小特集―音環境理解の近年の動向：音響イベント分析及び音響シーン分析―

深層学習に基づく音源分離*

亀岡弘和（日本電信電話株式会社）**

43.60.Fq, Np

1. はじめに

我々人間は日常的に様々な音源の音に接し，それらが混在する音からどのような音が鳴っているかを聞き分け，周囲で何が起こっているかを把握するための手がかりとしている。このような機能を計算機に備えさせる音環境理解の問題は，混合音を各音源に分解する音源分離，音源の位置を推定する音源定位，直接音と残響音を分離する残響除去，音が鳴っている時刻を推定する音声区間推定などの部分問題からなる。本稿では主に音源分離の問題に焦点を当て，モノラル信号及びマイクロホンアレーを用いて観測される多チャンネル信号を対象とした音源分離の問題に対するアプローチ，特に深層学習を用いた音源分離法についての最近の動向について解説する。

2. 音源分離問題へのアプローチ

音環境理解を目的とした音源分離の研究は古くから取り組まれており，これまで多くの音源分離手法が提案されているが，近年，深層学習に基づく手法の有効性が示され始めている。音源分離の問題は観測信号を分解する問題であるが，一つの観測信号に対し分解の仕方は無数に存在する。この点で単純な回帰分析やクラス識別などのように一つの観測データに対して唯一の解が対応することを前提としたものとは異なる。そのため深層学習に基づく音源分離法は，深層学習の特長を生かせるような形に音源分離の問題を定式化する工夫がなされている。

音源分離問題のアプローチは旧来のものも含め（例外はあるものの）音源推定アプローチと時間周波数マスク推定アプローチに大別される。前者は，各音源信号が混合されることにより観測信号が生成される過程をモデル化し，各音源信号に関するモデルや先験的知識に基づき観測信号から各音源信号を予測するアプローチである。一方で後者は，観測信号の時間周波数点でどの音源が支配的であるかを識別することで，同一音源が支配的な時間周波数点の成分のみを通過させる時間周波数マスクを得ることを目的とするアプローチである。以下，それぞれのアプローチの具体的な問題設定の例を示す。

2.1 音源推定アプローチ

まず，観測信号がモノラルの場合を考える。音源 j の信号の短時間フレーム n における短時間フーリエ変換（Short-Time Fourier Transform; STFT）を $s_j(f,n) \in \mathbb{C}$ とすると，J 個の音源の混合信号の STFT $x(f,n)$ は

$$x(f,n) = \sum_j s_j(f,n) \tag{1}$$

と表される。ただし，f は周波数のインデックスである。ここで，例えば，$s_j(f,n)$ が j, f, n に関し独立に

$$s_j(f,n) \sim \mathcal{N}_{\mathbb{C}}(s_j(f,n) \mid 0, v_j(f,n)) \tag{2}$$

のように平均が 0，分散が $v_j(f,n)$ の複素ガウス分布に従うと仮定すると，$x(f,n)$ は平均が 0，分散が $\sum_j v_j(f,n)$ の複素ガウス分布

$$x(f,n) \sim \mathcal{N}_{\mathbb{C}}\Big(x(f,n) \;\Big|\; 0, \sum_j v_j(f,n)\Big) \tag{3}$$

に従うことが示される。ただし，分散 $v_j(f,n)$ は，$\mathbb{E}[|s_j(f,n)|^2]$ を表すため音源 j のパワースペクトログラムに対応する。$\mathcal{X} = [x(f,n)]_{f,n}$, $\mathcal{V} = [v_j(f,n)]_{j,f,n}$ と表すと，観測信号 $\mathcal{X}$ が与え

* Deep learning approach to audio source separation.

** Hirokazu Kameoka (NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Atsugi, 243–0198) e-mail: kameoka.hirokazu@lab.ntt.co.jp

られた下での $\mathcal{V}$ の尤度関数 $p(\mathcal{X}|\mathcal{V})$ は

$$p(\mathcal{X}|\mathcal{V}) = \prod_{f,n} \mathcal{N}_{\mathbb{C}}\Big(x(f,n) \;\Big|\; 0, \sum_j v_j(f,n)\Big) \tag{4}$$

