音声言語情報処理 44－15
(2002. 12. 19)

音声・音楽識別を目的とした特徴量の検討

谷口 徹† 大川 茂樹†† 白井 克彦†

† 早稲田大学理工学部情報学科 〒169-8555 東京都新宿区大久保 3-4-1
†† 千葉工業大学情報科学部情報ネットワーク学科 〒275-0016 千葉県習志野市津田沼 2-17-1
E-mail: †{tani,shirai}@shirai.info.waseda.ac.jp, ††okawa@net.it-chiba.ac.jp

あらまし 音声・音楽識別は音響コンテンツへのインデキシングやその前処理など、様々な応用が期待されており、現在多くの手法が提案されている。本研究では特に音声・音楽識別に用いられる特徴量に注目し、先行研究で有効性が示されている4種の特徴量の評価を行った。評価には性別やBGMの有無、歌声と楽器音の重畳などを考慮し設定した7種のクラスによりラベル付けをしたデータセットを用い、各特徴量の誤認識の傾向を分析した。
キーワード 音声・音楽識別, 音響特徴量, 音響情報検索

A Study of Features for Speech/Music Discrimination

Toru TANIGUCHI†, Shigeki OKAWA††, and Katsuhiko SHIRAI†

† School of Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku, Tokyo 169-8555 Japan
†† Dept. Network Science, Chiba Institute of Technology
2-17-1 Tsudanuma, Narashino, Chiba 275-0016, Japan
E-mail: †{tani,shirai}@shirai.info.waseda.ac.jp, ††okawa@net.it-chiba.ac.jp

Abstract Speech/Music discrimination has been studied for various applications such as automatic indexing of audio data. In this paper, we focus on four acoustic features examined in related studies and evaluate these features with audio data sets classified into seven audio classes .
Key words Speech/Music discrimination, Acoustic feature, Audio retrieval

1. はじめに

近年、各種デジタル技術の進歩、コンピュータネットワークの広まりとその帯域の向上に伴い、音響・映像を含む高密度なマルチメディアコンテンツが日々増加し、利用可能になってきている。しかし従来のテキストコンテンツと比較し、マルチメディアコンテンツはその内容の解析や認識が困難であり、そのため、音響や映像、さらにそれらを統合したマルチメディアコンテンツに対する検索や認識といった情報処理技術の進歩が強く求められるようになってきている。そういった背景のもとで、音響コンテンツへのインデキシングやその前処理を目的とした音声・音楽識別が近年盛んに研究が行われている。

一般に識別問題は特徴量の選択、モデルの選択両者が識別器の性能を左右すると考えられる。音声・音楽識別問題について言えば、現状では有効な特徴量の検討が続けられている段階である。

先行研究[1][2][3]では、音声・音楽識別に有効な特徴量がいくつか提案されており、またこれらの組み合わせにより、識別率が向上することも確認されている。

しかし先行研究では、一連の音声・音楽信号に対して信号列全体で各々どのくらいの割合で正しい識別を行うかという側面での識別力の評価は行われているものの、どのような信号に対して正しい判別を行い、またどのような信号に対して誤った識別を行うかという識別力の内容については十分な検討がなされているとは言い難い。

そこで本研究では、先行研究で挙げられている各特徴量を評価するために、音声と音楽の音響シーンについて従来より細かなクラスを設定し、評価用のデータセットにそのクラスによるラベルを付与しておくこと

で、特徴量のより細かな評価を試みた。

2. 特 徴 量

今回評価する特徴量を以下に挙げる。

2.1 零交差数

音響信号波形の時間次元における零交差数は周波数次元上の重心と深い関係にあり、音声波形中の有声音・無声音の判別に有効であるとされている。音声波形と音楽波形を比較すると、音声波形は語の開始・終了箇所や子音の摩擦音、破裂音に対応する箇所で零交差数の著しい増加が見られる。一方音楽波形においては全体に零交差数は比較的安定しており、音声波形のような急激な増減は見られないことが多い[1]。

