Vol.2016-SLP-110 No.2
2016/2/5

混合コーシー２乗分布を用いた歌唱者識別

伊藤 彰則[1,a)]

概要：混合ガウス分布による確率モデル (GMM) は，様々な認識タスクに広く用いられている．GMM の確率計算を定義通りに行うためには，指数関数の計算が必要になり，小型のデバイスなどで計算を行うには計算量が多い．計算量を削減するため，addlog 計算が広く用いられている．このような計算法は，認識対象の分布をガウス分布の混合で表現するという定義通りに計算を行うための工夫であると言える．これに対し本稿では，そもそも対象の確率分布をもっと計算の簡単な分布で近似することにより，計算量を低減する方法を提案する．提案法では，コーシー分布から導出される「コーシー 2 乗分布」を利用する．この分布は四則演算だけで計算することが可能であり，確率を混合する場合にも addlog など特別な計算を必要としない．16 名の歌唱音声の歌唱者識別を対象として実験を行ったところ，速度は addlog を使った GMM の 1.2 倍高速であり，認識性能は GMM よりも高かった．

1. はじめに

混合ガウス分布による確率モデル (Gaussian Mixture Model)[1] は，あるクラスの事象の生起確率を混合ガウス分布で近似するモデルであり，対象の識別や変換などさまざまなタスクに広く利用されている．本稿では，Internet of Things (IoT)[2] などに代表される，環境に埋め込まれた多量のセンサ情報を利用した情報処理について検討する．

このようなセンサネットワークで利用できる通信の帯域は大きくないため，センサが取得した信号を分析するにあたり，すべての信号をサーバに集約して分析を行う方法は，多くの通信帯域を必要とするため望ましくない．帯域を節約するためには，各センサが信号の分析を行い，その結果のみをサーバに送信する方法が望ましい [3]．一方，センサが信号の分析をするためには，センサ自身が計算をする必要があり，多くの電力が必要となる．そのため，信号の分析をするにあたり，演算量の小さい分析手法が望ましい．

GMM は比較的演算量が少なく，高い識別性能が得られるため，音の認識 [4], [5], [6] だけでなく，電気機器の認識 [7]，映像認識 [8]，ウェアラブルデバイスでの行動認識 [9] などにも広く用いられている．

GMM の計算を高速化する手法はこれまで多く提案されており，ガウス分布の離散テーブル化，addlog 計算 [10] などがその代表である．これらの計算法は，いずれも分布がガウス分布の混合であることを前提としている．

本研究では，対象の確率密度をもっと演算量の低い確率分布で置き換える方法を検討した．この時の確率分布として，コーシー分布を元にして平均・分散が定義できるようにした「コーシー２乗分布」を利用する．この密度関数は，四則演算のみで計算することができ，また多次元ベクトルに対する対角共分散型の確率密度を高速に計算することができる．本稿では，安価な機器による歌唱者の識別をターゲットに，GMM と混合コーシー２乗分布による識別を行い，識別精度と計算速度を比較する，

2. GMM とその高速計算

2.1 GMM

入力ベクトルを $\boldsymbol{x} = (x_1, \ldots, x_D)$ とするとき，多次元ガウス分布による確率密度関数は

$$N(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}\right)}{(2\pi)^{\frac{D}{2}} \sqrt{|\Sigma|}} \qquad (1)$$

と表される，ここで $\boldsymbol{\mu}$ は平均ベクトル，Σ は共分散行列である．共分散行列が対角行列である場合，平均ベクトルの k 番目の要素を μ_k，共分散行列の k 番目の対角要素を σ_k^2 とすれば，上記の分布は

$$N(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \prod_{k=1}^{D} \frac{\exp\left(-\frac{(x_k-\mu_k)^2}{2\sigma_k^2}\right)}{\sqrt{2\pi\sigma_k^2}} \qquad (2)$$

と表せる．

混合ガウス分布は，ガウス分布を複数重み付き加算してできる分布であり，多数の分布を加算すればさまざまな分

[1] 東北大学 大学院工学研究科
Grad. Sch. Eng., Tohoku University, Sendai 980–8579, Japan
[a)] aito@spcom.ecei.tohoku.ac.jp