となる。また，$x(f,n)$ と $s_1(f,n),\ldots,s_J(f,n)$ の結合ガウス性より，$x(f,n)$ と $v_1(f,n),\ldots,v_J(f,n)$ が与えられた下での $s_j(f,n)$ の最小平均二乗誤差推定量 $\mathbb{E}[s_j(f,n)|x(f,n)]$ は

$$\mathbb{E}[s_j(f,n)|x(f,n)] = w_j(f,n)x(f,n) \tag{5}$$

$$w_j(f,n) = \frac{v_j(f,n)}{\sum_{j'} v_{j'}(f,n)} \tag{6}$$

となる（$w_j(f,n)$ をウィーナゲインと呼ぶ）。よって，$\mathcal{V}$ を推定することができれば，式 (5) より各音源信号を得ることができる。しかし，尤度関数 $p(\mathcal{X}|\mathcal{V})$ は

$$|x(f,n)|^2 = \sum_j v_j(f,n) \tag{7}$$

のときに最大になるという事実しか音源に関する手がかりを与えてくれず，当然，式 (7) を満たす $\mathcal{V}$ は無数に存在するため，尤度関数 $p(\mathcal{X}|\mathcal{V})$ を規準にするだけでは $\mathcal{V}$ を一意に決めることができない。従って，通常，各音源のパワースペクトログラムに関して何等かの先験的知識や仮定が必要となる。従来は，音源の特徴を反映したパラメトリックなモデル（非負値行列分解モデルなど）で $\mathcal{V}$ を表現した上でそのパラメータの最尤推定又は最大事後確率推定問題に帰着させる手法や，$\mathcal{V}$ に関する事前分布 $p(\mathcal{V})$ を仮定又は事前学習した上で $\mathcal{V}$ の最大事後確率推定値を探索する手法などが提案されている。これらの中に特定条件下において有効なものはあるものの，実世界の音源のパワースペクトログラムは極めて多様で，$\mathcal{V}$ をモデル化することこそが音源分離問題の難しさそのものであると言える。

次に，観測信号がマイクロホンアレーで収音された多チャンネル信号の場合を考える。J 個の音源信号の混合信号を I 個のマイクロホンで観測するとき，時不変な混合過程を仮定できる場合，各マイクロホンで観測される信号は時間周波数領域で

$$x_i(f,n) = \sum_j a_{i,j}(f)s_j(f,n) \tag{8}$$

と表される。ただし，$a_{i,j}(f)\in\mathbb{C}$ は音源 j からマイクロホン i までの音響経路の周波数応答，$x_i(f,n)$，$s_j(f,n)$ はそれぞれマイクロホン i の観測信号，音源 j の音源信号の STFT を表す。$\mathbf{a}_j(f)=[a_{1,j}(f),\ldots,a_{I,j}(f)]^{\mathsf{T}}$, $\mathbf{A}(f)=[\mathbf{a}_1(f),\ldots,\mathbf{a}_J(f)]$, $\mathbf{s}(f,n)=[s_1(f,n),\ldots,s_J(f,n)]^{\mathsf{T}}$ とすると式 (8) は

$$\mathbf{x}(f,n) = \sum_j \mathbf{a}_j(f)s_j(f,n) \tag{9}$$

$$= \mathbf{A}(f)\mathbf{s}(f,n) \tag{10}$$

と書ける。また，音源数 J とマイク数 I が等しく $\mathbf{A}(f)$ が逆行列 $\mathbf{A}(f)^{-1} = \mathbf{W}^{\mathsf{H}}(f)$ を持つ場合，式 (10) は以下のように書ける。

$$\mathbf{W}^{\mathsf{H}}(f)\mathbf{x}(f,n) = \mathbf{s}(f,n) \tag{11}$$

ここで，式 (2) と同様，$s_j(f,n)$ が独立に平均が 0 の複素ガウス分布に従う場合を考える。このとき，式 (9) と式 (11) のいずれの混合過程を仮定した場合も観測信号 $\mathbf{x}(f,n)$ は下記に従う。

$$\mathbf{x}(f,n) \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}(f,n)\,|\mathbf{0}, \mathbf{\Sigma}_{\mathbf{x}}(f,n)\right) \tag{12}$$

ただし，$\mathbf{\Sigma}_{\mathbf{x}}(f,n)$ は，式 (9), (11) の場合，

$$\mathbf{\Sigma}_{\mathbf{x}}(f,n) = \begin{cases} \sum_j v_j(f,n)\mathbf{R}_j(f) \\ \mathbf{W}^{\mathsf{H}}(f)^{-1}\mathbf{\Sigma}_{\mathbf{s}}(f,n)\mathbf{W}(f)^{-1} \end{cases} \tag{13}$$

