人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B302-02

ノンパラメトリックベイズモデルを用いた雑音ロバストな音響イベント同定

Noise-robust Acoustic Event Identification Based on a Nonparametric Bayesian Model

中村 圭佑, ゴメス ランディ, 中臺 一博

Keisuke NAKAMURA, Randy GOMEZ, Kazuhiro NAKADAI

（株）ホンダ・リサーチ・インスティチュート・ジャパン

Honda Research Institute Japan Co., Ltd.

keisuke@jp.honda-ri.com, r.gomez@jp.honda-ri.com, nakadai@jp.honda-ri.com

Abstract

本稿では，実環境応用のための雑音ロバストな音響イベント同定について述べる．既存の GMM などのパラメトリックモデルを用いた同定手法は，目的音に雑音や残響が混入した際の学習環境と音環境のミスマッチによって性能が劣化する問題があった．そこで本稿は Latent Dirichlet Allocation と Nested Pitman-Yor Process に基づくノンパラメトリックベイズモデルを用いて，大量の音環境データから音響イベント同定の統計的モデルを構築する手法を提案する．提案手法により，未知雑音が存在する環境下の音響イベント同定において，既存法より 5-18 pts の性能向上が確認できた．

1 序論

実環境での音を媒介としたシーン理解（音環境理解[1]）に関する研究が盛んに行われている．実環境における音環境は，音声だけでなく音楽や環境音などが含まれるため，音環境理解を実現するためには，特定の音に特化しない一般の音の理解（一般音理解）が必要不可欠である．音響イベント同定とは，音響信号の中からイベント（音源）を検出し，その種類や名前といった高次シンボルの抽出（同定）を行う一般音理解を実現するための要素技術であり，近年 ICASSP2013 での Special Session や WASPAA2013 での D-CASE Challenge [2] などで盛んに研究が行われるようになった．音響イベント同定における近年の主要問題は，音声に比べて時間的にも周波数的にも多様な特性を持つ音（反復音，突発音，音楽などの混合音など）に対応しうる特徴量抽出 [3; 4] と識別器設計 [5-22] であるといえる．多様な音に対応した識別器実現のためには，周波数だけでなく時間を陽に考慮した設計が重要となり，これまで，非負値行列因子分解を用いた手法 [7; 8; 20]や，重み付き有限状態トランスデューサを用いた手法 [13]，スペクトログラムを用いた手法 [14; 16]，隠れマルコフモデルを用いた手法 [15]，特徴量ヒストグラムを用いた手法 [19]，音声認識に倣った手法 [3; 21; 22]などが提案されてきた．これらの手法は，ガウス混合モデルを用いた手法 [5] などの周波数特性のみを用いた手法に比べ，時間を扱えるようになったことから，音声（話者同定）だけに限定しない多様な音に対応できるようになった．しかし，これらのパラメトリックモデルを用いた手法は，様々な音に対する最適なモデルを考慮することが難しいという課題がある．短い突発音や長い音楽など，各音源の長さや複雑さはそれぞれ異なるはずであり，パラメトリックモデル構築の際にあらかじめ決めなければならないモデルパラメータ（信号長，フレーム長，状態数，N グラムの次数など）は音の長さや複雑さに合わせて調節可能であることが望ましく，既存法は音源毎に異なるモデルを考慮しうるほど表現能力が十分でない．また，パラメトリックモデルを用いる場合，音環境が学習に用いた環境と適合しない場合は性能が低下してしまう問題があり，雑音や残響などが混入する実環境下の応用において課題を残している．

本稿ではこれらの，1) 音によって異なる長さや複雑さの考慮，2) 音環境と学習環境のミスマッチ問題に取り組み，実環境ロバストな音響イベント同定を実現することを目的とする．

1) に対して，本稿は音声認識に倣った音響イベント同定にノンパラメトリックベイズモデルを導入する．これまでの音声認識に倣った手法[3; 21; 22]は，音の長さをある程度表現することが可能であるが，パラメトリックモデルを用いているため，2) の問題解決が十分になされていない．また，モデルパラメータが固定されているため，音源毎の音の長さや複雑さに合わせた最適なモデルを得ることが難しかった．そこで，本稿はノンパラメトリックベイズ法の一つである Nested Pitman-Yor (NPY) 過程[26]を一般音の音響イベント同定モデル生成のために適用し，