布を近似することができる．混合ガウス分布による確率密度関数は，混合分布数を M とすると

$$p(\boldsymbol{x}) = \sum_{i=1}^{M} \lambda_i N(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i) \tag{3}$$

と表される．ここで λ_i は重みであり，$0 \leq \lambda_i \leq 1$ かつ

$$\sum_{i=1}^{M} \lambda_i = 1 \tag{4}$$

を満たす．

GMM のパラメータであるガウス分布の重み，平均，分散の推定には通常 EM アルゴリズムが用いられる [1]．分布の数は通常は事前に与える必要があるが，ベイズ理論に基づいて自動的に分布数を決定する手法も提案されている [11]．

2.2 Addlog 計算

GMM の高速計算法として addlog[10] が知られている．式 (2) の計算において，その対数は

$$\log N(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = -\sum_{k=1}^{D} \frac{(x_k - \mu_k)^2}{2\sigma_k^2} + C \tag{5}$$

$$C = -\sum_{k=1}^{D} \frac{1}{2} \log(2\pi\sigma_k^2) \tag{6}$$

となり，C は事前に計算しておけるため，実際の確率計算時には四則演算のみで対数確率を求めることができる．一方，式 (3) は真数領域での加算であるため，単純に計算するとすれば，対数確率をいったん指数関数によって真数領域に戻してから加算する必要がある．これを避けるため，addlog が使われる．

addlog は

$$\mathrm{addlog}(x, y) = \log(\exp(x) + \exp(y)) \tag{7}$$

を計算するための方法である．$x \geq y$ と仮定すると，

$$\begin{aligned} \mathrm{addlog}(x, y) &= \log(\exp(x) + \exp(y)) && (8)\\ &= \log(\exp(x)(1 + \exp(y-x)) && (9)\\ &= x + \log(1 + \exp(y-x)) && (10) \end{aligned}$$

と表せる．そこで，$x \leq 0$ に対して $\log(1 + \exp(x))$ の表を作っておき，この部分は表の参照だけで計算を行う．$\log(1 + \exp(x))$ のグラフを図 1 に示す．この図からわかるように，$|x|$ が大きくなると $\log(1 + \exp(x))$ の値は急速に小さくなるので，適当なところで計算を打ち切って 0 とすることができる．3 つ以上の数に関しては

$$\begin{aligned} &\mathrm{addlog}(x_1, x_2, \ldots, x_n) = \\ &\quad \mathrm{addlog}(\mathrm{addlog}(x_1, \ldots, x_{n-1}), x_n) \end{aligned} \tag{11}$$

により計算を行うことができる．最終的な混合分布の対数確率は

$$\log p(\boldsymbol{x}) = \underset{i=1,\ldots,M}{\mathrm{addlog}} \left(-\sum_{k=1}^{D} \frac{(x_k - \mu_{ik})^2}{2\sigma_{ik}^2} + C_i \right) \tag{12}$$

$$C_i = \log \lambda_i - \sum_{k=1}^{D} \frac{1}{2} \log(2\pi\sigma_k^2) \tag{13}$$

このような計算を行えば，式 (3) の計算を四則演算と表の参照だけで行うことができる．

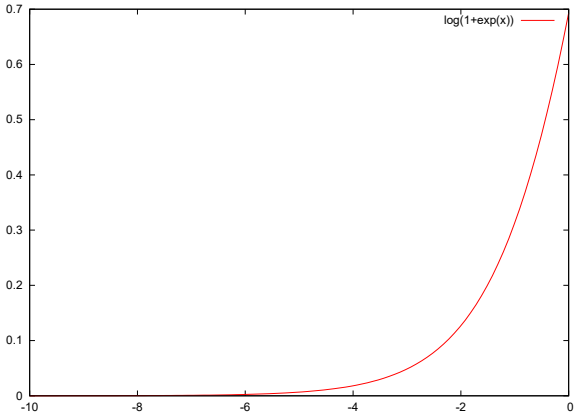

図 1 $\log(1 + \exp(x))$ for $x < 0$

3. コーシー 2 乗分布

コーシー分布は 1 次元の確率分布である．その確率密度関数は

$$p(x) = \frac{b}{\pi(b^2 + (x-a)^2)} \tag{14}$$

と表され，左右対称で裾の長い確率分布である．a は分布の中心，b は分布の広がりを表すパラメータであるが，コーシー分布には平均と分散は定義されない（それぞれを求める積分が発散する）ので，これらのパラメータは平均と分散ではない．コーシー分布のパラメータを推定することは理論的には簡単ではなく，いくつかの手法が提案されている [12], [13]．

