HPSS に基づく音響信号の多重解像度時間周波数解析*

☆橘秀幸, 小野順貴, 嵯峨山茂樹 (東大・情報理工)

1 はじめに

様々な時変な成分が混合する音響信号において，これらの各成分の変動の大きさには，しばしば重要な意味がある．このため，信号の各成分を変動の大きさごとに分離・分析・加工する技術が重要と考えられる．本稿では調波打楽器音分離 (Harmonic and Percussive Sound Separation, HPSS) [1] を多重的に用いることにより，音響信号を変動の大きさごとに分離する手法を提案する．これは，複数のフレーム長の短時間フーリエ変換 (STFT) 領域上での信号処理の枠組みとして我々が確立を目指している「多重解像度時間周波数解析」の一例と位置づけられる．

我々はこれまでにも，2 種類のフレーム長の STFT 上における HPSS によって，歌声やメロディにほぼ相当する成分を強調/抑圧する手法を提案している [2]. 本稿ではこれをさらに拡張し，より多くのフレーム長の STFT 上での HPSS によって，信号をより細かく分解する手法を提案する．また，これにより変動の大きさと概ね対応するパラメータを持った分離信号が得られること，およびそのパラメータを用いることにより時間周波数表現とは異なる新しい信号の 2 次元表現が可能となり，その領域上で従来の時間周波数領域での処理と同様の要領で信号の分析・加工ができることを示し，実際に分析・加工した例を示す.

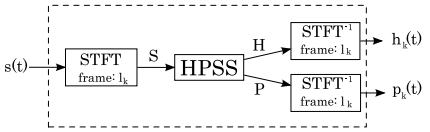

Fig. 1 Diagram of HPSS processing: HPSS separates a signal $s(t)$ into quasi-stationary/narrowband signal $h(t)$ and non-stationary/wideband signal $p(t)$.

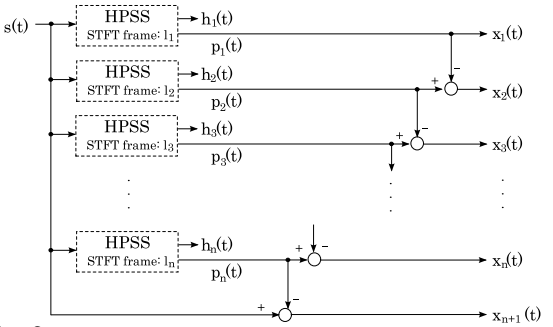

Fig. 2 Parallel n HPSSs separate a signal $s(t)$ into $2n$ components $\{h_k(t), p_k(t)\}_{1\le k\le n}$, $n-1$ of which are dependent on other $n+1$ components. A linear transformation of the $n+1$ signlas gives $n+1$ components $\{x_k(t)\}_{1\le k\le n+1}$, sum of which is equal to the original signal $s(t)$.

2 HPSS の並列的適用による信号の分離

2.1 調波打楽器音分離 (HPSS) の概要

HPSS では，信号 $s(t)$ を定常的・狭帯域（以下「H 的」）な成分 $h(t)$ と，非定常的・広帯域（以下「P 的」）な成分 $p(t)$ との和，すなわち $s(t)=h(t)+p(t)$ として表わすことを考える．ここで，各項の STFT スペクトログラム $\mathbf{S}$, $\mathbf{H}$, $\mathbf{P}$ には以下のような条件が課される：(1) $\mathbf{S}=\mathbf{H}+\mathbf{P}$, (2) $\mathbf{H}$ は時間軸方向に滑らか, (3) $\mathbf{P}$ は周波数軸方向に滑らか．HPSS は，これらの条件に基づいて，信号 $s(t)$ を Fig.1 のように $h(t)$ と $p(t)$ とに分離する．

2.2 複数の STFT 上での HPSS に基づく信号分離

HPSS の分離結果は STFT のフレーム長に依存する．例えば歌声は長時間観測した場合は非定常的であり P 的だが，短時間では定常的な正弦波の和と見なされうるため H 的である．すなわち，歌声信号は長いフレーム長の STFT 領域上では $p(t)$ へ，短いフレーム長の STFT 領域上では $h(t)$ へ，それぞれ分離されやすい [2].

一般に，$s(t)$ に対し，異なった n 通りのフレーム長 $l_1<\cdots<l_n$ のそれぞれの STFT 領域上で HPSS を適用することにより，n 通りの異なった分離信号 $\{h_k(t),p_k(t)\}_{1\le k\le n}$ が得られる．ところで HPSS では，$p_k(t)$ が決まれば同時に $h_k(t)$ も決まるから，$2n$ 次元の表現 $\{h_k(t),p_k(t)\}_{1\le k\le n}$ は冗長で，$n+1$ 次元の情報 $\{p_1(t),\cdots,p_n(t),s(t)\}$ で十分である．また，フレーム長 l_k の STFT 領域上で P 的である成分は，それよりも長いフレーム長 l_{k+1} の STFT 領域上でも P 的である可能性が高いと考えられることから，両者の P 的成分同士の差分 $\{p_{k+1}(t)-p_k(t)\}$ がより本質的な情報であると考えられる．そこで，Fig.2 のように $\{p_1(t),\cdots p_n(t),s(t)\}$ の k に関する差分情報 $\{x_k(t)\}_{1\le k\le n+1}$ を得て，以下ではこれを用いる．なお，$\{x_k(t)\}_{1\le k\le n+1}$ の各成分の和は原信号 $s(t)$ に一致し，$s(t)=\sum_{k=1}^{n+1}x_k(t)$ である．すなわち，$\{x_k(t)\}_{1\le k\le n+1}$ は $s(t)$ を分解した表現になっている.

