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

2006－SLP－64（3）
2006／12／21

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

藤本 雅清† 石塚健太郎† 加藤比呂子†

† 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, katohi}@cslab.kecl.ny.co.jp

あらまし 本研究では，音声と雑音両方の状態遷移過程を有する雑音に頑健な音声区間検出法を提案する．提案法では，事前にクリーン音声と無音の状態遷移モデルを構成し，観測信号が入力されると並列非線形カルマンフィルタにより雑音を推定し，状態遷移モデルを雑音環境に逐次適応させる．適応したモデルを用いて，音声状態（クリーン音声＋雑音）と非音声状態（無音＋雑音）の尤度比を算出することにより，音声／非音声識別を行う．また，時間順方向の前向き推定のみではなく，時間逆方向の後向き推定を導入することにより，さらなる改善が得られることを示す．

キーワード 音声区間検出，非定常雑音，状態遷移過程，並列非線形カルマンフィルタ／スムーザ，前向き後向き推定

A noise robust voice activity detection with state transition processes of speech and noise

Masakiyo FUJIMOTO†, Kentaro ISHIZUKA†, and Hiroko KATO†

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

Abstract This paper proposes a noise robust voice activity detection with state transition processes of speech and noise. The proposed method constructs a clean speech / silence state transition model beforehand, and sequentially adapts the model to noise environment by using a parallel non-linear Kalman filtering when the observed signal is given. Speech / non-speech discrimination is carried out by calculating the likelihood ratio of a speech (clean speech + noise) state to a non-speech (silence + noise) state with the adapted model. In addition, a backward techniques, i.e., a parallel Kalman smoother and a backward probability estimation, are used to estimate the noise and for the likelihood ratio calculation.

Key words voice activity detection, non-stationary noise, state transition process, parallel non-linear Kalman filter / smoother, forward-backward estimation

1. まえがき

連続して観測される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，雑音に対して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．だが，これらの特徴量を用いたとしても，信号対雑音比 (SNR: Signal to Noise Ratio) が低い環境下では，雑音の持つ大きなエネルギーにより音声の特徴が埋もれてしまい，音声／非音声の特徴の差が曖昧になることは避けられない．結果として，音声／非音声の識別性能が劣化し，VAD の性能もまた劣化する．この問題は，特徴抽出機構のみでは雑音環境下での十分な VAD 性能を得ることが難しいことを示唆している．ここで，特徴量自体の識別性能が曖昧なものとなったとしても，その後の音声／非音声識別機構が頑健なものであれば，VAD の性能改善が期待できる．以上から，本研究では VAD の音声／非音声識別機構について検討を行う．

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識別法がSohnらにより提案されており，雑音環境下における高い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，観測信号がそれぞれの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．しかし，Sohnらの手法では，音声状態と非音声状態の遷移モデルを定義しているものの実体は無く，実質的には観測信号の事前及び，事後のSNR[6]を用いて擬似的に尤度比を推定して音声／非音声識別を行っている．このような識別方法では，VADの性能がSNRの推定精度に大きく左右されるという問題がある．また，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をおいており，このような条件設定では，非定常的な雑音や雑音環境の変化に対応することができない．この雑音の問題に関して我々は以前，並列非線形カルマンフィルタを用いて雑音を逐次推定することにより，雑音が未知かつ，非定常的な場合でも頑健に動作するVADを提案した[7]．しかし我々の前手法においても尤度比の推定は，事前及び，事後のSNRを用いて行っており，SNRの推定誤りが問題として残っていた．この尤度比推定の問題において，SNRに基づいて擬似的な尤度比推定を行うのではなく，確率密度関数（確率分布）を用いて尤度比推定を行えば，音声信号の歪み等が生じてもそれらをある程度吸収できるため，柔軟な識別が行うことができると考えられる．よって，本研究では，音声と非音声の確率分布を明確に定義して，直接的に尤度比を推定する手法を提案する．

