修士論文

音光変換デバイス Blinky を用いた
複数音源の強調と定位

Multiple source enhancement and localization with sound-to-light conversion device Blinky

堀池 大樹

指導教員　　小野 順貴 教授

2021 年 2 月 19 日

東京都立大学大学院
システムデザイン研究科
情報科学域 博士前期課程

あらまし

本研究では，Blinky を用いた複数音源を目的とする音声強調と音源定位の手法を提案する．我々の研究室では，音光変換デバイス Blinky を開発し，ビデオカメラを用いて音響空間の音強度情報を取得する新しい枠組みを開発してきた．Blinky とは，マイクロホンと LED により音強度を光強度に変換する小型デバイスであり，これをのように音響空間に広く分散配置してビデオカメラを用いて観測することで，広範囲の音強度情報を簡単に取得することができる．Blinky を使用することで，従来のマイクロホンを用いた収録で必要な有線接続や無線通信を行わずに多チャネルの収録が可能となる．Blinky によって目的音源の音強度信号を取得すれば，これらの音響信号処理が可能だが，音源が複数存在する場合には，通常のマイクロホンアレイと同様に，音強度情報は混合し，個々の音源の音強度情報が独立には観測することができないといった問題が生じる．

この問題を解決するために，私は以前に，Blinky が観測した複数音源の混合音強度信号を，個々の音源音強度情報へ分離する手法を提案する．具体的には，観測信号を並べた音強度行列は，低ランクな時間空間構造を持つ制約により，非負値行列因子分解を適用することで，空間伝達関数と音源の音強度信号の行列の積に分解できる．この手法の性能を計算機シミュレーション及び実環境にて評価し，その有効性を確認した．本研究では，この Blinky の音強度分離を用いて従来困難であった複数音源を目的とする音声強調と音源定位の手法を提案する．

複数音声強調では，マイクロホンアレイと組み合わせることでビームフォーマを設計し実現する．具体的には，非負値行列因子分解により推定される各音源のアクティビティを用いて目的音源区間の検出を行い，マイクロホンアレイの観測信号から目的音と雑音それぞれの共分散行列を求めることで，信号対雑音比最大化ビームフォーマを設計する．各目的音源に対して強調フィルタを設計することで複数音源の音声強調を実現できる．提案するビームフォーマを計算機シミュレーションと実環境における実験にて性能を確認する．結果として，Blinky を用いた提案手法は，従来手法と比較して，特に信号対雑音比が低い場合に優れた性能を得られた．

また複数音源定位では，NMF により推定される各音源の空間伝達関数ゲインを音響特徴量とし，音源位置とのマッピングを深層ニューラルネットワークにて学習する．この手法は音強度分離により複数音源定位の問題を複数の単一音源定位の問題として扱うことができるため，DNN においては単一音源の音響特徴量と音源位置のデータセットさえあれば学習可能でありデータの作成コストを低下させるといった利点がある．提案する音源定位手法を計算機シミュレーションにて性能を確認する．結果として，提案手法は効果的であったことが確認できた．

目次

図目次

表目次

第 1 章

はじめに

1.1 研究の背景・目的

音源の位置を推定する音源定位 [1]，目的音源を強調する音源強調 [2]，混ざりあって観測される複数の信号を分離する音源分離 [3, 4]，これらの音響信号処理の技術は，幅広い分野で必要とされており，マイクロホンアレイ信号処理は音響信号処理の中でも活発に研究が行われている．マイクロホンアレイ信号処理とは，複数のマイクロホンで取得した多チャンネル信号を処理し，単一のマイクロホンでは困難な，音源定位，音源強調，音源分離等を，音源の空間情報を用いて行う枠組みである．一般的に，用いるマイクロホンの数が多いほど得られる音空間情報が多くなるため，制御できる指向性の自由度が増し，そして，マイクロホンを広範囲に配置することができるほどカバーできる範囲が広がり，音響信号処理の性能の向上が期待できる．しかし，従来のマイクロホンを広範囲に配置するには数量が必要となり，数を増やすほど有線接続や無線通信が困難になるため導入が難しくなるといった問題がある．

これを解決するために我々の研究室では，音光変換デバイス Blinky を開発し，ビデオカメラを用いて音響空間の音強度情報を取得する新しい枠組みを開発してきた．Blinky とは，図 1 のような，マイクロホンと LED により音強度を光強度に変換する小型デバイスであり，これを図 2 のように音響空間に広く分散配置してビデオカメラを用いて観測することで，広範囲の音強度情報を簡単に取得することができる [5, 6]．Blinky を使用することで，従来のマイクロホンを用いた収録で必要な有線接続や無線通信を行わずに多チャネルの収録が可能となる．また，現実では多くの場合，スマートフォンやテレカンファレンス等のようにビデオカメラはマイクロホンと同時に利用される機会が多く，ビデオに音響情報を埋め込む効果も期待できる．Blinky を用いたこれまでの研究として，目的音源付近に配置することでその音強度情報を利用したビームフォーミングや，空間に分散配置することで音源の音強度情報を利用した音源定位等の音響信号処理への応用が行われてきた．Blinky によって目的音源の音強度信号を取得すれば，これらの音響信号処理が可能だが，音源が複数存在する場合には，通常のマイクロホンアレイと同様に，音強度情報は混合し，個々の音源の音強度情報が独立には観測することができないといった問題が生じる．

この問題を解決するために，私は以前に，Blinky が観測した複数音源の混合音強度信号を，個々の音源音強度情報へ分離する手法を提案した [7]．具体的には，観測信号を並べた音強度行列は，低ランクな時間空間構造を持つ制約により，非負値行列因子分解 (non-negative matrix factorization; NMF) を適用することで，空間伝達関数と音源の音強度信号の行列の積に分解できる．この手法の

性能を計算機シミュレーション及び実環境にて評価し，その有効性を確認した．本研究では，このBlinky の音強度分離を用いて従来困難であった複数音源を目的とする音声強調と音源定位の手法を提案する [7,8].

複数音声強調では，マイクロホンアレイと組み合わせることでビームフォーマを設計し実現する．具体的には，NMF により推定される各音源のアクティビティを用いて目的音源区間の検出を行い，マイクロホンアレイの観測信号から目的音と雑音それぞれの共分散行列を求めることで，signal-to-interference-and-noise ratio (SINR) 最大化ビームフォーマを設計する．各目的音源に対して強調フィルタを設計することで複数音源の音声強調を実現できる．提案するビームフォーマを計算機シミュレーションと実環境における実験にて性能を確認する．

また複数音源定位では，NMF により推定される各音源の空間伝達関数ゲインを音響特徴量とし，音源位置とのマッピングを深層ニューラルネットワーク (deep neural network; DNN) にて学習する．この手法は音強度分離により複数音源定位の問題を複数の単一音源定位の問題として扱うことができるため，DNN においては単一音源の音響特徴量と音源位置のデータセットさえあれば学習可能でありデータの作成コストを低下させるといった利点がある．提案する音源定位手法を計算機シミュレーションにて性能を確認する．

1.2 関連研究

音響センシングの手段として光を使用する研究は，視覚化 [9] や通信 [10] を目的として過去にも行われてきた．近年では，カエルの合唱音声の観察を行うセンサーのデザインや，そのアルゴリズムの研究が行われている [11].

1.3 本論文の構成

第 2 章では，音光変換デバイス Blinky とビデオカメラを用いた音響センシングの詳細についての説明を行う．第 3 章では，音強度信号分離のアルゴリズムについて述べる．第 4，5 章では，複数音声強調，複数音源定位をそれぞれ提案する．第 6 章では，計算機シミュレーション及び実環境での実験により，提案手法の有用性と性能評価についての確認を行う．最後に第 7 章では，全体のまとめを行う．

図 1　音光変換デバイス Blinky

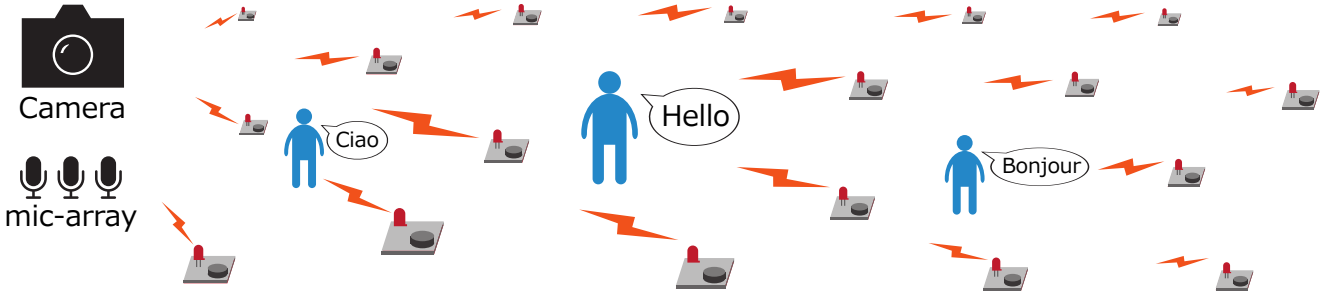

図 2　音光変換の構造

第 2 章

Blinky とビデオカメラによる音響センシング

本章では，音光変換デバイス Blinky のハードウェアと音光変換のシステム，そして，ビデオカメラで Blinky の音強度を取得する方法について説明を行う．

2.1 ハードウェアの構造

図 3 に Blinky の回路基板を示す．Blinky は system-on-a-chip (SoC) のマイクロコントローラーの ESP32 [12] を内臓しており，以下の特徴を持つ低コストな音響センサーとなっている．

- CPU：デュアルコア (最大 240 MH ｚで動作)
- メモリー: 520 KB SRAM
- 無線接続: Wi-Fi，Bluetooth
- 低消費電力
- C++，Arduino，MicroPython でプログラム可能
- MEMS マイクロホン × 2
- LED × 4

無線通信は，ファームウェアの更新とデバッグのための音響信号取得が主な目的である．マイクロホンはデジタル出力 MEMS マイクロホン (ICS-43432) を使用しており，これは ESP32 の I2S 規格により直接接続されている．また，4 つの LED は低消費電力で，それぞれ赤，緑，青，白の異なる色の LED を使用している．

2.2 Blinky による音光変換

Blinky は，ESP32 により自由にプログラムが可能なため，周波数帯域フィルタリングを行う等の特定の処理を施すことができる．ここでは，本研究の実環境実験で使用している Blinky について説明する．

Blinky は，マイクロホンの音響パワーから PWM (Pulse Width Modulation) への変換処理を行っている．観測信号からサンプル平均の音響パワーを求め，LED が駆動する 12-bit で表現される PWM へ対応させている．これは音声の音響パワーの多くの範囲を含むように，TIMIT コーパス [13] のデータセットの音声から作成した図 4 のような累積分布関数 (cumulative distribution

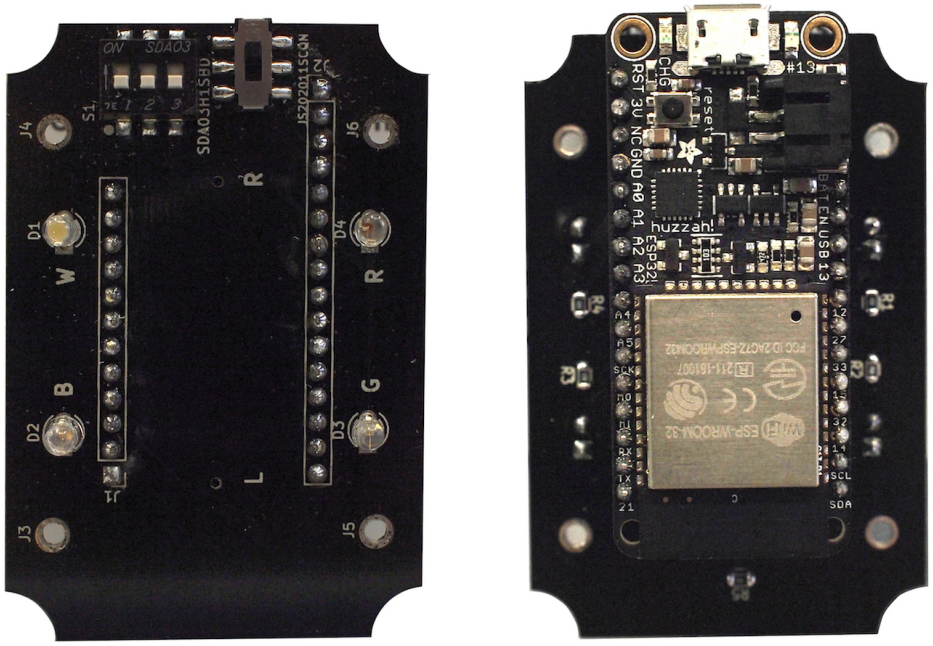

図 3　回路基板の上部 (左) と下部 (右)

function; CDF) をもとに PWM へと変換している．よって，Blinky の音光変換は，音響パワーから PWM に変換する関数と PWM の周期信号と画素値の非線形関数の逆関数の 2 つを用いており，図 5 のような非線形関数 $\varphi(\cdot)$ となっている．