2.3 信号の分析・加工

$x_k(t)$ は概ね $h_1(t),\cdots,h_k(t),p_{k+1}(t),\cdots p_n(t)$ に含まれる成分と考えられることから，$x_k(t)$ 成分はフレーム長 l_k 程度以下では H 的，l_k 程度以上では P 的な成分である可能性が高い．このため，l_k 程度のスケールで変動している成分であると考えられる．このことから，$\{x_k(t)\}_{1\le k\le n+1}$ のうちどの成分が強く現れているかを調べることにより，信号の変動に関する特徴を調べることができると考えられる.

また，$\{x_k(t)\}_{1\le k\le n+1}$ の添え字 k は，時間（フレームのインデックス）τ，周波数 ω などとは独立したパラメータである．このため，信号の STFT を時間周波数 (τ-ω) 領域にプロットすることで信号を分析することと同じ要領で，信号を τ-k 領域や k-ω 領域にプロットした表現を考えることができて，τ-k 領域上では $x_k(t)$ 成分がどの時刻 τ に強く現れているか，k-ω 領域上では $x_k(t)$ 成分がどの周波数 ω に強く現れているかを分析することができる．またこれらの領域上では，τ-ω 領域上での信号処理手法と同様の，例えばバンドパスフィルタやマスキングと類似した方法により，信号を加工することが可能である．以上のように，τ-k 領域，k-ω 領域は，信号分析・加工するための新しい領域として用いることができると考えられる.

3 信号の分析例

3.1 実験条件

本稿の手法によって，音響信号を概ね変動の大きさごとに分離し，さらに τ-k 領域表現，k-ω 領域表現上で信号が分析・加工できることを確認するため，様々

* Hideyuki Tachibana, Nobutaka Ono, Shigeki Sagayama, Graduate School of Information Science and Technology, The University of Tokyo, “HPSS-based Multi Resolution Time-frequency Analysis of Acoustic Signals.”

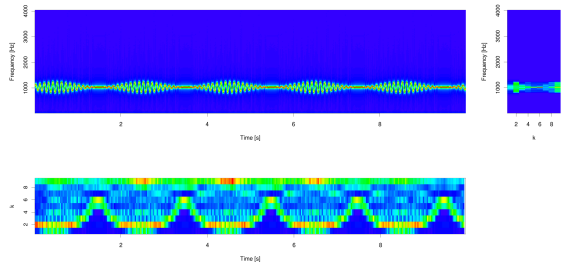

Fig. 3　The upper left plot shows the spectrogram (conventional τ-ω domain representation) of a FM wave, upper right plot shows its k-ω domain representation, and the lower shows its τ-k domain representation.

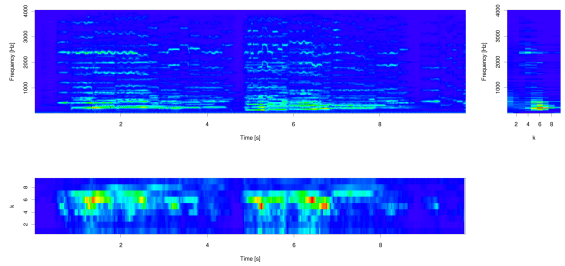

Fig. 4　τ-ω, τ-k, and k-ω representations of a sound of string quartet, similar to Fig. 3.

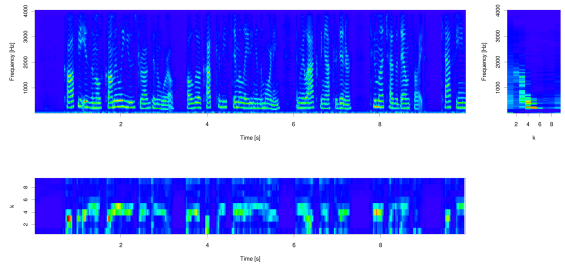

Fig. 5　τ-ω, τ-k, and k-ω representations of a female speech, similar to Fig. 3.

