音声言語情報処理 34－4
(2000. 12. 21)

音声の時間変化モデルに基づく音声信号推定法を用いた非定常雑音下での音声認識

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

あらまし本研究では，音声の時間変化モデルに基づいた非定常雑音に対する雑音除去法を提案する．提案手法では，音声の時間変化モデルをカルマンフィルタによる推定問題に適用することにより，音楽等のような非定常雑音が重畳した音声から，クリーンな音声信号を推定している．音声の時間変化モデルは，雑音重畳音声におけるクリーン音声の時間変動を，Taylor展開を用いることにより表現したモデルである．モデルの構成に必要なパラメータの1つである雑音の変動成分は，線形予測法により推定を行っている．提案手法の評価のために，3種類の音楽が重畳した音声を用いて大語彙連続音声認識を行ない，単語正解精度において，従来法であるParallel Model Combination(PMC)法と比較を行った．その結果，提案手法により，PMC法よりも高い単語正解精度が得られた．

キーワード　雑音環境下での音声認識，非定常雑音，音声の時間変化モデル，カルマンフィルタ

Speech Recognition under Non-stationary Noisy Environments Using Signal Estimation Method Based on Speech State Transition Model

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 non-stationary noise reduction method based on speech state transition model. Our proposed method estimates the speech signal under non-stationary noisy environments such as musical background by applying speech state transition model to Kalman filtering estimation. The speech state transition model represents the state transition of speech component in non-stationary noisy speech and is modeled by using Taylor expansion. In this model, the state transition of noise component is estimated by using linear predictive estimation. In order to evaluate the proposed method, we carried out large vocabulary continuous speech recognition experiments under 3 types of musics and compared the results with conventionally used Parallel Model Combination(PMC) method in word accuracy rate. As a result, the proposed method obtained word accuracy rate superior to PMC.

key words noisy speech recognition, non-stationary noise, speech state transition model, Kalman filter

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．これを受けて，背景雑音に頑健な音声認識システムを確立し，音声認識システムの完全な実用化を実現するために，様々な研究が行われている [1, 2].

雑音に頑健な音声認識システム確立のためのアプローチとして，認識システムを雑音に適応させる方法 (雑音適応)[3]-[6] と，雑音が重畳した音声から雑音成分を取り除き，クリーンな音声を抽出して認識を行う方法 (雑音除去) [7]-[10] の 2 種類が考えられる．

雑音適応の方法として，PMC(Parallel Model Combination)[3, 4] や，NOVO(VOice mixed with NOise)[5] に代表される HMM 合成法が提案されており，その有効性が報告されている．HMM 合成法において，合成する雑音 HMM の状態数，混合分布数を定常雑音の場合に比べて増加させることにより，非定常雑音への適応が可能であるといわれている．しかし，実環境下での適用を考えると，雑音 HMM を学習できるのは，発話が始まるまでの区間であり，発話が始まった後の雑音の変動は学習に反映されない．このようにして学習された雑音 HMM を音声 HMM に合成させると，認識の際に不整合が生じてしまうと考えられる．この問題を解決するために，初期の雑音モデルからの時間変動の残差を逐次的に適応させていく手法が提案されている [6].

HMM 合成法は合成のための計算量が比較的多く，大語彙連続音声認識に使われる Triphone モデルの HMM のように音素数，混合数の多いモデルに対して合成を行うと，非常に時間がかかってしまうという問題がある．

一方，雑音除去の観点では，Spectral Subtraction(SS) 法 [7] がよく知られている．しかし，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まねくという問題がある．また，SS 法で減算する雑音スペクトルは，雑音のみが存在する区間から得られた平均スペクトルであり，非定常雑音におけるスペクトルの時間変動が考慮されていない．

また，我々は以前カルマンフィルタを用いた雑音除去法 [10] を提案したが，この方法においても，カルマンフィルタの初期値推定に SS 法を用いており，雑音の時間変動を考慮にいれていなかった．

そこで，本研究では，これらの問題点をふまえて，非定常雑音が重畳した音声における，クリーン音声成分の時間変動の様子をモデル化し，得られたモデルに対してカルマンフィルタを適用することにより，非定常雑音が重畳した音声からクリーン音声を推定することを試みた．提案手法の評価には大語彙連続音声認識を用いており，単語正解精度において，従来法である PMC 法と比較を行った．