で与えられ，$\mathbf{R}_j(f) = \mathbf{a}_j(f)\mathbf{a}_j^{\mathsf{H}}(f)$, $\mathbf{\Sigma}_{\mathbf{s}}(f,n) = \mathrm{diag}([v_1(f,n),\ldots,v_J(f,n)]^{\mathsf{T}})$ である。式 (12) より観測信号 $\mathcal{X} = [x_i(f,n)]_{i,f,n}$ の確率分布は

$$p(\mathcal{X}|\mathcal{V},\mathcal{A}) = \prod_{f,n} \mathcal{N}_{\mathbb{C}}\left(\mathbf{x}(f,n)\,|\mathbf{0}, \mathbf{\Sigma}_{\mathbf{x}}(f,n)\right) \tag{14}$$

となり，これを未知パラメータ $\mathcal{V} = [v_j(f,n)]_{j,f,n}$, $\mathcal{A} = [\mathbf{A}(f)]_f$ の尤度関数とした最尤推定問題として多チャンネル音源分離を定式化することができる。ただし，式 (14) の対数は周波数 f ごとの項に分解されるため，以上の定式化では，各 f において分離信号のインデックス j にはパーミュテーションの任意性が残る。また，音源数がマイクロホン数より多い劣決定条件においては $\mathbf{A}(f)$ が決まったとしても $\mathbf{s}(f,n)$ を一意に決めることができない。このため，解を絞り込むためには更なる手がかりが必要である。モノラルの場合と同様，従

来音源の特徴を反映したパラメトリックなモデルで $\mathcal{V}$ を表現した手法や $\mathcal{V}$ に関する事前分布 $p(\mathcal{V})$ を仮定した手法が提案されているが，これらの手法で置かれる仮定が成立するのは特定条件のみで，多様な音源のスペクトログラムを包含するような汎用的なパラメトリックモデルや事前分布を手動で設計することの難しさは同様である。

2.2 時間周波数マスク推定アプローチ

実世界の音源は，時間周波数領域においてエネルギーがスパースに分布するものが多い。例えば音声や楽音のように局所的に周期性を有する信号は周波数領域において，基本周波数とその整数倍の周波数（高調波周波数）にエネルギーが集中する調波構造と呼ぶ特徴的な構造を持つ。このため，たとえ複数の音が同時に鳴っていたとしても各時間周波数点ではいずれか一つの音源のみのエネルギーが支配的になると仮定することができる。従って，各時間周波数点でどの音源が支配的かが分かれば，各音源の成分のみを通過させる時間周波数マスクを構成することにより信号を分離することができる。例えば観測信号が多チャンネル信号の場合，各時間周波数点の信号成分のチャネル間位相差などを手がかりに支配的な音源を推定することができる [1]。一方でモノラルの場合は音源の空間情報を手がかりにすることはできないが，画像中の各ピクセルを物体や人物ごとに対応する領域に分割する画像セグメンテーションの問題と類似し，各時間周波数点周辺のスペクトログラムから得られる情報を手がかりにすることができる。例えば，時間周波数点ごとに算出される特徴量をもとに各時間周波数点で音声と雑音のどちらが優勢かをサポートベクトルマシン（Support Vector Machine; SVM）により識別する手法が提案されている [2]。この手法では調波構造をなしているかどうかや各周波数成分の時間変化が同期しているかどうかといった音声のスペクトル構造に関する大まかな先験的知識を反映した特徴量を用いることで音声と雑音の識別を可能にしている。しかしこれは音声と雑音のようにスペクトル構造が大きく異なると仮定できるからこそ適用可能な方法であり，例えば音源がいずれも音声の場合，それぞれの話者のスペクトル構造の違いを反映したような特徴量を手動で設計する必要があり，容易には行かない。その他の例としてはスペクトルクラスタリングに基づく手法が提案されている [3]。これは，時間周波数点の各ペアにおいて同一音源が支配的かどうかを表した巨大な類似度行列を考え，これを用いて時間周波数点をクラスタリングする，という方法である。この手法も，各音源が満たすべき要件を反映した特徴量を手動で設計する必要がある点で前述の手法と共通している。

