社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

半教師付き正準密度推定法に基づく音響信号の自動タグ付けと検索

高木　潤† 大石 康智†† 木村 昭悟††
杉山　将† 山田　誠† 亀岡 弘和††

† 東京工業大学　大学院情報理工学研究科
†† 日本電信電話（株）NTT コミュニケーション科学基礎研究所

あらまし　本論文では音響信号自動タグ付け・検索問題に対し，半教師付き正準密度推定法 *SSCDE*(*Semi-supervised canonical density estimation*) の適用を試みる．SSCDE は正準相関分析にタグ無しサンプルの大域的分布構造を組み込んだ半教師型正準相関分析 (*SemiCCA*) によりトピックを表現する潜在変数の空間を構築し，カーネル密度推定法を半教師化した多クラス SSKDE によって潜在変数空間上のモデル学習を行う，トピックモデルに基づく半教師型の学習手法である．この手法は画像認識・検索の分野において提案されたものであるが，音響信号に対する適用もスムーズに行うことができる．実際の音楽データを用いた実験により，使用できるタグ付き音響信号が少ない状況下でも，SSCDE を用いて半教師型の学習を行うことにより，タグ付け性能が向上することを確認した．
キーワード　音響信号自動タグ付け・検索　半教師付き学習　トピックモデル推定　正準相関分析　カーネル密度推定

Automatic Audio Tagging and Retrieval Using Semi-Surpervised Canonical Density Estimation

Jun TAKAGI†, Yasunori OHISHI††, Akisato KIMURA††,
Masashi SUGIYAMA†, Makoto YAMADA†, and Hirokazu KAMEOKA††

† Graguate School of Computer Science, Tokyo Institute of Technology
†† NTT Communication Science Laboratories, NTT Corporation

Abstract　We apply SSCDE (semi-supervised canonical density estimation), a semi-supervised learning method based on topic modeling, to audio tagging and retrieval problems. SSCDE was originally proposed as an image annotaion and retireval method, but it can also be applied to audio data. The SSCDE method consists of two parts: 1) extraction of a low-dimentional latent space representing topics of sounds using a semi-supervised variant of canonical correlation analysis, and 2) learning a topic model using multi-class extention of semi-supervised kernel density estimation in the latent space. Audio tagging exrperiments with real-world data indicate that SSCDE improves the annotation accuracy even when only a small number of tagged sounds are available.
Key words　Audio tag classification, semi-supervised learning, topic model, canonical correlation analysis, kernel density estimation

1. はじめに

近年，インターネットを介して，ネットワーク上での膨大な音楽音響信号の蓄積・流通が行われるようになった．これに伴って，この膨大な音楽音響信号を効率的に蓄積し，検索するための技術が必要とされるようになっている．最も一般的な音楽音響信号の蓄積・検索方法は，作曲者やアーティスト，曲名，アルバムの発売日などの『メタデータ』（テキストタグ）を用いるものである．このテキストタグを入力クエリとすることにより容易に楽曲を検索することができ，また楽曲を入力すれば，その楽曲が蓄積されていた場合，付随するテキストタグを出力することができる．さらに，ジャンル名や使用楽器，歌詞，楽曲のレビューなどの情報もメタデータとして利用することにより，蓄積・検索の柔軟性を高めることができる [1], [2]．ただし，タグ付けされていない新規の音響信号を蓄積されている音響信号の中に加えたいときは，人手でタグ付けする必要があるため，大きなコストがかかるという欠点がある．

この問題に対処するため，音響信号とテキストタグを自動的

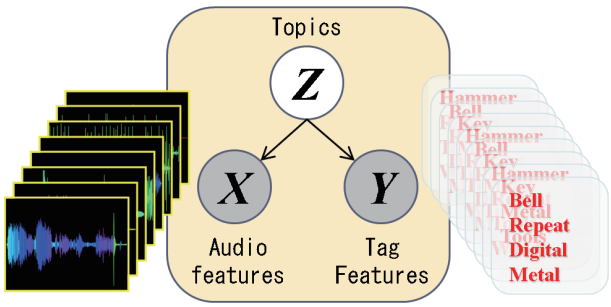

図 1　音響信号自動タグ付け・検索に対するトピックモデル

に関連付けるための研究が行われてきた [3]〜[9]．これらの研究で扱う問題は，未知の音響信号が与えられたときに，適切なテキストタグを自動的に付与する自動タグ付け問題と，テキストタグが与えられたときに，そのテキストタグに適合する音響信号を探して提示する検索問題からなる．本論文では，これら二つの双対問題を統一的に記述できる枠組みを用いるため，これらを総じて音響信号自動タグ付け・検索問題と呼ぶこととする．

音響信号に対する自動タグ付け問題の一般的なアプローチの一つとしては，線形回帰を用いてタグごとの 2 値識別器を構成する方法 [6] や，タグごとに音響特徴の分布を推定する方法 [7] などのように，各テキストタグに対応する識別器を用いる方法が挙げられる．音響信号の自動タグ付け問題は一つの音響信号に対して複数のタグが同時に与えられる多重ラベリング問題の一つとして考えられるため，タグ間の共起関係まで考慮してタグ付けを行うことができれば，さらに精度が向上すると期待される．しかし，識別的な手法を用いる場合，各タグを付与すべきかどうかの判断は個別に行うため，タグ間の共起関係を取り扱える形に拡張することは本質的に困難である．