よって、一定時間区間内での零交差数の変動度合いを比較することで、音声・音楽波形の識別ができると考えられ、変動度合いの指標として複数の連続したフレームから構成されるブロックと呼ぶ単位内での零交差数の分散（以下VZC）を音声・音楽識別器の特徴量として用いることが提案されている[2]。

2.2 Spectral Flux

Spectral Flux[2]は隣り合ったフレーム間における振幅スペクトルのユークリッドノルムであり、以下の式（1）のように定義される。

$$SF(n) = \|\mathbf{x}_n - \mathbf{x}_{n-1}\| \qquad (1)$$

ただし、$\mathbf{x}_n$は時刻nの振幅スペクトルベクトルを表し、$\|\cdot\|$は$\cdot$のユークリッドノルムを表す。

音楽は音声に比べスペクトル上の変動が激しく、フレーム間の差違も大きい。その為、Spectral Fluxは音声に比べより高い値を取ることが予想される。一方音声は母音から子音、子音から母音のように音素が切り替わる箇所のみSpectral Fluxが高くなり、Spectral Fluxの変動が大きいことが予想される。本研究では先行研究における評価に基づき、ブロック内での分散（以下VSF）を特徴量として用いる。

2.3 Cepstrum Flux

Cepstrum Flux[3]はSpectral Fluxを時間軸方向に拡張したもので、時刻nのCepstrum Fluxは以下の式（2）で定義される。

$$CF_J(n) = \frac{1}{J}\sum_{j=1}^{J}\|\mathbf{c}_n - \mathbf{c}_{n-j}\|^2 \qquad (2)$$

ここでJは窓のフレーム長、$\mathbf{c}_n$は時刻nのLPCケプストラムベクトルを表す。

Spectral Fluxが隣り合ったフレーム間の差違のみに注目しているのに対し、Cepstrum Fluxでは基準となるフレームのLPCケプストラムと、基準フレームから時間的に先行するJフレームのLPCケプストラムとの差違に着目している。よってSpectral Fluxと比較し、より安定した音声・音楽識別性能を持つことが予想できる。

さらに、ブロック内でCepstrum Fluxを平均することにより得られるBlock Cepstrum Fluxはより高精度の検出が可能であるとされている。Block Spectrum Flux(以下BSF)の定義を以下の式（3）に示す。

$$BCF_W(n) = \frac{1}{W}\sum_{i=0}^{W-1} CF_J(n-i) \qquad (3)$$

Wはブロック長を表す。

2.4 4Hz変調エネルギー

ここまで挙げた特徴量は全てスペクトルなど特徴量の時間変化に注目している。このような特徴量の時間変化を周波数次元で表したものを変調スペクトルと呼び、またその周波数次元を変調周波数と呼ぶ。この変調周波数軸上で見たとき、音声認識に必要な情報の大部分が1〜16Hzに存在しており、特に4Hz周辺の情報が最も重要であることが確認されている[4]。

この4Hzは音声の音節速度に対応しており、変調周波数4Hzの周辺には音声波形のエネルギーが集中することを予想できる。

先行研究[2]により提案されている4Hz変調エネルギー（4Hz modulation energy）は、メル周波数軸上で等間隔にフィルタを配置したフィルタバンク分析を用いて、信号波形をL個のチャネルに分割する。そして各チャンネルに中心周波数4Hzのバンドパスフィルタリングを行い、変調周波数4Hzのエネルギーを抽出する。時間n、チャネルlでのこのエネルギーを$m_{4Hz}(n,l)$としたとき、4Hz変調エネルギー（以下4HZM）は以下の式（4）のように定義される。

$$4HZM_W(n) = \frac{1}{W}\sum_{w=0}^{W-1}\frac{1}{E(n-w)}\sum_{l=1}^{L} m_{4Hz}(n-w,l) \qquad (4)$$

Wはブロック長、$E(n)$はフレームnのエネルギーを表す。

3. 評価用データセットの作成

各特徴量の評価を目的として、FMラジオ放送音源から音響データを採集し、音響シーン毎に人手によりラベリングを行うことで、評価用データセットを作成した。以下にその詳細を述べる。

