音声言語情報処理 29－13
(1999. 12. 20)

カルマンフィルタによる雑音除去法を用いた雑音環境下での音声認識

藤本 雅清　　有木 康雄

龍谷大学 理工学部

〒 520-2194 大津市瀬田大江町横谷1-5
Tel: 077-543-7427
E-mail: masa@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

あらまし 本報告では、雑音環境下音声認識の前処理を想定したカルマンフィルタによる雑音除去法を提案する。従来、カルマンフィルタは膨大な計算量を要するため、実時間向けの処理にはあまり使用されていなかった。そこで本研究では、大幅に計算量を削減した高速カルマンフィルタを用いることにより、ほとんど精度を低下させることなく実時間の1.5~2.0倍での処理を実現した。提案手法の評価のために雑音重畳音声から抽出されたクリーン音声を用いて単語認識実験を行い、従来のSpectral Subtraction法及びParallel Model Combination(PMC)法との単語認識精度の比較を行った。その結果、提案手法により雑音によってはPMC法と同等かそれ以上の単語認識精度を得ることができた。

キーワード 雑音環境下での音声認識、雑音除去、高速カルマンフィルタ、実時間向け処理

Noisy Speech Recognition Using Noise Reduction Method Based on Kalman Filter

Masakiyo Fujimoto　　Yasuo Ariki

Faculty of Science and Technology, Ryukoku University

1-5 Yokotani, Oe-cho, Seta, Otsu-shi, 520-2194 Japan
Tel: +81-77-543-7427
E-mail: masa@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

Abstract In this paper, we propose a noise reduction method based on Kalman filter for noisy speech recognition. Since Kalman filter needs a huge quantity of computation, it was never used for real time processing. We propose a noise reduction method using fast Kalman filter which can reduce a large quantity of computation and achieve processing in 1.5~2.0 times of real time, without loosing the accuracy. In order to evaluate the proposed method, we carried out experiments to extract clean speech signal from noisy speech and compared the results by our method with conventional Spectral Subtraction and Parallel Model Combination(PMC) in word recognition accuracy. As a result, the proposed method obtained word recognition rate equal or superior to PMC.

key words noisy speech recognition, noise reduction, fast Kalman Filter, real time processing

1 はじめに

ここ数年、数多くの音声認識手法が提案されており、また、音声認識システムを実装したソフトウェア、家電製品等が実際に商品化され、音声認識の実用化が進められている。しかし、それらの多くは比較的静かな環境を想定したものが大半を占めており、実環境で背景雑音の影響が大きい場合、認識率が極端に低下してしまうという問題があり、完全な実用化には至っていないのが現状である。

実環境を想定した音声認識システムとして、Parallel Model Combination(PMC)法[1, 2]のようにシステムを雑音に適応させる手法が提案されているが、環境がダイナミックに変化すると、背景雑音の学習をやり直さなければならないという問題がある。そこで本研究では、システムを雑音に適応させる手法とは逆に、雑音を除去した後に音声認識を行う手法を提案する。

雑音除去の従来手法として、Spectral Subtraction(SS)法があげられるが、SS法では雑音スペクトルの減算の際に、減算が足らずに雑音成分を残してしまったり、減算しすぎて目的とする音声のスペクトルが歪んでしまい、その結果認識率の低下を招くという問題がある。その問題を解決するために、2波形分離モデル[3, 4, 5]の考えを基に、カルマンフィルタにより音声信号の推定を行い、分離抽出された音声信号を用いて音声認識を行った。

また、我々はすでにカルマンフィルタを用いた雑音除去法[6]を提案しているが、処理に膨大な時間がかかり、雑音環境下における音声認識の前処理として利用するには問題があった。そこで、本研究ではカルマンフィルタの高速化を行い、実時間向けの雑音除去法について検討する。