Blinky の音響パワー $u[n]$ は，非線形関数 φ により B-bit の PWM$\ell[n] \in 0, \cdots, 2^B - 1$ に変換される．ここで，n は時間インデックスとする．これより，LED の光強度 $I[n]$ は，

$$I[n] = \frac{\ell[n]}{2^B - 1} I_{\max}, \quad \ell[n] = \varphi(u[n]), \tag{1}$$

のように表される．ただし，$I_{\max}$ は連続して駆動されている LED の強度の最大値である．

2.3　ビデオカメラによる Blinky の撮影

Blinky での音光変換後，LED の光は空中を伝搬し，ビデオカメラによって撮影される．ビデオカメラでの LED の光強度は角度や距離に依存する減衰係数 α と，周囲の光が加算されたバイアス β の影響を受ける．これらの理由から，ビデオカメラで観測される光強度 $v[n]$ は，減衰係数 α とバイアス β を用いて，

$$v[n] = \alpha I[n] + \beta. \tag{2}$$

と表される．

ビデオカメラのイメージセンサは光強度を撮影し，ビデオファイルにエンコードする．一般的に，消費者向けのビデオカメラはイメージセンサの出力を非線形な処理を経て画素値へ変換するが，産業

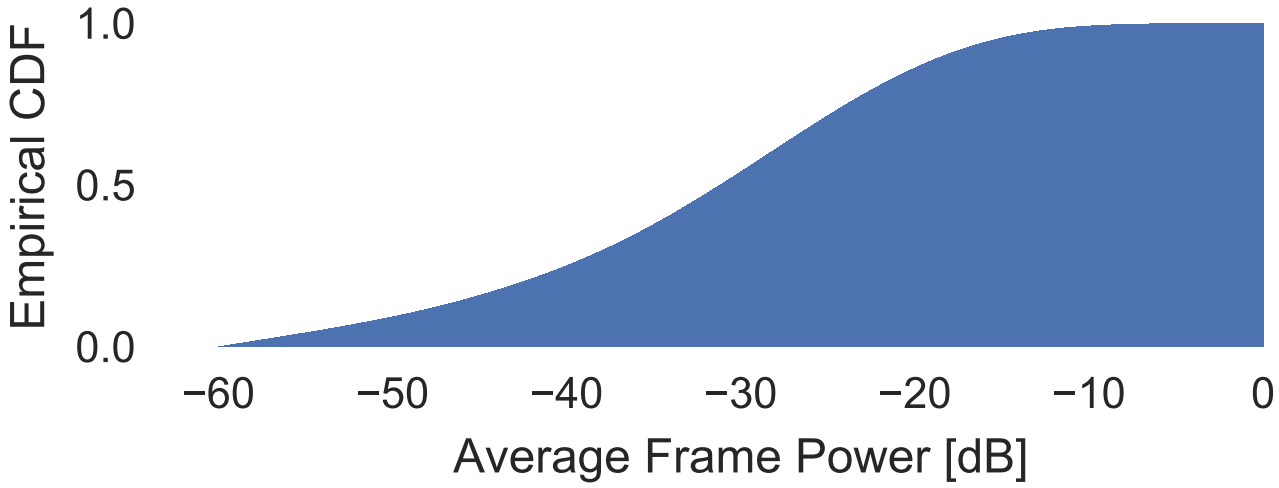

図 4　フレーム平均音響パワーの累積分布関数.

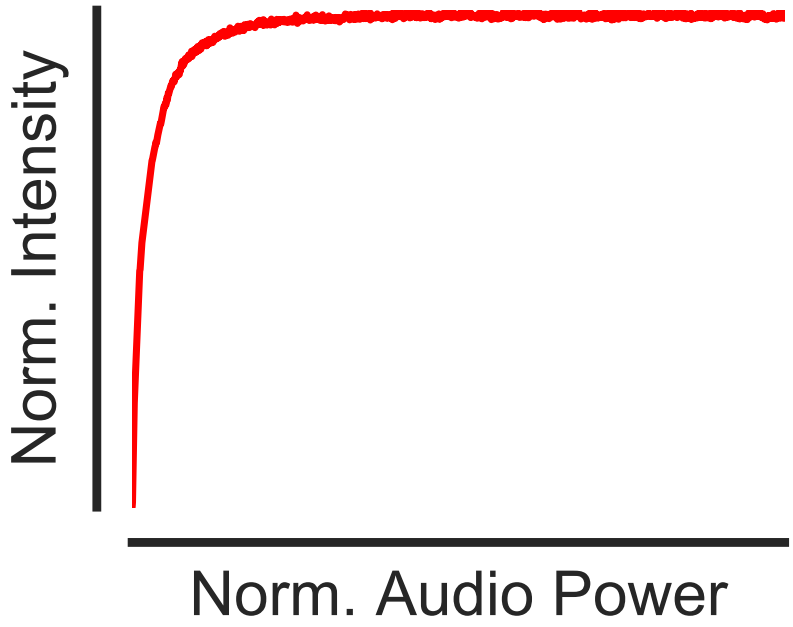

図 5　音響パワーから画素値への非線形関数 (Blinky 内部の処理).

用カメラはイメージセンサの出力を直接画素値へ反映させる．この典型的な非線形な処理の一つとして，*Gamma correction* がある．これはイメージセンサの出力 $v[n]$ を $p[n] = (v[n])^{1/\gamma}, \gamma = 2.2$ となるように変換させる非線形関数であり，ディスプレイ上に自然に表示するために，画像の輝度や彩度を調整する目的で設定されている [14]．この非線形な処理を避けるために，我々は産業用カメラを用いて $v[n]$ を取得する．これにより，本研究では $p[n] = v[n]$ と仮定する．

2.4 音強度の推定

式 (1) の非線形関数，式 (2) の伝搬により，画素値 $p[n]$ は Blinky での音響パワー $u[n]$ から変換されている．この音響パワー $u[n]$ を画素値 $p[n]$ から復元するために，α と β の両方を推定する必要がある．その復元方法を説明する．Blinky が既知のパイロット信号を送信するかどうかによってキャリブレーションを行う方法がいくつかあるが，本研究では，キャリブレーションのために 2 つ目の補助 LED を使用する方法について説明する．図 6 は，音響パワーから画素値への伝搬モデルをまとめたものである．ℓ_{sig} と ℓ_{ref} をそれぞれ信号用，キャリブレーション用の LED とする．α と β が同じとなるように 2 つの LED は十分に近いと仮定する．ここで，式 (1), 式 (2) より，ビデオカメラで得られる光強度は

$$p_{\mathrm{sig}}[n] = v_{\mathrm{sig}}[n] = \alpha \frac{\ell_{\mathrm{sig}}[n]}{2^B - 1} I_{\max}^{(\mathrm{sig})} + \beta, \tag{3}$$

$$p_{\text{ref-lo}} = v_{\text{ref-lo}} = \beta, \tag{4}$$

$$p_{\text{ref-hi}} = v_{\text{ref-hi}} = \alpha \frac{\ell_{\mathrm{ref}}}{2^B - 1} I_{\max}^{(\mathrm{ref})} + \beta, \tag{5}$$

と表される．ただし， v_{sig} は信号用の LED, $v_{\text{ref-lo}}$ と $v_{\text{ref-hi}}$ はそれぞれキャリブレーション用の LED の最小値と最大値である．式 (1)–式 (5) より，推定した α，β，そして音響パワー $\hat{u}_{\mathrm{sig}}$ は，

$$\alpha = (p_{\text{ref-hi}} - p_{\text{ref-lo}}) \frac{2^B - 1}{I_{\max}^{(\mathrm{ref})}} \frac{1}{\ell_{\mathrm{ref}}}, \tag{6}$$

$$\beta = p_{\text{ref-lo}}, \tag{7}$$

$$\begin{aligned} \hat{u}_{\mathrm{sig}}[n] &= \varphi^{-1}(\ell_{\mathrm{sig}}[n]) \\ &= \varphi^{-1}\left(\frac{p_{\mathrm{sig}}[n] - p_{\text{ref-lo}}}{p_{\text{ref-hi}} - p_{\text{ref-lo}}} \frac{I_{\max}^{(\mathrm{ref})}}{I_{\max}^{(\mathrm{sig})}} \ell_{\mathrm{ref}} \right). \end{aligned} \tag{8}$$

と表される．ただし，ビデオカメラのフレームレート制限により，周波数情報は復元されない．

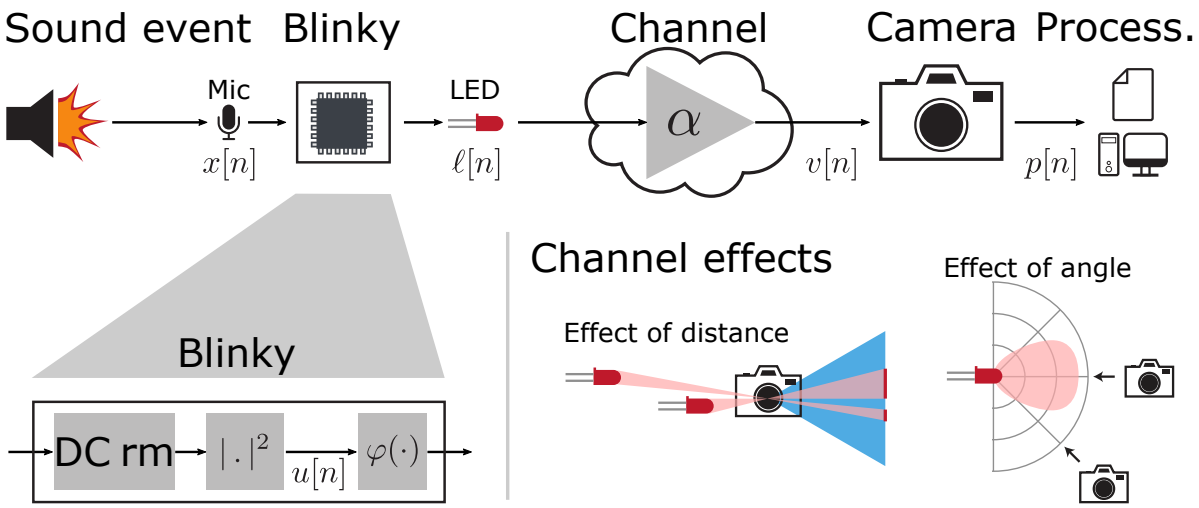

図 6　上: マイクロホンの観測信号からビデオファイルへの変換モデル．左下: Blinky の内部処理．右下: Blinky の LED からビデオカメラへの光の伝搬の影響．

第 3 章

非負値行列因子分解による音強度分離

本章では，複数音源の混合音強度信号分離における問題設定と混合モデル，NMF の適用による解法について説明を行う．

3.1 混合モデル

図 6 上に Blinky とビデオカメラによる音響パワー (以下，音強度とする) の取得モデルを示す．Blinky は，マイクロホンの観測信号から音強度を求め，これに対応した PWM で LED が駆動する．この LED の光強度をビデオカメラで撮影し，音強度を取得することで，Blinky とビデオカメラによる音響センシングを行う．しかし，複数の音源が同時に存在している場合には，通常のマイクロホンと同様に，各音源が混合するため，個々の音源の音強度を取得することは困難である．この問題を解決するために，以前提案した NMF による音強度分離の理論を本章ではまとめる．

図 7 のように K 個の音源を B 個の分散配置した Blinky で観測することを考える．Blinky$b$ のマイクロホンの観測信号を短時間フーリエ変換 (STFT) 領域で $x_b[f,n]$ とすると，

$$x_b[f,n] = \sum_{k=1}^{K} a_{bk}[f] s_k[f,n], \tag{9}$$

と表される．ここで，$a_{bk}[f]$ は音源 k の位置から Blinky b への伝達関数，$s_k[f,n]$ は音源信号 k，f は周波数，n は時間フレームを表す．

いま，各音源は無相関，周波数応答のゲインは周波数に依存せず一定，STFT のフレーム長は残響時間より十分長いと仮定すると，Blinky b が時間フレーム n で観測する音響パワー (以下，音強度と

