Vol.2009-SLP-79 No.12
2009/12/22

Dirichlet 事前分布を用いた音声区間検出の検討

藤本雅清[†1]　渡部晋治[†1]　中谷智広[†1]

本研究では，確率モデルに基づく音声区間検出法に Dirichlet 事前分布を導入して音声区間検出の性能改善が得られることを述べる．確率モデルに基づく音声区間検出法では，Switching カルマンフィルタを用いて観測信号の環境に適応した音声／非音声 GMM を逐次生成し，各 GMM の確率比に基づき音声／非音声の識別を行っている．生成される GMM には不要な分布と重要な分布が混在しており，不要な分布を取り除き，重要な分布のみを用いることにより VAD の性能改善が得られる．分布の削減を行うと，削減前の混合分布とは分布全体の形状が大きく異なり，分布の事前確率も大きく異なる．このため，本研究では，事前分布を Dirichlet 分布で定義し，分布選択後の混合重みを最適化することについて検討を行った．

Voice Activity Detection Using Dirichlet Prior

Masakiyo Fujimoto ,[†1] Shinji Watanabe [†1]
and Tomohiro Nakatani[†1]

This paper introduce the Drichlet prior into a statistical model-based voice activity detection (VAD), and shows its advantage. The statistical model-based VAD identify speech / non-speech period based on environmental adapted speech and non-speech GMMs which are constructed by the Switching Kalman filter. The constructed GMMs include important and unimportant Gaussian distributions. Thus, the performance of VAD can be improved by reducing unimportant Gaussian distribution. Here, prior probabilities of each remaining distribution may drastically change, because the distribution shape after the Gaussian reduction is much different from original GMM. Thus, we propose an optimization method of prior probabilities by using the Drichlet prior.

†1 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
NTT Communication Science Laboratories, NTT Corporation

1. はじめに

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．これらの処理の内，VAD は音声認識，雑音抑圧，残響除去のみならず，音声符号化などにも活用される，音声処理の基盤とも言える重要な技術である．このため，VAD の性能改善，特に雑音等に対する頑健性の確立は極めて重要な問題である．

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] の統合法を提案した[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．また，SKF に基づく VAD では，観測信号の環境に適応した音声／非音声 GMM (Gaussian Mixture Model) を逐次生成し，各 GMM の出力確率比の閾値処理により音声／非音声の識別を行っているが，出力確率計算の際に，各 GMM 内の不要な分布を取り除き，重要な分布のみを用いる方法の検討を行った[4]．

以前検討を行った GMM 内の分布削減法では，分布削減後の混合重み（分布の事前確率）を，選択された分布の混合重みの合計値で正規化するという単純な方法により求めていた．しかし分布の削減を行うと，削減前の混合分布とは分布全体の形状が大きく異なり，それに伴い，分布の事前確率も大きく異なることが考えられる．このため，本研究では，事前分布を Dirichlet 分布[5] で定義し，分布選択後の混合重みを最適化することについて検討を行った．

2. SKF に基づく VAD

ここでは，SKF に基づく VAD の概要を述べる．我々の提案する VAD は，PAR と SKF を統合した手法であるが，本研究での着目点は主として SKF に関わる内容であるため，PAR の詳細および，統合方法に関しては割愛する．PAR 及び，統合方法の詳細は文献[3] を参照されたい．

2.1 状態遷移モデルの定義

SKF に基づく VAD 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，SKF に基づく VAD では，事前にクリーン音声データを用いて無音状態 H_0 とク