以下、2章では提案する雑音除去法について述べ、3章では提案手法を用いた実験とその評価について述べる。

2 雑音除去手法

2.1 カルマンフィルタ

次の2式のように定義される、有限次元の線形システム

$$x_{k+1} = F_k x_k + G_k w_k \tag{1}$$

$$y_k = H_k x_k + v_k \tag{2}$$

において、w_k, v_kがそれぞれ独立な白色性ガウス雑音とし、F_k, G_k, H_k, Σ_{vk}, Σ_{wk} が既知であるとする(Σ_{vk}, Σ_{wk}はそれぞれv_k, w_kの共分散行列を表す)。この時、観測データ$\{y_0, y_1, \cdots, y_k\}$が与えられた時の最小分散推定量$\hat{x}_{k|k}$, $\hat{x}_{k|k-1}$を求める。この問題をカルマンフィルタリング問題といい、その解を与えるアルゴリズムをカルマンフィルタという[7, 8]。ここで、$\hat{x}_{k|k}$は時間kでのx_kの推定値、$\hat{x}_{k|k-1}$は時間$k-1$でのx_kの予測値である。また、v_kは観測雑音、w_kはシステム雑音と呼ばれ、式(1)は状態方程式、式(2)は観測方程式と呼ばれている。以下、式(3)～(9)にカルマンフィルタの定義式を、図1にカルマンフィルタの処理フローを、図2にカルマンフィルタのブロック図を示す。

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k \left(y_k - H_k \hat{x}_{k|k-1}\right) \tag{3}$$

$$\hat{x}_{k+1|k} = F_k \hat{x}_{k|k} \tag{4}$$

$$K_k = \hat{\Sigma}_{k|k-1} H_k^T \left[H_k \hat{\Sigma}_{k|k-1} H_k^T + \Sigma_{vk}\right]^{-1} \tag{5}$$

$$\hat{\Sigma}_{k|k} = \hat{\Sigma}_{k|k-1} - K_k H_k \hat{\Sigma}_{k|k-1} \tag{6}$$

$$\hat{\Sigma}_{k+1|k} = F_k \hat{\Sigma}_{k|k} F_k^T + G_k \Sigma_{wk} G_k^T \tag{7}$$

$$\hat{x}_{0|-1} = \overline{x}_0 \tag{8}$$

$$\hat{\Sigma}_{0|-1} = \Sigma_{x0} \tag{9}$$

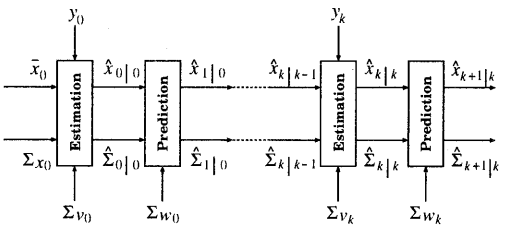

図 1: カルマンフィルタの処理フロー

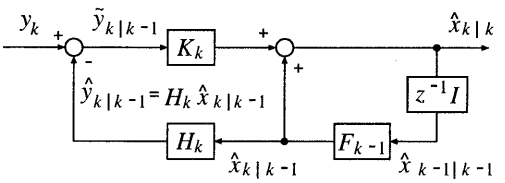

図 2: カルマンフィルタのブロック図

式(3), (5), (6)に式(8), (9)に示す初期値($\overline{x}_0$はx_0の平均値、Σ_{x0}はx_0 の共分散行列)を与えてやればx_kの推定値$\hat{x}_{k|k}$が逐次求まる。

$\overline{x}_0, \Sigma_{x0}$が未知の場合は$\hat{x}_{0|-1} = 0, \hat{\Sigma}_{0|-1} = \alpha_i \delta_{ij}$ (α_iは任意の定数、δ_{ij}はクロネッカーのデルタ) を与えてやればいいが、双方とも真値に近いほうが適切な推定値が得られる。

2.2 カルマンフィルタの適用

窓関数により切り出されたl番目の短時間フレーム内において、クリーン音声の複素スペクトルを$S(f,l)$、クリーン音声を$s(k,l)$、雑音を$v(k,l)$とすると雑音重畳音声$y(k,l)$は

$$
\begin{aligned}
y(k,l) &= s(k,l)+v(k,l) \\
&= \sum_{f=0}^{N-1} S(f,l)\exp\left(j2\pi\frac{fk}{N}\right)+v(k,l) \\
&= \underbrace{\begin{pmatrix} 1 \\ \exp\left(j2\pi\frac{k}{N}\right) \\ \vdots \\ \exp\left(j2\pi\frac{(N-1)k}{N}\right) \end{pmatrix}^{T}}_{H_k} \underbrace{\begin{pmatrix} S(0,l) \\ S(1,l) \\ \vdots \\ S(N-1,l) \end{pmatrix}}_{x_l} \\
&\quad +v(k,l) \\
&= H_k x_l + v(k,l) \qquad (10)
\end{aligned}
$$

