7Q−6

スペクトル変動を用いた楽音分離の試み

植田　護　　橋本　周司　　大照　完
早稲田大学　理工学部

1　はじめに

人間は、片耳でも複数の音源からの音をその音色やダイナミックスによって聞き分けることが出来る。

このように、多種類の音が重なっているものの分離認識を計算機上で実現することは、自動採譜[1]、会話認識に重要であるばかりでなく、信号処理の問題[2]としても興味深い。

ここでは、短時間スペクトルからの音源分離の問題を分類整理し、それぞれの場合での楽器音分離の試みの概要を報告する。

2　原理

まず、問題を単純にするために、次の仮定に基づいて考える。

「それぞれの音源の各周波数成分は、各時刻に同一の倍率で変動する。観測音の周波数成分は、各音源の周波数成分の和である。」

つまり、$X(w)$は観測音の短時間パワースペクトル,$Z_a(w)$,$Z_b(w)$は音源A、Bの短時間パワースペクトルとして、

$$X(w) = a\cdot Z_a(w) + b\cdot Z_b(w) \qquad (2\text{-}1)$$

と表される。ここでは、a, bを各音源の強度と呼ぶことにする。

即ち、音源分離とは、$X(w)$から$Z_a(w)$, $Z_b(w)$, a, b を推定することである。

また、ここでは、推定の基準を、平均自乗誤差にとる。つまり、次式のEを最小にする各変数 $Z_a(w)$, $Z_b(w)$, a, bを求めるのが問題である。

$$E = \sum_w \{(X(w) - a\cdot Z_a(w) - b\cdot Z_b(w)\}^2 \qquad (2\text{-}2)$$

但し、$Z_a(w)$, $Z_b(w)$に関する何らかの情報なしに、式(2-2)のみから各変数を一意に推定することは困難である。

そこで、$Z_a(w)$,$Z_b(w)$について、前もってパワースペクトル分布の情報がある程度既知であって、モデル化できる場合と、モデル化できない場合に分けて考える。

特にここでは、モデル化可能な例として、楽器音の音源分離を取り上げる。

3　モデル化可能な音源の分離

a) 線スペクトルモデル

一般的に音階楽器音の周波数スペクトルは、音階によって決まる周波数の整数倍の周波数（倍音）によって構成されており、その倍音の強さの比が楽器の特徴を表す。

したがって、楽器音の短時間パワースペクトルは、それぞれの固有な離散線スペクトルでモデル化できる。

そこで、あらかじめ測定しておいた各楽器の各音階のスペクトルモデルを用いて、観測音中に存在すると推測されるスペクトルモデル $Z_{am}(w)$, $Z_{bm}(w)$を選定し、式(2-2)の最小化によって、音源の強度 a, bを算出する。観測音中の各音源の周波数スペクトルは、

$$Z_a(w) = a\cdot Z_{am}(w) \qquad (3\text{-}1)$$

$$Z_b(w) = b\cdot Z_{bm}(w) \qquad (3\text{-}2)$$

となる。

以上のような方法で、音源分離のシステムを構築した。

ただし、モデルデータの選定は、二つの音源が異なる音階の音を出しているとして、まず観測音の短時間パワースペクトルに各音階の倍音のみを通過させるフィルタを掛け、その通過量の大きいもの二つを選び、各音源の音階を認識する。次に、それぞれの音階のフィルタを通過した観測音のデータと各楽器のスペクトルモデルデータとのマッチングにより、楽器の同定を行う方法をとった。

On Sound Signal Separation using short term Supectrum
Mamoru UEDA, Shuji HASHIMOTO,
Sadamu OHTERU
Waseda University

今回は、ＭＩＤＩ音源からのホルン、及びバイオリンの音について実験を行ったが、二音が同じ音階でない限り、９０％以上の正解率で二音の音階、楽器の同定に成功した。また、シュミレーションのために作成したデータで実験した結果、分離した観測音中の音源のスペクトルの誤差は、１０％程度であった。

b) 連続スペクトルモデル

パーカッション等では、音響が、倍音構造をとらず、連続スペクトルになる場合が多い。このような場合、音源モデルは w の適当な関数として、パラメータを推定する問題とする事ができる。

