2007－ＳＬＰ－69（２）
2007／12／20

複数の音声区間検出法の適応的統合の検討と考察

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，複数の音声区間検出法の適応的統合の検討を行う．本研究にて採用する音声区間検出法は，音声の周期性・非周期性成分比と Switching カルマンフィルタに基づく手法であり，各手法の統合はそれぞれが出力する尤度をフレーム単位で重み付け加算することにより行う．提案手法の評価は CENSREC-1-C を用いて行い，雑音環境下において高い音声区間検出性能が得られることを示す．また，提案手法における尤度の重み付け加算方法などについて，実験を伴った考察を行う．

キーワード 音声区間検出，周期性／非周期性成分比，Switching Kalman filter，適応的統合

A study and an examination on adaptive integration of multiple voice activity detection

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 The VAD method proposed in this paper integrates multiple speech features and a signal decision scheme, namely the speech periodic to aperiodic component ratio and a switching Kalman filter. The integration is carried out by using the weighted sum of likelihoods outputted from each VAD (stream). The stream weight is decided adaptively each short time frame. The evaluation is carried out by using a VAD evaluation framework, CENSREC-1-C. The evaluation results revealed that the proposed method significantly outperforms the baseline results of CENSREC-1-C as regards VAD accuracy in real environments. In addition, we examine the method of likelihoods weighting through the experiments.

Key words voice activity detection, periodic to aperiodic component ratio, switching Kalman filter, adaptive integration

1. まえがき

連続した観測信号から音声信号が存在する区間を検出する音声区間検出技術（VAD: Voice Activity Detection）は，音声認識のみならず，音声強調，音声符号化等，あらゆる音声情報処理の入り口に位置する，極めて重要な技術である．

一般に VAD は，音声／非音声を識別するための特徴の抽出部と，得られた特徴を基に音声／非音声の判定を行う識別部に大別される．VAD の特徴量として，かねてより音声／非音声のエネルギー比，ゼロ交差数等 [1] が用いられることが多いが，これらは背景雑音等が存在する場合には有効ではない．このため，雑音に頑健な特徴量が多数提案されている [2] [3]．これらの特徴量を用いることにより，雑音環境下における VAD の性能を改善することができる．

一方，雑音に頑健な音声／非音声識別機構として，確率モデルに基づく方法が Sohn らにより提案されており，雑音下における高い VAD 性能を示している [4]．Sohn らの手法は，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ており，観測信号が各状態に属する確率の比（尤度比）を求め閾値処理を行うことにより，音声／非音声の識別を行う．また，尤度比は過去のフレームの状態を考慮した前向き確率を用いて算出されており，単純なフレーム単位での識別に比べて頑健であることが示されている．

VAD は，用いる特徴や識別機構によってそれぞれ得意とする

雑音が異なり，一つの手法のみで実世界に存在するあらゆる雑音に対処することは困難である．そのため本研究では，複数の手法を組み合わせることにより，幅広い雑音環境に対応可能な VAD の検討を行う．組み合わせ手法として，本研究では，音声信号に含まれる周期性成分と非周期性成分の比（PAR: Periodic to Aperiodic component Ratio）[3] 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 [5] を採用する．手法の統合は尤度基準で行い，各々の手法独立で音声／非音声状態の尤度を計算した後に，フレーム単位で適応的に重み付け加算を行う．このような手法を導入することにより，各々の手法単体の場合に比べて，VAD 性能の改善が得られることを示す．この手法を以後，MUSCLE-VAD（MUlti Stream Combination of Likelihood Evolution for VAD）と呼ぶこととする．

また本研究では，MUSCLE-VAD の尤度重み付け方法，モデルパラメータ数，VAD 性能と処理時間の関係，特徴量の周波数帯域制限による影響について調査を行っており，その結果に対する考察を行う．

2. 各手法の比較

まず，各 VAD 手法の比較を行い，それぞれの手法が得意とする環境，不得意とする環境を確認する．表 1 は，各手法の様々な環境において期待される性能と，処理時間（RTF: Real Time Factor）を網羅している．RTF は，Intel Pentium 4 3.6GHz の CPU にて計測した．

これまでに，国際標準化規格である ITU-T 勧告 G.729B. [6]，欧州の標準化規格である ETSI 勧告 ES 202 050 [7] に加え，Sohn らの確率モデルに基づく方法 [4]，Ramirez らのスペクトル構造に基づく方法 [2] など，様々な手法が提案されている．しかし，表 1 に示したように，これらのほぼ全てが効果を発揮する雑音環境が定常なものに限定されており，非定常雑音や突発性雑音環境下での効果は期待できない．