同じ多重ラベリング問題として捉えることができる画像認識・検索の分野では，こうした背景から，probabilistic latent semantic analysis (pLSA) [10] や latent Dirichlet allocation (LDA) [11] など，トピックモデルと呼ばれる生成モデルに基づくアプローチが注目を集めている．このトピックモデルを音響信号の自動タグ付け・検索問題に当てはめると，図 1 のようになる．ここでは，潜在確率変数 $\boldsymbol{Z}$ を用いて 2 つの観測情報 $\boldsymbol{X}$ と $\boldsymbol{Y}$ を間接的に結びつけるため，タグ間の共起関係をも考慮できる構造となる．このようにトピックモデルを音響信号自動タグ付け・検索問題に応用し，タグ間の共起情報を使って精度の向上を図ることが，本論文の目的の一つである．

また，上記の手法はいずれも，あらかじめタグが付与された音響信号を学習に用いる，教師付き学習に基づくものであった．一般に，教師付き学習を行う場合，その性能を向上させるためには，信頼性の高いタグが付与された音響信号をより多く集めることが必要となる．Last.fm や MyStrands のような楽曲のレビューサイトを介して収集された大量のタグ情報を利用する研究も行われているが [3], [4], [6]，これらのサイトから収集したタグの信頼性は必ずしも高くない．手作業でタグ付けを行えば信頼度の高いタグが得られるが，大きなコストがかかる．例えば，交響曲などのクラシック音楽は一曲だけで 10 分を超える長さのものが多くあり，音楽データ全体を通して聴くだけでも相当な時間を要する．また，音楽を聞いたときの感じ方は個人差が大きいため，主観性を排除してさらに信頼度の高いタグを得るために，一つの音楽データに対して，複数の人がタグ付けすることも行われているが [7]，この場合にはさらにコストが増大する．したがって，タグ付きの音響信号を多く集める代わりに，タグ無しの音響信号を大量に使うことにより精度を向上させる半教師付き学習が，この問題に対して重要な役割を果たすと考えられる．

我々はこれまで，半教師付き学習によりトピックモデルを獲得する簡易な手法 *SSCDE*(*Semi-supervised canonical density estimation*) を提案し，画像の認識・検索問題に適用してきた [12]．この手法の技術的ポイントは，1) 独自に開発した半教師付き正準相関分析 (*SemiCCA*) [13] により，画像とタグの共起関係と大域的な分布構造とを同時に考慮した潜在変数空間を従来の正準相関分析と同等の計算で導出すること，及び 2) カーネル密度推定を用いた半教師付き事後確率推定法である SSKDE [14] を多重ラベリング問題に適用し，潜在変数空間における局所的な非線形構造を捉えたトピックモデルを導出すること，の 2 点である．また実データを用いた実験により，タグ付き画像が少数しかない状況下でも，大量のタグなし画像を活用することにより認識精度を大幅に向上できることを確認した．

本論文では，この SSCDE を音響信号自動タグ付け・検索問題に適用する．すなわち，SSCDE を適用可能にする音響信号の特徴表現を導入し，この音響特徴とタグ特徴とを関連付けるためのトピックモデルを学習する．評価実験では，SSCDE を音響信号自動タグ付け問題に適用した時の性能を評価し，その有効性を検証する．以下，第 2. 節では，SSCDE を音響信号自動タグ付け・検索問題に適用するための枠組みを述べ，第 3. 節では，音響信号の特徴表現の方法について説明する．第 4. 節では音響特徴とタグ特徴とを関連づけるための SSCDE の詳細を述べ，第 5. 節では評価実験内容とその結果を示す．

2.　SSCDE の枠組み

SSCDE を音響信号自動タグ付け・検索に適用する際の処理の流れを図 2，3 に示す．

第 1. 節でも述べたように，SSCDE は半教師型の学習手法であり，タグ付きのサンプルとタグ無しのサンプルとを学習に用いる．以下では，タグ付きのサンプルの個数を N，タグ無しのサンプルの個数を N_x とし，それぞれのサンプルの集合を $\boldsymbol{X}^{(T)} = \{\boldsymbol{x}_n\}_{n=1}^{N}$，$\boldsymbol{X}^{(U)} = \{\boldsymbol{x}_n\}_{n=N+1}^{N+N_x}$，これらを合わせた全サンプルの集合を $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^{N+N_X}$ とする．また，タグ付きサンプルに付与されているタグ（より正確にはタグをベクトル表現に直したもの）を $\boldsymbol{Y} = \{\boldsymbol{y}_n\}_{n=1}^{N}$ とする．音響信号自動タグ付け・検索では $\boldsymbol{X}$ が音響信号の特徴ベクトルの集合，$\boldsymbol{Y}$ がタグ特徴の集合となるが，それぞれの特徴量の抽出に関しては第 3. 節で詳述する．これらの学習用データの集合 $(\boldsymbol{X}, \boldsymbol{Y})$ から，トピックモデル推定を行う際の処理は以下の二つのステップに分けることができる．

