物理特徴量を用いたスピーチ, 音楽信号の分類*

◎原　晋吾　西　隆司（北九州市大）

1.INTRODUCTION

ディジタル音声データがあらゆる分野でその量を増加し続けている。特にインターネット上においては、その増加の度合いが著しく、データベースの構築、検索システムの向上が求められている。そこで、本研究は音声の自動分類システムの構築を目的とする。このシステムにより、人手によるテキストインデキシングを介さない、音声データベースの自動構築が可能となる。Juan Jose Burred[1],George Tzanetakis[2] らは音楽データに特化した分類法を報告している。従来、Tong Zhang[3] らが音声物理量を特徴量として用い分類を行ってきたが、音声の種類は違うものの特徴量が酷似した場合、分類が困難となる。本研究では、分類精度が強く要求される speech 信号と、その物理的特徴が酷似した drums 信号の分類アルゴリズムを提案する。信号の特徴量として、ZeroCrossingRate(ZCR) を用いることが有効であると報告されているが、drums 信号は speech 信号と酷似しているため ZCR をそのまま分類に用いるのは適当ではない。本研究は speech 信号と drums 信号を周波数解析により分類するアルゴリズムを提案する。

2.AUDIO FEATURE ANALYSIS

物理量の特徴抽出とそれを用た分類法を検討した。Table.1 に特徴抽出の条件を、Fig.1 に特徴抽出のフローチャートを示す。表中 Analysis window は音声をラベリングする単位の長さを表す。

Sampling rate	22050Hz
Frame length	30msec
frame shift	10msec
Analysis window	98frames/1sec
Window	Hamming

Table.1　Conditions for signal processing

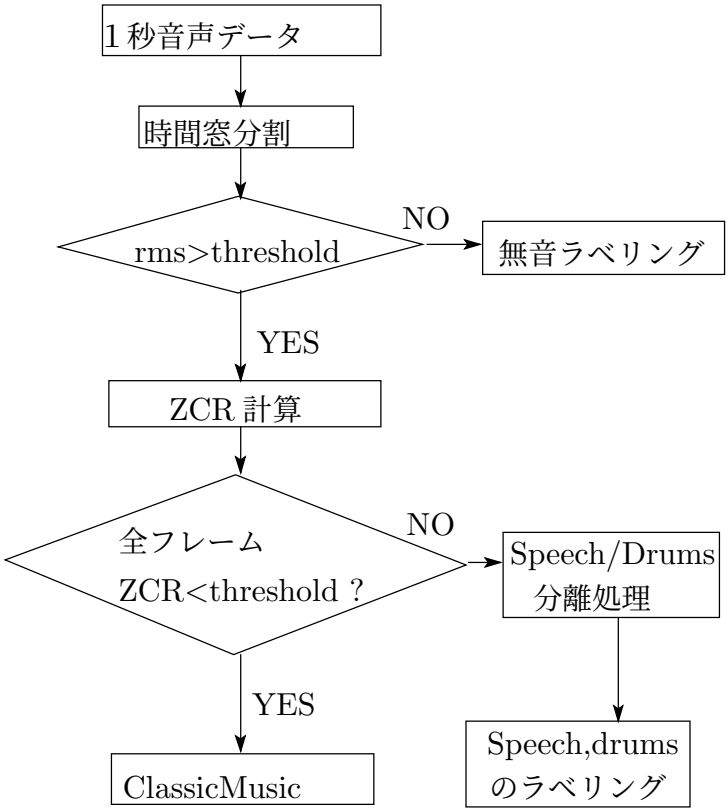

Fig.1　Flow chart of classification

2-1.ZeroCrossingsRate

ZeroCrossingsRate（以後 ZCR と記す）は、一定時間で振幅が 0 を通過する回数であり、信号に含まれる周波数成分と対応した物理量である。ZCR は次式で定義される。