コーシー分布は四則演算だけで計算できるため，これでデータの分布を近似すれば高速に分布が計算できる．しかし，コーシー分布には平均値がないため，コーシー分布に従うデータはいくら平均しても平均値が一定値に収束しないはずである．逆に言えば，多数の観測値を平均すると平均値が収束するデータはコーシー分布に従っていない．したがって，コーシー分布でそのようなデータの分布を近似するのは適切とは言えない．

そこで，平均と分散を持ち，四則演算のみで計算できる分布として，コーシー 2 乗分布を考案した．コーシー 2 乗分布の確率密度関数は次の式で表される．

$$p(x) = \frac{2\sigma^3}{\pi(\sigma^2 + (x-\mu)^2)^2} \tag{15}$$

この分布の平均は μ，分散は σ^2 である．3 次以上のモー

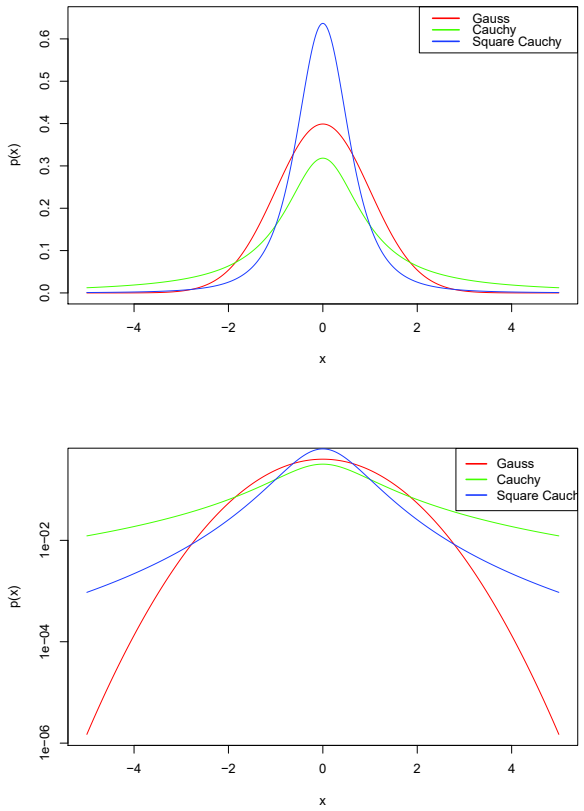

図 2 Probability density functions of Gaussian, Cauchy and Square Cauchy distributions (upper: anti-log domain, lower: log domain)

メントは定義されない．

ガウス分布，コーシー分布，コーシー 2 乗分布の確率密度関数の比較を図 2 に示す．コーシー 2 乗分布はガウス分布よりも中心付近の確率密度が高く，また裾が長いことがわかる．

多次元分布の場合，各次元が独立（多次元ガウス分布でいう対角共分散に相当）ならば分布関数は次のように表される．

$$p(\boldsymbol{x}) = \prod_{k=1}^{D} \frac{2\sigma_k^3}{\pi(\sigma_k^2 + (x_k - \mu_k)^2)^2} \tag{16}$$

ここで

$$K = \prod_{k=1}^{D} \frac{2\sigma_k^3}{\pi} \tag{17}$$

とおけば，

$$p(\boldsymbol{x}) = \frac{K}{\left(\prod_{k=1}^{D}(\sigma_k^2 + (x_k - \mu_k)^2)\right)^2} \tag{18}$$

となり，乗算 $2D$ 回，加減算 $2D$ 回，除算 1 回で確率計算が可能になる．この確率は真数領域で計算されているので，混合分布を計算する際には特別な演算は必要なく，最終的に対数確率を求める際には対数演算が 1 回必要になる．混合分布は次のように表される．