ステップ 1 では，SemiCCA を用いて潜在変数 $\boldsymbol{Z}$ ＝

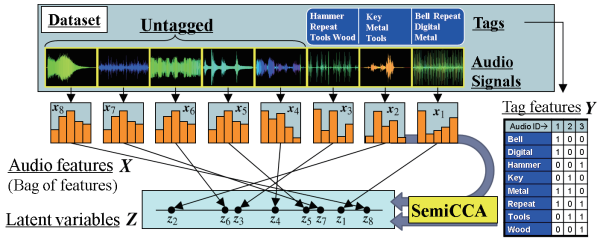

図 2　SSCDE の枠組み（ステップ 1）

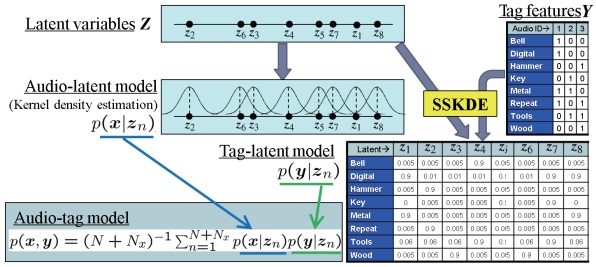

図 3　SSCDE の枠組み（ステップ 2）

$\{\boldsymbol{z}_n\}_{n=1}^{N+N_x}$ を生成する．具体的には，SemiCCA を $(\boldsymbol{X}, \boldsymbol{Y})$ に適用して，タグ付きサンプルとそれに対応するタグから潜在変数に変換する関数 $f_{xy} : \mathcal{R}^{D_x} \times \mathcal{R}^{D_y} \to \mathcal{R}^{D_z}$ と，タグ無しサンプルのみから潜在変数への変換を行う関数 $f_x : \mathcal{R}^{D_x} \to \mathcal{R}^{D_z}$ の組 (f_x, f_{xy}) を求め，これらの関数を用いて $(\boldsymbol{X}, \boldsymbol{Y})$ から $\boldsymbol{Z}$ を生成する．ここで，$\mathcal{R}$ は実数集合，D_x，D_y，D_z はそれぞれ $\boldsymbol{x}_n$，$\boldsymbol{y}_n$，$\boldsymbol{z}_n$ の次元である．ステップ 1 の詳細は 4.1 節で述べる．

ステップ 2 では，潜在変数 $\boldsymbol{z}_n$ が与えられたときのサンプルとタグ $\boldsymbol{x}$，$\boldsymbol{y}$ の条件付き確率からなるトピックモデルを，SSKDE を用いて形成する．

$$p(\boldsymbol{x}, \boldsymbol{y}) = (N + N_x)^{-1} \sum_{n=1}^{N+N_x} p(\boldsymbol{x}|\boldsymbol{z}_n) p(\boldsymbol{y}|\boldsymbol{z}_n) \tag{1}$$

ステップ 2 の詳細は第 4.2 節で述べる．

音響信号自動タグ付け・検索問題への応用では，ひとたびトピックモデルの構築が終わると，タグ付けと検索は両方とも最大事後確率 (MAP) 推定の枠組みで実現できる．タグ付けの際には，未知の音響信号 s が入力されたとき，そこから抽出した音響特徴 $\boldsymbol{x}^{(s)}$ を用いて，それに対するタグ特徴 $\hat{\boldsymbol{y}}$ を以下のように推定できる．

$$\begin{aligned}
\hat{\boldsymbol{y}} &= \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}|\boldsymbol{x}^{(s)}) && (2)\\
&= \arg\max_{\boldsymbol{y}} p(\boldsymbol{x}^{(s)}, \boldsymbol{y}) && (3)\\
&= \arg\max_{\boldsymbol{y}} \sum_{n=1}^{N+N_x} p(\boldsymbol{x}^{(s)}|\boldsymbol{z}_n) p(\boldsymbol{y}|\boldsymbol{z}_n) && (4)
\end{aligned}$$

あらかじめタグ特徴の各次元（$d = 1, 2, \ldots, D_y$）に対して閾値 θ_d を設定しておくことで，上記の式により得られたタグ特徴 $\hat{\boldsymbol{y}}$ の各要素 $\hat{y}_d$ がその閾値を上回るとき，第 d 次元に対応するテキストタグを音響信号 s に付与することができる．

検索も同様にテキストクエリ $\boldsymbol{w}$ が与えられたとき，そこからタグ特徴 $\boldsymbol{y}^{(w)}$ を抽出して，最も適合する音響特徴 $\hat{\boldsymbol{x}}$ を以下のように得ることができる．

$$\begin{aligned}
\hat{\boldsymbol{x}} &= \arg\max_{\boldsymbol{x} \in \boldsymbol{X}} p(\boldsymbol{x}|\boldsymbol{y}^{(w)}) && (5)\\
&= \arg\max_{\boldsymbol{x} \in \boldsymbol{X}} p(\boldsymbol{x}, \boldsymbol{y}^{(w)}) && (6)\\
&= \arg\max_{\boldsymbol{x} \in \boldsymbol{X}} \sum_{n=1}^{N+N_x} p(\boldsymbol{y}^{(w)}|\boldsymbol{z}_n) p(\boldsymbol{x}|\boldsymbol{z}_n) && (7)
\end{aligned}$$

