音楽番組からの歌謡曲認識システム

4M−8

中山 正樹　村岡 洋一
早稲田大学理工学部

1 はじめに

本研究は、自動編集のために、音楽番組の放送音から、歌謡曲を区間を認識することを目的としている。しかし、放送音には歌謡曲の他に、ナレーションや効果音などさまざまな音が途切れなく含まれているので、その認識は困難である。そこで、ステレオ信号で演奏される歌謡曲は、左右のチャンネルに音が散らばることが多い特性を利用し、左右のチャンネルの相関を調べる認識法を、本研究では提案する。実験の結果、本研究は、歌謡曲の認識に有効であることを確認した。

2 従来の研究とその問題点

従来、放送音を入力とした音楽、音声の分類システムがいくつか報告されていた。[1][2][3] これらは、音声認識の前処理として、音声区間の切り出しを目的としているので、これを歌謡曲認識に適用すると、以下の問題が生じる。

1. 音楽の定義が不十分
   従来のシステムでは、放送音を複数のカテゴリに分類していた。音声については、ＢＧＭ有り、ＢＧＭなし、など詳細なカテゴリを定義しているが音楽については、１種類しか定義していない。そのため、効果音の様な短い楽器演奏も音楽として認識されてしまい、歌謡曲かどうかは従来のシステムだけでは、認識不可能である。

2. 区間のセグメント
   従来のシステムでは、無音部分を利用し、放送音を区間に分類し、各区間でカテゴリに分類していた。しかし、ラジオの音楽番組では、歌謡曲とナレーションの区切れに、無音部分がない場合も多い。よって、無音部分のみをたよった従来のセグメント方法では、歌謡曲の切り出しができない場合もある。

3. 処理時間がかかる
   従来のシステムは、認識の前処理として、周波数解析 [1][2] やベクトル量子化の計算 [3] が必要であった。これらの計算量により、大量のデータを扱うと解析時間が長くなってしまう。よって、より速いアルゴリズムが必要である。

[1]Recognition of popular songs from a radio/TV music program
Masaki Nakayama,Yoichi Muraoka
School of Science & Engineering, Waseda University

3 歌謡曲の認識

本研究では、音響信号の波形ファイルを入力とし、歌謡曲を認識し、開始時刻と終了時刻を出力する。以下がその入力の制約である。

- 音楽番組であること
  ここでいう音楽番組とは、歌謡曲（洋楽，邦楽），ナレーション（ＢＧＭ入り，なし），トーク（ＢＧＭ入り，なし），効果音，コマーシャルが混在した番組のことである。
- ステレオ信号であること

下記 3.1,3.2 節でアルゴリズムを示す。

3.1 左右のチャンネルの相関の表を作成する

本節では、入力波形を 1.25 秒のフレームに区切り，各フレーム で 相関の値 $f_k(k=0,,,N-1)$ を計算する。N はフレーム総数である。
フレーム内左右の信号を $l_i,r_i(i=0,,,n-1)$ とする。n は フレーム内のサンプル数である。

1. 左右の音の差分を取る
   $s_i = l_i - r_i$
2. 左右チャンネルの音量の和 $(volLR_k)$、左右の音の差分の音量 $(volS_k)$ を計算する
   $volLR_k = \Sigma_{i=0}^{n-1}(|l_i| + |r_i|)$
   $volS_k = \Sigma_{i=0}^{n-1}(|s_i| + |s_i|)$
3. ２つの比を計算する
   $f_k = volS_k/volLR_k$

左右のチャンネルの信号の差分をとるので，f_k が大きいと，左右のチャンネルの信号がより独立していて，小さいと，より相関していることになる。

歌謡曲をステレオ信号で演奏する場合，各楽器が左右にちらばる傾向があり，左右のチャンネルはより独立した信号になる。よって，f_k が大きいものを時間的に追跡していけば，曲を認識できる。次節で，その追跡方法を述べる。

3.2 曲を時間追跡する

本節では、歌謡曲の演奏時間がコマーシャルや効果音に比べてある程度長いことを利用し，歌謡曲を認識する処理を行う。以下の4つのルールで，f_k を追跡する。

1. f_k がしきい値以上なら，音楽フレームとする
   今回の実装では経験則により、しきい値を 0.4 とした。
2. 音楽フレームが ある時間以上続くものを曲と認識する
   今回の実装では，90秒以上とした。音楽フレームではないフレームが来たところで，曲の終わりとするが，下記 3,4 を例外とする。
3. 音楽フレームに挟まれた，そうでないフレームは，曲の終わりにしない
   f_{k-1}, f_{k+1} が音楽フレームで，f_k が音楽フレームでない時，f_k は，曲の終わりとは認識しない。
4. 無音フレームが来たら曲の終わりとする
   $volLR_k$ が 最大音量の 10% 以下の場合，無音フレームとする。このフレームが来たら，曲の終わりとする。

4 計算量の評価

本研究の計算量を評価する。入力波形の総サンプル数を n とし、総フレーム数を N とする。

相関の表を作成するときは、比較 $3n$ 回、加算 $3n$ 回、乗算 N 回。曲を時間追跡するときは、比較 N 回である。比較を加算と考えると、本研究の計算量は、$6n + N$回 の加算と N 回の乗算である。

5 実験

サンプリング周波数 32000 Hz のデータで実験をした。以下が使用した音楽番組のデータである。

no	放送局	放送時間長	放送日
1	J-Wave(81.3Hz)	2:30	1/12 AM9:30
2	TVK(42ch)	1:00	1/11 PM3:00
3	TBS(6ch)	1:00	1/13 PM9:00
4	フジテレビ (8ch)	0:30	1/11 PM11:00

以下が実験結果である。

no	曲数	A	B	C	D
1	20	15	4	1	0
2	6	3	0	1	2
3	4	3	0	0	1
4	6	1	1	1	3

認識結果の記号は以下の通りである。

- A:曲を前後ともに認識成功 (5 秒までの余分なデータは認める)
- B:曲の前か後に，ＣＭなどの余分なデータも認識してしまった
- C:曲の途中に演奏音が弱い部分があり，それ以前までしか認識されなかった
- D:曲の演奏音が全体的に弱く，全く認識されなかった

番組 4 の認識がうまくいかなかったが，この番組では曲のＣＤがかかるのではなく，歌手は生演奏で，歌う。そして，その演奏がピアノ中心で，音量が小さかったため，本システムで認識できなかったと考えられる。また、曲でないのに，曲として認識されてしまったデータは１つもなかった。

6 まとめ

本研究では、放送音から、歌謡曲の特性を生かした認識を提案し、その実験結果を述べた。実験の結果、本研究は、音楽の中で、歌謡曲の認識に有効であることを確認した。今後は，認識結果の中で、認識出来なかったＢ，Ｃ，Ｄのデータを認識するように研究していく予定である。

References

[1] Michelle S.Spina,Victor W.Zue: "Automatic Transcription of General Audio Data:Preliminary Analyses" ICSLP'96 pp.594-597

[2] 河地吏司, 梶田将司, 武田一哉, 板倉文忠: "ＶＱ歪みに基づく放送音の自動分類" 電子情報通信学会 信学技報 SP97-50

[3] 木之下秀二, 有木康雄: "部分空間法を用いた音声・音楽・環境音の識別" 日本音響学会講演論文集 平成８年３月 2-5-4