と表される(k, Nはそれぞれl番目のフレーム内での時間、サンプル点数を表す)。ここでx_lはl番目のフレームの複素スペクトルを要素に持つベクトルであり、フレーム内での時間kでは時間不変なので、

$$x_l = x(k,l) = x(k+1,l) \qquad (11)$$

と表すことができる。また、$y(k,l)$は式(10)より

$$y(k,l) = H_k x(k,l) + v(k,l) \qquad (12)$$

と表すことができる。ここで式(11)を状態方程式、式(12)を観測方程式とすると、式(11)より$F_k = I$となり、これをカルマンフィルタの定義式である式(3)～(7)に適用すると最終的に以下のように簡略化される。

$$
\begin{aligned}
\hat{x}_{(k,l)|(k,l)} &= \hat{x}_{(k-1,l)|(k-1,l)} \qquad (13) \\
&\quad +K_{(k,l)}\left(y(k,l)-H_k\hat{x}_{(k-1,l)|(k-1,l)}\right) \\
K_{(k,l)} &= \hat{\Sigma}_{(k-1,l)|(k-1,l)}H_k^{*T} \qquad (14) \\
&\quad \times\left[H_k\hat{\Sigma}_{(k-1,l)|(k-1,l)}H_k^{*T}+\Sigma_{v(k,l)}\right]^{-1}
\end{aligned}
$$

$$\hat{\Sigma}_{(k,l)|(k,l)} = \hat{\Sigma}_{(k-1,l)|(k-1,l)} - K_{(k,l)}H_k\hat{\Sigma}_{(k-1,l)|(k-1,l)} \qquad (15)$$

$$\hat{x}_{(-1,l)|(-1,l)} = \{0_0, 0_1, \cdots, 0_{N-1}\} \qquad (16)$$

$$\hat{\Sigma}_{(-1,l)|(-1,l)} = \begin{pmatrix} P(0,l) & 0 & \cdots & 0 \\ 0 & P(1,l) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P(N-1,l) \end{pmatrix} \qquad (17)$$

式(13)～(15)は複素カルマンフィルタとなり($*T$は複素共役な転置行列を表す)、これにより$s(k,l)$の複素スペクトル$S(f,l)$の推定値$\hat{x}_{(k,l)|(k,l)} = (\hat{S}(0,l), \hat{S}(1,l), \cdots, \hat{S}(N-1,l))^T$が得られ、$\hat{x}_{(k,l)|(k,l)}$をIFFTすることにより推定クリーン音声$\hat{s}(k,l)$が得られる。

複素カルマンフィルタを駆動する際の初期値は式(16)、(17)に示した値に設定した。ここで、式(17)中の$P(f,l)$は2.4で述べる適応型Spectral Subtraction法により推定した$s(k,l)$のパワースペクトルを示す。また、今回は比較的定常な雑音の除去を行うので、雑音の分散$\Sigma_{v(k,l)}$は、雑音のみが存在する区間で求めた分散の値に設定した。

式(13)～(15)で表される複素カルマンフィルタに観測信号$\{y(0,l), y(1,l), \cdots, y(N-1,l)\}$を与えて実行すると、式(10)で示したベクトル$x_l$の推定が$N$回行われ、その過程の中で$x_l$の推定値$\hat{x}_{(k,l)|(k,l)}$が$N$個得られることになる。得られた$N$個の推定値から、$x_l$の推定値を決定するわけであるが、$k$の値が増大するに従って、$\hat{x}_{(k,l)|(k,l)}$は真値に向かって収束していくことを、予備実験により確認しているので、$k = N-1$の時の$\hat{x}_{(k,l)|(k,l)}$をx_lの最適な推定値とした。

