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

2008－SLP－74（33）
2008 / 12 / 10

# 音声の動的特徴のモデルを使った突発性雑音の除去

三宅 信之† 滝口 哲也† 有木 康雄†

† 神戸大学大学院工学研究科 〒657-8501 神戸市灘区六甲台町 1-1
E-mail: †miyake@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

**あらまし** 本稿では静的特徴と動的特徴の両方のモデルを用いた突発性雑音の除去法について述べる．突発性の雑音は出現する時が不明確で，また短時間しか続かないことも多いため推定しにくいが，音声に混入することで音声認識率が下がることは珍しくない．これまでに我々は，このような突発性の雑音に対し，フレームごとに雑音の検出を行い，雑音が重畳していると判定されたフレームはどのような雑音が重畳しているのか識別し，識別結果を元に静的特徴のモデルを使って雑音除去を適用した．しかしながら，雑音があるにも関わらず，雑音がないと判定されたフレームに対しては除去が行われないという問題があった．本稿ではその問題を解決するために，静的特徴の推定と併用して動的特徴のモデルと直前のフレームの推定値からの推定も行うことで，雑音未検出フレーム対してもある程度の補正をかけ音声認識率の向上を図る．また今まで雑音除去をかけてきたフレームに対しても，動的特徴も併用することで，より効果的な音声強調ができると考えられる．実験結果より，動的特徴のモデルも併用することで，静的特徴のみを利用した場合よりも認識率が高くなることが確認できた．

**キーワード** 突発性雑音，動的特徴量モデル，雑音除去，音声認識

# Sudden noise reduction using dynamic speech feature model

Nobuyuki MIYAKE†, Tetsuya TAKIGUCHI†, and Yasuo ARIKI†

† Graduate School of Engineering, Kobe University
1-1, Rokkodai-cho, Nada-ku, Kobe-shi, Hyogo, 657-8501 Japan
E-mail: †miyake@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

**Abstract** This paper describes a method for reducing sudden noise using a static and dynamic speech feature model. We have propoesed a method for reducing these noises using static feature model after the noisy frame is detected and what noise overlapped is classfied. But, the previous method has a problem that the noise reduction is not peroformed if the frame is not determined as noisy one although the frame has a noise. In this paper, we propose a noise reduction method using a dynamic feature model and last frame data in order to deal with such noisy frames. In our experiments, the proposed method achieved better performance for recognition of utterances overlapped by sudden noises than the previous method.

**Key words** model-based noise reduction, dynamic feature model, sudden noise, speech recognition

## 1. まえがき

音声認識技術を使用するとき，発話に雑音が重畳することで誤認識を引き起こすことが少なくない．そのためスペクトルサブトラクションを始めとした雑音を除去する研究が数多くなされている [1]．雑音の除去法には音声の特徴を混合ガウス分布 (GMM) といったモデルで保持しておき，その情報を利用して除去する手法 [2]～[4]，マイクロホンアレーを利用して雑音を除去するといった手法などが多く見られる [5], [6]．シングルチャネルでの雑音の推定には発話直前の雑音のみの区間や，その情報を元に追跡していくものなどが用いられる．雑音は時間的に緩やかに変化するものだと考えると，発話付近の雑音の情報を使用することで雑音抑圧は高い効果が得られると期待できる．

しかしながら，家の中のような実環境で音声認識を使用することを考えるとき，雑音にはドアの開閉音や電話の音など中には突然発生するものも少なくない．図 1 は発話区間のみに電話音が重畳しているが，このような雑音が発話中に発生した場合，発話していない区間から雑音を推定し，除去することは困難である．

突発的に発生する雑音の除去に関する研究はいくつか存在す

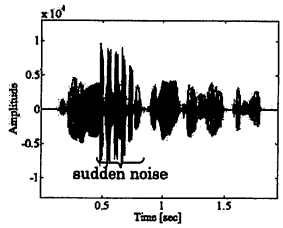


図 1　突発性雑音の例
Fig. 1　An example of sudden noise

る [7]〜[9]．我々も以前にこのような突発性の雑音に対する除去法の提案を行った [10]．これは，フレームごとに突発性の雑音があるかどうかの判定を行い，重畳している雑音の識別を行い，その結果を用いて雑音の SNR を推定しながら雑音除去するものであった．しかしながら，この手法では雑音除去に静的特徴量のモデルのみを使っており，たとえ雑音が存在しても，雑音があると判定されなかったフレームはそのまま認識に使用され認識に悪影響を及ぼすという問題点があった．