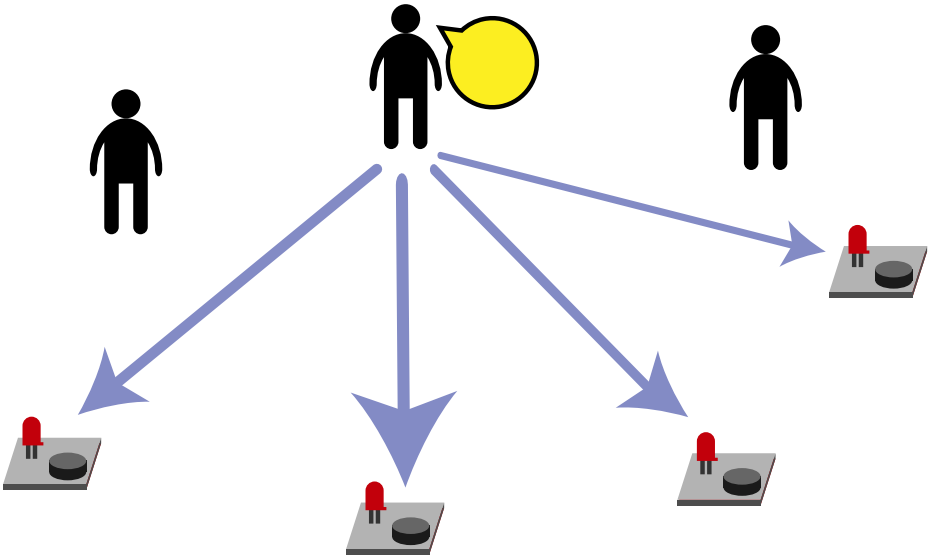

図 7　混合モデル

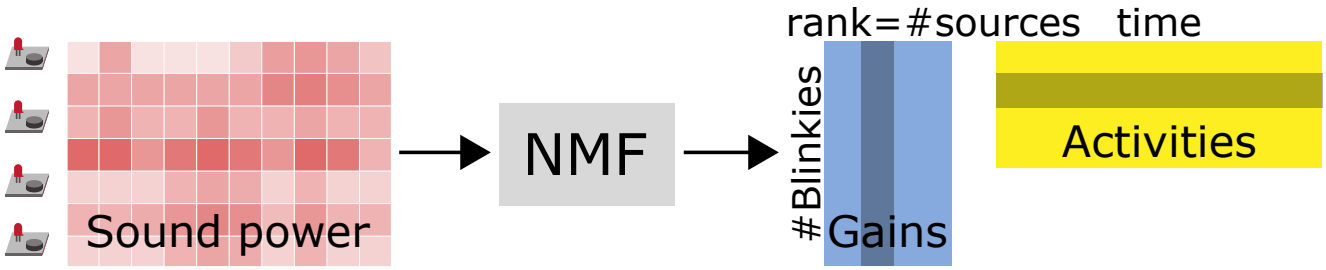

図 8　音強度信号の積和モデル化

呼ぶ) $u_b[n]$ は,

$$u_b[n] = \sum_{f=1}^{F} \left| \sum_{k=1}^{K} a_{bk}[f] s_k[f,n] \right|^2 \tag{10}$$

$$\approx \sum_{k=1}^{K} \sum_{f=1}^{F} |a_{bk}[f]|^2 |s_k[f,n]|^2 \tag{11}$$

$$= \sum_{k=1}^{K} g_{bk} \sum_{f=1}^{F} |s_k[f,n]|^2 \tag{12}$$

$$= \sum_{k=1}^{K} g_{bk} \sigma_k^2, \tag{13}$$

と表される [7]. ただし, $g_{bk} = |a_{bk}[f]|^2$, $\sigma_k^2[n] = \sum_{f=1}^{F} |s_k[f,n]|^2$ であり, それぞれ伝達関数ゲインと音源アクティビティを示している. 図 8 ように, 音強度は伝達関数ゲインと音源アクティビティの積和としてモデル化することができる.

3.2 非負値行列因子分解による音強度信号分離

実世界にはパワースペクトル，画素値，頻度などの非負値で表されるデータが多い．このような非負値のデータを加法的な構成成分に分解することを目的とした多変量解析手法の 1 つに NMF がある [15]．本節では，NMF のモデルとアルゴリズムについて説明する．

ある観測値からなる $M \times T$ の非負値行列 $\boldsymbol{V}$ は，低ランクの場合，図 9 のように $M \times K$ の非負値行列 $\boldsymbol{W}$ と $K \times T$ の非負値行列 $\boldsymbol{H}$ に近似的に分解することができると考え，NMF は $\boldsymbol{W}$ と $\boldsymbol{H}$ を推定する問題となる．ただし，K は NMF の基底数，それぞれの行列の要素 $v_{m,n}, w_{m,k}, h_{k,n}$ は非負値である．分解した行列の要素からなる推定値 $\hat{v}_{m,n}$ は

$$\hat{v}_{m,n} = \sum_{k=1}^{K} w_{m,k} h_{k,n}, \tag{14}$$

であり，一般的に観測値 $v_{m,n}$ と誤差が発生するため，これらの距離を定義し，最小化することで分解後の行列を推定する．$\boldsymbol{V}$ と $\boldsymbol{WH}$ の距離 $D(\boldsymbol{V}, \boldsymbol{WH})$ を，

$$D_{\text{cost}}(\boldsymbol{V}, \boldsymbol{WH}) = \sum_{m=1}^{M} \sum_{n=1}^{N} d_{\text{cost}}(v_{m,n}, \hat{v}_{m,n}), \tag{15}$$

と定義する．ただし，cost はコスト関数を表しており，NMF では，ユークリッド距離 (EUC)，Kullback-Leibler ダイバージェンス (KL)，Itakura-Saito ダイバージェンス (IS) の 3 種類が広く用いられている．d_{cost} はそれぞれ，

$$d_{\text{EUC}}(v_{m,n}, \hat{v}_{m,n}) = (v_{m,n} - \hat{v}_{m,n})^2, \tag{16}$$

$$d_{\text{KL}}(v_{m,n}, \hat{v}_{m,n}) = v_{m,n} \log \frac{v_{m,n}}{\hat{v}_{m,n}} - v_{m,n} + \hat{v}_{m,n}, \tag{17}$$

$$d_{\text{IS}}(v_{m,n}, \hat{v}_{m,n}) = \frac{v_{m,n}}{\hat{v}_{m,n}} - \log \frac{v_{m,n}}{\hat{v}_{m,n}} - 1, \tag{18}$$

となる．NMF の基底数 K は，分解する成分数を定めるものであり，アルゴリズムを実行する際に決定しなければならない．

D_{EUC}，D_{KL}，D_{IS} それぞれを最小化するアルゴリズムは様々あるが，本節では広く用いられている Multiplicative update rules を説明する．3 種類の距離の更新式はそれぞれ，
ユークリッド距離

$$h_{k,n} \leftarrow h_{k,n} \frac{\sum_m w_{m,k} v_{m,n}}{\sum_m w_{m,k} \hat{v}_{m,n}}, \qquad w_{m,k} \leftarrow w_{m,k} \frac{\sum_t v_{m,n} h_{k,n}}{\sum_t \hat{v}_{m,n} h_{k,n}}, \tag{19}$$

KL ダイバージェンス

$$h_{k,n} \leftarrow h_{k,n} \frac{\sum_m w_{m,k} v_{m,n} \hat{v}_{m,n}^{-2}}{\sum_m w_{m,k}}, \qquad w_{m,k} \leftarrow w_{m,k} \frac{\sum_t v_{m,n} \hat{v}_{m,n}^{-2} h_{k,n}}{\sum_t h_{k,n}}, \tag{20}$$

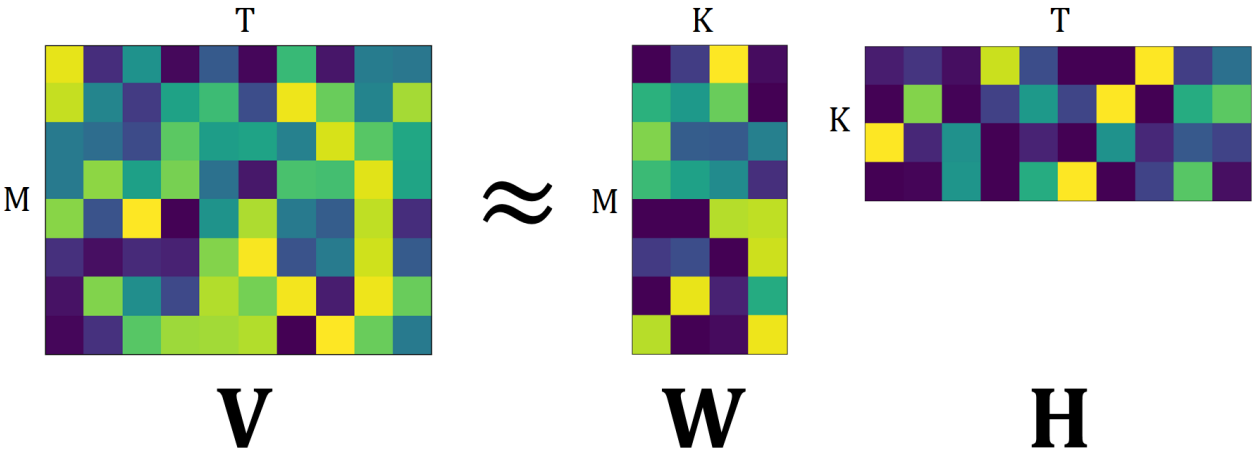

図 9　NMF のモデル

IS ダイバージェンス

$$h_{k,n} \leftarrow h_{k,n}\sqrt{\frac{\sum_m w_{m,k} v_{m,n} \hat{v}_{m,n}^{-2}}{\sum_m w_{m,k} \hat{v}_{m,n}^{-1}}}, \qquad w_{m,k} \leftarrow w_{m,k}\sqrt{\frac{\sum_t v_{m,n} h_{k,n} \hat{v}_{m,n}^{-2}}{\sum_t h_{k,n} \hat{v}_{m,n}^{-1}}}, \tag{21}$$

となる [15, 16]．ランダムな非負の初期値を代入した行列 $\boldsymbol{W}$，$\boldsymbol{H}$ に上記の更新式を繰り返し適用することで，分解後の行列 $\boldsymbol{W}$，$\boldsymbol{H}$ を得られる．式 (19)–式 (21) は補助関数法を用いたアルゴリズムで導出しており，その導出方法は付録 A に記述する．

式 (13) は，NMF と同様のモデルであり，Blinky の音強度信号を各チャネルごとに並べた時間空間行列は，伝達関数ゲインと各音源のアクティビティに分解できると考えられる．すなわち，$u_b[n]$，$g_{b,k}$，$\sigma_k^2[n]$ それぞれを要素に持つ行列，

$$\boldsymbol{P} = \begin{pmatrix} u_1[0] & \dots & u_1[N] \\ \vdots & \ddots & \vdots \\ u_B[0] & \dots & u_B[N] \end{pmatrix},$$
$$\boldsymbol{G} = \begin{pmatrix} g_{1,1} & \cdots & g_{1,K} \\ \vdots & \ddots & \vdots \\ g_{B,1} & \cdots & g_{B,K} \end{pmatrix},$$
$$\boldsymbol{H} = \begin{pmatrix} \sigma_1^2[0] & \dots & \sigma_1^2[N] \\ \vdots & \ddots & \vdots \\ \sigma_K^2[0] & \dots & \sigma_K^2[N] \end{pmatrix},$$

を用いて，

$$\boldsymbol{U} \approx \boldsymbol{G}\boldsymbol{H}, \tag{22}$$

のように NMF のモデルで表すことができる．ここで，$\boldsymbol{U} \in \mathbb{R}_+^{B\times N}$ は Blinky が取得した音強度信

号，$\boldsymbol{G} \in \mathbb{R}_{+}^{B \times K}$ は伝達関数ゲイン，$\boldsymbol{H} \in \mathbb{R}_{+}^{K \times N}$ は音源アクティビティ，N は音強度信号のフレーム総数である．

第 4 章

Blinky とマイクロホンアレイを用いた複数音声強調

本章では，複数音源に対する音声強調手法について説明を行う．

4.1 問題設定

M 個のマイクロホンと B 個の分散配置した Blinky を用いて K 個の音源を強調することを考える．すなわち，短時間 Fourier 変換 (STFT) 領域でのマイクロホン m の観測信号 $x_m[f,n]$ から目的音源 k を強調するビームフォーマの出力信号 $y_k[f,n]$ を推定する

提案する音強度分離に基づく音声強調手法では，SINR 最大化ビームフォーマを用いる．このビームフォーマは，目的音源と非目的音源の共分散行列を用いることで，明示的な音源の方向情報を必要とせずにフィルタベクトルを設計できる．一般的に，これら共分散行列の推定は事前情報として目的音源の有声区間，すなわち，音声区間検出 (voice activity detection; VAD) が必要であるが，Blinky を用いることでこれを推定し，強調フィルタを設計することができる．しかし，複数の音源が同時に存在している場合には，各音源が混合するため，各目的音源の有声区間を推定することは困難である．この問題を解決するために，各音源のアクティビティを得られる音強度分離を用いた音声強調手法を提案する．