$$Z_n = \frac{1}{2}\sum_m |sgn[x(m)w(n-m)] - sgn[x(m-1)w(n-m)]|$$

$$x(m) > 0 .....sgn[x(m)] = 1$$
$$x(m) < 0 .....sgn[x(m)] = -1$$

Classification of speech/musics by physical parameters

$x(m)$ は長さ N の信号列、m はサンプル番号、n はフレーム番号、Z_n は n 番目のフレームの ZCR、w は窓関数である。

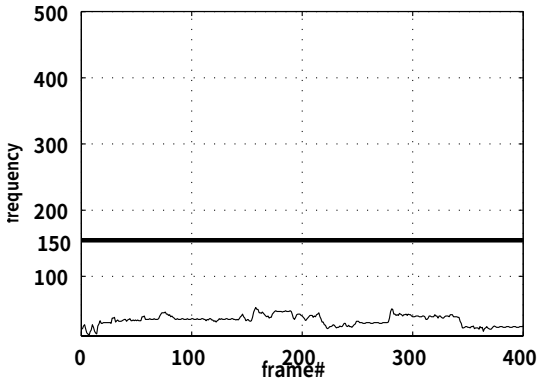

Fig.2　ZCR-Classic(cello-orchestra)

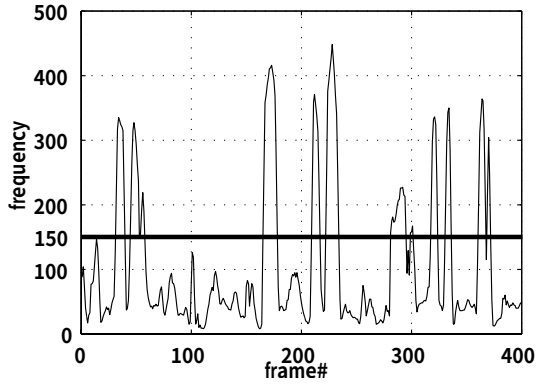

Fig.3　ZCR-FemaleSpeech

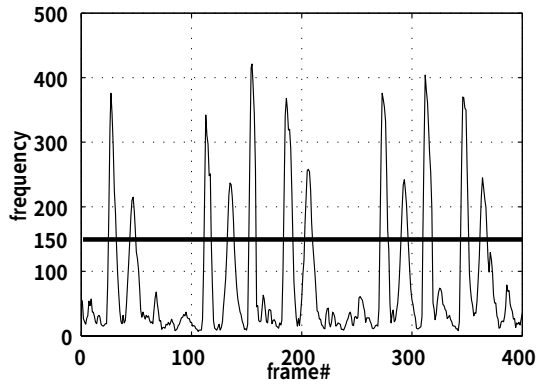

Fig.4　ZCR-Drums

Fig.2,3,4　に 、Classic、FemaleSpeech、Drums のフレーム毎の ZCR の時間経過を示す。Fig.2 より、Classic は他の 2 つと比較して、ZCR が低い値を示すことが分かる。Fig.3,4 より、FemaleSpeech,Drums は部分的に ZCR の値が大きくなり、互いに時間経過パターンが酷似していることが分かる。speech 信号の特徴として、子音部分で ZCR が大きくなり、母音部分でそれが小さくなる特徴が現れる。そこで、Classic と Speech/Drums 2 つを分類するために、ZCR の閾値を ZCR=150 と定めた。ZCR は speech と ClassicMusic の分類に有効であるが、drums との分類に関しては両者のパターンが酷似しているため適していないことが分かる。そこで、speech の場合の子音と比較してエネルギーが大きく、波形に差異が見られる母音部分を分析することにより、speech/drums 両信号の特徴の違いをさらに調べた。ZCR<150 の部分を母音的特徴部分とし、この部分の周波数解析を行った。