検索結果を複数提示したい場合には，事後確率 $p(\boldsymbol{x}|\boldsymbol{y}^{(w)})$ により音響信号 $\boldsymbol{x}$ をランク付けして提示すればよい．

3. 音響信号自動タグ付け・検索への SSCDE の適用

従来の音響信号自動タグ付け・検索においては，音響信号を表現する際に，音響信号を短い時間のフレームに分割し，各フレームからメル周波数ケプストラム係数（Mel-frequency cepstrum coefficients: MFCC）を取り出すことがよく行われる [7]．MFCC は人間の聴覚特性を考慮した特徴量であり，他の音声信号処理の分野でも標準的に用いられる．本論文では音響信号の特徴量としては始めの 13 次元の MFCC を用いる．さらに，MFCC の動的な変化を見るために，一次と二次微分の近似値である ΔMFCC，ΔΔMFCC を計算し，先の MFCC のベクトルに付け加える．したがって，結局各フレームから 39 次元の特徴ベクトルが取り出されることになる．

ただし，この音響信号の表現のままでは SSCDE を適用できる形にはならない．第 2. 節で述べたように，SSCDE は後述する SemiCCA を用いて潜在変数空間の構築を行うが，この際，一つの音響信号は一つの特徴ベクトルで表現されている必要がある．したがって，SSCDE を適用する際には，一つの音響信号の全てのフレームから取り出した 39 次元のベクトルを全てまとめて，一つの特徴ベクトルとして表現する必要がある．本論文では，代表的な画像特徴量として知られる bag-of-features のように，ベクトル量子化を用いてベクトルの集合をヒストグラムにしたものを特徴量として用いる．このようなヒストグラムを用いた特徴表現は，音響信号検索に対して非常に識別性能が高いことが実験的に証明されている [15], [16].

ベクトル量子化を用いてヒストグラムを作成する際の具体的な処理は以下のようになる．まず，特徴ベクトルをクラスタリングし，ベクトル量子化のためのコードブックを作成する．ただし，一つの音響信号からは大量の MFCC のベクトルが抽出されるので，全音響信号の全ての特徴ベクトルを用いてクラスタリングを行おうとすると計算量が大きくなる．そのため，各音響信号から等しい数の特徴ベクトルをそれぞれランダムサンプリングし，それらを全て合わせたベクトルの集合を使ってコードブックを作成することで，計算を効率化する．次に，音響信号から抽出された全てのベクトルを，コードブックを用いて符号化し，各符号語の数をヒストグラムにする．音響信号の長さによって，そこから得られる MFCC の特徴ベクトルの数は異なり，ヒストグラムをそのまま用いるとスケールが不揃いになってしまうため，最後にこのヒストグラムに対して正規化

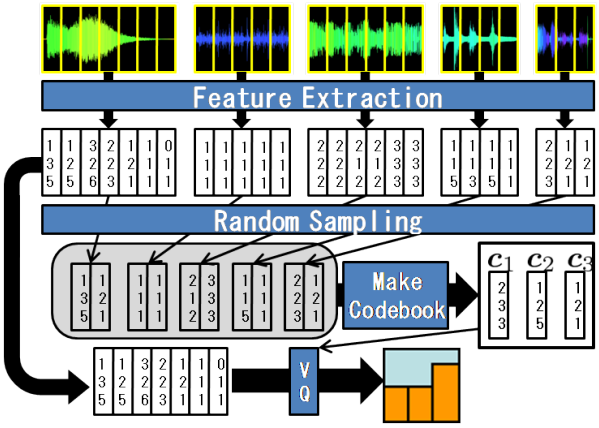

図 4　音響特徴の表現

を行う．このようにして得られた最終的な特徴ベクトルの次元数 D_x は，ベクトル量子化を行う際の符号語の数に等しい値となる．

上記をまとめると，特徴量の抽出の流れは以下のようになる．

（1） 音響信号を短いフレームに区切り，各フレームから 39 次元の特徴ベクトルを抽出する．

（2） 各音響信号から 500 個の特徴ベクトルをランダムサンプリングし，それを全て集めてコードブックを作成する．

（3） 得られたコードブックを元に，各音響信号の全ての特徴ベクトルを量子化して正規化ヒストグラムを作成し，1024 次元の特徴ベクトルとする．

この様子を図にしたものを図 4 に示す．

なお，タグ特徴は先行研究 [12] にならい，タグの種類と同じ次元の 2 値ベクトルを用いて表現する．すなわち，対応するタグが付いている要素の値に 1，付いていない要素の値に 0 を持つベクトルで，音響信号に付けられたタグを表す．

4. SSCDE の詳細

第 2. 節で述べたように，SSCDE は SemiCCA を用いたトピックモデルの構築と，SSKDE を用いた確率密度推定の二つの部分からなる．この節では，これらの詳細について述べる．