表 1 Comparison of floating point operations of the three distributions

分布	Gauss	Gauss(addlog)	Square Cauchy
加減算	$MD + M - 1$	$2MD$	$2MD + M - 1$
比較演算	0	$M - 1$	0
表参照	0	$\leq M - 1$	0
乗除算	$2MD$	$2MD$	$2MD$
指数関数	MD	0	0
対数関数	1	0	1

$$p(\boldsymbol{x}) = \sum_{i=1}^{M} \frac{K_i}{\left(\prod_{k=1}^{D}(\sigma_{ik}^2 + (x_k - \mu_{ik})^2)\right)^2} \tag{19}$$

$$K_i = \lambda_i \prod_{k=1}^{D} \frac{2\sigma_{ik}^3}{\pi} \tag{20}$$

混合ガウス分布，addlog を使った混合ガウス分布および混合コーシー 2 乗分布での演算回数を表 1 に示す．M は混合分布数，D は特徴量の次元数である．混合コーシー 2 乗分布では，addlog による混合ガウス分布と比較して，addlog 計算時に必要になる表参照が省略できることがわかる．

4. 実験 1

4.1 実験概要

GMM とコーシー 2 乗分布の識別性能および計算速度を比較するため，歌唱者識別の実験を行った．実験の目的は GMM と混合コーシー 2 乗分布モデルの性能を比較することであり，高性能な歌唱者識別を行うことではないので，手法としては単純な識別手法を利用している [14]．実験手順は以下のとおりである．

(1) 学習データを用意し，全学習データを使って GMM による UBM を学習する．
(2) 評価する歌唱者の歌声データを用意し，UBM を初期値として各歌唱者の GMM を学習する．通常は MAP 適応を使うべきところだが，ここでは EM 学習を行っている．
(3) 入力歌唱データの各フレームに対して GMM，混合コーシー 2 乗分布モデルおよび混合コーシー分布モデルで各歌唱者モデルによる確率を計算し，対数確率の全フレームの総和が最大となる歌唱者を識別結果とする．

混合コーシー 2 乗分布および混合コーシー分布では，混合ガウス分布の重み・平均・分散をそのまま利用した．しかし，コーシー分布やコーシー 2 乗分布は分布形状がガウス分布と異なるため，推定した分散の値がガウス分布と同じになる保証はない．そこで，ガウス分布の分散に係数をかけて性能を比較する実験も行う．

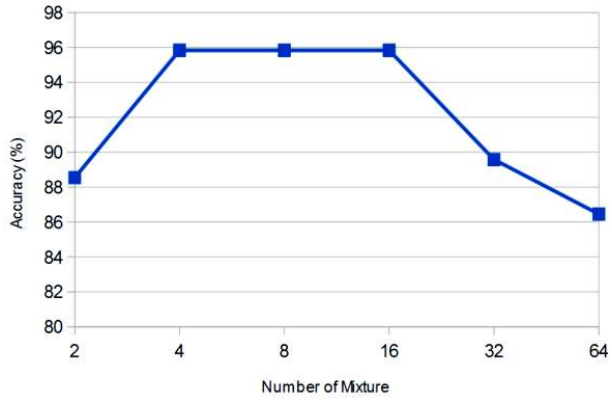

図 3　Recognition result using GMM

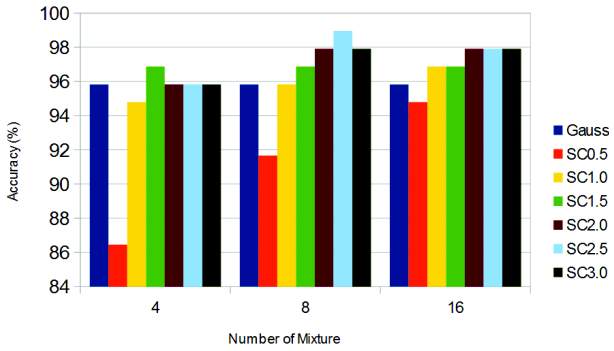

図 4　Recognition result using Squared Cauchy with different magnification of variance