Vol.2009-SLP-79 No.12
2009/12/22

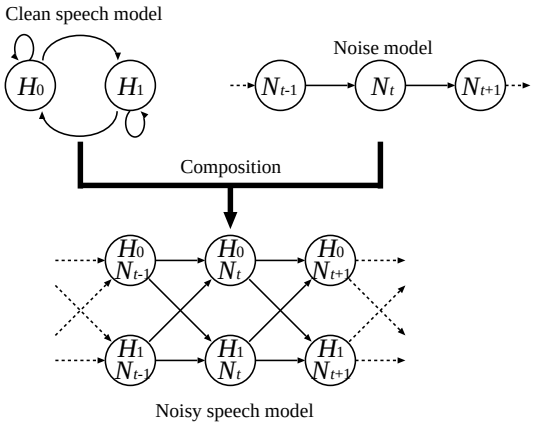


図 1 音声／非音声状態遷移モデル（H_0: 無音状態，H_1: クリーン音声状態，N_t: 雑音状態系列）

リーン音声状態 H_1 の GMM を学習し，図 1 の Clean speech model のような状態遷移モデルを構成する．また，雑音は図 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され，観測信号が与えられると，SKF により Noise model の状態 N_t が逐次更新される．この際，同時に図 1 の Noisy speech model が Clean speech model と Noise model との合成によって得られる．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．

なお，Noisy speech model の各状態は GMM で構成され，上段の各状態の GMM を非音声 GMM，下段の各状態の GMM を音声 GMM を呼ぶこととする．

2.2 状態遷移モデルの定式化と尤度比の算出

図 1 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 での観測信号の特徴量ベクトル $\mathbf{O}_t$（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，ベイズの定理より次式で与えられる．

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

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 $p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t})$ は次の再帰式で表現される．

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

$p(q_t|q_{t-1})$，$p(\mathbf{N}_t|\mathbf{N}_{t-1})$，$p(\mathbf{O}_t|q_t, \mathbf{N}_t)$ 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，$p(q_t = H_j|q_{t-1} = H_i) = a_{i,j}$，$p(\mathbf{N}_t|\mathbf{N}_{t-1}) = c_{t,t-1}$，$p(\mathbf{O}_t|q_t = H_j, \mathbf{N}_t) = b_j(\mathbf{O}_t)$ 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 $\alpha_{j,t}$ に相当する．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．

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

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[6]．

$$q_t = \begin{cases} H_0 & R_t < \text{Threshold} \\ H_1 & R_t \geq \text{Threshold} \end{cases} \tag{4}$$

上記の方法において，雑音状態の逐次更新，音声／非音声 GMM の生成は SKF により行われる．詳細については文献[2] を参照されたい．

2.3 重要な確率分布の選択

SKF により逐次生成される音声／非音声 GMM には，それぞれ K 個の正規分布が含まれているが，各時刻の観測信号 $\mathbf{O}_t$ との確率計算においては，全ての正規分布が重要ではなく，幾つかの不要な分布が存在する．それら不要な分布を取り除いて重要な分布のみを用いて出力確率 $b_j(\mathbf{O}_t)$ を得，式 (3) の前向き確率を求めることにより，音声／非音声の識別性能が改善すると考えられる．このような考えに基づき，本研究では各分布の事後確率の累積値を利用して GMM 内の重要な分布を取り出す方法について検討する．

まず，音声／非音声 GMM の各分布の事後確率 $w_{t,j,k}$ を次式により求める．

$$w_{t,j,k} = \frac{w_{j,k} \cdot \mathcal{N}\left(\mathbf{O}_t; \boldsymbol{\mu}_{O,t,j,k}, \boldsymbol{\Sigma}_{O,t,j,k}\right)}{\sum_{k'=1}^{K} w_{j,k'} \cdot \mathcal{N}\left(\mathbf{O}_t; \boldsymbol{\mu}_{O,t,j,k'}, \boldsymbol{\Sigma}_{O,t,j,k'}\right)} \tag{5}$$

式中，$w_{j,k}$，$\boldsymbol{\mu}_{O,t,j,k}$，$\boldsymbol{\Sigma}_{O,t,j,k}$ はそれぞれ，状態 j（$j = 0$: 非音声状態，$j = 1$: 音声状態），分布 k の混合重み，平均ベクトル，対角共分散行列である．