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混合正規分布モデル(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．雑音が重畳した観測信号が与えられると，並列非線形カルマンフィルタにより雑音を推定し，クリーン音声と無音の状態遷移モデルを雑音環境に適応させる．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的に更新する．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状態遷移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，音声信号の歪み，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．さらに雑音推定及び，尤度比推定を時間順方向の前向き推定のみで行うのではなく，時間逆方向の後向き推定を含めて行うことにより，VADの性能がより改善されることを示す．

2. 確率モデルに基づくVADと状態遷移モデル

2.1 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．

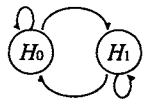


図1　音声／非音声状態遷移モデル（H_0: 非音声状態，H_1: 音声状態）

図2　雑音の状態遷移モデル

Sohnらの手法は効果的な手法として知られている．しかし，図1の状態遷移モデルが仮定されているものの実体は無く，実際には確率密度関数（確率分布）を用いて直接的に尤度比を計算するのではなく，事前及び，事後のSNR[6]を推定して間接的に尤度比を計算している．しかし，このようなSNRに基づく手法では，VAD性能がSNRの推定精度に大きく依存し，僅かな推定誤りであっても重大なVAD性能の劣化を招く可能性がある．一方，確率分布を用いた尤度計算はSNRに基づく手法に比べて柔軟であり，ある程度の信号の歪み等を吸収することができる上，分布を構成する特徴量を自由に選択できる．これらの利点をふまえて，本研究では，GMMを用いて音声／非音声状態の実質的な確率分布を定義し，尤度比を算出する．

また，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GMM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GMM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GMM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．

2.2 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(注1)．次に，音声はShon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(注2)．

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分布であるGMMを事前に学習しておくことが可能である．なお，

(注1)：定常雑音は非定常雑音の特殊なケースであり，非定常雑音は定常雑音を包含するという仮定のもと，雑音の定常性，非定常性に関わらず逐次推定を行う．
(注2)：雑音が既知で学習データが与えられるならば，モデル構造を定義して学習することにより，雑音の状態遷移モデルも図1のような離散的，Ergodicなモデルとして表現することが可能である．

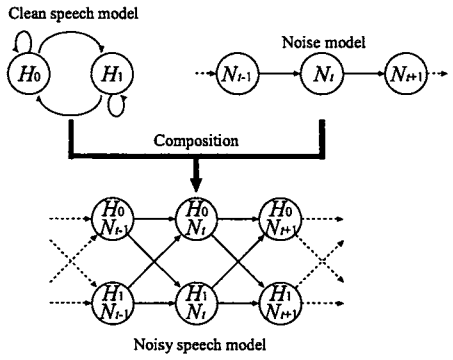


図 3　雑音の時間変化を考慮した音声／非音声状態遷移モデル

状態遷移確率には任意の値を与えるものとする．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，後述する並列非線形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，並列非線形カルマンフィルタの枠組で解決できることを後に示す．

2.3　状態遷移モデルの定式化と尤度比の算出

まず，図 1 の状態遷移モデルを用いた，雑音が存在を考慮しない場合における音声／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 と定義すると，$\mathbf{O}_t$ が音声／非音声状態のどちらに属するかを推定する問題は，$\mathbf{O}_{0:t}=\{\mathbf{O}_0,\cdots,\mathbf{O}_t\}$ が与えられたときの状態 q_t を求めることに相当する．つまり，次式の確率 $p(q_t|\mathbf{O}_{0:t})$ に基づき q_t を決定する．

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

$p(\mathbf{O}_{0:t},q_t)$ は，1 次マルコフ連鎖に基づき式 (2) の再帰式により表現され，時間に対して順方向で得られる前向き確率 $\alpha_{j,t}=p(\mathbf{O}_{0:t},q_t=H_j)$ に相当する．

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

上式において，$p(q_t=H_j|q_{t-1}=H_i)$ と $p(\mathbf{O}_t|q_t=H_j)$ はそれぞれ音声／非音声の状態遷移確率，各状態における出力確率であり，$p(q_t=H_j|q_{t-1}=H_i)=a_{i,j}$，$p(\mathbf{O}_t|q_t=H_j)=b_j(\mathbf{O}_t)$ と定義する．これらの定義から，式 (2) は次式で表現される．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．