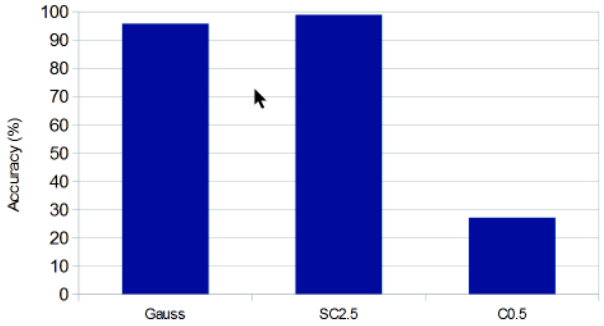

図 5　Maximum accuracy by GMM, Square Cauchy mixture and Cauchy mixture models

4.2　コーパスと分析方法

実験に利用したコーパスは，素人による歌唱であり，複数の歌唱者が「通常」と「熱唱」の 2 つの歌唱スタイルで歌った歌唱音声が収録されている [15]．UBM の学習には，11 名の男性歌唱者がそれぞれ異なる曲を歌唱したデータを用いた．このデータではそれぞれの歌唱者が同じ曲を 2 回歌唱しており，それぞれの曲から異なるフレーズ 4 か所を抽出している．各フレーズの平均長は 4.45 秒である．適応と識別実験には，UBM の学習データの歌唱者を含む男性 16 名の歌唱を用いた．この歌唱は，楽曲「いとしのエリー」を歌唱したデータの一部であり，この曲は UBM の学習データには含まれていない．各歌唱者が同楽曲をそれぞれ 2 回歌唱しており，1 曲の中に 4 回現れる「エリー」の歌唱部分を抽出している．データの平均長は 2.1 秒である．一人につき 8 つの歌唱データがあるうち，2 回の歌唱から 1 回ずつ，計 2 つのデータを適応用とし，それ以外の 6 つのデータを評価用とした．適応と評価に用いた歌唱音声は同じ歌詞・メロディなので，話者認識で言えばテキスト依存の認識と同じ枠組みである．

特徴量は MFCC12 次元と対数パワーおよび Δ，Δ Δ 特徴量を合わせた 39 次元である．分析窓長は 40ms，フレームシフトは 10ms，窓関数はハミング窓である．特徴量計算および GMM 学習には SPTK を用いた．

4.3　識別精度

まず最初に通常の GMM を用いて識別実験を行った．混合数を変化させて識別を行った時の結果を図 3 に示す．横軸は混合分布数，縦軸は認識率である．単純なタスクなので識別精度は全体に高く，混合数 4 から 16 の場合に性能が頭打ちになる．

次に，混合コーシー 2 乗分布モデルによる識別を行った．前述のとおり，GMM の学習で得られた分散に係数をかけ，それを変化させて識別性能の推移を確かめた．実験結果を Fig. 4 に示す．横軸は混合分布数，縦軸は認識率である．図中の “SC1.0” などの表記は混合コーシー 2 乗分布における分散の倍率を示しており，倍率を 0.5 から 3.0 まで変化させている．図からわかるとおり，倍率が 1.0 以上の場合に良好な認識性能が得られており，混合数 8 で倍率 2.5 の場合には GMM よりも 3 ポイント高い認識性能であった．

図 5 は，混合数 8 における GMM，混合コーシー 2 乗分布，混合コーシー分布での認識結果である．混合コーシー 2 乗分布と混合コーシー分布では，認識率が最大になる分散の倍率を事後的に選んでいる．この結果から，混合コーシー分布ではほとんど認識ができていないことがわかる．なお，これは GMM で学習した平均・分散をそのまま適用した場合の結果なので，混合コーシー分布であっても，パラメータを適切に設定することができれば，もっと高い認識性能が得られると思われる．

最後に計算時間を比較した．計測したのはすべての評価データの確率計算に要した CPU 時間である．入出力および特徴量変換の時間は含まれていない．利用した計算機は，CPU が Intel Xeon L5640 2.2GHz 2CPU 12core で，搭載メモリは 64GByte，OS は Ubuntu Linux である．CPU 時間の計測にあたっては，実験を 10 回繰り返し，得られた CPU 時間の平均値を用いた．

計算結果を Fig. 6 に示す．縦軸は CPU 時間（秒），横軸は混合数である．計算方法は，そのまま計算した GMM，addlog 計算を使った GMM およびコーシー 2 乗分布による計算である．GMM に addlog 計算を使うことにより，確率計算は平均して 7.4 倍高速になるが，コーシー 2 乗分布