3. 深層学習に基づく手法

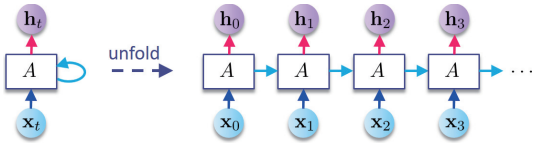
以上のようにいずれのアプローチにおいても個々のアルゴリズムの成否は音源のモデルや特徴量の選択に依るところが大きかったが，深層学習に基づく手法は音源の生成モデルや特徴量抽出器をニューラルネットワーク（Neural Network; NN）に担わせた上で NN のパラメータを学習により決定することで，NN の関数としての高い表現能力をいかしつつ従来の音源モデルや特徴量の手動設計におけるヒューリスティクスをできるだけ排除することを動機としている。音声など実世界の音響信号の特徴は特に時間方向及び周波数方向の依存関係に強く現れるため，このような依存関係を捉えられるような構造を持つ NN を用いることが重要である。以下ではまず音声や音響信号のモデルとしての有用性が既に知られている幾つかの NN モデルの例を挙げた上で，これらを用いた音源分離手法の例を紹介する。

3.1 ニューラルネットワークモデル

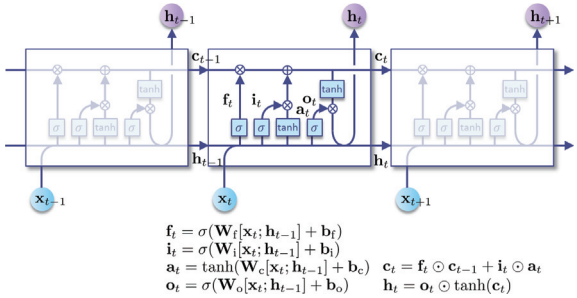
音は時系列信号で表され，ほぼ例外なく時間的な相関を持つ。特に音声は発声器官の物理制約（ダイナミクス）によってもたらされる局所的な相関構造から発話全体にわたる大域的な相関構造まで，様々な時間スケールの相関構造を有している。このため例えば，入出力系列間の時間依存関係を学習可能な再帰型 NN（Recurrent Neural Network; RNN）は，音源の生成モデルや音の特徴量抽出器として用いることに適した NN の一つである。RNN は内部状態が時間発展する NN であり，隠れ層の出力を次の時間ステップにおける隠れ層の入力とする形をとることで，入力値及びその履歴をもとに出力値を逐次予測する能力を持つ NN にすることができる。図–1 のように隠れ層を時間ステップごとに展開すると，例えば x_0 と h_3 の間に四つの隠れ層があるのと同等であることが分かることから RNN は深い層の NN と見なすことができる（ただ

し，同じ色の矢印はパラメータが共有される点が異なる）。従ってこの展開された NN の層の深さは，出力値の予測において考慮する入力値の履歴の長さに相当する。深い層の NN では誤差逆伝搬（Backpropagation）法による学習プロセスにおいてしばしば勾配消失（Vanishing Gradient）が問題となるが，通常の RNN も例外ではなく勾配消失が問題となる場合が多い。これは各パラメータの勾配が連鎖律により出力層までの層数分の活性化関数の微分の積が乗じられた形になるためであり，微分が 0 の領域又は 0 に近い領域のある活性化関数を用いた場合，層数が増えるほど勾配の絶対値が小さくなり易くなる（勾配法において勾配の絶対値が小さいパラメータはほとんど更新されない）。この問題を克服するため考案されたのが長短期記憶ネットワーク（Long Short-Term Memory; LSTM）である [4]。図–2 に示すように，LSTM はメモリセル，忘却ゲート，入力ゲート，出力ゲートの 4 層からなる再帰モジュールを採用している。LSTM では，メモリセルを介して情報 c_t（メモリと呼ぶ）を継承していく仕組みがとられており，メモリの継承においては活性化関数を通過させないようにすることで勾配消失を起こしにくくしている。