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

それぞれの状態における $\alpha_{j,t}$ の比 $R_t=\alpha_{1,t}/\alpha_{0,t}$ を次式で閾値処理することにより，時刻 t の状態を識別する [5]．

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

次に，図 3 の状態遷移モデルを用いて，雑音の影響を考慮する．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，$\mathbf{O}_{0:t}$，$\mathbf{N}_{0:t}=\{\mathbf{N}_0,\cdots,\mathbf{N}_t\}$ 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．

$$\begin{aligned}p(q_t|\mathbf{O}_{0:t},\mathbf{N}_{0:t}) &= p(\mathbf{O}_{0:t},q_t,\mathbf{N}_{0:t})/p(\mathbf{O}_{0:t},\mathbf{N}_{0:t})\\ &\propto p(\mathbf{O}_{0:t},q_t,\mathbf{N}_{0:t})\end{aligned} \tag{5}$$

確率 $p(\mathbf{O}_{0:t},q_t,\mathbf{N}_{0:t})$ の再帰表現は次式のようになり，

$$\begin{aligned}p(\mathbf{O}_{0:t},q_t,\mathbf{N}_{0:t}) = \sum_{q_{t-1}}&p(q_t,\mathbf{N}_t|q_{t-1},\mathbf{N}_{t-1})\,p(\mathbf{O}_t|q_t,\mathbf{N}_t)\\ &\times p(\mathbf{O}_{0:t-1},q_{t-1},\mathbf{N}_{0:t-1})\end{aligned} \tag{6}$$

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，

$$\begin{aligned}p(\mathbf{O}_{0:t},q_t,\mathbf{N}_{0:t}) = \sum_{q_{t-1}}&p(q_t|q_{t-1})\,p(\mathbf{N}_t|\mathbf{N}_{t-1})\,p(\mathbf{O}_t|q_t,\mathbf{N}_t)\\ &\times p(\mathbf{O}_{0:t-1},q_{t-1},\mathbf{N}_{0:t-1})\end{aligned} \tag{7}$$

と表現される．また，雑音の状態遷移確率と各状態における出力確率をそれぞれ，$p(\mathbf{N}_t|\mathbf{N}_{t-1})=c_{t,t-1}$，$p(\mathbf{O}_t|q_t=H_j,\mathbf{N}_t)=b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ と定義し，前向き確率を $\alpha_{j,t}=p(\mathbf{O}_{0:t},q_t=H_j,\mathbf{N}_{0:t})$ と再定義する．これらの定義から，式 (7) は次式で表現される．

$$\alpha_{j,t}=\sum_{i=0}^{1}[a_{i,j}\alpha_{i,t-1}]\,b_{j,\mathbf{N}_t}(\mathbf{O}_t)\,c_{t,t-1} \tag{8}$$

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(注3)．よって，上式は次式のように簡略化される．

$$\alpha_{j,t}=\sum_{i=0}^{1}[a_{i,j}\alpha_{i,t-1}]\,b_{j,\mathbf{N}_t}(\mathbf{O}_t) \tag{9}$$

一方，図 2 に示された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起式で表現され，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 [8]．

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

ここで，上式に確率変数 q_t を加えると，式 (7) と一致する．すなわち式 (7) 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」という点とも符合する．

3.　カルマンフィルタによる雑音推定と尤度計算

3.1　並列非線形カルマンフィルタ

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

(注3)：雑音にも自己ループを許容することは可能であり，その場合は自己ループと状態変化のそれぞれに状態遷移確率を割り当てる必要がある．

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．

$$N_{t+1,l} = N_{t,l} + W_{t,l} \tag{11}$$

$$W_{t,l} \sim \mathcal{N}\left(0, \sigma^2_{W_l}\right) \tag{12}$$

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．

$$\begin{aligned} O_{t,l} &= S_{t,l} + \log\left(1 + \exp\left(N_{t,l} - S_{t,l}\right)\right) \\ &= f\left(S_{t,l}, N_{t,l}\right) \end{aligned} \tag{13}$$

