

Switching dynamical system とパーティクルフィルタを用いた 非定常雑音の逐次推定

藤本 雅清[†] 中村 哲[†]

[†] ATR 音声言語コミュニケーション研究所

〒 619-0288 京都府相楽郡精華町光台 2-2-2

E-mail: {masakiyo.fujimoto, satoshi.nakamura}@atr.jp

あらまし 本研究では、パーティクルフィルタを用いた非定常雑音の逐次推定法及び抑圧法を提案し、従来法と比較して非定常雑音下での音声認識性能改善に効果的である事を示す。提案手法では、パーティクルフィルタの根幹に相当する状態空間モデルを Polyak averaging と feedback により定義しており、さらに switching dynamical system へと拡張することによって、非定常雑音の時間変動をより効果的に表現する。このような状態空間モデルを導入することにより、非定常雑音環境下における音声認識性能を大幅に改善できることを示す。

キーワード 雑音下音声認識, 非定常雑音, パーティクルフィルタ, switching dynamical system

Sequential Non-Stationary Noise Tracking Based on Switching Dynamical System and Particle Filtering

Masakiyo FUJIMOTO[†] and Satoshi NAKAMURA[†]

[†] ATR Spoken Language Communication Research Laboratories

2-2-2, Hikaridai, Seika-cho, Souraku-gun, Kyoto, 619-0288, Japan

E-mail: {masakiyo.fujimoto, satoshi.nakamura}@atr.jp

Abstract This paper addresses a speech recognition problem in non-stationary noise environments: the estimation of noise sequences. To solve this problem, we present a particle filter-based sequential noise estimation method for the front-end processing of speech recognition. In the proposed method, the particle filter is defined by a dynamical system based on Polyak averaging and feedback. We also introduce a switching dynamical system into the particle filter to cope with non-stationarity of the environment noise. In the evaluation results, we observed that the proposed method improves speech recognition accuracy in the results of non-stationary noise environments by a noise compensation method with stationary noise assumptions.

Key words noisy speech recognition, non-stationary noise, particle filter, switching dynamical system

1. ま え が き

雑音下での音声認識性能の改善は、音声認識技術に課せられた重要な問題の一つである。この問題において、時間的に定常的な性質を持つ雑音を含む音声の認識については、様々な研究成果が報告されており、高い技術水準に達しているといえる [1]~[3]。しかし、実環境で観測される雑音の多くは、時間的に変動する非定常的な性質を持っており、実環境下における頑健な音声認識技術を確立するためには、このような非定常雑音への対処が必要不可欠となる。

一般に、非定常雑音下での音声認識性能を改善するためには、雑音の時間変動を正確に逐次推定する必要がある。しかし、音

声認識を行う際に観測できる信号は通常、雑音が重畳した音声のみであり、クリーン音声のみならず雑音もが非定常的な性質を持つ場合、各信号がそれぞれどのような時間推移を行うかを推定することは困難な問題である。この問題に関して、逐次 EM アルゴリズムを用いた非定常雑音の逐次推定法が提案され、その有効性が報告されている [4]~[6]。しかし逐次 EM アルゴリズムでは、1 フレーム毎に EM アルゴリズムの繰り返し推定を行ってパラメータを収束させる必要があるため、計算量が膨大なものとなり実時間処理に向かないという問題がある。よって実用面の観点から、高速かつ高精度な逐次推定法が望まれる。

この問題に関して、近年、ベイズ推定法的一种であるパーティクルフィルタ [7], [8] に基づく逐次推定法が注目されてお

り、様々な研究分野で応用されている。パーティクルフィルタは逐次モンテカルロ法に基づく手法であり、逐次 EM アルゴリズムのような繰り返し推定法を必要としないので計算量が少なく、実時間向け処理に適した手法であるという利点がある。以上の点を踏まえて、本研究では、パーティクルフィルタによる非定常雑音の逐次推定について検討する。さらに、パーティクルフィルタ推定された雑音を MMSE (Minimum Mean Square Error) 推定に基づく雑音抑圧法 [3] に適用し、非定常雑音下での音声認識精度の改善に効果があることを示す。

