ステレオ多楽器音中のドラムパートの抽出

2011SE144 倉内 慎　2011SE154 増田 大輝
指導教員 後藤邦夫

1　はじめに

近年，楽譜の普及が増えたが，ドラム譜だけを見ても正確には書かれておらず，また，私がドラムを演奏するにあたってコピーしようとしたアーティストのものがないということが多々あった．そのとき，私は動画サイトを見て，実際に演奏している姿を真似してコピーをしてきた．しかし，それは効率が悪く素人である私にとってとても難しいことであった．

先行研究 [4] では多楽器音中からベースギターパートの譜面を作成することを目的としており，抽出したパートを音声化することはしていない．ドラムは他の楽器とは異なり，1 つ 1 つの音がはっきりしている．そのため，音のみを聞けばどこを叩いているのかがわかりやすい楽器であるので，ドラム音のみを音声化することによってドラムを演奏する人にとって大きな手助けとなる．当研究ではフーリエ変換やメディアンフィルタ，ステレオフィールド内の panpot などを使用し，周波数の分析をして，多楽器で演奏された楽曲からドラム音のみを抽出する．ステレオ音源ファイルを作成したプログラムに出力し，ドラム音のみがどれだけきれいに聞こえるかどうかで評価する．文中に出てくる percussive な音というのは瞬間的に出る音，harmonic な音というのはある程度継続する音である．主に，倉内慎は panning による音の分離を担当し，増田大輝がメディアンフィルタによる音の分離を担当した．

2　周波数分離と計算方法

ここでは，ドラムパートを抽出するときの問題を定義し，その解決策を示す．

2.1　周波数による音の分離

図 1 は縦軸を周波数，横軸を panpot で表したものである．図 1 から見ても分かるように周波数領域だけでは楽器ごとに被ってしまっており，特定した音のみを分離することは不可能である．よって本研究では percussive な音の分離，panning による音の分離を用いてドラムパートを抽出していく（対象とするパートはドラム，ヴォーカル，ギター，ベースギター，キーボードとする）．

2.2　STFT

STFT（short-time Fourier transform）は短時間フーリエ変換のことを表し，関数に窓関数をずらしながらかけて, それにフーリエ変換することである．当研究では，音声を対象としているため 1024 のサンプル数でフーリエ変換する．音声を対称としているため，合成する際に少しのずれがあってはならない．つまり，細かく区切ってフーリエ変換する必要がある．1024 は時間にして約 0.02 秒ほどであり，音声がずれてはいけないという 0.03 秒より少ないため，当研究ではこのサンプル数でフーリエ変換している．STFT する例を図 2 に示す．

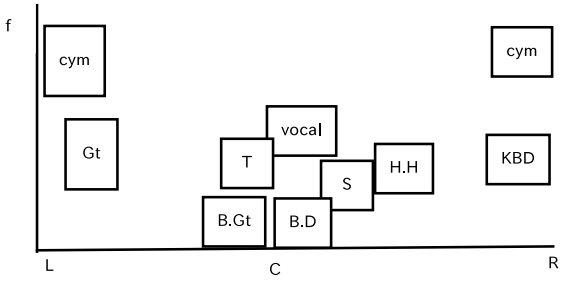

図 1　各パートごとの周波数と定位

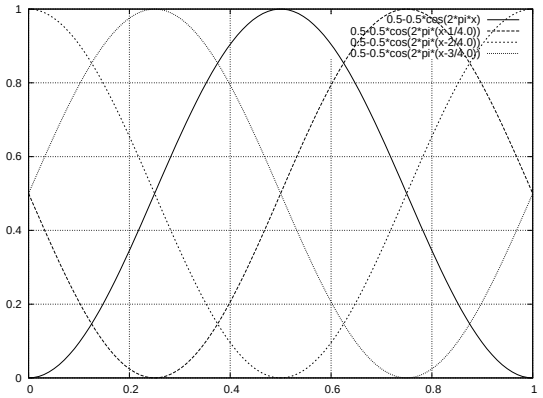

図 2　1/4 ずつずらした STFT の例

3　panning による音の分離 [1]

panpot を用いて，ステレオフィールド内の左右の音声出力バランスの差に基づいて分離する．panpot の領域で取り出したい音の周波数に依存する場所を明かにするため，ゲインの大きさを調整し位相を消す．その後，周波数の大きさを推定し，再合成する．

