DNN-GMM と連結特徴量を用いた音響シーン識別の検討 *

☆高橋玄，山田武志（筑波大），小野順貴（NII / 総研大），牧野昭二（筑波大）

1　はじめに

人間の行動や周囲の状況を自動認識しようとする取り組みがなされている．例えば，高齢者の見守りシステムや動画への自動タグ付け，ライフログの収集などの応用システムが考えられている．これらのシステムを実現する要素技術の一つとして，環境音認識が注目されている．環境音認識は大きく二つに分けられる．一つは音響イベント検出，もう一つは音響シーン識別である．音響イベント検出は「いつ，何の音がしたか」を検出するものであり，音の種類としてはドアの開閉音，咳をする音，マウスのクリック音などの短い単発的な音が多い．一方，音響シーン識別は数秒から数十秒程度の長い音から「どんな場所か，どんな状況か」を識別するものであり，音の種類としてはバスの中，公園，人込みなどがある．本研究では音響シーン識別に焦点を当てる．

近年，音響シーン識別の研究が盛んに行われている．例えば環境音認識に関するワークショップである DCASE 2013 （Detection and Classification of Acoustic Scenes and Events 2013）が開催され，そこでは音響シーン識別と音響イベント検出のタスクが用意された [1]．音響シーン識別タスクに対して提案された手法 [2][3] では，特徴量に MFCC（Mel frequency spectral coefficients）やメル周波数スペクトル，RQA（Recurrence Quantification Analysis）などが用いられ，また識別器には GMM（Gaussian Mixture Model）や SVM（Support Vector Machine）が用いられた．識別精度は最も高いもので 70%程度に留まった．

一方で近年，多層のニューラルネットワークである DNN（Deep Neural Network）が注目されている．一般的に，DNN は局所解に陥りがちであり，かつ非現実的な学習時間を要求する．しかしながら，適切な初期値を与える事前学習法 [4] と GPU（Graphics Processing Unit）での高速計算が確立された．その結果，DNN は現在様々な識別問題に適用されている．音声認識においては，DNN-HMM という DNN と HMM（Hidden Markov Model）を統合したものが提案されている [4]．HMM における確率分布は一般的に混合ガウス分布によって表現される．それに対して DNN-HMM では HMM における確率分布を DNN によってより精密に表現している．音声認識の性能は DNN-HMM を用いることによって著しく改善されることが報告されている [4].

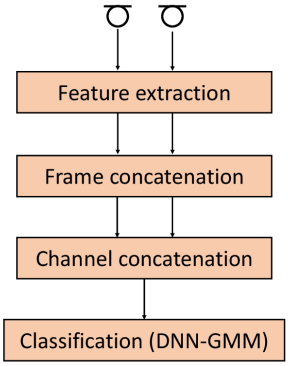

Fig. 1　Process flow of the proposed method.

そこで本稿では，DNN-GMM を用いた音響シーン識別手法を提案する [5]．ここで，DNN-GMM は，DNN-HMM における HMM をよりシンプルな GMM に置き換えたものである．提案手法で識別に用いる特徴量は，音声認識で広く採用されている MFCC とその一次差分，二次差分である．各時間フレームの特徴量には，時間的な関係性を捉えるために時間的に離れたフレームの特徴量を連結する．この特徴量を DNN-GMM に与えることによって音響シーン識別を行う．DCASE 2016 Challenge[6] における音響シーン識別のタスクを用いて提案手法の有効性を評価する．

2　提案手法

2.1　提案手法の処理の流れ

Fig. 1 は提案手法の処理の流れを示している．提案手法では 2 チャネルの入力を想定しており，まず左右チャネルそれぞれに対して各時間フレームにおける特徴量（MFCC とその一次差分，二次差分）を計算する．MFCC は人の聴覚特性を考慮した周波数特性の一表現であり，提案手法で用いている特徴量は DCASE 2016 のベースラインシステムと同じである．

* Acoustic scene classification using DNN-GMM and concatenated acoustic feature. by Gen TAKAHASHI, Takeshi YAMADA (University of Tsukuba), Nobutaka ONO (NII / SOKENDAI), Shoji MAKINO (University of Tsukuba)

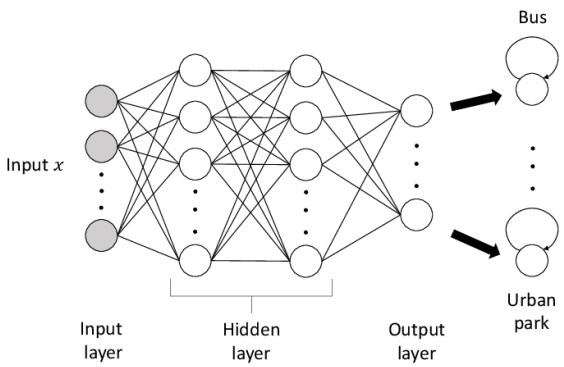

Fig. 2　Example of the DNN-GMM.