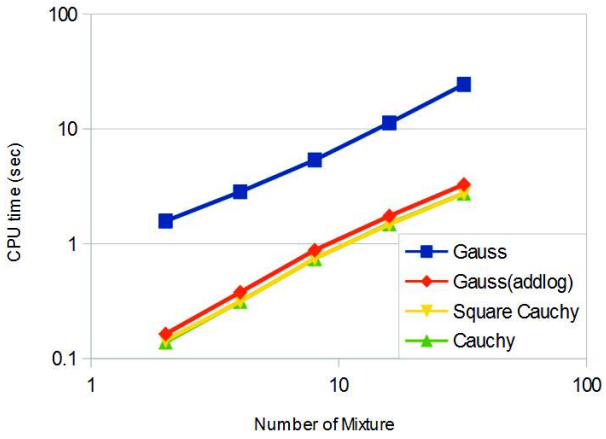

図 6 Comparison of CPU time

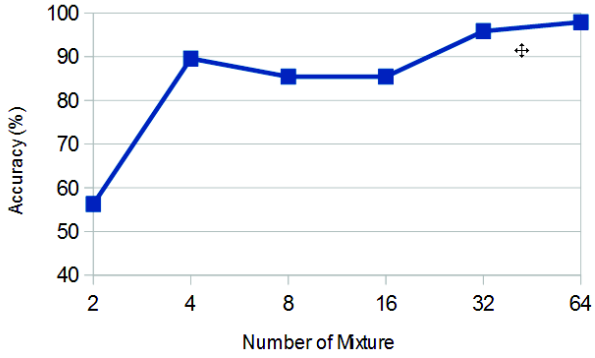

図 7 Recognition result using GMM (phrase independent)

を利用することでさらに約 1.2 倍高速に計算できるようになった．

5. 実験 2

前節での実験はテキスト（フレーズ）依存の認識であったが，次にフレーズ独立の認識で性能を比較した．実験の枠組みは前節とほぼ同じであり，適応のデータのみが異なる．実験 2 において適応に用いたデータは，「いとしのエリー」から評価データと同じ歌詞・メロディーを含まない歌唱であり，1 歌唱者当たり 2 回歌唱したそれぞれの歌唱音声から 4 カ所ずつ，一人当たり 8 つのフレーズである．1 フレーズ当たりの長さは平均 5.8 秒であり，適応用音声としては一人当たり 46.7 秒になる．

GMM を用いた識別結果を図 7 に示す．前節よりも難しいタスクなので，混合数が 32〜64 で認識性能が高くなる．これに対し，混合コーシー 2 乗分布による結果と比較したものを図 8 に示す．分散の倍率が 1.5〜2.0 の時，認識性能は GM よりも高くなる．

なお，評価データは前節の実験と同じなので，計算にかかる時間は図 6 と全く同じである．

6. まとめ

高速に確率密度分布を計算する方法として，コーシー 2 乗分布を利用する方法を提案した．コーシー 2 乗分布は単純な四則演算のみで計算でき，確率を混合する際にも特別な計算を必要としない．歌唱者識別タスクでは，提案法は従来の GMM よりも性能が高く，計算速度は addlog 計算の 1.2 倍であった．

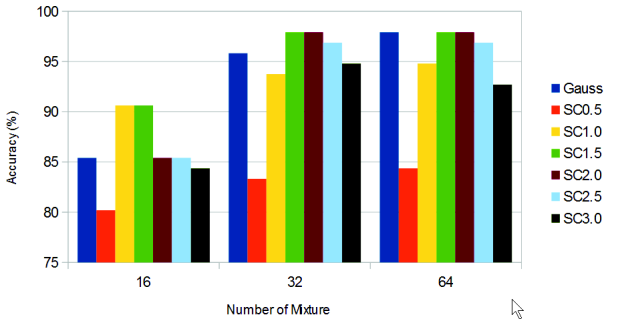

図 8 Recognition result using Squared Cauchy with different magnification of variance (phrase independent)

今回は GMM の学習で得られた平均と分散をそのまま用いてコーシー 2 乗分布を計算したが，本来はコーシー 2 乗分布自体を用いた学習が望ましい．これにより最適な学習結果が得られると同時に，学習に要する計算量が低減できると期待される．今後は学習法の開発が必要である．

また，今回は識別が容易なタスクに提案法を適用したが，音声認識など，より識別が難しいタスクに提案法を適用した場合にも GMM と同等な結果が得られるのかどうかは興味深い問題である．今後研究を進めていきたい．

参考文献

[1] Redner, R. A. and Walker, H. F.: Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, Vol. 26, No. 2, pp. 195–239 (1984).