2.3 カルマンフィルタの高速化

カルマンフィルタには膨大な計算量を必要とするという問題がある。この問題を解決するために、以下のようにして計算量の削減を行った。

カルマンフィルタの膨大な計算量の原因は、主に共分散行列$\hat{\Sigma}_{(k,l)|(k,l)}$の計算にあり、その全ての要素を用いてカルマンフィルタを実行すると、膨大な計算量を要することになってしまう。そこで、計算量を削減するために、$\hat{\Sigma}_{(k,l)|(k,l)}$の対角成分のみを用いてカルマンフィルタを実行したところ、ほとんど精度を落さずに、実時間での処理を実現することができた。この処理速度はワークステーション、SGI, Origin200(R12000, 250MHz CPU)により確認した。

2.4 適応型Spectral Subtraction

2.2で定式化した複素カルマンフィルタの初期値の1つである$\hat{\Sigma}_{(-1,l)|(-1,l)}$は、以下に示す適応型Spectral Subtraction法(ASS)[9]により推定した。

雑音重畳音声のパワースペクトルを$P_X(f)$、クリーン音声の推定パワースペクトルを$\hat{P}_S(f)$、雑音の平均推定パワースペクトルを$\overline{P}_N(f)$としたとき、SS法は以下のように示される(α, βはそれぞれサブトラクション係数、フロアリング係数を表す)。

$$\hat{P}_S(f) = \max\left[P_X(f)-\alpha\overline{P}_N(f), \beta P_X(f)\right] \qquad (18)$$

ここで、雑音が比較的定常であっても、局所的なSNRはクリーン音声のパワーによって常に変化しており、局所的なSNRであるsnr_lに応じてサブトラクション係数αの値を、以下のような決定関数fを定義して設定する必要がある。

$$\alpha = f(snr_l) \qquad (19)$$

次にsnr_lの推定法について述べる。雑音重畳音声の短時間RMSパワーをPow_X、クリーン音声の推定短時

間RMSパワーをPow_S、雑音の平均推定短時間RMSパワーを$\overline{Pow}_N$としたとき、snr_lは以下のように推定される。

$$snr_l = \begin{cases} 10\log_{10}\dfrac{Pow_S}{\overline{Pow}_N} & \sqrt{Pow_S} > 0 \\ \gamma(=-10) & \sqrt{Pow_S} \leq 0 \end{cases} \tag{20}$$

$$\sqrt{Pow_S} = \sqrt{Pow_X} - \sqrt{\overline{Pow}_N} \tag{21}$$

$\sqrt{Pow_S}$が負の値を持つとき、snr_lを計算できないので、定数γを代入した。以下、図3に本研究で用いたサブトラクション係数決定関数fを示す。

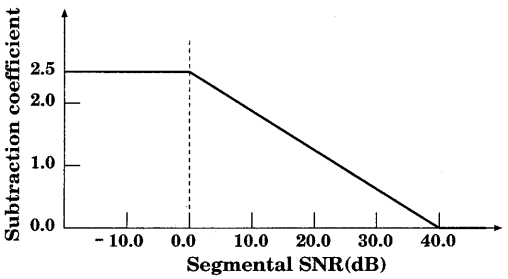

図 3: サブトラクション係数決定関数 f

2.5 雑音除去法の改良

本研究での雑音除去処理の中核は2.2, 2.3で述べたカルマンフィルタであり、その初期値の1つ$\hat{\Sigma}_{(-1,l)|(-1,l)}$は2.4で述べたASS法により推定する。ここで、雑音除去精度の向上のために、以下の処理を行った。

1. カルマンフィルタによりクリーン音声の複素スペクトルを推定する。
2. 推定された複素スペクトルからパワースペクトルを求める。
3. 得られたパワースペクトルを初期値$\hat{\Sigma}_{(-1,l)|(-1,l)}$に設定して再度カルマンフィルタを実行する。

これらの処理を行うことによって、より高い雑音除去精度を得ることができると思われる。また、これらの処理は2回のカルマンフィルタを必要とするが、実時間の1.5~2.0倍で実行可能である。以下、図4に処理フローを示す。

