している。スペクトル時系列データでのバッファリング対象のフレームは発話区間検出によって決定される。発話開始前のフレーム時刻と発話終了後のフレーム時刻 (Fig. 11 の A 地点に一致) からその間のフレーム長 T を決定する。このとき切り出されたスペクトルを $X(1), X(2), \dots, X(T)$ と与える (Fig. 12(a))。発話検出開始直後と終了直前のフレームは音声信号では無音に該当する部分であるのでこのフレームに該当するデータを用いて発話検出区間外のデータを補完しても影響が少ない。そこで Fig. 12(a) のフレームの前後の補完データをフィルタ次数の半分ずつ並べてフレームを拡張する (Fig. 12(b))。拡張後のスペクトルを $E(1), E(2), \dots, E(T+Q)$ と再構築したフレーム拡張方式での計算式は以下の通りとなる。

$$Y(k) = \sum_{i=0}^Q h(i)E(Q+k-i) \quad \text{for } 1 \leq k \leq T \quad (9)$$

フレーム拡張方式を用いない場合、群遅延により発話後から 1 秒以上経過しなければ応答が得られなかったのに対し、この手法を導入することで、群遅延問題が解消され応答時間は

Fig. 11 Relation between time delay and group delay of Linear-Phase FIR filter

Fig. 12 The frame expanding method

Fig. 13 The circuit of the robust process

Table 5 Comparison of power consumption

	Proposed Architecture	DSP
Specification	0.18 μ m CMOS ASIC 340K gates	TMS320C6203B fixed-point DSP
Operating Frequency	25MHz (maximum 128MHz)	300MHz / 2400MIPS
Voltage Supply(V)	1.8	1.7
Power Consumption(mW)	136.1	867

計算時間のみとなる。また、フレーム拡張方式によって得られたフィルタリング後の波形とフレーム拡張方式を用いない場合のフィルタリング後の波形を同じポイント数で比較したところほとんど差が無いことから、認識性能にも大きな影響は無いと考えられる。ただし、Fig. 11 の A 地点の時刻からフィルタ計算を一括で処理するため、直接計算方式よりも高い演算処理能力が要求されるため、FIR 計算専用回路で処理を行なうのが有効である。Fig. 13 はロバスト処理回路の構成である。固定小数点演算の加算、乗算器と 2 個のメモリで構成される。Memory1 は FIR フィルタ係数の参照、Memory2 は入力 $E(k)$ の参照と出力 $Y(k)$ の記憶に使用される。図中にある Address Generator はフレーム拡張前の入力 $X(k)$ から拡張後の入力 $E(k)$ へ写像するためのインデックス計算を行なう。

5.2 音声認識トータルシステムと検証ボード

Fig. 14 は音声認識トータルシステムの構成である. 前に

述べたロバスト処理回路を含むモジュールのほかに、音声分析 (MFCC)、音声認識 (HMM) の専用計算モジュールを実装した。ロバスト処理回路の計算時間が最大 20ms、認識計算時間が最大 160ms であり、発話区間の検出終了後からの応答時間が最大 180ms 以内となるように設計されている。認識計算はシステムの中で最も高い計算処理能力を必要とし、この計算モジュール内でパイプライン処理と並列分散処理を実装した。Table 5 は専用回路と DSP の比較において同程度の処理能力、電源電圧の条件下で 800 単語認識タスクの認識計算処理を実行させた場合での想定される消費電力を示している。専用ハードウェアの消費電力値は DSP に比べて約 1/6 と大幅な削減が期待できる。

Fig. 15 は音声認識トータルシステムの動作を検証するための FPGA(Field Programmable Gate Array) を用いた検証ボードである。三菱電機 PowerMedusa を用いて検証ボードを構築した。音声認識トータルシステムに加えて A/D 変換器、マイクロフォン、メモリ、PC インターフェイスが接続される。10MHz のクロック周波数で動作し、スイッチを押すことにより発話区間の開始と終了のタイミングを決め、約 0.3 秒の無音区間が発話前後にあるようなタイミングでマイクを通し発話した。A/D 変換は 12bit 量子化、サンプリング周波数 11.025kHz で行なった。認識結果は該当する単語を示す番号が LED に表示される。この検証ボードを用いて単語音声認識実験を行ない計算機シミュレーションとの性能(同じ位の SNR 環境下)を比較した結果、ほぼ同等の認識性能を示し、量子化ビットの変更や固定小数点演算による精度の低下に伴う性能の劣化がほとんど無いことが確認された。