表 1　音声・音楽クラスとサブクラス

クラス	サブクラス	説明
M	M-A	音楽, 歌声のみ
(音楽)	M-I	音楽, 楽器音のみ
	M-V	音楽, 歌声＋楽器音
S	S-B-F	音声,BGM あり, 女性
(音声)	S-B-M	音声,BGM あり, 男性
	S-NB-F	音声,BGM なし, 女性
	S-NB-M	音声,BGM なし, 男性

3.1　評価用音響データの収録

FM ラジオ放送を市販の FM ラジオチューナにより受信し、音響データの採集を行った。受信状況による影響を避ける為、放送局は採集箇所で受信状況が良好な放送局を 1 局選択し、話者、音楽のジャンル等でデータの内容に偏りがでないように適宜採集時間帯の変更を行った。収録は量子化方式 16bit PCM、標本化周波数 16kHz、モノラルで行った。

内容は男女複数パーソナリティーの発話、ポップス、ロックなどの音楽、CM など通常のラジオ放送の構成を反映したものとなっている。

3.2　評価用データセットの構成

収録したデータに対し、波形の目視と聴取を行いながら人手にて音響シーンラベルの付与を行った。ラベルは 3.2.1 章で定義する音響クラス・サブクラスを用いた。どのクラスにも分類できない区間は “U” のラベルを付与し、評価には使用しなかった。

ラベル付けされた波形データより、1 秒のブロック単位で評価用のサンプルを抽出し、評価用のデータセットとした。

3.2.1　クラス分類

音響信号を表 1 の音声・音楽の 2 クラスに分類し、さらに音声信号中の BGM が存在する箇所、音楽信号中の歌声がある箇所を分析するため、BGM の有無、歌声の有無、音声に関しては話者の性別に着目して音声 4 種類、音楽 3 種類のサブクラスに分類した。

表 2 に本データセットのクラス・サブクラス毎のサンプル長を示す。

4.　特徴量の分布

3. 章で作成したデータセットを用い、各特徴量におけるクラス毎の分布の調査を行った。図 1〜4 に各特徴量におけるクラス “M”（音楽）、“S-B”（BGM 有り音声）、“S-NB”（BGM なし音声）の分布としてヒストグラムを示した。各クラス毎にサンプル数が異なるため、ヒストグラムの度数は各クラス毎にクラス内のサンプル数で正規化している。音響分析条件を表 3 に示す。各パラメータは予備実験や先行研究の知見により設定した。

ここには示さないが、音声おける話者の性別に関して、“S-NB-F” と”S-NB-M”、“S-B-F” と”S-B-M” のクラス間で分布に大きな違いは見られなかった。また、音楽に関して、“M-V”、“M-I” の間にも顕著な差は見られなかった。よって、図 1〜4 には音楽は 3 種のクラス”M-A”、“M-I”、“M-V” を合わせた “M” を、音声については”S-NB-F” と “S-NB-M”、“S-B-F” と “S-B-M” をそれぞれ併せて一つのクラスとした “S-NB”、

表 2　クラス毎のサンプル長

クラス	時間長 (s)	
M	8085	(48.3 %)
M-A	95	(0.6 %)
M-I	1772	(10.6 %)
M-V	6218	(37.1 %)
S	8655	(51.7 %)
S-B-F	4578	(27.3 %)
S-B-M	2897	(17.3 %)
S-NB-F	627	(3.7 %)
S-NB-M	553	(3.3 %)

