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

2008－SLP－74（3）
2008 / 12 / 9

音声区間検出と雑音抑圧の統合法を用いた雑音下音声認識

藤本 雅清† 石塚健太郎† 中谷 智広†

† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
〒 619–0237 京都府相楽郡精華町光台 2–4
E-mail: {masakiyo, ishizuka, nak}@cslab.kecl.ntt.co.jp

あらまし 本研究では，雑音下音声認識における頑健なフロントエンド処理について述べる．提案するフロントエンド処理は音声区間検出（VAD: Voice Activity Detection）と雑音抑圧を統合した処理となっており，(1) 確率モデルの共有，(2) 音声／非音声状態確率を用いた Wiener フィルタ設計，(3) 雑音抑圧音声を用いた VAD 性能の改善の 3 点が手法を構成する大きな要素となっている．また提案手法は逐次処理によりフレーム遅延無しで処理を行うことが可能である．本研究では提案手法を用いることにより，連続発話音声の認識をフレーム遅延無しで実行し，かつ認識性能の大幅な改善が得られることを示す．また，CMN と音響モデルの逐次適応との併用による評価についても述べる．
キーワード 統合的フロントエンド処理，音声区間検出，雑音抑圧，逐次処理，音声認識

Noisy speech recognition using integrated method of statistical model-based voice activity detection and noise suppression

Masakiyo FUJIMOTO†, Kentaro ISHIZUKA†, and Tomohiro NAKATANI†

† NTT Communicaition Science Laboratories, NTT Corp.
2–4, Hikaridai, Seika–cho, Souraku–gun, Kyoto, 619–0237, Japan
E-mail: {masakiyo, ishizuka, nak}@cslab.kecl.ntt.co.jp

Abstract This paper addresses robust front-end processing for automatic speech recognition in noise. The proposed method integrates voice activity detection (VAD) and noise suppression, and consists of three core techniques, i.e., (1) statistical model sharing, (2) Wiener filter design by using speech / non-speech probabilities, and (3) VAD improvement by using enhance speech. In addition, the proposed method can perform sequential processing without frame delay. In an evaluation, the proposed method significantly improves accuracy of concatenated speech recognition without frame delay. In addition, we investigate to combine cepstrum mean normalization and sequential acoustic model adaptation with the proposed method.
Key words integrated front-end processing, voice activity detection, noise suppression, sequential processing, speech recognition

1. まえがき

実環境で音声認識を頑健に行うためには，音声区間検出（VAD: Voice Activity Detection），雑音抑圧，残響除去など，様々な処理が必要となる．

まず，雑音に頑健な VAD として我々はこれまでに，音声信号中の周期性成分と非周期性成分の比（PAR: Periodic to Aperiodic component Ratio）[1] と，確率モデルと Switching カルマンフィルタ（SKF: Switching Kalman Filter）に基づく方法 [2] を統合した，MUSCLE-VAD（MUlti Stream Combination of Likelihood Evolution for VAD）を提案した [3]．提案した VAD により，様々な雑音環境下において高い VAD 性能が得られ，加えて連続発声音声の認識性能を改善することを示した．しかし VAD のみでは十分な雑音下音声認識の性能改善を得ることは難しく，雑音抑圧技術を適用することが必要不可欠である．

ここで，VAD は雑音抑圧に用いるフィルタの設計においても重要な役割を果たす．逆に，雑音抑圧を行って SNR を改善することは VAD の性能改善につながる．このことから，これらの技術は別個のものと捉えるべきではなく，一つの大きな音声認識向けフロントエンド処理として捉える必要がある．すなわち，Fig. 1 の (a) のように，従来通り VAD と雑音抑圧を別

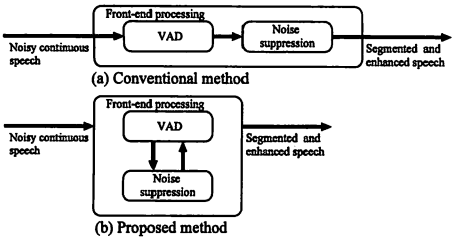

図 1　音声認識のフロントエンド処理

個の技術として単純に連結するのではなく，(b) のように，二つの手法を統合し，相互に情報のやり取りが可能となる手法が必要となる．

以上を踏まえて本研究では，VAD と雑音抑圧を統合的に扱う頑健な音声認識フロントエンドの検討を行う．提案手法において雑音抑圧には，確率モデルに基づく方法 [4] を採用する．MUSCLE-VAD もまた確率モデルに基づく手法であり，VAD と雑音抑圧において確率モデルを共有し [5]，二つの手法間での情報のやり取りを密にすることが，提案手法の狙いである．また，雑音抑圧の結果を VAD にフィードバックすることにより VAD の性能がさらに改善し，音声認識性能の改善に貢献することを示す．

