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

2007－ＳＬＰ－69（34）
2007／12／20

パーティクルフィルタを統合した複数雑音モデル合成による雑音抑圧手法

實廣 貴敏† 鳥山 朋二† 小暮 潔†

† ATR 知識科学研究所 〒619-0288 京都府「けいはんな学研都市」光台 2-2-2
E-mail: †{takatoshi.jitsuhiro,toriyama,kogure}@atr.jp

あらまし 実環境下での音声認識では，認識対象の音声，異なる種類の周囲雑音，および，それらが重なり合ったものが入力される．雑音抑圧の従来法では一つの雑音源と仮定することが多く，複数種類の雑音が存在する場合には対応が困難であった．このような実環境下での雑音に対応するため，我々は以前に複数種類の雑音モデルおよびその合成モデルを用いた雑音抑圧手法を提案した．しかし，学習データから推定された雑音モデルをベースにしているため，雑音の変動や学習データに含まれない未知の雑音分布への対応が不十分であった．そこで，パーティクルフィルタを取り入れ，入力に応じた雑音の推定機構を組入れる．具体的には，前フレームから推定される雑音パーティクルとともに，複数雑音合成モデルにより検出された雑音に対するパーティクルを用いる．検出された時点でその雑音モデルを事前分布としてパーティクルをサンプリングする．その雑音が継続して検出されるフレームでは，前フレームから推定されたパーティクルを用いる．これにより，変動した雑音や未知の雑音も近似的に推定することができる．E-Nightingale タスクにおいて，従来法に対し，確実な精度向上を確認できた．

キーワード 音声認識，雑音抑圧，モデル合成，パーティクルフィルタ，E-Nightingale プロジェクト

Noise Suppression Using Multi-noise Model Compositions Integrating Particle Filtering

Takatoshi JITSUHIRO†, Tomoji TORIYAMA†, and Kiyoshi KOGURE†

† ATR Knowledge Science Laboratories 2-2-2, Hikaridai, "Keihanna Science City," Kyoto, 619–0288 Japan
E-mail: †{takatoshi.jitsuhiro,toriyama,kogure}@atr.jp

Abstract We propose a noise suppression method based on multi-model compositions using particle filtering. In real environments, input speech for speech recognition includes many kinds of noise signals. For such noisy speech, we have proposed Multi-Model Noise Suppression (MM-NS) that uses many kinds of noise models and their compositions obtained from training data. However, MM-NS only uses static property of noise models, and it is difficult to handle unknown noise distributions. We introduce a particle filter into MM-NS. The distributions of noise models is used as prior distributions of particle filtering. It makes more accurate estimation of noise signals for input data. We evaluated this method using the E-Nightingale task, which contains voice memoranda spoken by nurses during actual work at hospitals. The proposed method obtained higher performance than the original MM-NS.

Key words speech recognition, noise suppression, model composition, particle filter, E-Nightingale project

1. まえがき

本研究所では E-Nightingale プロジェクトと称して，日常の活動や状況を理解する基礎技術の確立を目指している [1]．特に医療現場に焦点を合わせ，現場での看護師による各自の作業についての音声メモを集め，業務分析に役立てている [2]．現在，この業務分析の自動化を目指し，音声メモの音声認識を試みている．しかし，この音声メモは実環境下および実作業中での収録であるため，多くの雑音を含んだ自由発話に近い音声となっており，音声認識としては大変困難なタスクである．

雑音重畳音声の認識のために，これまで多くの耐雑音音声認識手法が提案されてきた．定常雑音に対しては，Spectral Subtraction [3] や Parallel Model Combination (PMC) [4] がある．Gaussian Mixture Model (GMM) を用いた Minimum Mean-Squared Error (MMSE) に基づく手法 [5] はフレーム同期で処理を行うことで，入力音声に対する変動に対応できる．さらに，最近では，非定常的な雑音に対する研究が盛んになっている ([6] [7] [8] など)．これらの手法は一般に一種類の雑音の

みを考慮，または一つのモデルで雑音をモデル化できるという仮定がある．しかし，実環境下では，定常的な雑音だけでなく，突発的な雑音も多い．また，その他に，どのように入力から実際の雑音を推定するかという問題もある．