大量の音環境データから，音響イベント同定の統計的モデルを構築する．これにより，任意の長さのセグメント（単語）と N-gram 言語モデルの次数を教師無し学習で推定でき，音の長さや複雑さをモデルに反映することができる（2.3 章）．これまでにもノンパラメトリックベイズモデルを用いた音響イベント同定手法は提案されてきた[18; 19; 20]が，同定に用いる信号の時系列長を固定していたため，音源毎の音の長さを陽に考慮することが困難であった．提案法は時系列長が可変なモデルを用いるため，音の長さをより柔軟に表現することが可能である．

2) に対し，2 つのアプローチを提案する．まず，Latent Dirichlet Allocation (LDA) [24] を一般音に適用し，音響特徴量を符号化する際に必要な音の基本単位（符号）を雑音ロバストとなるように選択することで，実環境下の残響や雑音で生じる音環境と学習環境とのミスマッチを吸収する（2.2 章）．2 つ目に，音の基本単位の相互距離に基づくあいまい検索を導入して音響イベント同定の雑音ロバスト性向上を図る（2.4 章）．

提案手法を学習環境とミスマッチする雑音環境下における音響イベント同定に適用し，その有効性を示す．

2　提案手法

本章では，音響特徴量抽出について触れた後，ノンパラメトリックベイズモデルを用いた手法の詳細について述べる．

2.1　音響特徴量抽出

ある音源からのモノラル信号入力に対する短時間フーリエ変換 $u_\tau(\omega)$ を，目的音 $s_\tau(\omega)$ と雑音 $n_\tau(\omega)$ が混合した以下のモデルとして定義する．

$$u_\tau(\omega) = s_\tau(\omega) + n_\tau(\omega) \tag{1}$$

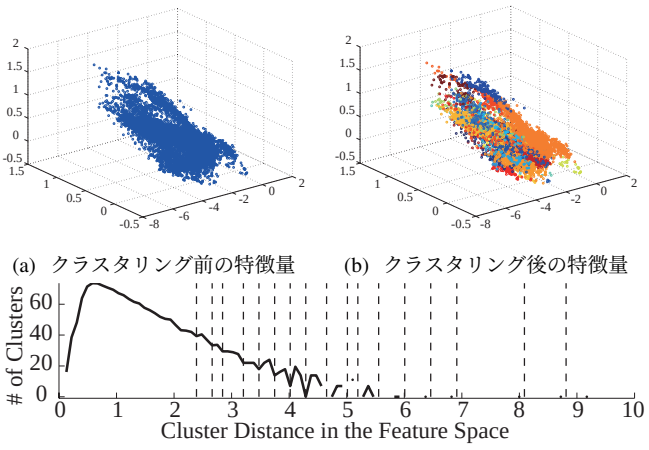
ここで，τ はフレーム番号を表す．$u_\tau(\omega)$ から音響特徴量を抽出し，それを $x_{d\tau}$ と表す．ここで，$u_\tau(\omega)$ には，D 個の音響イベントが含まれるとし，d $(1 \le d \le D)$ はそのインデックスとする．また，d 個目の音響イベントを構成するフレーム総数を N_d とする．ゆえに，$1 \le \tau \le N_d$ に対して，$x_d = \{x_{d1}, \ldots, x_{dN_d}\}$ となる．

2.2　音響イベントの基本単位の推定

音響イベントの基本単位の推定は，音響特徴量のクラスタリングを行う際に，クラスタリングの閾値を LDA によって最適に設定することによって実現する．

2.2.1　凝集型階層クラスタリング

音声認識における音声信号の音素のように，音響イベントにおける一般音の基本単位を「音ユニット」と定義し，音響イベントを音ユニットに分解する．音ユニットを定義するため，音響特徴量の凝集型階層クラスタリング



