

セグメントレベル特徴量を用いた楽曲のジャンル分類

赤羽 慎^{1,a)} 藤田 侑介^{2,b)} 王 龍標^{2,c)} 甲斐 充彦^{1,d)}

概要: 近年 IT 技術の発展により, デジタル音楽は様々なメディアから入手可能となった. その音楽の量は膨大であるため, 何らかの手がかりにより効率的に音楽にアクセスすることが必要である. それは音楽情報検索 (MIR) により実現することができる. 本研究では MIR のタスクの一つである楽曲のジャンル分類に関して取り扱う. ジャンル分類は, デジタル化された楽曲データの音情報から特徴量を抽出し, あらかじめジャンルのタグ付けされた事例データから識別器を構築することで実現できる. 音声認識において有効な特徴量は通常, 10-100[ms] 程のフレームから抽出されるため, 音楽の短期的な性質しか捉えることができない. 音楽に使われている楽器などの情報を得るためには, 音楽をもっと長期的なスパンで捉えることが必要である. 先行研究 [1], [3] では, 統計量に基づく楽曲レベル特徴量を用いることで音楽を長期的に捉える事を試みている. 本研究では, フレームより長い楽曲より短いセグメントレベルの特徴量を用いた楽曲ジャンル推定方法を提案する.

キーワード: 音楽情報検索, ジャンル分類, 楽曲レベル特徴量, セグメントレベル特徴量

1. はじめに

ジャンル分類はクラシック, ロックなどの楽曲のジャンルにより音楽进行分类することであり, MIR のタスクの一つである. ジャンル分類は, デジタル化された楽曲データの音情報から特徴量を抽出し, 一般的なパターン認識の問題と同様に, あらかじめジャンルのタグ付けされた事例データから識別器を構築することで実現できる.

音声認識で用いられる特徴量は, 10-100[ms] 程のフレーム (音楽のスライス) から抽出されるため, 音楽の短期的な性質しか捉えることができない. これに対して, 音楽に使われている楽器や旋律などの情報を得るためには, 音楽のもっと長期的な性質を捉えることが重要であると考えられる.

文献 [2] によれば, 先行研究 [1] では, フレームレベルで抽出した特徴量に対して, 隣接する複数フレーム単位で統計量を計算する $MuVar$ という手法を用いることで音楽を長期的に捉える事を試みている. また [3] では, $MuVar$ により得られた統計量に対して, 更に統計量を計算する

$MuVar^2$ という手法を用いることで 1 楽曲を 1 つの特徴 (統計) ベクトルで表すことを試みている.

2 章ではベースラインジャンル分類実験として, $MuVar^2$ を用いた楽曲レベルの特徴量による識別実験について述べる. 続いて 3 章では提案手法によるジャンル分類実験として, $MuVar_N^2$ を用いたセグメントレベルの特徴量による識別実験について述べる.

2. ベースラインジャンル分類

本章ではベースラインシステムの概要, 統計量に基づく特徴量 (手法: $MuVar$, $MuVar^2$), $MuVar^2$ を用いた楽曲レベルの特徴量による識別実験について述べる.

2.1 ベースラインシステム

ジャンル分類とは, クラシック, ロックなどの楽曲のジャンルにより音楽进行分类する手法である. 図 1 に, ベースラインシステムの概要を示す.

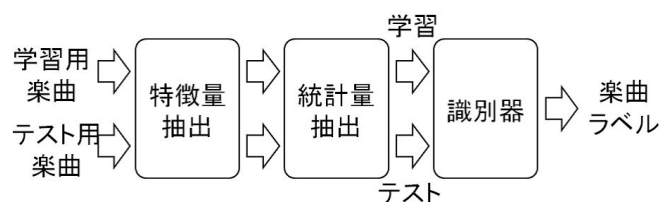


図 1 ベースラインシステムの概要

¹ 静岡大学
Shizuoka University
² 長岡技術科学大学
Nagaoka University of Technology
a) akahane@spa.sys.eng.shizuoka.ac.jp
b) s133167@stn.nagaokaut.ac.jp
c) wang@vos.nagaokaut.ac.jp
d) kai@sys.eng.shizuoka.ac.jp