提案手法は VAD と雑音抑圧の双方を逐次的に処理することが可能であり，原理的にはフレーム遅延無しでの処理を実現している．フレーム遅延無しの処理はオンライン音声認識の応答性を高めるために重要な要素である．また，音声認識の性能改善を得るための重要技術として，VAD と雑音抑圧以外には，CMN (Cepstrum Mean Normalization) [6] と音響モデル適応 [7] がある．本研究では，フレーム遅延無しという条件の下，これら 2 つの手法を併用し，音声認識性能の更なる改善が得られることを示す．

2.　音声区間検出

まず，雑音に頑健な VAD である MUSCLE-VAD について述べる．MUSCLE-VAD は，音声信号中の周期性，非周期性成分の比と，確率モデルと Switching カルマンフィルタに基づく方法を統合した手法であり，様々な雑音環境において頑健に動作する [3]．

2.1　信号の周期性，非周期性成分比 PAR

音響信号は周期性成分と非周期性成分に分離可能であり，この両成分を分離して併用する表現形式は，音声合成や音楽信号の分析合成において従来その効果が確認されている．ここで，音声信号，特に有声音は周期性成分を多く含む信号であるため，音声信号中の周期性成分と非周期性成分のパワー比 PAR_t は，音声／非音声を識別する有効な特徴量となり得る．尚，パワー比 PAR_t の推定方法の詳細は，文献 [1] を参照されたい．

2.2　Switching カルマンフィルタ

2.2.1　状態遷移モデルの定義

提案手法では，観測信号が音声状態と非音声状態を遷移する信号であると仮定し，観測信号が各状態に属する確率の比に基づき，音声／非音声の識別を行う．

まず提案手法は，事前にクリーン音声データを用いてクリーン音声と無音の GMM (Gaussian Mixture Model) を学習し，それぞれの GMM を用いて，クリーン音声と無音の状態遷移モデルを構成しておく（Fig. 2 の Clean speech model）．また，雑音を Fig. 2 の Noise model のように常に状態遷移を伴う信号であると定義し，観測信号が与えられると，Switching カルマンフィルタにより雑音状態を逐次更新する．その後，Clean speech model との合成により，Noisy speech model（環境適応モデル）を得る．このような状態遷移モデルを用いることにより，音声信号の多様性，雑音の時間変化に対して頑健な VAD を実現できる．

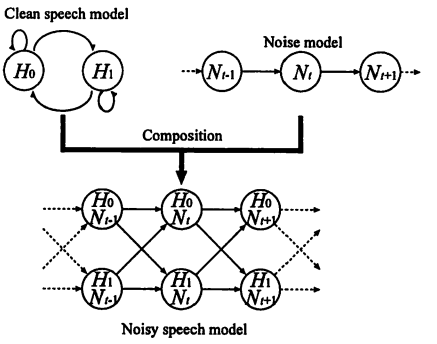

図 2　音声／非音声状態遷移モデル

2.2.2　状態遷移モデルの定式化と尤度比の算出

Fig. 2 に基づく，雑音の非定常性を考慮した音声／非音声状態の識別方法について延べる．

時刻（フレーム）t での観測信号 $\mathbf{O}_t$（L 次元の対数メルスペクトルベクトル）の状態を q_t と定義し，雑音の L 次元対数メルスペクトルベクトルを $\mathbf{N}_t$ とすると，$\mathbf{O}_{0:t} = \{\mathbf{O}_0, \cdots, \mathbf{O}_t\}$，$\mathbf{N}_{0:t} = \{\mathbf{N}_0, \cdots, \mathbf{N}_t\}$ が与えられたときの状態 q_t の確率 $p(q_t|\mathbf{O}_{0:t}, \mathbf{N}_{0:t})$ は次式で与えられる．

$$p(q_t|\mathbf{O}_{0:t}, \mathbf{N}_{0:t}) \propto p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) \tag{1}$$

q_t と $\mathbf{N}_t$ の状態遷移がそれぞれ独立と仮定すると，確率 $p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t})$ の次の再帰式で表現される．

$$\begin{aligned} &p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) \\ &= \sum_{q_{t-1}} p(q_t|q_{t-1})\, p(\mathbf{N}_t|\mathbf{N}_{t-1})\, p(\mathbf{O}_t|q_t, \mathbf{N}_t) \\ &\quad \times p(\mathbf{O}_{0:t-1}, q_{t-1}, \mathbf{N}_{0:t-1}) \end{aligned} \tag{2}$$