これまで我々は，複数モデル合成を用い，マルチパス探索で最尤ラベル系列を得て，複数雑音モデルにより雑音抑圧を行う手法 (Multi-Model Noise Suppression, MM-NS) を提案してきた [9] [10]．学習データから得られた雑音モデルおよびそれらを組み合せた合成モデルを用い，マルチパス探索による雑音ラベル認識により雑音ラベル系列を得て，それを元に GMM を用いた MMSE 雑音抑圧手法 [5] を複数合成モデルへ拡張した手法により，雑音抑圧を行う．雑音重畳音声モデルを学習ではなく，モデル合成により得ることで，学習データにはない SNR での音声と雑音の合成された分布をも作成できる．さらに，GMM による MMSE 雑音抑圧手法により，入力音声に対する事後確率により各分布に重み付けされ，雑音が推定される．したがって，入力雑音重畳音声を反映した雑音モデルで雑音抑圧処理が行われることになる．ただし，各雑音モデルの混合分布数を増加することで性能向上を図ることができることから，各雑音モデルの持つ粒度は詳細なほどよいことが分かる．また，この手法では，学習データに存在しない未知の雑音に対して近い分布が利用される可能性はあるが，厳密な対処は難しい．入力雑音の変動，さらには未知雑音へのより積極的な対応が必要と考えられる．

そこで，今回，さらに本提案手法にパーティクルフィルタを導入する．変動する雑音や，雑音モデルにおいてモデル化できていない未知雑音の分布に対応することを可能にする．従来研究として，推定した雑音を音響モデルに合成して認識する手法 [6]，雑音を差し引く前処理的手法 [8] がある．これらの手法では基本的に一種類の変動する雑音のみを考慮している．パーティクルフィルタは基本的に時間的に非線形に変動する分布を逐次推定していく手法である．非線形とはいえ，時間的な変動の推定を行うため，全く確率分布の異なる，突発的な雑音のような大きな変動を推定することは必ずしも容易でない．そこで，雑音ラベル認識結果を利用し，ある雑音が検出されたフレームでは，その雑音モデルを事前分布としてパーティクルをサンプリングすることにする．同じ雑音が継続する間は従来法と同様，拡張カルマンフィルタにより，各パーティクルにおいて前フレームからの推定値を利用して推定する．雑音モデルからのサンプリングにより，突発的な変動に対してもより正確な推定を行うことが可能になる．

以下，2. では，本研究の動機となる E-Nightingale プロジェクトの音声認識タスクについて概要を述べる．3. では，提案法について述べ，4. では，実際に病院で収録されたデータを用いた評価を行う．5. でまとめを述べる．

2. E-Nightingale プロジェクト

近年，医療事故が大きな社会的問題となっている．本プロジェクトでは，看護業務を支援するため，ウェアラブル・コンピュータやセンサー・ネットワークを用いた技術の研究開発を行っている．その一つとして，日常の看護活動を分析するため，看護師による業務内容に関する音声メモを収集している．業務内容を短い文で，胸に取り付けた小型マイクを通して IC レコーダに録音してもらう．この発話を元に，各看護師の作業内容や時間を計測する．ヘッドセットマイクロホンは業務の邪魔になり，現場の要望もあって，現在はピンタイプのマイクを用いている．そのため，収録音声の SN 比が悪いものも多く，平均 10dB 以下である．図 1 に収録音声の例を示す．録音開始ボタンを押してもらい，10 秒間録音する．発話を促すビープ音の後，「業務調整終了」と認識対象となる発話が録音され，その直後，同僚に話しかけられ，本人と同僚の声が録音されている．背景雑音だけでなく，状況に応じて様々な音が収録される．また，認識対象発話自体は比較的丁寧なものが多いが，小声や曖昧な発声もある．さらに対象外音声は自由発話であり，方言なども含む．音声認識としては困難なタスクであり，一度に解決は難しい．本研究では，雑音抑圧を重点的に検討する．

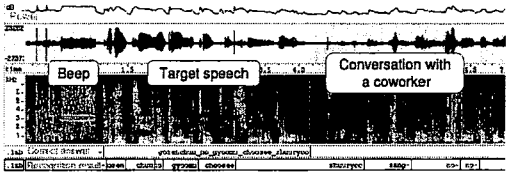

図 1 認識対象音声を含んだ音声サンプル．ビープ音により音声入力を促す．音声メモを録音後，同僚と話している．

Fig. 1 Wave sample including a target speech. Beep prompts speech input. Speaker talked with her coworker after recording her voice memorandum.