4.2 Blinky を用いたビームフォーマ設計

Blinky が分散配置されており，マイクロホンアレイとビデオカメラを使用できる状況を考える．マイクロホン m の観測信号を短時間 Fourier 変換 (STFT) 領域で $x_m[f,n]$ とすると，

$$x_m[f,n] = \sum_{k=1}^{K} a_{mk}[f] s_k[f,n], \tag{23}$$

と表される．ここで，$a_{mk}[f]$ は音源 k からマイクロホン m への伝達関数，$s_k[f,n]$ は音源信号 k，f は周波数，n は時間フレームを表す．目的音源 k を強調するビームフォーマの出力信号 y_k は，

$$y_k[f,n] = \boldsymbol{w}_k^{\mathrm{H}}[f]\boldsymbol{x}[f,n], \tag{24}$$

$$\boldsymbol{x}[f,n] = [x_1[f,n], \cdots, x_M[f,n]]^{\mathrm{T}}, \tag{25}$$

$$\boldsymbol{w}_k[f] = [w_{k1}[f], \cdots, w_{kM}[f]]^{\mathrm{T}}, \tag{26}$$

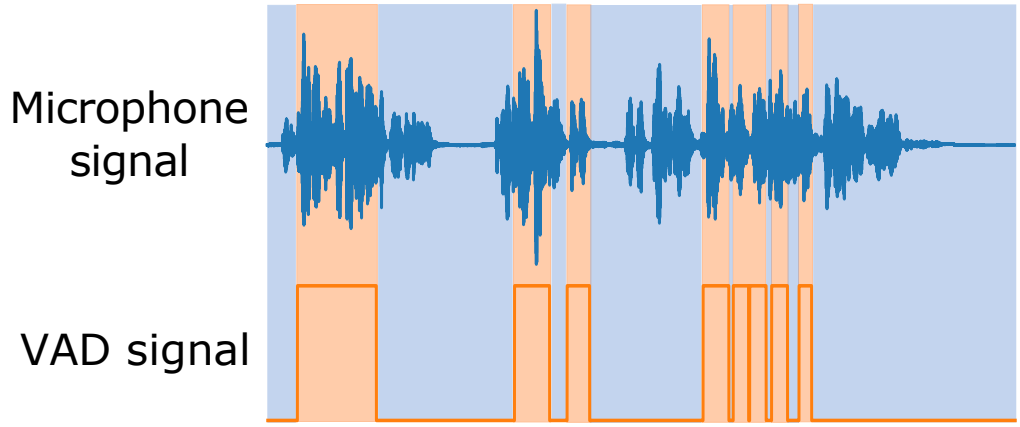

図 10　VAD 信号による有声区間検出

と表される．ただし，$\boldsymbol{w}_k$ は音源 k を強調するフィルタベクトルであり，$(\cdot)^{\mathrm{T}}$，$(\cdot)^{\mathrm{H}}$ はそれぞれ転置，エルミート転置を示している．SINR 最大化ビームフォーマのフィルタベクトルは以下のように設計できる．

$$\boldsymbol{w}_k^* = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^{\mathrm{H}} \mathbf{C}_{\mathrm{S}}^{(k)} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{H}} \mathbf{C}_{\mathrm{IN}}^{(k)} \boldsymbol{w}}. \tag{27}$$

ただし，$\mathbf{C}_{\mathrm{S}}^{(k)}$，$\mathbf{C}_{\mathrm{IN}}^{(k)}$ はそれぞれ目的音源と非目的音源の共分散行列であり，一般化固有値問題 $\mathbf{C}_{\mathrm{S}}^{(k)} \boldsymbol{w} = \lambda \mathbf{C}_{\mathrm{IN}}^{(k)} \boldsymbol{w}$ を解くことで，$\boldsymbol{w}_k^*$ を求めることができる．一般的に，$\mathbf{C}_{\mathrm{S}}^{(k)}$，$\mathbf{C}_{\mathrm{IN}}^{(k)}$ の推定には事前情報として図 10 のような目的音源の有声区間が必要であるため実装が困難であるが，Blinky の音強度分離から得られる各音源のアクティビティを用いることで推定し，以下のように共分散行列を導出できる．

$$\mathrm{VAD}_k[n] = \begin{cases} 1 & \text{if } r_{kn} \geq \rho \\ 0 & \text{otherwise} \end{cases}, \tag{28}$$

$$\mathbf{C}_{\mathrm{S}}^{(k)}[f] = \sum_{n=1}^{N} \mathrm{VAD}_k[n] \boldsymbol{x}[f,n] \boldsymbol{x}[f,n]^{\mathrm{H}}, \tag{29}$$

$$\mathbf{C}_{\mathrm{IN}}^{(k)}[f] = \sum_{n=1}^{N} (1 - \mathrm{VAD}_k[n]) \boldsymbol{x}[f,n] \boldsymbol{x}[f,n]^{\mathrm{H}}. \tag{30}$$

この VAD 信号は，図 11 のように音源 k のアクティビティ r_{kn} に対して事前に設定した閾値 ρ に従い設計する．最後に，提案する枠組みを図 12 に示す．

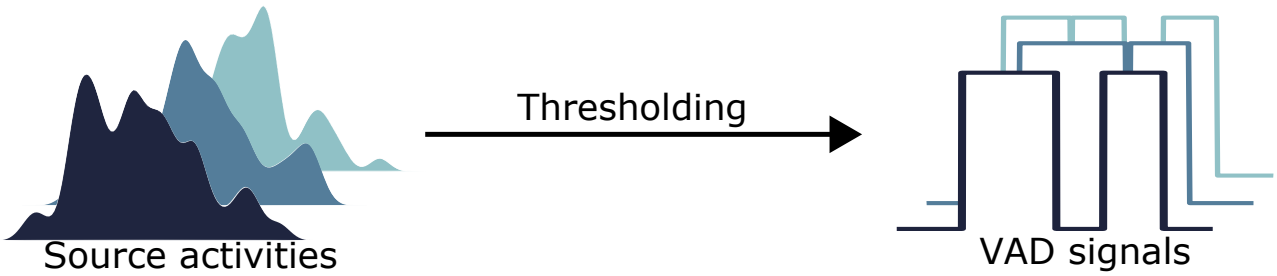

図 11　音源アクティビティから VAD 信号の設計

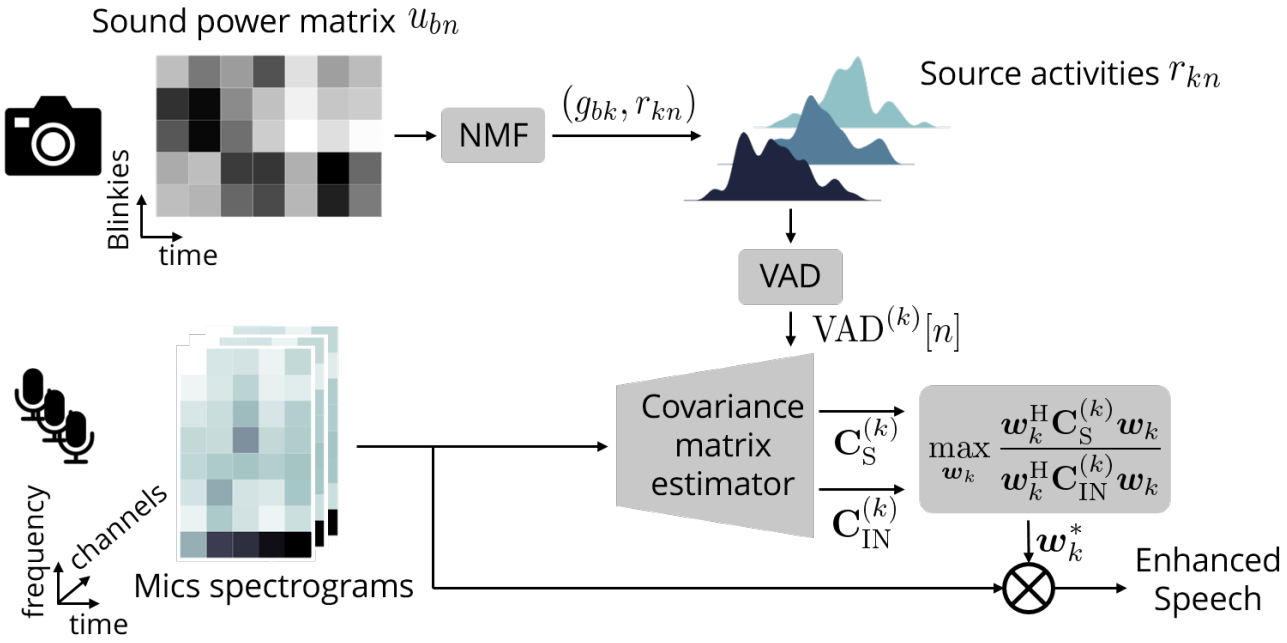

図 12　提案する SINR 最大化ビームフォーマのフローチャート

第 5 章

Blinky を用いた複数音源定位

5.1 問題設定

K 個の目的音源の位置を B $(> M)$ 個の分散配置した Blinky を用いて推定することを考える．すなわち，音源位置 $\mathbf{r}_k^{(s)} \in \mathbb{R}^3$, $k = 1, 2, \cdots, K$ を $\mathbf{r}_m^{(b)} \in \mathbb{R}^3$, $m = 1, 2, \cdots, B$ に位置する Blinky の音強度 $\hat{u}_b[n]$ から推定する．

音源定位では、エネルギーベースの定位手法が提案されており，例えば，Chen ら [17] が提案している．この手法では，センサで観測される音響パワーは音源からの距離に反比例しており，これがエネルギーベースの定位の基礎となっている．しかし，複数の音源が同時に存在している場合には，各音源が混合するため，音強度 $u_b[n]$ を取得することは困難である．このような状況では，エネルギーベースの定位手法を複数音源定位に直接適用することはできない．この問題を解決するために，各音源の音強度を分離するために NMF を用いる音源定位手法を提案する．

提案する音強度分離に基づく音源定位手法は，複数音源定位の問題を単一音源定位の問題として扱う．$u[n]$ を要素とする音強度行列は，音強度分離により伝達関数ゲインと音源アクティビティの行列それぞれに分離することができる．伝達関数ゲインは，音源から Blinky までの距離，すなわち空間内の音強度の分布を含んでいる．したがって，NMF により得られた伝達関数ゲインを用いて，エネルギーベースの音源定位アルゴリズムにより各音源位置を推定する．

5.2 Blinky を用いた DNN による複数音源定位

NMF から得られる伝達関数ゲイン g_{bk} を音源定位に利用する方法について説明する．提案するアルゴリズムは，伝達関数ゲイン g_{bk} から音源位置 $\mathbf{r}_m^{(b)}$ への写像 Φ をニューラルネットワークを用いて学習する．全結合ニューラルネットワーク (fully connected neural network; FCNN) と残差接続を用いた全結合ニューラルネットワーク (fully connected neural network with residual connection; FCNN w/ RC) の性能を比較する．損失関数は平均絶対誤差 (mean absolute error; MAE)

$$\mathrm{MAE} = \frac{1}{K} \sum_k \left\| \hat{\mathbf{r}}_k^{(s)} - \mathbf{r}_k^{(s)} \right\|_1, \tag{31}$$

を用いた．ただし，$\|\cdot\|_1$ は $\ell 1$-ノルム，$\hat{\mathbf{r}}_k^{(s)}$ は推定した音源 k である．提案する複数音源定位の流れを図 13 に示す．ニューラルネットワークモデルは伝達関数ゲインから音源位置への写像を単一音源で学習する (図 13 上)．この学習済みのモデルに，各音源の伝達関数ゲインを入力することで，複数

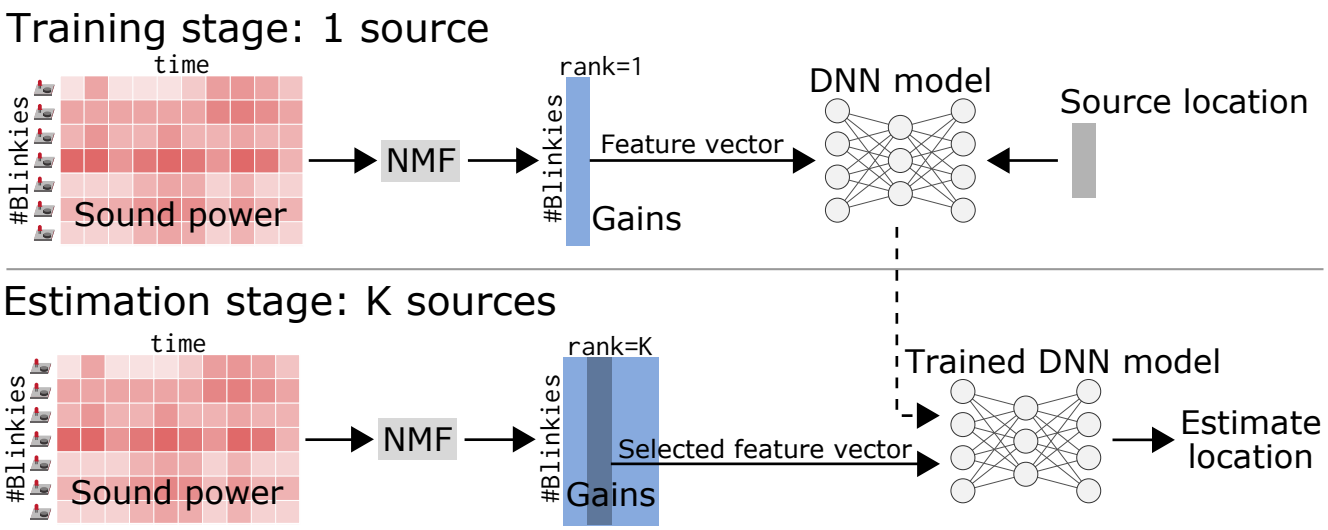

図 13　提案する複数音源定位のフローチャート.