$p(q_t|q_{t-1})$，$p(\mathbf{N}_t|\mathbf{N}_{t-1})$，$p(\mathbf{O}_t|q_t, \mathbf{N}_t)$ は，それぞれ音声／無音の状態遷移確率，雑音の状態遷移確率，各状態における出力確率であり，$p(q_t = H_j|q_{t-1} = H_i) = a_{i,j}$，$p(\mathbf{N}_t|\mathbf{N}_{t-1}) = c_{t,t-1}$，$p(\mathbf{O}_t|q_t = H_j, \mathbf{N}_t) = b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ と定義する．また，$p(\mathbf{O}_{0:t}, q_t = H_j, \mathbf{N}_{0:t})$ は前向き確率 $\alpha_{j,t}$ に相当し，本研究では雑音が常に状態遷移をするという前提をおいているので，$c_{t,t-1} = 1$ となるため，式 (2) は次式で表現される．なお，時刻 $t = 0$ の場合は，初期値 $\alpha_{0,0} = 1$，$\alpha_{1,0} = 0$ を与える．

$$\alpha_{j,t} = \sum_{i=0}^{1} (a_{i,j}\alpha_{i,t-1})\, b_{j,\mathbf{N}_t}(\mathbf{O}_t) \tag{3}$$

それぞれの状態における $\alpha_{j,t}$ の比 $R_t = \alpha_{1,t}/\alpha_{0,t}$ を次式で閾値処理して，時刻 t の状態を識別する [8].

$$q_t = \begin{cases} H_0 & R_t < Threshold \\ H_1 & R_t \geqq Threshold \end{cases} \quad (4)$$

なお，Switching カルマンフィルタに基づく雑音状態の更新方法，提案手法の詳細については文献 [2] を参照されたい.

2.3 複数手法の統合

2.1 と 2.2 にて述べた，PAR と SKF の統合について述べる．本研究では尤度計算レベルでの統合を行っており，文献 [1] の方法により得た PAR の尤度 $b_{j,PAR}(PAR_t)$ と，2.2.2 の方法により得た SKF の尤度 $b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ の直積を次式のように求めることにより，結合尤度 $b_j(\mathbf{O}_t, PAR_t)$ を得る.

$$b_j(\mathbf{O}_t, PAR_t) = b_{j,\mathbf{N}_t}(\mathbf{O}_t) \cdot b_{j,PAR}(PAR_t) \quad (5)$$

上式により得られた尤度 $b_j(\mathbf{O}_t, PAR_t)$ を用いて式 (3) の前向き確率を計算し，識別を行う.

3. 確率モデルに基づく雑音抑圧

次に，確率モデルに基づく雑音抑圧法について述べる．本手法は Wiener filter に基づく手法であり，メル周波数軸上での Wiener filter のフィルタゲインを確率モデルのパラメータを用いて最適推定する [4].

一般に Wiener filter のフィルタゲイン $G_{t,l}$ は，次式のように得られる.

$$G_{t,l} = \frac{\exp(S_{t,l})}{\exp(O_{t,l})} \simeq \frac{\exp(S_{t,l})}{\exp(S_{t,l}) + \exp(N_{t,l})} \quad (6)$$

上式において，$S_{t,l}$ は，音声の対数メルスペクトル，$O_{t,l}$，$N_{t,l}$ は，それぞれ $\mathbf{O}_t$，$\mathbf{N}_t$ の第 l 要素である.

ここで $S_{t,l}$ は，$O_{t,l}$ が観測された時点では未知であるので，事前に学習した確率モデルの平均値を用い，$N_{t,l}$ は，2.2.2 及び，文献 [2] に述べられた方法により得られた値を用いる．また，$S_{t,l}$ の確率モデルは GMM で構成されるため，K 個の正規分布を持つ．よって，各正規分布 k の平均ベクトルの対応する要素値 $\mu_{S_{k,l}}$ を用いて，複数のフィルタゲイン $\hat{G}_{t,k,l}$ を以下のように構成できる.

$$\hat{G}_{t,k,l} = \frac{\exp(\mu_{S_{k,l}})}{\exp(\mu_{O_{t,k,l}})} \quad (7)$$

上式において，$\mu_{O_{t,k,l}}$ は $O_{t,l}$ の平均ベクトル $\boldsymbol{\mu}_{O_{t,k}}$ の第 l 要素であり，$\mu_{S_{k,l}}$ と $N_{t,l}$ を用いて，HMM 合成法 [9] 等により生成する.