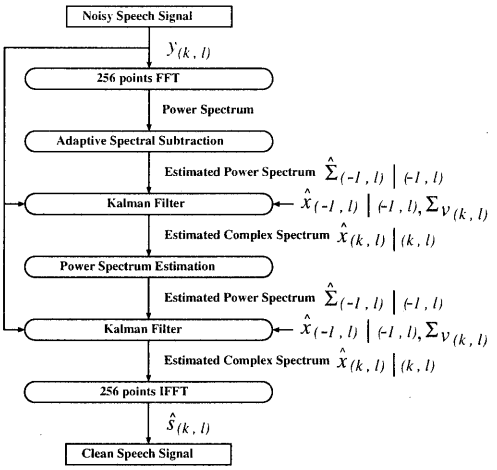

図 4: 雑音除去の処理フロー

3 実験

3.1 雑音除去実験

2の手法に従って雑音除去実験を行った。

3.1.1 雑音除去実験条件

音声データは、ATR音声データベースAセット5240単語より、ランダムに選出した男性話者1名の100単語音声を使用した。雑音の重畳はSNRを調整した後に計算機により行った。実験条件を表1に示す。

表 1: 雑音除去実験条件

標本化周波数	12kHz
音響パラメータ	256点FFTスペクトル
分析区間長	21.3ms
分析周期	21.3ms
時間窓	Hamming Window
重畳雑音	白色ガウス雑音, ピンク雑音, 計算機雑音
SNR	20,10,5,0dB

また、雑音除去音声の品質評価の為に、以下の式で表されるスペクトル歪み(SD)[10]を用いて評価を行った($A(f)$, $\hat{A}(f)$はそれぞれ原信号の振幅スペクトル、推定信号の振幅スペクトルを表す)。

$$SD = \sqrt{\frac{1}{N}\sum_{f=0}^{N-1}\left(20\log_{10}\frac{A(f)}{\hat{A}(f)}\right)^2} \tag{22}$$

3.1.2 雑音除去実験結果

図5, 6に提案手法により雑音除去された音声波形とスペクトログラムの1例を示す。

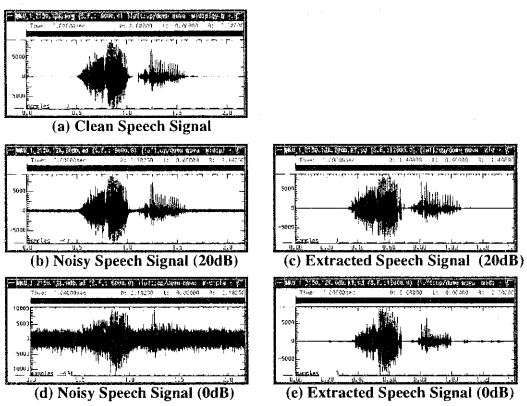

図 5: 雑音除去音声の波形(男性話者 /SHUUKYOU/)

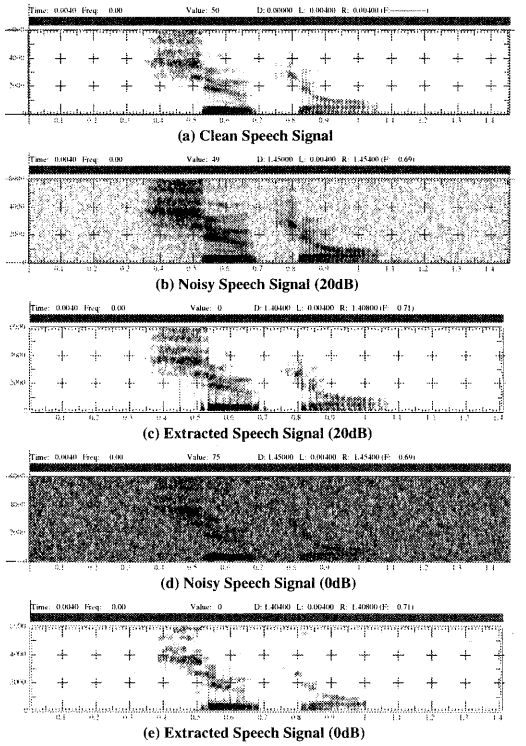

図 6: 雑音除去音声のスペクトログラム(男性話者 /SHUUKYOU/)