Vol.2009-SLP-79 No.12
2009/12/22

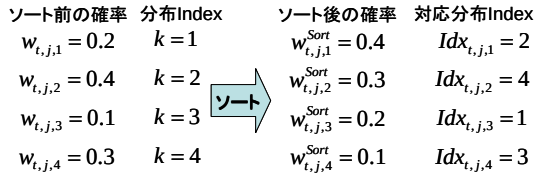


図 2 $w_{t,j,k}$ の降順ソートの例

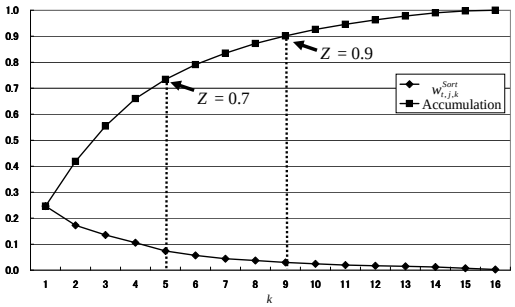


図 3 分布数 $M_{t,j}$ の決定例

その後，$w_{t,j,k}$ の値が高い分布を重要な分布であると仮定して $w_{t,j,k}$ を降順ソートし，ソート後の確率 $w^{Sort}_{t,j,k}$ と，ソート前とソート後の対応分布 Index$Idx_{t,j,k}$ を求める（以後，表記の簡単化のため，$m = Idx_{t,j,k}$ とする）．図 2 は，$K = 4$ としたときの降順ソートの例である．

要／不要な分布の選択は，降順ソートの結果から上位にあたる分布を選択することにより行い，実際には，次式のように $w^{Sort}_{t,j,k}$ の累積値が閾値 Z 以上となる最小の分布数 n を求めることにより必要な分布数 $M_{t,j}$ を決定する．

$$M_{t,j} = \arg\min_{n} \left\{ \sum_{k=1}^{n} w^{Sort}_{t,j,k} \geq Z \right\} \quad (0 < Z \leq 1) \tag{6}$$

図 3 は，分布数 $M_{t,j}$ の決定例（$K = 16$）であり，$Z = 0.7$ では $M_{t,j} = 5$，$Z = 0.9$ では $M_{t,j} = 9$ と決定される．

分布数 $M_{t,j}$ が決定された後，上位 $M_{t,j}$ 個の分布のみを用いて出力確率 $b_j(\mathbf{O}_t)$ を再計算する．まず，式 (7) により混合重み $w_{j,m}$ を $M_{t,j}$ 個のみの値を用いて正規化し，新たに得られた混合重み $\tilde{w}_{j,m}$ と式 (8) により，出力確率 $b_j(\mathbf{O}_t)$ が得られる．

なお，ここで述べた分布の選択はフレーム毎に行う処理であり，選択される分布はフレーム毎に異なる．

$$\tilde{w}_{j,m} = \frac{w_{j,m}}{\sum_{m'=1}^{M_{t,j}} w_{j,m'}} \tag{7}$$

$$b_j(\mathbf{O}_t) = \sum_{m=1}^{M_{t,j}} \tilde{w}_{j,m} \cdot \mathcal{N}\left(\mathbf{O}_t; \boldsymbol{\mu}_{O,t,j,m}, \boldsymbol{\Sigma}_{O,t,j,m}\right) \tag{8}$$

3. Dirichlet 事前分布の導入

2.3 節の方法では正規分布の削減後，式 (7) を用いて混合重みを正規化している．しかしながら，GMM から正規分布が削減されると，削減前の GMM の分布に比べて分布全体の形状が大きく変化することとなり，混合重み，すなわち事前確率もまた大きく変化することが考えられる．よって，式 (7) のような単純な正規化ではなく，式 (9) のように事前分布 $P(\boldsymbol{w}_j)$ を定義して最適な混合重みを再推定し，出力確率 $b_j(\mathbf{O}_t)$ を再計算することが必要となる．なお式中，$\boldsymbol{w}_j = \left\{w_{j,1}, \cdots, w_{j,m}, \cdots, w_{j,M_{t,j}}\right\}$ である．

$$b_j(\mathbf{O}_t) \Leftarrow \hat{\boldsymbol{w}}_j = \arg\max_{\boldsymbol{w}_j} \{P(\boldsymbol{w}_j) \cdot b_j(\mathbf{O}_t)\} \tag{9}$$

このような事前分布 $P(\boldsymbol{w}_j)$ として，本研究では Dirichlet 分布[5] を採用する．Dirichlet 分布は，ベイズ学習等において混合重みの共役事前分布として頻繁に用いられる分布である．

