2007－SLP－67 (13)
2007／7／21

音声の周期性・非周期性成分比とSwitching Kalman filterに基づく雑音下音声区間検出

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，音声の周期性・非周期性成分比と Switching Kalman filter に基づく雑音に頑健な音声区間検出法について検討を行う．提案手法は，音声特徴抽出部（音声の周期性・非周期性成分比）と，音声／非音声識別部（Switching カルマンフィルタに基づく識別）の双方に頑健性を有しており，それぞれを統合的に扱うことにより，雑音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，検出された音声信号の音声認識評価を行い，提案法が音声認識性能の改善に寄与することを示す．

キーワード: 音声区間検出，周期性・非周期性成分比，Switching カルマンフィルタ，CENSREC-1-C

Noise Robust Voice Activity Detection Based on Periodic to Aperiodic Component Ratio and Switching Kalman Filter

Masakiyo Fujimoto Kentaro Ishizuka 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 investigates a noise robust voice activity detection based on periodic to aperiodic component ratio and switching Kalman filter. The proposed method has the robustness for both discriminative feature extraction part (speech periodic to aperiodic component ratio) and speech / non-speech discrimination part (discrimination based on switching Kalman filter), and shows significant improvement of voice activity detection in noise. In addition, we carried out the evaluations of speech recognition by using detected speech signal, and confirmed that the proposed method contributes to improvement of speech recognition accuracy.

Keywords: voice activity detection, periodic to aperiodic component ratio, switching Kalman filter, CENSREC-1-C

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][4]．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 [5]．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．

よって，VAD においては特徴抽出部と識別部がそれぞれ重要な役割を担っており，双方の特性を有効に利用することにより，より高い性能を示すことができると言える．このことから本研究では，特徴抽出部と識別部それぞれに頑健性をもつ VAD 構築の第一次検討を行い，特徴量には，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）[4] を，識別部として，確率モデルに基づく方法 [6] を採用する方法について報告する．

特徴量については，確率モデルに基づく識別機構で用いる特徴量である対数メルスペクトルに，PAR を追加する形

式になっており，単純にベクトルの要素にPARを追加する方法と，各々の特徴量独立で音声／非音声状態の尤度を計算した後に，重み付け加算する方法の比較を行った．つまり，シングルストリームの識別と，マルチストリームの識別の比較を行っており，結果として，マルチストリーム識別により，最良の音声区間検出性能が得られることを示す．

また，検出後の音声信号を用いて音声認識による評価を行い，提案法が音声認識性能の改善に寄与することを示す．

2 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[7]．ここで，音声信号，特に母音部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．

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[8]により求める．

一方，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 の左右対称な分析窓関数 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[4]を参照されたい．

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

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

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．

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

3 確率モデルに基づくVADと状態遷移モデル

3.1 確率モデルに基づくVAD

Sohnらにより提案された確率モデルに基づくVAD[5]では，図1のような状態遷移を持つ確率モデルを定義し，当該フレームにおける非音声状態（H_0）と音声状態（H_1）の尤度比を求め，閾値処理を行う．尤度比が閾値以上であれば，当該フレームが音声フレーム，そうでなければ，非音声フレームと判別される．

Sohnらの手法では，雑音が既知かつ定常的であるという前提条件がおかれている．この条件下では，既知の雑音とそれを重畳した音声データを用いて，音声（音声＋雑音），非音声（雑音）状態の確率モデルを事前に学習することが可能である．しかし多くの場合，現実環境で観測される雑音は未知かつ非定常的であり，事前にこれらの確率モデルを学習しておくことは難しく，環境の変化に対応できないという問題が生じる．よって本研究では，音声／非音声状態の確率モデルを適応的，かつ逐次的に構成し，環境の変化に対して頑健なVADの構築を検討する．

3.2 状態遷移モデルの定義

まず，雑音が非定常的であることから，図2に示すような，常に状態遷移を伴う信号であると仮定する．次に，音声はSohnらの手法と同様に，図1の状態遷移モデルを持つものとする．ただし，雑音の無いクリーン音声状態と無音状態を持つものとする．これら2種類の状態遷移モデルを定義し，それぞれを合成することにより，雑音環境下での信号の状態遷移モデルを図3のように定義することができる．すなわち，音声，雑音それぞれに状態遷移を持つ確率過程となり，音声は離散的，雑音は連続的な状態遷移過程を有している．