4.1 SemiCCA：半教師型正準相関分析

正準相関分析（CCA）[17] はサンプルの集合 $\{\boldsymbol{x}_n\}_{n=1}^N$ とタグの集合 $\{\boldsymbol{y}_n\}_{n=1}^N$ とを，それぞれ射影ベクトル $\boldsymbol{w}_x$ と $\boldsymbol{w}_y$ に射影するとき，射影後のサンプル集合同士の相関を最大にする射影軸 $\boldsymbol{w}_x, \boldsymbol{w}_y$ を求めるものである．この CCA は第 1. 節において述べたトピックモデルに基づく学習手法の一つである，ガウシアンを用いた pLSA の近似として捉えることができる [18]．そのため，CCA をトピックモデルの構築に用いることは計算が効率的であるだけでなく，理論的な裏付けも兼ね備えているといえる．ただし，サンプルが少数しか与えられない状況下では，CCA は過学習を起こす可能性がある．これを避けるため，SSCDE では CCA を半教師型に拡張した SemiCCA を用いて潜在変数空間の構築を行う．

SemiCCA のアイデアは，タグ無しサンプルを含めた大域的な分布構造を用いて射影軸 $\boldsymbol{w}_x, \boldsymbol{w}_y$ を補正するというものである．この際，サンプルの大域的な分布構造には，全サンプルに対する主成分分析（PCA）を用いる．この様子を図 5 に示す．SemiCCA では，CCA と PCA と統一的かつ効率的に解くために，これらを以下に示す一つの固有値問題に帰着させている．

$$\boldsymbol{B}\begin{pmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\end{pmatrix} = \lambda \boldsymbol{C}\begin{pmatrix}\boldsymbol{w}_x\\ \boldsymbol{w}_y\end{pmatrix} \tag{8}$$

$$\boldsymbol{B} = \beta\begin{pmatrix}\boldsymbol{0} & \boldsymbol{S}_{xy}^{(T)}\\ \boldsymbol{S}_{yx}^{(T)} & \boldsymbol{0}\end{pmatrix} + (1-\beta)\begin{pmatrix}\boldsymbol{0} & \boldsymbol{S}_{xy}^{(T)}\\ \boldsymbol{S}_{yx}^{(T)} & \boldsymbol{0}\end{pmatrix} \tag{9}$$

$$\boldsymbol{C} = \beta\begin{pmatrix}\boldsymbol{S}_{xx}^{(T)} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{S}_{yy}^{(T)}\end{pmatrix} + (1-\beta)\begin{pmatrix}\boldsymbol{I}_{D_x} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_{D_y}^{(T)}\end{pmatrix} \tag{10}$$

$$\boldsymbol{S}_{xx} = (N+N_x)^{-1}\sum_{n=1}^{N+N_x}\boldsymbol{x}_n\boldsymbol{x}_n^\top \tag{11}$$

$$\boldsymbol{S}_{yy} = N^{-1}\sum_{n=1}^{N}\boldsymbol{y}_n\boldsymbol{y}_n^\top \tag{12}$$

ここで，$\boldsymbol{S}_{xy}$ は $(\boldsymbol{X}, \boldsymbol{Y})$ の間の散布図行列，$\boldsymbol{S}_{xy}^{(T)}$ はタグ付きサンプルのみの散布図行列，β は 0 以上 1 以下のトレードオフパラメータである．上記の式は $\beta = 1$ の時に CCA の固有値問題，$\beta = 0$ の時に PCA の固有値問題となる．β の値をそれ以外の範囲で変化させることで，大域的構造と対サンプルの共起情報に対する重みを任意に制御することができる．上記の一般化固有値問題の上位の D_z 個の固有値に対応する固有ベクトルを列ベクトルに用いることで，D_z 次元の写像 $\boldsymbol{W}_x$ と $\boldsymbol{W}_y$ を得ることができる．さらに，CCA の確率的な解釈（probabilistic CCA）[18] の導入により，第 2. 節で述べた 2 つの関数 f_x，f_{xy} は次のように導かれ，その出力として潜在変数 $\boldsymbol{Z}$ を得ことができる．

$$f_x(\boldsymbol{x}) = \boldsymbol{M}_x^\top \boldsymbol{W}_x^\top \boldsymbol{x} \tag{13}$$

$$f_{xy}(\boldsymbol{x}, \boldsymbol{y}) = \begin{pmatrix}\boldsymbol{M}_x\\ \boldsymbol{M}_y\end{pmatrix}^\top \begin{pmatrix}\tilde{\boldsymbol{\Lambda}} & -\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Lambda}\\ -\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Lambda} & \tilde{\boldsymbol{\Lambda}}\end{pmatrix}\begin{pmatrix}\boldsymbol{W}_x\boldsymbol{x}\\ \boldsymbol{W}_y\boldsymbol{y}\end{pmatrix} \tag{14}$$

$$\tilde{\boldsymbol{\Lambda}} = (\boldsymbol{I}_{D_x} - \boldsymbol{\Lambda}^2)^{-1} \tag{15}$$