それぞれの図において、(a)が雑音の重畳しないクリーン音声、(b), (d)がそれぞれ20dB, 0dBの白色ガウス雑音が重畳した音声、(c), (e)がそれぞれ雑音除去を行った音声を表している。

20dBでの結果は波形、スペクトログラムともにクリーン音声の特徴をほぼ忠実に再現しているが、0dBでは波形に歪みが生じ、僅かに雑音成分が残るという結果になった。また、以下に式(22)によって計算されたSDの値を示す(図中、NoNR, KFはそれぞれ雑音除去無し、提案手法の結果を示す。)。

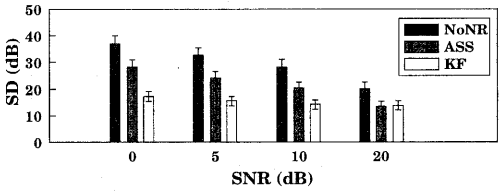

図 7: スペクトル歪み(白色ガウス雑音重畳)

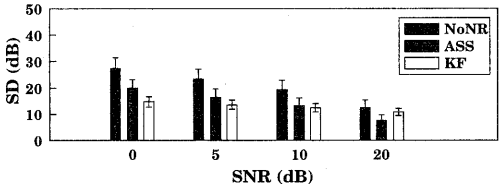

図 8: スペクトル歪み(ピンク雑音重畳)

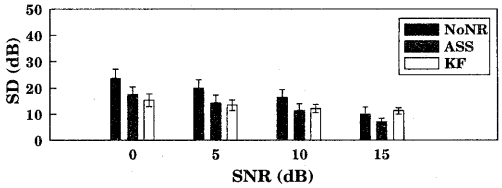

図 9: スペクトル歪み(計算機雑音重畳)

それぞれの結果において、0dB, 5dBでは提案手法によりSD値の改善が顕著に見られるが、10dB, 20dBではあまり改善が得られなかった。これはカルマンフィルタによる推定の際に推定誤差が大きかったためであると考えられる。

3.2 単語認識実験

3.1の実験により得られた雑音除去音声を用いてHMMにより単語認識実験を行った。

3.2.1 単語認識実験条件

認識に用いた音響モデルは、男性不特定話者音素HMMで、音素数41のmonophoneモデルである。学習には、日本音響学会新聞記事読み上げ音声コーパスのうち、男性話者137人分の21782発話を用いており、それぞれのデータに対してCepstrum Mean Normalization(CMN)を行っている。音響分析の条件、HMMの構造を表2, 3に示す。また、今回PMC法との認識精度の比較を行うので、合成する雑音HMMの構造もまた表3に示す。

表 2: 音響分析条件

標本化周波数	12kHz
音響パラメータ	12次MFCC + Power + 12次ΔMFCC + ΔPower + 12次ΔΔMFCC + ΔΔPower
分析区間長	20ms
分析周期	5ms
時間窓	Hamming Window

表 3: HMMの構造

HMM	音素HMM	雑音HMM
状態数	5状態3ループ	3状態1ループ
混合数	8	1
音素数／雑音数	41	1
HMMのタイプ	Left-to-Right HMM	Left-to-Right HMM

3.2.2 単語認識実験結果

表4～6に白色ガウス雑音、ピンク雑音、計算機雑音それぞれが重畳した場合の単語認識実験の結果を示す。また、図10に全ての結果をプロットしたグラフを示す。

表 4: 白色ガウス雑音重畳音声での結果(%)

SNR	∞dB	20dB	10dB	5dB	0dB
NoNR	98.0	78.0	15.0	1.0	0.0
ASS	98.0	85.0	80.0	33.0	5.0
PMC	98.0	87.0	81.0	77.0	68.0
KF	98.0	91.0	88.0	81.0	71.0

表 5: ピンク雑音重畳音声での結果(%)

SNR	∞dB	20dB	10dB	5dB	0dB
NoNR	98.0	81.0	37.0	6.0	2.0
ASS	98.0	93.0	91.0	74.0	30.0
PMC	98.0	94.0	89.0	88.0	84.0
KF	98.0	88.0	88.0	84.0	79.0

表 6: 計算機雑音重畳音声での結果(%)