2-2.Frequency components of voiced sound

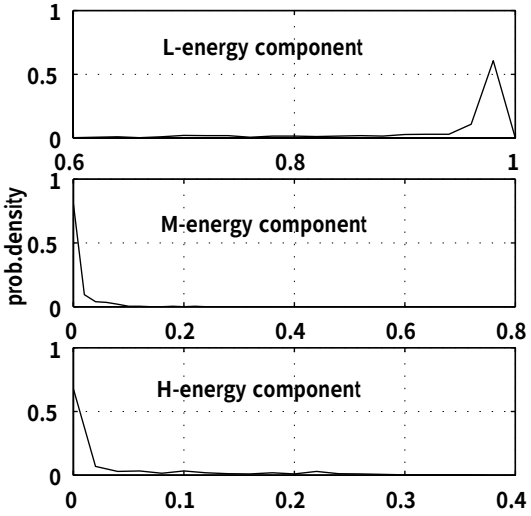

Fig.5　frequency components of drums(ZCR<150)

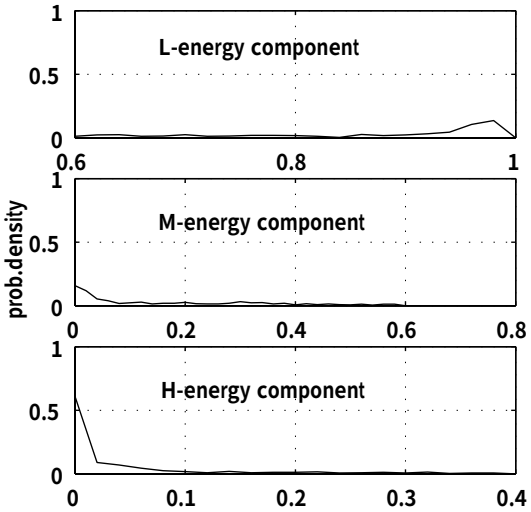

Fig.6 frequency components of speech(ZCR<150)

Fig.5,6 は ZCR<150 でフィルタリングされたフレームの周波数分析の結果である。周波数を Lowfrequency（<500Hz）,Mid-frequency（500Hz〜2kHz）,Highfrequency（2kHz〜5kHz）に分割し、下記の式で LowFreqeuencyEnergyComponent（L-comp）, MidFreqeuencyEnergyComponent（M-comp）,HighFreqeuencyEnergyComponent（H-comp）を求めた。

$$X(k) = FFT[x(m)]$$

$$T - eng = \sum_{k=1}^{L} X(k)^2$$

$$L - eng = \sum_{k=1}^{k_1} X(k)^2$$

$$M - eng = \sum_{k=k_1+1}^{k_2} X(k)^2$$

$$H - eng = \sum_{k=k_2+1}^{k_3} X(k)^2$$

$$L - comp = L - eng/T - eng$$

$$M - comp = M - eng/T - eng$$

$$H - comp = H - eng/T - eng$$

L は周波数の最終 bin 番号、k_1,k_2,k_3 はそれぞれ 500Hz,2kHz,5kHz に対応した bin 番号である。これらより、ZCR が 150 より小さいフレームの speech 信号と drums 信号は、L-comp 及び、M-comp の確率密度の違いがあることが分かった。drums 信号は speech 信号に比べ、L-comp が大きい割合の確率が高く、M-comp は小さい割合の確率が高いことが分かった。L-comp,M-comp の違いに着目し、両者を軸とする 2 次元分布を求めた。

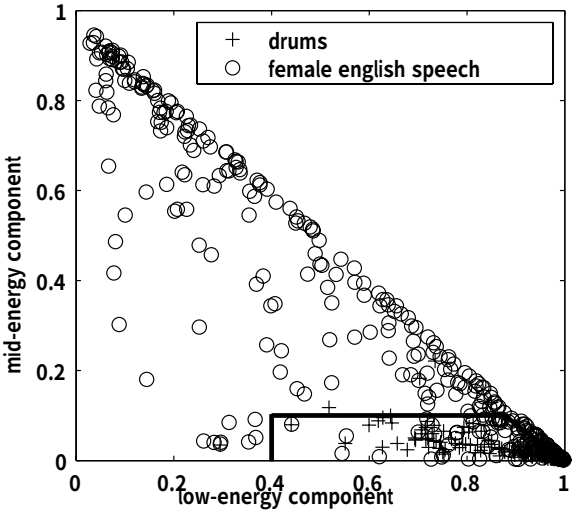

Fig.7 Low and Mid components(speech & drums)