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 (13) 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．

$$O_{t,l} = f\left(\mu_{S_{j,k,l}}, N_{t,l}\right) + V_{t,j,k,l} \tag{14}$$

$$V_{t,j,k,l} \sim \mathcal{N}\left(0, \sigma^2_{S_{j,k,l}}\right) \tag{15}$$

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 (11)，(14) 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，後述する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次式により与えられる．

$$N_{t|t-1,j,k,l} = \hat{N}_{t-1,l} \tag{16}$$

$$\sigma^2_{N_{t|t-1,j,k,l}} = \hat{\sigma}^2_{N_{t-1,l}} + \sigma^2_{W_l} \tag{17}$$

$$\mu_{O_{t|t-1,j,k,l}} = f\left(\mu_{S_{j,k,l}}, N_{t|t-1,j,k,l}\right) \tag{18}$$

$$\begin{aligned} &\sigma^2_{O_{t|t-1,j,k,l}} \\ &\quad = F_{t|t-1,j,k,l}\sigma^2_{N_{t|t-1,j,k,l}}F_{t|t-1,j,k,l} + \sigma^2_{S_{j,k,l}} \end{aligned} \tag{19}$$

$$F_{t|t-1,j,k,l} = \partial\mu_{O_{t|t-1,j,k,l}} / \partial N_{t|t-1,j,k,l} \tag{20}$$

$$G_{t,j,k,l} = \sigma^2_{N_{t|t-1,j,k,l}}F_{t|t-1,j,k,l} \Big/ \sigma^2_{O_{t|t-1,j,k,l}} \tag{21}$$

$$N_{t,j,k,l} = N_{t|t-1,j,k,l} + G_{t,j,k,l}\left(O_{t,l} - \mu_{O_{t|t-1,j,k,l}}\right) \tag{22}$$

$$\sigma^2_{N_{t,j,k,l}} = \left(1 - G_{t,j,k,l}F_{t|t-1,j,k,l}\right)\sigma^2_{N_{t|t-1,j,k,l}} \tag{23}$$

$$\mu_{O_{t,j,k,l}} = f\left(\mu_{S_{j,k,l}}, N_{t,j,k,l}\right) \tag{24}$$

$$\sigma^2_{O_{t,j,k,l}} = F_{t,j,k,l}\sigma^2_{N_{t,j,k,l}}F_{t,j,k,l} + \sigma^2_{S_{j,k,l}} \tag{25}$$

$$F_{t,j,k,l} = \partial\mu_{O_{t,j,k,l}} / \partial N_{t,j,k,l} \tag{26}$$

上式において，$t|t-1$ は時刻 $t-1$ の値からの予測値を示しており，$N_{t,j,k,l}$ と $\sigma^2_{N_{t,j,k,l}}$ はそれぞれ GMM j，分布 k のパラメータを用いて構成された非線形カルマンフィルタによる雑音の推定値及び，誤差分散であり，$\hat{N}_{t-1,l}$ と $\hat{\sigma}^2_{N_{t-1,l}}$ はそれぞれ時刻 $t-1$ で得られた最終的な雑音の推定結果である．また，$\mu_{O_{t|t-1,j,k,l}}$ と $\sigma^2_{O_{t|t-1,j,k,l}}$ はそれぞれ時刻 $t-1$ の値から得られる観測信号 $O_{t,l}$ の平均と分散の予測値，$\mu_{O_{t,j,k,l}}$ と $\sigma^2_{O_{t,j,k,l}}$ はそれぞれ $O_{t,l}$ の平均と分散の推定値である．