その後，$\hat{G}_{t,k,l}$ を事後確率 $p(k|\mathbf{O}_t)$ で重み付き平均をとることにより，最適なフィルタゲイン $\hat{G}_{t,l}$ を得る（式 (9) の $\boldsymbol{\Sigma}_{O_{t,k}}$ は，$O_{t,l}$ の共分散行列）.

$$\hat{G}_{t,l} = \sum_{k=1}^{K} p(k|\mathbf{O}_t)\, \hat{G}_{t,k,l} \quad (8)$$

$$p(k|\mathbf{O}_t) = \frac{\mathcal{N}\left(\mathbf{O}_t; \boldsymbol{\mu}_{O_{t,k}}, \boldsymbol{\Sigma}_{O_{t,k}}\right)}{\sum_{k'=1}^{K} \mathcal{N}\left(\mathbf{O}_t; \boldsymbol{\mu}_{O_{t,k'}}, \boldsymbol{\Sigma}_{O_{t,k'}}\right)} \quad (9)$$

ここで文献 [4] の方法では，無音／音声区間の区別無しに学習された GMM を用いてフィルタゲインの推定を行っている．しかし，無音と音声では音響特徴が大きく異なり，それぞれを区別して GMM を学習し，無音／音声区間で GMM を使い分けることにより雑音抑圧性能が改善することが報告されている [10].

本研究では，2.2.2 にて述べたように，無音 ($j = 0$) と音声 ($j = 1$) の GMM を用いて逐次的に雑音環境に適応した非音声，音声の GMM を推定しており，得られた GMM のパラメータを用いてフィルタゲインを以下のように構成することが可能である.

$$\hat{G}_{t,j,k,l} = \frac{\exp(\mu_{S_{j,k,l}})}{\exp(\mu_{O_{t,j,k,l}})} \quad (10)$$

上式において，$\mu_{O_{t,j,k,l}}$ は Switching カルマンフィルタにより更新された，非音声，音声 GMM の平均ベクトルの要素値である.

その後，文献 [4] と同様に事後確率を用いて重み付け平均を行うが，本研究では非音声，音声両方の GMM を用いているため，以下のように式 (3) で得られる前向き確率 $\alpha_{j,t}$ を用いて，さらに重み付け平均を行うことによりフィルタゲインを推定する．これにより，無音と音声の音響特徴の差異を考慮した，最適なフィルタゲインの推定が可能となる．またこのことは，VAD との情報の共有にもつながる.

$$\hat{G}_{t,l} = \sum_{j=0}^{1} \alpha_{j,t} \sum_{k=1}^{K} p(k|\mathbf{O}_t, j)\, \hat{G}_{t,j,k,l} \quad (11)$$

$$p(k|\mathbf{O}_t, j) = \frac{\mathcal{N}\left(\mathbf{O}_t; \boldsymbol{\mu}_{O_{t,j,k}}, \boldsymbol{\Sigma}_{O_{t,j,k}}\right)}{\sum_{k=1}^{K} \mathcal{N}\left(\mathbf{O}_t; \boldsymbol{\mu}_{O_{t,j,k'}}, \boldsymbol{\Sigma}_{O_{t,j,k'}}\right)} \quad (12)$$

フィルタゲインの推定後，Mel-warped DCT [11] によりインパルス応答に変換し，観測信号波形に畳み込むことにより，雑音抑圧された音声信号を得る.

4. 雑音抑圧結果のフィードバック

雑音抑圧結果の VAD へのフィードバックについて検討を行う．観測信号は，3. で述べた雑音抑圧により SNR の改善が得られるため，雑音抑圧結果を用いて再度 VAD を行うことにより，VAD の性能を改善させる可能性がある．しかし，雑音抑圧を行うことにより SNR が改善されたとしても抑圧後の音声に歪みが生じてしまい，結果として確率モデルとのミスマッチが大きくなり，尤度を低下させる恐れがある．これは低 SNR 環境において顕著に発生すると考えられる.

よって本研究では一度の処理で雑音を全て抑圧するのではなく，段階的に雑音を抑圧することにより，音声歪みを低減させる手法について検討を行う.

まず，式 (10) のフィルタゲインを次式で再定義する.

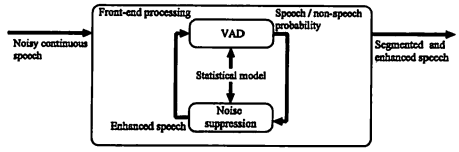

図 3 確立モデルの共有と雑音抑圧音声のフィードバック

$$\hat{G}_{t,j,k,l} = \frac{\exp\left(\mu_{S_{j,k,l}}\right) + \epsilon \cdot \exp\left(N_{t,l}\right)}{\exp\left(\mu_{O_{t,j,k,l}}\right)} \tag{13}$$

上式において，ϵ ($0.0 \leqq \epsilon \leqq 1.0$) は雑音の抑圧量を制御する変数であり，$\epsilon = 0.0$ は通常の雑音抑圧，$\epsilon = 1.0$ は雑音抑圧を行わないことを示す．