畳み込み NN（Convolutional Neural Network; CNN）もまた音源の生成モデルや音の特徴量抽出器として用いることに適した NN である。CNN は特に画像やコンピュータビジョンの分野で早くからその高い効果や能力が知られていたが，近年は時系列データにおける時間軸を画像における座標軸と同等に見なした利用方法も広まってきており，特に言語モデル [5] や音声波形の生成モデル [6] としての高い能力が示されている。CNN では層を深くするほど受容野を広くすることができるが，CNN も深い層のものになると勾配消失が生じ易くなる。しかし CNN においても RNN における LSTM のように勾配消失問題を回避する工夫がなされている。その一つがゲート付き CNN（Gated CNN; GCNN）である [5, 6]。通常の CNN における畳み込み層は $\mathsf{h}_{l+1} = \phi(\mathsf{W} * \mathsf{h}_l + \mathsf{b})$ と表されるが，GCNN の畳み込み層は $\mathsf{h}_{l+1} = (\mathsf{W} * \mathsf{h}_l + \mathsf{b}) \odot \mathrm{sigmoid}(\mathsf{V} * \mathsf{h}_l + \mathsf{c})$ と表される [5]。ただし，h_l, h_{l+1}, W, V, b, c は畳み込み層 l の入出力，カーネル，バイアスを表し，ϕ, sigmoid は任意の非線形活性化関数，シグモイド関数，$\odot$ は要素ごとの積を表す。LSTM においてメモリの継承が線形的に行われていくのと同様，GCNN では線形層の出力 $(\mathsf{W} * \mathsf{h}_l + \mathsf{b})$ が乗法的に深い層へ継承されていくスタイルをとることで，勾配消失を起こしにくくしている。



図–1 再帰型ニューラルネットワーク（RNN）



図–2 長短期記憶（LSTM）

3.2 音源推定アプローチ

音源照合スコア最大化法

モノラル信号を対象とした音源推定アプローチとしては，音源照合器（音声の場合は話者照合器）を用いた手法が提案されている [7]。以下，簡単のため 2 音源の場合を考える。もし混合信号に含まれる各音源の種類が既知ならば，分離信号の最良解は当該音源の特徴に最も適合しているときのはずである。また，同時に，分離信号の和は混合信号と一致していなければならない。そこで，分離信号が音源 1 又は音源 2 の特徴に適合しているかどうかをチェックする音源照合器 $p = \varphi_j(\mathbf{s}) \in [0, 1]$ $(j = 1, 2)$ を NN でモデル化することを考える。この照合器を各音源の振幅スペクトログラムの学習サンプルを用いて事前学習した上で，分離信号の振幅スペクトログラム $\mathbf{s}_1$, $\mathbf{s}_2$ の和が観測信号の振幅スペクトログラム $\mathbf{x}$ と等しくなる制約の下，$\mathbf{s}_1$, $\mathbf{s}_2$ を入力したときの音源照合器のスコアが最大になるように $\mathbf{s}_1$, $\mathbf{s}_2$ を推定することで，各音源の振幅スペクトログラムを得ることができる。あとは式 (5) を用いて分離信号を得ることができる。

この手法では目的関数の中に $\varphi_j(\mathbf{s}_{j'})$ という項が含まれるが，φ_j は NN となっているため，目的関数の $\mathbf{s}_1$ と $\mathbf{s}_2$ に関する勾配を計算することが可

能である。従って，$\mathbf{s}_1$ と $\mathbf{s}_2$ を推定には（NN の学習と同様）勾配法を用いることができる。

多チャンネル変分自己符号化器法

多チャンネル信号を対象とした音源推定アプローチとしては，式 (14) において，$\mathcal{V}_j = [v_j(f,n)]_{f,n}$ を NN によりモデル化する方法が提案されている [8–13]。例えば多チャンネル変分自己符号化器 (Multichannel Variational Autoencoder; MVAE) 法 [8–10] では，$\mathcal{V}_j = [v_j(f,n)]_{f,n}$ を

$$v_j(f,n) = g_j \cdot \sigma^2_\theta(f,n;\mathbf{z}_j,c_j) \qquad (15)$$

のように音源クラスインデックス c_j[1] を補助入力とし，$\boldsymbol{\sigma}^2_\theta(\mathbf{z}_j,c_j) = [\sigma^2_\theta(f,n;\mathbf{z}_j,c_j)]_{f,n}$ を出力する NN によりモデル化する。ただし，g_j はパワースペクトログラムのスケールを表す変数，$\sigma^2_\theta(f,n;\mathbf{z},c)$ は NN の出力 $\boldsymbol{\sigma}^2_\theta(\mathbf{z},c)$ の第 (f,n) 要素を表す。MVAE 法ではスペクトログラムの要素間の相関構造を学習できるようにするため，スペクトログラムをチャネル数が F，サイズが $1\times N$ の画像と見なし，CNN を用いてスペクトログラム全体をモデル化している点が特徴である。