次に，各非線形カルマンフィルタから得られた K 種類の推定結果を以下のように重み付け平均する．

$$N_{t,j,l} = \sum_{k=1}^{K} w_{N_{t,j,k}} \cdot N_{t,j,k,l} \tag{27}$$

$$\sigma^2_{N_{t,j,l}} = \sum_{k=1}^{K} w_{N_{t,j,k}} \cdot \sigma^2_{N_{t,j,k,l}} \tag{28}$$

$$w_{N_{t,j,k}} = \frac{w_{S_{j,k}}\mathcal{N}\left(\mathbf{O}_t; \mu_{O_{t,j,k}}, \mathbf{\Sigma}_{O_{t,j,k}}\right)}{\sum_{k'=1}^{K} w_{S_{j,k'}}\mathcal{N}\left(\mathbf{O}_t; \mu_{O_{t,j,k'}}, \mathbf{\Sigma}_{O_{t,j,k'}}\right)} \tag{29}$$

上式において，$w_{N_{t,j,k}}$ は各推定結果に対する重み，$N_{t,j,l}$ と $\sigma^2_{N_{t,j,l}}$ は音声及び，無音 GMM それぞれに対応する，重み付け平均の結果である．また，$w_{S_{j,k}}$ 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，$\mathbf{\Sigma}_{O_{t,j,k}}$ は $\sigma^2_{O_{t,j,k,l}}$ を対角要素に持つ行列である．すなわち，確率 $\sum_{k=1}^{K} w_{j,k}\mathcal{N}\left(\mathbf{O}_t; \mu_{O_{t,j,k}}, \mathbf{\Sigma}_{O_{t,j,k}}\right)$ は非音声状態 ($j=0$: 無音＋雑音) 及び，音声状態 ($j=1$: 音声＋雑音) における出力確率 $b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ に相当する．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．

各 GMM から重み付け平均 $N_{t,j,l}$ が得られるが，それらを次式の様に正規化した $b_{j,\mathbf{N}_t}(\mathbf{O}_t)$ で重み付け平均することにより，時刻 t における最終的な推定値を得，次時刻の推定に用いる．

$$\hat{N}_{t,l} = \sum_{j=0}^{1} \frac{b_{j,\mathbf{N}_t}(\mathbf{O}_t)}{\sum_{j'=0}^{1} b_{j',\mathbf{N}_t}(\mathbf{O}_t)} N_{t,j,l} \tag{30}$$

$$\hat{\sigma}^2_{N_{t,l}} = \sum_{j=0}^{1} \frac{b_{j,\mathbf{N}_t}(\mathbf{O}_t)}{\sum_{j'=0}^{1} b_{j',\mathbf{N}_t}(\mathbf{O}_t)} \sigma^2_{N_{t,j,l}} \tag{31}$$

3.2 後向き確率の導入

式 (7) の状態遷移過程は，過去と現在の時刻 $0, \cdots, t$ の影響のみを考慮していたが，本研究では，次式のように未来の時刻 $t+1, \cdots, T$ の影響も考慮することで，より正確な尤度比計算及び，雑音推定を実現する．

$$p(\mathbf{O}_{0:T}, q_t, \mathbf{N}_{0:T}) = p(\mathbf{O}_{0:t}, q_t, \mathbf{N}_{0:t})\, p(\mathbf{O}_{t+1:T}, \mathbf{N}_{t+1:T}|q_t, \mathbf{N}_t) \quad (32)$$

確率 $p(\mathbf{O}_{t+1:T}, \mathbf{N}_{t+1:T}|q_t, \mathbf{N}_t)$ の再帰式は次式で与えられ，

$$p(\mathbf{O}_{t+1:T}, \mathbf{N}_{t+1:T}|q_t, \mathbf{N}_t) = \sum_{q_{t+1}} p(q_{t+1}|q_t)\, p(\mathbf{N}_{t+1}|\mathbf{N}_t)\, p(\mathbf{O}_{t+1}|q_{t+1}, \mathbf{N}_{t+1}) \times p(\mathbf{O}_{t+2:T}, \mathbf{N}_{t+2:T}|q_{t+1}, \mathbf{N}_{t+1}) \quad (33)$$