一方，我々のこれまでの提案手法である，PARADE（Periodic to Aperiodic component RAtio-based DEtection）[3] と SKF [5] は，幅広い雑音環境の適応範囲を持っており，それぞれの手法を効果的に組み合わせることにより，あらゆる雑音環境に対応することが可能である（表 1 の MUSCLE-VAD）．また処理時間においては，組み合わせにより計算量がわずかに増加するのみで，実用的な演算量にて処理が可能である．

提案手法は確率モデルを用いているため，モデルパラメータ数に応じて演算量と性能が変化する．この点については，後ほど実験，考察を行う(注1)．

3. 信号の周期性，非周期性成分比

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 [8]．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は，音声／非音声を識別する有効な特徴量となり得る．

表 1　VAD 手法の比較

	音声検出精度				RTF
	無雑音	雑音環境下			
	環境	定常性	非定常性	突発性	
ITU-T G. 729B	○	×	×	×	0.06
ETSI ES 202 050	△	○	△	×	0.06
Sohn	○	○	×	×	0.07
Ramirez	○	○	△	×	0.05
PAR	○	○	△	○	0.06
SKF	○	○	○	△	0.10
MUSCLE-VAD	○	○	○	○	0.11

最初に，観測信号から周期性成分のパワーを推定する方法について述べる．なお，ここで周期性成分とは，ある基本周波数（F0）とその倍音成分から成る調波複合音成分を指す．まず，観測信号 o_τ が，周期性成分 o_{p_τ} と非周期性成分 o_{a_τ} の和で与えられると仮定し，それぞれの短時間フレーム t におけるパワースペクトル密度（PSD: Power Spectral Density）を $O^{PSD}_{t,m}$，$O^{PSD}_{p_{t,m}}$，$O^{PSD}_{a_{t,m}}$，フレーム内のパワーを ρ_t，ρ_{p_t}，ρ_{a_t} として，以下を仮定する．

$$O^{PSD}_{t,m} = O^{PSD}_{p_{t,m}} + O^{PSD}_{a_{t,m}} \tag{1}$$

ここで m は周波数ビンを示し，$\rho_t = \frac{1}{M}\sum_{m=1}^{M} O^{PSD}_{t,m}$ と上式の仮定により，次式を得る（ただし，$m = 1, \cdots, M$）．

$$\rho_t = \rho_{p_t} + \rho_{a_t} \tag{2}$$

次に，フレーム t における周期性成分の F0 と倍音成分の数をそれぞれ f_{0_t}，H_t とし，f_{0_t} の整数倍の周波数ビンに含まれる非周期性成分の平均パワーは，非周期性成分の全周波数ビンにおける平均パワーと等しいとみなし，次式の仮定を導入する．

$$\frac{1}{M}\sum_{m=1}^{M} O^{PSD}_{a_{t,m}} = \frac{1}{H_t}\sum_{h=1}^{H_t} O^{PSD}_{a_{t,[hf_{0_t}]}} \tag{3}$$

上式において，$[hf_{0_t}]$ は，第 h 倍音の周波数ビンを求める演算子であり，本研究では，F0 の推定は，自己相関法 [9] により求める．

一方，周期性成分のパワーは純音とその倍音成分の合算により求まると仮定している．ここで，純音のエネルギー ρ_{s_t} は，純音のパワースペクトル密度 $O^{PSD}_{s_{t,m}}$ と，時間長 T の分析窓関数 g_τ から求まる，

$$\eta = \left(2\sum_{\tau=1}^{T} g_\tau^2\right) \Big/ \left(\sum_{\tau=1}^{T} g_\tau\right)^2 \tag{4}$$

から，$\rho_{s_t} = \eta O^{PSD}_{s_{t,m}}$ により求めることができる．

従って，式 (1)，(2) より，観測信号 o_τ に含まれる周期性成分と非周期性成分の推定パワー ρ_{p_t}，ρ_{a_t} を以下のようにして求める．式変形の詳細は文献 [3] を参照されたい．

$$\rho_{p_t} = \eta \frac{\sum_{h=1}^{H_t} O^{PSD}_{t,[hf_{0_t}]} - H_t\rho_t}{1 - \eta H_t} \tag{5}$$