パーティクルフィルタを適用する際には、状態空間モデル (Dynamical system) と呼ばれる信号モデルを定義する必要がある。一般に、状態空間モデルは目的信号の時間 (状態) 遷移過程を表現した状態方程式と、観測信号の生成過程を表現した観測方程式から構成されている。このような状態空間モデルの定義において、我々は以前、状態方程式に Polyak averaging と feedback [6], [9], [10] を導入した。ここで、Polyak averaging には 3 つの制御パラメータがあり、以前はそれらのパラメータを時間固定のものとして扱っていた [11]。しかし、非定常雑音の時間変化の様相は時々刻々と変化するため、これらのパラメータも時間固定ではなく、時間変動させる必要があるといえる。この点に関しては、状態空間モデルを複数の Dynamical system から構成される Switching dynamical system に拡張し、フレーム毎に最適な Dynamical system を選択する方法を導入する。このような状態空間モデルを用いる事によって、非定常雑音の時間変動をより効果的に表現でき、音声認識性能の更なる改善が得られることを示す。

2. パーティクルフィルタによる非定常雑音の抑圧

図 1 は、提案する非定常雑音の抑圧法の概要を示しており、パーティクルフィルタに基づく非定常雑音の推定部と MMSE 推定に基づく雑音抑圧部の二つに大きく分かれている。

非定常雑音の推定部において、本研究で用いるパーティクルフィルタは、

- (1) Sequential importance sampling [7] によるパラメータサンプリング
- (2) 再サンプリング
- (3) マルコフ連鎖モンテカルロ (Metropolis-Hastings 法 [12]) によるサンプリング

の 3 つの要素技術により構成されている。以下、各部の処理内容について述べる。

2.1 パーティクルフィルタによる逐次推定

Sequential importance sampling

t 番目の短時間フレームにおいて、雑音重畳音声、雑音の対数メルスペクトルを要素に持つベクトルをそれぞれ X_t , N_t と定義する。次に、 X_t が観測されたときの N_t の事後確率分布が、次式のようなマルコフ連鎖に従うと仮定すると、 N_t の推定問題は、事後確率 $p(N_{0:t}|X_{0:t})$ 最大にするような信号列 $N_{0:t}$ を推定する問題に帰着する (ただし、 $N_{0:t} = \{N_0, \dots, N_t\}$ and $X_{0:t} = \{X_0, \dots, X_t\}$ とする)。

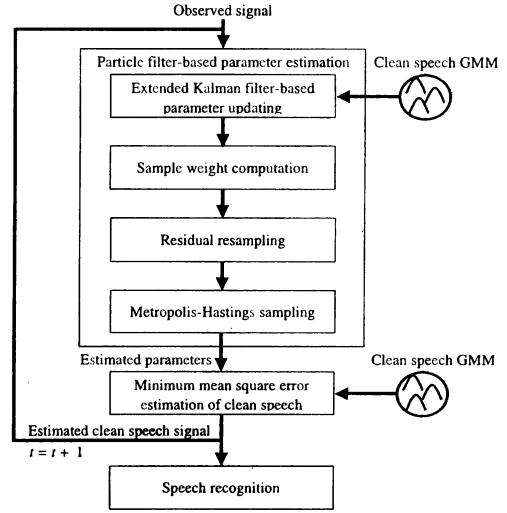


図 1 提案手法の概要

$$p(N_{0:t}|X_{0:t}) = p(N_0|X_0) \prod_{t'=1}^t p(N_{t'}|N_{t'-1})p(X_{t'}|N_{t'})(1)$$

パーティクルフィルタでは、時刻 t の事後確率分布を次式のようなモンテカルロサンプリングにより近似する。

$$\begin{aligned} p(N_{0:t}|X_{0:t}) &\simeq \frac{1}{J} \sum_{j=1}^J \delta(N_{0:t} - N_{0:t}^{(j)}) \\ &\simeq \sum_{j=1}^J w_t^{(j)} p(N_{0:t}^{(j)}|X_{0:t}) \end{aligned} \quad (2)$$

上式において、 j はサンプル番号、 J はサンプルの総数、 $\delta(\cdot)$ は Dirac-delta 関数、 $w_t^{(j)}$ は各時刻におけるサンプル j の重みであり ($\sum_{j=1}^J w_t^{(j)} = 1$)、 $w_t^{(j)}$ は次式により与えられる。

$$w_t^{(j)} \propto \frac{p(N_{0:t}^{(j)}|X_{0:t})}{q(N_{0:t}^{(j)}|X_{0:t})} \quad (3)$$

$q(N_{0:t}^{(i)}|X_{0:t})$ は、importance density と呼ばれる、サンプル $N_{0:t}^{(i)}$ を出力する確率分布であり、以下の連鎖モデルで表現されるものとする。

$$q(N_{0:t}|X_{0:t}) = q(N_t|N_{0:t-1}, X_{0:t})q(N_{0:t-1}|X_{0:t-1}) \quad (4)$$