後向き確率 $\beta_{j,t} = p(\mathbf{O}_{t+1:T}, \mathbf{N}_{t+1:T}|q_t = H_j, \mathbf{N}_t)$ に相当する．よって，式 (33) は式 (8) の表現及び，$p(\mathbf{N}_{t+1}|\mathbf{N}_t)p = c_{t+1,t} = 1$ より，次式のように表現される．なお，時刻 $t = T$ の場合は初期値 $\beta_{0,T} = \beta_{1,T} = 1$ を与える．

$$\beta_{i,t} = \sum_{j=0}^{1} a_{i,j} b_{j,\mathbf{N}_t}(\mathbf{O}_{t+1})\, \beta_{j,t+1} \quad (34)$$

従って，$p(\mathbf{O}_{0:T}, q_t = H_j, \mathbf{N}_{0:T}) = \alpha_{j,t} \cdot \beta_{j,t}$ となり，前向き後向き推定により，尤度比 R_t は次式で与えられる．

$$R_t = \frac{p(\mathbf{O}_{0:T}, q_t = H_1, \mathbf{N}_{0:T})}{p(\mathbf{O}_{0:T}, q_t = H_0, \mathbf{N}_{0:T})} = \frac{\alpha_{1,t} \cdot \beta_{1,t}}{\alpha_{0,t} \cdot \beta_{0,t}} \quad (35)$$

加えて，カルマンフィルタの後向き推定（平滑化）手法である，カルマンスムーザ [9] を用いて，雑音推定の精度を向上させる．ここで，カルマンスムーザにおいても，前節の並列非線形カルマンフィルタと同様に複数のカルマンスムーザを実行し，得られた複数の推定結果に対して重み付け平均を行う．この手法を本研究では，並列カルマンズムーザと呼ぶ．並列カルマンスムーザによる後向き推定は次式のように行う．

$$J_{t,j,k,l} = \sigma^2_{N_{t,j,k,l}} \Big/ \sigma^2_{N_{t+1|t,j,k,l}} \quad (36)$$

$$\tilde{N}_{t,j,k,l} = N_{t,j,k,l} + J_{t,j,k,l}\left(\tilde{N}_{t+1,j,k,l} - N_{t+1|t,j,k,l}\right) \quad (37)$$

$$\tilde{\sigma}^2_{N_{t,j,k,l}} = \sigma^2_{N_{t,j,k,l}} + J_{t,j,k,l}\left(\tilde{\sigma}^2_{N_{t+1,j,k,l}} - \sigma^2_{N_{t+1|t,j,k,l}}\right) J_{t,j,k,l} \quad (38)$$

上式において，$\tilde{N}_{t,j,k,l}$ と $\tilde{\sigma}^2_{N_{t,j,k,l}}$ はそれぞれ並列カルマンスムーザにより得られる雑音の平滑化推定値及び，平滑化誤差分散である．また，それぞれの平滑値は，式 (27)〜(31) と同様の方法により，重み付け平均を行う．

一般に，後向き推定はデータの終端から推定を行うが，VAD では終端が未知であるので，現在の時刻 t から tb フレーム未来の時刻，すなわち時刻 $T = t + tb$ から遡って推定を行う．また，$tb = 0$ の場合は後向き推定を行わないことを意味する．

4. 実　　験

4.1 実験条件

評価データは，ATR 旅行会話データベース 2,292 発話（178 話者）を 8 kHz にダウンサンプリングした後，空港ロビー，及び街頭で収録した雑音をそれぞれ SNR 0, 5, 10 dB で加算して作成した．音響分析は，フレーム長 20 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，$a_{i,j} = \{0.8, 0.2, 0.1, 0.9\}$ とし，駆動雑音の分散は $\sigma^2_{W_l} = 0.0001$ とした．また，無音及び，音声 GMM の学習データは，5,050 発話の ATR 音素バランス文 (101 話者) であり，GMM の混合分布数はそれぞれ 32 である．