の音源が存在する場合においても，音源位置を推定することができる (図 13 下).

第 6 章

評価実験

計算機シミュレーション及び実環境にて，提案する複数音声強調と複数音源定位それぞれの性能を評価する実験を行った．

6.1 複数音声強調: 計算機シミュレーション

従来の音声強調法では困難な拡散雑音下にて，提案手法の性能を評価する実験を行った．$15 \times 15 \times 2.6$ m^3，残響時間 0.45 s の部屋を Python パッケージの Pyroomacoustics [18] を用いてシミュレートした．ただし，シミュレーションであるため Blinky はマイクロホンで代用し，$B = 30$ 個用意した．目的音源は $K = 1, 2, 3$ 個，マイクロホンアレイのサイズは $M = 2, 4, 8$ と変化させた．また，壁際に $Q = 16$ 個の音源を配置することで拡散雑音を作成した．

参照マイクロホンに対して，signal-to-noise ratio (SNR) と SINR をそれぞれ $\mathrm{SNR} = \frac{1}{K}\sum_{k=1}^{K}\sigma_k^2/\sigma_n^2$，$\mathrm{SINR} = \sum_{k=1}^{K}\sigma_k^2/(Q\sigma_i^2 + \sigma_n^2)$ と定義し，SNR = 60 dB，SINR = 0 dB とした．ただし，σ_k^2，σ_i^2，σ_n^2 はそれぞれ目的音源，拡散雑音源，白色雑音の分散である．サンプリング周波数は 16 kHz，STFT はフレーム長を 4096 サンプル，フレームシフトをフレーム長の 1/2，分析窓に Hann 窓を用いた．ビデオカメラで観測した画素値から音強度を推定できたものと仮定し，Blinky の代わりとして用いたマイクロホンの観測信号から直接音強度信号を求めた．音声は日本音響学会新聞記事読み上げコーパス (JNAS) [19] を用い，異なる話者の組み合わせ 20 通りで混合信号を作成した．VAD 信号は事前に設定した閾値に基づいて作成した．比較手法として，効率的なブラインド音源分離である AuxIVA [20] と，AuxIVA と Blinky を組み合わせた Blink-IVA [21] を用いた．

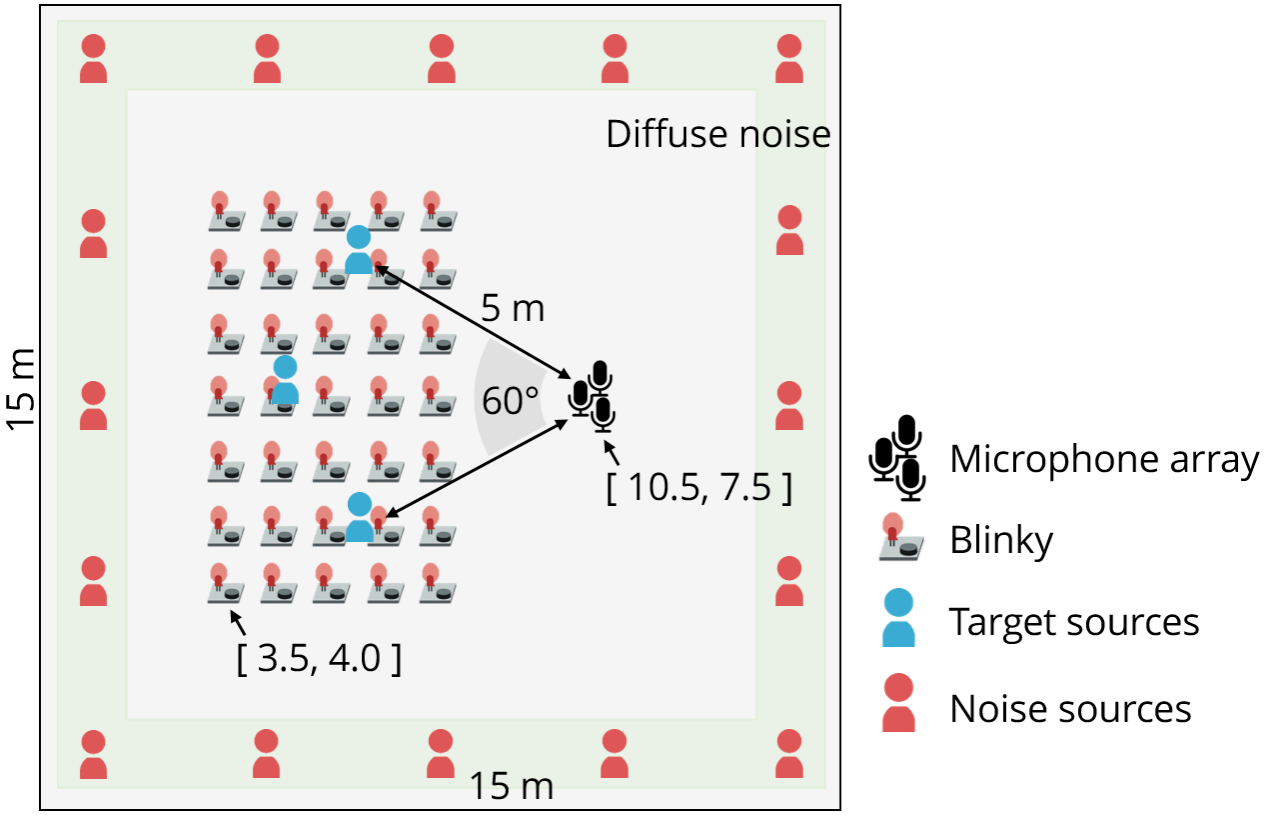

図 14　計算機シミュレーションにおける音源，マイクロホンアレイ，Blinky の位置．

図 15 に NMF のコスト関数の減衰例を示す．これにより，目的関数の誤差は最小化されており，反復回数は十分であることが確認できる．また，図 16 に目的音源が 2 つの場合における音源アクティビティと推定された VAD 信号，音源信号をそれぞれ収録し推定した正解信号となる音源アクティビティを目的音源別に示す．これにより，NMF による音源アクティビティの推定精度は VAD 信号を設計する上では十分であり，目的音源の有声区間を不足なく含んでいることが確認できる．SN 比最大化ビームフォーマを設計する上では，目的音源の有声区間を取り除くことのないように共分散行列を推定することで強調性能を高めることができるため，VAD 信号は十分に有声区間を含むように設計することが重要である．

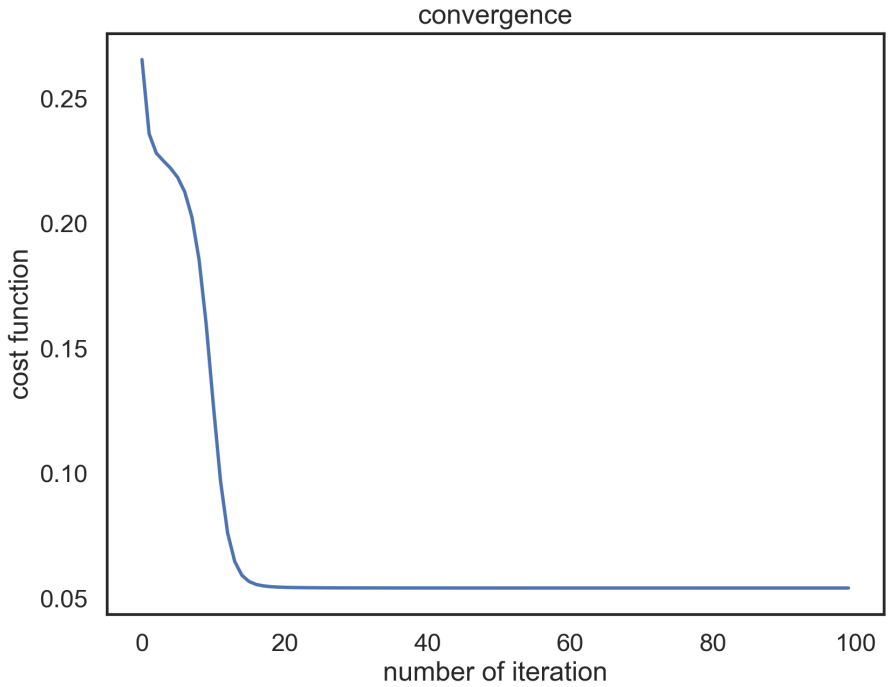

図 15　NMF のコスト関数の収束.

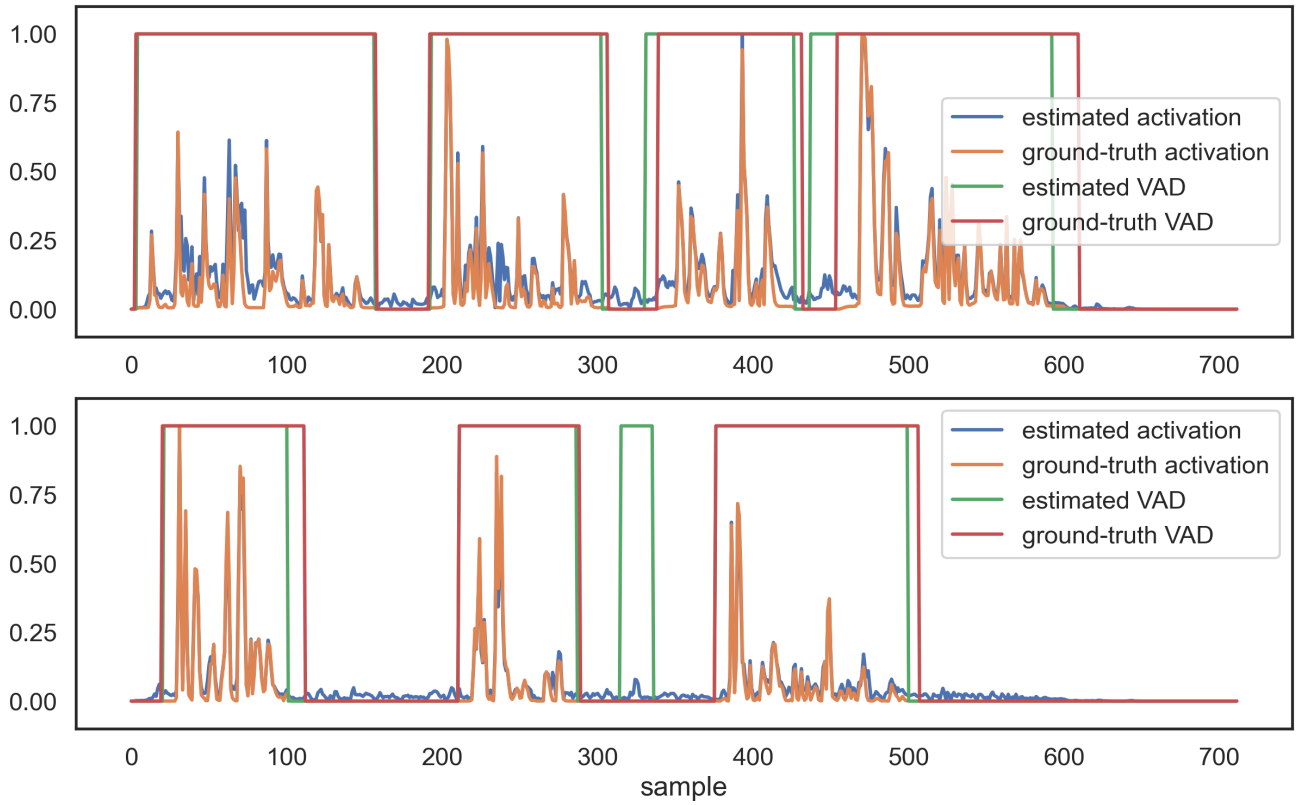

図 16　音源アクティビティと VAD 信号.