この方法は，SNR が低く通常の雑音抑圧では大きな音声歪みが生じてしまう可能性が高い場合に，$\epsilon = 0.5$〜0.9 程度の値を用いて，完全ではないものの SNR の改善と音声歪み軽減を両立させる．逆に SNR が高く音声歪みの発生が深刻で無い場合は，$\epsilon = 0.0$〜0.5 程度の値を用いて一度に大きな SNR 改善を行うことが可能である．このような雑音抑圧処理により SNR が改善された観測信号を用いて，再度 Switching カルマンフィルタによる（残留）雑音の推定，VAD 処理を行う．すなわち，Fig. 3 に示すように，VAD，VAD 結果を用いた雑音抑圧，雑音抑圧音声のフィードバックを繰り返すことにより，段階的に VAD と雑音抑圧両方の性能を向上させることが狙いである．

なお本研究では，このような手法の第一次検討として，繰り返し回数を 1 回とした場合について実装，評価を行う．つまり VAD と雑音抑圧を 2 回ずつ行うことになり，2 回目の雑音抑圧では $\epsilon = 0.0$ を用いて残留雑音を発生させないような抑圧を行う．処理の繰り返し回数および，ϵ の値の最適化については今後の課題とし，まずは提案手法の有効性について調査を行う．

5. CENSREC-1-C による評価

5.1 CENSREC-1-C と実験条件

提案手法の評価は，VAD の評価用に設計されたデータベース CENSREC-1-C [12] を用いて行う．CENSREC-1-C は，人工的に作成したシミュレーションデータと，実環境で収録した実データの 2 種類のデータを含んでおり，本研究では，実環境における音声品質劣化の影響（雑音及び，発声変形の影響等）を調査するため，実データを用いて評価を行う．

CENSREC-1-C の実データの収録は，学生食堂 (Rest.) と高速道路付近 (St.) の 2 環境で行われており，SNR はそれぞれ，High SNR (Hi.: 騒音レベル 60 dB(A) 前後) と Low SNR (Lo.: 騒音レベル 70 dB(A) 前後) である．音声データは，1 名の話者が 1〜12 桁の連続数字を 9 もしくは 10 回，約 2 秒間隔で発話した音声を 1 ファイルとして収録しており，各環境において話者 1 名あたり 4 ファイルを収録している．発話者は 10 名（男女各 5 名）である．収録機材等の詳細については文献 [12] を参照されたい．

音響分析は，フレーム長 20 ms，シフト長 10 ms で行い，特徴量は，対数メルスペクトル 12 次元，及び PAR1 次元である．音声の状態遷移確率は，$\{a_{0,0}, a_{0,1}, a_{1,0}, a_{1,1}\} = \{0.90, 0.10, 0.45, 0.55\}$ とした．無音及び，音声 GMM の学習は，AURORA-2J [13] のクリーン学習データ 8,440 発話を用いて行い，GMM の混合分布数はそれぞれ 32 である．

評価は発話単位の検出性能で行い，評価尺度は区間検出正解率 *Corr* と区間検出正解精度 *Acc* である．

$$Corr = N_c/N \times 100\ [\%] \tag{14}$$

$$Acc = (N_c - N_f)/N \times 100\ [\%] \tag{15}$$

上式の N は総発話区間数，N_c は正解発話区間検出数，N_f は誤発話区間検出数である．*Corr* は，発話区間をどれだけ多く検出できるかを評価する尺度であり，*Acc* は，発話区間をどれだけ過不足なく検出できるかを評価する尺度である．

式 (4) の閾値は，全評価環境における *Corr* と *Acc* の平均値が最良となるように調整した．

5.2 VAD の評価結果

Table 1 は，VAD の評価結果を示しており，“Baseline” は CENSREC-1-C のベースライン結果（パワー比＋適応閾値），“Sohn” は Sohn らの確率モデルに基づく VAD [8]，“AFE” は ETSI Advanced-frontend [11]，“Integ. w/o FB” はフィードバック無し（MUSCLE-VAD 単体での結果と等価），“FB ($\epsilon = 0.0$)”，“FB ($\epsilon = 0.7$)” は提案手法による結果である．

表 1 より，“Integ. w/o FB” 以降の提案手法は，従来手法である “Sohn”，“AFE” に比べて高い VAD 性能を示すことがわかる．特に “AFE” は，*Corr* に比べて *Acc* が極端に低いことから湧き出し誤りが多く，音声区間と非音声区間の識別性能が不十分であることがわかる．

“Integ. w/o FB” 以降の結果において，“FB ($\epsilon = 0.0$)” の結果は “Integ. w/o FB” より劣化しており，通常の雑音抑圧を行って VAD にフィードバックすると音声歪みにより尤度が劣化して音声／非音声の誤りを生じさせやすくなることが分かる．一方，“FB ($\epsilon = 0.7$)” の結果は “w/o FB” より改善している．これは雑音抑圧量の適切な制御により，段階的に SNR 改善を行いつつ，音声歪みによる尤度の劣化を防ぐことができたため，雑音抑圧音声の VAD へのフィードバックが効果的に作用したと言える．