次に，左右チャネルそれぞれに対して，各フレームの特徴量に時間的に離れたフレームの特徴量を連結する．音響シーン識別の場合，音声認識に比べてより時間的に離れた音の関係性が識別に関わってくると考えられる．そのため，このような特徴量の連結を行うことによって識別精度の向上が見込めると考えられる．更に各フレームにおいて左右チャネルの特徴量を連結することで空間情報を持った特徴量とする．このようにして得られた高次元特徴量を DNN-GMM に与えることで，音響シーン識別を行う．以下では DNN-GMM について詳しく述べる．

2.2　DNN-GMM

Fig. 2 は DNN-GMM の例である．DNN-GMM は多層のニューラルネットワークと個々の音響シーンに対応する GMM からなる．GMM は直感的には 1 状態の HMM とみなすことができ，DNN-GMM では，DNN-HMM における HMM の代わりに GMM を用いている．

音響シーン識別は特徴量の時系列 $\mathbf{X}$ から対応する音響シーン $\hat{s}$ を識別する問題であり，以下のように表される．

$$\hat{s} = \underset{s_k}{\operatorname{argmax}}\, P(s_k|\mathbf{X}) \tag{1}$$

ここで，s_k は個々の音響シーンである．この式の $P(s_k|\mathbf{X})$ をベイズの定理を用いて変形することにより次式を得る．

$$P(s_k|\mathbf{X}) = \frac{P(\mathbf{X}|s_k)}{P(\mathbf{X})} P(s_k) \tag{2}$$

ここで，$P(\mathbf{X})$ は s_k に依らないため定数とみなすことができる．また各音響シーンの出現確率が一様であると仮定すると，$P(s_k)$ に関しても無視することができる．よって以下の式を解けばよいことになる．

$$\hat{s} = \underset{s_k}{\operatorname{argmax}}\, P(\mathbf{X}|s_k) \tag{3}$$

式 (3) を解くためのモデルとして音響シーン識別では GMM がよく用いられている．$\boldsymbol{x}_t$ を時間フレーム t における特徴量ベクトル，s_k^t を時間フレーム t における状態だとすると GMM を用いて $P(\mathbf{X}|s_k)$ は以下の式で求められる．

$$P(\mathbf{X}|s_k) = \prod_t P(\boldsymbol{x}_t|s_k^t) P(s_k^t|s_k^{t-1}) \tag{4}$$

これは音響シーン s_k の GMM から特徴量系列 $\mathbf{X}$ が生成される確率を表している．

出力確率 $P(\boldsymbol{x}|s_k)$ の真の分布を求めることは一般には難しいので，GMM では以下の混合ガウス分布により近似表現する．

$$p(\boldsymbol{x}|s_k) = \sum_{i=1}^{I} \pi_i N(\boldsymbol{x}|\mu_i, \Sigma_i) \tag{5}$$

ここで $N(\cdot)$ は正規分布であり，μ_i と Σ_i はそれぞれ正規分布の平均と分散である．また I は混合に用いる分布の数であり，π_i は分布ごとの重みである．この出力確率を，混合ガウス分布の代わりに DNN を用いて表現したものが DNN-GMM である．DNN で出力確率を表現するために式 (4) 中の $P(\boldsymbol{x}_t|s_k^t)$ をベイズの定理により以下のように変形する．

$$p(\boldsymbol{x}_t|s_k^t) = \frac{P(s_k^t|\boldsymbol{x}_t)}{P(s_k^t)} P(\boldsymbol{x}_t) \tag{6}$$

ここで $P(\boldsymbol{x}_t)$ は s_k に依らない値であるため定数とみなすことができる．また，各音響シーンの出現確率が一様であると仮定すると，$P(s_k^t)$ に関しても無視することができる．DNN の入力ベクトルを式 (6) 中の $\boldsymbol{x}_t$，出力層の各ノードを $P(s_k^t|\boldsymbol{x}_t)$ とすることで DNN と GMM を統合している．

本稿で採用した DNN-GMM の学習方法について述べる．まず従来の GMM を使って各音響シーンのモデルを教師あり学習する．次にその GMM を用いて，各学習データの各時間フレームにおける特徴量と GMM の状態番号からなる教師データを作成する．ただし，音響シーン識別の場合は 1 つの学習データが 1 つの音響シーンに対応するので，このステップは簡略化できる．そして，DNN の初期パラメータを教師なし学習する事前学習により決定する．事前学習では各層を RBM（Restricted Boltzmann Machine）とみなして CD（Contrastive Divergence）法 [7] を用いて学習を行う．最後に，教師データを用いて教師あり学習をする本学習を行う．本学習では乱数によって初期化したソフトマックス層をネットワークに追加し，確率的勾配降下法を用いた誤差逆伝播法を実行する．

Table 1　Overview of the development dataset and evaluation dataset.