3.1　panpot [3]

panpot（panoramic potentiometer）とは音の左右の定位のこと．ステレオスピーカの場合の音の出力先を左右に割り振る際に使用する．範囲は 0 から 127 までで，64 がちょうどセンターにあたる．当研究では 0 をセンターとし，左寄りをマイナス，右寄りをプラスとしている．

3.2　分離方法

右から聞こえる音と左から聞こえる音をそれぞれ $L(t)$, $R(t)$ とし，式 (1)，(2) を使いそれぞれのフーリエ変換

$\mathcal{F}L(k), \mathcal{F}R(k)$ を計算する．

$$\mathcal{F}L(k) = \sum_{n=0}^{N-1} L(n) w_n^{kn} \tag{1}$$

$$\mathcal{F}R(k) = \sum_{n=0}^{N-1} R(n) w_n^{kn} \tag{2}$$

$1 \leq k \leq N$, $w_n^{kn} = e^{-j2\ \pi/N}$. a, b は $L(t)$ と $R(t)$ の比率と考えるので，$0 \leq a \leq 1, 0 \leq b \leq 1$．このとき
右の音が大きい場合

$$F_R(k) = |\mathcal{F}L(k) - a\mathcal{F}R(k)| \tag{3}$$

左の音が大きい場合

$$F_L(k) = |\mathcal{F}R(k) - b\mathcal{F}L(k)| \tag{4}$$

として，指定した位置に近い音を取り出すので，それぞれ式 (3)，(4) が最小になるときの a, b を求める．参考文献 [1] では全通り計算して a, b の値を出していたが，それは効率が悪いので $F_R(k), F_L(k)$ を 2 乗して出てきた二次関数の最小，最大になるときの a，b を求めた．この二次関数の軸を d とおく．$\mathcal{F}L(k) = x_L + iy_L$, $\mathcal{F}R(k) = x_R + iy_R$ とおくと，$absL = \sqrt{x_L^2 + y_L^2}$, $absR = \sqrt{x_R^2 + y_R^2}$, $absLR = \sqrt{(x_L - x_R)^2 + (y_L - y_R)^2}$ となる．ここで，$\mathcal{F}L(k)\overline{\mathcal{F}R(k)}$ の実部（内積）を dp，虚部（外積）を cp とおくと，$dp = x_L x_R + y_L y_R$, $cp = x_R y_L - x_L y_R$ となる．$d = \frac{dp}{(absR)^2}$ と表すことができる．右の音が大きい場合，

- $d < 0$ のとき

$$\begin{cases} min : absL & (a = 0) \\ max : absLR & (a = 1) \end{cases}$$

- $0 \leq d < \dfrac{1}{2}$ のとき

$$\begin{cases} min : \dfrac{abscp}{absR} & (a = d) \\ max : absLR & (a = 1) \end{cases}$$

- $\dfrac{1}{2} \leq d < 1$ のとき

$$\begin{cases} min : \dfrac{abscp}{absR} & (a = d) \\ max : absL & (a = 0) \end{cases}$$

- $1 \leq d$ のとき

$$\begin{cases} min : absLR & (a = 1) \\ max : absL & (a = 0) \end{cases}$$

次に最大値から最小値引いて，その時間での周波数の大きさを推定する．また，その時間以外での周波数を 0 とする．

- $d < 0$ のとき

$$max - min = absLR - absL \tag{5}$$

- $0 \leq d < \dfrac{1}{2}$ のとき

$$max - min = absLR - \frac{abscp}{absR} \tag{6}$$

- $\dfrac{1}{2} \leq d < 1$ のとき

$$max - min = absL - \frac{abscp}{absR} \tag{7}$$

- $1 \leq d$ のとき

$$max - min = absL - absLR \tag{8}$$

この一連の手順を，左の音が大きい場合についても同様に計算する

3.3 再合成

3.2 で求めた周波数の大きさを式 (9) で足し合わせていく．$i =$ 方位, $c =$ 取り出す位置, $H =$ 予備空間 として，

$$G_R(k) = \int_{i=c-H/2}^{i=c+H/2} F_R(k) \tag{9}$$

綺麗に再合成するため，取り出したい方位の前後に幅を持たせる．次に，元の位相の偏角を式 (10) を使い実数と虚数の 2 つに分解する．

$$J(k) = \left\{ \begin{array}{l} ReJ(k) = G(k) cos\angle(\mathcal{F}R(k)) \\ ImJ(k) = G(k) sin\angle(\mathcal{F}R(k)) \end{array} \right\} \tag{10}$$