ここで，$\boldsymbol{I}_D$ は $D \times D$ 単位行列，$\boldsymbol{\Lambda}$ は d 番目に大きい固有値 λ_d $(d = 1, 2, \ldots, D_x)$ を d 番目の対角成分とする対角行列，$\boldsymbol{M}_x, \boldsymbol{M}_y$ は $\boldsymbol{M}_x\boldsymbol{M}_y = \boldsymbol{\Lambda}$ を満たし，かつスペクトルノルムが 1 より小さい行列である．

4.2 SSKDE によるトピックモデルの設計

SSCDE では，SemiCCA によって潜在変数空間を構築した後，カーネル密度推定の考え方に基づき，トピックモデルを事例ベースで以下のように設定する．

$$p(\boldsymbol{x}|\boldsymbol{y}) = (N+N_x)^{-1}\sum_{n=1}^{N+N_x} p(\boldsymbol{x}|\boldsymbol{z}_n)p(\boldsymbol{y}|\boldsymbol{z}_n) \tag{16}$$

$$p(\boldsymbol{x}|\boldsymbol{z}_n) = \kappa(f_x(\boldsymbol{x}) - \boldsymbol{z}_n) \tag{17}$$

$$p(\boldsymbol{y}|\boldsymbol{z}_n) = \prod_{d=1}^{D_y} p(y_d|\boldsymbol{z}_n) \tag{18}$$

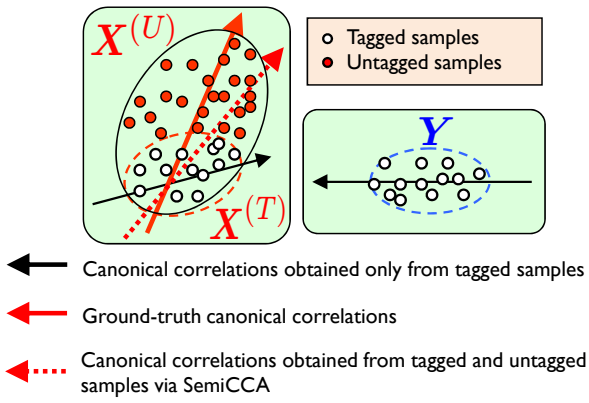

図 5　SemiCCA におけるタグ無し音響信号の影響

$$p(y_d|\boldsymbol{z}_n) = \mu\delta(y_d - y_{n,d}) + (1-\mu)N_d/N \tag{19}$$

ここで，潜在変数 $\boldsymbol{z}_n$ は以下のように計算される．

$$\boldsymbol{z}_n = \begin{cases} f_{xy}(\boldsymbol{x}_n|\boldsymbol{y}_n), & n = 1,\ldots,N \\ f_x(\boldsymbol{x}_n), & n = N+1,\ldots,N_x \end{cases} \tag{20}$$

$\kappa(\cdot)$ はカーネル幅 γ のガウスカーネル，$\delta(\cdot)$ はディラックデルタ，$y_{n,d}$ は $\boldsymbol{y}_n$ の d 番目の要素，N_d は d 番目のタグを持つサンプルの個数，$\mu(0<\mu<1)$ はタグの信頼度を表すパラメータである．

注意すべきは，タグ無しサンプルに対してはタグ $\boldsymbol{y}_n$ が定義されないため，従来の KDE では式（19）の計算を行うことができず，トピックモデル $p(\boldsymbol{x}|\boldsymbol{y})$ の推定にタグ無しサンプルを用いることができないことである．すなわち，この確率密度推定の精度はタグ付きサンプルの数に大きく依存することになる．SSCDE では KDE を半教師型に拡張した，半教師型カーネル密度推定 (semi-supervised kernel density estimation:SSKDE) を用いて，確率密度推定にもタグ無しサンプルが利用できるようにしている．

式 (19) によって与えられる条件付確率 $p(y_d|\boldsymbol{z}_n)$ は，特徴 $\boldsymbol{z}_n$ が与えられたとき，それが d 番目のクラスに属する事後確率としてみることもできる．この事後確率を用いてタグの付いていないサンプルにもタグを伝播していくというのが SSKDE の基本的なアイデアである．具体的には，これを行うために以下の行列表現を導入する．

$$\boldsymbol{P} = \{P_{n,m}\}_{n,m=1}^{N+N_x}, \boldsymbol{F} = \{F_{n,d}\}_{n=1,d=1}^{N+N_x,Dx} \tag{21}$$

$$P_{n,m} = \frac{\kappa(\boldsymbol{z}_n - \boldsymbol{z}_m)}{\sum_{m'=1}^{N+N_x}\kappa(\boldsymbol{z}_n - \boldsymbol{z}_{m'})} \tag{22}$$

$$F_{n,d} = p(y_d|\boldsymbol{z}_n) \tag{23}$$

ここで，$\boldsymbol{P}$ はサンプル間の類似度行列であり，$\boldsymbol{F}$ は全サンプルに対する全クラスの事後確率の行列表現である．求めるべきは $\boldsymbol{F}$ であるが，これは行列演算の繰り返しによって得ることができるため，効率的な計算が可能である（詳細は [14] を参照）．この SSKDE の手続きを簡単に表したものを図 6 に示す．