2 音声の時間変化モデル

本研究における非定常雑音は，打楽器を含まない比較的ゆったりとした音楽を対象としている．まず，音声と音楽のスペクトルの変動の度合いが，それぞれどのような特徴をもっているかを調べるために，各フレームにおいて，音声と音楽それぞれのパワースペクトルを特徴ベクトルと考える．隣接するフレーム間で，このベクトルの正規化された内積値，つまりそれぞれのベクトルの余弦値を計算した．図 1に余弦値の時間推移の例を示す．

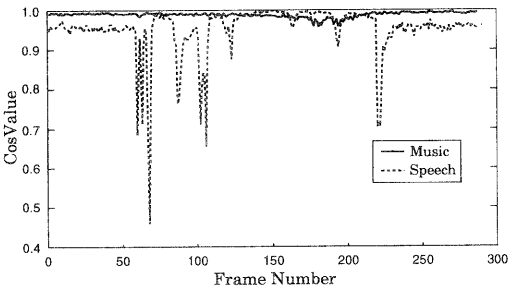

図 1: 音声，音楽の余弦値の時間推移

図 1より，音楽のパワースペクトルの時間変化は比較的緩やかであり，音声のパワースペクトルの時間変化は激しいことがわかる．この特徴をふまえて，以下，音声のパワースペクトルの時間変化モデルを定義する．

k 番目の短時間フレームにおいて，雑音重畳音声に含まれているクリーン音声のパワースペクトルは以下のように表される．

$$S(k) = \exp\left(X^l(k)\right) - \exp\left(N^l(k)\right) \qquad (1)$$

ここで，$X(k)$，$S(k)$，$N(k)$ は，雑音重畳音声のパワースペクトル，クリーン音声のパワースペクトル，雑音のパワースペクトルベクトルであり，添字 l は対数パワースペクトル領域を表す．

式 (1) において，$S(k)$ から $S(k+1)$ への時間変化は以下のように表される．

$$
\begin{aligned}
S(k+1) &= S(k) + \Delta S(k) \\
&= \exp\left(X^l(k) + \Delta X^l(k)\right) \\
&\quad - \exp\left(N^l(k) + \Delta N^l(k)\right)
\end{aligned} \tag{2}
$$

$$
\Delta X^l(k) = X^l(k+1) - X^l(k) \tag{3}
$$

$$
\Delta N^l(k) = N^l(k+1) - N^l(k) \tag{4}
$$

ここで，式(2)に対して1次のTaylor展開を適用することにより，$S(k)$から$S(k+1)$への時間変化を以下のように線形近似することができる．

$$
\begin{aligned}
S(k+1) &\simeq S(k) + \frac{\partial S(k)}{\partial X^l(k)}\Delta X^l(k) + \frac{\partial S(k)}{\partial N^l(k)}\Delta N^l(k) \\
&= S(k) + X(k)\Delta X^l(k) - N(k)\Delta N^l(k) \\
&= S(k) + (S(k) + N(k))\,\Delta X^l(k) \\
&\quad -N(k)\Delta N^l(k) \\
&= \left(1 + \Delta X^l(k)\right) S(k) \\
&\quad +N(k)\left(\Delta X^l(k) - \Delta N^l(k)\right) \\
&= F_k S(k) + G_k W(k)
\end{aligned} \tag{5}
$$

$$
\frac{\partial S(k)}{\partial X^l(k)} = \frac{\partial (X(k) - N(k))}{\partial X(k)} \cdot \frac{\partial X(k)}{\partial X^l(k)} = \dot{X}(k) \tag{6}
$$

$$
\frac{\partial S(k)}{\partial N^l(k)} = \frac{\partial (X(k) - N(k))}{\partial N(k)} \cdot \frac{\partial N(k)}{\partial N^l(k)} = -N(k) \tag{7}
$$

$$
F_k = 1 + \Delta X^l(k) \tag{8}
$$

$$
G_k = N(k) \tag{9}
$$

$$
W(k) = \Delta X^l(k) - \Delta N^l(k) \tag{10}
$$

式(5)~(7)において，$\frac{\partial S(k)}{\partial X^l(k)}$，$\frac{\partial S(k)}{\partial N^l(k)}$はベクトルの要素は無相関として，要素毎に独立に偏微分を行う事を意味する．

式(5)において，$N(k)$，$\Delta N^l(k)$が未知パラメータとして存在するが，これらの値は$N(k)$の時間変化が比較的緩やかなことを利用して推定を行う．これらの値の推定は後述の3.3節にて述べる．