クラスラベルつきの学習サンプルを用いて事前学習された上記音源モデルは，様々なクラスの音源のスペクトログラムを表現可能な生成モデルとなっており，c_j が音源クラスを表すパラメータ，$\mathbf{z}_j$ がクラス内変動を表すパラメータと見なせる。未知パラメータ $\mathcal{A}, \mathcal{Z} = \{\mathbf{z}_j\}_j, \mathcal{C} = \{c_j\}_j, \mathcal{G} = \{g_j\}_j$ の尤度関数は式 (14) と同形となるので，この尤度関数に基づき $\mathcal{A}, \mathcal{Z}, \mathcal{C}, \mathcal{G}$ を推定する最適化アルゴリズムを導くことができる。なお，（パワースペクトログラムではないが）パワースペクトルを NN でモデル化するアイディアはモノラル音源分離にも適用されている [14–17]。

3.3 時間周波数マスク推定アプローチ

深層クラスタリング法

NN を用いた時間周波数マスク推定法としては入力スペクトログラムに対し各時間周波数点の音源ラベルを直接予測する手法（例えば [18]）が主流であったが，この手法では各学習データ間で音源ラベルが一貫していない場合（例えば，音源 1 と音源 2 からなる混合信号のスペクトログラム A とスペクトログラム B を学習データとして，スペクトログラム A では音源 1 にクラスラベル 1，音源 2 にクラスラベル 2 が付与され，逆にスペクトログラム B では音源 1 にクラスラベル 2，音源 B にクラスラベル 1 が付与されている場合），学習された識別器は当然音源を識別する能力を持ちえない。このため，学習データを準備する際は，各スペクトログラム間で一貫したラベルを慎重に付与する必要があり，利用場面によっては難点となりえる。これに対し，前述のスペクトラルクラスタリング法と同様類似度行列の利用に着目した深層クラスタリング（Deep Clustering; DC）法と呼ぶ方法が提案されている [19]。DC 法は，混合信号のスペクトログラムの時間周波数点ごとに埋め込みベクトルを考え，同一音源が支配的な時間周波数点の埋め込みベクトルが互いに近接するように時間周波数点から埋め込みベクトルへの写像を NN を用いて学習することで，テスト時に埋め込みベクトルにクラスタリングを行うことにより各音源の時間周波数マスクを推定する方法である [19]。

J 個の音源からなる混合信号のスペクトログラムの全要素をベクトル化したものを $\mathbf{x} = [x_1,\ldots,x_k,\ldots,x_K]^\mathsf{T} \in \mathbb{R}^K$ とする。ただし，k は時間周波数点 (f,n) に対応するインデックスを表し，K は時間周波数点の総数 $F\times N$ である。ここで，スペクトログラムの各点 k ごとにノルムが 1 の D 次元埋め込みベクトル $\mathbf{v}_k = [v_{k,1},\ldots,v_{k,D}]$ を考え，$\mathbf{V} = [\mathbf{v}_1;\ldots;\mathbf{v}_K] \in \mathbb{R}^{K\times D}$ とする。ただし，; は行列の改行を表す。DC 法では，同一音源が支配的な時間周波数点の埋め込みベクトルが互いに接近するように非線形写像 $\mathbf{V} = \varphi_\theta(\mathbf{x})$ を学習することが目標である。埋め込みベクトルがスペクトログラムの大域的な構造を手がかりにして決定される仕組みにするため，非線形写像 φ_θ は RNN や CNN を用いてモデル化される [19, 20]。θ はそのパラメータを表す。$\mathbf{x}$ の各時間周波数点 k で支配的な音源ラベルを示した one-hot ベクトル（行ベクトル）を $\mathbf{y}_k \in \{0,1\}^{1\times J}$ とし，$\mathbf{Y} = [\mathbf{y}_1;\ldots;\mathbf{y}_K] \in \{0,1\}^{K\times J}$ とすると，

$$\begin{aligned} J(\theta) &= \|\mathbf{V}\mathbf{V}^\mathsf{T} - \mathbf{Y}\mathbf{Y}^\mathsf{T}\|^2_F \qquad (16)\\ &= \|\mathbf{V}^\mathsf{T}\mathbf{V}\|^2_F - 2\|\mathbf{V}^\mathsf{T}\mathbf{Y}\|^2_F + \|\mathbf{Y}^\mathsf{T}\mathbf{Y}\|^2_F \end{aligned}$$