(注1)：表 1 に示した結果では，確率モデルに 32 混合分布の Gaussian Mixture Model を用いている．

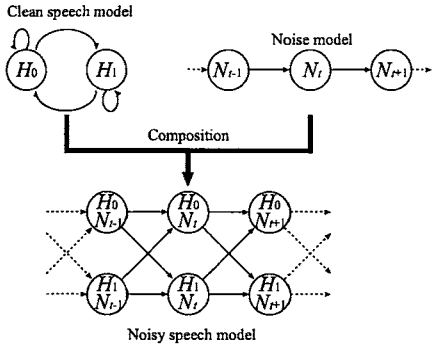

図 1　非定常性雑音を考慮した音声／非音声状態遷移モデル

$$\rho_{a_t} = \rho_t - \rho_{p_t} \tag{6}$$

上式により得られた，ρ_{p_t}，ρ_{a_t} より，周期性成分と非周期性成分のパワー比 PAR_t を次式により得る．

$$PAR_t = \frac{\rho_{p_t}}{\rho_{a_t}} \tag{7}$$

4. 確率モデルに基づく音声区間検出

4.1 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（図 1 の Clean speech model）．また，雑音を図 1 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，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．つまり，雑音環境に適応した音声（クリーン音声＋雑音）と非音声（無音＋雑音）の状態遷移モデルを生成し，さらにこのモデルを逐次更新する．言い換えれば，音声（クリーン音声⇔無音）と雑音（環境変化）両方の状態遷移過程を有する確率モデルを生成することとなる．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 を実現することができる．

4.2 状態遷移モデルの定式化と尤度比の算出

図 1 の状態遷移モデルに基づく，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．

$$p(q_t|\mathbf{O}_{0:t}, \mathbf{N}_{0:t}) \propto p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) \tag{8}$$

確率 $p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t})$ の再帰表現は次式で与えられ，

$$\begin{aligned} &p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) \\ &= \sum_{q_{t-1}} p(q_t, \mathbf{N}_t|q_{t-1}, \mathbf{N}_{t-1})\, p(\mathbf{O}_t|q_t, \mathbf{N}_t) \\ &\times p(\mathbf{O}_{0:t-1}, q_{t-1}, \mathbf{N}_{0:t-1}) \end{aligned} \tag{9}$$

q_t と $\mathbf{N}_t$ の状態遷移がそれぞれ独立事象と仮定すると，

$$\begin{aligned} &p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t}) \\ &= \sum_{q_{t-1}} p(q_t|q_{t-1})\, p(\mathbf{N}_t|\mathbf{N}_{t-1})\, p(\mathbf{O}_t|q_t, \mathbf{N}_t) \\ &\times p(\mathbf{O}_{0:t-1}, q_{t-1}, \mathbf{N}_{0:t-1}) \end{aligned} \tag{10}$$

と表現される．

$p(q_t|q_{t-1})$，$p(\mathbf{N}_t|\mathbf{N}_{t-1})$，$p(\mathbf{O}_t|q_t, \mathbf{N}_t)$ 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 (10) は次式で表現される．

$$\alpha_{j,t} = \sum_{i=0}^{1} (a_{i,j}\alpha_{i,t-1})\, b_{j,\mathbf{N}_t}(\mathbf{O}_t) \tag{11}$$

なお，時刻 $t = 0$ の場合は当該フレームが非音声フレームであるとみなして，初期値 $\alpha_{0,0} = 1$，$\alpha_{1,0} = 0$ を与える．

それぞれの状態における $\alpha_{j,t}$ の比 $R_t = \alpha_{1,t}/\alpha_{0,t}$ を次式で閾値処理することにより，時刻 t の状態を識別する．

$$q_t = \begin{cases} H_0 & R_t < Threshold \\ H_1 & R_t \geqq Threshold \end{cases} \tag{12}$$

一方，図 1 の雑音の状態遷移モデルにのみ着目すると，雑音の状態遷移もまた以下の再帰式で表現され，さらに次式はカルマンフィルタの確率的表現とも完全に一致する [10]．

$$\begin{aligned} &p(\mathbf{O}_{0:t}, \mathbf{N}_{0:t}) \\ &= p(\mathbf{N}_t|\mathbf{N}_{t-1})\, p(\mathbf{O}_t|\mathbf{N}_t)\, p(\mathbf{O}_{0:t-1}, \mathbf{N}_{0:t-1}) \end{aligned} \tag{13}$$