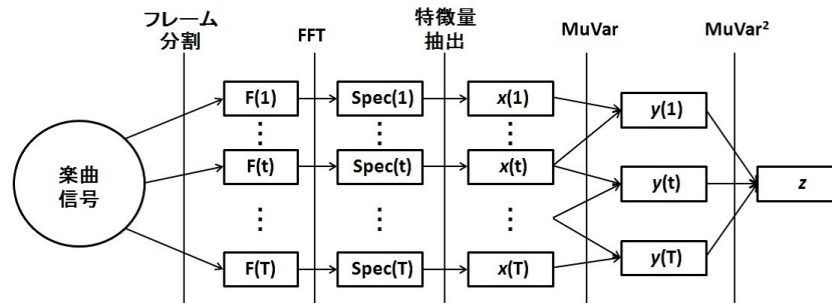


図 2 MARSYAS[3] による特徴量抽出のしくみ

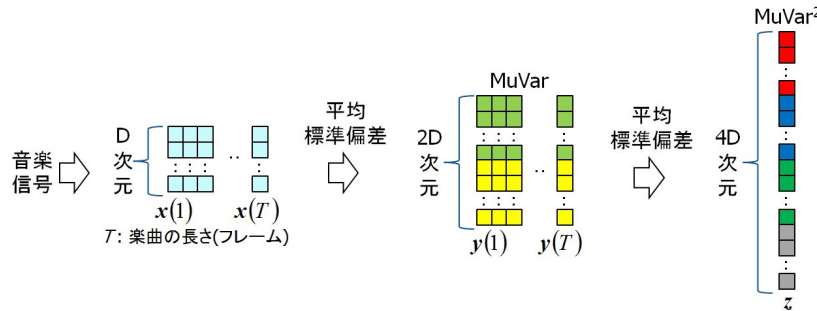


図 3 MARSYAS[3] による特徴量抽出のしくみ (次元)

まず学習用のデジタル音楽から特徴量を抽出，それに対して統計量を抽出し，その統計量をもとに識別器を構築する．次に学習用と同様の手順でテスト用のデジタル音楽から統計量を抽出し，識別器に通すことで楽曲のジャンルを推定する．

2.2 短時間の特徴量

まず，特徴量抽出の概要について説明する．音楽から特徴量を抽出する過程を図2に示す．

音データは，時間によって周波数成分が変化しない定常信号と変化する非定常信号に分類される．音声（音楽）は後者の非定常信号である．したがって，周波数分析を音声全体で行う長時間フーリエ変換 (Long-term Fourier transform) より，フレーム単位 (短い区間で音声を分割したもの) で行う短時間フーリエ変換 (Short-term Fourier transform) の方が適している．また，計算を簡単にするために周波数分析には高速フーリエ変換 (Fast Fourier Transform) を使う．FFT では周波数分析を行うサンプル数を2のべき乗にすることでフーリエ変換を高速に行うことができる．周波数分析により得られたスペクトルから様々な特徴量を計算することができる．

本研究で用いた特徴量は，特徴量抽出ツール MARSYAS[3] により抽出した．本研究では周波数分析を行うサンプル数を512とし，オーバーラップはなしとした．例えば，サンプリングレート22050[Hz]で30[s]の音声については，音声は約1300フレームに分割され特徴量が抽出される．したがって1フレームは音声の約23[ms]に

相当する．以下，このようにして D 次元の特徴ベクトル時系列データ $\mathbf{x}(t)$ が楽曲から抽出されるものとする．

2.3 統計量に基づく特徴量 (手法: MuVar , MuVar^2)

一般的に音声認識の分野で，特徴量はフレーム単位 (音楽をより小さな塊に分割したもの) で抽出される．これは通常10-100[ms]程のフレームであり，音声の短期的な性質しか捉えることができない．しかし，楽曲のジャンル分類などの音楽分類のタスクにおいては，音楽固有の性質 (各種楽器の性質，メロディ等) を捉えることが重要である．そのためには音楽の長期的な性質を捉えなければならない．

これを可能にする方法として，フレーム単位で抽出された特徴量の統計量を計算するという考えがある．フレーム単位で抽出された特徴量に対し，いくつかのまとまりで統計量 (平均，分散等) を計算して，これらを要素とする新たな特徴ベクトル列を生成することで，より長期的な性質を捉えることを試みる．