(c) ヒストグラムと極小点

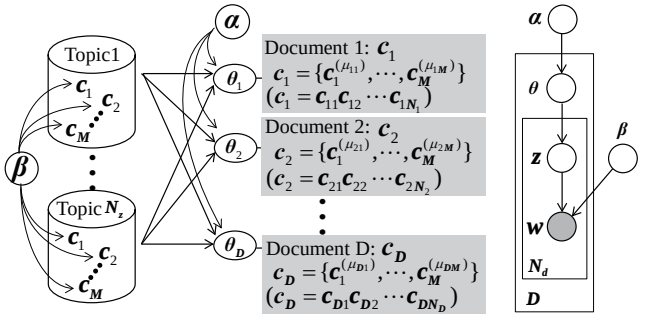
Figure 1: 凝集型階層クラスタリングの例

[23] を行う．ここで，便宜上，$u_\tau(\omega)$ に含まれる音響イベントのフレーム特徴量 $x_{d\tau}$ を時系列に並べ，インデックスを振り直し，x_n $(1 \le n \le N)$ と表記する．凝集型階層クラスタリングを Figure 1(a) のように x_n に対して行う．ここで $N = \sum_{d=1}^{D} N_d$ である．距離関数 Δ とクラスタ間距離は，クラスタ重心のユークリッド距離として定義した．したがって，i 番目と j 番目のクラスタの距離関数は以下で定義される．

$$\Delta_{ij} = \left\| \frac{1}{N_i} \sum_{n \in i} x_n - \frac{1}{N_j} \sum_{n \in j} x_n \right\| , \tag{2}$$

ここで，N_i と N_j はそれぞれ，i 番目と j 番目のクラスタに属するサンプル数である．全ての i と j $(i \neq j)$ の組の中で Δ_{ij} が最小となる組を凝集させることをクラスタ数が十分に小さくなる（1 になる）まで繰り返す．音ユニットは得られた各クラスタに対して定義される．クラスタ間距離に対するクラスタ数のヒストグラムを Figure 1(c) に示す．横軸は Δ_{ij} の距離範囲を 100 に分割したビンを表す．縦軸はそれぞれのビンに対応した距離範囲において統合されたクラスタ数を表す．Figure 1(c) は点線で示された極小を複数持ち，これらの極小となる距離のいずれかで分割したクラスタから音ユニットを構成する．Figure 1(a) と 1(b) にクラスタリング前の特徴量分布と，ヒストグラムが極小となる距離でクラスタリングした特徴量分布を示す．図では，x_n を特異値分解を用いて三次元に次元削減を行ったものをプロットしている．$C = \{C_1, C_2, \ldots, C_M\}$ を得られたクラスタの重心の集合とする．ここで，M はクラスタ数を表す．また，m 番目のクラスタ C_m に対応する音ユニットを c_m と定義する $(1 \le m \le M)$．従って，音ユニットの集合は，$c = \{c_1, c_2, \ldots, c_M\}$ となる．$x_{d\tau}$ に対応した音ユニット（以降，$c_{d\tau}$ と表記する）は，$m = \underset{1 \le m \le M}{\operatorname{argmin}} ||C_m - x_{d\tau}||$ を満たす c_m として決定される．最終的に，d 番目の音響イベントの音ユニット系列は以下のように符号化できる．

$$c_d = c_{d1} c_{d2} \ldots c_{dN_d} , \tag{3}$$



(a) Illustration of LDA (b) Graphical Model

Figure 2: LDA の生成過程とグラフィカルモデル

ここで，c_d と $c_{dn}(1 \le n \le N_d)$ は，次節の LDA における文書と単語とみなして処理を行う．

2.2.2 LDA に基づく音ユニットの選択

雑音ロバストな音ユニットを得るためには，Figure 1(c) が極小となる複数の距離候補から最適なものを選択することが求められる．直感的には，音ユニット同士の距離が近い場合は，音の変化に対する感度が向上するものの，定常音に対して符号化された系列が揺れてしまう場合があり，遠い場合は雑音ロバストな音ユニットが得られるものの，異なる音響イベントを表現しうるほどの感度を満たせない可能性がある．そこで，最適な距離候補の選択に LDA [24] を用いる．LDA は文書モデルの一種で，N_z 個の潜在トピック ($z = \{z_1, z_2, \ldots, z_{N_z}\}$) でコーパス上の文書が表せると仮定した確率的生成モデルである．本稿では，LDA を 式 (3) に適用する．すなわち，コーパスは D 個の文章からなる文章集合 W ($W = \{c_1, c_2, \ldots, c_D\}$) を持ち，$d$ 番目の文書は，音ユニットから定義される N_d 個の単語 ($c_d = c_{d1}c_{d2}\ldots c_{dN_d}$) から構成されるとする．ここで，語彙数は M となり ($c = \{c_1, c_2, \ldots, c_M\}$)，音ユニットの種類数と一致する．LDA のため，単語 c_m $(1 \le m \le M)$ に対して，d 番目の文書中の c_m の個数を上付き文字で $c_m^{(\mu_{dm})}$ と表す．ここで，μ_{dm} は d 番目の文書中に存在する単語 c_m $(1 \le m \le M)$ の数，すなわち $N_d = \sum_{m=1}^{M} \mu_{dm}$ である．LDA では，z を生成するための N_z 次元の確率 ($\theta_d = \{\theta_{d1}, \theta_{d2}, \ldots, \theta_{dN_z}\}$) がディリクレ分布 $\mathrm{Dir}(\theta_d|\alpha)$ に従うと仮定する．W を生成するための確率は以下で表される．