Dirichlet 分布の確率密度関数は式 (10) により定式化され，式中 β_m はハイパーパラメータであり $\boldsymbol{\beta} = \left\{\beta_1, \cdots, \beta_m, \cdots, \beta_{M_{t,j}}\right\}$ と定義し，$B(\cdot)$ は，式 (11) で表される多変量ベータ分布である．また式 (11) 中 $\Gamma(\cdot)$ はガンマ分布である．

$$P(\boldsymbol{w}_j) = \frac{1}{B(\boldsymbol{\beta})} \prod_{m=1}^{M_{t,j}} w_{j,m}^{\beta_m - 1} \tag{10}$$

$$B(\boldsymbol{\beta}) = \frac{\prod_{m=1}^{M_{t,j}} \Gamma(\beta_m)}{\Gamma\left(\sum_{m=1}^{M_{t,j}} \beta_m\right)} \tag{11}$$

以上の定義に基づき，式 (9) の右辺を最大にする混合重み $\hat{w}_{j,m}$ を式 (12) で与えられる Q 関数を用いて推定する．なお，式 (12) において const は，$\hat{w}_{j,m}$ に依存しない定数項で

ある．

$$Q\left(\boldsymbol{w}_j, \hat{\boldsymbol{w}}_j\right) = \sum_{m=1}^{M_{t,j}} w_{t,j,m} \left(\log \hat{w}_{j,m} + \log \mathcal{N}\left(\mathbf{O}_t; \boldsymbol{\mu}_{O,t,j,m}, \boldsymbol{\Sigma}_{O,t,j,m}\right)\right) + \sum_{m=1}^{M_{t,j}} (\beta_m - 1) \log \hat{w}_{j,m} + \log B\left(\boldsymbol{\beta}\right) = \sum_{m=1}^{M_{t,j}} (w_{t,j,m} + \beta_m - 1) \log \hat{w}_{j,m} + \text{const} \quad (12)$$

ここで $\sum_{m=1}^{M_{t,j}} \hat{w}_{j,m} = 1$ であることから，ラグランジェの未定乗数法により式 (12) は次式のように表現される．

$$Q\left(\boldsymbol{w}_j, \hat{\boldsymbol{w}}_j\right) = \sum_{m=1}^{M_{t,j}} (w_{t,j,m} + \beta_m - 1) \log \hat{w}_{j,m} - \lambda \left(\sum_{m=1}^{M_{t,j}} \hat{w}_{j,m} - 1 \right) + \text{const} \quad (13)$$

式 (13) を $\hat{w}_{j,m}$ で変微分すると，

$$\frac{\partial Q\left(\boldsymbol{w}_j, \hat{\boldsymbol{w}}_j\right)}{\partial \hat{w}_{j,m}} = \frac{w_{t,j,m} + \beta_m - 1}{\hat{w}_{j,m}} - \lambda \quad (14)$$

となり，$\frac{\partial Q\left(\boldsymbol{w}_j, \hat{\boldsymbol{w}}_j\right)}{\partial \hat{w}_{j,m}} = 0$ とすることにより，

$$w_{t,j,m} + \beta_m - 1 = \lambda \cdot \hat{w}_{j,m} \quad (15)$$

$$\hat{w}_{j,m} = \frac{w_{t,j,m} + \beta_m - 1}{\lambda} \quad (16)$$

が得られる．ここで $\sum_{m=1}^{M_{t,j}} \hat{w}_{j,m} = 1$ であることから式 (15) は，

$$\sum_{m=1}^{M_{t,j}} (w_{t,j,m} + \beta_m - 1) = \lambda \sum_{m=1}^{M_{t,j}} \hat{w}_{j,m} = \lambda \quad (17)$$

と変形され，これを式 (16) に代入することにより，

$$\hat{w}_{j,m} = \frac{w_{t,j,m} + \beta_m - 1}{\sum_{m'=1}^{M_{t,j}} \left(w_{t,j,m'} + \beta_{m'} - 1\right)} \quad (18)$$