本稿ではその問題点を解決するため，静的特徴のモデルだけではなく，動的特徴のモデルと前フレームも考慮し推定を行う手法 [3] を使い，今までは雑音がないと判定され，除去が行われなかったフレームに対してもある程度補正をかけることを試みる．

雑音は識別結果からは SNR はわからないため，[3] で述べられている手法に，雑音の強さを表す定数を用意し，EM アルゴリズムによる値の推定を組み込み使用する．

また，この手法は雑音の未検出フレームに効果的なだけでなく動的特徴も併用することで使える情報が増えるため，雑音除去の効果もより高くなることが期待できる．

## 2. 雑音のクラスタリング

雑音には様々なものがあるが，本稿では RWCP 非音声データベース [11] のすべての雑音を取り扱うものとする．このデータベースには 105 種類の雑音がある．このままでは種類が非常に多いため，これらを識別しようと考えたとき，後に述べる識別器の学習や識別そのものに非常に時間がかかる．そのためあらかじめクラスタリングする．しかし多くの雑音をひとつのクラスにまとめてしまうと，除去時は雑音のデータとしてそのクラスの平均ベクトルを使用するために，雑音の特徴を捉えられず除去がうまく働かないと考えられる．そこで，上段では荒く，下段では細かくクラスタリングされたツリーを構成する．

### 2.1 クラスタリング手法

クラスタリング手法として k-means 法を使用する．k-means 法はクラス数は手動で与える必要がある．本稿ではツリーを構成するときに各ノードでどのような値を設定すればいいかわからないため，それぞれのデータからクラスの中心までの距離の最大値を与えることで，クラス数を自動的に決定しながらクラスタリングを行う．

まずクラスの中心から距離の最大値 $d_{max}$ を手動で決める．そして，k-means 法を用いてクラスタリングする．その後，データとデータが属すクラスの中心との距離 $d$ を測り，距離が指定した値以上 ($d > d_{max}$) ならばそのデータが属すクラスを 2 つに分割する．そして再び k-means 法を行う．すべてのデータとクラスの中心の距離が $d_{max}$ を下回るまでこれを繰り返す．

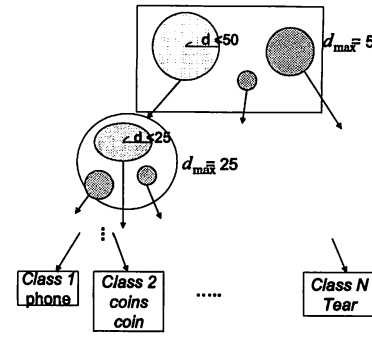


図 2　雑音の種類によるツリーの例
Fig. 2　An example of tree of noises

### 2.2 ツ リ ー

上記の手法をそのまま適用すると，距離の最大値 $d_{max}$ の値を小さくしたときにクラス数が大きくなりすぎる．そこで上記のクラスタリング手法を用いて図 2 のようなツリーを形成する．

上段ではこの中心までの距離の最大値を大きくすることで荒くクラスタリングを行う．下段では，上段で分けられたクラスをより小さな $d_{max}$ を設定することでより細かく分ける．$d_{max}$ の値は何段目かによって決めておく．こうすることで，上段では雑音が荒くクラスタリングされ，下段では細かく分けられているツリーが形成される．

## 3. 雑音の検出と識別

本稿で扱う雑音はそのほとんどが短時間しか継続せず，またいつ起こるかわからないものである．そのためまず除去を行う前に雑音が重畳しているかどうか判定を行い，またそれがどのような雑音であるか識別する [10]．

### 3.1 雑音の検出

雑音の検出には AdaBoost を用いる．AdaBoost は Boosting の一種であり，多数の弱識別器を使うことで，非線形な識別器を作成することができる [12]．