$$p(W|\alpha,\beta) = \prod_{d=1}^{D} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{k=1}^{N_z} p(z_{dk}|\theta_d) p(c_{dn}|z_{dk},\beta) \right) d\theta_d ,$$

ここで，α と β が LDA で推定するパラメータとなる．$\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_{N_z}\}$ は $\mathrm{Dir}(\theta_d|\alpha)$ のためのパラメータであり，$\beta \in \mathbb{R}^{\mathbb{N}_F \times \mathbb{M}}$ は，トピック $z_k(1 \le k \le N_z)$ の中の語彙 $c_m(1 \le m \le M)$ のユニグラムの確率 $p(c_m|z_k)$ である．Figure 2 に生成過程のイメージとグラフィカルモデルを示す．α と β の推定のため，本稿では変分ベイズを用いた．

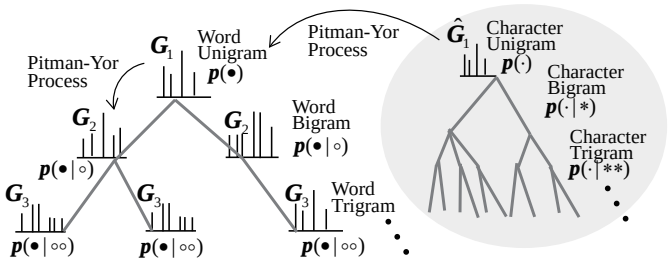


Figure 3: NPY 過程での単語・文字 N グラムモデル

最適な距離の選択のため，Figure 1(c) の N_l 個の極小に対して，それぞれ式 (3) の c_d を算出し，LDA を行う．$\beta = [\beta_1, \ldots, \beta_M]$ を各語彙（音ユニット）に対応した列ベクトルとする．i 番目と j 番目の列ベクトル β_i と β_j の向きが十分に近い場合，それらに相当する音ユニットである c_i と c_j はトピック空間内で同じものを表すと考えられるため，統合する必要がある．従って，N_l 個の候補の中で，これらの列ベクトルの分散が最大となるように距離を選択することで，最適な音ユニットを選択する．具体例を以下に示す．

1) $x_{d\tau}$ の凝集型階層クラスタリングを行う．
2) クラスタ間距離-クラスタ数のグラフを算出し，N_l 個の極小を得る．
3) $1 \le n \le N_l$ に対して以下を繰り返す．
   - n 番目の極小となる距離でクラスタ C を形成する．
   - 全ての d に対して音ユニット系列 c_d を計算する．
   - c_d を用いて LDA の β を推定する．
4) β の列ベクトルの分散が最大となる n を選択する（$\hat{n}$ とする）．
5) $\hat{n}$ 番目の極小に対応する音ユニット $\hat{c} = \{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_M\}$ を用いて音ユニット系列 $\hat{c}_d = \hat{c}_{d1}\hat{c}_{d2}\ldots\hat{c}_{dN_d}$ を求める．

2.3 NPY 過程によるノンパラメトリックな時間統合

自然言語処理では，単語が経験的にわかっているので，音素（音声信号の基本単位）を単語ごとに区切ることができる．一方，非音声を含む一般音の音響イベントでは，そのような分割を経験的な知見から，事前情報として得ることは難しい．事前情報無しに分割を行うため，自然言語の形態素解析に用いられる NPY 過程 [26] を用いる．NPY 過程は，言語を単語 N グラムと文字 N グラムのネスト構造でモデル化する．二つの N グラムモデルの推定には，未知語に対してロバストな Hierarchical Pitman-Yor (HPY) 過程 [25] を用いる．HPY 過程はディリクレ過程の階層的な拡張である．単語 N グラムモデルでは，単語列 $h = w_{t-n}, \ldots, w_{t-1}$ の次の単語 w の生成確率が以下で表される．