が得られ，正規分布数を削減した GMM の混合重みを最適化することができる．

以上の方法により得られた混合重み $\hat{w}_{j,m}$ を用いて次式のように出力確率 $b_j\left(\mathbf{O}_t\right)$ を得，式 (3) の前向き確率を計算する．

$$b_j\left(\mathbf{O}_t\right) = \sum_{m=1}^{M_{t,j}} \hat{w}_{j,m} \cdot \mathcal{N}\left(\mathbf{O}_t; \boldsymbol{\mu}_{O,t,j,m}, \boldsymbol{\Sigma}_{O,t,j,m}\right) \quad (19)$$

4. CENSREC-1-C による評価

4.1 CENSREC-1-C と実験条件

評価実験は，VAD の評価用に設計されたデータベース CENSREC-1-C[7] 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 (Restaurant) と高速道路付近 (Street) 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[7] 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．無音及び，音声 GMM の学習は，AURORA-2J[8] 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 K はそれぞれ 32 である．他の条件については，文献[3] と同様である．

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．

$$Corr = N_c/N \times 100\ [\%] \quad (20)$$

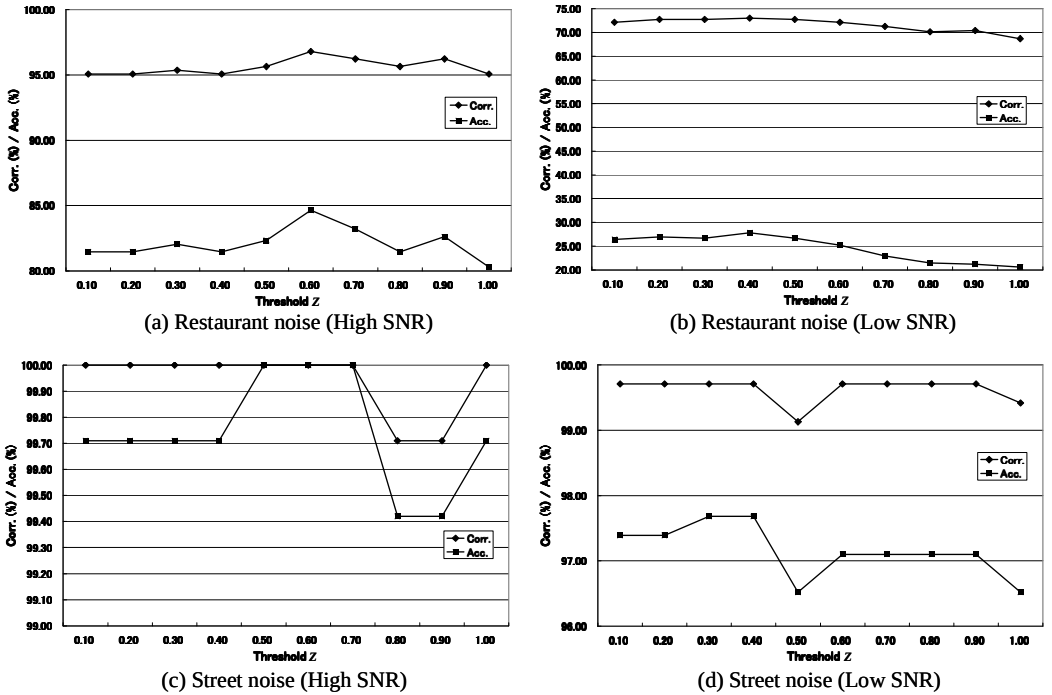
$$Acc = (N_c - N_f)/N \times 100\ [\%] \quad (21)$$

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．

4.2 確率分布選択の評価結果

まず 2.3 節で述べた，確率分布の選択方法の評価を行う．図 4 は，閾値 Z を 0.1〜1.0 まで変化させた場合の各雑音環境の VAD 評価結果であり，$Z = 1.0$ の場合は，分布の選択を行わず従来通り GMM 中の全ての分布を用いて確率計算を行うことを示している．

Vol.2009-SLP-79 No.12
2009/12/22



(a) Restaurant noise (High SNR)　(b) Restaurant noise (Low SNR)

(c) Street noise (High SNR)　(d) Street noise (Low SNR)

