音声特徴量に基づくテレビ番組のジャンル分類*

◎立花伸元　西隆司（北九市大）

1　まえがき

近年のハードディスクレコーダの普及によって、アナログ放送のテレビ番組をデジタルデータとして扱うことができるようになった。それにより、テレビ番組をコンピュータ上で編集できるようになり、大量のテレビ番組をハードディスクに保存できるようになった。そこで、我々はその大量のテレビ番組を自動的に分類できるシステムが必要であると考え、音声信号のみを使って、テレビ番組のジャンル分類を行うための基礎的なアルゴリズムについて検討する。また、音声のスペクトログラムで視覚的に分類を行うことが出来るかについても検討する。ジャンル分類にはニューラルネットワークを使用してその有効性を検討する。

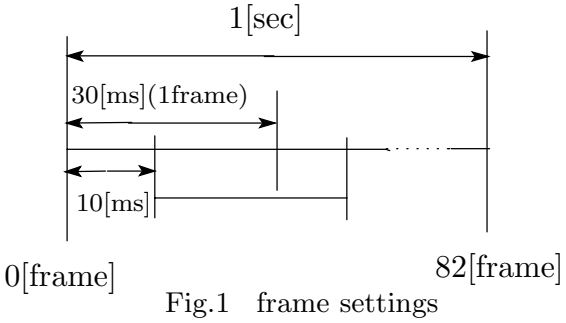

Fig.1　frame settings

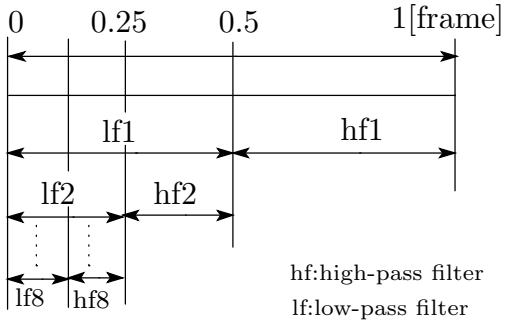

Fig.2　bandpass filtering by DWT

2　ジャンル自動分類のために用いた音声特徴量

本研究ではジャンル自動分類の基礎検討を行うため、ニュース番組とスポーツ番組（野球とサッカー）の分類を対象とした。ニュース番組とサッカー番組の分類及びニュース番組と野球番組の分類を以下に示す手順で行った。

1. テレビ番組から抽出した、ジャンルが未知の音声信号に離散ウェーブレット変換を行い、2 乗平均振幅を求めた。
2. 離散ウェーブレット変換後の音声データに主成分分析を行い、情報量を圧縮した。
3. 圧縮した音声データをあらかじめ学習させていたニューラルネットワークに通して、いずれかのジャンルに分類させ、その正解率を評価した。

ニュース、サッカー及び野球のテレビ番組からそれぞれ時間長 50[sec]、サンプリング周波数 22.5[kHz] で無作為に抽出して、学習用音声データとして使用した。さらに、抽出した音声を 1[sec]（82 フレーム）毎に区切り、1 秒内の特徴量を導出した。本研究ではまず、1[sec] 内の音声を Fig.1 に示すように、30[ms] のフレーム毎に分割し、離散ウェーブレット変換 (DWT)[1] を行い、帯域分割する。それぞれの離散ウェーブレット成分から、周波数成分の平均値の時間変化を求めた。各フレーム信号はハミング窓を使用して抜き出した。離散ウェーブレット変換後の音声データを時

Genres classification of TV program based on audio features
By Nobumoto Tachibana and Takashi Nishi
University of Kitakyushu

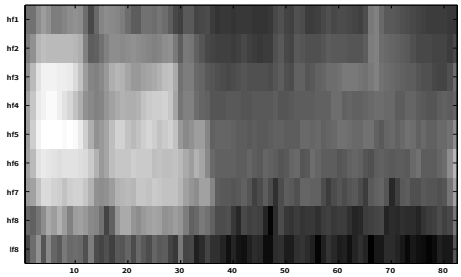

Fig.3 audio spectrogram extracted from news program

間－周波数変化パターン（以後、スペクトログラムと呼ぶ）で表示した例を Fig.3 に示す。

以上の方法で導出した音声データは周波数成分と時間成分を持つ行列と考える。この研究では、学習データを作るために、この音声データの行列を以下のようにベクトルに変換する。

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a_1} & \cdots & \boldsymbol{a_n} \end{pmatrix}$$