また、式 (1) の事後確率分布は、ベイズ則により次式のように表されるため、

$$\begin{aligned} p(N_{0:t}|X_{0:t}) &= \frac{p(N_t|N_{t-1})p(X_t|N_t)}{p(X_t|X_{0:t-1})}p(N_{0:t-1}|X_{0:t-1}) \\ &\propto p(N_t|N_{t-1})p(X_t|N_t)p(N_{0:t-1}|X_{0:t-1}) \end{aligned} \quad (5)$$

式 (4), (5) より、 $w_t^{(j)}$ は次式により与えられる。

$$w_t^{(j)} \propto w_{t-1}^{(j)} \frac{p(N_t^{(j)}|N_{t-1}^{(j)})p(X_t|N_t^{(j)})}{q(N_t^{(j)}|N_{0:t-1}^{(j)}, X_{0:t})} \quad (6)$$

ここで、 $p(N_t^{(j)}|N_{t-1}^{(j)}) = q(N_t^{(j)}|N_{0:t-1}^{(j)}, X_{0:t})$ と仮定するこ

とにより、次式が得られる。

$$w_t^{(j)} \propto w_{t-1}^{(j)} p(\mathbf{X}_t | \mathbf{N}_t^{(j)}) \quad (7)$$

また、式 (2) において、各サンプルに対する事後確率分布 $p(\mathbf{N}_{0:t}^{(j)} | \mathbf{X}_{0:t})$ は、拡張カルマンフィルタを用いて過去の分布 $p(\mathbf{N}_{0:t-1}^{(j)} | \mathbf{X}_{0:t-1})$ より更新される。

一般に、以上に述べたパーティクルフィルタは、Sequential Importance Sampling (SIS) アルゴリズムと呼ばれている [7]。

再サンプリング (Residual resampling)

重み $w_t^{(j)}$ は、各サンプルごとに割り当てられるが、 $w_t^{(j)}$ の値が微小であるサンプルは、事後確率分布を近似するサンプルとして対応しない。よって図 2 に示すように微小な $w_t^{(j)}$ を持つサンプルを破棄する。また、大きな $w_t^{(j)}$ を持つサンプルを幾つかの同じ値を持つ子サンプルに分割して、親サンプルに割り当てることにより、サンプルの総数を維持する (Residual re-sampling) [7]。これは、大きな $w_t^{(j)}$ を持つサンプルを重要サンプルと見なし、重点的に利用することを意味している。なお、割り当てられる子サンプルの数は $w_t^{(j)}$ の値に依存する。

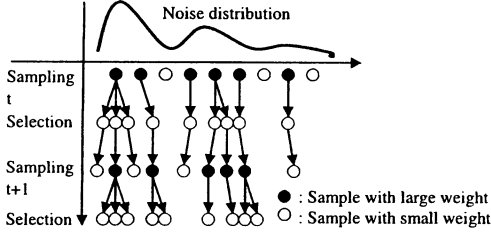


図 2 再サンプリングの概念図

マルコフ連鎖モンテカルロ

再サンプリングによる重要サンプルの割り当てにおいて、時間が経過するに従い、1つの親サンプルに多くの子サンプルが割り当てられる場合がある。最悪の場合、1つの親サンプルに全ての同じ子サンプルが割り当てられ、分布の近似精度が低下する。この問題において、本研究では、マルコフ連鎖モンテカルロ法の Metropolis-Hastings サンプリング [12] を用いて新たな子サンプルを生成し、1つの親サンプルに全ての同じ子サンプルが割り当てられるような状況を回避している。

Metropolis-Hastings サンプリングではまず、サンプル j のパラメータセットを $\Psi_t^{(j)} = (w_t^{(j)}, \mathbf{N}_t^{(j)})$ と定義し、同一の親サンプルから SIS アルゴリズムにより新たなパラメータセット $\Psi_t^{*(j)}$ を発生させる。次に、次式により定義される許容確率 ν と一様乱数 $u \sim U_{[0,1]}$ を発生させる ($U_{[0,1]}$ は 0~1 の範囲の一様分布)。

$$\nu = \min \left\{ 1, w_t^{*(j)} / w_t^{(j)} \right\} \quad (8)$$

その後、次式のように ν と u を比較して、 $\Psi_t^{*(j)}$ を受容するか否かを決定する。

$$\Psi_t^{(j)} = \begin{cases} \Psi_t^{*(j)} & \text{if } u \leq \nu \text{ (accept state transition)} \\ \Psi_t^{(j)} & \text{otherwise (reject state transition)} \end{cases} \quad (9)$$