3. パーティクルフィルタを用いた複数モデル雑音抑圧手法

3.1 複数モデル雑音抑圧手法

複数モデル雑音抑圧 (MM-NS) のキーポイントは重畳されている雑音を表現するためモデル合成を用いることである．そのモデルを用い，認識処理により音声および雑音種類の検出と，その情報に基づいてモデルベースの雑音抑圧を行う．図 2 に例を示す．図中，上半分は収録音声に対してラベルを付与したものを模式図で示す．時間軸上で重なりを表現するために多層化する．図の下半分のように，このラベルを元に重なり区間を別々にモデル化する．同じ “target utterance” 区間であっても “beep” や “other's utterance” との重畳区間は別のラベル（モデル）とする．このラベルを「マルチラベル」，それに対応するモデルを「マルチモデル」と呼ぶことにする．重なりのない部分は音声データからパラメータを推定し，重畳区間はモデル合成でモデル化する．図 3 に本雑音抑圧手法の流れ図を示す．音響モデルとして前述のマルチモデルを用い，マルチラベルを単位とした辞書および N-gram モデルを学習データから得て，一般的な音声認識と同様なマルチパス探索を行うことで得られた最尤マルチラベル系列とその時刻情報から，各フレームにラベルに対応するモデルが割り当てられる．そのモデルを用いる

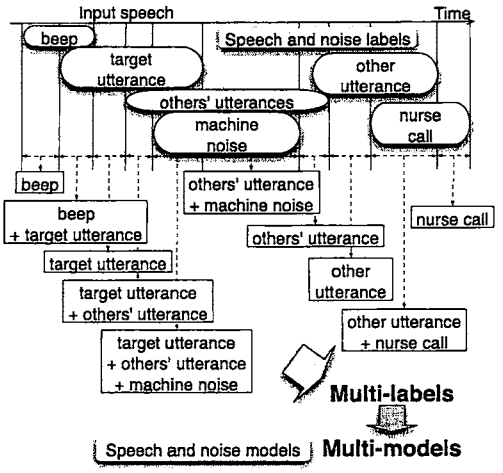

図 2 多層化ラベルおよびマルチラベルの例

Fig. 2 Examples of multi-layered labels and multi-labels

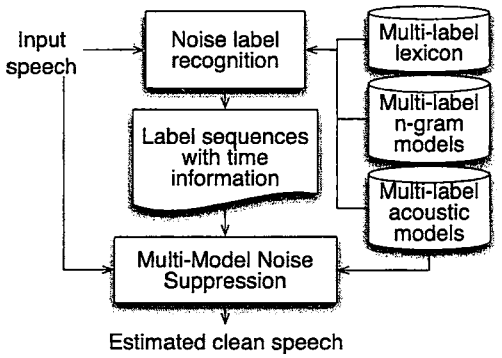

図 3 複数モデル雑音抑圧手法の概要

Fig. 3 Overview of Multi-Model Noise Suppression

ことで，GMM に基づく雑音抑圧手法を適用できる．

3.2 雑音抑圧処理

この節では，雑音抑圧処理自体の具体的な手法について述べる．対数メルスペクトル領域において，第 t フレームでの観測雑音重畳音声ベクトルを $\mathbf{x}_t$，クリーン音声ベクトルを $\mathbf{s}_t$，第 n 雑音ベクトルを $\mathbf{n}_t(n)$ とする．まず，これらの状態空間モデルを定義する．観測モデルは下記のように表される．

$$\begin{aligned}\mathbf{x}_t &= \mathbf{s}_t + \log\left[\mathbf{I} + \exp\left\{\log\left(\sum_{n=1}^{N}\exp\left(\mathbf{n}_t(n)\right)\right) - \mathbf{s}_t\right\}\right] \\ &\quad + \mathbf{v}_t \\ &= \mathbf{s}_t + \log\left\{\mathbf{I} + \exp(\mathbf{n}_t - \mathbf{s}_t)\right\} + \mathbf{v}_t \\ &= \mathbf{s}_t + g(\mathbf{s}_t, \mathbf{n}_t) + \mathbf{v}_t \qquad (1) \\ &= f(\mathbf{s}_t, \mathbf{n}_t) + \mathbf{v}_t \qquad (2)\end{aligned}$$

$$\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{x}}) \qquad (3)$$

ここで，$\mathbf{n}_t$ は t フレームでのすべての雑音を含む合成雑音である．$g(\mathbf{s}_t, \mathbf{n}_t)$ はクリーン音声 $\mathbf{s}_t$ と雑音重畳音声 $\mathbf{x}_t$ とのミスマッチ成分である．$\mathbf{v}_t$ は観測ノイズを示し，$\boldsymbol{\Sigma}_{\mathbf{x}}$ は $\mathbf{x}_t$ の共分散行列である．