図3のような状態遷移モデルを定義するにあたり，音声の状態遷移モデルに関しては，クリーン音声データを用い

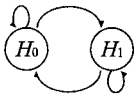

図 1: 音声／非音声状態遷移モデル

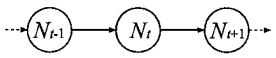

図 2: 雑音の状態遷移モデル

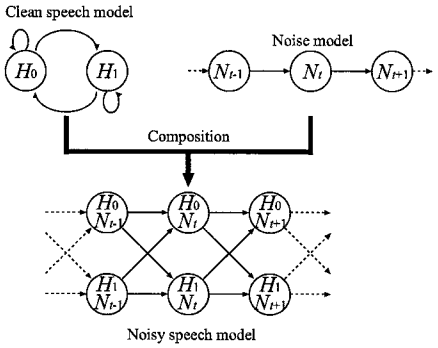

図 3: 非定常性雑音を考慮した音声／非音声状態遷移モデル

ることにより，音声（有音）／非音声（無音）それぞれの状態の確率モデル（本研究では，GMM（Gaussian Mixture Model）を利用）を事前に学習しておくことが可能である．一方，雑音の状態遷移モデルに関しては，雑音が事前に未知であることから，カルマンフィルタによる逐次推定を行う．また，2つの状態遷移モデルの合成，尤度計算を含めて，カルマンフィルタの枠組で解決できることを後に示す．

3.3 状態遷移モデルの定式化と尤度比の算出

図 3 の状態遷移モデルに基づき，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，

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

q_t と $\mathbf{N}_t$ の状態遷移がそれぞれ独立の事象であるとすると，

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

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，式 (10) は次式で表現される．

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

上式において，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．よって，上式は次式のように簡略化される．

$$\alpha_{j,t} = \sum_{i=0}^{1} (a_{i,j}\alpha_{i,t-1})\, b_{j,\mathbf{N}_t}(\mathbf{O}_t) \tag{12}$$

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．

$$q_t = \begin{cases} H_0 & R_t < Threshold \\ H_1 & R_t \geq Threshold \end{cases} \tag{13}$$

一方，図 2 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 [9]．

$$\begin{aligned} &p(\mathbf{O}_{0:t}, \mathbf{N}_{0:t}) \\ &\quad = p(\mathbf{N}_t|\mathbf{N}_{t-1})\, p(\mathbf{O}_t|\mathbf{N}_t)\, p(\mathbf{O}_{0:t-1}, \mathbf{N}_{0:t-1}) \end{aligned} \tag{14}$$

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 カルマンフィルタの確率表現と一致することを意味している．また前節にて述べた，「状態遷移モデルの合成，尤度計算をカルマンフィルタの枠組みで解決できる」という点とも符合する．

3.4 カルマンフィルタによる雑音推定と尤度計算

カルマンフィルタによる雑音推定について述べる．まず，カルマンフィルタの適用に必要な，状態空間モデルを構成する．状態空間モデルは，目的信号の状態遷移を表現した状態方程式と，観測信号の生成機構を表現した観測方程式から構成され，本研究では状態方程式に次式の Random walk 過程を用いる．

$$N_{t+1,l} = N_{t,l} + W_{t,l} \tag{15}$$

$$W_{t,l} \sim \mathcal{N}\left(0, \sigma^2_{W_l}\right) \tag{16}$$

上式において，$N_{t,l}$ はベクトル $\mathbf{N}_t$ の第 l 要素，$W_{t,l}$ は駆動雑音と呼ばれる平均 0，分散 $\sigma^2_{W_l}$ の白色ガウス雑音である．

一方，観測方程式には次式を用いる．

$$\begin{aligned} O_{t,l} &= S_{t,l} + \log\left(1 + \exp\left(N_{t,l} - S_{t,l}\right)\right) \\ &= f\left(S_{t,l}, N_{t,l}\right) \end{aligned} \tag{17}$$