その後，式 (11) を使い J(a) を逆フーリエ変換して音声に戻す．

$$J(n) = \frac{1}{N} \sum_{k=1}^{N} J(k) W_n^{-kn} \tag{11}$$

4 percussive な音の分離 [2]

メディアンフィルタを用いてモノラル音声信号の harmonic な音と percussive な音を分離する．この技術は，メディアンフィルタの連続するフレームにおいて harmonic な音を高めるために percussive な音を抑制し，また逆のこともする．得られた二つのメディアンフィルタリングスペクトログラムは harmonic な音と percussive な音を分離するために，元のスペクトログラムに適用するマスクを生成するために使用される．私たちは youtube の音源から素材を抽出し，リミックスする．

4.1 メディアンフィルタ

与えられたサンプルの中央値をとり，その値を出力すること．$y(n) =$ 中央値の出力, $x(n) =$ 入力ベクトル，$\alpha =$ サンプルの数, $k = (\alpha - 1)/2(\alpha$: 奇数), $k = \alpha/2(\alpha$: 偶数) として，出力する値は式 (12) から導く．

$$y(n) = median\{x(n-k : n+k)\} \tag{12}$$

メディアンフィルタは画像処理において広く使用されている．

4.2 前提条件

- 音声サンプルは wav
- メディアンフィルタの長さ=17
- FFT の長さ=4096
- 1024（4096/4）ずつずらす
- 周波数 44.1kHz

参考文献 [2] では，メディアンフィルタの長さ 15 から 30 未満の数字で実験をしており，その結果，劇的には変化しなかったため長さ 17 で実験をしている．周波数 44.1kHz に対し FFT の長さ 4096Hz（約 0.1 秒）の長さであり，窓関数を 1024Hz（0.025 秒）に設定したのは音楽が 0.03 秒ずれると致命的な問題となり，また短すぎると低音の分析ができなくなってしまうからである．

4.3 分離の流れ

- 音声サンプルを STFT して，時間軸から周波数軸に変換する．
- 出てきた周波数の共役複素数をスペクトログラムに表し，図 3 のように harmonic 要素の集まり（$H_{k,i}$）と percussive 要素の集まり（$P_{k,i}$）にメディアンフィルタをかける（それぞれ $MH_{k,i}$，$MP_{k,i}$ とする）．

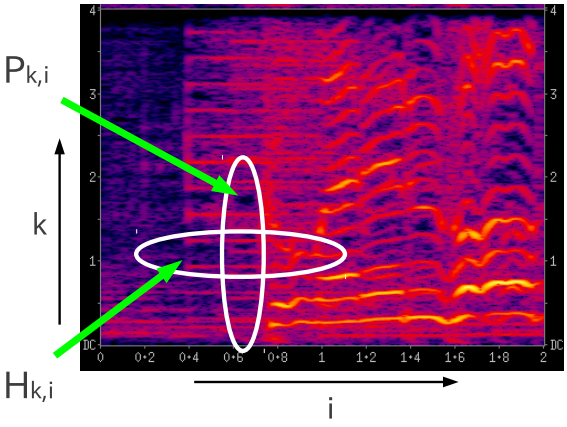

図 3　2 秒間のスペクトログラム

- 図 4 は図 3 のメディアンフィルタをかけた部分を拡大したものである．$H_{k,i}$, $P_{k,i}$ それぞれの要素を $h_{k,i}$, $p_{k,i}$ と表している．
- $MH_{k,i}$，$MP_{k,i}$ の掛け比率は式 (13)，(14) のように表すことができる．