以上より，式(5)を音声の時間変化モデルと定義し，このモデルをカルマンフィルタによる信号推定問題に適用させた．

3 カルマンフィルタによる推定

図2に提案する手法の概念図を示す．

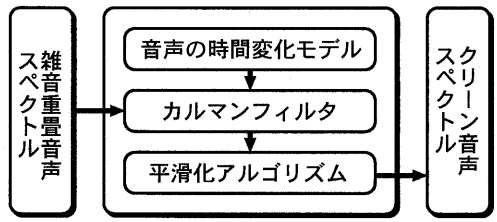

図 2: 提案手法の概念図

図2において，カルマンフィルタのフィルタ方程式は，前述の音声の時間変化モデルに基づいて定義され，得られたフィルタ方程式により推定されたクリーン音声のスペクトルは平滑化アルゴリズムにより，再推定される．以下，図2の各要素について述べる．

3.1 状態空間モデル

カルマンフィルタを用いて信号を推定するためには，状態空間モデルと呼ばれる信号モデルを定義する必要がある[11, 12]．状態空間モデルは状態方程式，観測方程式と呼ばれる2つの方程式により構成され，それぞれ目的信号の状態変化（時間変化）と観測信号が目的信号から生成される過程を表現している．本研究では，$S(k)$をカルマンフィルタにより推定するために，以下のような状態空間モデルを定義した．

$$
S(k+1) = F_k S(k) + G_k W(k) \tag{11}
$$

$$
X(k) = S(k) + N(k) \tag{12}
$$

上式において，式(11)が状態方程式であり，2節で定義した音声の時間変化モデルを適用させている．また，式(12)が観測方程式である．

3.2 カルマンフィルタ

3.1節で定義された状態空間モデルにより，以下のようなカルマンフィルタのフィルタ方程式が得られる．

$$
\hat{S}(k) = F_{k-1}\hat{S}(k-1) + K_k\left(X(k) - F_{k-1}\hat{S}(k-1)\right) \tag{13}
$$

$$
K_k = \frac{Q_k}{Q_k + \Sigma_{N(k)}} \tag{14}
$$

$$
\begin{aligned}
Q_k &= F_{k-1}(I - K_{k-1})Q_{k-1}F_{k-1}^T \\
&\quad +G_{k-1}\Sigma_{W(k-1)}G_{k-1}^T
\end{aligned} \tag{15}
$$

式(13)～(15)において，$k=0,1,\cdots,N$(Nは最終フレーム)であり，$\hat{S}(k)$は$S(k)$の推定値，Q_kは誤差の共分散行列である．$\hat{S}(k)$，Q_kの初期値はそれぞれ以下のように設定した．

$$\hat{S}(0)=\mathbf{0} \tag{16}$$

$$Q_0=\mathbf{0} \tag{17}$$

式(15)の$\Sigma_{W(k)}$は$W(k)$の対角共分散行列であり，$W(k)$は平均零のガウス過程であると仮定することにより，以下のようにして求められる．

$$\Sigma_{W(k)}=W(k)W(k)^T \tag{18}$$

また，式(14)の$\Sigma_{N(k)}$は$N(k)$の対角共分散行列であり，$W(k)$同様，平均零のガウス過程であると仮定して，以下のようにして求めた．

$$\Sigma_{N(k)}=N(k)N(k)^T \tag{19}$$

3.3 線形予測法による$N(k)$の推定

式(13)～(15)において，$\Sigma_{W(k)}$，$\Sigma_{N(k)}$の計算にはパラメータ$N(k)$が必要となるが，実際に観測可能なのは$X(k)$のみであり，$N(k)$は未知パラメータである．ここで，$N(k)$の時間変動は図1より，比較的緩やかであり，$N(k)$は過去の値$N(k-1),N(k-2)\cdots$と高い相関を持つと考えられる．このことより，$N(k)$を以下のようなp次の線形予測法により推定した．

$$N_j(k)=\begin{cases} X_j(k) & 0\le k<p \\ \sum_{i=1}^{p} a_{ij}N_j(k-i) & k\ge p \end{cases} \tag{20}$$

式(20)において，jはFFT分析におけるチャネル番号，a_{ij}は線形予測係数である．式(20)において，$0\le k<p$のときは，発話が始まっていない，雑音のみが存在している区間と見なして，$N_j(k)=X_j(k)$とした．また，$k\ge p$の時は，$N(k)$は線形予測法により推定する．本研究では，線形予測係数の次数は$p=12$とした．この次数は実験的に求めた値である．

3.4 平滑化アルゴリズム

カルマンフィルタは観測信号$X(0),X(1),\cdots,X(k)$を用いて時刻tにおける目的信号$S(t)$の最適な推定値を求めるアルゴリズムである．ここで，$S(t)$の最適な推定値を求める問題は，$S(t)$と$X(0),X(1),\cdots,X(k)$の時間的関係から，以下のように3つの場合に分類される．