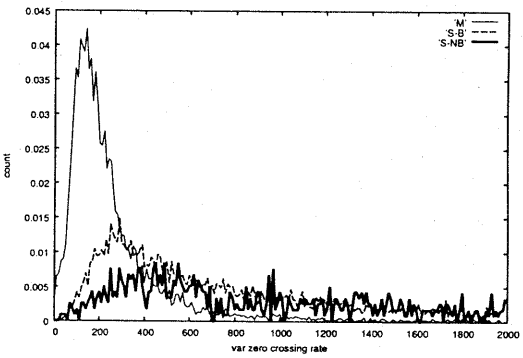

図 1　零交差数の分散（VZC）の分布

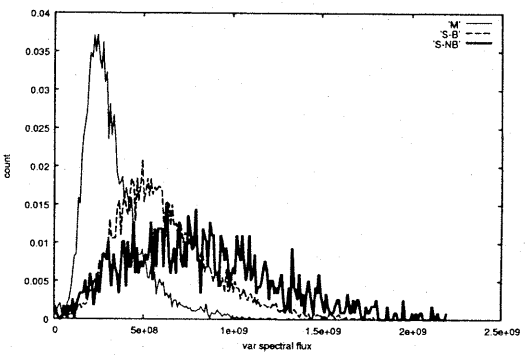

図 2　Spectral Flux の分散（VSF）の分布

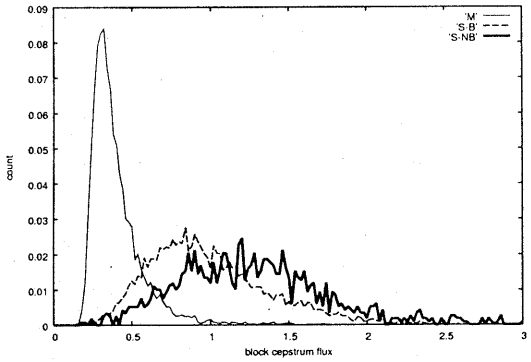

図 3　Block Cepstrum Flux（BCF）の分布

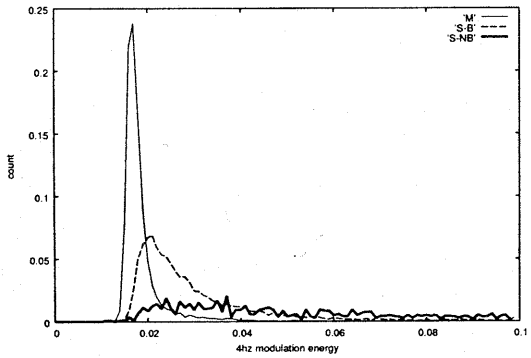

図 4　4Hz 変調エネルギー（4HZM）の分布

“S-B” として表示した。

4.1 考　　察

それぞれの特徴量の性質から予め予測されたように、音楽はより低い値、音声はより高い値を取っている。また音楽の分布は分散が小さく、一方音声の値は広い範囲に分布している。

識別に重要な分布の重なりであるが、“M” と “S-NB” に注目すると、図 1 の VZC、図 2 の VSF と比較し、図 3 の BCF、図 4 の 4HZM は分布の重なりが少ない為、識別も高精度であると予想できる。

BGM 有りの音声 “S-B” に注目すると、どの特徴量においても BGM なしの音声 “ S-NB” と比べ、分布が音楽に近づいている。広い分布を持つのは”S-NB” と同様だが、包絡のピークが音楽と近くなっており、識別は困難であると考えられる。4 つの特徴量の中では図 3 の BCF が比較的 “S-NB” の分布と近くなっている為、4 つの特徴量の中では BGM 有り音声に対して比較的高い検出力を持つことが予想される。

表 3　各特徴量の音響分析条件

VZC	
フレーム長	380 点 (24ms)
分析周期	160 点 (10ms)
ブロック長	100 フレーム (1000ms)
VSF	
フレーム長	512 点 (32ms)
分析周期	160 点 (10ms)
ブロック長	100 フレーム (1000ms)
窓関数	ハミング窓
BCF	
フレーム長	256 点 (16ms)
分析周期	256 点 (16ms)
LPC 分析次数	14
Cepstrum 分析次数	16
Cepstrum Flux 窓長 (J)	10 (160ms)
BCF 窓長 (W)	62 (992ms)
4HZM	
フレーム長	512 点 (16ms)
分析周期	256 点 (8ms)
フィルタバンク数 (L)	40
バンドパスフィルタ次数	41
ブロック長 (W)	128 (1000ms)

5. 識別実験による評価

2. 章の特徴量の音声・音楽識別力を評価するため、3. 章で述べた FM ラジオ放送音源より作成したデータセットを用いて識別実験を行った。

5.1 識 別 手 法

正規分布モデルを用い、音声と音楽の識別を行う。まず音声・音楽の各クラスについて、用意した学習データから最尤推定により正規分布のパラメータを推定する。具体的には学習データサンプルから標本平均と分散を求めることにより、式 (5) に挙げた μ、σ の 2 つのパラメータを推定する。式 (5) はサンプル x に関するクラス C_i の尤度である。

$$p(x|C_i) = p(x|\mu_i, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (5)$$