$$MH_{k,i} = \frac{H_{k,i}^2}{H_{k,i}^2 + P_{k,i}^2} \tag{13}$$

$$MP_{k,i} = \frac{P_{k,i}^2}{H_{k,i}^2 + P_{k,i}^2} \tag{14}$$

- 求めた $MH_{k,i}$, $MP_{k,i}$ それぞれの要素 $Mh_{k,i}$, $Mp_{k,i}$ を式 (15)，(16) を使い，行列計算をして足し合わせていく．$\hat{S}$ は元のフーリエ変換．

$$\hat{H}_{k,i} = \hat{S}_{k,i} \otimes MH_{k,i} \tag{15}$$

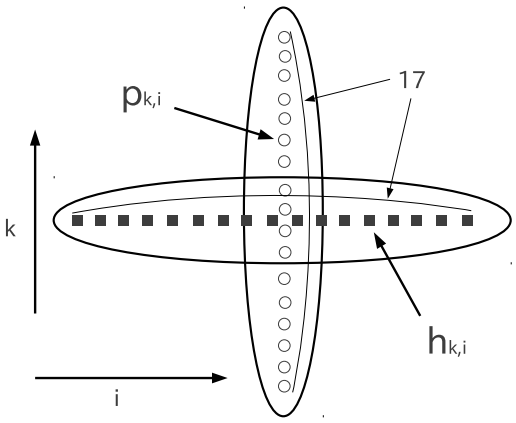

図 4　拡大図

$$\hat{P}_{k,i} = \hat{S}_{k,i} \otimes MP_{k,i} \tag{16}$$

- 最後に $\hat{H}$ と $\hat{P}$ を逆フーリエ変換（Inverse Fast Fourier Transform，IFFT）して，音声に戻す．

5 実験結果

5.1 panning による音の分離

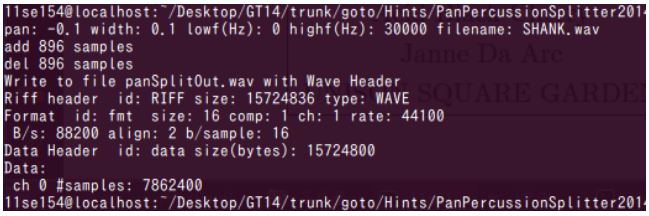

図 5　panning のコマンド

- youtube-dl を使って youtube から音楽ファイルを取り入れる．
- このプログラムは拡張子が wav でないとできないため，wav でない場合拡張子を変換する．
- 「./panSplit ファイル名」とコマンドを入力すると，図 5 のように分離した音が生成される．この時，最後に入力する値は panpot の位置であり，0 がセンター，1 が右端，-1 が左端と設定してある．

生成される音は入力した panpot の位置で抽出されたものと，それ以外の音の 2 種類である．ステレオ音源におけるドラムの位置は曲や構成によって大きく変化しないためドラム音のみの音源を作成し panpot の位置を調べた．具体的な値を入力したとき，どの音がどの位置から発しているか表 1 に示す．

以上が実験結果である．ドラムの大体の位置を特定することができたので多楽器の音源を分離する際もその位置周りをとることで良い結果が得られると考える．

5.2 percussive な音の分離

- 1：youtube-dl を使って youtube から音楽ファイルを取り入れる．

表 1　入力結果

楽器	pan 位置
スネアドラム	0.0
バスドラム	0.1
ハイタム	-0.6
ロータム	0.9
フロアタム	-0.9
シンバル	1.0(-1.0)
ハイハット	-1.0

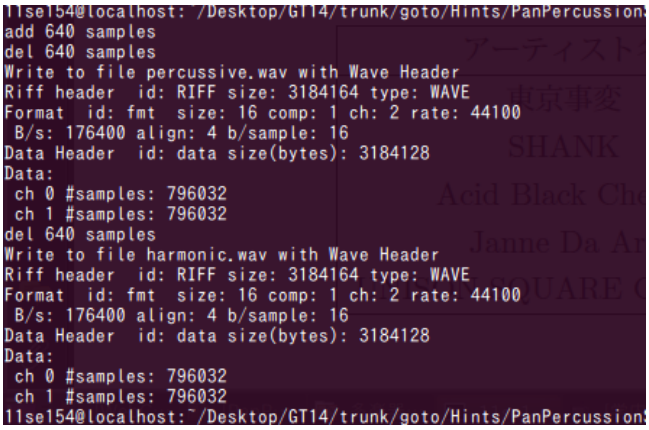

図 6　percussive のコマンド

- 2：取り込んだ音楽ファイルの拡張子が mp3 となっているため wav ファイルに変換する．
- 3：「./panSplit ファイル名」とコマンドを入力すると，図 6 のように実行される．

