音楽データの音源分離
楽曲からのボーカル抽出

2012SE151 三島隆大朗
指導教員：大石泰章

1　はじめに

複数の音源から生じた音声を，混合した状態からそれぞれの音源ごとに分離し，目的の音声のみを抽出する技術を音源分離という．

音楽において，複数の楽器の音をそれぞれ聞き分けることは困難であるが，各楽器ごとに分離して聞くことができれば，細かいテンポやピッチのずれを調整することが容易になると考えられる．音源分離はこのようなときに有効な手段である．また，会議のような複数の人が同時に喋っている場合における特定の人の声の抽出や，スマートフォンの音声認識に用いられるなど，音楽データのみではなく人の声への応用も考えられる．

音源分離には，独立成分分析（ICA）という手法 [1] を用いることが多かった．しかし，この手法を用いた場合，音源の数と同じ数のマイクロホンが必要となる．つまり，音源の数より観測信号の数のほうが多いか，もしくは同じ数でなければいけないという条件がある．

そこで近年，研究が進められているのが，スパース性を用いた音源分離法である．この手法は音源の数よりマイクロホンの数が少なくても使える手法である．

音楽データからのボーカル抽出は，様々な方法が研究され，実用化されている．例えば，CD がステレオ音源であり，ボーカルが中央に定位していることを利用して抽出を行う方法などがある．

本研究では，最近のシングル CD が，1 曲目にメインの曲，2 曲目にカップリング曲，3 曲目にメイン曲の Instrumental（楽器の音のみでボーカルが入っていない曲），4 曲目にカップリング曲の Instrumental という構成になっていることに着目する．このことを利用して，ある楽曲からボーカルのみを抽出する方法について研究を行った．

2　抽出の手法

2.1　基本的な考え方

今回行った実験は，波の重ね合わせの原理を利用した方法である．波の重ね合わせとは，ある波とそれと全く同じ位相（同位相）の波を足し合わせると，合わさった波の山が倍に膨れ上がり，大きな波となる．これを音で考えると，音が大きくなるということである．

それに対して，ある波とそれと逆の位相（逆位相）の波を足し合わせると，合わさった波の山が打ち消しあい，波がなくなる．これを音で考えると，音が消えるということになる．

ボーカルのみの音を $s_1(n)$，楽器音のみの音を $s_2(n)$ とする．n は時間を表す．このとき，通常の楽曲は $s_1(n)+s_2(n)$ と考えられる．これに楽器音のみの音を逆位相にして足し合わせると，$s_1(n)+s_2(n)+(-s_2(n))=s_1(n)$ となり，ボーカルのみの音が抽出できると考える．

抽出が不完全な場合は，さらにバイナリマスキング法 [2] を用いて精度を高める．以上が，本研究で試みる抽出法の基本的な考え方である．

2.2　対象とする音声データ

本研究では，ボーカルと楽器音の混ざっている普通の楽曲の音源と，楽器音のみの音源を用いて実験を行った．これらを使い，Matlab，Scilab を用いて処理を行った．Matlab では扱える音声の形式が決まっているので，それに合わせて形式の変換を行った．

まず，通常 CD は MP3，もしくは MP4（動画ファイル）の形式であることが多く，またステレオ音源である．Matlab では MP3 や MP4 の形式のデータは取り扱いができないので，Matlab でも扱える Wave 形式への変換を行った．音源はステレオ音源からモノラル音源へ変換して用いた．またメモリの関係上，それぞれのデータを 5 秒程度に短縮して用いた．

3　抽出実験

3.1　波の重ね合わせを用いた結果

図 1 は，ボーカルと楽器音の混ざった通常楽曲のグラフである．

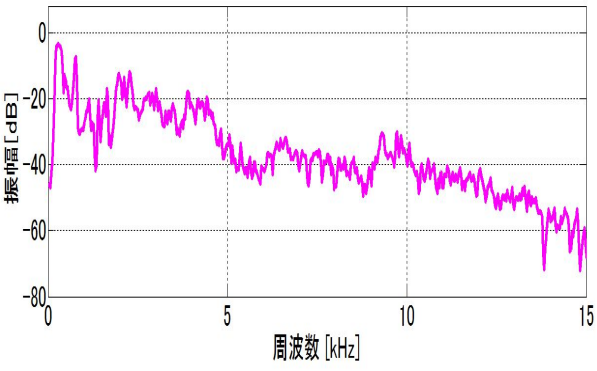

図 1　通常楽曲の音源 $s_1(n)+s_2(n)$

図 2 は，楽曲の楽器音のみのグラフである．

図 3 は，通常楽曲の音源に楽器音のみの音源を逆位相にして足し合わせたもののグラフで，この信号を $x_1(n)$ と書く．理想的には，ボーカルのみが抽出されるはずである．