学習する雑音データの雑音重畳音声を作成し，それらすべてとクリーン音声を用いて AdaBoost で，以下の識別器 $f(\mathrm{x})$ の学習を行う．

$$f(\mathbf{x}) = \frac{1}{||\beta||} \sum_t \beta_t h_t(\mathbf{x}) \qquad (1)$$

ここで，x は入力する特徴量，$h(\mathrm{x})$ は弱識別器であり $h(\mathrm{x}) = \{-1, 1\}$，$\beta_t$ は弱識別器の重み，$||\beta||$ は重みの正規化項になっている．弱識別器には decision stump を使う．

雑音除去時は，この識別器を用いてフレーム単位で雑音の検出を行う．具体的には特徴量を入力したときの式 (1) の正負でクリーン音声か雑音が重畳しているかの判定を行う．そしてこの識別器によって雑音が重畳していると判定されたフレームに

**Input:** $m$ examples $\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_m, y_m)\}$

$$y_i = \{1, \cdots, K\}$$

**Do for** $k = 1, \cdots, K$

1. Set labels

$$y_i^k = \begin{cases} +1, & \text{if } y_i = k \\ -1, & \text{otherwise} \end{cases} \quad (2)$$

2. Learn $k$-th classifier $f^k(\mathbf{x})$ using AdaBoost for data set $Z^k = (\mathbf{x}_1, y_1^k), \cdots, (\mathbf{x}_m, y_m^k)$

**Final classifier:**

$$\hat{k} = \underset{k}{\operatorname{argmax}} f^k(\mathbf{x}) \quad (3)$$

図 3 one-vs-rest 法による AdaBoost のマルチクラス化

Fig. 3 one-vs-rest multi-class algorithm for AdaBoost

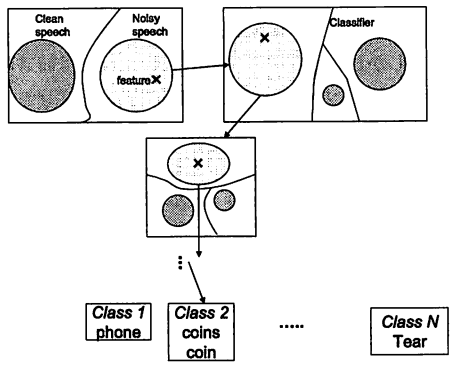


図 4 検出・識別の例

Fig. 4 An example of detection and classification

対して雑音の識別が行われる．

### 3.2 雑音の識別

検出において雑音と判定されたフレームに対し雑音の識別を行う．2.2 節で作成したツリーの各ノード毎に，AdaBoost を用いて識別器を作成し，上段から順に識別を行う．この識別器の作成は各クラスに属す雑音重畳音声を使用して行う．AdaBoost は二値判別の識別器しか作成できないため，実際には他クラスを識別するために図 3 のように one-vs-rest 法を用いて他クラス判別ができるように拡張している [13]．これは AdaBoost によって作られる識別器である式 (1) を 1 つのクラスとその他のクラスを分離できるように作成し，その中で $f(\mathrm{x})$ が最大になるものを識別結果とする手法である．この識別器を用いて入力されたフレームを上段から順に分類していく．図 4 は検出・識別の例である．識別された結果，最終的に重畳している雑音はフレームごとに 1 つのクラスに分類される．雑音除去はこのクラスに存在する雑音の平均ベクトルを使用して行われる．

## 4. 雑音重畳音声

フレーム $t$ における雑音重畳音声の $k$ 次元目のフーリエ変換特徴量 $X_t(k)$ は

$$X_t(k) = S_t(k) + N_t(k) \quad (4)$$

ただし，本稿では SNR は未知なので，それを表すパラメータ $\alpha$ を雑音 $N_t(k)$ に掛け合わせる．

$$X_t(k) = S_t(k) + \alpha \cdot N_t(k) \quad (5)$$

この時の $l$ 次元目のメルフィルタバンク特徴量 $M_x(l)$ は以下のように書き表せる．

$$\begin{aligned} M_x(l) &= \sum_k w_{l,k} |X_t(k)|^2 \\ &= \sum_k w_{l,k} \{ |S_t(k)|^2 + |\alpha N_t(k)|^2 \\ &\qquad + 2\alpha |S_t(k)||N_t(k)| \cos\theta_k \} \\ &= M_s(l) + \alpha^2 M_n(l) + \lambda(l) \end{aligned} \quad (6)$$