2.2 Polyak averaging と feedback を用いた状態空間モデル

前節の SIS 処理において、各サンプルに対する事後確率分布のパラメータは、状態空間モデルと呼ばれる信号モデルに基づいて導出された拡張カルマンフィルタにより、過去のパラメータから更新される。一般に、状態空間モデルは目的信号の時間 (状態) 遷移過程を表現した状態方程式と、観測信号の生成過程を表現した観測方程式から構成されている。

まず、クリーン音声 $\mathbf{S}_t$ の背後に確率モデルとして、GMM (Gaussian Mixture Model) が存在し、時刻 t において、GMM 内のある要素分布 k_t (平均 $\mu_{\mathbf{S}, k_t}$, 分散 $\Sigma_{\mathbf{S}, k_t}$) から、パラメータ $\mathbf{S}_{k_t, t}^{(j)}$ が出力されると仮定する。この時、 $\mathbf{X}_t$ の観測方程式は、雑音のサンプル $\mathbf{N}_t^{(j)}$ と誤差 $\mathbf{V}_t^{(j)}$ を伴い、次式のように表現される。

$$\mathbf{X}_t = \mathbf{S}_{k_t, t}^{(j)} + \log \left(\mathbf{I} + \exp \left(\mathbf{N}_t^{(j)} - \mathbf{S}_{k_t, t}^{(j)} \right) \right) + \mathbf{V}_t^{(j)} \quad (10)$$

$$\mathbf{V}_t^{(j)} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{S}^{(j)}, k_t}) \quad (11)$$

一方、 $\mathbf{N}_t^{(j)}$ の時間推移 (状態方程式) は多くの場合、ランダムガウス雑音 $\mathbf{W}_t^{(j)}$ を伴い、次式のような Random walk 過程により表現される。

$$\mathbf{N}_{t+1}^{(j)} = \mathbf{N}_t^{(j)} + \mathbf{W}_t^{(j)} \quad (12)$$

$$\mathbf{W}_t^{(j)} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{W}}^{(j)}) \quad (13)$$

しかし、Random walk 過程はパラメータの時間推移をランダム雑音により規定しているので、パラメータの時間推移を正確に表現できないという問題がある。状態空間モデルに基づいて対象のパラメータを正確に逐次推定するためには、状態方程式の定義が極めて重要であるため、本研究では Random walk 過程に代わって、Polyak Averaging と feedback [9], [10] に基づく、次のような状態方程式を導入する。

$$\mathbf{N}_{t+1}^{(j)} = (1-\alpha)\mathbf{N}_t^{(j)} + \alpha\hat{\mathbf{N}}_t + \alpha\beta \left(\mu_{\mathbf{N}_t}^{(j)} - \mathbf{N}_t^{(j)} \right) + \mathbf{W}_t^{(j)} \quad (14)$$

式 (14) の $\hat{\mathbf{N}}_t$ は、式 (15) により計算されるサンプル $\mathbf{N}_t^{(j)}$ の加重平均であり、 α は忘却係数である。式 (14) の第 1, 第 2 項はサンプル $\mathbf{N}_t^{(j)}$ を平均値に近づけていることを意味しており、サンプルの散らばりを抑制する効果がある。これにより、真値とはかけ離れた値を持つ無意味なサンプルの出現を防ぐことができる。

$$\hat{\mathbf{N}}_t = \sum_{j=1}^J w_t^{(j)} \mathbf{N}_t^{(j)} \quad (15)$$

次に、式 (14) の $\mu_{\mathbf{N}_t}^{(j)}$ は、式 (16) により計算される過去 T 点のサンプルの平均 (Polyak average [9]) である。式 (14) の第 3 項は、Polyak average のフィードバックを示しており、過去の平均値との差分を組み込むことにより、パラメータの時間変化量を表現している [6], [10]。なお、係数 β は、フィードバックのスケーリング係数である。

$$\mu_{N_t}^{(j)} = \frac{1}{T} \sum_{s=t-T+1}^t N_s^{(j)} \quad (16)$$

図3は、Polyak averaging と feedback の概念図を示しており、図中 (a) のように $N_t^{(j)}$ が過去に緩やかな動きをしている場合は、Polyak average $\mu_{N_t}^{(j)}$ と $N_t^{(j)}$ の差分が小さくなる。この場合、式 (14) より $N_t^{(j)}$ から $N_{t+1}^{(j)}$ への変化量は小さいものとして予測される。一方、図中 (b) のように $N_t^{(j)}$ が過去に激しい動きをしている場合は、 $\mu_{N_t}^{(j)}$ と $N_t^{(j)}$ の差分が大きくなる。よって、 $N_t^{(j)}$ から $N_{t+1}^{(j)}$ への変化量は大きいものとして予測される。このように、過去の信号の変化度合いを考慮する Polyak averaging と feedback を導入することにより、Random walk 過程の場合に比べてパラメータの時間推移に対する拘束条件が強化され、より正確なパラメータの逐次推定を行うことができる。