評価は，人手により作成した音声始終端の時間ラベルとフレーム単位で照合することにより行った．評価尺度は，式 (39)，(40) に示す，FAR (False Acceptance Rate) と FRR (False Rejection Rate) である．FAR と FRR は，互いにトレードオフの関係にあるので尤度比判定の敷値 *Threshold* を調整して複数の実験結果を得，ROC (Receiver Operating Characteristics) 曲線を描くことにより評価を行う．

$$FAR = \frac{\text{非音声区間を音声区間と誤識別したフレーム数}}{\text{正解非音声フレーム数}} \quad (39)$$

$$FRR = \frac{\text{音声区間を非音声区間と誤識別したフレーム数}}{\text{正解音声フレーム数}} \quad (40)$$

4.2 実験結果

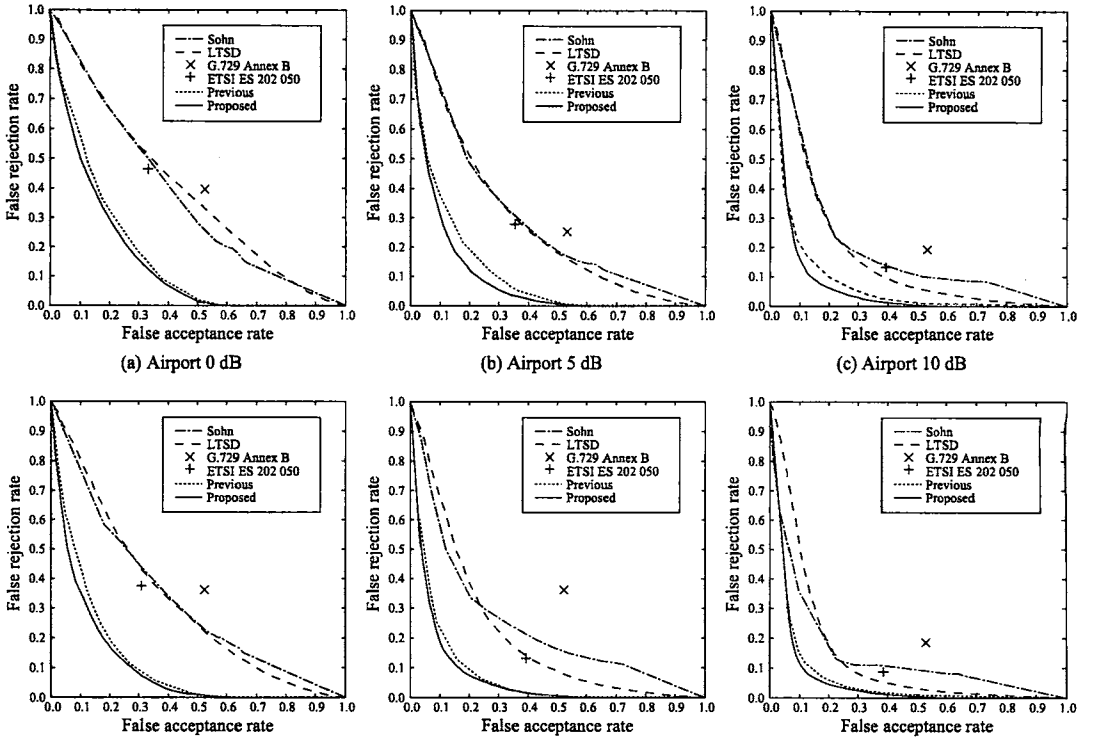
図 4 に，提案手法 (Proposed) 及び，従来手法である Sohn らの手法 [5]，LTSD (Long-Term Spectral Divergence) [3]，ITU-T G.729 Annex B [10]，ETSI ES 202 050 [11]，我々の旧手法 (Previous) [7] の結果を示す．なお，文献 [10], [11] の手法はパラメータが固定されているため一点の結果のみを示しており，提案手法及び，旧手法の結果では，後向き推定を行わない場合 ($tb = 0$) の結果を示している．