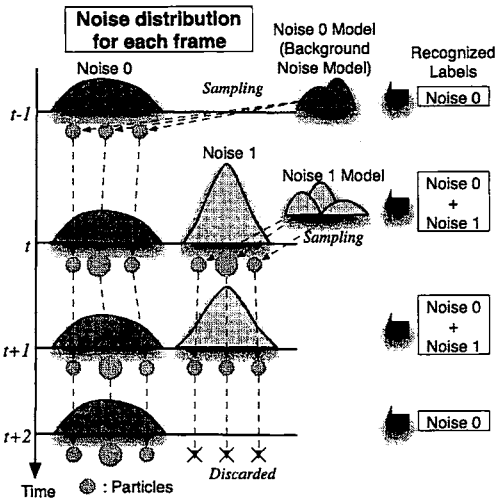

図 4 複数モデル雑音抑圧手法におけるパーティクルフィルタ

Fig. 4 Particle filtering for MM-NS

この手法は音声区間検出を含むので，雑音抑圧は各区間に対し別々に行うことができる．各混合分布に対するミスマッチ成分を下記のように定義する．

$$g(\mathbf{s}_{t,l}, \mathbf{n}_{t,l}) = \begin{cases} \mu_{\mathbf{x},l} - \mu_{\mathbf{s},l} & \text{for target utterances} \\ \mu_{\mathbf{x},l} - \varepsilon & \text{for the others} \end{cases} \qquad (4)$$

ここで，$\mu_{\mathbf{x},l}$ は雑音重畳音声の第 l 平均ベクトルで，対象発話に対して，クリーン音声の第 l 平均ベクトル $\mu_{\mathbf{s},l}$ と推定された雑音モデルの合成で生成される．ε は小さい正の値であり，雑音抑圧後の残差パワーを制御できる．

さらに，雑音 $\mathbf{n}_t(n)$ の第 l 混合分布 $\mathbf{n}_{t,l}(n)$ のシステムモデルがランダムウォーク仮定で下記のようにモデル化できると仮定する．

$$\mathbf{n}_{t+1,l}(n) = \mathbf{n}_{t,l}(n) + \mathbf{w}_t \qquad (5)$$

$$\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{n}(n),l}) \qquad (6)$$

ここで，$\mathbf{w}_t$ はシステムノイズであり，$\boldsymbol{\Sigma}_{\mathbf{n}(n),l}$ は雑音 $\mathbf{n}_{t,l}(n)$ の共分散行列である．

状態空間モデルを式 (2) と (5) で表し，文献 [8] と同様にパーティクルフィルタを定義，MM-NS に統合する．提案手法は同様に拡張カルマン・パーティクルフィルタ，residual sampling および Markov Chain Monte Carlo (MCMC) を用いる．MM-NS 法へ導入するために，雑音モデル分布を各パーティクルの事前分布として用いる．図 4 に MM-NS でのパーティクルフィルタリングの概要を示す．はじめに，雑音分布に対するパーティクルを背景雑音モデルからサンプリングする．この段階では従来法 [8] と同様である．次に，雑音ラベル認識により時刻 t において新たな雑音 "Noise 1" が検出されたフレームでは，あらか

じめ学習データから作成し，保持している "Noise 1" のモデルから新たなパーティクルをサンプリングする．次フレームでは，"Noise 1" が継続して検出されているので，同じような分布を持つ雑音が存在していると考えられ，時刻 t の "Noise 1" パーティクルから推定しやすいと考えられる．そこで，時刻 $t+1$ の "Noise 1" パーティクルは従来法と同様，拡張カルマンフィルタを用いて推定される．時刻 $t+2$ にて "Noise 1" が検出されなくなったとき，そのパーティクルも推定に用いるパーティクルから外す．以上のようにすることで，提案法は突発的な雑音を扱うことが可能になる．また，各フレームでパーティクルフィルタリングを行う際，入力雑音重畳音声とより近いものを推定するために，クリーン音声と雑音間のパーティクルの組合せを考慮し，推定に用いることで推定精度向上を図る．

下記に本手法の雑音抑圧手順を示す．

パーティクルフィルタリングに基づく複数モデル雑音抑圧手法：

I. 初期化

(a) フレーム番号: $t=0$

(b) 各パーティクル $\mathbf{n}_0^{(i)}(i=1,\ldots,I)$ を背景雑音モデルからサンプリング．

II. 各フレームに対して，$t=1,2,\ldots,T$,

(a) Importance sampling step (パーティクルフィルタ):

i. もし原フレームが音声区間ならば，
クリーン音声モデルから音声パーティクル $\mathbf{s}_t^{(i)}$ をサンプリング．

$$\mathbf{s}_t^{(i)} \sim \sum_{l=1}^{L_{\mathbf{s}}} w_{\mathbf{s},l}\mathcal{N}(\mu_{\mathbf{s},l},\Sigma_{\mathbf{s},l}),\quad i=1,\ldots,I$$