本研究では，MARSYASを利用してこれを実現した．MARSYASによる特徴量抽出のしくみを図2，図3に示す．

前節「短時間の特徴量」で示したように，MARSYASにより，楽曲からフレーム単位で D 次元の特徴ベクトル列 $\mathbf{x}(t)$ を抽出する．MARSYASでは， $\mathbf{x}(t)$ に対し，隣接する複数フレーム単位で平均・標準偏差を計算して，これらを要素とする $2D$ 次元の新たな特徴ベクトル列 $\mathbf{y}(t)$ を生成する．この手法を MuVar と呼ぶ．

$$\mathbf{y}(t) = \text{mean, std}[\mathbf{x}(t-m+1), \dots, \mathbf{x}(t)] \quad (1)$$

なお， t はフレームの番号 ($t = 1, \dots, T$) であり， T はフ

表 1 ジャンル分類実験

特徴量	識別器	識別率 (%)
$LPCC \times MuVar^2$	SVM(RBF)	53.2
$CHROMA \times MuVar^2$	SVM(RBF)	55.0
$STFT \times MuVar^2$	SVM(RBF)	57.7
$SCF \times MuVar^2$	SVM(RBF)	67.5
$SFM \times MuVar^2$	SVM(RBF)	71.1
$LSP \times MuVar^2$	SVM(RBF)	71.5
$MFCC \times MuVar^2$	SVM(RBF)	74.9
$(STFT + MFCC + CHROMA) \times MuVar^2$	SVM(RBF)	79.2
$(STFT + MFCC + CHROMA + SCF) \times MuVar^2$	SVM(RBF)	80.7
$(STFT + MFCC + CHROMA + SFM) \times MuVar^2$	SVM(RBF)	82.4
$(STFT + MFCC + CHROMA + LPCC) \times MuVar^2$	SVM(RBF)	82.5
$(STFT + MFCC + CHROMA + LSP + LPCC) \times MuVar^2$	SVM(RBF)	82.9
$(STFT + MFCC + CHROMA + SFM + SCF) \times MuVar^2$	SVM(RBF)	83.0
$(STFT + MFCC + CHROMA + LSP) \times MuVar^2$	SVM(RBF)	83.3
$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar^2$	SVM(RBF)	84.5
$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar^2$	DNN	81.9

レームの総数, m は統計量を求めるフレーム数を表す. 本研究では $m = 20$ とし, 約 0.5[s] 単位で統計量を求めることに対応する.

$MuVar$ により得られた特徴量 (統計量) に関し, 更に楽曲の全体で平均・標準偏差を計算して, これらを要素とする 4D 次元の新たな特徴ベクトル $\mathbf{z}$ を生成する. この手法を $MuVar^2$ と呼ぶ.

$$\mathbf{z} = \text{mean, std}[\mathbf{y}(1), \dots, \mathbf{y}(T)] \quad (2)$$

$MuVar^2$ に関して, 特徴量が 1 楽曲あたり 1 ベクトルで表わされることから, 楽曲レベル特徴量と呼ぶ.

2.4 ベースラインジャンル分類実験

実験データとして GTZAN Genre Collection^{*1} を用いた. データセットは Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae および Rock の 10 ジャンルで, 各ジャンル 100 曲, 合計 1000 曲の wav ファイルでできている. 全ての楽曲は 30[s] で, サンプリングレート 22050[Hz], 16 ビットのモノラル信号である. 特徴量として 4 次元の Short-time Fourier transform(STFT), 13 次元の Mel-frequency cepstral coefficients(MFCC), 14 次元の CHROMA, 24 次元の Spectral Flatness Measure(SFM), 24 次元の Spectral Crest Factor(SCF), 18 次元の Line Spectral Pair(LSP), 12 次元の Linear Prediction Cepstral Coefficients(LPCC) に対し $MuVar^2$ により求めた 436 次元の楽曲レベル特徴量を使用した. また, 各種特徴量を組み合わせた場合も試した. 識別器には Support Vector Machine(SVM)[4], Deep Neural Network(DNN)[5][6] を使用した. SVM については, Radial Basis Function (RBF) カーネルを用いる. 評価には 10 分割交差検証を用いた.

実験結果を表 1 にまとめた.