上式において，$O_{t,l}$ は $\mathbf{O}_t$ の第 l 要素，$S_{t,l}$ はクリーン音声の対数メルスペクトルである．式 (17) の構成には，クリーン音声の対数メルスペクトル $S_{t,l}$ が必要だが，雑音の推定を行う時点で $S_{t,l}$ は未知であるため，次式のように音声及び，無音の K 混合分布 GMM のパラメータで代用する．

$$O_{t,l} = f\left(\mu_{S_{j,k,l}}, N_{t,l}\right) + V_{t,j,k,l} \tag{18}$$

$$V_{t,j,k,l} \sim \mathcal{N}\left(0, \sigma^2_{S_{j,k,l}}\right) \tag{19}$$

上式において，$\mu_{S_{j,k,l}}$ は GMM j ($j=0$: 無音 GMM，$j=1$: 音声 GMM) に含まれる要素分布 k の平均ベクトルの第 l 要素である．また，$V_{t,j,k,l}$ は $S_{t,l}$ と $\mu_{S_{j,k,l}}$ 間の誤差信号であり，平均 0，分散 $\sigma^2_{S_{j,k,l}}$ (GMM j，要素分布 k の共分散行列の第 l 対角要素) の白色ガウス雑音であるとする．

以上より，本研究では，式 (15)，(18) で構成される非線形な状態空間モデルを用いてカルマンフィルタを構成する．なお，このような非線形な状態空間モデルから構成されるカルマンフィルタは非線形カルマンフィルタと呼ばれる．また，各 GMM には K 種類の正規分布が含まれているため，GMM 毎に K 種類のカルマンフィルタを構成し，K 種類の推定結果を得ることができる．得られた K 種類の推定結果は，重み付け平均により，1 つの推定結果に集約される．このような手法を，本研究では，並列非線形カルマンフィルタと呼ぶ．なお，$K=1$ の場合は，従来の非線形カルマンフィルタと等価である．各非線形カルマンフィルタによる雑音の逐次推定は，文献 [6] を参照されたい．

カルマンフィルタの過程において，$O_{t,l}$ の平均と分散の推定値，$\mu_{O_{t,j,k,l}}$ と $\sigma^2_{O_{t,j,k,l}}$ がそれぞれ次式で得られる．

$$\mu_{O_{t,j,k,l}} = f\left(\mu_{S_{j,k,l}}, N_{t,j,k,l}\right) \tag{20}$$

$$\sigma^2_{O_{t,j,k,l}} = F_{t,j,k,l}\sigma^2_{N_{t,j,k,l}}F_{t,j,k,l} + \sigma^2_{S_{j,k,l}} \tag{21}$$

$$F_{t,j,k,l} = \partial\mu_{O_{t,j,k'}}/\partial N_{t,j,k,l} \tag{22}$$

上式において，$N_{t,j,k,l}$ と $\sigma^2_{N_{t,j,k,l}}$ はそれぞれ時刻 t において，GMM j，分布 k のパラメータを用いて構成されたカルマンフィルタによる雑音の推定値及び，誤差分散である．

式 (20)，(21) のパラメータを用いて，次式により各状態の出力確率 $b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ を求める．

$$b_{j,\mathbf{N}_t}(\mathbf{O}_t) = \sum_{k=1}^{K} w_{S_{j,k}} \mathcal{N}\left(\mathbf{O}_t; \mu_{O_{t,j,k}}, \Sigma_{O_{t,j,k}}\right) \tag{23}$$

上式において，$w_{S_{j,k}}$ は音声及び，無音 GMM の混合重みであり，$\mu_{O_{t,j,k}}$ は $\mu_{O_{t,j,k,l}}$ を要素に持つベクトル，$\Sigma_{O_{t,j,k}}$ は $\sigma^2_{O_{t,j,k,l}}$ を対角要素に持つ行列である．