ここで，上式に確率変数 q_t を加えると，式 (10) と一致する．すなわち式 (10) が，ある状態変数に基づき状態空間表現を変化させる Switching カルマンフィルタ（SKF）の確率表現と一致することを意味している．なお，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 [5] を参照されたい．

5. 複数手法の統合

3. と 4. 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について検討を行っており，PAR と SKF で尤度計算を独立に行い，それぞれの尤度を重み付け加算する統合方法となっている．すなわち，事前に PAR 単体の音声，無音 GMM を学習しておき，次式にて PAR を用いた観測信号の尤度（出力確率）$b_{j,PAR}(PAR_t)$ を計算する．

$$b_{j,PAR}(PAR_t) = \sum_{k=1}^{K} w_{PAR_{j,k}} \mathcal{N}\left(PAR_t; \mu_{PAR_{t,j,k}}, \sigma^2_{PAR_{t,j,k}}\right) \tag{14}$$

その後，SKF の尤度 $b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ を次式のように，重み γ_t を用いて加算する．

$$b_j(\mathbf{O}_t, PAR_t) = \gamma_t b_{j,\mathbf{N}_t}(\mathbf{O}_t) + (1-\gamma_t) b_{j,PAR}(PAR_t) \tag{15}$$

上式により得られた尤度 $b_j(\mathbf{O}_t, PAR_t)$ を用いて式 (11) の前向き確率を計算し，識別を行う．重みの値は $0 \leqq \gamma_t \leqq 1$ であり，$\gamma_t = 0$ の場合は PAR 単体での識別，$\gamma_t = 1$ の場合は SKF 単体での識別となる．

重み γ_t は，フレーム単位で適応的に決定する．まず，SKF の尤度 $b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ と PAR の尤度 $b_{j,PAR}(PAR_t)$ の双方に対して，非音声状態 ($j=0$) と音声状態 ($j=1$) の尤度の合計が 1 になるよう正規化を行う．その後，非音声状態もしくは音声状態のどちらに属するかという確信度 $C^x_{y,t}$ を求める ($y = PAR$ or SKF)．$C^x_{y,t}$ の求め方としては，以下の絶対値差分 ($x = Diff$) とエントロピー ($x = Ent$) に基づく方法について比較検討を行う．すなわち線形尺度（絶対値差分）と非線形尺度（エントロピー）の比較を行う．

[絶対値差分]

$$C^{Diff}_{SKF,t} = |b_{0,\mathbf{N}_t}(\mathbf{O}_t) - b_{1,\mathbf{N}_t}(\mathbf{O}_t)| \tag{16}$$

$$C^{Diff}_{PAR,t} = |b_{0,PAR}(PAR_t) - b_{1,PAR}(PAR_t)| \tag{17}$$

[エントロピー]

$$C^{Ent}_{SKF,t} = 1 - (ent(b_{0,\mathbf{N}_t}(\mathbf{O}_t)) + ent(b_{1,\mathbf{N}_t}(\mathbf{O}_t))) \tag{18}$$

$$C^{Ent}_{PAR,t} = 1 - (ent(b_{0,PAR}(PAR_t)) + ent(b_{1,PAR}(PAR_t))) \tag{19}$$

$$ent(p) = -p\log_2(p) \tag{20}$$

$C^x_{y,t}$ は値が大きければ，非音声状態もしくは音声状態のどちらかに属する確信度が高く，値が小さければ，確信度が低いことを示す．本研究では，確信度が高い手法を積極的に利用するように，重み γ^z_t を決定する．また，重み γ^z_t の決定においても単純に正規化を行う線形尺度 ($z = Norm$) とエントロピーを用いた非線形尺度 ($z = Ent$) の比較を行う．それぞれの尺度は以下のように得られる．

[正規化]

$$\gamma^{Norm}_t = \frac{C^x_{SKF,t}}{C^x_{SKF,t} + C^x_{PAR,t}} \tag{21}$$

[エントロピー]

$$\gamma^{Ent}_t = \begin{cases} \gamma & C^x_{SKF,t} \geqq C^x_{PAR,t} \\ 1-\gamma & C^x_{SKF,t} < C^x_{PAR,t} \end{cases} \tag{22}$$

$$\gamma = \frac{ent(\gamma^{Norm}_t) + ent(1-\gamma^{Norm}_t)}{2} \tag{23}$$

上式は，それぞれ確信度が高い手法ほど重みが増加することを示しており，フレーム単位に現時刻の環境に適した手法を効果的に選択することができる．

6. CESNREC-1-C による評価

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 [11] を用いて行う．また，5. にて述べた重み決定法の効果及び，モデルパラメータ数と性能の関係，さらに特徴の周波数帯域の制限による影響について調査，比較を行う．