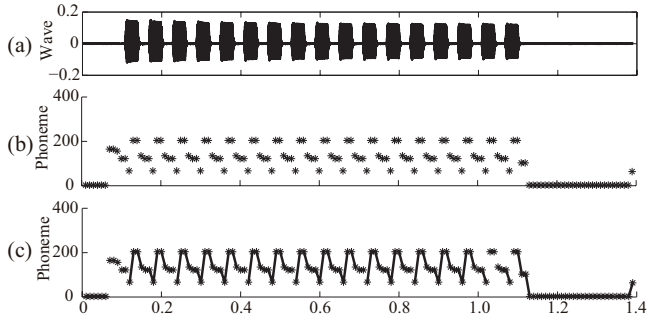


Fig.4.1. Repetitive sound : Phone ring

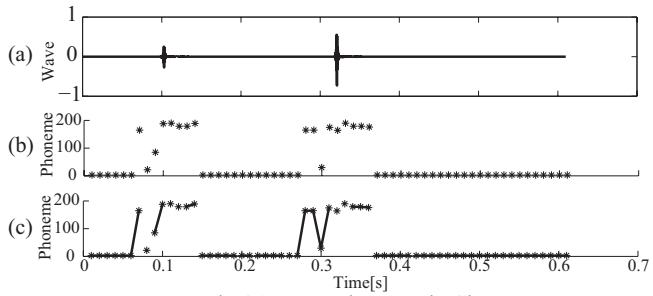


Fig.4.2. Percussive sound : Clap

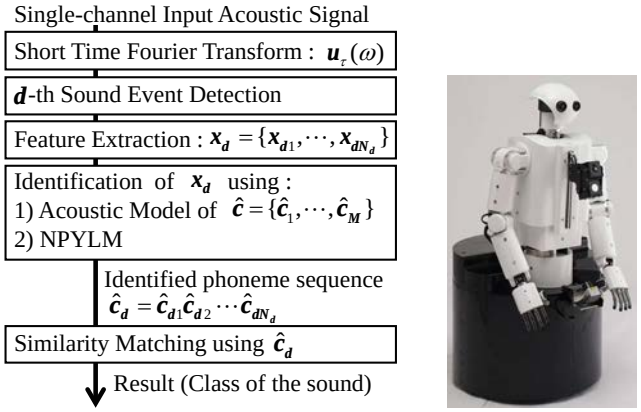
Figure 4: NPY 過程の例: (a) 音響信号, (b) 分割前音ユニット系列, (c) 分割後音ユニット系列

$$p(w|h) = \frac{\gamma(w|h) - \eta t_{hw}}{\xi + \gamma(h)} + \frac{\xi + \eta t_{hw}}{\xi + \gamma(h)} p(w|h') \,, \qquad (4)$$

ここで， $h' = w_{t-n-1}, \ldots, w_{t-1}$ は h から前単語 w_{t-n} を除いた文字列である．ゆえに，$p(w|h')$ は $(n-1)$ グラムの系列 h' から w を生成する確率となる．$\gamma(w|h)$ は h の w の数であり，$\gamma(h) = \sum_w \gamma(w|h)$ である．t_{hw} は h' から生成された w の数であり，$t_h = \sum_w t_{hw}$ である．η と ξ は HPY 過程のパラメータであり，Gibbs sampling を用いて推定する．

HPY 過程は次数を一つ減らした $p(w|h')$ を $p(w|h)$ の推定に用いるため，単語ユニグラムの指定に使用する基底測度 $p(w|h')$ を事前知識として与えられた語彙から得る．しかし，セグメントの情報が与えられていない場合は基底測度を得ることができない．そこで NPY 過程では文字 N グラムを階層的に設け，基底測度を HPY 過程を用いて推定することでこの問題を解決する．本稿における NPY 過程は $\hat{c}_d$ を用いて文字 N グラムと単語 N グラムを Figure 3 のように同時に推定する．ここで，2.2.2 節の LDA では各音ユニットが単語として定義されたが，NPY 過程における単語 w は音ユニット系列として定義される．

Figure 4 に電話の着信音（定常的な反復音:Figure 4.1(a)）と拍手（突発音:Figure 4.2(a)）から得られた音ユニット系列と単語列を示す．横軸と縦軸はそれぞれ時間フレーム τ と音ユニット番号 $\hat{c}_{dN_d}$ を表す．Figure 4.1(b) と Figure 4.2(b) に分割前の音ユニット系列を点線で，Figure 4.1(c) と Figure 4.2(c) に分割後の単語系列を実線で示す．図より，電話の着信音は繰り返しを単位として同じ単語に，拍手は突発音の前後で一単語として区切られたことがわかる．NPY 過程で得られた単語を用いることで，音響信号の時間情報を考慮した分割を実現することができる．