つまり，前述の「状態遷移モデルの合成，尤度計算を並列非線形カルマンフィルタの枠組みで解決できる」という点を体現している．また，パラメータを取得する GMM の種別，つまり (クリーン) 音声の状態 q_t によりカルマンフィルタのフィルタ方程式及び，推定結果が変化するため，Switching カルマンフィルタとしての特性も備えている．

4 特徴抽出部と識別部の統合

2 と 3 にて述べた，PAR と Switching カルマンフィルタの統合について述べる．本研究では，ベクトルレベルでの統合と，尤度計算レベルの 2 種類の統合方法について検討を行っており，まず，前者について述べる．

3 にて述べた方法は，L 次元の対数メルスペクトルを特徴ベクトルとして音声，無音の GMM を学習していたが，このベクトルの要素に PAR を追加する．つまり，観測信号ベクトルを $\mathbf{O}_t = \{O_{t,0}, \cdots, O_{t,L-1}, PAR_t\}$ と定義し，PAR を音声認識におけるパワー項と同様に扱うこととなる．ただし，3.4 にて述べた雑音の推定を行う際には，PAR 項を除いた要素のみで処理を行う．識別部における尤度計算は，式 (23) により行うが，モデルパラメータはそれぞれ，$\mu_{O_{t,j,k}} = \left\{\mu_{O_{t,j,k,0}}, \cdots, \mu_{O_{t,j,k,L-1}}, \mu_{PAR_{t,j,k}}\right\}$，$\Sigma_{O_{t,j,k}} = diag\left\{\sigma^2_{O_{t,j,k,0}}, \cdots, \sigma^2_{O_{t,j,k,L-1}}, \sigma^2_{PAR_{t,j,k}}\right\}$ であり，$\mu_{PAR_{t,j,k}}$，$\sigma^2_{PAR_{t,j,k}}$ は，事前に学習した音声，無音 GMM の PAR 項の平均値と分散値である．

次に，尤度計算レベルでの統合について述べる．この方法は，PAR と Switching カルマンフィルタ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である．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．

$$\begin{aligned} & b_{j,PAR}(PAR_t) \\ &= \sum_{k=1}^{K} w_{PAR_{j,k}} \mathcal{N}\left(PAR_t; \mu_{PAR_{t,j,k}}, \sigma^2_{PAR_{t,j,k}}\right) \end{aligned} \tag{24}$$

その後，式 (23) の Switching カルマンフィルタの尤度 $b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ を次式のように，重み γ を用いて加算する．

$$b_j(\mathbf{O}_t, PAR_t) = \gamma b_{j,\mathbf{N}_t}(\mathbf{O}_t) + (1-\gamma) b_{j,PAR}(PAR_t) \tag{25}$$

上式により得られた尤度 $b_j(\mathbf{O}_t, PAR_t)$ を用いて式 (12) の前向き確率を計算し，識別を行う．重みの値は $0 \leq \gamma \leq 1$ であり，$\gamma = 0$ の場合は PAR 単体での識別，$\gamma = 1$ の場合は Switching カルマンフィルタ単体での識別となる．

以上，二種類の統合方法を述べたが，ベクトルレベルでの統合は，尤度計算機構が一系統であることから，シングルストリームによる識別，尤度計算レベルでの統合は尤度計算機構が二系統であることから，マルチストリームによる識別と位置づけることができる．

5 CESNREC-1-Cによる評価

5.1 CESNREC-1-C

提案手法の評価は，VADの評価用データベースCENSREC-1-C [10]を用いて行う．CESNREC-1-Cは，人工的な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，実環境での音声品質劣化の影響を調査するため，実データを用いて評価を行う．

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（騒音レベル60 dB(A)前後）とLow SNR（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[10]を参照されたい．

5.2 VADの評価結果

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$L = 24$とした．音声の状態遷移確率は，$a_{i,j} = \{0.90, 0.10, 0.45, 0.55\}$とし，カルマンフィルタにおける駆動雑音の分散値は，$\sigma^2_{W_i} = 0.0001$とした．無音及び，音声GMMの学習は，AURORA-2J [11]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．尚，4にて述べた，シングルストリーム識別の評価を行う際には，対数メルスペクトル24次元にPAR1次元を加えた25次元の特徴量を用いて学習したGMMを用いる．一方，マルチストリーム識別の評価を行う際には，対数メルスペクトル24次元で学習したGMMとPAR1次元で学習したGMMの2種類を利用する．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．

$$Corr = N_c/N \times 100\ [\%] \tag{26}$$

$$Acc = (N_c - N_f)/N \times 100\ [\%] \tag{27}$$

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(13)の閾値と式(25)のγ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．