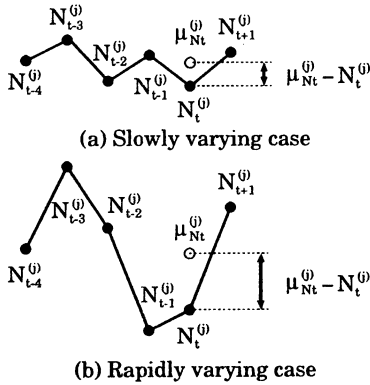


図3 Polyak averaging と feedback の概念図

2.3 Switching dynamical system への拡張

前節にて述べた Polyak averaging に基づく状態空間モデルは、過去の $N_t^{(j)}$ の時間変化状況に応じて未来のパラメータを予測しており、以前の報告 [11] では、Polyak averaging の3つの制御パラメータ (α , β , T) を時間固定のものとして扱っていた。しかし、非定常雑音の時間変化の様相は時々刻々と変化するため、これらのパラメータも時間固定ではなく、時間変動させる必要があるといえる。本稿ではこの問題を、状態空間モデルを Switching dynamical system に拡張することにより解決する。

Switching dynamical system は、それぞれ異なるパラメータ設定を行った複数の Dynamical system から構成されており、フレーム毎に未来予測を行うのに最適な Dynamical system を選択する。本研究では、以下のような方法により Switching dynamical system を構成し、非定常雑音の時間変化をより効果的に表現した状態空間モデルを定義する。

まず、Polyak averaging のパラメータ (α , β , T) がそれぞれ独立に複数の状態をもち、各状態からパラメータの値が出力されるものとする。次に、フレーム毎に現時刻の (パラメータ) 状態 $m_t^{(j)}$ から未来 (遷移先) の状態 $n_{t+1}^{(j)}$ を、状態遷移確率

$a_{mn,t}^{(j)}$ に基づいてモンテカルロサンプリングにより選定し、遷移先の状態からパラメータの値を出力する。その後、出力されたパラメータを用いて Dynamical system (Polyak averaging) を構成し、2.1 節にて述べたパーティクルフィルタ (SIS) による逐次推定を行う。

図4は、パラメータ α の状態遷移の例を示しており、 α が4つの状態 (1, 2, 3, 4) を持ち、それぞれの状態からパラメータ値 (0.05, 0.10, 0.15, 0.20) が出力されるものとする。また、状態遷移は自己ループを含め、全ての状態への遷移を許容する。

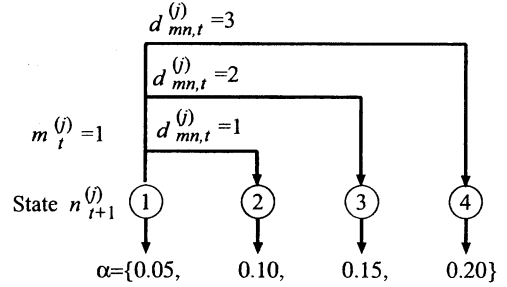


図4 パラメータ α の状態遷移の例

状態遷移確率 $a_{mn,t}^{(j)}$ は次式により定義され、

$$a_{mn,t}^{(j)} = \gamma d_{mn,t}^{(j)} \quad (17)$$

$d_{mn,t}^{(j)}$ は、状態 $m_t^{(j)}$ と $n_{t+1}^{(j)}$ の状態番号の絶対値差分 $d_{mn,t}^{(j)} = |n_{t+1}^{(j)} - m_t^{(j)}|$ であり、 γ ($1 \geq \gamma > 0$) は、各状態への遷移確率を制御する係数である。また、 $\sum_n a_{mn,t}^{(j)} = 1$ となるよう正規化を行う。

式 (17) より、状態遷移確率 $a_{mn,t}^{(j)}$ は、 $d_{mn,t}^{(j)}$ が大きくなるほど減少するため、遷移先が現在のパラメータと大きく異なる値を持った状態になるほど $a_{mn,t}^{(j)}$ が減少し、自己ループの場合に最大となることを意味している。 γ は各状態に割り当てられる $a_{mn,t}^{(j)}$ を制御しており、小さな値になるほど遠隔の状態への遷移確率が減少し、近傍の状態への遷移確率が増加する。また、 $\gamma = 1$ の場合は全ての状態に対して等確率の状態遷移確率が割り当てられる。