図中の + 記号は, 使用された特徴量の組み合わせを, ×

記号はそれら特徴量に対する処理を示している. 実験結果に関して 1 行目を例に説明すると, LPCC 特徴量がフレームレベルで抽出され, $MuVar^2$ により統計的処理を行って楽曲レベル特徴量を得て, それを元に SVM(RBF) により識別を行うことを意味している.

実験結果より, 単独で良い性能を示したのが SFM(71.1 %), LSP(71.5 %), MFCC(74.9 %) であった.

しかし, 8 行目以降を見ると分かるように, 単独で良い性能を示さなかった特徴量でも, 組み合わせで使用することで高い性能を示すことが分かった. 例えば STFT, CHROMA は単独では 55 % 前後であるが, 8 行目にあるように MFCC と組み合わせることで 79.2 % という高い識別率を得た.

何通りか特徴量の組み合わせを試したが, 結果的に最終行の STFT, MFCC, CHROMA, SFM, SCF, LSP, LPCC を組み合わせた結果 84.5 % が最も高い識別率となった.

したがって, 最終行の 84.5 % をベースラインとし, 識別の内訳を表すコンフュージョン (混合) 行列を表 2 に示す.

表 2 ベースライン (識別率:84.5 %) のコンフュージョン行列

		識別器が出力したジャンル									
入力データのジャンル		cl	co	di	hi	ja	ro	bl	re	po	me
	cl	97	2	0	0	1	0	0	0	0	0
	co	0	88	2	0	2	5	1	0	2	0
	di	0	2	79	3	1	8	0	5	2	0
	hi	0	1	2	82	0	2	1	7	2	3
	ja	5	1	0	0	92	0	1	0	1	0
	ro	1	7	8	3	1	67	2	4	4	3
	bl	0	1	0	0	2	4	92	0	0	1
	re	0	3	4	4	2	6	1	77	3	0
	po	0	7	3	4	0	4	0	2	80	0
	me	0	0	0	1	0	6	2	0	0	91

^{*1} Available at, <http://marsyas.info/download/data-sets/>

なお、表の値は楽曲数である。表の見方は、正解が classical の楽曲を正しく classical と識別できた数が 97, 誤って country と識別した数が 2, 誤って jazz と識別した数が 1, となる。

また、DNN による識別率が SVM に比べ大幅に低いのは、楽曲レベル特徴量を用いたことにより、学習データが大幅に不足してしまったことが原因だと考えられる。

3. 提案手法によるジャンル分類

本章では統計量に基づく特徴量 (提案手法: $MuVar_N^2$), $MuVar_N^2$ を用いたセグメントレベルの特徴量による識別実験について述べる。

3.1 $MuVar_N^2$ による統計量に基づく特徴量

ベースラインジャンル分類実験においては、 $MuVar^2$ によって得られた楽曲レベル特徴量を用いた。しかし、これでは 1 つの音楽を表すのが 1 つの特徴量ベクトル (統計量) だけになってしまうので、楽器・音色など音楽にとって重要な情報が失われてしまう可能性がある。それに対して、ある程度の時間単位で特徴量をまとめたセグメントレベル特徴量を提案する。

セグメントレベル特徴量の抽出過程を図 4 に示す。

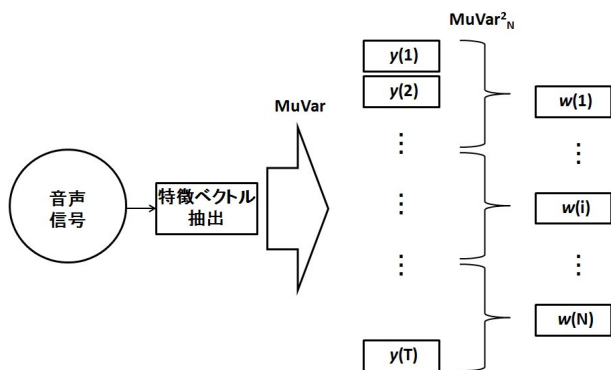


図 4 セグメントレベル特徴量抽出の流れ

まず、MARSHAS を用いて D 次元の特徴ベクトル $x(t)$ を抽出し、その平均・標準偏差を計算する ($MuVar$)。楽曲レベル特徴量の場合では、 $MuVar$ により得られる全ての特徴量ベクトル時系列に対してもう一度平均・標準偏差を計算 ($MuVar^2$) したが、ここでは、 $MuVar$ により得られる特徴ベクトル時系列を N 等分し、区分化されたベクトル列 (セグメント) 毎に平均・標準偏差を計算して、これらを要素とする $4D$ 次元の新たな特徴ベクトル列 $w(i)$ を生成する。この手法を $MuVar_N^2$ と呼ぶ。

$$w(i) = mean, std[y((i-1) * \frac{T}{N} + 1), \dots, y(i * \frac{T}{N})] \quad (3)$$

