多重HPSS法によるモノラル音楽音響信号に対するボーカル抑圧*

◎橘秀幸, 小野順貴, 嵯峨山茂樹 (東大院・情報理工)

1 はじめに

音楽音響信号には通常，多種の音が混合されている．このような混合信号の中から，特にボーカル成分を抑圧する技術は，カラオケなど音楽市場での需要があり，すでに多数の実装が知られている．これらの多くは，現代のレコーディング・ミキシングにおいてボーカルを中央に定位させるという慣習に基づき，ステレオ信号の左の波形から右の波形を減算することによりボーカル成分を相殺する手法を用いている．しかしこの手法は，ライブ録音などのようにボーカルの定位が一定しない場合などにはあまり効果が見込めない．

これに対し，より広範囲な信号に対して適用可能なボーカル抑圧/抽出の手法として，これまでに様々なものが提案されている [1, 2, 3, 4].

本稿では，スペクトログラムの時間周波数分解能を変えたときの，ボーカル成分音と楽器音の形状の変化という新しい着眼点によって両者を分離する多重 HPSS 法を提案し，それによりボーカル成分を抑圧する．

2 多重 HPSS 法によるボーカル成分抑圧

2.1 調波打楽器音分離 (HPSS)

調波打楽器音分離 (Harmonic/Percussive Sound Separation: HPSS)[5, 6, 7, 8] は，本研究室で開発された，モノラル音楽音響信号から調波楽器音と打楽器音を分離する手法である．

調波楽器と打楽器の混合信号を考え，それを短時間フーリエ変換 (STFT) したとする．そのとき，調波楽器は比較的長時間にわたり一定の周波数が鳴り続けるため，時間方向の筋となって現れる．一方，打楽器は瞬間的に多くの帯域を占めるため，周波数方向の筋となって現れる．HPSS ではこのことを利用して，個々の楽器音の調波構造に関する事前知識を用いることなく，調波楽器的な成分 (H) と打楽器的な成分 (P) に関する以下の 3 つの指針

- H の時間方向の変化が可能な限り小
- P の周波数方向の変化が可能な限り小
- H と P の和は原信号に可能な限り一致

のもとで目的関数を設計し，その最適化により H，P を求める手法である．

これまでの研究では，楽器毎の特徴として，ピアノ，ギター，バスドラム，ボーカルが H に分離される傾向にあり，スネアドラムやハイハットが P に分離される傾向にあることがわかっている．

2.2 時間周波数分解能とボーカル成分の形状

このような楽器毎の傾向は，通常の時間周波数分解能で計算したスペクトログラム上での HPSS におけるものであり，一般の時間周波数分解能では必ずしも成り立たない．特に，STFT のフレーム長を通常よりも長く与えた場合，ボーカル成分が H ではなく P に分離されるようになるという性質がある．これは，ボーカル成分のピッチやパワーなどのゆらぎ (短時間変動) に由来すると考えられる．

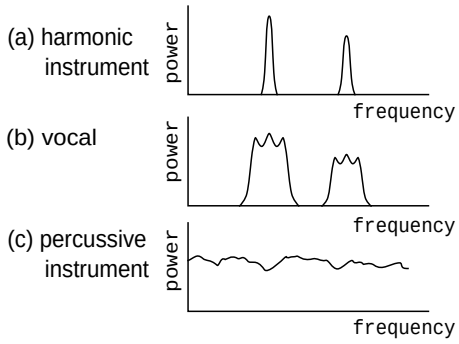

Fig. 1 Vocal spectrum occupies broader bandwidth than harmonic instrument on long-framed STFT domain, which is regarded as “Percussive sound” by HPSS.

フレーム長が 10[ms] 程度である場合，フレーム内でのボーカル成分のゆらぎは無視できる程度であり，短時間スペクトル上でエネルギーは特定の周波数に局在する．これに対し，フレーム長が 100[ms] から 1[s] 程度であると，フレーム内にボーカル成分のゆらぎが収まる．すなわち振幅や周波数の変調が短時間スペクトル上に現れ，周波数と振幅が一定の無変調信号に比べ広い帯域にエネルギーが分散される．言い換えれば，ボーカル成分は，ピアノやギターなどの調波楽器音よりも広い帯域を占有する (Fig. 1).