一般に、音源Ａ、Ｂのスペクトルモデルをそれぞれ未知のパラメータ（α_a, α_b, β_a, β_b, ･･･）を含む形 $F_{am}(w, \alpha_a, \beta_a, \cdots)$、$F_{bm}(w, \alpha_b, \beta_b, \cdots)$で与えられる場合、

$$E = \sum_w \{X(w) - aF_{am}(w, \alpha_a, \beta_a, \cdots) - bF_{bm}(w, \alpha_b, \beta_b, \cdots)\}^2 \quad (3\text{-}3)$$

を、最小化するパラメータセット、a, b, α_a, α_b, β_a, β_b, ･･･を求めれば良い。

例えば、シンバルのクラッシュ音と、スネアドラムのロール音のスペクトルは、$\exp(-\alpha w)$ のαの相違である程度の近似が可能である。

この場合、両者の強度とαを推定するのは、放射能の減衰曲線から、核種を推定する問題と同等である。

4　モデル化なしでの分離

3.では楽器音を例にとって、モデル化できる信号の分離が可能であることを示した。ここで、一般的なモデル化できない信号についての分離について考察する。

式(2-1)のみから各変数を分離することは不可能である。

そこで、w を離散化したn点のスペクトル$(w_1, w_2, \ldots. w_n)$のデータがあるとして、時間方向には、m個のサンプリング点$(t_0, t_1, t_2, \ldots..t_{m-1})$をとり、強度 $a(t_j), b(t_j)$を考える。時刻 t_0に於いて、

$$a(t_0) = b(t_0) = 1 \quad (4\text{-}1)$$

とすれば、式(2-2)は、以下の式で表せる。

$$X(w_1, t_0) = Z_a(w_1) + Z_b(w_1) \quad (4\text{-}2)$$

$$X(w_1, t_1) = a(t_1)\cdot Z_a(w_1) + b(t_1)\cdot Z_b(w_1) \quad (4\text{-}3)$$

$$X(w_1, t_2) = a(t_2)\cdot Z_a(w_1) + b(t_2)\cdot Z_b(w_1) \quad (4\text{-}4)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$X(w_n, t_m) = a(t_{m-1})\cdot Z_a(w_n) + b(t_{m-1})\cdot Z_b(w_n) \quad (4\text{-}5)$$

観測音のデータ数は、n･m 個，求める変数の数は、2(m-1)+2n 個である.解を一意的に決めるには、上に示したデータの数と、変数の数が等しいことが必要である。これを満たす整数 n,mの最小のものは n = 4, m = 3 である。即ち、少なくとも w について４点、t について３点の１２個のデータを必要とする。

今回、式(4-2),(4-3),(4-4)を各周波数において満たす解の算出を最小自乗誤差基準で試みたが、非線形の逆問題となるので、一意的な解を得るための条件を得るまでには至っていない。

5　あとがき

式(2-1)の仮定の下に、音源分離問題を、モデル化可能な場合と、不可能な場合に整理して実験結果の一部と共に述べた。楽音をモデル化して分離するには、音階のある楽器を対象とする線スペクトルモデルと、音階の不明な楽音を対象とする連続スペクトルモデルの２種類が考えられる。しかし、モデル化が不可能な場合は、ブラインド分離の最も一般的な問題であるが、非線形になり、解析的に解くことは困難である。これについては、現在適当な制限を加えた解法を検討中である。

参考文献

[1] 柏野 邦夫、田中 英彦：“音源分離同定システムについての考察”、情報処理学会第４３回全国大会，7C-1 (1991)

[2] Barry Vercoe, David Cumming:"Connection Machine Tracking of Polyphonic Audio", ICMC Proceedings (1988)