(a) 処理の流れ　　(b) ロボット

Figure 5: 処理の流れとハードウェア

2.4 あいまい検索による雑音補償

NPY 過程を一般音に適用することによる問題は，もともと符号化された自然言語を対象とするため，雑音が混入したことによる入力系列の揺れを扱うことができないことである．こうした揺れを補償するため，本稿は音ユニットの相互距離に基づくあいまい検索を導入する．

前節の凝集型階層クラスタリングから得られた C に対して，クラスタ重心の相互距離（ユークリッド距離）を計算し，あるクラスタから N_Δ 番目に近いクラスタまでを同じクラスタ，つまり同じ音ユニットとみなして処理を行う．例えば，x_d が $\hat{c}_d = c_1c_2c_3$ と推定されたとする．$N_\Delta = 1$ で，(c_1, c_2) と (c_3, c_4) の組が近いと判定された場合，あいまい検索では，これらの音ユニットを入れ替えた $c_1c_1c_3$, $c_2c_1c_3$, $c_1c_2c_3$, $c_2c_2c_3$, $c_1c_1c_4$, $c_2c_1c_4$, $c_1c_2c_4$, $c_2c_2c_4$ も $\hat{c}_d$ として用いる．従って，長さ N_d の $\hat{c}_d$ を同定する場合は，$N_d^{N_\Delta+1}$ 個の候補系列が検索対象となる．

3 評価実験

提案手法と既存手法の音響イベント同定の性能比較を行う．Figure 5(a) に音響イベント同定の処理の流れを示す．評価では Figure 5(b) のロボットの頭部額の位置に設置されたマイクを用いて，音響イベントを 1m の距離で発生させて録音を行った．ロボットは残響時間 0.2 秒の部屋の中央に配置した．音響信号は 16bit，16kHz でサンプリングした．音響特徴量抽出のフレーム長とシフト長は 512, 160 とした．同定に用いるモデルの学習と評価は，Table 1 に示されるデータセットを用いて 5 分割交差検定を行った．手法の雑音ロバスト性を評価するため，学習は雑音の無い環境のデータのみを用いて行った．評価では，実環境下の同定性能評価のため，音声認識で一般的に用いられるバブル雑音とロボット背面のファンの雑音を用い，これらの雑音の SN 比を変化させて行った．SN 比は SNR[dB] $= 20\log_{10}(\pi_s/1-\pi_s)$ $(0 \le \pi_s \le 1)$ と定義した．

ここで，音響イベント信号 $s_\tau(\omega)$ と雑音 $n_\tau(\omega)$ は

Table 1: 学習・評価用データセット

Speech	Dataset : ATR dataset (216 words by 5 male and 5 female speakers) # of cls : 2 (male and female)
Music	Dataset : RWC-MDB-G [28] (32 genres of music for approx. 5 minutes) # of cls : 32 (ex. popular, ballad, etc.)
Environment	Dataset : RWCP [29] (92 kinds of sounds for approx. 4 minutes) # of cls : 92 (ex. phone ring, clap, etc.)

$u_\tau(\omega) = \pi_s s_\tau(\omega) + (1-\pi_s) n_\tau(\omega)$ となるよう混合した．

π_s={1,0.95,0.9,0.85,0.8,0.7,0.5,0.3} とし，SNR[dB] = $\{\infty, 12.8, 9.5, 7.5, 6.0, 3.7, 0.0, -3.7\}$ となった．

音響特徴量には 41 次元の Mel Scale Log Spectrum(MSLS) [27]を用いた (13 次元 MSLS + Δ + Δ^2 + ΔE + Δ^2E).