識別時はこの尤度の対数を取った対数尤度をクラス毎に比較し、尤度が最大となるクラスを x が属するクラスとして識別を行う。すなわち、

$$j = \arg\max_i \log p(x|C_i) \quad (6)$$

なるクラス C_j を x が属するクラスと識別する。

5.2 識別実験の方法

3. 章で作成したデータセットのうち、70 %をモデルのパラメータ推定用の学習データとして用い、残りの

表 4　各特徴量毎の正解識別率

	VZC	VSF	BCF	4HZM
M-A	0.657	0.717	0.699	0.825
M-I	0.958	0.849	0.935	0.945
M-V	0.967	0.911	0.956	0.995
S-B-F	0.418	0.627	0.795	0.203
S-B-M	0.332	0.577	0.773	0.073
S-NB-F	0.689	0.785	0.934	0.762
S-NB-M	0.642	0.686	0.879	0.641

30 %を評価用データとした。この組み合わせを変えることで5つの実験データセットを作成し、評価に用いた。なお、各セットの学習データ、評価用データ間で独立性が保てるように、放送上で連続していたサンプルが両者に含まれないよう配慮した。

識別の流れは、まず学習データのうち、音楽クラス“M” に属するデータと音声クラス “S” に属するデータから、それぞれのクラスの正規分布モデルパラメータの推定を行い、モデルを作成した。次にその2つのモデルを用いて評価用データの識別を行った。

各特徴量の各クラスのデータサンプルについて正解識別率を算出した。これを5 つの実験データセットについて行い、5回の平均を算出した。

各特徴量の音響分析条件は4. 章で用いた表3の条件に従っている。

5.3　実験結果と考察

結果の正解識別率を表4に示す。

特徴量毎に見ていくと、分布の考察から予想されたとおりBCF が音楽の “ M-{I,V}” で94 %近く、音声の “S-NB-{F,M}” で90 %前後と高い識別率を出している。分布の考察から4HZM も高い正解識別率が予想されたが、音楽の “M-{I,V}” で高い識別率を出しているのに対し、音声の “S-NB-{F,M}” が60〜70 %台と低い識別率にとどまった。特徴量VZC とVSF も同様の傾向である。これは分布の形状によるものと推測できる。

クラス毎に見ていくと、予想されたようにBGM 有りの音声 “S-B-{F,M}”、歌声の “M-A” の識別率が低くなっている。ただし、4HZM の”S-B-{F,M}” は20 %以下となっており、これは逆に音楽性をうまく検知していると言える。

6. ま と め

先行研究にて提案されている代表的な4種の特徴量の評価を行った。評価には7種のサブクラスによりラベル付けされた音響データセットを用い、特徴量毎の識別結果の差違を分析した。

評価した特徴量では音声と音楽が重畳した区間や、歌声のみの区間の識別には困難であり、そういった区間についても正しく検出する手法を検討していく必要がある。

謝　辞

本研究の一部は、早稲田大学理工学総合研究センターの研究課題「マルチモーダル情報空間における総合的ヒューマンインタフェースに関する研究」によるものである。ここに記して謝意を表する。

文　献

[1] J.Sanders, “Real-Time Discrimination of Broadcast Speech/Music”, Proc. of 1996 ICASSP, pp.993-996, 1996

[2] E.Scheirer,M.Slaney, “Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator”, Proc. 1997 ICASSP, vol.2, pp.1331-1334, 1997

[3] 内田, 山下, 杉山, “Cepstrum Flux を用いた音声と音楽のセグメンテーション”, 信学技報, SP2000-17, pp.9-16, 2000

[4] 金寺, H.Hermansky, 荒井, 船田, “ロバストな音声認識実現を目的とした変調スペクトル特性の検討”, 信学技報, SP97-70, pp.15-22, 1997