5. 評価実験

本節では，音響信号自動タグ付けに対して行った実験と，その結果に関して述べる．

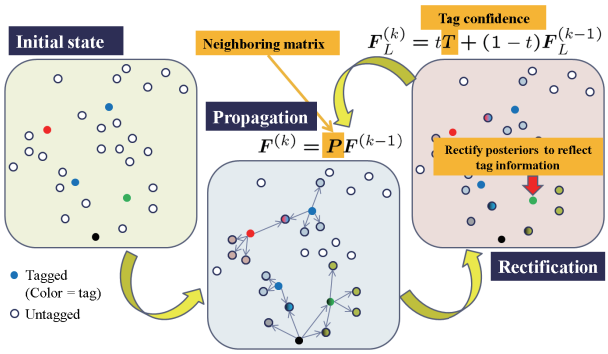

図 6　半教師型カーネル密度推定

本論文では，『Freesound』から得たタグ付き音響信号を用いた実験を行った．Freesound Project はクリエイティブ・コモンズ・ライセンスに基づき，音響信号を共有するプロジェクトである [19]．ただし，Freesound 上にある音響信号は様々なデータ形式，ビットレート，ビット深度を持つ音響信号が存在するため，今回はその中からデータ形式が WAV，ビットレート 44.1kHz，ビット深度が 16 のもののみに限定して使用することとした．また，ステレオの音響信号は左右のチャネルの平均をとることで全てモノラルに統一した．これにより 2012 個の音響信号が得られた．

Freesound 上ではユーザが音響信号をアップロードする際に任意のタグを付与することができる．したがって，Freesound 上の音響信号に付与されたタグの種類は膨大であり，中には一つの音響信号に対してしか付けられていないタグも多く存在する．そのため，今回は用いる音響信号に付与されているタグの中で，付与されている音響信号が多い方から 230 種類のタグを用いることとした．

第 3. 節で述べた特徴量を音響信号から抽出する際には，MFCC などを抽出するフレームの幅を約 23ms，フレームの移動幅をその半分とし，Δ パラメータは，前後 2 フレームから計算される回帰係数とした．また，コードブック作成のためにランダムサンプリングするベクトルは各音響信号から 500 個ずつ，最終的な特徴ベクトルの次元は $D_x = 1024$ とした．

実験では 2012 個の音響信号の中から，100 個を評価用に使用し，残りの 1912 個を学習用音響信号として使用した（$N + N_x = 1912$）．パラメータ D_z, β, γ, μ の値は経験的に決定し，それぞれ 100，0.99，0.8，0.99 とした．評価指標としては，正答率 (precision)，再現率 (recall)，及び F 値を用いた．

SSCDE を用いて行った実験では，学習用の音響信号 1912 個のうち 1000 個をタグ付き音響信号，残りの 912 個をタグ無し音響信号として使用した（$N = 1000, N_x = 912$）．SSCDE との比較のため，同数のタグ付き音響信号のみを使った CCA ベースの教師付き学習 ($N = 1000$) と，1912 個の学習用音響信号全てをタグ付きとして CCA ベースの教師付き学習（$N = 1912$），SemiCCA のみを用いた半教師付き学習（$N = 1000, N_x = 912$），さらに既存手法の階層的ガウス混合分布（H-GMM）を用いた教師付き学習（$N = 1912$) [7] によ

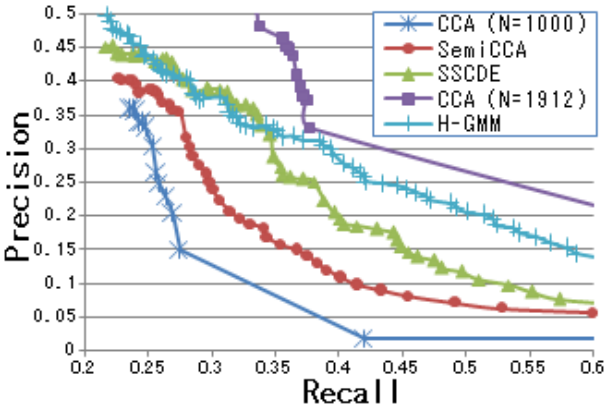

図 7　自動タグ付けの実験結果

表 1　Precision, Recall, F 値

	Precision	Recall	F 値
CCA (N=1000)	0.355	0.240	0.287
SSCDE (N=1000,N_x=912)	0.359	0.333	0.345
SemiCCA (N=1000,N_x=912)	0.355	0.274	0.310
CCA (N=1912)	0.462	0.356	0.402
H-GMM (N = 1912)	0.311	0.385	0.344

る自動タグ付けの実験も行った．閾値 θ の値を変えながら行った実験の結果を図 7 に示す．表 1 は各手法において，最も F 値が高かった点の precision と recall をまとめたものである．表からも分かる通り，SSCDE は CCA（N = 1000）に比べてタグ付け精度が大きく向上しており，CCA（N = 1912）に近づいていることが分かる．このことから，SSCDE はタグ無し音響信号を効果的に利用できていることが分かる．また，SSCDE で用いたタグ付きデータの数は H-GMM の半分程度であるにもかかわらず，SSCDE は H-GMM と同程度の精度を達成している．