# of scenes	15
# of sound data of development dataset	1170 (=15 scenes×78 data)
# of sound data of evaluation dataset	390 (=15 scenes×26 data)
data length	30 s
# of channels	2 (left and right)
sampling frequency	44.1 kHz
quantization bits	16 bit

3　提案手法の有効性の評価

3.1　実験条件

本章では，DCASE 2016 Challenge の開発用データセットと評価用データセットを用いて提案手法の有効性を評価する．Table 1 は開発用データセットと評価用データセットの概要を示している．音響シーンの数は 15 であり，例えばバスやレストラン，公園などがある．各音響シーンには開発用データセットにおいて 78 個，評価用データセットにおいて 26 個の音響データがあり，それぞれ 30 秒のステレオ信号である．評価用データセットは，開発用データセットとは異なるオープンな音響データである．サンプリング周波数は 44.1kHz で量子化ビットは 16 ビットである．

Table 2 は特徴量と DNN-GMM の条件を示している．特徴量は 20 次元 MFCC とその一次差分，二次差分の計 60 次元をベースとする．フレーム分析におけるフレーム長とフレーム周期はそれぞれ 40ms と 20ms である．これは DCASE 2016 Challenge のベースラインで用いられているものと同じである．時間的に離れた n フレームの連結，及び左右 2 チャネルの連結によって特徴量の次元は $60 \times n$ (フレーム) $\times 2$ (チャネル) になる．フレーム連結に用いるフレームは，m ms 間隔で選択する．この実験では，フレーム連結数 n は 1，3，5，7，フレーム連結間隔 m は 20，100，200，500，1000，2000 ms で変化させ，それぞれの効果を比較する．

DNN の隠れ層の数（入力層と出力層を除く）は 2，3，4，5，隠れ層の次元は 256，512，1024，2048 である．各隠れ層の次元は同じとする．隠れ層の RBM では CD-1 法を用いて，入力層の RBM では CD-2 法を用いて事前学習を行った．また予備実験の結果から RBM の学習率を 0.4，DNN の学習率を 0.008，ドロップアウト率を 0.0 に設定した．識別器を学習する際には，まず開発用データセットで 4-fold クロスバリデーション（データオープン，音響シーンクローズド）を行い，DNN の層数と隠れ層の次元を最適化する．次に，最適化した隠れ層の数と隠れ層の次元において，評価用の識別器を学習する．識別器の学習には開発用データセットにおける音響データを全て用いた．学習した識別器と評価用データセットを用いて，提案手法の有効性を評価する．なお，提案手法の実装には Kaldi[8] を用いた．

Table 2　Condition of the acoustic features and the DNN-GMM.

feature	20 dimensions MFCC $+\Delta+\Delta\Delta$ (60 dimensions) $\times n$ frame $\times 2$ channel
frame length	40 ms
frame period	20 ms
# of concatenated frames n	1, 3, 5, 7
frame concatenation interval m (ms)	20, 100, 200, 500, 1000, 2000
# of hidden layers	2, 3, 4, 5
dimension of hidden layer	256, 512, 1024, 2048
dimension of input layer	$120\times n$
dimension of output layer	15

3.2　実験結果

Table 3 は実験の結果を示している．表の列はフレーム連結数 n，行はフレーム連結間隔 m，各値は隠れ層の数と隠れ層の次元を最適化して得られた識別精度である．ここで，括弧書きされているのは最適化した隠れ層の数（左）と隠れ層の次元（右）である．提案手法の識別精度は n と m の値によって変化しており，$n=3$, $m=20$ のときに最も高く，$n=1$ のときと比較すると 2.82%改善している．このときの識別精度は 86.67%であり，これは隠れ層数が 2，各層の次元が 2048 の DNN-GMM によって達成された．ま