3.1 LDA に基づく音ユニットの雑音ロバスト性

提案法による音ユニットの雑音ロバスト性を評価するため，音ユニットを用いて学習した GMM(GMM-D) と，手動ラベルを用いて学習した GMM (GMM-S[5]) の同定性能を比較した．GMM は雑音の無いデータを用いて学習した混合数 16 のものを用いた．

手動ラベル (GMM-S) では，x_d の全区間において $\hat{c}_d = c_d c_d \ldots c_d$ と一様にラベル付けを行い，音ユニットの種類数は Table 1 のデータベースの正解クラス数と同じ 126(=2+32+92) とした．一方 GMM-D では，音ユニットは LDA によって抽出されるため，音ユニットの種類数は自動的に決められ，本評価では 96 種類となった．また，ラベル付けは自動で行われる．評価には各音響イベントの平均フレーム正解率 (Frame Correct Rate (FCR)) を用いた．

Figure 6 の GMM-S と GMM-D を比較すると，SNR = ∞ に対しては GMM-S が GMM-D より高い性能を示した．GMM-D では音ユニットが自動的に決められるため学習データに対して過剰適合してしまったと考えられる．クリーン環境における性能向上は今後の課題である．一方，SNR ≠ ∞ では，GMM-D が GMM-S より 3-9 pts 性能が高いという結果が得られた．このことから提案法によって自動的に抽出された音ユニットの雑音ロバスト性を確認できた．

3.2 NPY 過程に基づく言語モデルとあいまい検索

最後に，2.2 節で得られた音ユニット系列を用いて，NPY 過程によって得られた言語モデルを音響イベント同定に適用した．言語モデルの語彙数は 351 となった．言語モデルの有効性検証のため，前節の GMM から，モノフォン HMM の音響モデルへ拡張した．ここで状態数と混合数はそれぞれ 1，16 とし，提案する音ユニットに基づいて学習を行った．まず，音響モデルを拡張したことによる影響を確認するため，Figure 6 の MONO-D にユニグラム

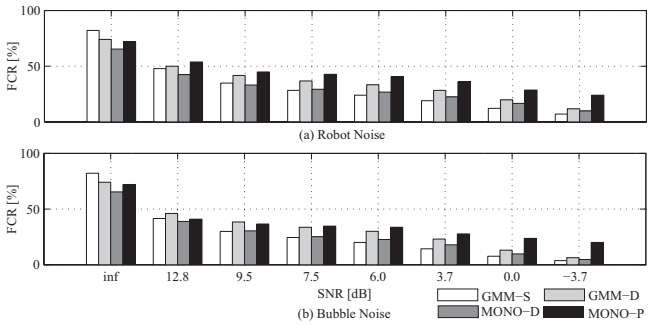


Figure 6: 平均フレーム正解率による同定結果

の言語モデルを用いた時の結果を示した．モノフォンモデルは GMM に比べて次状態の音ユニットへの遷移を抑制するため，GMM-D より性能が劣化したと考えられる．

次に，MONO-D と同じ音響モデルを用いて，ユニグラム言語モデルと後段処理に NPY 過程とあいまい検索を適用した時の結果を Figure 6 に示す．3.1 節で議論した GMM-D と同じ問題が影響し，クリーンな環境においては，MONO-P は GMM-S ほど高い性能を示さなかった．一方，クリーン環境以外の全ての場合において，MONO-P は GMM-S や GMM-D に比べて，それぞれ 5-18 pts, 5-13 pts 高い性能を示しており，雑音ロバスト性の向上を確認できた．実環境での音響イベント同定では，未知の雑音環境での動作が求められるため，未知雑音へのロバスト性が高いことは，実環境への適用性が高いと言える．

4 結論

本稿は雑音存在下の一般音の音響イベント同定について述べた．一般音の音響イベント同定を実環境応用する際に問題となる，1) 音によって異なる長さや複雑さの考慮と，2) 音環境と学習環境のミスマッチ問題に取り組んだ．1) に対し，音声認識に倣った音響イベント同定に対して，NPY 過程に基づくノンパラメトリックベイズモデルを導入した．2) に対し，LDA に基づく雑音ロバストな音ユニットの抽出と，音ユニットの相互距離に基づくあいまい検索を提案した．実環境収録データで評価実験を行った結果，一般的な GMM と比較し，学習環境とミスマッチする雑音環境下で 5-18 pts の性能向上が得られ，提案法の有効性を示すことができた．

参考文献

[1] D. Rosenthal and H. G. Okuno, “Computational Auditory Scene Analysis”, Lawrence Erlbaum Associates, Mahwah, New Jersey, pp. 399+xiii, 1998.

[2] D. Giannoulis *et al.*, “Detection and classification of acoustic scenes and events: an IEEE AASP challenge”, in *IEEE WASPAA*, 2013.