を θ に関してできるだけ小さくすることが DC 法の学習目標となる。ただし，$\|\cdot\|^2_F$ はフロベニウスノルムを表す。$\mathbf{Y}\mathbf{Y}^\mathsf{T}$ は，k 行 k' 列目の要素が

[1] ここで音源クラスインデックスは，例えば音声の場合は話者 ID を表すインデックスを表す。

時間周波数点 $k=(f,n)$ と $k'=(f',n')$ において同一音源が支配的のときに 1，そうでないときに 0 であるような $K\times K$ のバイナリ行列で，一種の類似度行列と見なせる。つまり DC 法では，時間周波数点のペアの埋め込みベクトルの内積を要素にした行列がこの類似度行列とできるだけ一致するように φ_θ を決めていることになる。φ_θ の学習後，入力信号のスペクトログラム $\mathbf{x}$ に対し $\mathbf{V}$ を算出し，$\mathbf{V}$ の各行ベクトルをデータベクトルとしてクラスタリング（k 平均法など）を行うことで，同一音源が支配的な時間周波数点の集合を得ることができる。これにより，各音源を抽出するための時間周波数マスクを構成することができる。

DC 法では学習データとして，各時間周波数点に付与される音源ラベル Y を必要とせず，代わりに，各スペクトログラムの時間周波数点のペアごとに支配的な音源が同一かどうかを示す類似度行列 $\mathbf{YY}^\mathsf{T}$ を用いる手法となっている。このようなラベルの付与にかかる労力は，全データ間で一貫した音源ラベルを付与する労力に比べて小さく済むため，実用上のメリットが大きい。なお，音源クラスを直接予測するタイプの手法においても，音源とラベルの全通りの対応付けを考え，その中で最小となる学習ロスを最小化することで，音源ラベルが一貫していなくともクラス識別器を適切に学習できる方法も提案されている [21]。また，DC 法は各音源のスペクトログラムの特徴を捉えようとするのではなく単一音源のスペクトログラムで構造的に共通するパターン（調波構造など）を捉えようとする手法になっていると考えられ，学習データに含まれない話者の音声分離も高精度に行えることが実験的に示されている。近年は DC 法の多チャンネル拡張の試みも行われている [22]。

深層アトラクタネットワーク法

DC 法では各時間周波数点から埋め込みベクトルを算出するプロセスと，各点の埋め込みベクトルに対してクラスタリングを行い時間周波数マスクを構成するプロセスは独立しており，写像 φ_θ の学習規準が時間周波数マスクにより得られる分離信号そのものが最適となるような規準となっていなかった。DC 法の考え方をベースにしつつこの点を改良したのが深層アトラクタネットワーク（Deep Attractor Network; DANet）法である [23]。

音源 j の学習サンプルのスペクトログラムを $s_j(f,n)$ とし，$s_1(f,n),\ldots,s_J(f,n)$ を混合して作られた混合信号を $x(f,n)$ とすると，DANet 法では，$x(f,n)$ から $s_1(f,n),\ldots,s_J(f,n)$ をできるだけ高精度に復元する時間周波数マスク $m_j(f,n)$ を得るプロセスを学習することが目標となる。そこでまず，

$$\mathcal{J}(\theta)=\sum_{j,f,n}|s_j(f,n)-m_j(f,n)x(f,n)|^2 \tag{17}$$

のような規準を考え，これを小さくすることを目標設定する。ここで，時間周波数マスク $m_j(f,n)$ は埋め込みベクトル $\mathbf{v}_k$ から

$$m_j(f,n)=\mathrm{sigmoid}\left(\sum_d \alpha_{j,d}v_{k=(f,n),d}\right)$$
$$\text{where}\quad \alpha_{j,d}=\frac{\sum_k v_{k,d}y_{k,j}}{\sum_k y_{k,j}} \tag{18}$$