ここで，$w_{l,k}$ はフィルタ係数であり，$\theta_k$ は $X(k)$ と $N(k)$ が為す角である．この時，対数メルフィルタバンク $x_t(l) = \log M_x(l)$ は

$$\exp(x_t(l)) = \exp(s_t(l)) + \exp(n_t(l)) + \lambda(l)$$

$$\begin{aligned} x_t(l) &= s_t(l) + \log[1 + \alpha^2 \exp(n_t(l) - s_t(l))] + r(l) \\ &= s_t(l) + g_l(s_t, n_t, \alpha) + r(l) \end{aligned} \quad (7)$$

$$r(l) = \log\left[1 + \frac{2\alpha \sum_k w_{l,k} |S_t(k)||N_t(k)| \cos\theta_k}{M_s(l) + \alpha^2 M_n(l)}\right] \quad (8)$$

となる．ここで $r$ は位相に依存するため，ランダムな値と仮定できる．そこで，$r$ の確率分布を平均 0，分散 $\Phi_n$ の正規分布で仮定する．

$$p(r) = \mathcal{N}(r; 0, \Phi_n) \quad (9)$$

分散 $\Phi_n$ はあらかじめ学習データから求める．ノイズによって多少差が出ることが考えられるので，本稿では雑音のクラス毎に使う値を変えている．

## 5. 音声特徴量の推定

音声特徴量の推定を行うため，雑音除去を行う．本研究は主に突発的な雑音を対象としているため，3.2 節で述べたように，雑音除去はあらかじめ雑音をある程度判別してから行う．しかしながら，雑音区間が検出できない場合も少なからずあり，今まではその区間に除去をかけることができなかった．本稿では，静的特徴と動的特徴の両方を使った雑音除去法 [3] を，SNR を補正するパラメータ $\alpha$ を入れた形で展開しなおし，今までは除去がかけられなかったフレームに対しても雑音除去を行う．

まず学習データより音声信号の対数メルフィルタバンク特徴量とその動的特徴量の GMM を用意する．

$$p(s_t, \Delta s_t) = \sum_m c_m \mathcal{N}(s_t; \mu_m^s, \Sigma_m^s) \mathcal{N}(\Delta s_t; \mu_m^{\Delta s}, \Sigma_m^{\Delta s}) \quad (10)$$

ここでは，$\Delta s_t = s_t - s_{t-1}$ である．また $\mu_m^s, \mu_m^{\Delta s}$ は平均ベク

トルである．$\Sigma^s_m, \Sigma^{\Delta s}_m$ は分散共分散行列であり，本稿では対角成分のみを用いている．

音声の推定は以下のように条件付期待値として求める．

$$\hat{s}_t = \int s_t p(s_t|x_t, n_t, \alpha, s_{t-1}) ds_t \tag{11}$$

ここで，雑音 $n_t$ が出てくるが，これは 3.2 において識別されたクラスの平均値であり，雑音が未検出だった場合は $n_t = 0$ としている．この時 $p(s_t|x_t, n_t, \alpha, s_{t-1})$ を GMM とし，混合要素のインデックス $m$ で周辺化すると，式 (11) は次のように書ける．

$$\hat{s}_t \approx \sum_m p(m|x_t, n_t, \alpha) \int s_t p(s_t|x_t, n_t, \alpha, s_{t-1}, m) ds_t \tag{12}$$

この時，$p(m|x_t, n_t, \alpha)$ は以下のように近似を行う．

$$\begin{aligned} p(m|x_t, n_t, \alpha) &= \frac{c_m \mathcal{N}(x_t; \mu^x_m, \Sigma^x_m)}{\sum_m c_m \mathcal{N}(x_t; \mu^x_m, \Sigma^x_m)} \\ \mu^x_m &\approx \mu^s_m + g(\mu^s_m, n_t, \alpha) \\ \Sigma^x_m &\approx \Sigma^s_m \end{aligned} \tag{13}$$