なお、 i はセグメントの番号 ($i = 1, \dots, N$) である。1 つの音楽を N 個のセグメントで表すことができ、それぞれ 1

つ 1 つのセグメントは一定の時間単位でまとめられることになる。 $MuVar_N^2$ を N 分割のセグメントレベル特徴量と呼ぶ。

例えば、30[s] の音楽に対し、10 分割のセグメントレベル特徴量を計算した場合、その 1 つ 1 つのセグメントは音楽を 3[s] 間でまとめたものであるため、音楽の 3[s] 間の性質を表すことができると考えられる。どの程度の時間単位で音楽を捉えれば、よりその音楽を表現することができるのかという問題は、ジャンル分類に関して重要であるかもしれない。

3.2 セグメントレベルの識別結果を用いた楽曲ジャンル推定

問題になるのは、 N 分割のセグメントレベル特徴量を用いて識別を行うと、1 楽曲に対して N 通りの識別結果が得られることである。 N 通りの識別結果を用いて 1 楽曲に対するラベル (ジャンル) を推定する方法を提案する。

1 目のアプローチとして、得られた N 通りの識別結果の多数決をとり、最終的なラベルを推定する方法を majority とする。

2 目のアプローチとして、確率値を用いたラベル推定を考える。図 5 に確率値を用いたラベル推定法を示す。

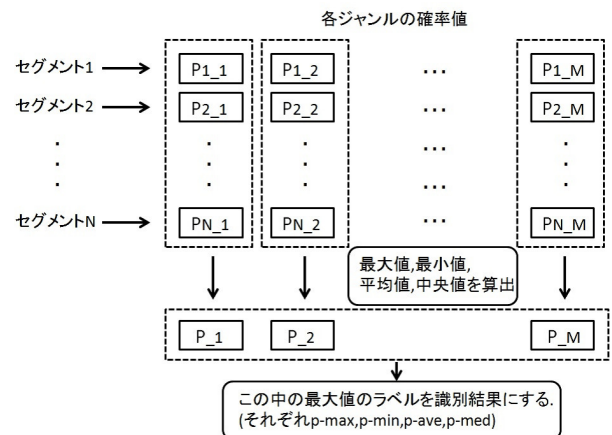


図 5 セグメントレベルの確率値によるラベル推定

LIBSVM[7], DNN には確率推定機能が搭載されており、識別結果と共に各ラベルの確率値を得ることができる。例えば、ジャンル数が M の場合、ジャンル 1 である確率、ジャンル 2 である確率、..., ジャンル M である確率がそれぞれ得られる。つまり、1 楽曲に対して N 通りの識別結果が得られ、それぞれに関し M 通りの確率値が得られる。そこで、 N 通りの各ラベルの確率値の最大値、最小値、平均値、中央値を計算し、その最大値のラベルをその楽曲の最終的なラベルとする。これらの手法をそれぞれ p-max, p-min, p-ave, p-med とする。

3.3 提案手法によるジャンル分類実験

実験データとして GTZAN Genre Collection を用いた。特徴量として STFT(4), MFCC(13), CHROMA(14), SFM(24), SCF(24), LSP(18), LPCC(12) を組み合わせた 109 次元の特徴量に対し, $MuVar_N^2$ により求めた 436 次元の N 分割のセグメントレベル特徴量を使用した。識別器には SVM(RBF), DNN を使用した。評価には 10 分割交差検証を用いた。

実験結果を表 3 に示す。表中の+記号は、使用された特徴量の組み合わせを、×記号はそれら特徴量に対する処理を示している。表 3(1) を例に説明すると、STFT, MFCC, CHROMA, SFM, SCF, LSP, LPCC 特徴量がフレームレベルで抽出され, $MuVar_3^2$ により統計的処理を行ってセグメントレベル特徴量を得て、それを元に SVM(RBF) により識別を行うことを意味している。なお、この識別結果は前章の図 5 に示した確率値を用いたラベル推定法とは異なり、セグメント全体の識別結果で評価している。

