複素 NMFD による音声抽出マスクの設計と背景音楽抑圧への応用*

◎北野佑 (東大院・情報理工), 亀岡弘和, 柏野邦夫 (NTT CS 研),
小野順貴, 嵯峨山茂樹 (東大院・情報理工)

1 はじめに

地上波やインターネットを介して入手可能なマルチメディアコンテンツに含まれるオーディオデータには，音声や音楽が混在していることが多い．このようなオーディオデータをめぐり，音楽コンテンツなどの著作物の使用状況をモニタリングする著作権管理技術，音声・動画コンテンツをテキストクエリにより検索可能にするメディア検索技術が求められている．前者の場面では，音声ナレーションが「雑音」になりうる一方で，後者の場面では，背景音楽が「雑音」になりうるため，音声と音楽のうちどちらが目的情報でありどちらが「雑音」であるかは場面に応じて異なる．ここで言う「雑音」は，定常性や Gauss 性など，従来の多くの雑音抑圧手法において仮定される雑音の性質を必ずしも満たさないため，音声と音楽が混在する音響信号から音声，音楽を強調する用途にこうした従来法を適用しても有効に動作しない可能性が高い．

Fig. 1 のように，音声と音楽のスペクトログラムを見比べてみると，各々を構成しているパターンには大きな差異が見られる．例えば音楽においては，同じ音高が一定時間保たれたり，複数の音が同時刻に一斉に立ち上がる傾向があるため，時間軸に平行に連なるスペクトル成分のパターンや，周波数軸に平行に連なるスペクトル成分のパターンが多く見られる点が特徴的である．一方音声の場合は，基本周波数やフォルマントが連続的に時間変化することによって描かれる特有なパターンをもつ点が特徴的である．

音声と音楽におけるスペクトログラムの構成パターンの違いに着目し，本研究では音声と音楽の混合信号を分離する方法を検討する．これを実現するため，(i) 音声および音楽の学習データからスペクトログラムの構成パターンを学習するステージ，(ii) 学習したスペクトログラムパターンを用いて観測信号に含まれる音声と音楽のパワースペクトルを推定するステージ，(iii) 推定した両者のパワースペクトルから Wiener フィルタを構成して音声を抽出するステージ，からなる処理系を構成する．この処理系の (i) と (ii) を行うための枠組を以下で提案する．

2 複素 NMFD

2.1 観測モデル

音響信号のスペクトログラムを構成する特徴的な基底スペクトルパターンを抽出したい．これは，観測スペクトログラムを K 個の基底パターンのみで構成されるモデルで最もよく表現できるような基底パターンを決定することにより，実現できると考えられる．

そのような手法の中で代表的なものとして NMF(Nonnegative Matrix Factorization)[1] がある．これは任意の音響信号の時間周波数成分を $F_{x,t}$ とし

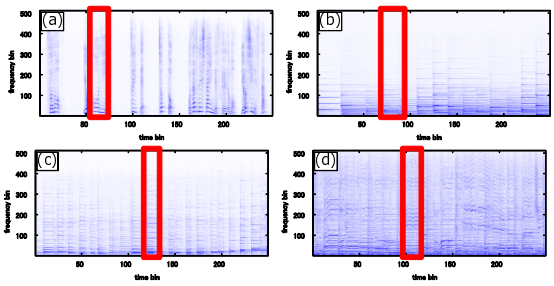

Fig. 1 (a) 音声と音楽 ((b) ピアノ，(c) ギター，(d) ポップス音楽) のスペクトログラム．各々を構成するスペクトログラム素片 (赤) のパターンの違いは明らか．

たとき，振幅スペクトルの加法性を仮定し，

$$|F_{x,t}| = \sum_{k=1}^{K} H_{k,x} U_{k,t} \tag{1}$$

というモデルをたてた時，k 番目の振幅スペクトルパターン $H_{k,x}$ とその係数 $U_{k,t}$ が非負値であるという制約のみで特徴的な振幅スペクトルパターンを抽出することができる手法である．ただし，x は周波数に，t は時刻に対応するインデックスである．NMF では一次元ベクトルの振幅スペクトルを基底としているが，これを時間軸方向に次元を増やし，二次元のスペクトログラム素片に拡張したものを NMFD(Nonnegative Matrix Factor Deconvolution)[2] という．つまり $F_{x,t}$ がスペクトログラム素片 $H_{k,x,t}$ によって構成されているとすると，

$$|F_{x,t}| = \sum_{k=1}^{K} \sum_{\tau=0}^{M-1} H_{k,x,\tau} U_{k,t-\tau} \tag{2}$$