SNR	∞dB	20dB	10dB	5dB	0dB
NoNR	98.0	87.0	47.0	13.0	8.0
ASS	98.0	95.0	86.0	64.0	22.0
PMC	98.0	97.0	92.0	88.0	82.0
KF	98.0	91.0	88.0	86.0	75.0

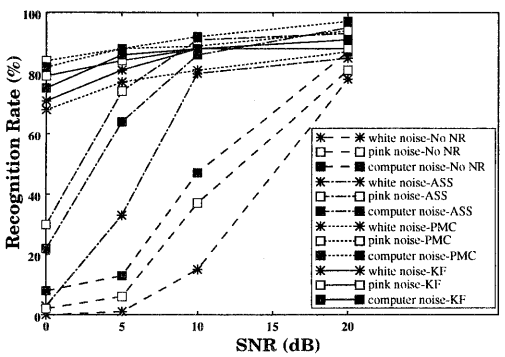

図 10: 単語認識結果

白色ガウス雑音が重畳した音声での結果に注目すると、全てのSNRにおいて、PMC法を上回る結果が得られた。しかし、ピンク、計算機雑音が重畳した音声での結果では僅に下回る結果となった。これは、カルマンフィルタの性質によるものであると考えられる。カルマンフィルタは、観測雑音が白色ガウス性であれば最適フィルタを構成するが、白色ガウス性以外の場合は準最適フィルタという位置づけになり、白色ガウス性の場合に比べて精度が劣ってしまう[7, 8]。この性質が雑音除去精度および雑音除去音声の認識率に影響したものと考えられる。また、それぞれの結果において20dB, 10dBで一部ASS法に劣るものがあるが、これは図7～9で示した結果と同様に推定誤差が大きかったためであると考えられる。この問題は隣接するフレーム間での相関を考慮して推定値を修正し、推定誤差を小さく押さえることにより解決できると考えられる。

4 おわりに

カルマンフィルタにより雑音除去を施した音声で音声認識を行い、良好な結果が得られることを示した。また、カルマンフィルタの計算量を大幅に削減して高速化を行うことにより、実時間の1.5～2.0倍での処理を実現した。今後は雑音除去精度の向上及び、音楽、他者の音声等、非定常雑音への対応について検討する予定である。

参考文献

[1] M.J.F.Gales, S.J.Young: "An Improved Aproach to the Hidden Markov Model Decomposition of Speech and Noise", *ICASSP*, I-233-236(1992)

[2] M.J.F.Gales, S.J.Young: "Robust Continuous Speech Recognition Using Paralell Model Combination", *IEEE Trans. Speech and Audio Processing*, Vol.4, No.5, pp.352-359, Sep.(1996)

[3] M.Unoki, M.Akagi: "A Method of Signal Extraction from Noisy Signal Based on Auditory Scene Analysis", *Speech Communication 27*, pp.261-279(1999)

[4] D.C.Popescu, I.Zejiković: "Kalman Filtering of Colored Noise for Speech Enhancement", *ICASSP*, II-997-1000(1998)

[5] Z.Goh, K.Tan, B.T.G.Tan: "Kalman-Filtering Speech Enhancement Method Based on Voiced-Unvoiced Speech Model", *IEEE Trans. Speech and Audio Processing*, Vol.7, No.5, pp.510-524, Sep.(1999)

[6] 藤本雅清, 有木康雄: "カルマンフィルターを用いた雑音環境下における音声認識の検討— 雑音適応と雑音除去 —", 音響講論, 1-1-16, pp.31-32(1999)

[7] 有本 卓: "カルマン・フィルター", 産業図書.

[8] 中野道雄 監修 西山 清 著: "パソコンで解くカルマンフィルタ", 丸善.

[9] 山本寛樹, 山田雅章, 小森康弘, 大洞恭則: "推定Segmental SNRに基づく適応的Spectral Subtraction法による音声認識", 信学技報, SP94-50, pp.17-24(1994)

[10] M.Mizumachi, M.Akagi: "An Objective Distortion Estimator for Hearing Aids and its Application to Noise Reduction", *Eurospeech'99*, vol.6, pp.2619-2622, Sep.(1999)