6.1 CESNREC-1-C と実験条件

CESNREC-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．

CESNREC-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 8〜10 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 [11] を参照されたい．

音響分析は，フレーム長 25 ms，シフト長 10 ms で行い，対数メルスペクトルの次元は $L = 24$ とし，音声の状態遷移確率は，$a_{i,j} = \{0.90, 0.10, 0.45, 0.55\}$ とした．無音及び，音声 GMM の学習は，CENSREC-1 (AURORA-2J) [12] のクリーン学習データ 8,440 発話のデータを用いて行い，GMM の混合分布数はそれぞれ 32 である．特徴量は，対数メルスペクトル 24 次元，及び PAR1 次元であり，それぞれ独立に学習する．

評価は発話単位の検出性能にて行う．評価尺度は次式の区間検出正解率 $Corr$ と区間検出正解精度 Acc である．

$$Corr = N_c/N \times 100\ [\%] \tag{24}$$

$$Acc = (N_c - N_f)/N \times 100\ [\%] \tag{25}$$

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 (12) 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．

6.2 重み決定方法の比較

まず 5. にて述べた，重み決定法の比較を行う．本研究では，確信度 $C^x_{y,t}$ と重み γ^z_t の決定方法に，それぞれ線形 ($x = Diff$, $z = Norm$)，非線形 ($x = Ent$, $z = Ent$) の尺度を定義し

表 2 各重み付け方法による評価結果 (%)

確信度尺度 x	重み計算尺度 z	*Corr* (%)					*Acc* (%)				
		Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.
Diff	*Norm*	**93.04**	**70.72**	**100.00**	97.97	90.43	72.75	**19.71**	**99.13**	94.78	**71.59**
Diff	*Ent*	**93.04**	**70.72**	99.42	**98.84**	**90.51**	73.62	19.42	95.94	95.94	71.23
Ent	*Norm*	**93.04**	68.70	99.42	**98.84**	90.00	73.04	13.91	97.10	**96.23**	70.07
Ent	*Ent*	**93.04**	69.57	**100.00**	97.97	90.15	**74.20**	17.39	**99.13**	94.49	71.30

表 3 モデルパラメータ数と性能及び，RTF の関係 (%)

混合分布数	*Corr* (%)					*Acc* (%)					RTF
	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.	
1	83.48	57.39	91.01	61.74	73.41	44.35	-8.41	77.97	20.58	33.62	0.004
2	91.30	60.87	95.94	84.64	83.19	69.86	-0.58	87.25	63.77	55.08	0.007
4	92.17	66.96	98.26	91.88	87.32	69.86	7.54	90.72	80.58	62.18	0.014
8	91.01	70.14	98.84	93.62	88.40	66.38	19.42	97.68	85.80	67.32	0.029
16	93.04	**71.59**	99.42	96.23	90.07	71.59	**21.16**	97.97	91.30	70.51	0.060
32	93.04	70.72	**100.00**	97.97	90.43	**72.75**	19.71	**99.13**	94.78	**71.59**	0.112
64	93.33	70.43	**100.00**	98.84	90.65	71.59	17.10	98.26	96.52	70.87	0.216
128	**94.20**	70.14	**100.00**	**99.42**	**90.94**	72.46	17.97	97.97	**97.97**	**71.59**	0.424
256	93.91	68.70	99.42	99.13	90.29	71.88	16.52	96.23	96.81	70.36	0.887
512	91.30	68.99	99.42	99.13	89.71	68.99	13.91	96.23	97.10	69.06	1.818

ており，全ての組み合わせによる評価を行う．

表 2 は重み決定方法の比較結果を示しており，確信度 $C^x_{y,t}$ と重み γ^z_t の決定方法として，それぞれ線形尺度（$x = Diff$, $z = Norm$）を用いた場合に，平均 *Corr* と平均 *Acc* の双方を考慮した最良の結果が得られた．また，$C^x_{y,t}$, γ^z_t の決定尺度は，双方とも線形もしくは非線形に統一した方が性能が高くなる傾向が見られた．