と表すことができ，同様に非負値制約のみで各パラメータを求めることにより，信号分解を行うというものである．ここで M はスペクトログラム素片の長さである．

しかし実際には振幅スペクトルの加法性は成立しない．そこでこの問題を加法性の成り立つ複素スペクトル領域へ拡張し，

$$F_{x,t} = \sum_{k=1}^{K} \sum_{\tau=0}^{M-1} H_{k,x,\tau} U_{k,t-\tau} e^{j\phi_{k,x,t,\tau}} \tag{3}$$

というモデルを考える．ここで $\phi_{k,x,t,\tau}$ は位相スペクトルである．このモデルを用いた信号分解を本稿では複素 NMFD と呼ぶことにする．NMF を複素スペクトル領域へ拡張したものとして複素 NMF[3] があるが，この複素 NMFD は，NMFD を複素スペクトル領域へ拡張したものである．今回はモデルの自由度を下げる目的で，$\phi_{k,x,t,\tau} = \phi_{k,x,t}$ とした．このとき $H \equiv (H_{k,x,t})_{K\times X\times M}$，$U \equiv (U_{k,t})_{K\times T}$，$\phi \equiv (\phi_{k,x,t})_{K\times X\times T}$ を求める際に，できるだけ $U_{k,t}$ をスパースに，かつ，観測信号とのモデル化誤差を小さく

*Wiener Filtering Steered by Complex NMFD with Application to Background Music Suppression. by KITANO Yu(The University of Tokyo), KAMEOKA Hirokazu, KASHINO Kunio(NTT), ONO Nobutaka, SAGAYAMA Shigeki(The University of Tokyo)

するようにパラメータ推定を行いたい．これはすなわち，適当な位相が付与された少ないスペクトログラム素片だけで観測時間周波数成分を良く表そうとすることに相当し，これにより，観測時間周波数成分の中に繰り返し生起する特徴的なスペクトログラムのパターンほど $H_{k,x,t}$ の解として選ばれやすくなると考えられる．次節にてパラメータ推定法を述べる．

2.2 目的関数とパラメータ推定法

得られた観測音響信号の時間周波数成分を $Y_{x,t}$ とすると，観測とモデルの間に，

$$Y_{x,t} = F_{x,t} + \epsilon_{x,t} \tag{4}$$

なる関係があると仮定し，モデル化誤差 $\epsilon_{x,t}$ は複素 Gauss 分布に従う白色雑音とする．またスパース性を表す U の事前分布を一般正規分布と仮定すると，パラメータの最大事後確率推定は，

$$\Phi(H,U,\phi) \equiv \sum_{x,t} |Y_{x,t} - F_{x,t}|^2 + 2\lambda \sum_{k,t} |U_{k,t}|^p \tag{5}$$

の最小化と等価となる．ここで λ, p は定数で，$0 < p < 2, \lambda > 0$ を満たす．なおここで分解のスケールの任意性を防ぐために，

$$\sum_{x,\tau} H_{k,x,\tau} = 1 \tag{6}$$

とする．従って，(6) の条件下で (5) を最小化することが目的となる．

この目的関数は補助関数法 [3] により局所最小化することができる．紙面の都合上，補助関数法の原理や補助関数の具体形は省略せざるを得なかったため，以下に各パラメータの更新式の導出結果のみ示す．

$$H_{k,x,\tau} \leftarrow H_{k,x,\tau} \left(1 + \frac{\sum_t U_{k,t-\tau} \mathrm{Re}\left[\zeta_{k,x,t}\right]}{\sum_t U_{k,t-\tau} \eta_{x,t}} \right) \tag{7}$$

$$U_{k,t} \leftarrow U_{k,t} \frac{\sum_{x,\tau} H_{k,x,t+\tau} (\zeta_{k,x,t+\tau} + \eta_{x,t+\tau})}{\lambda p U_{k,t}^{p-2} + \sum_{x,\tau} H_{k,x,t+\tau} \eta_{x,t+\tau}} \tag{8}$$

$$\phi_{k,x,t} \leftarrow \arg\left((\eta_{x,t} e^{j\phi_{k,x,t}} + Y_{x,t} - F_{x,t}) \gamma_{k,x,t} \right) \tag{9}$$