また，提案手法は VAD と雑音抑圧を発話単位でのバッチ処理では無く，フレーム単位での逐次処理により行っているため，現在のフレームの処理を行うのに，未来のフレームのデータを必要としない．すなわち，原理的にフレーム遅延が発生しない処理機構となっている．よって，音声認識のみならず，フレーム遅延の影響が如実に現れるオンラインでの音声通話等においても有効な手法であると言える．

5.3 音声認識の評価結果

次に音声認識による評価を行う．HMM の学習データ，実験条件は AURORA-2J [13] と同様であり，HMM はクリーンモデル（CMN 無し）である．

Table 2 は認識結果であり，“w/o VAD” は VAD 無しで連続発生音声を認識した結果，“Ideal” は正解の音声／非音声ラベルを用いて音声区間には音声 GMM のみを，非音声区間には非音声 GMM のみを用いて雑音抑圧を行った場合の結果である

表 1　音声区間検出結果 (%)

	Corr (%)					*Acc* (%)				
	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Avg.	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Avg.
Baseline	74.20	56.52	39.42	41.45	52.90	21.45	-43.48	-15.65	-33.91	-17.90
Sohn	72.75	57.10	97.39	78.55	76.45	45.51	-6.38	94.49	57.39	47.75
AFE	44.35	80.58	47.25	72.17	61.09	-82.03	-245.51	-101.16	-168.70	-149.35
Integ. w/o FB	**93.04**	70.72	**100.00**	97.97	90.43	72.75	19.71	99.13	94.78	71.60
FB ($\epsilon = 0.0$)	91.88	67.54	**100.00**	93.91	88.33	70.14	17.97	98.54	85.80	68.11
FB ($\epsilon = 0.7$)	92.75	**71.88**	**100.00**	**99.71**	**91.09**	**75.36**	**24.06**	**99.42**	**97.97**	**74.20**

表 2　音声区間検出後の音声認識結果（単語正解精度 (%)）

	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Avg.
w/o VAD	45.17	1.28	34.43	25.23	26.53
Baseline	44.16	18.12	29.96	21.62	28.47
Sohn	37.45	-3.81	33.41	29.58	24.16
AFE	42.71	-37.98	22.31	27.14	13.54
Integ. w/o FB	71.31	21.68	76.96	56.83	56.70
FB ($\epsilon = 0.0$)	70.13	17.67	74.86	53.46	54.03
FB ($\epsilon = 0.7$)	**73.86**	**28.87**	**83.97**	**62.39**	**62.27**
Ideal	*88.89*	*49.09*	*86.25*	*60.66*	*71.22*

(理想状態)．

Table 2 の結果より，VAD の場合と同様に，“FB ($\epsilon = 0.0$)” は性能が劣化し，“FB ($\epsilon = 0.7$)” は改善することが分かる．この結果より，提案するフロントエンド処理が雑音抑圧により生じる音声歪みから大きな影響を受け，処理の性能を改善させるためには音声歪みを軽減する，もしくは歪みを吸収できるような枠組みが必要不可欠であると言える．

また，VAD の結果に誤りの無い “Ideal” と提案手法を比較すると，音声認識性能に約 9%の差がある．提案手法と “Ideal” の大きな違いは VAD の性能であり，雑音抑圧から VAD への情報のフィードバックの必要性，提案手法の方針の妥当性を示唆している．

以上より，VAD と雑音抑圧の繰り返し回数，ϵ の値の最適化などについて今後検討を行う必要がある．

5.4　CMN の適用

5.3 では CMN を行わずに音声認識実験を行っていたが，CMN はパラメータ空間の正規化に伴い，入力マイク特性の正規化等を行える極めて有用な技術である．また，処理量が少ないにもかかわらず効果が高いため，音声認識においては必須ともいえる技術である．しかし，CMN は入力データのケプストラムの長時間平均を求める必要があるため，CMN を行ったデータが音声認識器に入力されるまでに遅延が生じる．この遅延はオフライン処理での音声認識では問題はないが，オンライン処理では音声認識の応答性を低下させる．

この遅延の問題に対して，本研究では VAD が検出した直前の発話からケプストラム平均 (CM: Cepstrum Mean) を求め，現在の発話の CM 系列から減算することにより，原理的にフレーム遅延の生じない CMN を実現する．評価においては，図 4 に示す通り，

(1)　観測データ全体からの CM (CMN whole)
(2)　現在の発話からの CM (CMN current，従来の CMN)
(3)　直前の発話からの CM (CMN previous)