以上の状態遷移確率の割り当て及び、遷移先状態の選択は Polyak averaging の3つパラメータ α , β , T に対してそれぞれ独立で行う。

2.4 MMSE 推定による雑音抑圧

以上に述べた手法で逐次推定された雑音を用いて、MMSE 推定法 [3] に基づくクリーン音声の推定 (雑音抑圧) を行う。パーティクルフィルタにより得られた1サンプルのパラメータセット $\Psi_t^{(j)}$ を用いた MMSE 推定結果は次式により得られる。

$$\hat{S}_t^{(j)} = X_t - \sum_{k=1}^K P(k|X_t, (j)) \left(\mu_{X_{k,t}^{(j)}} - \mu_{S,k} \right) \quad (18)$$

上式において、 K は、クリーン音声 GMM の混合分布数で

ある。また、 $\mu_{\mathbf{X}_{k,t}^{(j)}}$ は $\mathbf{X}_t$ の平均ベクトル、 $P(k|\mathbf{X}_t, (j))$ は $\mathbf{X}_t$ の事後確率であり、それぞれ

$$\mu_{\mathbf{X}_{k,t}^{(j)}} = \mu_{S,k} + \log \left(\mathbf{I} + \exp \left(\mathbf{N}_t^{(j)} - \mu_{S,k} \right) \right) \quad (19)$$

$$P(k|\mathbf{X}_t, (j)) = \frac{P_{S,k} \mathcal{N} \left(\mathbf{X}_t; \mu_{\mathbf{X}_{k,t}^{(j)}}, \Sigma_{\mathbf{X}_{k,t}^{(j)}} \right)}{\sum_{k'=1}^K P_{S,k'} \mathcal{N} \left(\mathbf{X}_t; \mu_{\mathbf{X}_{k',t}^{(j)}}, \Sigma_{\mathbf{X}_{k',t}^{(j)}} \right)} \quad (20)$$

により与えられる。 $\Sigma_{\mathbf{X}_{k,t}^{(j)}}$ は $\mathbf{X}_t$ の共分散行列であるが、本研究では $\Sigma_{\mathbf{X}_{k,t}^{(j)}} \simeq \Sigma_{S,k}$ と近似する。

最終的に、推定クリーン音声 $\hat{\mathbf{S}}_t$ は、次式により得られる。

$$\hat{\mathbf{S}}_t = \sum_{j=1}^J w_t^{(j)} \hat{\mathbf{S}}_t^{(j)} \quad (21)$$

3. 実験

3.1 実験条件

実験に用いた雑音重畳音声は、AURORA-2J [13] のクリーン音声 1001 文に、実環境で収録した雑音 [15] を人工的に加算して作成した。使用した雑音は、工場雑音と道路工事雑音であり、それぞれ非定常的な性質が強い雑音である。また、SNR は 20~0dB とした。

音声認識の際の特徴量は 0 次を含む 13 次 MFCC 及び、1 次、2 次の回帰係数を含む 39 次元の特徴量 (CMS 処理有り) であり、音響モデルは、AURORA-2J 標準の HMM (16 状態、20 混合分布) を用いている。音響モデルの学習、認識は HTK ver. 3.2 [14] にて行った。

パーティクルフィルタによる雑音推定及び、MMSE 推定に基づく雑音抑圧法で用いるクリーン音声 GMM は、AURORA-2J のクリーン学習データを用いて学習しており、混合分布数は 512 である (特徴量は 23 次対数メルスペクトル)。式 (13) のパラメータ $\Sigma_{\mathbf{w}}$ は、 $\Sigma_{\mathbf{w}} = \text{diag}(0.0001)$ に設定し、サンプルの総数は、 $J = 50$ とした。また Polyak averaging と feedback で用いる 3 種類パラメータには、それぞれ 4 つの状態を設定し、各状態における出力値は $\alpha = \{0.05, 0.10, 0.15, 0.20\}$ 、 $\beta = \{0.5, 1.0, 1.5, 2.0\}$ 、and $T = \{5, 10, 15, 20\}$ とした。