式 (12) における積分は，$p(s_t|x_t, n_t, \alpha, s_{t-1}, m)$ を $s_t$ の正規分布と仮定すると，その平均が解となる．そこで $p(s_t|x_t, n_t, \alpha, s_{t-1}, m)$ の平均を考える．ベイズの定理より，

$$p(s_t|x_t, n_t, \alpha, s_{t-1}, m) \approx \frac{p(x_t|s_t, n_t, \alpha, m) p(s_t|s_{t-1}, m)}{p(x_t|n_t, \alpha, m)} \tag{14}$$

分母の $p(x_t|n_t, \alpha, m)$ は正規化項であり，正規分布の平均を求める上で考慮する必要は無い．

$s_t, n_t, \alpha$ が与えられた場合の $x_t$ 条件付確率は，式 (9) より，

$$p(x_t|s_t, n_t, \alpha, m) = \mathcal{N}(x_t; s_t + g(s_t, n_t, \alpha), \Phi_n) \tag{15}$$

となるが，この形では式 (12) を解くことはできないため，以下のような近似を行う．

$$g(s_t, n_t, \alpha) \approx g(\mu^s_m, n_t, \alpha) \tag{16}$$

$$p(x_t|s_t, n_t, \alpha, m) = \mathcal{N}(x_t; s_t + g(\mu^s_m, n_t, \alpha), \Phi_n) \tag{17}$$

また，$p(s_t|s_{t-1}, m)$ はベイズの定理で展開することにより，以下の式が導かれる．

$$\begin{aligned} p(s_t|s_{t-1}, m) &\propto p(s_t, s_{t-1}|m) \\ &\approx p(s_t, s_t - s_{t-1}|m) \\ &= \mathcal{N}(s_t; \mu^s_m, \Sigma^s_m) \mathcal{N}(\Delta s_t; \mu^{\Delta s}_m, \Sigma^{\Delta s}_m) \end{aligned} \tag{18}$$

この 2 つの正規分布の積を $s_t$ に対する 1 つの正規分布に書き直すと，

$$\begin{aligned} &p(s_t|s_{t-1}, m) \approx \mathcal{N}(s_t; \eta_m, \varphi_m) \\ &\eta_m = \frac{\Sigma^s_m}{\Sigma^s_m + \Sigma^{\Delta s}_m} \mu^s_m + \frac{\Sigma^s_m}{\Sigma^s_m + \Sigma^{\Delta s}_m} (\hat{s}_{t-1} + \mu^{\Delta s}_m) \\ &\varphi_m = \frac{\Sigma^s_m \Sigma^{\Delta s}_m}{\Sigma^s_m + \Sigma^{\Delta s}_m} \end{aligned} \tag{19}$$

同様に，式 (15,19) をひとつの正規分布で書き表すと，

$$\begin{aligned} p(s_t|x_t, n_t, \alpha, s_{t-1}, m) &\approx \mathcal{N}(s_t; \mu_m, \Sigma_m) \\ \mu_m &= w_1 \eta_m + w_2 [x_t - g(s_t, n_t, \alpha)] \\ \Sigma_m &= \frac{\Phi_n \varphi_m}{\Phi_n + \varphi_m} \\ w_1 &= \frac{\Phi_n}{\Phi_n + \varphi_m}, \; w_2 = I - w_1 \end{aligned} \tag{20}$$

となる．ただし，重み $w_1, w_2$ は一定の範囲に収まるように値を制限する．本稿では $0.4 < w_1 < 0.6$ とした．また雑音が検出されなかった場合 $\Phi_n = \varphi_m$ として計算しており，そのため重みは $w_1 = 0.5, w_2 = 0.5$ となる．式 (20) より，音声特徴の推定値は以下のようになる．

$$\hat{s}_t \approx \sum_m p(m|x_t, n_t, \alpha) \Big( w_1 \eta_m + w_2 [x_t - g(\mu^s_m, n_t, \alpha)] \Big) \tag{21}$$

雑音が未検出の場合，$n_t = 0$ として計算されるため，第 2 項は $x_t$ となる．しかしながら，第 1 項に含まれる $\eta_m$ は前フレームの推定値から考えた場合の推定値であるため，前フレームの推定が正しければ，$\eta_m$ の推定も実際の音声に近いものになると考えられる．その推定値と実際に観測されたフレームとの平均をとることで，たとえ雑音がないと判定されたフレームに雑音が重畳していても，ある程度雑音を抑圧できると考えられる．