表1において，Single-PARとMulti-PARが提案手法の結果を示しており，比較のため，Sohnらの方法，PARを対数パワーに置き換えた結果，対数パワー，PAR，Switchingカルマンフィルタ単体での結果も同様に示している．

結果より，提案手法であるMulti-PARにより最良の結果が得られた．もう一つの提案手法であるSingle-PARは，Switchingカルマンフィルタ単体の結果よりも劣化している．シングルストリーム識別では，異なる特徴量を一本のベクトルにまとめている．しかし一般に特徴量が異なると，ダイナミックレンジや時間的な変動量がそれぞれ異なるため，尤度計算時にある特定の特徴量の影響が強く現れ，他の特徴量が意味を成さなくなる恐れがある．一方，マルチストリーム識別では，尤度計算を特徴量間で独立に行い，重み付け加算によりそれぞれの特徴量による尤度のバランス調整を行っているため，特定の特徴量の影響を強く受けるという問題を回避できる．このような効果により，Multi-PARがSingle-PARよりも高い性能を示したと言える．特に，PARを対数パワーに置き換えたMulti-Pow，Single-Powでは，この効果が顕著に現れている．対数パワー単体でのVAD性能は低くなっており，Single-Powではベクトル内のパワー項の影響を強く受けたため，Multi-Powに比べて大きく性能が劣化したと言える．また，Multi-PARとMulti-Powでは，PARと対数パワーの性能差が現れており，PARが対数パワーよりも雑音に頑健な特徴量であることがわかる．

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．音声認識評価において，HMMの学習データ，実験条件はAURORA-2J [11]と同様であり，HMMはクリーンモデル（CMN無し）である．

表2は認識結果を示しており，表中，w/o VADはVADを行わずに複数発話音声を認識した結果，Ideal VADは，真の発話境界情報を用いた場合の結果であり，Relative improvementは，w/o VADからの誤り改善率を示す．

表2の結果より，提案手法Multi-PARが最良の認識結果を示しており，平均認識率においてIdeal VADとほぼ同等の認識率が得られた．この結果より，提案手法が雑音下における連続発声音声の認識性能改善に寄与することが分かる．現状，雑音抑圧処理等を行っていないため絶対的な認識性能は低いが，音声区間検出を正確に行うことにより，非音声区間での単語湧き出し等を防ぐことができ，認識性能を改善させることができる．今後，雑音抑圧処理等を加えることにより，更なる認識性能改善が期待できる．

6 むすび

本研究では，音声特徴抽出部（音声の周期性・非周期性成分比）と，音声／非音声識別部（Switchingカルマンフィルタに基づく識別）の双方に頑健性を有するVADを提案

表 1: VAD 評価結果（Baseline: CENSREC-1-C ベースライン，Sohn: Sohn らの方法，Single-Pow: シングルストリーム識別（パワー），Multi-Pow: マルチストリーム識別（パワー，$\gamma = 0.8$），Single-PAR: シングルストリーム識別（PAR），Multi-PAR: マルチストリーム識別（PAR，$\gamma = 0.8$），Pow: パワー（Multi-Pow，$\gamma = 0$），PAR: PAR（Multi-PAR，$\gamma = 0$），SKF: Switching カルマンフィルタ（Multi-PAR，$\gamma = 1$））