3.2 実験結果

表 2, 3 に音声認識結果 (単語正解精度) を示す。表中、“HTK”、“ETSI”、“MMSE”、“SEM” は、それぞれ HTK ベースライン (雑音抑圧無し)、ETSI Advanced front-end、MMSE 推定による雑音抑圧 (雑音の逐次推定無し)、逐次 EM アルゴリズムによる雑音の逐次推定を用いた結果を示している。“PF 1”、“PF 2”、“PF 3” は、パーティクルフィルタによる雑音の逐次推定を用いた結果であり、それぞれ状態方程式に、Random walk 過程、Polyak averaging (固定パラメータ)、Switching dynamical system を用いている。また、“SEM”、“PF 1”、“PF 2”、“PF 3” は、雑音の逐次推定後に MMSE 推定による雑音抑圧を行っている。“PF 2”において、Polyak averaging のパラメータを、工場雑音環境では $\alpha = 0.20$ 、 $\beta = 0.5$ 、 $T = 10$ 、

道路工事雑音環境では $\alpha = 0.20$ 、 $\beta = 0.5$ 、 $T = 20$ と設定した。“PF 3”において、状態遷移確率の制御パラメータを、工場雑音環境では $\gamma = 0.5$ 、道路工事雑音環境では $\gamma = 0.6$ と設定した。

表 1 単語正解精度 (工場雑音) (%)

表 2 Word accuracy of factory noise environments (%)

SNR	HTK	ETSI	MMSE	SEM	PF 1	PF 2	PF 3
20 dB	93.61	92.88	96.41	96.50	96.13	96.71	97.54
15 dB	81.12	86.86	88.92	90.55	90.02	91.74	93.43
10 dB	54.81	76.73	74.27	75.07	75.87	82.13	85.08
5 dB	29.47	53.18	50.94	55.60	54.50	64.02	68.41
0 dB	18.73	23.15	24.72	26.48	28.92	38.66	43.20
Average	55.55	66.56	67.05	68.88	69.09	74.65	77.48

表 3 単語正解精度 (道路工事雑音) (%)

SNR	HTK	ETSI	MMSE	SEM	PF 1	PF 2	PF 3
20 dB	96.68	96.90	99.20	99.10	98.34	99.26	99.72
15 dB	89.93	94.81	97.61	97.90	95.61	98.28	98.83
10 dB	70.28	89.81	91.77	92.43	89.84	94.66	96.93
5 dB	38.81	76.02	71.57	77.20	75.28	81.79	86.15
0 dB	22.29	48.48	43.60	51.00	49.43	58.00	63.89
Average	63.60	81.20	80.75	83.53	81.70	86.40	89.10

それぞれの結果より、Switching dynamical system を用いた “PF 3” が最良の結果を示しており、Polyak averaging のパラメータを固定にした “PF 2” に比べて、全体的に高い改善性能を示している。このことから、Switching dynamical system を用いて Polyak averaging のパラメータを時間変動させる事の有効性が確認できる。つまり、雑音の非定常的な特徴を、より詳細にモデル化することが逐次推定において極めて重要な要素であることを示唆している。

また “PF 3” は、クロック周波数 3.2 GHz の CPU (Intel Xeon processor) を用いてほぼ実時間で動作した (“PF 1”、“PF 2” も同等)。一方、“SEM” の処理時間は実時間の 2~4 倍であった^(注1)。このことから、提案手法である “PF 3” は、従来の逐次 EM アルゴリズムに比べて高速で動作し、かつ逐次推定の性能が高い手法であると言える。

表 4, 5 は Switching dynamical system のパラメータ γ を変化した場合の認識結果を示している。表より、 γ を変化させても、雑音種別、SNR に関わらず音声認識性能の変動が 1% 以下であるため、音声認識性能に対する γ の依存性は低いと言える。

提案手法では、Switching dynamical system の状態を直前の状態のみに依存して決定していたが、雑音は連続的な信号であるため、過去の信号の影響を強く受ける。このため、直前の状態のみでなく、より過去の状態の影響について考慮した Switching dynamical system について検討する必要がある。

(注1)：“SEM” は繰り返し推定が行われており、結果が収束するまでの繰り返し回数はデータに依存する。

表 4 Switching dynamical system のパラメータ γ を変化させた場合の単語正解精度 (工場雑音) (%)

SNR	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 1.0$
20 dB	97.39	97.45	97.42	97.45	97.54	97.39	97.30	97.36	97.61	97.82
15 dB	92.91	93.09	92.88	92.78	93.43	93.15	93.55	93.49	93.55	93.52
10 dB	84.43	84.31	84.43	85.08	84.80	84.99	84.77	84.56	84.71	85.05
5 dB	67.95	67.76	68.07	67.70	68.41	68.31	68.10	67.67	68.25	67.82
0 dB	42.31	42.16	42.55	42.31	43.20	42.80	42.43	42.74	43.14	42.43
Average	77.00	76.95	77.07	77.06	77.48	77.33	77.23	77.16	77.45	77.33