## 6. EM アルゴリズムよる雑音の強さの推定

式 (21) を使って音声特徴量を推定するためには雑音の強さとして設定したパラメーター $\alpha$ の値が必要になる．そこで本稿ではこのパラメーターの値を EM アルゴリズムを使って求める．$p(s_t|x_t, n_t, \alpha, s_{t-1})$ を最大化するように雑音の強さを推定する．この時 $m$ を隠れ変数として EM アルゴリズムを使って最大化する．E-step として次のような Q 関数を設定する．

$$\begin{aligned} Q(\alpha^{(k)}, \overline{\alpha}) &= \sum_m p(s_t, m|x_t, n_t, \alpha, s_{t-1}) \log p(s_t, m|x_t, n_t, \alpha, s_{t-1}) \\ &= \sum_m p(m|x_t, n_t, \alpha^{(k)}) p(s_t|x_t, n_t, \alpha^{(k)}, s_{t-1}, m) \\ &\quad \cdot \log p(m|x_t, n_t, \overline{\alpha}) p(s_t|x_t, n_t, \overline{\alpha}, s_{t-1}, m) \end{aligned} \tag{22}$$

そして M-step でこの関数を最大化する．

$$\alpha^{(k+1)} = \arg\max_{\overline{\alpha}} Q(\alpha^{(k)}, \overline{\alpha}) \tag{23}$$

これらを収束するまで繰り返すことで最適な重み係数を求める．ここで，式 (23) を解くには関数 $Q$ を $\overline{\alpha}$ で微分すればよいが，式が複雑になるため解析的に解くのは困難である．そこで本稿ではニュートン法を用いて求めている．

また，式 (23) を解くためには $s_t$ の値が既知である必要がある．そこで $s_t = \hat{s}_t$ として $\alpha$ を求め，求めた $\alpha$ を式 (21) に代入することで $\hat{s}_t$ を更新する．すなわち，収束するまで以下の 2 式を交互に計算する．

表 1 実験条件

Table 1 Experimental condition

| | |
|---|---|
| **Making tree** | |
| Feature parameters | 24 – log Mel filter bank |
| Tree depth | 5 |
| Upper limit (in order of depth level) | 50, 25, 12, 6 |
| **Detection and Classification** | |
| Feature parameters | 24 – log Mel filter bank |
| The number of weak learners of AdaBoost | 200 |
| **Noise reduction** | |
| Feature parameters | 24 – log Mel filter bank |
| The number of components of GMM (experiment 3patterns) | 32, 64, 128 |
| **Speech recognition** | |
| Feature parameters | 12-MFCC + Δ + ΔΔ with CMN |
| Acoustic models | Phoneme HMM 5 states 12 mixtures |
| Lexicon | 500 words |

$$\alpha^{(k+1)} = \arg\max_{\overline{\alpha}} Q(\alpha^{(k)}, \overline{\alpha}) \tag{24}$$

$$\hat{s}_t^{(k+1)} \approx \sum_m p(m|x_t, n_t, \alpha^{(k)}) \cdot \left(w_1\eta_m + w_2\left[x_t - g(\mu_m^s, n_t, \alpha^{(k)})\right]\right) \tag{25}$$

## 7. 実　　験

### 7.1 実験条件

ATR の特定話者単語データベースを用いて実験を行った．男性話者 5 名，女性話者 5 名を用い，各話者 2,000 発話を学習に，500 発話をテストに使用した．雑音は RWCP 非音声ドライソースを使用した [11]．このデータベースには 105 種類の雑音が各種類 100 データ存在し，そのうち 50 データを学習用に残りの 50 データをテストデータとして利用した．このデータベースの雑音重畳時間は 20〜300 msec 程度になっている．この 105 種類の雑音のパワーをそろえ，24 次元の対数メルフィルタバンクを作成，種類ごとに平均ベクトルを算出し，クラスタリングを行った．クラスタリングの結果 105 種類の雑音は 37 個のクラスに分割された．検出・識別時の識別器に使用する学習データはこの 2 つのデータベースから各クラスの雑音重畳音声を作成し，対数メルフィルタバンクに変換したものを用いた．この時 SNR は-5〜5 dB になるように調整した．テストデータに対しては，発話ごとに SNR を調整して雑音を 1〜5 つ重畳させた．この時の SNR は 5 dB, 0 dB, -5 dB になっている．このテストデータに対して雑音除去を適用し，認識実験を行った．この時 EM アルゴリズム，ニュートン法の初期値は 0 としており，$\hat{s}_t^{(k)}$ の初期値は $\hat{s}_t^{(0)} = x_t$ とした．そのほかの実験条件は表 1 に示す．