ここで，$\mu_{\mathbf{s},l},\Sigma_{\mathbf{s},l}$ はそれぞれ音声の第 l 平均ベクトル，共分散行列．$L_{\mathbf{s}}$ は混合分布数．

ii. もし新たな雑音 $\mathbf{n}(n)$ が検出されたら，雑音パーティクルをサンプリング．

$$\mathbf{n}_{t-1}^{(j)} \sim \sum_{l=1}^{L_{\mathbf{n}(n)}} w_{\mathbf{n}(n),l}\mathcal{N}(\mu_{\mathbf{n}(n),l},\Sigma_{\mathbf{n}(n),l}),\quad j=1,\ldots,J$$

ここで，$\mu_{\mathbf{n}(n),l},\Sigma_{\mathbf{n}(n),l}$ はそれぞれ第 n 雑音モデルの第 l 平均ベクトル，共分散行列．$L_{\mathbf{n}(n)}$ は混合分布数．

iii. $\mathbf{s}_t^{(i)}$ と $\mathbf{n}_{t-1}^{(j)}$ との合成で雑音重畳音声パーティクルの組を作成．総パーティクル数は $K=I\times(J+1)$.

iv. For $k=1,\ldots,K$
拡張カルマンフィルタにより各パーティクルを更新．$\hat{\mathbf{n}}_t^{(k)}$ と $\hat{\Sigma}_{\mathbf{n}_t}^{(k)}$ を推定．

v. For $k=1,\ldots,K$, $w_t^{(k)} \propto w_{t-1}^{(k)} p(\mathbf{x}_t|\mathbf{n}_t^{(k)})$.

vi. For $k=1,\ldots,K$, 正規化重み $\hat{w}_t^{(k)}$ 計算．

(b) Residual sampling step:
高/低 importance weight を持つパーティクルを増殖/抑圧．

(c) MCMC step:
Metropolis-Hastings sampling を適用．

(d) 雑音事後分布の推定:
パーティクルから事後分布を推定する．

$$p(\mathbf{n}_{0:t}|\mathbf{x}_{0:t}) \simeq \sum_{k=1}^{K} \hat{w}_t^{(k)} p(\mathbf{n}_{0:t}^{(k)}|\mathbf{x}_{0:t}) = \mathcal{N}(\mu_{\hat{\mathbf{n}}_t},\Sigma_{\hat{\mathbf{n}}_t}),$$

ここで，$\mu_{\hat{\mathbf{n}}_t},\Sigma_{\hat{\mathbf{n}}_t}$ はそれぞれ雑音モデルの推定平均ベクトル，共分散行列．

(e) クリーン音声の GMM に基づく MMSE 推定:
ミスマッチ成分を用いてクリーン音声の推定を行う．

$$\hat{\mathbf{s}}_t = \mathbf{x}_t - \sum_{l=1}^{L_{\mathbf{s}}} P(l|\mathbf{x}_t) g(\mathbf{s}_{t,l},\hat{\mathbf{n}}_{t,l})$$

$$P(l|\mathbf{x}_t) = \frac{w_{\mathbf{s},l}\mathcal{N}(\mathbf{x}_t;\mu_{\mathbf{x},l},\Sigma_{\mathbf{x},l})}{\sum_m^{L_{\mathbf{s}}} w_{\mathbf{s},m}\mathcal{N}(\mathbf{x}_t;\mu_{\mathbf{x},m},\Sigma_{\mathbf{x},m})}$$

ここで，$\mu_{\mathbf{x},l},\Sigma_{\mathbf{x},l}$ はそれぞれ雑音重畳音声モデルの平均ベクトル，共分散行列．これらは 1 次 Taylor 級数展開 [11] によりクリーン音声と推定された雑音モデル $\mathcal{N}(\mu_{\hat{\mathbf{n}}_t},\Sigma_{\hat{\mathbf{n}}_t})$ から推定される．

この手法は現在の雑音分布を推定できる．つまり，単一分布が得られる．従来の MM-NS では雑音モデルに対して混合分布の計算が必要であったが，本手法では必要がなく，この点では演算が簡単になっている．

4. 実　　験

4.1 実験条件

評価実験に用いるデータはある病院にて，看護師が実作業を行いつつ，録音したものである．初日分を評価データとし，2 日目分を雑音モデル学習および話者適応データとした．各音声データは 10 秒間の長さで，認識対象となる発話を含む．表 1 に詳細な実験条件を示す．病院にて 32kHz サンプリング周波数，16 bit で収録後，16kHz へダウンサンプリングしたものを本実験で用いた．勤務シフトの関係で，評価データと話者適応データをそろえると，評価話者は女性 8 名となった．