HPSS は，打楽器ほど広帯域を占有していない音でも，周波数方向にある程度のクラスタをなしている音は P に分離する．したがって，ボーカル成分のこのような性質を検出するのに適し，実際にボーカル成分の多くは P に分離されやすくなる傾向にある．

2.3 多重 HPSS 法によるボーカル成分抽出

以上のようなボーカル成分の性質を利用すれば，HPSS を 2 段階で用いることにより，ボーカルに相当する成分を強調することができる．すなわち，長いフレーム長 l_{long} でのスペクトログラム上での HPSS により H（調波楽器）と P（打楽器+ボーカル）を分離し，次に P を一度波形に戻してから再び短いフレーム長 l_{short} で STFT し，そのスペクトログラム上で HPSS して H（ボーカル）と P（打楽器）を分離することで，調波楽器，ボーカル，打楽器の混合信号からボーカル成分を抽出することができる．

この手法を，多重 HPSS 法と呼ぶこととする．

---

*Vocal Sound Suppression in Monaural Audio Signals by Multi-stage Harmonic-Percussive Sound Separation(HPSS). by TACHIBANA Hideyuki, ONO Nobutaka and SAGAYAMA Shigeki (Graduate School of Information Science and Technology, The University of Tokyo).

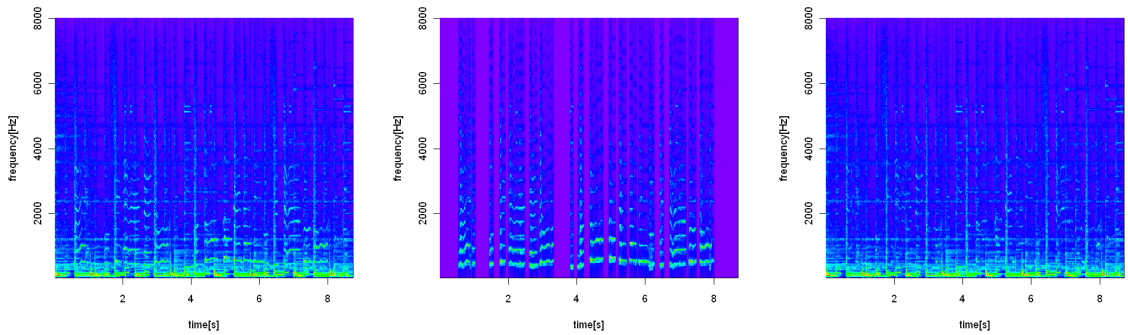

Fig. 2 Left: Original spectrogram of "tell me"(RWC-MDB-P-2001 No. 25). Center: Extracted vocal component ($\mathbf{V}'$). Right: Vocal-suppressed spectrogram.

2.4 時間周波数マスクによるボーカル抑圧

多重 HPSS 法で抽出したボーカル成分 $\mathbf{V} = \{V_{t,\omega}\}_{1\leq t\leq T, 1\leq\omega\leq\Omega}$ では，ボーカル成分が他の成分よりも強調されている．したがって，ボーカル成分 $\mathbf{V}$ のうち特にパワーが大きい成分を用いてバイナリマスクを設計すると，それを原信号に掛けることでボーカルを抑圧することができる．

スペクトログラム $\mathbf{V}$ の各成分のうち，ある閾値を超えたものを 0, 超えないものを 1 とするような直接的な 2 値化によりマスクを得るのも一つの方法だが，ここではボーカル成分の持つ倍音構造と，時刻上でボーカル成分が存在する区間と休止区間があることを利用して，より適切なマスクを設計することを考える．

倍音構造を持つマスクを設計するため，まず各時刻で $\mathbf{V}$ のピッチを抽出する．これにより，F_0 および倍音のみを抑圧するような各時刻のマスク $\mathbf{m}_t$ が設計できる．また，それらを並べることにより，時間周波数マスク $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_T]$ を作ることができる．