### 7.2 実験結果

まず，フレーム単位の検出・識別の結果を表 2 に示す．結果を見てみると再現率が低く，正しく検出できていないフレームも多いことがわかる．また識別率もあまりよくない．この結果を利用して除去した場合の認識結果を図 5 に示す．ここで，Previous とは以前我々が使用していた手法 [10] であり，式 (21) における重みを $w_1 = 0, w_2 = 1$ とした場合と等価である．Proposed は本稿で我々が提案した手法である．また baseline は雑音除去を行わないときの認識率，Oracle Label とは検出・識別の結果を手動で与えた場合であり，これは 128 混合での実験結果である．結果を見ると，どの場合でも以前の手法に比べ，提案手法を使ったほうが認識率の大きな改善が得られることが分かる．混合ごとの違いを見ると，混合数が多いほうが高い認識率となった．また，検出・識別を行った場合と Oracle Label の認識率を見ると，Previous よりも提案手法の方が認識率の差

表 2 検出・識別結果

Table 2 Detection and classification results

| | 5 dB | 0 dB | -5 dB |
|---|---|---|---|
| Recall | 0.850 | 0.908 | 0.942 |
| Precision | 0.861 | 0.868 | 0.871 |
| Classification | 0.290 | 0.382 | 0.406 |

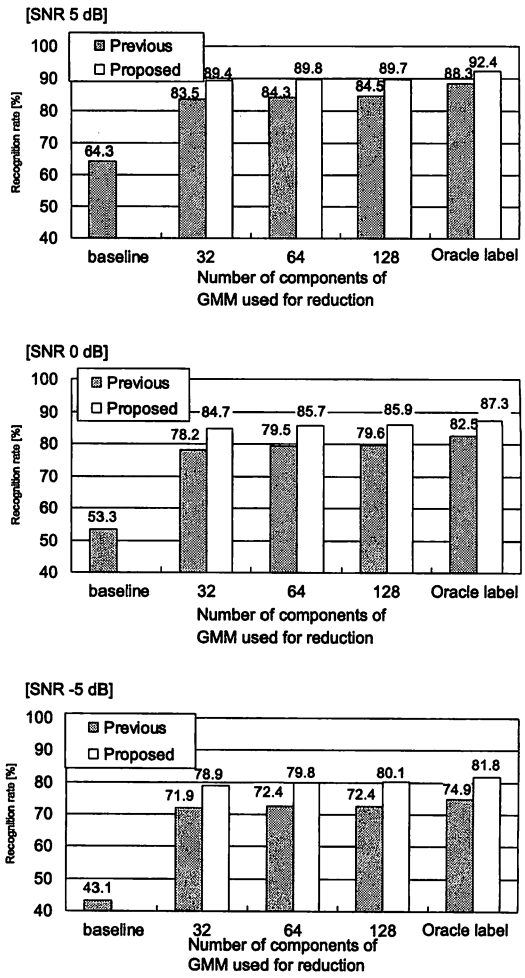


図 5 5 dB, 0 dB, -5 dB での認識結果

Fig. 5 Recognition results at SNRs of -5 dB, 0 dB and 5 dB

表 3　未知雑音に対する検出の結果
Table 3　Results of detection for unknown noises

| | 5 dB | 0 dB | -5 dB |
|---|---|---|---|
| Recall | 0.831 | 0.886 | 0.926 |
| Precision | 0.849 | 0.856 | 0.860 |