$x_1(n)$ の音声を聴いてみると，楽器音は小さくなったも

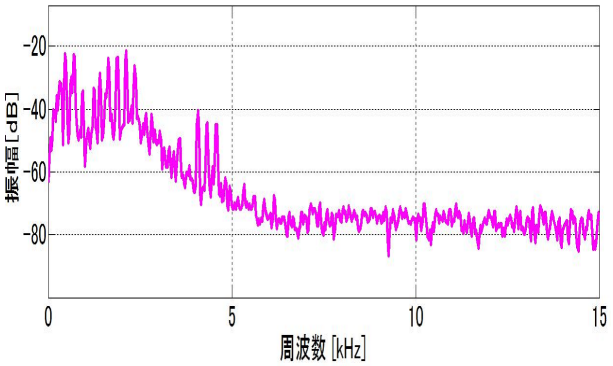

図 2　楽器音のみの音源 $s_2(n)$

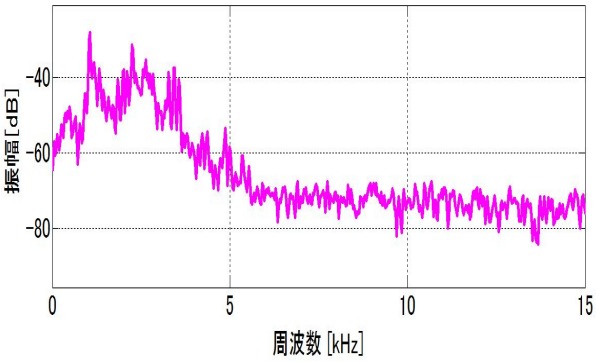

図 3　波の重ねあわせで得た音声 $x_1(n)$

のの，完全には消すことができなかった．波の重ね合わせは全く同じ位相のものを逆にしたものを足し合わせているので，少しでもずれが生じると精度が落ちてしまう．今回の実験では，そのずれが生じたことが原因であると考える．

そこで，バイナリマスキング法を用いて高い精度の抽出を試みた．

3.2　バイナリマスキング法を用いた結果

前節で得た音声 $x_1(n)$, 楽器音のみの音源 $s_2(n)$ を $x_2(n)$ として，それぞれに短時間フーリエ変換を行ったものを，$X_1(n,\ f)$, $X_2(n,\ f)$ とする．このとき，f は周波数を表す．

各 n，各 f で $|X_1(n,f)|$ と $|X_2(n,f)|$ を比較する．$|X_1(n,f)| \geq |X_2(n,f)|$ のとき，出力 $Y_1(n,f)$，$Y_2(n,f)$ を

$$\begin{cases} Y_1(n,f) = X_1(n,f), \\ Y_2(n,f) = 0 \end{cases}$$

のように定める．また，$|X_1(n,f)| < |X_2(n,f)|$ のときは，

$$\begin{cases} Y_1(n,f) = 0, \\ Y_2(n,f) = X_2(n,f) \end{cases}$$

のように定める．

そして，$Y_1(n,f)$ と $Y_2(n,f)$ を逆フーリエ変換してもとに戻すと，それぞれ $y_1(n)$，$y_2(n)$ となる．このときの $y_1(n)$ はボーカルのみの音声に近いと考えられる．図 4 がバイナリマスキング法を用いて抽出した音声 $y_1(n)$ のグラフである．

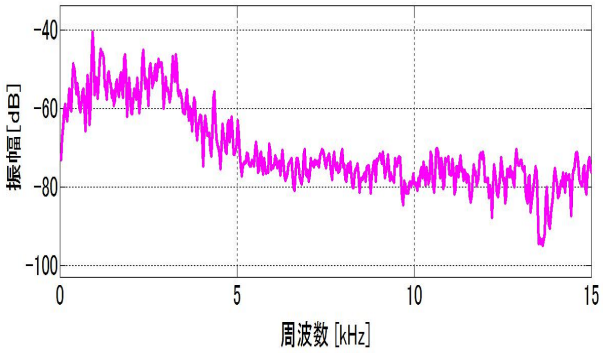

図 4　バイナリマスキングを用いて得た音声 $y_1(n)$

$y_1(n)$ の音声を聴いてみると，$x_1(n)$ よりも高い精度でのボーカル抽出ができた．

4　おわりに

本研究では，音楽 CD を用いて，通常の楽曲とその楽器音のみの音源を使い，ボーカルの抽出を試みた．

波の重ね合わせの原理を用いた結果は，楽器音は小さくなったものの，完全には消すことができなかった．その音源に対して，バイナリマスキング法を適用し，より高い精度でのボーカル抽出を行い成果を得た．

ボーカル抽出ソフトとして実用化されているものは，周波数カットのフィルタ [3] を使用しているものが多いので，今後，バイナリマスキング法を用いて行う手法 (本研究の手法) との精度の比較を行いたいと考える．また，楽器音のみの音源が収録されていない楽曲から，ボーカルのみを抽出するためのシステムを考えることが今後の課題としてあげられる．

5　参考文献

[1] 澤田宏・荒木章子・牧野昭二：「音源分離技術の最新動向」．電子情報通信学会誌， Vol. 91， No. 4， 2008.

[2] 川村新・尾知博：「特定の音声を抽出する方法：マイクロホンによる観測信号から複数の音源を分離する技術」，『音声＆画像処理の常識： AV 機器で必須の MP3 も MPEG-4 もよく分かる』，CQ 出版社，東京，2010，pp. 51–63.

[3] 川村新・尾知博：「音声信号処理の基礎理論：音声圧縮，ノイズ除去，音源分離で用いられる理論」，『音声＆画像処理の常識： AV 機器で必須の MP3 も MPEG-4 もよく分かる』，CQ 出版社， 東京， 2010， pp. 101–119.