ここで得たマスクを単純に原信号に掛ければ，ボーカル成分を抑圧した信号を得ることができるが，$\mathbf{M}$ ではボーカルが休止している場合でも何らかの成分が遮断されるため，ここで得られた信号では前奏などでも一部の音が抑圧されてしまう．これを防ぐため，抽出したボーカル成分にマスクを反転して掛けたスペクトログラム $V'_{t,\omega} = (1 - M_{t,\omega})V_{t,\omega}$ では，ボーカル以外の成分がほとんど抑圧されていると仮定し，エネルギーがある閾値 θ に満たない時刻 $\{t|\sum_\omega (V'_{t,\omega})^2 < \theta\} =: \Theta$ ではボーカルが歌っていないものとみなした．これにより，マスクを $M'_{t,\omega} = M_{t,\omega}(t \notin \Theta), 1(t \in \Theta)$ と再設計し，これを原信号に掛けることで，より適切なボーカル抑圧信号が得られると考えられる．

以上を、本稿で提案するボーカル抑圧手法とする．

3 実音楽信号による実験

3.1 実験条件

実験には，RWC 研究用音楽データベース [9] よりポピュラー音楽，ジャズ，著作権切れ音楽を 16 kHz のモノラル音楽音響信号に変換して用いた．多重 HPSS 法の STFT のフレーム長は $l_{\text{long}} = 512$[ms], $l_{\text{short}} = 16$[ms], フレームシフトはフレーム長の半分，窓関数は分析窓，合成窓ともにハニング窓の平方根とした．ピッチ抽出にはラグ窓を用いた方法 [10] を使用し，抽出したピッチを f_t としたとき，マスクの各成分を

$$M_{t,\omega} = \begin{cases} 0 & \text{if } |\omega - nf_t| \leq 62\text{Hz } (n = 1, 2, \cdots) \\ 1 & \text{otherwise} \end{cases}$$

とした．なお，62 Hz は，時間分解能が 16[ms] のときのスペクトログラムの周波数分解能である．マスクの再設計に用いる閾値は，$\mathbf{V}'$ のエネルギーの冪乗変換 $\{\sum_\omega (V'_{t,\omega})^2\}^{0.1}$ に対する判別分析により決定した．

3.2 実験結果

RWC 研究用音楽データベースのポピュラー音楽より，"tell me"(RWC-MDB-P-2001 No. 25) に対して提案法を適用した結果を Fig 2 に示す．原信号のスペクトログラム（左図）の 2[s], 4.5 [s], 7[s] 付近に顕著に見られるボーカル成分が，$\mathbf{V}'$（中央図）に抽出され，右図では抑圧されている様子が観察できる．また聴感上も，ボーカル成分が抑圧されたことがこの楽曲を含む複数の楽曲において確認された．一方，子音は抑圧されにくく，トランペットなどの楽器が抑圧されやすいなどの傾向が見られた．

4 まとめ

本稿では，スペクトログラムの時間周波数分解能を変えた 2 段階の分析によりボーカル成分を抽出し，それにより設計した時間周波数マスクでボーカル成分を抑圧する手法を提案，および聴感的な検証をした．今後は，マスクの改良や，子音の扱いの検討，定量的な性能評価などを進めていく予定である．

謝辞 本研究の一部は科学技術振興機構 CREST プロジェクトの補助を受けて行われた．

参考文献

[1] Lagrange *et al.*, *IEEE Trans. ASLP,* Vol.16, No.2, pp.278-290, 2008.
[2] Li, Wang, *IEEE Trans. ASLP,* Vol.15, No.4, pp.1475-1487, 2007.
[3] Ozerov *et al.*, *Proc. of WASPAA,* pp.90-93, 2005.
[4] You, Sun, *Proc. of ICSP,* pp.1711-1714, 2002.
[5] Ono *et al.*, *Proc. of ISMIR,* pp.139-144, 2008
[6] Ono *et al.*, *Proc. of EUSIPCO,* 2008.
[7] 宮本他, 音講論 (春), pp.903-904, 2008.
[8] 宮本他, 音講論 (秋), pp.825-826, 2007.
[9] Goto *et al.*, *Proc. of ISMIR,* pp.287-288, 2002.
[10] 嵯峨山, 古井, 信学総大, Vol.5, 1235, 1978.