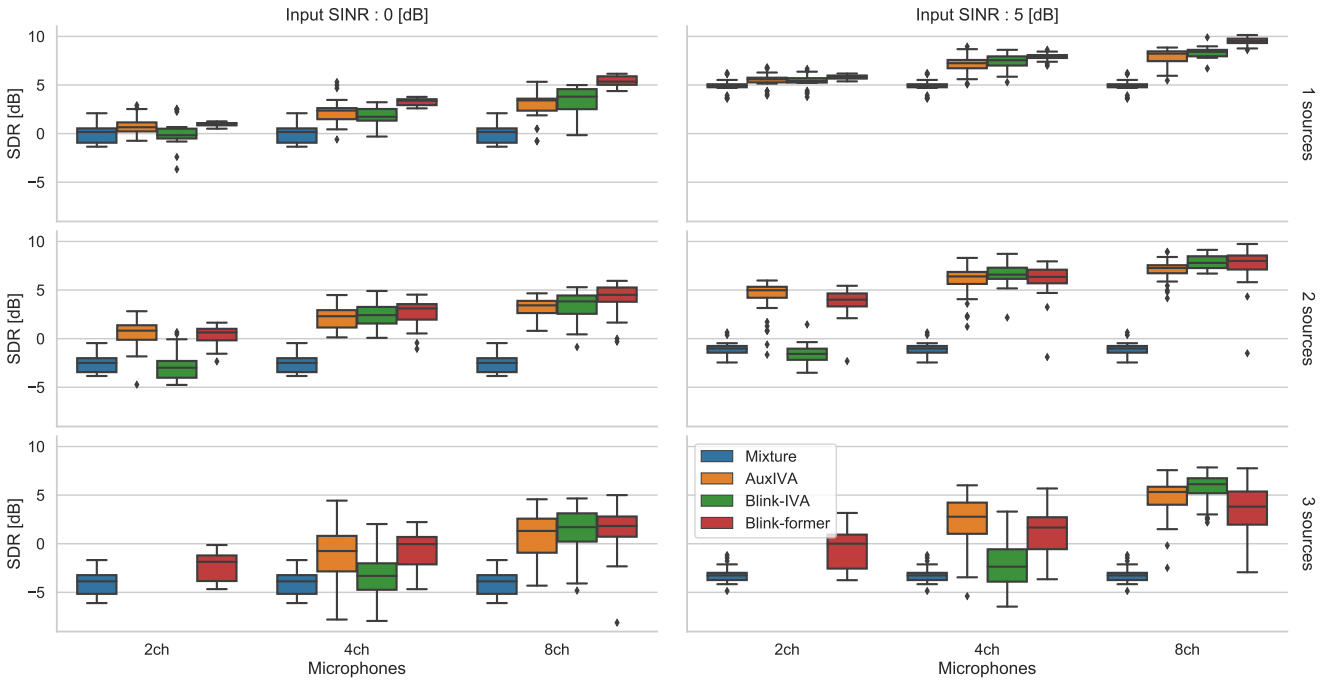

図 17　計算機シミュレーションにおける分離信号の signal-to-distortion ratio (SDR) の箱ひげ図.

分離信号の強調性能を signal-to-distortion ratio (SDR) [22] により評価する．SDR は Python パッケージの mir_eval toolbox [23] を用いて計算した．提案手法 (Blink-former)，AuxIVA，及び Blink-IVA での分離信号の SDR の分散を図 17 に示す．全体的に，提案したビームフォーマは AuxIVA と Blink-IVA それぞれと比較して同等かそれ以上の性能といえる．目的音源が 1 個の場合には，提案したビームフォーマの 25 パーセンタイルが比較手法の 75 パーセンタイルより高く，最も良い性能である．また，目的音源が 2 つ及び 3 つの場合にはその性能は同程度以上である．これにより，Blinky は十分に音源アクティビティ情報を推定できていることが確認できる．しかし，SNR が高い場合には，提案法よりも AuxIVA が良い性能を示している．これは，提案法では音源のアクティビティを推定する際に Blinky のみの情報を用い，マイクロホンの情報を用いていないことが原因の 1 つとして考えられる．そのため，マイクロホンの観測信号の SNR が良い場合には，AuxIVA や Blink-IVA がよいと考えられる．

6.2　複数音声強調: 実環境実験

実環境にて提案手法の性能を評価する実験を行った．8.9 × 7.4 × 2.6 m^3，残響時間 0.55 s の部屋に 2 つのスピーカーと 8 ch のマイクロホンアレイと 40 個の Blinky を図 18 のように配置した．Blinky の 1 つはキャリブレーション用として既知の光強度で発光させた．SONY HDR-CX535 (フレームレート 29.97 fps，フルH D 1980 × 1080) で Blinky の LED を撮影し，観測した画素値から逆特性を用いて音強度を推定した [7]．STFT はフレーム長を 4096 サンプル，フレームシフトをフレーム長の 1/2，分析窓に Hann 窓を用いて行った．図 19 に分離信号の SDR の改善量を示す． マイクロホンの観測信号から 2 ch と 8 ch を用いて分離信号を推定した．図 19SDR はそれぞれ 5.9, 8.4 dB 程度の向上しており，実環境でも十分に提案手法の性能を確認した．

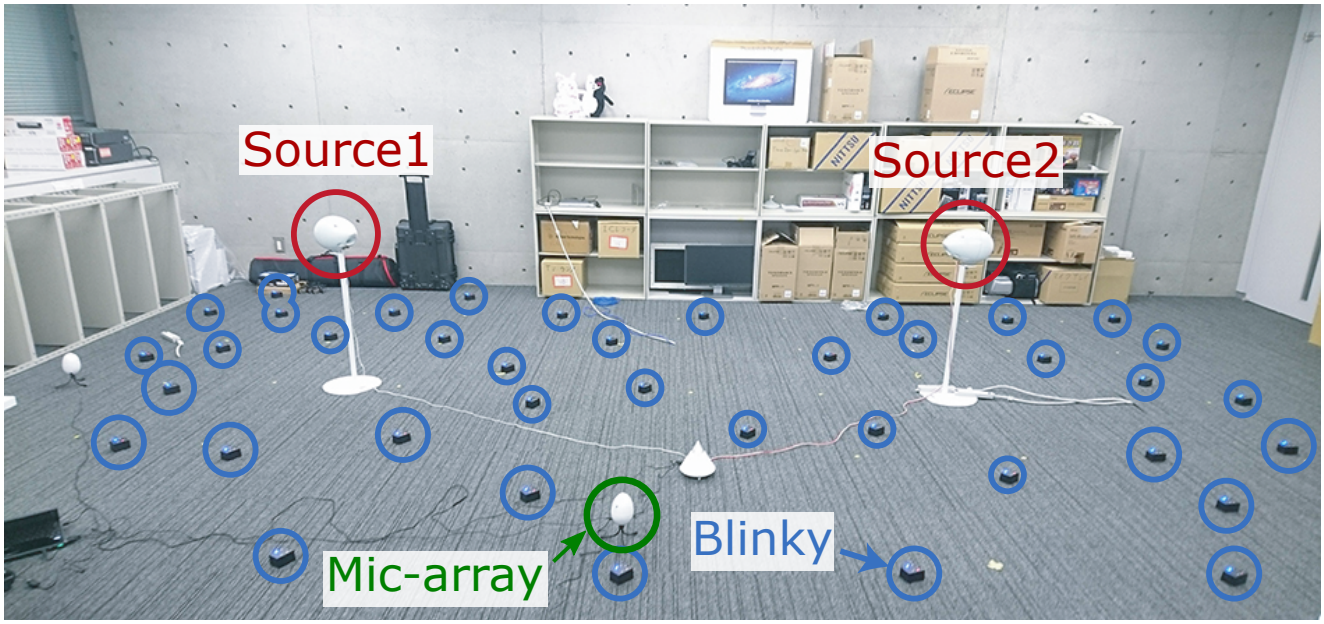

図 18　実環境実験における音源，マイクロホンアレイ，Blinky の位置．

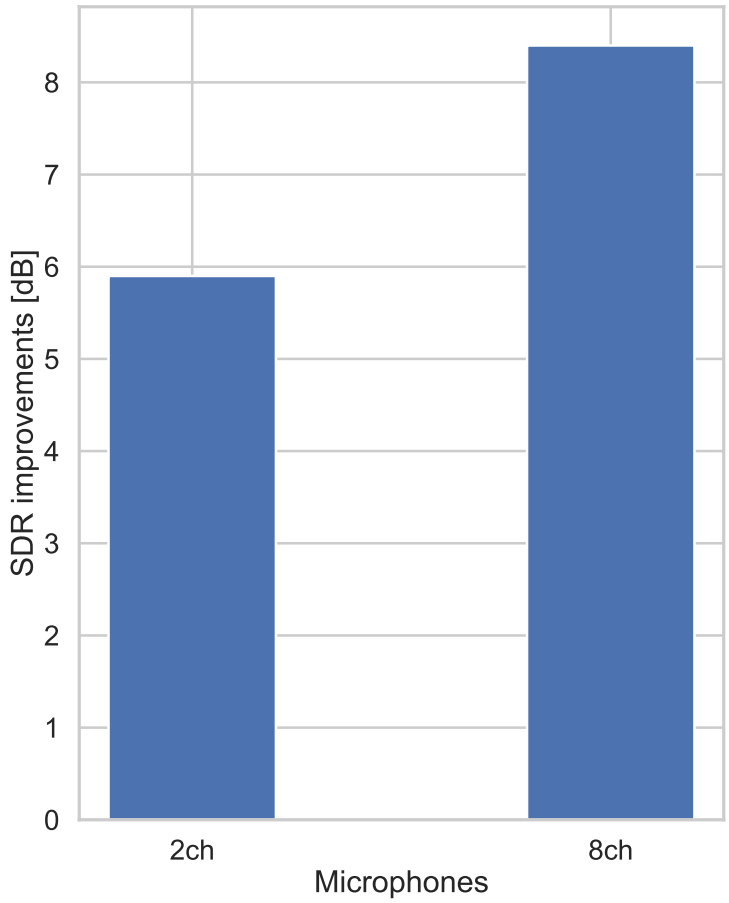

図 19　実環境実験における分離信号の signal-to-distortion ratio (SDR) の箱ひげ図.

6.3　複数音源定位: 計算機シミュレーション

提案する複数音源定位手法の推定性能を回帰問題にて推定する計算機シミュレーションを行った. $5 \times 6 \times 2.5$ m^3，残響時間 200 ms の部屋を Python パッケージの Pyroomacoustics [18] を用いて鏡像法によりシミュレートした. Blinky は $M = 20$ 個を格子状に，音源はその範囲内となるように配置した. ただし，シミュレーションであるため Blinky の観測信号はマイクロホンの観測信号の 2 乗パワーで代用した. サンプリング周波数は 16 kHz，音源は CMU Arctic database [24] から，20 s の信号を訓練とテストそれぞれ 1000，200 通りとなるように生成した. 図 20 に音源，マイクロホンアレイ，Blinky の位置を示す.

Blinky の観測信号の signal-to-noise ratio (SNR) を $\mathrm{SNR} = \frac{1}{K}\sum_{k=1}^{K} \sigma_k^2/\sigma_{\mathrm{noise}}^2$ と定義し，訓練データは $\mathrm{SNR} = 5, 10, 60$ dB，テストデータは $\mathrm{SNR} = 60$ dB となるように白色雑音を付加した. ただし，$\sigma_{\mathrm{noise}}^2$ は白色雑音の分散である. 訓練データにテストデータよりも SNR の高い白色雑音を付加することで NMF の推定誤差に対して頑健となることを期待する. テストデータに対する NMF は基底数を音源数の 2，反復更新を 100 回，コスト関数をユークリッド距離で適用した. ニューラルネットワークは， Python パッケージの Pytorch [25] を用いて学習した. 層の数，要素数，最適化