6. ま と め

本論文ではトピックモデルを半教師型で効率的に学習する手法 SSCDE を提案し，それを音響信号の自動タグ付け・検索に適用した．SSCDE の特徴は，サンプルの大域的分布構造を考慮した半教師的な潜在変数空間の推定（SemiCCA）と，潜在空間内の非線形を考慮したノンパラメトリックな半教師型の確率密度推定（SSKDE）とを用い，トピックモデルの半教師学習を簡易に実現している点にある．さらに，約 2000 個の音響信号を用いた実験により本手法の有用性を示した．

さらなる研究課題としては，現在は ΔMFCC により局所的な音響信号の変化を表しているが，音響信号全体の変化の情報をも用いる，さらに性能を向上させることができると期待される．また，学習用のタグ付き，あるいはタグ無しの音響信号を新しく追加したり，新たな種類のタグを加えたときに，効率的にトピックモデルを更新するように拡張すること，なども今後の課題として挙げられる．

謝辞　本研究の実験用データの収集にご協力いただきましたアリゾナ州立大学 GordonWichern 博士に感謝します．また，実験用のプログラムに関して多大なる御協力をいただいた東京大学 中野拓帆氏にに感謝します．本研究に対し真摯にご議論いただき有益なご助言を頂いた NTT CS 研 坂野鋭博士，前田英作博士，石黒勝彦博士，柏野邦夫博士，永野秀尚博士に深謝します．本研究は第 1 著者が NTT コミュニケーション科学基礎研究所に実習生として在籍中に行ったものであり，本実習中に快適な研究環境を提供していただき，研究に対する多くの助言をいただいた NTT コミュニケーション科学基礎研究所メディア情報研究部 メディア認識研究グループの方々に深謝いたします．

文　献

[1] P.Knees et al., “Artist classification with Web-based data”, in Proc. WWW, 2000.
[2] W.W.Cohen, W.Fan, “Web-collaboratetive filtering: recommending music by crawling the Web”, Computer Networks 33, pp. 685-689 (2000).
[3] M.Slaney, ”Semantic–audio Retrieval”, in IEEE International Conference on Acoustics, Speech and Signal Processing, 2002.
[4] B.Whitman and R.Rifkin, “Musical query-by-description as a multiclass learning problem”, in Proc. IEEE Multimedia Signal Processing, pp.153-156 (2002).
[5] P.Knees et al. “A music search engine built upon audio-based and web-based similarity measures”, in Proc SIGIR, 2007.
[6] B.Whiteman and D.P.W.Ellis, “Automatic record revies”, in Proc. ISMIR, 2004.
[7] D.Torres et al. “Semantic annotation and retrieval of music and sound effects”, in IEEE Trans. on Audio, Speech, and Language Processing, vol. 16, No. 2, 2008.
[8] D.Torres et al. “Identifying words that are musically meaningful”, in Proc ISMIR 2007.
[9] L.Barrington et al. “Combining feature kernels for semantic music retrieval”, in Proc. ISMIR 2008.
[10] K.Barnard et al., “Matching words and pictures”, J.Mach.Learn.Res., vol. 3, pp. 1107-1135, 2003.
[11] Li Fei-Fei and P.Pietro, “A Baysian hierarchical model for learning natural scene categories”, in Proc. CVPR 2005.
[12] 木村昭悟，中野拓帆，杉山将，亀岡弘和，前田英作，坂野鋭，“SSCDE:画像認識検索のための半教師付き正準密度推定法”, 画像の認識・理解シンポジウム, MIRU 2010.
[13] A.Kimura et al., “SemiCCA: Efficient semi-supervised learning of canonical correlations”, in Proc. ICPR 2010.
[14] W.Wang et al. “Semi-supervised kernel density esimation for video annotation”, Computer Vision and Image Understanding, vol. 113, no.3, pp. 384-396, 2009.
[15] K.Kashino, T.Kurozumi, and H.Murase: ”A Quick Search Method for Audio and Video Signals Based on Histogram Pruning”, IEEE Transactions on Multimedia, vol.5, no.3, pp.348-357 (Sep. 2003).
[16] 柏野 邦夫, ガビン スミス, 村瀬 洋: ”ヒストグラム特徴を用いた音響信号の高速探索法 - 時系列アクティブ探索法 -”, 電子情報通信学会論文誌, vol.J82-D-II, no.9, pp.1365-1373 (Sep. 1999).
[17] H.Yanai and S.Puntanen, “Partial canonical correlation associated with the inverse and some generalized inverse of a partitioned dispersion matrix”, in Proc. the Third Pacific Area Statistical Conference on Statistical Sciences and Data Analysis, 1993, pp. 253-264.
[18] F.R.Bach and M.Jordan, “A probablistic interpretaion of canonical correlation analysis”, Tech. Reo. 688, Department of Statistics, University of Calfornia, Berkeley, 2005.
[19] The Freesound Project, “http://www.freesound.org/”.