雑音抑圧では，HTK Ver.3.3 を特徴量抽出と GMM 学習に用いた．24 次対数メルフィルタバンクの出力 "FBANK" を特徴量として用いた．雑音ラベル認識には，FBANK モデルから変換して得た MFCC モデルを用いた．クリーン音声モデルとして話者適応 GMM を用いた．その他の音声や雑音モデルには 4, 8, 12 混合分布を持つ GMM を用いた．共分散行列としてすべて対角共分散行列を用いた．学習データから 32 種類の雑音モデルが得られ，合成モデルも含むモデルの合成数は 194 であった．合成モデルにおいて，一つのモデルに合成されたモデルの数は最大 3 個であった．また，背景雑音は各入力音声に対して推定され，すべてのモデルに合成された．これら 194 モデルをマルチラベル辞書のエントリーとした．未知マルチラベル率は 3.77%であった．

また，パーティクルフィルタに関して，従来法 [8] では 300 パーティクル，提案法では 110 パーティクルを用いた．これは雑音抑圧処理がほぼ同程度となる設定である (Intel Pentium-D 3.2 GHz での計測で Real Time Factor が約 2.5)．提案法では，雑音ラベル認識の処理時間も含んでいる．

音声認識器などのツールには，ATR 音声言語コミュニケー

表 1 実験条件
Table 1 Experimental conditions

共通	
録音装置	IC レコーダ: COWON Japan 製 iAUDIO G3 マイクロホン: RASTA BANANA RBENS02 (携帯電話用イヤホンマイク) 帯域: 100Hz〜10kHz
分析条件	16kHz サンプリング周波数, 16 bit フレーム周期 10 ms, フレーム長 20 ms
評価データ	女性 8 名 (208 発話, 1,051 単語)
雑音ラベル認識系	
ツール	HTK Ver.3.3 (GMM 特徴量・学習) ATRASR Ver.3.6 (認識器, N-gram 学習)
特徴量	24 次元メルフィルタバンク出力 (FBANK) (探索時および雑音抑圧時)
	12 次元 MFCC および 0 次 MFCC (探索時のみ)
音響モデル	32 基本音声または雑音 GMM 学習データ: 約 1 時間 (評価データ以外) 162 合成モデル
	クリーン音声 GMM: 話者独立 (SI): 512 混合分布 話者依存 (SD): 約 200 混合分布
	※探索特徴量が MFCC の場合, モデル上で FBANK から MFCC へ変換
言語モデル	マルチラベル bigram, マルチラベル trigram テストセット・パープレキシティ: bigram: 8.08, trigram: 6.47 未知語率: 3.77%
	学習データ: 354 発話
辞書	194 マルチラベル
パーティクル数	従来法 PF: 300 提案法 PF-MM-NS: 110
音声認識系	
ツール	ATRASR Ver.3.6
特徴量	12 次元 MFCC, 12 次元Δ MFCC, Δ対数パワー 発話単位 CMS(Cepstral Mean Subtraction)
音響モデル	音素 HMM: 2,086 状態, 5 混合分布 無音 HMM: 3 状態, 10 混合分布
音響モデル 学習データ	構造学習: 37 時間 ATR 旅行会話 (TRA), 音素バランス文
	再学習: 21 時間 (上記から女性音声のみ)
言語モデル	単語 bigram, 単語 trigram (姓, 名のみそれぞれクラス化)
	テストセット・パープレキシティ: bigram: 39.4, trigram: 39.3 未知語率: 2.36%
言語モデル 学習データ	E-Nightingale データ 9 日分, 9,936 発話 (評価データ以外)
辞書	2,636 単語

ション研究所で開発された ATRASR 大語彙音声認識システム Ver.3.6 を用いた. 音声認識用音響モデルの構造学習には, MDL-SSS [12] を用いた. この実験では話者が女性だけのため, 再学習で作成した 5 混合分布の女声モデルのみを用いた. 話者適応手法として, 音声認識系音響モデルには MAP-VFS [13] を用いた.