ただし，$\gamma_{k,x,t} = \sum_\tau H_{k,x,\tau} U_{k,t-\tau}$，$\zeta_{k,x,t} = (Y_{x,t} - F_{x,t}) e^{-j\phi_{k,x,t}}$，$\eta_{x,t} = \sum_k \gamma_{k,x,t}$ とした．ただし，(7) の更新後に (6) の条件を満たすように規格化を行うことにする．

3 音声強調のための処理系

前章のパラメータ推定法により，音声・音楽の学習データからそれぞれスペクトログラムパターンを学習することができる．そして観測信号に対して，(3) の H に学習されたそれぞれのスペクトログラムパターンを代入し，H 固定で (8)(9) を繰り返し，U, ϕ を推定後，それらを再合成することにより，音声と音楽の時間周波数成分を推定することができる．

しかし学習用データを構成するスペクトログラムパターンと観測信号を構成するスペクトログラムパターンは必ずしも一致しない．音声の例だと様々な音声のスペクトログラムパターンの重ね合わせで表現しようとしているが，必ずしもそれぞれの位相スペクトルが正しく推定できているわけではない．しかし，音声・音楽のパワースペクトルをおおよそ表現することは可能である．そのため，この推定されたスペクトルを用いて時間周波数マスキングを行う．

そこで推定された音声，音楽の振幅スペクトル $S_{x,t}, N_{x,t}$ を用いて，Wiener フィルタ

Table 1 分離信号の SN 比 ([dB])

音楽信号の種類	音声	音楽
Piano	2.64	2.65
Guitar	2.21	2.24
Pops	4.43	4.01

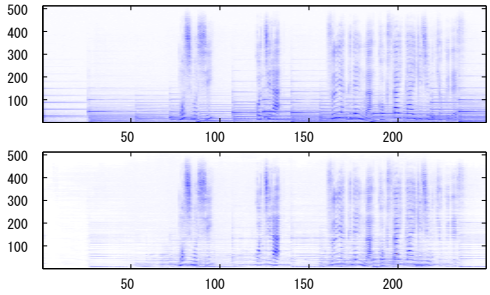

Fig. 2 観測信号と抽出された音声のスペクトログラムの例．音楽信号はピアノ演奏である．なお縦軸，横軸はそれぞれ周波数，時刻にあたるインデックスである．

$$\mathrm{mask}(x,t) = \frac{|S_{x,t}|^2}{|S_{x,t}|^2 + |N_{x,t}|^2} \tag{10}$$

に基づいて音声抽出マスクを設計する．これは目的信号と妨害信号が無相関であるとき，理想的な時間周波数マスクである．

4 シミュレーション実験

今回の提案法の性能を検証するために，実験を行った．音声に関しては ATR 音声データベースの音声データ (女性話者)，音楽に関しては RWC 音楽データベースからピアノ演奏，ギター演奏，ポップス音楽の三種類を用いた．用いた信号は 16 kHz サンプリングで，フレーム長 64 ms，フレーム周期 32 ms の条件で STFT することにより，時間周波数領域に変換した．音声，音楽の学習用データは同じ話者，同じ楽器の演奏のものとした．また，前章に述べた時間周波数マスクによる音声抽出だけでなく，(10) の音声と音楽のスペクトルを逆にすることにより，音楽抽出も行った．実験結果を Table 1 に示す．なお混合信号の時点の各 SN 比は 0.0 dB とした．また観測信号と抽出された音声のスペクトログラムの例を Fig. 2 に示す．これらの結果をみると，音声だけでなく音楽に関しても SN 比が改善されていることがわかる．

5 おわりに

本稿では，音声と音楽を構成するスペクトログラムパターンの違いに着目し，モノラル信号を対象とした混合信号中の音声と音楽の分離手法を，複素 NMFD の原理に基づいて，提案した．そして推定された音声・音楽のスペクトルから音声抽出マスクを設計し，シミュレーション実験により音声抽出を行った．その結果，本手法が有効であることが示された．また同様に音楽抽出も行い，その有効性を確認した．今後はパラメータチューニング，学習用データの拡大による精度向上やケプストラム歪みによる性能評価をしていく予定である．

謝辞 本研究は東大と NTT との共同研究の一環として，NTT での夏期実習を通して行われた．

参考文献

[1] Lee & Seung, Nature, 384, pp.607–609, 1996.
[2] Smaragdis, Proc. ICA2004, pp.494–499, 2004.
[3] 亀岡他, 音講論 (秋), pp.657–660, 2008.