上記の手順を踏まえ実行すると，percussive な音と harmonic な音が別々に生成され，それぞれ再生することができる．表 2 に示したように構成がバラバラな曲を用いて実験をした．このとき percussive な音を再生するとドラムパートの他にヴォーカル，ギター，キーボードの音が入っている．これらの音が入ってしまった理由は，percussive は打楽器という意味なのでギターを弾く，ヴォーカルが発声する，キーボードを叩くというアタックが percussive として認識されてしまっているためであると考えている．

表 2　percussive 分離

アーティスト名	タイトル名	構成	perccusive な音
東京事変	群青日和	女 Vo*1,Gt*1,Key*1,Ba*1,Dr*1	Vo,Gt,Dr,Key
SHANK	Cigar store	男 BaVo*1,Gt*1,Dr	Vo,Gt,Dr
Acid Black Cherry	イエス	男 Vo*1,Gt*2,Ba*1,Dr*1,Key*1	Vo,Gt,Dr,Key
Janne Da Arc	ダイヤモンドヴァージン	男 Vo*1,Gt*1,Ba*1,Dr*1,Key*1	Vo,Gt,Dr,Key
UNISON SQUARE GARDEN	オリオンをなぞる	男 Vo*1,Gt*1,Ba*1,Key*1	Vo,Gt,Dr,Key

5.3　合成プログラム

panning プログラム，percussive プログラムの 2 つのプログラムを合わせてドラムパート抽出をする．panning プログラム→ percussive プログラムの順番で実行する．以下の手順で進める．

- 1：-1 から 1 の範囲の pan 位置を 0.1 ずつ区切り，21 個に分割．
- 2：分割したものを 1 個ずつ panning プログラムにかける．
- 3：できた音源をまた 1 個ずつ percussive プログラムにかける．
- 4：2 つのプログラムにかけてできた 21 個の音源を再合成．

しかし，panning で分割した音を percussive のプログラムで処理しようとするとエラーが起きてしまい解決することができなかった．エラーが起きた理由は，ステレオ音源を panning で処理した時にモノラル音源に変わり，percussive プログラムがステレオ音源用になっているためチャンネルの数が合わないからである．また再合成プログラムもできていないため，分割した音源をつなげることができなかった．

6　おわりに

当研究では多楽器で演奏された楽曲からドラム音のみを抽出し，音声化することを試みた．メディアンフィルタ，ステレオフィールド内の panpot を使用し，周波数分析してそれぞれのプログラム実行，音声化は成功した．当研究の成果は，

- メディアンフィルタによる percussive な音と harmonic な音の大まかな分離と音源の再生．
- pan 位置によって分割した音源の再生．

しかし，2 つのプログラムを合わせることができず，ドラムパート以外の音も入ってしまう結果となってしまった．今後の課題は，

- モノラル音源に対応したプログラムの作成．
- pan 位置によって分割した音源を再合成するプログラムの作成．
- ドラムパート以外の音を消すための精度向上．

参考文献

[1] Barry, D.: Sound Source Separation: Azimuth Discrimination and Resynthesis, *7th International Conference on Digital Audio Effects* (2004). http://eprints.maynoothuniversity.ie/696/.

[2] Fitzgerald, D.: Harmonic/Percussive Separation using Median Filtering, *13th International Conference on Digital Audio Effects* (2010). http://arrow.dit.ie/argcon/67/.

[3] 小原肇晃：定位の基礎知識 (accessed Oct. 2014). http://www.hikari-ongaku.com/study/panpot.html.

[4] 竹内雅浩，鵜飼立樹：多楽器音のベースギターパート自動採譜，南山大学情報理工学部システム創成工学科 2013 年度卒業論文 (2014).