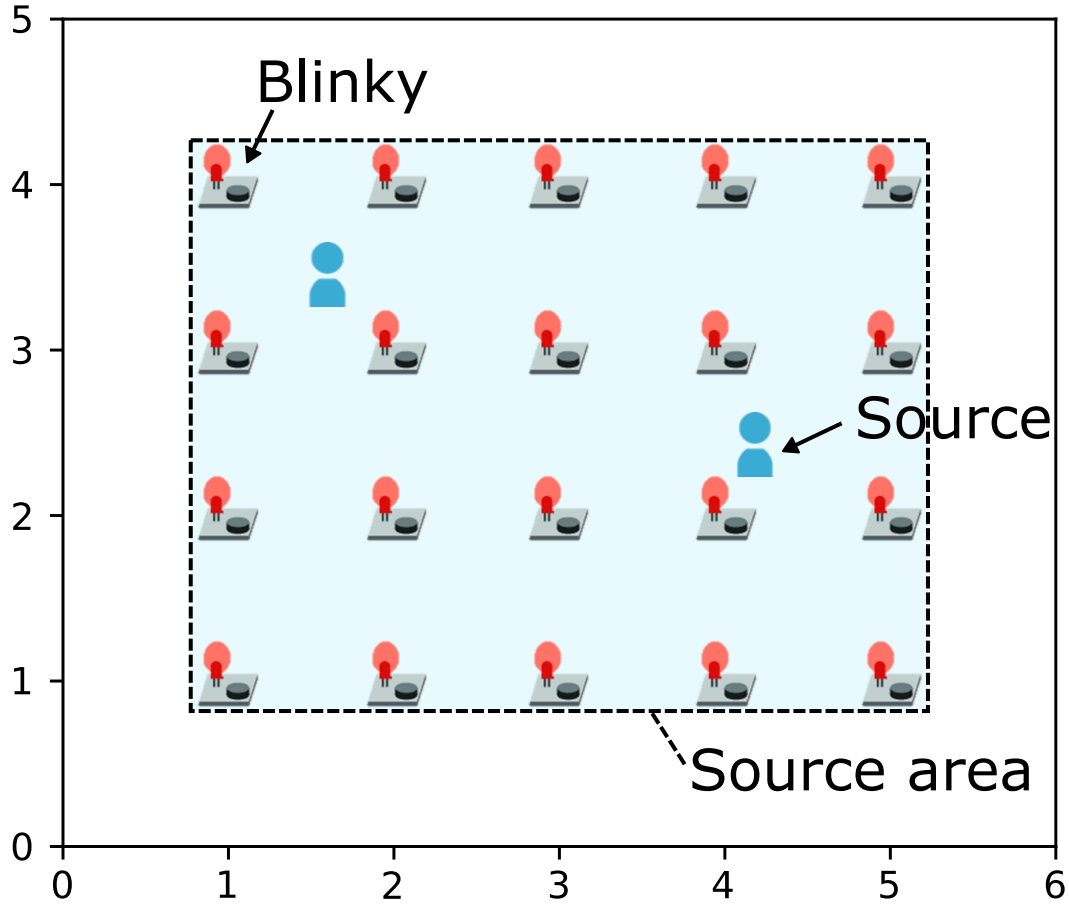

図 20　計算機シミュレーションにおける音源，Blinky の位置.

アルゴリズム [26, 27]，ドロップアウト [28] それぞれを Optuna [29] を用いて最適化した．エポック数は 1000 回，ミニバッチ数は 16 とした．層構造の詳細は表 1 に示す．また比較手法として，観測した音強度が最大となる Blinky の位置を推定位置とするアルゴリズム (Baseline) を用いた．推定誤差を平均平方二乗誤差 (root mean square error; RMSE) で評価した.

表 1　FCNN と FCNN w/ RC それぞれのネットワーク構造.

	FCNN	**FCNN w/ RC**
layers	20-15-10-5-3	20-20-{bottleneck}-20-3
bottleneck architecture	–	20-3-20
number of bottleneck	–	3
dropout	–	0.1

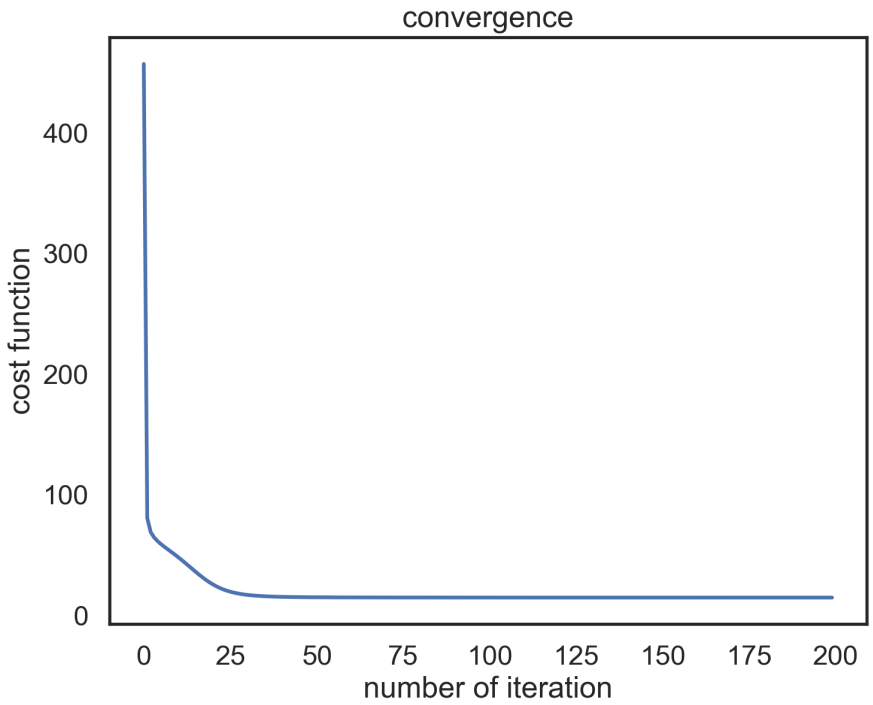

図 21　NMF のコスト関数の収束.

図 21 に NMF のコスト関数の減衰例を示す．これにより，目的関数の誤差は最小化されており，反復回数は十分であることが確認できる．また，図 22 に伝達関数ゲインと，音源信号をそれぞれ収録し推定した正解値となる伝達関数ゲインの例を各 Blinky において目的音源別に示す．これにより，NMF による伝達関数ゲインの推定精度は高いことが確認できる．しかし，音源数が増えるほどこの伝達関数ゲインの推定精度は下がってしまうことも考えられる．これを考慮するために，雑音源を加えて低い SNR で収録し推定した伝達関数ゲインを特徴量として学習させることで，この誤差に対して頑健となることを期待する．

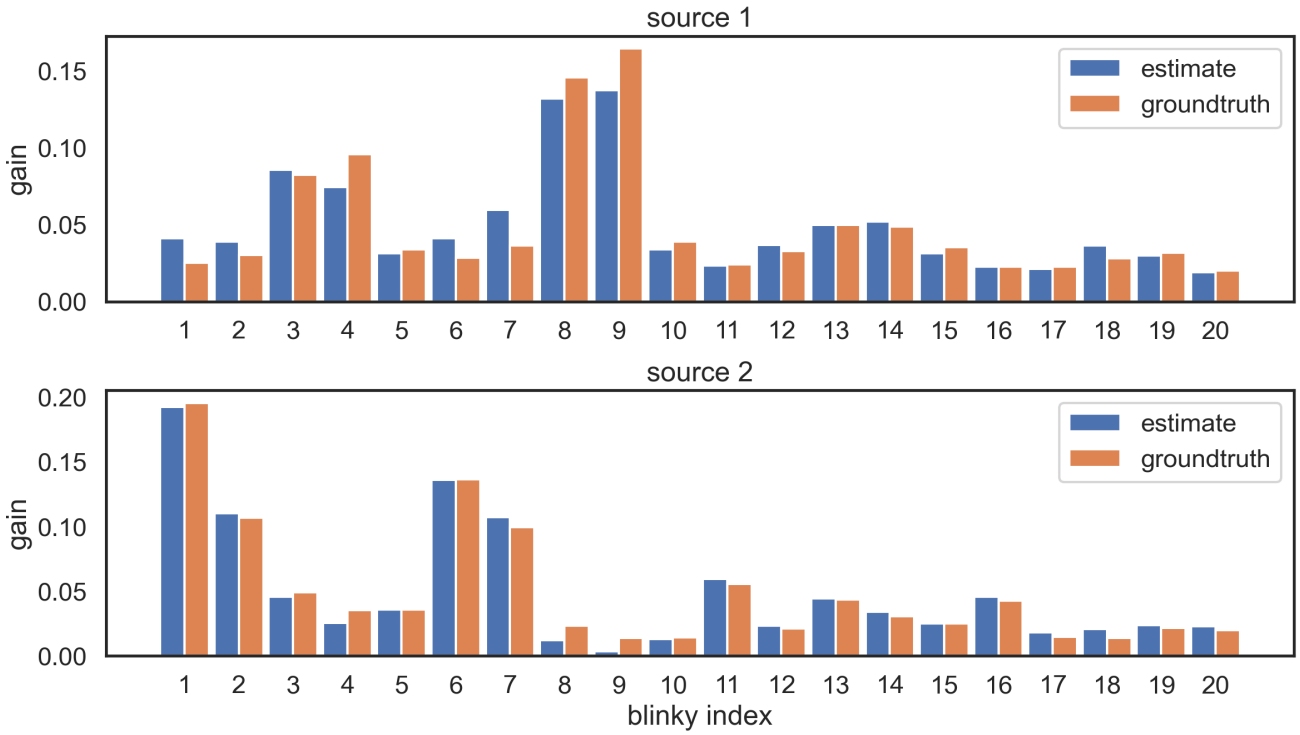

図 22　伝達関数ゲインの推定値.

図 23 は，テストデータに対する 1 音源と 2 音源に対する推定誤差の RMSE の分布を箱ひげ図で示したものである．全体的に，提案手法は全ての条件において Baseline より誤差の範囲が小さい．単一音源定位の場合には，学習時の SINR の差は性能に影響を与えなかった．しかし，2 音源の場合には，雑音を加えた学習を行うことで，誤差が大幅に減少することが確認できた．特に FCNN w/RC が最も優れた性能を示しており，2 音源定位に対する性能は単一音源定位に対する性能と同等であった．これは，提案手法が複数音源定位を単一音源定位と同様に扱うことができることを意味する．また，雑音を加え低 SINR で学習させることで性能が向上しており，NMF による伝達関数ゲインの推定誤差に対して頑健となることが確認できた．以上の結果から，Blinky を用いた複数音源定位が有効であることが確認された.

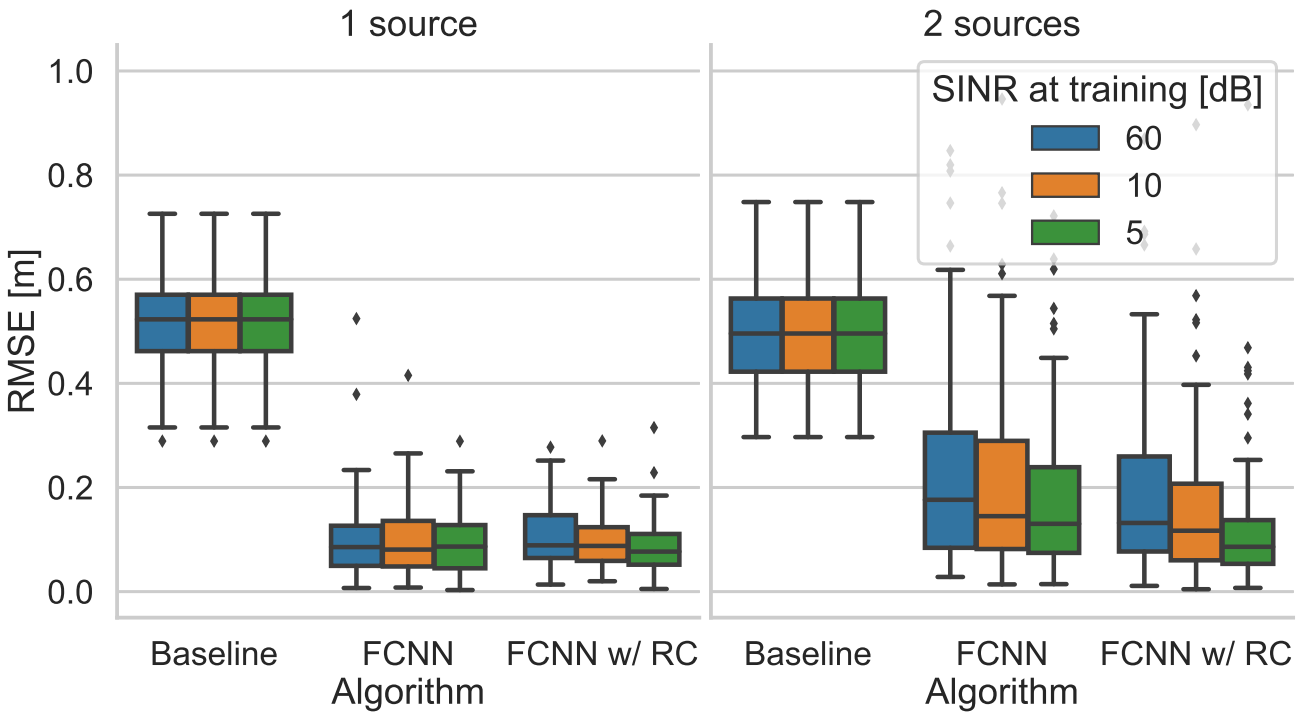

図 23　1 音源と 2 音源に対する推定した音源位置の RMSE の箱ひげ図.