表 3 より、セグメントレベル特徴量の分割数を上げるほど識別率が下がることが分かった。おそらく、30 秒の音楽の中でそのジャンルらしさを表しているのは一部分であり、それは 1 秒間に現れていたり、2 秒間に現れていたりすると思われる。したがって、分割数を上げたことで、そのジャンルらしさを表していない部分が多く現れたことが識別率低下の原因だと考えられる。

続いて、上記ジャンル分類により得られた識別ラベルおよび各ラベルの確率値を用いて、前章の図 5 に示した確率値を用いたラベル推定法により、ラベルの再推定を行った。

識別結果を表 4、表 5 に示す。表 4(1) を例に説明すると、STFT, MFCC, CHROMA, SFM, SCF, LSP, LPCC 特徴量がフレームレベルで抽出され, $MuVar_3^2$ により統計的処理を行ってセグメントレベル特徴量を得て、それを元に SVM(RBF) による識別を行い、得られた識別ラベルおよび各ラベルの確率値を用いて、識別手法 majority, p-max, p-min, p-ave, p-med による識別を行うことを意味している。

表 4、表 5 の実験結果より、majority, p-ave, p-med が他より良い識別性能となっている。majority に関しては、分割数の増加により多数決の信頼度が上昇したのだと考えられる。p-ave, p-max では確率値の平均、中央値をとることで、曲全体を通した各ジャンルの確率値を得られたのではないと思う。(個々の結果では大きくずれてしまったりすることが考えられるが、平均、中央値をとることでずれを減らせる)

中でも表 4(4) の 30 分割のセグメントレベル特徴量による識別手法 p-ave が 87.9 % と最高であった。

表 4(4) の 30 分割のセグメントレベル特徴量による識別手法 p-ave(識別率:87.9 %) の識別の内訳を表 6 に、表 5(2) の 30 分割のセグメントレベル特徴量による識別手法

p-ave(識別率:86.5 %) の識別の内訳を表 7 に示す。

どちらの表に関しても言えることであるが、ベースラインに対してほぼ全てのジャンルに関し正解数が 3~5 程増えている。理由として、30 分割のセグメントレベル特徴量すなわち約 1[s] の単位で楽曲を表現することで、より多くのセグメントがそのジャンルを捉えることができるようになったためだと考えられる。

表 3 セグメントレベルでのジャンル分類実験

番号	特徴量	識別器	識別率 (%)
(1)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_3^2$	SVM(RBF)	81.6
(2)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_6^2$	SVM(RBF)	80.3
(3)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_{10}^2$	SVM(RBF)	78.7
(4)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_{30}^2$	SVM(RBF)	75.0
(5)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_{50}^2$	SVM(RBF)	73.1
(6)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_{10}^2$	DNN	77.3
(7)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_{30}^2$	DNN	73.1

表 4 SVM の識別ラベル、各ラベルの確率値を用いたジャンル分類実験

番号	特徴量	識別手法	識別率 (%)
(1)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_3^2$	majority	83.3
		p-max	85.7
		p-min	85.8
		p-ave	85.4
		p-med	84.4
(2)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_6^2$	majority	85.6
		p-max	85.5
		p-min	85.8
		p-ave	86.6
		p-med	86.1
(3)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_{10}^2$	majority	86.4
		p-max	85.6
		p-min	85.0
		p-ave	86.8
		p-med	87.1
(4)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_{30}^2$	majority	86.9
		p-max	84.1
		p-min	84.4
		p-ave	87.9
		p-med	87.5
(5)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_{50}^2$	majority	87.2
		p-max	83.2
		p-min	83.4
		p-ave	87.2
		p-med	87.0

表 5 DNN の識別ラベル, 各ラベルの確率値を用いたジャンル分類実験

番号	特徴量	識別手法	識別率 (%)
(1)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_{10}^2$	majority	86.2
		p-max	85.1
		p-min	86.1
		p-ave	86.2
		p-med	85.8
(2)	$(STFT + MFCC + CHROMA + SFM + SCF + LSP + LPCC) \times MuVar_{30}^2$	majority	86.5
		p-max	82.5
		p-min	83.4
		p-ave	86.5
		p-med	86.6