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{a_1}^T & \cdots & \boldsymbol{a_n}^T \end{pmatrix}^T \qquad (1)$$

$$n = 9, m = 82$$

ここで、n は帯域数、m は 1 秒内のフレーム数である。(1) 式で作られたベクトル $\boldsymbol{B}$ は 1 秒間の音声特徴ベクトル（$n \times m$ 次元）である。この列ベクトル$\boldsymbol{B}$を 1 秒間の音声特徴量として用いた。

3 ニューラルネットの学習

3.1 学習データ作成

音声特徴ベクトル$\boldsymbol{B}$から学習データを作成する。本研究ではニュースとサッカー及びニュースと野球番組の 2 つのジャンル分類を行うため、ニュース番組、サッカー番組及び野球番組から 48 秒ずつ、3 つの音声サンプルを合計 144[秒/ジャンル]、学習データ作成の為に使用した。ジャンル分類において、入力される信号は未知の音声信号を対象としているので、用意した音声信号から導出した特徴ベクトルを列ベクトルとして持つ行列 C を 2 つのジャンル分類に対してそれぞれ作成した。この時、行列 C は (2) 式のように表すことが出来る。

Table.1 table of genres

j	Pattern1	Pattern2
1〜144	news	news
145〜288	soccer	baseball

$$C = \begin{pmatrix} c_{11} & \cdots & c_{1j} \\ \vdots & \ddots & \vdots \\ c_{i1} & \cdots & c_{ij} \end{pmatrix} \qquad (2)$$

$$i = m \times n = 738, j = 288$$

ここで、c_{ij} は各音声サンプルから抽出した音声特徴ベクトルの要素。i は 1 枚のスペクトログラムを表現する要素数であり、j はスペクトログラムの枚数である。また、スペクトログラムの枚数はそれぞれ Table.1 に示すテレビ番組のジャンルに対応している。

(2) 式から行列 C のサイズは $i \times j$ である。このサイズでニューラルネットワークに学習させるには膨大な時間が掛かり、コンピュータの演算処理の限界も超えてしまうので、情報量圧縮を行う必要がある。このため、我々は主成分分析 [2] を使用して情報量を圧縮した。以下に、その手順を示す。

1. C から共分散行列 D を

$$D = (C - \boldsymbol{C_m})(C - \boldsymbol{C_m})^T \qquad (3)$$

から求めた。ここで、$\boldsymbol{C_m}$ は行列 C の列方向の平均ベクトルである。

2. 共分散行列 D を使って、

$$D \cdot e_l = \lambda_l \cdot e_l \qquad (4)$$

から、固有値 λ_l と固有ベクトル e_l を求めた。ここで、$l = 1, 2, \cdots, 738$ である。

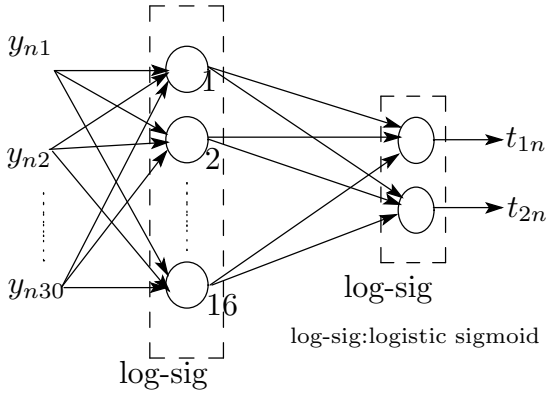

Fig.4　structure of neural network

3. 求まった固有値から累積負荷量を 0.9 として、削減する次元数を算出し、主成分を求めた。この結果、738 次元から次元数は 30 次元に圧縮された。
4. 30 個の固有ベクトル
   $u_k = (\boldsymbol{e_1}, \boldsymbol{e_2}, \cdots, \boldsymbol{e_{30}})$
   を使って、

$$y = {u_k}^T (C - \boldsymbol{C_m}) \qquad (5)$$

   により行列 C より次元数が削減された主成分 y を求めることが出来る。

これにより、y のサイズは $30 \times j$ となった。

主成分分析により求めた主成分 y を学習データとして用いた。

3.2　ニューラルネットワークトレーニング

学習データを使ってニューラルネットワークをトレーニングさせる。ここで、今回使用したニューラルネットワークの構造は Fig.4 に示すように 3 層構造（入力層 30, 中間層 16, 出力層 2）とした。図中の t_{1n}, t_{2n} はニューラルネットワークに出力させたい n 枚目の出力値である。この出力値と誤差の 2 乗が最小になるようにニューラルネットワークは反復を繰り返して最適な重みを得るべくトレーニングする。Table.2 に、定義した出力値とテレビ番組のジャンルとの対応を示す。また、ニューラルネットワークの学習アルゴリズムとして Levenberg-Marquardt アルゴリズム [3] を使用した。