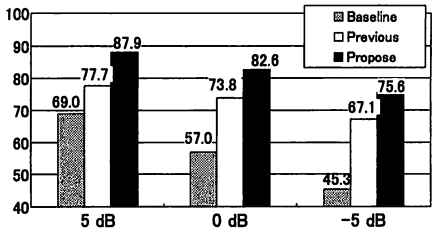


図 6　未知雑音が重畳した場合の認識率
Fig. 6　Recognition results for words utterances mixed unknown noises

が小さいことがわかる．これは本手法が雑音が検出できなかったフレームに対してもある程度補正をかけたためと思われる．しかしながら，それでもラベルを手動で与えた場合のほうが大きく改善していることがわかる．なお，雑音が重畳していない場合の認識率は 98.9 % であった．

### 7.3　未知雑音に対する実験

上記の実験はデータはオープンであるが，学習データの雑音の種類は等しかった．そこで 10 fold クロスバリデーションを用いて未知の雑音に対して実験を行う．まず 105 種類を 10 セットに分割し，そのうち 9 セットを学習用に，1 セットをテスト用に使用する．128 混合の GMM を使って雑音を除去し，その場合の認識率を確認した．そのほかの条件は上記の実験と同様である．まず検出結果を表 3 に示す．雑音が検出されたフレームに対しては，学習データで作られた識別器によって，何かしらのラベルがつくことになる．しかし，テスト用のデータに対してはどのラベルをつけるのが最適かは分からないので，識別率は計算していない．このときの認識率を図 6 に示す．未知雑音に対しても以前の手法に比べて改善しており，また雑音除去を行わなかった場合と比べても大幅な改善が見られた．

## 8. おわりに

静的特徴のモデルと動的特徴量のモデルを併用した突発性雑音の除去法について提案を行った．

雑音が未検出のフレームに対しても補正をかけることができ，また，除去の効果も以前より高かったため，認識率を以前の手法に比べて大幅に改善することができた．しかしながら，それでも手動でラベルを与えた場合と比べると認識率は低くなってしまうため，検出精度の改善が今後の課題に挙げられる．

## 文　献


[1] N. Evans, J. Mason, W. Liu and B. Fauve, "An assessment on the fundamental limitations of spectral subtraction," ICASSP2006, Vol.1, pp. 145-148, 2008.

[2] P. J. Moreno, B. Raj and R. Stern, "A Vector Taylor Series Approach for Environment Independent Speech Recognition, " Proc. ICASSP-1996, pp. 733-736, 1996.

[3] L. Deng, J. Droppo and A. Acero, "A Bayesian approach to speech feature enhancement using the dynamiccepstral prior," ICASSP2002, Vol. 1, pp. I-829-32, 2002.

[4] L. Deng, J. Droppo and A. Acero, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," IEEE Trans. SAP, vol. 12, pp. 133-143, 2004.

[5] K. Tachibana, H. Saruwatari, Y. Mori, S. Miyabe, K. Shikano, A. Tanaka, "Efficient Blind Source Separation Combining Closed-Form Second-Order ICA and Nonclosed-Form Higher-Order ICA," ICASSP2007, Vol. 1, pp. 45-48, 2007.

[6] R. Weiss, M. Mandel, D. Ellis, "Source separation based on binaural cues and source model constraints," Interspeech2008, pp. 439-442, 2008.

[7] 野口賢一ら，"通信会議における 1 チャネル突発性雑音抑圧，" 電子情報通信学会技術研究報告. EA, Vol.105, No.403, pp. 31-36 , 2005.

[8] K. Manohar and P. Rao, "Speech enhancement in nonstationary noise environments using noise properties," Speech Communication, 48(1), pp. 96-109, 2006.

[9] T. Jitsuhiro, T. Toriyama and K. Kogure, "Robust speech recognition using noise suppression based on multiple composite models and multi-pass search," ASRU2007, pp. 53-58 , 2007.

[10] N. Miyake, T. Takiguchi and Y. Ariki, "Sudden Noise Reduction Based on GMM with Noise Power Estimation," Interspeech2008, pp.403-406, 2008.

[11] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura and T. Yamada,"Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands- Free Speech Recognition, " 2nd ICLRE, pp. 965- 968, 2000.

[12] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," Journal of Comp. and System Sci., 55, pp. 119-139, 1997.

[13] E. Alpaydin, "Introduction to Machine Learning," The MIT Press, October 2004.