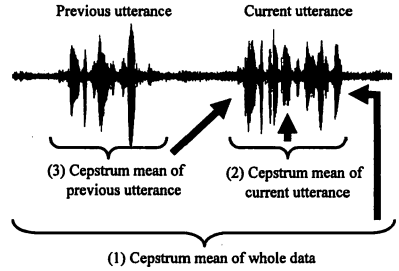

図 4　Cepstrum mean の算出方法

表 3　各 CMN を用いた音声認識結果（単語正解精度 (%)）

	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Avg.
w/o CMN	73.86	28.87	83.97	62.39	62.27
CMN whole	74.68	26.68	**87.34**	**68.67**	64.34
CMN current	**76.50**	**29.05**	86.07	68.31	**64.98**
CMN previous	76.23	28.78	86.07	68.31	64.85

の 3 種類の CM 算出方法での CMN を比較しており，表 2 の “FB ($\epsilon = 0.7$)” の結果に対して各 CMN を適用する．ここで，(1) の処理では，評価データが CENSREC-1-C の連続発話データであるため，9 もしくは 10 発話分のデータを用いて CM を求めることとなる．また (3) の処理において，1 番目の発話の CM は，観測データの始点から VAD が検出した発話の始点までの非音声区間から算出する．

表 3 は，各手法による音声認識結果であり，“w/o CMN” は CMN を行わない場合の結果である．表より，各 CMN を行うことにより，行わない場合に比べて音声認識性能が改善している．また，従来の手法である “CMN current” が最良の結果を示したが，“CMN previous” との差はわずかであり，直前の発話の CM を用いても十分な性能が得られることがわかる．原理的にフレーム遅延が生じないという観点からも，“CMN previous” は有用な手法であると言える．

5.5　音響モデルの逐次適応

次に，音声認識に用いる音響モデルの適応について検討を行う．適応手法には教師無し MLLR (Maximum Likelihood Linear Regression) [7] を用い，図 5 に示す通り，

(1)　観測データ全体を用いて適応 (MLLR whole)
(2)　現在の発話を用いて逐次適応 (MLLR current)
(3)　直前の発話を用いて逐次適応 (MLLR previous)

の 3 種類の手法を比較する．これらは，5.4 の CMN の場合と同様の方法であるため，(3) の処理は原理的にフレーム遅延の生じない手法となっている．また各適応手法は，表 3 の “CMN previous” の結果に対して適用し，(3) の処理において 1 番目の

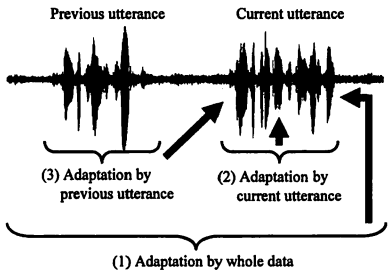

図 5 適応データの取得方法

表 4 各 MLLR 適応を用いた音声認識結果（単語正解精度 (%)）

	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Avg.
w/o MLLR	76.23	28.78	86.07	68.31	64.85
MLLR whole	75.96	26.14	**87.16**	**70.67**	64.98
MLLR current	79.96	**37.98**	86.25	68.76	68.24
MLLR previous	**80.51**	37.89	86.16	68.67	**68.31**

発話に対しては適応処理を行わない．

表 4 は，各手法による音声認識結果であり，“w/o MLLR” は適応を行わない場合の結果である．表より，“MLLR previous” は，“MLLR current” とほぼ同等の結果となり，遅延を生じさせずに適応処理による改善を得ることができた．一方，“MLLR whole” では適応処理の効果が現れなかった．これは適応データのラベル精度の問題により生じた問題である．ラベルは，適応データを一度認識した結果より得られ，“MLLR current” と “MLLR previous” の逐次適応は，1 発話毎に適応処理が行われるため，発話が進むにつれて認識性能とラベル精度が改善する．これにより適応の効果も逐次的に向上する．しかし，“MLLR whole” は観測データ全体を用いて適応を行うことから，初期の適応無し音響モデルを用いて観測データ全体を認識してラベルを作成する必要があり，逐次適応の場合と異なりラベル精度の改善が得られない．逐次適応に比べて適応データが多くとも，ラベル精度が低いために十分な適応性能が得られず，逐次適応手法に比べて性能が低くなったと言える．

音響モデル逐次適応手法としては，巨視的な時間発展系を利用した手法 [14] が提案されている．この手法では，適応の処理機構に Kalman filter を適用しており，我々の提案する Switching Kalman filter に基づく VAD との親和性は高いと言える．今後，これらの手法の統合について検討を行う必要がある．