	Corr (%)					*Acc* (%)				
	Rest. High	Rest. Low	St. High	St. Low	Average	Rest. High	Rest. Low	St. High	St. Low	Average
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
Single-PAR	89.57	67.54	99.13	95.94	88.05	63.48	18.55	95.36	89.86	66.81
Multi-PAR	**93.04**	**69.86**	**100.00**	99.42	**90.58**	**75.94**	**18.84**	**97.97**	**97.68**	**72.61**
Single-Pow	91.59	1.45	96.52	26.96	54.13	73.33	-13.91	80.58	-8.70	32.83
Multi-Pow	88.41	65.80	**100.00**	**99.71**	88.48	68.99	16.52	96.81	**97.68**	70.00
Pow	90.14	0.00	64.64	10.43	41.30	70.43	10.43	26.38	-10.14	19.06
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

表 2: 音声認識結果（w/o VAD: VAD 無し，Baseline: CENSREC-1-C ベースライン，Ideal VAD: 真の VAD 結果，Sohn: Sohn らの方法，Single-Pow: シングルストリーム識別（パワー），Multi-Pow: マルチストリーム識別（パワー，$\gamma = 0.8$），Single-PAR: シングルストリーム識別（PAR），Multi-PAR: マルチストリーム識別（PAR，$\gamma = 0.8$），Pow: パワー（Multi-Pow，$\gamma = 0$），PAR: PAR（Multi-PAR，$\gamma = 0$），SKF: Switching カルマンフィルタ（Multi-PAR，$\gamma = 1$））

	Word accuracy (%)					Relative improvement (%)				
	Rest. High	Rest. Low	St. High	St. Low	Average	Rest. High	Rest. Low	St. High	St. Low	Average
w/o VAD	45.17	1.28	34.43	25.23	26.53	0.00	0.00	0.00	0.00	0.00
Baseline	44.16	18.12	29.96	21.62	28.47	-1.84	17.06	-6.82	-4.83	2.64
Ideal VAD	*52.67*	*29.17*	*41.25*	*29.50*	*38.15*	*13.68*	*28.25*	*10.40*	*5.71*	*15.82*
Sohn	37.45	-3.81	33.41	29.58	24.16	-13.83	-6.22	-1.46	5.55	-3.31
Single-PAR	43.36	20.60	44.68	32.79	35.36	-3.30	19.57	15.63	10.11	12.02
Multi-PAR	49.13	**22.31**	**47.54**	**33.42**	**38.10**	7.22	**21.30**	**19.99**	**10.95**	**15.75**
Single-Pow	51.65	0.18	44.08	24.59	30.13	11.82	-1.11	14.72	-0.86	4.90
Multi-Pow	44.06	9.84	45.74	32.79	33.11	-2.02	8.67	17.25	10.11	8.96
Pow	**52.60**	1.73	39.07	24.85	29.56	**13.55**	0.46	7.08	-0.51	4.13
PAR	39.76	8.89	39.16	24.08	27.97	-9.87	7.71	7.21	-1.54	1.97
SKF	43.75	12.50	46.99	33.15	34.10	-2.59	11.37	19.16	10.59	10.30

し，それぞれ単体の技術を用いる場合に比べて，VAD，音声認識双方に性能改善が得られることを示した．また，シングルストリーム識別とマルチストリーム識別の比較を行い，マルチストリーム識別が特徴量の統合に有効であることを示した．今後，マルチストリーム識別の重み，閾値等のパラメータ適応的決定法について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 AURORA-2J を使用した．

参考文献

[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.*, "A feature for voice activity detection derived from speech analysis with the exponential autoregressive model," Proc. ICASSP '06, Toulouse, France, Vol. I, pp. 789–792, May 2006.

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

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

[6] 藤本他, "音声と雑音両方の状態遷移過程を有する雑音下音声区間検出," 信学技報, SP2006–87, pp. 13–18, Dec. 2006.

[7] Xavier S., *et al.*, "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition," Computer Music Journal, Vol. 14, No. 4 pp. 12–24, Dec. 1990.

[8] Hess, W., "Pitch Determination of Speech Signals," Springer–Verlag, Dec. 1983.

[9] Balakrishnan, A.V., "Kalman Filtering Theory," Springer–Verlag, Feb. 1984.

[10] 北岡他, "雑音下音声区間検出手法評価基盤の構築," 音響講論, 1–P–27, pp. 103–104, Sept. 2006.

[11] 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.