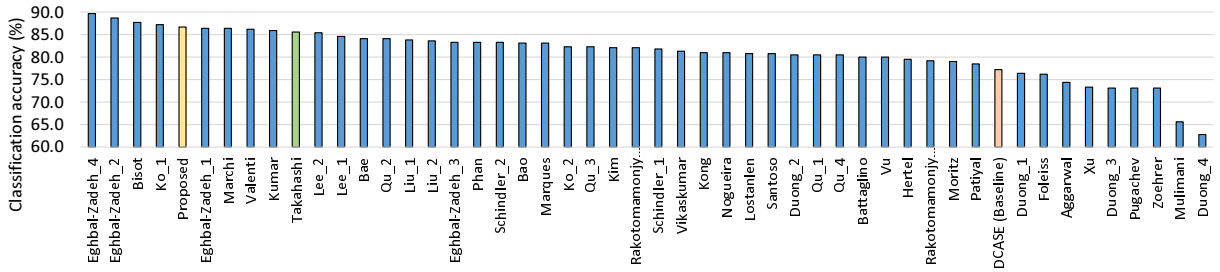

Fig. 3　Results of DCASE 2016 Challenge.

Table 3　Classification accuracy for each combination of n and m.

m \ n	1	3	5	7
20 ms	83.85 (3, 1024)	86.67 (2, 2048)	85.64 (2, 1024)	85.38 (3, 1024)
100 ms	–	85.13 (2, 1024)	86.41 (2, 2048)	83.59 (3, 512)
200 ms	–	85.38 (2, 512)	85.38 (2, 2048)	84.87 (3, 2048)
500 ms	–	85.64 (3, 2048)	85.38 (2, 2048)	84.36 (2, 1024)
1000 ms	–	85.64 (3, 512)	82.82 (2, 256)	84.62 (2, 1024)
2000 ms	–	84.10 (2, 512)	83.85 (2, 1024)	81.54 (2, 256)

た，表中の色が濃いほど識別精度が高いことを表しており，n が 5 以下，m が 500ms 以下のときに，比較的安定して高い識別精度が得られている．一方，n あるいは m を大きくしすぎると識別精度が急激に低下することが分かる．

次に，DCASE 2016 Challenge の結果 [9] を Fig. 3 に示す．本稿の提案手法（$n = 3, m = 20$）は黄色，DCASE 2016 Challenge のベースライン（GMM）はオレンジである．ベースラインと比較すると，提案手法（$n = 3, m = 20$）では識別精度が 9.47%改善した．その改善の内訳を述べる．まず，提案手法の $n = 1$（フレーム連結なし）ではベースラインと比較して識別精度が 6.65%改善した．これは GMM を DNN-GMM に変更したことによる効果である．更に，提案手法（$n = 1$）と比べて，提案手法（$n = 3, m = 20$）では識別精度が 2.82%改善した．これは連結特徴量を用いたことの効果である．また，著者らが DCASE 2016 Challenge に提出した結果は緑色で示している．これは，提案手法（$n = 5, m = 100$）に相当し，特徴量が非可逆圧縮されている点のみが異なる．これと比べると，提案手法（$n = 3, m = 20$）は識別精度が 1.07%改善し，49 種類のアルゴリズムの中で 5 位になった．

4　おわりに

本稿では DNN-GMM と連結特徴量を用いた音響シーン識別手法を提案した．DCASE 2016 Challenge の開発用データセットと評価用データセットを用いて実験を行い，提案手法の有効性を確認した．提案手法におけるフレーム連結数 n を 5 以下，フレーム連結間隔 m を 500ms 以下に設定することで，比較的安定して高い識別精度が得られることが分かった．

謝辞　本研究は国立情報学研究所自由提案公募型共同研究の助成を受けた．

参考文献

[1] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, M. Plumbley, “Detection and Classification of Acoustic Scenes and Events: An IEEE AASP Challenge,” Proc. WASPAA 2013, Oct. 2013.

[2] G. Roma, W. Nogueira, P. Herrera, “Recurrence quantification analysis features for auditory scene classification,” Proc. WASPAA 2013, Oct. 2013.

[3] J. Nam Z. Hyung, K. Lee, “Acoustic scene classification using sparse feature learning and selective max-pool by event detection,” Proc. WASPAA 2013, Oct. 2013.

[4] 久保陽太郎, “ディープラーニングによるパターン認識,” 情報処理学会, Vol. 54, No. 5, May 2013.

[5] G. Takahashi, T. Yamada, S. Makino, N. Ono, “Acoustic scene classification using deep neural network and frame-concatenated acoustic feature,” DCASE 2016 Challenge Technical Report, Sep. 2016, http://www.cs.tut.fi/sgn/arg/dcase2016/documents/challenge_technical_reports/Task1/Takahashi_2016_task1.pdf.

[6] http://www.cs.tut.fi/sgn/arg/dcase2016/.

[7] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” Neural Computation, Vol. 14(8), pp. 1771-1800, 2002.

[8] http://kaldi-asr.org/.

[9] http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-acoustic-scene-classification.