のようなプロセスによって算出されるものと仮定する。中間変数の $\alpha_{j,d}$ は，音源 j が支配的とラベル付けされた全時間周波数点における埋め込みベクトルの重心に相当する。従って，$m_j(f,n)$ は，ある時間周波数点 $k=(f,n)$ における埋め込みベクトルがその重心と近いほど 1 に近くなり，遠いほど 0 に近くなるものとなっているため，音源 j が支配的な時間周波数点における埋め込みベクトルが重心 $\alpha_{j,d}$ 周辺に集中しているとき $m_j(f,n)$ は音源 j の成分だけを通過させる時間周波数マスクとなることが分かる。以上より DANet 法は，式 (18) の関係式により埋め込みベクトルと時間周波数マスクが関連付けられたことで，式 (17) の復元誤差を直接小さくするように埋め込みベクトルへの写像関数 φ_θ を学習できる方式となっている。

4. おわりに

本稿では音源分離の問題に焦点を当て，特に深層学習を用いた音源分離法についての最近の動向を紹介した。本分野は近年急速に発展をとげており，今後も動向が注目される。

文　献

[1] O. Yılmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Trans. Signal Process.*, **52**, 1830–1847 (2004).

[2] K. Han and D. Wang, “An SVM based classification approach to speech separation,” *Proc. ICASSP 2011*, pp. 4632–4635 (2011).

[3] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," *J. Mach. Learn. Res.*, **7**, pp. 1963–2001 (2006).
[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, **9**, 1735–1780 (1997).
[5] Y. N. Dauphin, A. Fan, M. Auli and D. Grangier, "Language modeling with gated convolutional networks," *Proc. ICML*, pp. 933–941 (2017).
[6] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv:1609.03499* (2016).
[7] E. M. Grais, M. U. Sen and H. Erdogan, "Deep neural networks for single channel source separation," *Proc. ICASSP 2014*, pp. 3734–3738 (2014).
[8] H. Kameoka, L. Li, S. Inoue and S. Makino, "Semi-blind source separation with multichannel variational autoencoder," *arXiv:1808.00892* (2018).
[9] S. Seki, H. Kameoka, L. Li, T. Toda and K. Takeda, "Generalized multichannel variational autoencoder for underdetermined source separation," *arXiv:1810.00223* (2018).
[10] L. Li, H. Kameoka and S. Makino, "Fast MVAE: Joint separation and classification of mixed sources based on multichannel variational autoencoder with auxiliary classifier," *arXiv:1812.06391* (2018).
[11] S. Mogami, H. Sumino, D. Kitamura, N. Takamune, S. Takamichi, H. Saruwatari and N. Ono, "Independent deeply learned matrix analysis for multichannel audio source separation," *Proc. EUSIPCO*, pp. 1557–1561 (2018).
[12] K. Sekiguchi, Y. Bando, K. Yoshii and T. Kawahara, "Bayesian multichannel speech enhancement with a deep speech prior," *Proc. APSIPA-ASC*, pp. 1233–1239 (2018).
[13] S. Leglaive, L. Girin and R. Horaud, "Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization," *arXiv:1811.06713* (2019).
[14] P. Smaragdis and S. Venkataramani, "A neural network alternative to non-negative audio models," *Proc. ICASSP 2017*, pp. 86–90 (2017).
[15] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," *Proc. ICASSP 2018*, pp. 716–720 (2018).
[16] Y. Subakan and P. Smaragdis, "Generative adversarial source separation," *Proc. ICASSP 2018*, pp. 26–30 (2018).
[17] S. Leglaive, L. Girin and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," *Proc. MLSP*, pp. 1–6 (2018).
[18] Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, **21**, 65–68 (2014).
[19] J. R. Hershey, Z. Chen, J. Le Roux and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," *Proc. ICASSP 2016*, pp. 31–35 (2016).
[20] L. Li and H. Kameoka, "Deep clustering with gated convolutional networks," *Proc. ICASSP 2018*, pp. 16–20 (2018).
[21] D. Yu, M. Kolbak, Z. Tan and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," *Proc. ICASSP 2017*, pp. 241–245 (2017).
[22] Z. Wang, J. Le Roux and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," *Proc. ICASSP 2018*, pp. 1–5 (2018).
[23] Z. Chen, Y. Luo and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," *Proc. ICASSP 2017*, pp. 246–250 (2017).