表 6 30 分割セグメントレベル特徴量を用いた SVM による p-ave 手法 (識別率:87.9 %) のコンフュージョン行列

		識別器が出力したジャンル									
入力データのジャンル		cl	co	di	hi	ja	ro	bl	re	po	me
	cl	99	0	0	0	1	0	0	0	0	0
	co	0	91	1	0	2	2	2	0	2	0
	di	1	1	83	3	0	4	2	2	4	0
	hi	0	2	2	86	0	0	0	5	3	2
	ja	3	1	0	0	95	0	1	0	0	0
	ro	0	8	7	1	0	72	3	4	3	2
	bl	0	0	0	0	2	3	95	0	0	0
	re	0	3	2	7	0	3	1	79	5	0
	po	0	4	1	3	0	4	0	1	86	1
	me	0	0	1	0	0	4	2	0	0	93

表 7 30 分割セグメントレベル特徴量を用いた DNN による p-ave 手法 (識別率:86.5 %) のコンフュージョン行列

		識別器が出力したジャンル									
入力データのジャンル		cl	co	di	hi	ja	ro	bl	re	po	me
	cl	96	1	0	0	1	2	0	0	0	0
	co	0	90	1	0	1	3	2	2	1	0
	di	1	2	79	4	0	4	2	3	5	0
	hi	0	2	2	83	0	1	0	5	5	2
	ja	3	1	0	0	94	0	1	0	1	0
	ro	0	12	7	2	0	69	4	3	1	2
	bl	0	2	0	0	1	2	95	0	0	0
	re	0	3	2	7	1	3	1	80	3	0
	po	0	8	2	1	0	2	0	2	84	1
	me	0	0	0	0	0	5	0	0	0	95

4. おわりに

本論文では楽曲のジャンル分類に対し、ベースラインとして 1 楽曲を 1 つのベクトルで表す楽曲レベル特徴量を用いた識別実験, 提案手法として統計量を基にした, 1 楽曲を複数の特徴ベクトルで表わすセグメントレベル特徴量を用いた実験を行った。また, 楽曲レベル・セグメントレベルの特徴量による識別結果の比較を行った。

セグメントレベル特徴量を用いたジャンル分類実験ではベースラインを最高で 3.4 ポイントと大きく上回った。中でも平均値, 中央値 (p-ave, p-med) を用いた場合の識別結果が高くなった。これはセグメント全体でその楽曲を評価したため, その楽曲を通して平均的なジャンルを推定できたためだと考えられる。しかし, 闇雲にセグメントの数を増やせばいいというものではなく, 1 つ 1 つのセグメントがそのジャンルを捉えていることが重要となる。本研究結果では, 分割数を 3-50 まで変化させたとき, 分割数 30 が最も良かったことから, 1[s] 程度の単位での統計量が識別情報をより多く持っている可能性があることが分かった。

今後の課題についてだが, セグメントレベル特徴量による識別で得られた確率値により, 1 楽曲に対するラベルを推定する際に, 信頼度の高いセグメントだけを選択する方法を考えたい。

また, 表 6, 表 7 から分かるように, rock に関する識別性能がとりわけ低い。この部分に関して性能を向上させれば, 全体の識別率を大きく上げることができると思われる。

参考文献

- [1] T. Li, M. Ogiwara, "Toward intelligent music information retrieval," IEEE Transactions on Multimedia, Vol.8, No.3, pp.564-574, 2006.
- [2] Z. Fu, G. Lu, K. Ting, D. Zhang, "A Survey of Audio-based Music Classification and Annotation," IEEE Transactions on Multimedia, Vol.13, No.2, pp.303-319, 2011.
- [3] G. Tzanetakis, "Marsyas-0.2: A case study in implementing music information retrieval systems," in Proc. Intelligent Mobile Information Systems.
- [4] C. Burges, "A tutorial on support vector machines for pattern recognition," Knowledge Discovery and Data Mining, vol. 2, no. 2, pp.121 -167, 1998.
- [5] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, Vol. 313, no. 5786, pp. 504 - 507, 2006.
- [6] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Process. Mag., Vol.29, No.6, pp.82-97, 2012.
- [7] R. Fan, P. Chen and C. Lin, "Working set selection using the second order information for training SVM." Journal of Machine Learning Research 6, 1889-1918, 2005