図 4　確率分布選択の実験結果

図 4 の結果から，全ての雑音環境下で Z の値を変化させることにより，$Z = 1.0$ の場合より高い $Corr$，Acc が得られることがわかる．図中，(c)，(d) の高速道路雑音の結果については，$Z = 1.0$ のときの性能がほぼ上限に達しているため，Z の変化による絶対的な効果は小さいが，(a)，(b) の学生食堂雑音の結果では，Z の変化による効果が顕著に現れており，確率分布選択の有効性が確認できる．

また，(a)，(c) では $Z = 0.6$，(b)，(d) では $Z = 0.4$ の時に最良の結果が得られている．このことから最適な Z の値は，SNR によって変化することが考えられ，実際に図 4 の結果では，High SNR の場合に $Z = 0.5$〜0.7 程度，Low SNR の場合に $Z = 0.2$〜0.4 程度の値で性能が高くなっている．すなわち，SNR が劣化するほど分布数を削減するような Z を選択すれば良いこととなる．これは SNR が劣化すると，雑音状態の推定誤差などが伴って音声／非音声 GMM 自身の性能が劣化し，特徴量空間の表現能力が劣化した分布，つまり不要な分布が増加するためであると考えられる．

4.3 Dirichlet 事前分布の評価結果

次に 3 章で述べた，Dirichlet 事前分布を用いた混合重み最適化の評価を行う．表 1 は，様々な手法による結果を示しており，“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[6]，“w/o Selection” は，前節における $Z = 1.0$ の結果，“Selection” は，前節における $Z = 0.6$ (High SNR)，$Z = 0.4$ (Low SNR) の結果，”Selection+Dirichlet Prior” は “Selection” に対して 3 章の Dirichlet 事前分布を用いた混合重み最適化を行った場合の結果である．なお Dirichlet 分布のハイパーパラメータには $\beta_m = 0.9$ を与えた．

表 1 Dirichlet 事前分布を用いた実験結果及び，他手法との比較 (%)

	Corr					*Acc*				
Noise	Restaurant		Street		Avg.	Restaurant		Street		Avg.
SNR	High	Low	High	Low		High	Low	High	Low	
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
w/o Selection	95.07	68.70	**100.00**	99.42	90.80	80.29	20.58	99.71	96.52	74.28
Selection	96.81	73.04	**100.00**	**99.71**	92.39	**84.64**	27.83	**100.00**	97.68	77.54
Selection+Dirichlet prior	**97.10**	**74.20**	**100.00**	**99.71**	**92.75**	82.32	**32.46**	**100.00**	**98.55**	**78.33**

表 1 の結果より，混合重み最適化を行うことで全体的な VAD 性能の改善が得られており，その効果が確認できる．この結果は，分布の削減を行うことにより混合重み，すなわち分布の事前確率が大きく変化していることを示唆しており，式 (7) のような単純な正規化では，混合重みとして不十分であることの証明となる．また，提案手法は SKF により観測信号の環境に適応した GMM を逐次生成しており，GMM 内の各分布の形状がフレーム単位で変化している．この分布形状の逐次的な変化も混合重みの変化の要因となると考えられる．これらのことから，事前分布を仮定して混合重みを最適化を行うことは妥当な手段であると言える．

本研究における評価では，それぞれの手法のパラメータである Z と β_m を実験的に求めているが，分布の削減数と混合重み $w_{j,m}$，もしくは分布の削減数とハイパーパラメータ β_m との関係について様々な雑音環境，SNR での実験を通じて検証し，各パラメータを自動で最適化する方法について検討を行う必要がある．

5. おわりに

本研究では，確率モデルに基づく VAD において，Dirichlet 事前分布を用いて，分布数が削減された GMM の混合重みを最適化する方法について見当を行い，効果的な処理であることを示した．今後，様々な雑音環境，SNR での実験での評価，検証を行い，分布選択のパラメータ Z と Dirichlet 分布のハイパーパラメータ β_m の最適化法について検討を行う．

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) 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) 藤本 雅清, 中谷 智広, “確率モデルに基づく音声区間検出法における確率分布選択と確率重み付けの検討,” 日本音響学会, 平成 21 年度秋季研究発表会, 1-1-14, pp. 43-46, Sept. 2009.
5) Bishop, C. M., “Pattern recognition and machine learning,” Springer, 2008.
6) 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.
7) 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.
8) 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.