しかしながら，各手法の性能差は微々たるものであるため，どの手法が最適であるかを決定することは困難である．よって本研究では，性能に大きな差は見られないものの，計算量の観点から確信度 $C^x_{y,t}$ と重み γ^z_t の決定にそれぞれ線形尺度（$x = Diff$, $z = Norm$）を用いる手法を採用する．

6.3 モデルパラメータ数と性能の関係

次に，モデルパラメータ数と性能，RTF との関係について調査を行った．表 1 の技術対比表と，6.1 の実験条件では，無音及び，音声 GMM の混合分布数を 32 と定義したが，これを 1 から 512 まで変化させて評価を行う．すなわち，分布のベクトル次元等を変化させるのではなく，混合分布数を可変とすることによりモデルパラメータ数を変化させる．なお，重みの決定方法には，前節にて述べた線形尺度を用いる．

表 3 は，モデルパラメータ数を変化させた場合の評価結果であり，平均 *Corr* と平均 *Acc* ともに混合分布数 32 の段階でほぼ性能が上限に達している．また，RTF は混合分布数の増加に比例して，ほぼ線形的に増加している．一般に，モデルパラメータ数が増加すると性能が向上する傾向にあるが，混合分布数 256，512 などの場合では性能が劣化する傾向にある．これは，学習データの不足から生じる確率モデルの過学習が原因であると考えられる．

結果として，混合分布数 128 の時に平均 *Corr* と平均 *Acc* ともに最良の結果が得られているが，32 の場合に比べて性能はほぼ同等であるものの，計算量に約 4 倍の差がある．よって本研究では，計算量を考慮にいれて，混合分布数 32 の場合を最適な結果として採用する．

6.4 帯域制限の効果

特徴抽出時に，周波数低域と高域（上限 4,000 Hz）の両方に帯域制限を適用し，性能がどのように変化するかを調査した．帯域制限は PAR と SKF 双方に対して行い，SKF において対数メルスペクトルの次元は有効周波数帯域に関わらず 24 に固定した．重みの決定方法は線形尺度，無音及び，音声 GMM の混合分布数は 32 である．

表 4 は帯域制限による評価結果を示しており，遮断周波数が低域 0 Hz，高域 4,000Hz の場合は帯域制限を行わないことを示している．結果として，平均的には帯域を制限することにより性能が劣化しているが，雑音環境が高速道路付近（St.）の場合は，低域を 250〜500 Hz 遮断することにより性能改善が得られている．これは，高速道路付近の主な雑音源が自動車の走行音であり，雑音のエネルギーが低域に集中していることに起因する．この帯域を遮断することにより，雑音抑圧的な効果が得られており，雑音の影響が少ない高域の情報を用いることにより性能改善が得られている．逆に，高域を遮断して低域のみを利用した場合は，雑音が低域に持つ強いエネルギーの影響で性能が劣化する．この傾向が表 4 の結果に顕著に現れている．

一方，雑音環境が学生食堂（Rest.）の場合は，主な雑音源が周囲の人の話し声であるため，雑音のエネルギー分布が目的音声とほぼ同一であり，全帯域の情報を用いて，総合的に識別を行う必要があると考えられる．実際に，学生食堂にて帯域制限を行うと大きく性能が劣化することが表 4 の結果より読み取れる．

以上のことから，雑音の種類によっては周波数の帯域制限が効果的に作用するが，逆効果となる雑音も存在する．本研究では，より汎用的な手法を目指しているため，特徴量の帯域制限は行わず，全帯域の情報を用いることとする．

表 4　帯域制限を行った場合の評価結果 (%)

遮断周波数 低域 (Hz)	遮断周波数 高域 (Hz)	*Corr* (%)					*Acc* (%)				
		Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.
0	1,000	84.49	53.62	97.39	88.99	81.12	65.22	-26.09	89.57	69.28	49.50
0	2,000	88.12	61.74	96.23	85.51	82.90	55.36	-3.48	88.12	66.38	51.60
0	3,000	90.43	62.61	99.13	93.33	86.38	64.93	5.22	95.94	84.64	62.68
0	4,000	**93.04**	**70.72**	**100.00**	97.97	**90.43**	**72.75**	**19.71**	99.13	94.78	**71.59**
250	4,000	90.72	69.28	99.71	99.13	89.71	61.16	16.52	**99.71**	**99.13**	69.13
500	4,000	87.83	65.22	99.71	**99.71**	88.12	63.77	10.14	97.97	98.26	67.54
1,000	4,000	88.70	65.80	99.42	98.55	88.12	68.41	11.01	96.81	96.52	68.19
2,000	4,000	87.25	62.61	99.42	97.68	86.74	65.80	8.70	95.94	92.17	65.65
3,000	4,000	86.38	62.32	97.97	96.23	85.73	53.04	7.83	93.04	90.14	61.01