6. む す び¹¹

本論文では、対数スペクトル領域フィルタリングに用いる FIR 型フィルタの特性と認識性能について考察し、200 次以上の高い次数が必要であることがわかった。また新たに、パワースペクトル領域フィルタリング、ケプストラム振幅最大値規格化法を提案し、その効果を孤立単語音声認識実験を行ない検証した。従来法である SS+RASTA 法と比較し提案手法は安定した高い認識率を得た。しかし、フィルタリング時に起こる群遅延や演算量の問題をフレーム拡張方式、専用 LSI の実現により解消した。FPGA を用いた検証の結果、応答時間は発話終了から最大 180ms 以内であり、認識性能は演算精度の低下による劣化はほとんどなくコンピュータシミュレーションの結果とほぼ同等であることが確認できた。しかし、雑音環境下における認識精度は人間に比べ依然として低く、さらなる改善が要求され、新たな雑音ロバスト手法について検討して行きたい。

謝辞 本研究に関して多くのご助言を頂いた山武開発本部の皆様へ深く感謝いたします。また、本研究の一部は半導体理工学センター (STARC) の援助による。

参 考 文 献

- 1) 吉沢, 宮永, 吉田: 一括並列処理による HMM 高速化手法及びその VLSI 設計, 電子情報通信学会論文誌, **J85-A-12**, 1440/1450 (2002)
- 2) S.F.Boll: Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. on Speech and Signal Processing, **ASSP-27-2**, 113/120 (1979)
- 3) M.J.F.Gales and S.J.Young: Robust continuous speech recognition using prallel model combination, IEEE Trans. on Speech and Audio Processing, **4**, 352/359 (1996)
- 4) B.S.Atal: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, J.Acoust. Soc. Am., **55**, 1304/1312 (1974)
- 5) H.Hermansky and N.Morgan: RASTA processing of speech, IEEE Trans. Speech and Audio Processing, **2**, 578/589 (1994)
- 6) 金寺, 荒井, 船田: 変調スペクトルの重要な成分のみを選択的に用いた雑音に強い音声認識, 電子情報通信学会論文誌, **J84-D-II-7**, 1261/1269 (2001)
- 7) 早坂, 和田, 吉沢, 宮沢: ロバスト音声認識とその LSI 化について, 計測自動制御学会システムインテグレーション部門講演会, 講演論文集 (II), 67/68 (2002)
- 8) 中川聖一: 音声認識研究の動向, 電子情報通信学会論文誌, **J83-D-II-2**, 433/457 (2000)
- 9) フィルタ設計と解析の GUI マニュアル, サイバネットシステム (株), <http://www.cybernet.co.jp/matlab/support/helpdesk/r14/toolbox/signal/fdtool.shtml>
- 10) Naoya Wada, Yoshikazu Miyanaga, Norinobu Yoshida and Shingo Yoshizawa: A consideration about an extraction of features for isolated word speech recognition in noisy environments, ISPACS2002, DSP2002-33, 19/22 (2002)
- 11) 吉沢, 和田, 早坂, 宮永: RSF/DRA 手法を用いた雑音ロバスト音声認識システムのハードウェア実現, 電子情報通信学会技術研究報告, CAS2003-22, 127/132 (2003)

[著 者 紹 介]

早 坂 昇 (学生会員)

2002年北海道大学工学部電子工学科卒業。2004年同大学大学院工学研究科修士課程了。同年同大学大学院情報科学研究科博士課程入学, 現在に至る。音声認識の研究に従事。

吉 澤 真 吾

2001年北海道大学工学部電子工学科卒業。2003年同大学大学院工学研究科修士課程了。同年同研究科博士課程入学, 現在に至る。音声認識, 無線通信, デジタル信号処理の VLSI 化に関する研究に従事。

和 田 直 哉

2001年北海道大学工学部電子工学科卒業。2003年同大学大学院工学研究科修士課程了。同年同研究科博士課程入学, 現在に至る。音声信号処理並びに音声認識の研究に従事。

宮 永 喜 一

1979年北海道大学工学部電子工学科卒業。81年同大学大学院工学研究科修士課程了。84年米国イリノイ大学客員研究員。87年北海道大学工学部電子工学科講師。88年同助教授。現在, 北大学院情報科学研究科教授。主として, 並列信号処理, 並列計算機アーキテクチャ, 適応信号処理の研究に従事。工博。

畑 岡 信 夫

1976年東北大学工学部通信工学科卒業。78年同大学大学院工学研究科修士課程了。同年日立製作所中央研究所入所。88年米国カーネギーメロン大学客員研究員。89年日立ヨーロッパ社ダブリン研究所長。93年日立製作所中央研究所。現在, 同研究所主管研究員。主として, 音声認識の研究に従事。工博。