6. むすび

本研究では，VAD と雑音抑圧の統合方法について述べ，雑音抑圧結果の VAD へのフィードバックの際に，雑音抑圧量の制御による音声歪みの抑制が重要であることを示した．今後，VAD と雑音抑圧の繰り返し回数，雑音抑圧量の制御係数 ϵ の最適化などについて検討を行う予定である．

また，VAD 結果を活用して直前の発話を用いた，CMN と音響モデル適応を行い，原理的にフレーム遅延を生じさせずに従来の手法をほぼ同等の音声認識性能を得た．今後，Kalman filter に基づく逐次音響モデル適応手法との統合についても検討する予定である．

謝辞 本研究では，IPSJ SIG-SLP 雑音下音声認識評価 WG により作成された雑音下音声区間検出評価環境 CENSREC-1-C と雑音下音声認識評価環境 AURORA-2J を使用した．

文 献

[1] Ishizuka, K. and Nakatani, T., “Study of noise robust voice activity detection based on periodic component to aperiodic component ratio,” Proc. SAPA ’06, Pittsburgh, PA, USA, pp.65–70, Sept. 2006.

[2] Fujimoto, M. and Ishizuka. K., “Noise Robust Voice Activity Detection Based on Switching Kalman Filter,” IEICE Trans. on Info. & and Syst., Vol. E91–D, No. 3, pp. 467–477, March. 2008.

[3] Fujimoto, M., Ishizuka, K., and Nakatani, T., “A Voice Activity Detection Based on the Adaptive Integration of Multiple Speech Features and a Signal Decision Scheme,” Proc. ICASSP ’08, Las Vegas, NV, USA, pp. 4441–4444, Apr. 2008.

[4] Segura, J. C., Torre, A. de la, Benitez, M. C., and Peinado, A. M., “Model-based compensation of the additive noise for continuous speech recognition. experiments using AURORA II database and tasks,” Proc. of EuroSpeech ’01, Aalborg, Denmark, Vol. I, pp. 221–224, Sept. 2001.

[5] Fujimoto, M., Ishizuka, K., and Nakatani, T., “Study of Integration of Statistical Model-Based Voice Activity Detection and Noise Suppression,” Proc. Interspeech ’08, Brisbane, Australia, pp. 2008–2011, Sept. 2008.

[6] Liu, F. H. and Stern, R. M., “Efficient Cepstral Normalization for Robust Speech Recognition,” Proc. ARPA Workshop on Human Language Technology, pp. 69–74, March 1993.

[7] Leggetter, C. L., and Woodland, P. C., “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models,” Computer Speech and Language, Vol. 9, No. 2, pp. 171–185, Apr. 1995.

[8] Sohn, J., Kim, N. S., and Sung, W., “A statistical model-based voice activity detection,” IEEE SP Letters, Vol. 6, No. 1, pp. 1–3, Jan. 1999.

[9] Gales, M. J. F., “Model-Based Techniques for Noise Robust Speech Recognition,” Ph.D Thesis, Cambridge University, Sep. 1995.
http://mi.eng.cam.ac.uk/~mjfg/thesis.ps.gz

[10] 藤本 雅清, 有木 康雄, “GMM と EM アルゴリズムを用いた加法性雑音及び乗法性歪みの抑圧,” 信学論, Vol. J88–D–II, No. 7, pp. 1093–1102, July 2005.

[11] ETSI standard document, “Speech processing, Transmission and Quality aspects (STQ), Advanced Distributed Speech Recognition; Front-end feature extraction algorithm; Compression algorithms,” ETSI ES 202 050 v.1.1.4, Nov. 2005.

[12] Kitaoka, N., Yamada, T., Tsuge, S., Miyajima, C., Nishiura, T., Nakayama, M., Denda, Y., Fujimoto, M., Yamamoto, K., Takiguchi, T., Tamura, S., Kuroiwa, S., Takeda, K., and Nakamura, S., “Development of VAD Evaluation Framework CENSREC-1-C and Investigation of Relationship Between VAD and Speech Recognition Performance,” Proc. ASRU ’07, Kyoto, Japan, pp. 607–612, Dec. 2007.

[13] Nakamura, S., Takeda, K., Yamamoto, K., Yamada, T., Kuroiwa, S., Kitaoka, N., Nishiura, T., Sasou, A., Mizumachi, M., Miyajima, C., Fujimoto, M., and Endo, T., “AURORA-2J, An evaluation framework for Japanese noisy speech recognition,” IEICE Trans. on Inf. & Syst., Vol. E88–D, No. 3, pp. 535–544, March 2005.

[14] 渡部 晋治, 中村 篤, “巨視的な時間発展系に基づく逐次追従型音声認識,” 音講論, 2–P–9, pp. 129–130, Sept. 2008.