Table.2　definition of output number

pattern1	genre	output number
t_{1n}	news	$(10)^T$
t_{2n}	soccer	$(01)^T$

pattern2	genre	output number
t_{1n}	news	$(10)^T$
t_{2n}	baseball	$(01)^T$

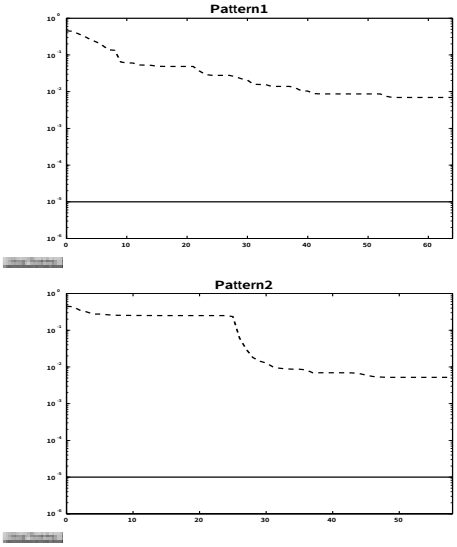

Fig.5　learning nueual network of 'pattern1' and 'pattern2'

Table.2 内で定義した 2 パターンのニューラルネットワークのトレーニング結果を Fig.5 に示す。ここで、実線は定義した収束限界値、点線は試行回数毎の誤差を表している。また、y 軸は誤差値、x 軸は試行回数を表している。誤差はどちらも 10^{-5} に達していないが、ある程度の収束は見せているのでこの学習結果を用いてジャンル分類を行った。

4　ジャンル分類実験

トレーニング後のニューラルネットワークを使用してテレビ番組のジャンル分類を行った。この実験で使用した音声信号は、学習データに使用したテレビ番組と同じテレビ番組から、学習データに使用していない区間を各ジャンル 2 区間（48[sec] の信号を 2 つ）、異なるテレビ番組から各ジャンル 2 区間の計

4 区間を抽出した。実験は以下に示す手順で行った。

1. 抽出した音声信号から音声特徴ベクトル $\boldsymbol{V}$ を算出する。
2. 音声特徴ベクトル $\boldsymbol{V}$ を、

$$y = {u_k}^T (\boldsymbol{V} - \boldsymbol{Cm}) \qquad (6)$$

で、主成分分析を行い、情報量を削減する。ここで、u_k と $\boldsymbol{Cm}$ は学習時に使用した固有ベクトル及び平均ベクトルである。

3. 主成分分析によって求められた主成分 y をトレーニングしたニューラルネットに通してジャンル分類を行う。

Table.4 にパターン 1 及びパターン 2 の場合のジャンル分類シミュレーションの結果を示す。Table.4 から、パターン 1 の場合は正解率は 80[%] を超え、実験は良い結果が得られた。しかし、パターン 2 では、サッカー番組の音声信号を使った場合に正解率が 50[%] を切っている。この原因として、学習データの偏りがあったことが挙げられる。サッカー番組の学習データとして使用した音声信号が全て同じ様な信号であったので、少し違う信号が入ってくると対応できない事が分かった。

5 むすび

今回の結果からジャンル分類を行う手法の一つとして、スペクトログラムを使って音声を視覚的に捉え、これにニューラルネットを用いて行う手法の有効性が明らかになった。ジャンル分類に関してはスペクトログラムのみでは全てのジャンルを分類することが困難であることも分かった。分類の幅を広げたり、分類の正解率をさらに上げるため、スペクトログラム以外の音声特徴量を用いる方法について今後検討する。

Table.3 resuct of classification

pattern1		
input	pacentage of correct	
genre	news	soccer
news	86	14
soccer	2.0	98

pattern2		
input	pacentage of correct	
genre	news	baseball
news	92.7	7.3
baseball	50.5	49.5

参考文献

[1] Wavelet Toolbox User's Guide.
Michel Misiti,Yves Misiti,Georges Oppenheim,Jean-Michel Poggi.
The Math Works

[2] Digital Image Processing second edition.
Rafael C.Gonzalez,Richard E.Woods.
Prentice Hall.

[3] Neural Network Toolbox User's Guide(1998).
Howrd Demuth,Mark Beale.
The MATH WORKS Inc.

[4] AUDIO FEATURE EXTRACTION AND ANALYSIS FOR SCENE SEGMENTATION AND CLASSIFICATION.
Zhu Liu and Yao Wang, Tsuhan Chen.
Polytechnic University Image Processing Lab.
(http://vision.poly.edu:8080/paper/jvsp98.pdf)