Fig.7 は speech 信号及び drums 信号の L-comp,M-comp の 2 次元平面における分布図である。これより、drums 信号と speech 信号の領域を指定し、分類のための学習データとした。範囲指定は下記の通りである。

$$\text{drums} \begin{cases} \text{L-comp:0.4 以上} \\ \text{M-comp:0.1 以下,y 以下} \\ (y = 0.82x + 0.82) \\ y = M - comp, x = L - comp \end{cases}$$

Fig.7 中の実線で囲まれた部分は Drums の領域を示す。

speech:それ以外

3.EXPERIMENT AND RESULTS

ClassicMusic-990sec、Speech-570sec、Drums-30sec の長さの音声データを用意した。各音声データは同一の曲、人の声ではなく、別々のものを収集した。本アルゴリズムを適用し、音声データ 1sec ごとに、Classic/Speech/Drums のラベルをつけ、分類精度を検証した。さらに、部分的なエラーに対処するため、エラー修正を適用した。エラー修正のアルゴリズムは、5sec 単位の多数決によるものとした。結果を Table.2 に示す。表中 Cl、Sp、Dr、はそれぞれ Classic、Speech、Drums を表す。

	分類結果			
	Cl	Sp	Dr	正答率 (%)
Cl(修正前)	904	76	2	99.2
Cl(修正後)	972	10	0	99.9
Sp(修正前)	131	370	68	65
Sp(修正後)	51	482	36	84.7
Dr(修正前)	1	0	29	96.6
Dr(修正後)	0	0	30	100

Table.2　Result

4.CONCLUSION AND FUTURE WORK

本研究によって、ClassicMusic 信号を入力した際、99.9 %の精度で分類が可能であった。また、Speech 信号を入力した際、84.7 %の精度で分類が可能であった。これは、母音が 1sec 以上連続する時、ZCR の閾値を一度も超えないため本アルゴリズムでは Speech であると判別されないことが原因として挙げられる。これには、ZCR 以外のパラメータを用い、分類精度向上を図る必要がある。Drums 信号を入力した際、100 %の精度で分類が可能であった。

５秒間の多数決を用いたエラー修正の結果、Classic では 0.7 %、Speech では 19.7 %、Drums では 3.4 %の精度向上を達成した。エラー修正アルゴリズムに関しては、今後の重要な課題の一つである。さらに今後は、Fig.8 に示す分類の階層化、精度の向上を進めて行く方針である。

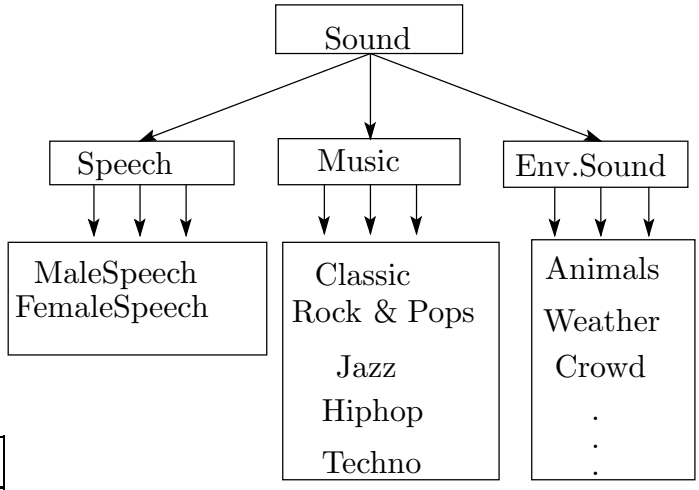

Fig.8　Genre Hierarcy

参考文献

[1] Juan Jose Burred and Alexander Lerch「Hierarchical Automatic Audio Signal Classification」

[2] George Tzanetakis,Georg Essl,Perry Cook「Automatic Musical Genre Classification Of Audio Signals」

[3] Thong Zhang and C.-C. Jay Kuo「Heuristic Approach for Generic Audio Data Segmentation and Annotation」