表 2 雑音モデルでの平均総分布数
Table 2 Average total # of distributions in noise models

#mix. for each model	4	8	12
Total # distributions	70,871	252,547	553,678

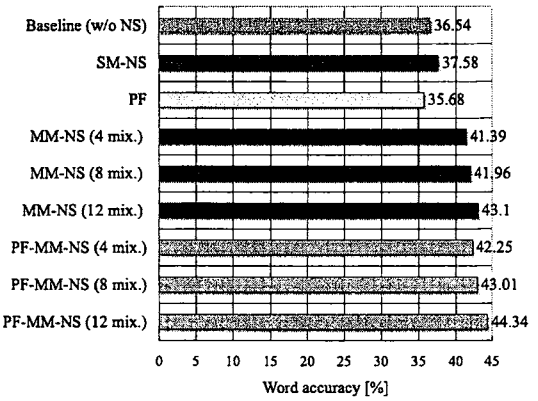

図 5 単語認識率
Fig. 5 Word accuracy

4.2 異なる混合分布数の雑音モデルにおける比較

表 2 は, MM-NS 法で用いた雑音抑圧用モデルの総分布数の評価話者に対する平均を示している. 分布数は各雑音モデルの混合数に対して指数関数的に増加することが分かる. 今回用いた音声認識用の音響モデルに含まれる総分布数が 10,460 であることを考えると, それぞれ大変大きな分布数であるといえる. 図 5 に従来法, 提案法による単語認識率を示す. "SM-NS" は雑音モデルとして単一分布を用いた "Single-Model Noise Suppression" 手法を示す. この単一分布としては各入力音声の開始 100 ms から推定されたものを用いた. "PF" はパーティクルフィルタを用いた従来法 [8] を示す. この方法はこのタスクでは baseline より良い精度を得られなかった. 多くの挿入エラーが生じたためであり, 突発的な雑音に対して追跡することが困難であると考えられる. "PF-MM-NS (4 mix.)", "PF-MM-NS (8 mix.)", "PF-MM-NS (12 mix.)" は, それぞれ提案法において, 各雑音モデルの混合分布数が 4, 8, 12 である場合を示す. これらのパターンは同じモデルを用いた従来法の MM-NS, "MM-NS (4 mix.)", "MM-NS (8 mix.)", "MM-NS (12 mix.)", それぞれの精度を 1%程度, 上回った. "PF-MM-NS (12 mix.)" は baseline に対し, 12.3%のエラー改善率を得た. また, "PF-MM-NS (4 mix.)" と "MM-NS (8 mix.)", あるいは, "PF-MM-NS (8 mix.)" と "MM-NS (12 mix.)" との比較では, 提案法が同程度の性能を少ない分布数 (それぞれ, 28%, 46%) で得られることが分かった.

4.3 提案法の効果を見るためのより詳細な比較

次に, より詳細に提案法の効果を見るために, 中間的な処理パターンを 2 つ評価した. 1 つは, MM-NS における音声区間

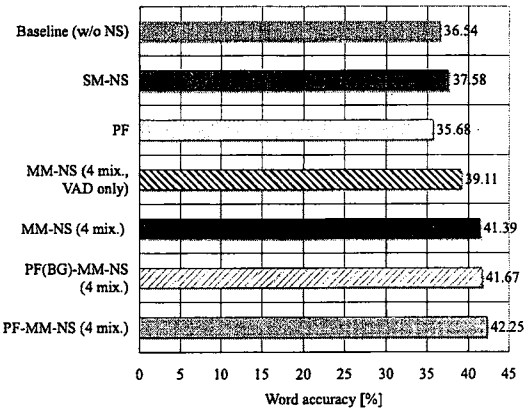

図 6 PF-MM-NS 法の効果を見るための比較

Fig. 6 Comparison to confirm effectiveness of PF-MM-NS.

検出だけの効果を見るために，ラベル認識による音声区間検出を行い，音声区間以外では雑音抑圧処理を行うが，音声区間では何もしない，つまり，雑音抑圧を行わない場合 "MM-NS (4 mix., VAD only)" である．もう 1 つは，提案法 "PF-MM-NS (4 mix.)" のパーティクルフィルタにおいて，雑音に応じたパーティクルを用いない，すなわち，背景雑音モデルから得たパーティクルのみで逐次雑音を推定・抑圧する場合 "PF(BG)-MM-NS (4 mix.)" である．なお，このときのパーティクル数は提案法 "PF-MM-NS (4 mix.)" と同じ 110 とした．

図 6 にそれぞれのパターンに対する単語認識率を示す．上記の 2 つ以外は図 5 に含まれるものと同じ結果である．"MM-NS (4 mix., VAD only)" では，音声区間検出を行うことで "Baseline (w/o NS)" からの改善があったが，雑音抑圧処理を行う "MM-NS (4 mix.)" よりは精度が低かった．複数モデルによる雑音抑圧の効果があることがわかる．"PF(BG)-MM-NS (4mix.)" は逐次雑音推定を行うことで，"MM-NS (4 mix.)" に若干上回る精度を得たが，提案法 "PF-MM-NS (4 mix.)" には及ばなかった．パーティクルフィルタにおいて，雑音種類に応じたモデルをパーティクルの事前分布として用いる効果があるといえる．