表 5　他手法との比較結果 (%)

手法	*Corr* (%)					*Acc* (%)				
	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.	Rest. Hi.	Rest. Lo.	St. Hi.	St. Lo.	Ave.
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
PAR	70.72	57.10	87.25	80.58	73.91	24.35	-6.67	64.35	54.49	34.13
SKF	89.57	66.96	**100.00**	**97.97**	88.63	68.41	12.46	97.68	93.62	68.04
MUSCLE-VAD	**93.04**	**70.72**	**100.00**	**97.97**	**90.43**	**72.75**	**19.71**	**99.13**	**94.78**	**71.59**

6.5 他手法との比較

表 5 は，VAD の評価結果を示しており，"Baseline" は，CESNREC-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，"PAR"，"SKF" は，それぞれ PAR，SKF 単体での評価結果，"MUSCLE-VAD" は提案手法（重み決定: 線形，混合分布数: 32，帯域制限無し）の結果である．

表 5 の結果から，提案手法により従来方と比べて全ての雑音環境で性能改善が得られていることがわかる．SKF は単体でも高い VAD 性能を示しているが，PAR を組み合わせることにより，それぞれの手法の得手不得手とする環境での性能を補いあい，より高い VAD 性能を示したといえる．

また，表 1 の技術対比表では提案手法は他手法に比べて RTF が増大していたが，表 3 の結果より，モデルパラメータを半減し，RTF を削減しても他手法よりも高い性能を示している．よって提案手法は，他手法よりも高速，高性能な VAD を実現していることとなる．

7. むすび

本研究では，音声の周期性・非周期性成分比と，Switching カルマンフィルタの統合手法である MUSCLE-VAD について述べ，それぞれ単体の技術を用いる場合に比べて，VAD の性能改善が得られることを示した．また，MUSCLE-VAD の尤度重み付け方法，モデルパラメータ数，VAD 性能と処理時間の関係，特徴量の周波数帯域制限による影響について調査を行い，CENSREC-1-C での評価を通じて最適な動作条件を見出した．今後，様々な雑音環境下で統合手法の評価を行い，今回検討した項目及び，その他の点についてより一般的な評価を行い，妥当性の検証を行う予定である．

謝辞　本研究では，IPSJ SIG-SLP 雑音下音声認識評価ワーキンググループ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 CENSREC-1 (AURORA-2J) を使用した．

文　献

[1] Rabiner, L. R. *et al.*, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, Vol. 54, No. 2, pp. 297–315, Feb. 1975.

[2] Ramirez, J. *et al.*, "Efficient voice activity detection algorithm using long-term speech information," *Speech Communication*, Vol. 42, pp. 271–287, Apr. 2004.

[3] Ishizuka, K. *et al.*, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio," *Proc. of SAPA '06*, pp.65–70, Sept. 2006.

[4] Sohn, J. *et al.*, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, Vol. 6, No. 1, pp. 1–3, Jan. 1999.

[5] Fujimoto, M. *et al.*, "Noise robust voice activity detection based on switching Kalman filter," *Proc. of Intespeech '07*, pp. 2933-2936, Aug. 2007.

[6] ITU-T Recommendation G.729 Annex B., "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," Nov. 1996.

[7] 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.

[8] Xavier S., *et al.*, "Spectral modeling zynthesis:A sound analysis / synthesis based on a deterministic plus stochastic decomposition," Computer Music Journal, Vol. 14, No. 4 pp. 12–24, Dec. 1990.

[9] Hess, W., "Pitch determination of speech signals," Springer–Verlag, New York, 1983.

[10] Balakrishnan, A.V., "Kalman filtering theory," Springer–Verlag, Feb. 1984.

[11] CENSREC-1-C Web site, http://sp.shinshu-u.ac.jp/CENSREC/en/CENSREC/CENSREC-1-C/

[12] Nakamura, S., *et al.*, "AURORA-2J, An Evaluation Framework for Japanese Noisy Speech Recognition," IEICE Trans. on Information and Systems, Vol. E88–D, No. 3, pp. 535–544, March 2005.