[3] X. Zhuang *et al.*, "Real-world acoustic event detection", *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[4] L. Ballan *et al.*, "Deep networks for audio event classification in soccer videos", in *ICME*, pp. 474–477, 2009.

[5] D. A. Reynolds *et al.*, "Robust text-independent speaker identification using gaussian mixture speaker models", *IEEE TSAP*, vol. 3, no. 1, pp. 72–83, 1995.

[6] K. Nakamura *et al.*, "Intelligent Sound Source Localization and Its Application to Multimodal Human Tracking", in *Proc. of IEEE/RAS IROS*, pp. 143-148, 2011.

[7] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection", in *Proc. of IEEE WASPAA*, pp. 69–72, 2011.

[8] M. L. Chin *et al.*, "Audio event detection based on layered symbolic sequence representations" in *Proc. of ICASSP*, pp. 1953–1956, 2012.

[9] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments", *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[10] A. Mesaros, T. Heittola, and A. Klapuri, "Latent semantic analysis in sound event detection", in *Proc. of 19th EUSIPCO*, pp. 1307–1311, 2011.

[11] B. Schauere *et al.*, " "Wow!" Bayesian surprise for salient acoustic event detection", in *Proc. of ICASSP*, pp. 6402–6406, 2013.

[12] K. H. Lin *et al.*, "Improving faster-than-real-time human acoustic event detection by saliency-maximized audio visualization", in *Proc. of ICASSP*, pp. 2277–2280, 2012.

[13] D. Rybach *et al.*, "Silence is golden: Modeling non-speech events in WFST-based dynamic network decoders", in *Proc. of ICASSP*, pp. 4205–4208, 2012.

[14] Y. Sasaki *et al.*, "Daily sound recognition using Pitch-Cluster-Maps for mobile robot audition", in *Proc. of IEEE/RAS IROS*, pp. 2724-2729, 2009.

[15] V. Ramasubramanian *et al.*, "Continuous audio analytics by HMM and Viterbi decoding", in *Proc. of ICASSP*, pp. 2396–2399, 2011.

[16] C. Bauge *et al.*, "Representing environmental sounds using the separable scattering transform", in *Proc. of ICASSP*, pp. 8667–8671, 2013.

[17] M. Espi *et al.*, "A tandem connectionist model using combination of multi-scale spectro-temporal features for acoustic event detection", in *Proc. of ICASSP*, pp. 4293–4296, 2013.

[18] Y. Sasaki *et al.*, "Nested Infinite Gaussian Mixture Model for Environmental Audio Signal Recognition", in Proc. of SIG-Challenge 2012, B202-07.

[19] T. Nakamura, T., T. Nagai, and N. Iwahashi, "Multimodal categorization by hierarchical dirichlet process", in *Proc. of IEEE/RAS IROS*, pp. 1520–1525, 2011.

[20] Y. Ohishi *et al.*, "Bayesian Semi-supervised Audio Event Transcription based on Markov Indian buffet Process", in *Proc. of ICASSP*, pp. 3163–3167, 2013.

[21] S. Chaudhuri, M.Harvilla, and B. Raj, "Unsupervised Learning of Acoustic Unit Descriptors for Audio Content Representation and Classification", in *Proc. of INTERSPEECH*, pp. 2265-2268, 2011.

[22] A. Kumar *et al.*, "Audio event detection from acoustic unit occurrence patterns", in *Proc. of ICASSP*, pp. 489–492, 2012.

[23] W. H. Press *et al.*, *Numerical Recipes in C: the Art of Scientific Computing*, 2nd ed., Cambridge University Press, 1998.

[24] D. M. Blei *et al.*, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[25] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes", in *Proc. of ICCL and ACL*, vol. 44, pp. 985–992, 2006.

[26] D. Mochiahshi *et al.*, "Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling", in *Proc. of the Joint Conf. of ACL and AFNLP*, vol. 1, pp. 100–108, 2009.

[27] Y. Nishimura *et al.*, "Noise-robust speech recognition using multiband spectral features", in *Proc. 148th Acoustical Soc. of America Meet.*, San Diego, CA, no. 1aSC7, 2004.

[28] M. Goto, "Development of the RWC Music Database", in *Proc. of ICA*, pp. 553–556, Apr. 2004.

[29] S. Nakamura *et al.*, "Sound Scene Database in Real Acoustic Environments", in *Proc. of Oriental COCOSDA Workshop*, pp. 103–106, 1998.