[2] Gubbi, J., Buyya, R., Marusic, S. and Palaniswami, M.: Internet of Things (IoT): A vision, architectural elements, and future directions, *Future Generation Computer Systems*, Vol. 29, No. 7, pp. 1645 – 1660 (online), DOI: http://dx.doi.org/10.1016/j.future.2013.01.010 (2013).

[3] Shen, C., Choi, H., Chakraborty, S. and Srivastava, M.: Towards a rich sensing stack for IoT devices, *Proc. IEEE/ACM Int. Conf. on Computer-Aided Design (ICCAD)*, pp. 424–427 (online), DOI: 10.1109/ICCAD.2014.7001386 (2014).

[4] Cowling, M. and Sitte, R.: Comparison of techniques for environmental sound recognition, *Pattern Recognition Letters*, Vol. 24, No. 15, pp. 2895 – 2907 (online), DOI: http://dx.doi.org/10.1016/S0167-8655(03)00147-8 (2003).

[5] Chu, S., Narayanan, S. and Kuo, C.-C.: Environmental Sound Recognition With Time-Frequency Audio Features, *IEEE Transactions on Audio, Speech, and Language Processing,*, Vol. 17, No. 6, pp. 1142–1158 (online), DOI: 10.1109/TASL.2009.2017438 (2009).

[6] Ntalampiras, S., Potamitis, I. and Fakotakis, N.: Automatic Recognition of Urban Soundscenes, *New Directions in Intelligent Interactive Multimedia*, Vol. 142, Springer Berlin Heidelberg, pp. 147–153 (2008).

[7] Lai, Y.-X., Lai, C.-F., Huang, Y.-M. and Chao, H.-C.: Multi-appliance recognition system with hybrid SVM/GMM classifier in ubiquitous smart home, *Information Sciences*, Vol. 230, pp. 39–55 (2013).
[8] Ou, S.-H., Lee, C.-H., Somayazulu, V. S., Chen, Y.-K. and Chien, S.-Y.: Video Sensor Node with Dis1t ributed Video Summary for Internet-of-Things Applications, *Proc. Int. Conf. on Consumer Electronics-Taiwan* (2015).
[9] Meharia, P. and Agrawal, D. P.: The able amble: gait recognition using Gaussian mixture model for biometric applications, *Proc. the 12th ACM Int. Conf. on Computing Frontiers* (2015).
[10] Sagayama, S. and Takahashi, S.: On the use of scalar quantization for fast HMM computation, *ICASSP-95*, Vol. 1, pp. 213–216 vol.1 (online), DOI: 10.1109/ICASSP.1995.479402 (1995).
[11] Roberts, S., Husmeier, D., Rezek, I. and Penny, W.: Bayesian approaches to Gaussian mixture modeling, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 20, No. 11, pp. 1133–1142 (online), DOI: 10.1109/34.730550 (1998).
[12] Rothenberg, T. J., Fisher, F. M. and Tilanus, C. B.: A Note on Estimation from a Cauchy Sample, *Journal of the American Statistical Association*, Vol. 59, No. 306, pp. 460–463 (1964).
[13] Freue, G. V. C.: The Pitman estimator of the Cauchy location parameter, *Journal of Statistical Planning and Inference*, Vol. 137, pp. 1900–1913 (2007).
[14] Reynolds, D. and Rose, R.: Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE Trans. on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72–83 (online), DOI: 10.1109/89.365379 (1995).
[15] Daido, R., Ito, M., Makino, S. and Ito, A.: Automatic evaluation of singing enthusiasm for karaoke, *Computer Speech & Language*, Vol. 28, No. 2, pp. 501–517 (2014).