ROC 曲線は，原点に近い曲線ほど高い性能を示しており，図 4 の結果より，全ての雑音環境において提案法が最良の結果を示していることがわかる．提案手法及び，我々の旧手法と Shon らの手法との相違点は雑音の推定問題であり，Shon らの手法では雑音の逐次推定を行っていないのに対し，提案手法及び，旧手法では非線形カルマンフィルタによる逐次推定を行っている．この結果より，非線形カルマンフィルタによる雑音の逐次推定が VAD 性能の改善において極めて重要な役割を果たしていることがわかる．また，提案法と旧手法の相違点は，尤度比の計算を確率分布により直接的に行うか，SNR により間接的に行うかである．これについても，提案手法は旧手法よりも高い性能を示しており，当初の狙い通り確率分布を用いた頑健な尤度比計算，音声／非音声識別が行われていることが確認できる．

次に，提案法において後向き推定を行った場合の結果を表 1 に示す．この際，後向き推定に要するフレーム数は $tb = 0, 5, 10, 15, 20$ であり，評価尺度には等誤り率（FAR と FRR が同値となる誤り率）を用いた．表 1 の結果より，後向き推定を行うことにより性能改善が得られることがわかる．全ての雑音環境において，$tb = 10$ の時に等誤り率が最良となり，それ以後は却って劣化することが分かる．このことから，あまりに遠い未来の情報を用いると，現在の時刻の情報に対して却って悪影響を与えることが確認できる．よって，後向き推定を行う場合には，現在の時刻から遠く離れた時刻の情報を用いないか，忘却係数等を用いて影響を弱めることが必要である．

5. むすび

本研究では，音声と雑音両方の状態遷移過程を有する状態遷



(a) Airport 0 dB　(b) Airport 5 dB　(c) Airport 10 dB

(d) Street 0 dB　(e) Street 5 dB　(f) Street 10 dB

図 4　ROC 曲線による実験結果

表 1　後向き推定を用いた場合の等誤り率 (%)

Noise	Airport			Street		
SNR	0 dB	5 dB	10 dB	0 dB	5 dB	10 dB
$tb = 0$	23.82	16.51	12.12	18.82	13.30	10.64
$tb = 5$	23.11	15.99	11.94	18.66	12.97	10.50
$tb = 10$	**22.75**	**15.62**	**11.58**	**18.28**	**12.89**	**10.49**
$tb = 15$	22.90	15.76	11.79	18.42	13.11	10.68
$tb = 20$	22.91	16.25	12.32	19.00	13.66	11.11

移モデルに基づく VAD を提案し，従来法に比べて大幅な性能改善が得られることを示した．また，提案手法が，並列非線形カルマンフィルタ／スムーザの枠組で全て構成できることを明らかにし，Switching カルマンフィルタとの関連性についても述べた．今後，様々な特徴抽出法との組み合わせや，様々なパラメータの適応的決定法について検討する予定である．

文　献

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

[2] Nemer, E., Goubran, R., and Mahmoud, S., "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Trans. on Speech and Audio Processing*, Vol. 9, No. 3, pp. 217–231, March 2001.

[3] Ramirez, J., Segura, J.C., Benitex, C., de la Torre, A., and Rubio, A., "Efficient voice activity detection algorithm using long-term speech information," *Speech Communication*, Vol. 42, pp. 271–287, Apr. 2004.

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

[5] Sohn, J. , Kim, N. S., and Sung, W., "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, Vol. 6, No. 1, pp. 1–3, Jan. 1999.

[6] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *Trans. on Acoust., Speech, Signal Processing*, Vol. ASSP–32, pp. 1109–1121, Dec. 1984.

[7] 藤本 雅清, 石塚健太郎, 加藤 比呂子, "音声／非音声状態遷移モデルに基づく音声区間検出," 日本音響学会, 平成 18 年度春秋研究発表会, 1–2–17, pp. 33–34, Sept. 2006.

[8] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking," *IEEE Trans. SP*, Vol. 50, No. 2, pp. 174–188, Feb. 2002.

[9] Balakrishnan, A.V., "Kalman Filtering Theory," *Optimization Software*, 1987.

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

[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.