1. 未来の$S(t)$ $(t>k)$の推定値を求める予測(prediction)問題
2. 現在の$S(t)$ $(t=k)$の推定値を求めるろ波(filtering)問題
3. 過去の$S(t)$ $(t<k)$の推定値を求める平滑(smoothing)問題

一般的なカルマンフィルタのアルゴリズムでは1.の予測問題及び，2.のろ波問題が取り込まれており，3.の平滑問題は扱われていない．平滑問題は，予測問題，ろ波問題で得られた推定値を用いて，時間に対して後ろ向きに推定を行うことによって，平滑化された推定値を求めるアルゴリズムである[12]．よって，平滑問題は一般的なカルマンフィルタにより求められた推定値を再推定するアルゴリズムであると位置づけることができ，この平滑問題を取り入れることにより，目的信号の推定精度を高めることが可能である．以下に式(13)～(15)に対する平滑化アルゴリズムを示す．

$$\tilde{S}(k)=\hat{S}(k)+C_k\left(\tilde{S}(k+1)-F_k\hat{S}(k)\right) \tag{21}$$

$$C_k=\frac{(I-K_k)Q_kF_k^T}{Q_{k+1}} \tag{22}$$

式(21)，(22)において$k=N-1,N-2,\cdots,0$であり，$\tilde{S}(k)$は$\hat{S}(k)$の平滑化推定値である．

4 実験

提案手法により推定された音声信号に対して，大語彙連続音声認識実験を行い，PMC法との比較を行った．

4.1 実験条件

評価用データには，IPA-98-TestSetのうち，男性23名が発声したデータ100文を用いている．また，重畳させる音楽は，ピアノソロ曲3曲(Piano1, Piano2, Piano3)を用いた．音楽の重畳はそれぞれの音楽データからランダムに切り出した100区間分のデータを，式(23)を用いてSNR(= 20, 10, 0dB)を調整した後に，それぞれ音声データに計算機を用いて重畳させた．

$$x(t)=s(t)+\frac{Pow_s}{10^{SNR/20}Pow_n}\cdot n(t) \tag{23}$$

ここで，$x(t)$，$s(t)$，$n(t)$はそれぞれ雑音重畳音声，音声，雑音を表し，Pow_s，Pow_nはそれぞれ音声，雑音のRMS(Root Mean Square)パワーを表す．

音響モデルには，話者独立なmonophone HMMを用いた．HMMの学習には，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1, 2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の構造を表3に示す．

ここで，認識における特徴パラメータにはMFCCを用いているが，本研究では，提案手法により推定されたパワースペクトルをメルフィルタバンク，DCTを用いることにより直接MFCCを算出している．また，得られたMFCCについても学習時と同様にCMNを行っている．

表 1: 音響分析条件

標本化周波数	16kHz
高域強調	$1-0.97z^{-1}$
特徴パラメータ (雑音除去)	512点FFTスペクトル
特徴パラメータ (認識)	12次MFCC + log Power $+\Delta+\Delta\Delta$
分析区間長	20ms
分析周期	10ms
時間窓	Hamming Window
SNR	20,10,0dB

表 2: 音素HMMの構造

状態数	5状態3ループ
混合数	12
音素数	41
タイプ	Left-to-Right HMM

表 3: 雑音HMMの構造

状態数	3状態1ループ
混合数	1
雑音数	1
タイプ	Left-to-Right HMM

言語モデルには，1st-passにbigram，2nd-passにtrigramを用いており，IPAモデル98年度版のうち，語彙数20k，cut-offはbigram，trigramそれぞれに対して4-4のモデルを用いている．言語モデルの学習データは，毎日新聞記事75ヶ月分である．

4.2 実験結果

表4〜6にそれぞれの手法による認識結果を示す．それぞれの手法において，上段は単語正解率 ($Corr$)，下段は単語正解精度 (Acc) を示す．

$$Corr(\%) = \frac{N-S-D}{N} \times 100 \quad (24)$$

$$Acc(\%) = \frac{N-S-D-I}{N} \times 100 \quad (25)$$

S : 置換誤り単語数

D : 脱落誤り単語数

I : 挿入誤り単語数

N : 全単語数

表 4: 認識結果 (Piano 1)(%)
(上段: $Corr$, 下段: Acc)

SNR	∞dB	20dB	10dB	0dB
雑音処理無し	88.78	83.26	69.88	35.64
	86.49	79.52	61.57	20.04
PMC	88.78	81.42	69.06	37.22
	86.49	78.50	63.47	30.94
提案手法	88.14	85.23	75.52	50.35
	85.16	81.17	67.91	36.33