な音響信号に関して $\{x_k(t)\}_{1\leq k\leq n+1}$ を求め，それぞれの信号の τ-k 領域と k-ω 領域での表現を得る実験を行った．

実験に使用した信号は，著者らが計算機で生成した FM 波，および建築と環境のサウンドライブラリ [3] より抜粋した 2 データで，いずれも簡単のためサンプリング周波数 8kHz のモノラル信号とした．STFT のフレーム長は $l_k = 64 \times 2^k$ [samples] $(1 \leq k \leq 8)$, すなわち $128, \cdots, 16384$ [samples] $(16, \cdots, 2048$ [ms]) とした．これらの条件で各信号に関して $\{x_k(t)\}_{1\leq k\leq 9}$ を求め，τ-k, k-ω 領域上にそれぞれプロットし，分析を行った．また，τ-k 領域上でのマスキングにより信号を加工する実験を行った．

3.2　実験結果

実験に用いた FM 波, 弦楽四重奏 (s12202.wav)，英語女性朗読 (s13102.wav) の STFT，および各々の信号の k-ω 表現と τ-k 表現を Fig. 3, 4, 5 に示す．また，英語女性朗読の τ-k 領域でのマスキング処理を Fig.6 に示す．

FM 波（Fig. 3）の τ-k 領域表現では，変調が深い 0.5 [s] 付近などの時刻に低次の $k = 1$ で，変調が浅い 1.5 [s] 付近などの時刻は高次の $k = 6$ で，$x_k(t)$ に強い成分が現れやすく，間の時刻では $k = 2, \ldots, 5$ の間を段階的に推移することが観察できる．また k-ω 領域では，$k = 2, \ldots, 6$ のとき，概ね k が大きいほど $x_k(t)$ は狭帯域であることが観察できる．

弦楽四重奏（Fig. 4）の τ-k 領域表現では，概ね $k = 5, 6, 7$ など高次の k が強い．また，5-7 [s] 付近のように高域にゆらぎがあるときは，より低次の $k = 3, 4$ にも強い成分が現れていることが観察できる．k-ω 領域では，比較的定常的な低周波成分は高次の $k = 5, 6$ に，ゆらぎを含むような高周波成分は比較的低次の $k = 4$ に強く現れていることが観察できる．

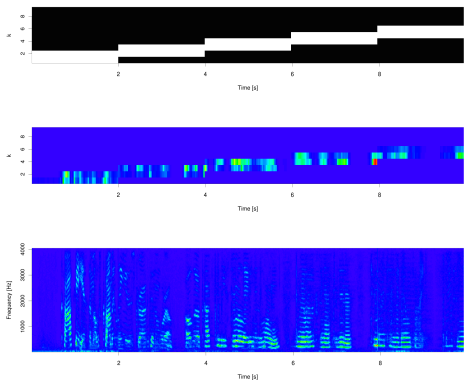

Fig. 6　The top plot shows a τ-k mask, and the middle plot shows the masked τ-k representation of the female speech: the product of the mask and the τ-k representation Fig. 5 lower. The bottom plot shows the spectrogram of the reconstructed signal from the masked τ-k representation.

英語女性音声（Fig. 5）の τ-k 領域表現では，音声は概ね $k = 3, 4, 5$ に強く現れ，特に高域に非定常的な成分がある時刻では低次の k に強い成分が現れやすいことが観察できる．k-ω 領域では，$k = 1, 2, 3$ では広帯域な成分が，$k = 4, 5$ では低域の成分が現れていることが観察できる．また，τ-k 領域において Fig.6 上段のようなマスクを乗算して新しい τ-k 表現（Fig. 6 中段）を得て，それに基づき信号を再合成した．その結果，初めは低次の k の $x_k(t)$ の性質を反映して無声音的で，次第に高次の k の $x_k(t)$ の性質を反映して有声音的かつ平坦となっていくように，音声を加工することができた (Fig. 6 下段)．

4　まとめ・今後の展望

本稿では，フレーム長の異なる複数の STFT 上で，定常的・狭帯域（H 的）成分と非定常的・広帯域（P 的）成分とを分離する手法 HPSS を並列的に適用することにより，信号を概ね変動の大きさに対応して分離する，新しい信号の分離手法を提案した．また，ここで得られる分解信号を用いることで信号の新しい 2 次元表現が可能であること，およびその 2 次元領域上において信号が分析・加工できることを示す実験例を示した．

本稿では τ-k, k-ω 領域上の 2 次元表現を得たが，τ-ω-k の 3 次元表現に基づいた分析・加工も可能であると考えられる．これは今後の検討課題となる．また，雑音環境下での本手法の頑健性も検討課題であり，本手法によって得られた $\{x_k(t)\}_{1\leq k\leq n+1}$ を多チャンネル信号とみなし，多チャンネル信号処理の手法を適用することを検討している．また，$\{x_k(t)\}_{1\leq k\leq n+1}$ の各成分のゼロ交差率や MFCC などの特徴量を用いて特徴量ベクトルを構成し，音声認識などの認識問題へ応用することも，今後の研究課題である．

謝辞　本研究の一部は日本学術振興会科研費特別研究員奨励費 (22-6961) の助成を受けて行われた．

参考文献

[1] Ono *et al.*, "Harmonic and Percussive Sound Separation and its Application to MIR-related Tasks," Springer 274, pp.213-236, 2010.
[2] Tachibana *et al., Proc. ICASSP,* pp. 425-428, 2010.
[3] 日本建築学会編 建築と環境のサウンドライブラリ SMILE2004