表 5 Switching dynamical system のパラメータ γ を変化させた場合の単語正解精度 (道路工事雑音) (%)

SNR	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 1.0$
20 dB	99.72	99.72	99.66	99.72	99.66	99.72	99.63	99.69	99.75	99.72
15 dB	98.99	98.89	98.99	98.93	99.08	98.83	99.14	98.86	98.96	99.02
10 dB	97.05	96.96	97.18	97.18	97.11	96.93	96.96	97.02	97.18	96.87
5 dB	86.18	86.03	86.12	85.75	85.97	86.15	86.37	85.54	85.94	85.88
0 dB	63.37	63.83	63.34	63.43	63.43	63.89	62.97	63.40	63.59	63.37
Average	89.06	89.09	89.06	89.00	89.05	89.10	89.01	88.90	89.08	88.97

4. む す び

本研究では、パーティクルフィルタを用いた非定常雑音の逐次推定及び、抑圧について検討を行った。また、状態空間モデルの状態方程式に Switching dynamical system を導入して、必要なパラメータを時間に応じて変化させることにより、非定常雑音の時間変化を詳細に表現した。提案手法の評価の結果、制御パラメータを時間不変とした Polyak averaging に比べて、大幅な音声認識性能の改善が得られた。

今後、直前の状態だけではなく、より過去の状態を考慮した Switching dynamical system の検討、空間伝達特性（特に移動音源の特性）の逐次推定への応用について検討を行う予定である。

謝辞

本研究は、情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。本研究では、IPJS SIG-SLP 雑音下音声認識評価ワーキンググループが作成した雑音下音声認識評価環境 (AURORA-2J) を利用した。

文 献

- [1] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. on ASSP*, Vol. 27, No. 2, pp. 113-120, Apr. 1979.
- [2] ETSI ES 202 050 V1.1.3, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms," Nov. 2003.
- [3] J. C. Segura, A. de la Torre, M. C. Benitez, and A. M. Peinado, "Model-Based Compensation of the Additive Noise for Continuous Speech Recognition. Experiments Using AURORA II Database and Tasks," *Proc. EuroSpeech '01*, Vol. I, pp. 221-224, Sept. 2001.
- [4] M. Afify and O. Siohan, "Sequential Estimation with Optimal Forgetting for Robust Speech Recognition," *IEEE Trans. on SAP*, Vol. 12, No. 1, pp. 19-26, Jan. 2004.
- [5] K. Yao, K. K. Paliwal, and S. Nakamura, "Noise Adaptive Speech Recognition Based on Sequential Noise Parameter Estimation," *Speech Communication*, Vol. 42, Issue 1, pp. 5-23, Jan. 2004.
- [6] T. A. Myrvoll and S. Nakamura, "Online Cepstral Filtering Using a Sequential EM Approach with Polyak Averaging and Feedback," *Proc. ICASSP '05*, Vol. I, pp. 261-264, March, 2005.
- [7] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking," *IEEE Trans. SP*, Vol. 50, No. 2, pp. 174-188, Feb. 2002.
- [8] K. Yao and S. Nakamura, "Sequential noise compensation by sequential Monte Carlo method," *Proc. NIPS '01*, pp. 1205-1212, Dec. 2001.
- [9] B. T. Polyak and A. B. Juditsky, "Acceleration of Stochastic Approximation by Averaging," *SIAM J. Contr. Optim.*, Vol. 30, No. 4, pp. 838-855, July 1992.
- [10] H. J. Kushner and J. Yang, "Stochastic Approximation with Averaging and Feedback: Rapidly Convergent "On-Line" Algorithm," *IEEE Trans. on AC*, Vol. 40, No. 1, pp. 24-34, Jan. 1995.
- [11] M. Fujimoto and S. Nakamura, "Particle Filtering and Polyak Averaging-based Non-stationary Noise Tracking for ASR in Noise," *Proc. ASRU '05*, Nov. 2005. (to appear)
- [12] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, Vol. 57, No. 1, pp. 97-109, Jan. 1970.
- [13] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, "AURORA-2J, An Evaluation Framework for Japanese Noisy Speech Recognition," *IEICE Transactions on Information and Systems*, Vol. E88-D, No. 3, pp. 535-544, Mar. 2005.
- [14] HTK Web site, <http://htk.eng.cam.ac.uk/>
- [15] 遠藤 俊樹, 堀内 俊治, 清水 徹, 中村 哲, "ATR 実環境雑音 DB - ATRANS -を用いた雑音重畳音声認識実験," 情報処理学会研究報告, SLP-57-8, pp. 43-48, July 2005.