表 5: 認識結果 (Piano 2)(%)
(上段: $Corr$, 下段: Acc)

SNR	∞dB	20dB	10dB	0dB
雑音処理無し	88.78	82.69	66.01	34.43
	86.49	79.62	56.44	20.04
PMC	88.78	80.66	67.47	38.62
	86.49	77.62	61.64	31.90
提案手法	88.14	82.75	71.15	46.48
	85.16	78.19	62.08	31.96

表 6: 認識結果 (Piano 3)(%)
(上段: $Corr$, 下段: Acc)

SNR	∞dB	20dB	10dB	0dB
雑音処理無し	88.78	83.45	70.13	37.54
	86.49	79.45	61.95	21.81
PMC	88.78	81.36	68.36	38.11
	86.49	78.38	63.35	32.53
提案手法	88.14	85.10	75.90	48.45
	85.16	81.23	67.53	33.61

それぞれの表において，PMC法と比較した結果，提案手法により，全ての条件下でAccの改善が見られた．しかし，全体的にAccの改善量は小さい．一方，$Corr$においては，Accに比べて大きい改善が得られた．このことから，提案手法により，置換誤り及び，脱落誤り単語数は減少したが，挿入誤り単語数が増加したことがわかる．

挿入誤り単語数が増加した理由として，式(20)の線形予測法を用いて$N_j(k)$を推定する際に，$N_j(k)$の推定誤差が大きくなり，$N(k)$の十分な推定精度が得られなかったためであると考えられる．この$N(k)$の推定精度の低さが$\tilde{S}(k)$の推定精度に影響を与え，認識時に$\tilde{S}(k)$から得られたMFCCパラメータを用いて計算された尤度が，

低くなってしまったと考えられる．また，デコーディングがこの尤度計算結果に基づいて行われるため，誤った単語が連結されてしまったものと考えられる．この問題において，$N(k)$として真値を与えることができれば，大幅なAccの改善が得られることを予備実験により確認している（例えば，Piano1，0dBの環境下で，約78%のAccが得られる）．このことより，$N(k)$を可能な限り正確に推定する必要があると言える．

また，式(5)において，式(2)に対して1次のTaylor展開を適用しているが，1次のTaylor展開では近似精度が荒らく，高次項を含めたTaylor展開が必要であると考えられる．

さらに，$\tilde{S}(k)$の推定精度の低さから生じる尤度の曖昧性を利用して，信頼尺度を組み込んだデコーダーを用いることにより，信頼度の低い単語を棄却する方法についても今後検討する必要があると考えられる．

5 おわりに

本研究では，非定常雑音下における音声認識手法として，音声スペクトルの時間変動をモデル化し，得られたモデルに対してカルマンフィルタを適用して，音声スペクトルを推定する手法について検討した．3種類の音楽が重畳した音声に対して，音声スペクトルの推定を行い，大語彙連続音声認識実験を行った結果，提案手法により単語正解率及び，単語正解精度の改善が見られ，特に．単語正解率の改善が大きいことを確認した．

今後，雑音成分の推定精度が重要なことから，雑音成分の時間変化の高精度な推定法について検討する予定である．また，今回の実験では，非定常雑音に打楽器等を含まない，時間変動が比較的緩やかな音楽を用いていたが，今後打楽器等を含む変動が激しい音楽についても検討する予定である．

参考文献

[1] 中川聖一: “ロバストな音声認識のための音響信号処理”, 音響誌, 53巻, 11号, pp.864-871(1997)

[2] 松本 弘: “音声認識における環境適応技術”, 信学技報, SP99-111, pp.109-114(1999)

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

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

[5] F.Martin, K.Shikano, Y.Minami, Y.Okabe: “Recognition of Noisy Speech by Composition of Hidden Markov Models”, 信学技報, SP92-96, pp.9-16(1992)

[6] Kaisheng Yao, Bertram E. Shi, Pascale Fung, Zhigang Cao: “Residual Noise Compensation for Robust Speech Recognition in Nonstationary Noise”, *ICASSP*, II-1125-1128(2000).

[7] S.F.Boll: “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, *IEEE Trans. Acoustic Speech Signal Processing*, Vol.27, No.2, pp.113-120(1979)

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

[9] 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)

[10] M.Fujimoto, Y.Ariki: “Noisy Speech Recognition Using Noise Reduction Method Based on Kalman Filter”, *ICASSP*, III-1723-1726(2000)

[11] 有本 卓: “カルマン・フィルター”, 産業図書.

[12] 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.