5. ま と め

我々が従来提案してきた複数モデル雑音抑圧手法 (Multi-Model Noise Suppression, MM-NS) では，実環境で音声データと同時に収録される複数種類の雑音を扱うことが可能である．しかし，あらかじめ学習データから得られた雑音モデルやその組合せで得られる合成モデルを元に雑音抑圧をするため，学習データで得られない組合せや未知の雑音分布に対して対応することができなかった．そこで，本報告ではパーティクルフィルタに基づく複数モデル雑音抑圧手法を提案した．パーティクルフィルタにより，入力音声から雑音分布を推定することができる．ただし，本タスクで多く見られる突発的な雑音に対応するために，雑音ラベル認識結果を利用し，新たな雑音が検出されたら，その雑音モデルを事前分布として新たなパーティクルをサンプリングする．同じ種類の雑音が継続するうちは拡張カルマンフィルタにより推定し，検出できなくなったら，そのパーティクルは消去する．これにより，学習データから得られた先験的知識を利用し，かつ，入力音声から得られる情報を用い，動的に雑音推定を行うことが可能になる．実験結果から，本提案手法は同じ雑音モデルを用いた従来型の MM-NS 法より高い精度を得ることができることがわかった．また，詳細な比較実験により，MM-NS 法での音声区間検出による効果，複数雑音モデルを用いる効果，パーティクルの事前分布として雑音に応じたモデルを用いる効果を示すことができた．

謝辞 本研究は独立行政法人 情報通信研究機構の研究委託「日常行動・状況理解に基づく知識共有システムの研究開発」により実施したものである．実験に協力していただいた医療機関の方々，また，音声認識ツールの提供およびご助言をいただく ATR-SLC の諸氏，およびデータベース作成や日頃ご議論いただく ATR-KSL の諸氏に感謝する．

文 献

[1] K. Kogure, "Toward a knowledge sharing system based on understanding everyday activities and situations – Introduction to the E-Nightingale Project –," Proc. of the Workshop on Knowledge Sharing for Everyday Life 2006 (KSEL2006), pp. 1–8, 2006.

[2] H. Ozaku, A. Abe, K. Sagara, N. Kuwahara, K. Kogure, "A task analysis of nursing activities using spoken corpora," Advances in Natural Language Processing (Ed. A. Gelbukh), Research in Computing Science 18, pp. 125–136, Instituto Politecnico Nacional, Mexico, 2006.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol. 27, no. 27, pp. 113–120, 1979.

[4] M. F. J. Gales, "Model-based techniques for noise robust speech recognition," PhD thesis, University of Cambridge, 1995.

[5] J. C. Segura, A. de la Torre, M. C. Benitez, A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks," Proc. of EUROSPEECH2001, vol. 1, pp. 221–224, 2001.

[6] K. Yao, S. Nakamura, "Sequential noise compensation by sequential Monte Carlo method," Advances in Neural Information Processing Systems 14. MIT Press, 2002.

[7] K. Yao, K. K. Paliwal, S. Nakamura, "Noise adaptive speech recognition based on sequential noise parameter estimation," Speech Communication, vol. 42, no. 1, pp. 5–23, 2004.

[8] M. Fujimoto, S. Nakamura, "A non-stationary noise suppression method based on particle filtering and Polyak averaging," IEICE Trans. Inf. & Syst., vol. E89-D, no. 3, 2006.

[9] 實廣, 鳥山, 小暮, "複数の雑音合成モデルを用いた探索に基づく雑音抑圧手法," 信学技報, vol. SP2007-16, pp. 49–54, 2007.

[10] 實廣, 鳥山, 小暮, "複数雑音合成モデルによるマルチパス探索に基づく雑音抑圧手法," 音講論, pp. 151–154, 2007.

[11] P. J. Moreno, "Speech recognition in noisy environments," PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, 1996.

[12] T. Jitsuhiro, T. Matsui, S. Nakamura, "Automatic generation of non-uniform HMM topologies based on the MDL criterion," IEICE Trans. on Information and Systems, vol. E87-D, no. 8, pp. 2121–2129, 2004.

[13] M. Tonomura, T. Kosaka, S. Matsunaga, "Speaker adaptation based on transfer vector field smoothing using maximum a posteriori probability estimation," Computer Speech and Language, vol. 10, pp. 117–132, 1996.