第 7 章

おわりに

本論文では，複数音源が存在する状況で，多数の Blinky からビデオカメラにより音強度情報を求め，NMF による音強度分離を用いることで，複数音源に対する音声強調と音源定位を提案した．複数音源強調では，拡散雑音下での計算機シミュレーションを行い，その強調性能を評価した．結果として，SINR が低い条件では従来の音源分離手法と同程度以上の性能を確認し，提案手法の有効性を確認した．加えて実環境でも実装しその有効性も確認した．複数音源定位では，計算機シミュレーションにより，その推定性能を評価した．結果として，単一音源定位に対する性能と同等であり，その有効性を確認した．今後は複数音源定位において，実環境で実装し，その推定性能を評価する．

謝辞

本研究を行うに当たり，ご指導を頂いた小野教授に感謝致します．本研究に対して懇切な御指導，御教授を賜りました木下特任助教，若林特任助教，並びに，ロビン シャイブラー氏に感謝致します．最後になりましたが，日常，有益な議論をして頂いた研究室の皆様に感謝致します．

付録 A

補助関数法を用いた NMF の更新式の導出

以下では，補助関数法を用いた NMF の更新式の導出方法について述べる．

NMF で最小化する距離 $D_{\mathrm{cost}}(\boldsymbol{V}, \boldsymbol{WH})$ を展開し，$\boldsymbol{W}$，$\boldsymbol{H}$ に依らない項を省略したものを目的関数 $J_{\mathrm{cost}}(\boldsymbol{W}, \boldsymbol{H})$ とすると，

$$J_{\mathrm{EUC}}(\boldsymbol{W}, \boldsymbol{H}) = \sum_{m,n}(\hat{v}_{m,n}^2 - 2v_{m,n}\hat{v}_{m,n}), \tag{32}$$

$$J_{\mathrm{KL}}(\boldsymbol{W}, \boldsymbol{H}) = \sum_{m,n}(\hat{v}_{m,n} - v_{m,n}\log\hat{v}_{m,n}), \tag{33}$$

$$J_{\mathrm{IS}}(\boldsymbol{W}, \boldsymbol{H}) = \sum_{m,n}(\frac{v_{m,n}}{\hat{v}_{m,n}} + \log\hat{v}_{m,n}), \tag{34}$$

となる．しかし，この目的関数には非線形関数項が存在するため，直接最小化することは難しい．そこで代わりに以下の条件を満たす補助関数 $Q(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R})$ を導入し，これを最小化することで間接的に目的関数を最小化する．

$$J(\boldsymbol{W}, \boldsymbol{H}) = \min_{\boldsymbol{R}} Q(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}) \tag{35}$$

ここで，目的関数と異なる変数 $\boldsymbol{R}$ は補助変数と呼ばれる．式 (35) は，

- 任意の補助変数 $\boldsymbol{R}$ に対して，$J(\boldsymbol{W}, \boldsymbol{H}) \leq Q(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R})$
- 任意のパラメータ $\boldsymbol{W}$，$\boldsymbol{H}$ に対して，$J(\boldsymbol{W}, \boldsymbol{H}) = \min_{\boldsymbol{R}} Q(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R})$

を満たしている．

目的関数の非線形関数項に対して上限関数を設計し，補助関数を導入するために，以下の Jensen の不等式を用いる．任意の凸関数 f，$r_k > 0$，$\sum_{k=1}^{K} r_k = 1$ を満たすものとすると，

$$f(\sum_{k=1}^{K} r_k x_k) \leq \sum_{k=1}^{K} r_k f(x_k), \tag{36}$$

となる．ただし，$r_{m,k,n} > 0$，$\sum_{k=1}^{K} r_{m,k,n} = 1$ であり，$x_1 = x_2 = \cdots = x_K$ のとき等号は成立する．また，IS ダイバージェンスの $\log\hat{v}_{m,n}$ の項に関しては，凹関数であり Jensen の不等式では上限関数が作れないため，以下の不等式を用いる．任意の微分可能な凹関数 g に対して，

$$g(\hat{v}_{m,n}) \leq g(u_{m,n}) + (\hat{v}_{m,n} - u_{m,n})g'(u_{m,n}), \tag{37}$$

が成り立つ．ただし，$u_{m,n} > 0$ であり，$\hat{v}_{m,n} = u_{m,n}$ のとき等号は成立する．式 (36)，式 (37) より，式 (32)，式 (34) の補助関数はそれぞれ，

$$Q_{\mathrm{EUC}}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}) = \sum_{m,n}(\sum_{k} \frac{(w_{m,k}h_{k,n})^2}{r_{m,k,n}} - 2v_{m,n}\hat{v}_{m,n}), \tag{38}$$

$$Q_{\mathrm{KL}}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}) = \sum_{m,n}(\hat{v}_{m,n} - v_{m,n}\sum_{k} r_{m,k,n} \log \frac{w_{m,k}h_{k,n}}{r_{m,k,n}}), \tag{39}$$

$$Q_{\mathrm{IS}}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}, \boldsymbol{U}) = \sum_{m,n}(v_{m,n}\sum_{k} \frac{r_{m,k,n}^2}{w_{m,k}h_{k,n}} + \log u_{m,n} + \frac{\hat{v}_{m,n} - u_{m,n}}{u_{m,n}}). \tag{40}$$

となる．補助関数の最小化は以下のように，補助関数をパラメータ $\boldsymbol{W}$，$\boldsymbol{H}$ と補助変数 $\boldsymbol{R}$ についての最小化を反復することにより行われる．

$$\boldsymbol{R} \leftarrow \mathrm{argmin}_{\boldsymbol{R}} Q(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}), \tag{41}$$

$$\boldsymbol{H} \leftarrow \mathrm{argmin}_{\boldsymbol{H}} Q(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}), \tag{42}$$

$$\boldsymbol{W} \leftarrow \mathrm{argmin}_{\boldsymbol{W}} Q(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{R}, \boldsymbol{U}), \tag{43}$$

補助関数は式 (36)，式 (37)) の等号が成立するときに $\boldsymbol{R}$ に関して最小化され，その条件は，

$$r_{m,k,n} = \frac{w_{m,k}h_{k,n}}{\sum_k w_{m,k}h_{k,n}} = \frac{w_{m,k}h_{k,n}}{\hat{v}_{m,n}}, \tag{44}$$

$$u_{m,n} = w_{m,k}h_{k,n}, \tag{45}$$

となる．続いて，補助関数を $w_{m,k}$，$h_{k,n}$ により偏微分し 0 とおくことで，$\boldsymbol{W}$，$\boldsymbol{H}$ に関して最小化され，

ユークリッド距離

$$\frac{\partial Q_{\mathrm{EUC}}}{\partial w_{m,k}} = 2w_{m,k}\sum_{n} \frac{h_{k,n}^2}{r_{m,k,n}} - 2\sum_{n} v_{m,n}h_{k,n} = 0, \tag{46}$$

$$\frac{\partial Q_{\mathrm{EUC}}}{\partial h_{k,n}} = 2h_{k,n}\sum_{m} \frac{w_{m,k}^2}{r_{m,k,n}} - 2\sum_{m} v_{m,n}w_{m,k} = 0, \tag{47}$$

$$w_{m,k} = \frac{\sum_n v_{m,n}h_{k,n}}{\sum_n \frac{h_{k,n}^2}{r_{m,k,n}}}, h_{k,n} = \frac{\sum_m w_{m,k}v_{m,n}}{\sum_m \frac{w_{m,k}^2}{r_{m,k,n}}}, \tag{48}$$

KL ダイバージェンス

$$\frac{\partial Q_{\mathrm{KL}}}{\partial w_{m,k}} = \sum_{n} h_{k,n} - \sum_{n} \frac{v_{m,n}r_{m,k,n}}{w_{m,k}} = 0, \tag{49}$$

$$\frac{\partial Q_{\mathrm{KL}}}{\partial h_{k,n}} = \sum_{m} w_{m,k} - \sum_{m} \frac{v_{m,n}r_{m,k,n}}{h_{k,n}} = 0, \tag{50}$$

$$w_{m,k} = \frac{\sum_n v_{m,n} r_{m,k,n}}{\sum_n h_{k,n}}, h_{k,n} = \frac{\sum_m v_{m,n} r_{m,k,n}}{\sum_m w_{m,k}}, \tag{51}$$

IS ダイバージェンス

$$\frac{\partial Q_{\mathrm{IS}}}{\partial w_{m,k}} = 2 w_{m,k} \sum_k \frac{h_{k,n}^2}{r_{m,k,n}} - 2 \sum_k v_{m,n} h_{k,n} = 0, \tag{52}$$

$$\frac{\partial Q_{\mathrm{IS}}}{\partial h_{k,n}} = 2 h_{k,n} \sum_k \frac{w_{m,k}^2}{r_{m,k,n}} - 2 \sum_k v_{m,n} w_{m,k} = 0, \tag{53}$$

$$w_{m,k} = \frac{\sum_t v_{m,n} h_{k,n}}{\sum_t \frac{h_{k,n}^2}{r_{m,k,n}}}, h_{k,n} = \frac{\sum_m w_{m,k} v_{m,n}}{\sum_m \frac{w_{m,k}^2}{r_{m,k,n}}}, \tag{54}$$

となる．式 (44)，式 (45) を代入し整理することで，更新式を導出することができる．

参考文献

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, 1996.

[2] H. L. Van Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 2004.

[3] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *International Conference on Independent Component Analysis and Signal Separation (ICA)*. Springer, 2006, pp. 601–608.

[4] T. Kim, H. T. Attias, S.-Y. Lee, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Audio Speech, Language Process.*, vol. 15, no. 1, pp. 70–79, 2007.

[5] R. Scheibler, D. Horiike, and N. Ono, "Blinkies: Sound-to-light conversion sensors and their application to speech enhancement and sound source localization," in *Proc. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1899–1904.

[6] R. Scheibler and N. Ono, "Blinkies: Open source sound-to-light conversion sensors for large-scale acoustic sensing and applications," *IEEE Access*, vol. 8, pp. 67 603–67 616, 2020.

[7] D. Horiike, R. Scheibler, Y. Wakabayashi, and N. Ono, "Blink-former: Light-aided beamforming for multiple targets enhancement," in *Proc. 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 2019, pp. 1–6.

[8] D. Horiike, R. Scheibler, Y. Kinoshita, Y. Wakabayashi, and N. Ono, "Energy-based multiple source localization with blinkies," in *Proc. 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 443–448.

[9] W. E. Kock, *Seeing Sound*. Wiley, 1971. [Online]. Available: https://books.google.co.jp/books?id=uAlRAAAAMAAJ

[10] G. P. Nava, H. D. Nguyen, Y. Kamamoto, T. G. Sato, Y. Shiraki, N. Harada, and T. Moriya, "A high-speed camera-based approach to massive sound sensing with optical wireless acoustic sensors," *IEEE Trans. Comp. Imaging*, vol. 1, no. 2, pp. 126–139, 2015.

[11] I. Aihara, T. Mizumoto, T. Otsuka, H. Awano, K. Nagira, H. G. Okuno, and K. Aihara, "Spatio-temporal dynamics in collective frog choruses examined by mathematical modeling and field observations," *Scientific reports*, vol. 4, no. 3891, 2014.

[12] Espressif Systems, "ESP32 datasheet," 2018, [Online; accessed 2021 年 2 月 19 日]. [Online]. Available: https://www.espressif.com/sites/default/files/documentation/esp32_datasheet_

en.pdf

[13] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993. [Online]. Available: https://hdl.handle.net/11272.1/AB2/SWVENO

[14] J. Watkinson, *The MPEG handbook*. Focal Press, 2012.

[15] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[16] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with β-divergence," in *Proc. 2010 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2010, pp. 283–288.

[17] M. Chen, Z. Liu, L.-W. He, P. Chou, and Z. Zhang, "Energy-based position estimation of microphones and speakers for ad hoc microphone arrays," in *Proc. IEEE WASPAA*, 2007, pp. 22–25.

[18] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 351–355.

[19] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *JASJ*, vol. 20, no. 3, pp. 199–206, 1999.

[20] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *Proc. 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 189–192.

[21] R. Scheibler and N. Ono, "Multi-modal blind source separation with microphones and blinkies," in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 366–370.

[22] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[23] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "mir_eval: A transparent implementation of common MIR metrics," in *Proc. 15th International Conference on Music Information Retrieval (ISMIR)*. Citeseer, 2014.

[24] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Tech. Rep. CMU-LTI-03-177, 2003.

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin,

N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.

[26] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. International conference on machine learning*, 2013, pp. 1139–1147.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv.org*, Dec. 2014.

[28] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[29] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.

発表文献

[1] 堀池 大樹, シャイブラー ロビン, 若林 佑幸, 小野 順貴. 音光変換デバイス blinky と非負値行列因子分解を用いた音強度信号分離の理論と実験. 音響学会秋季研究発表会 講演論文集, pp. 145–146, September 2018.

[2] 堀池 大樹, シャイブラー ロビン, 若林 佑幸, 小野 順貴. ブリンキーと非負値行列因子分解を用いた混合音声の音強度信号分離. 音響学会春季研究発表会 講演論文集, pp. 277–278, March 2019.

[3] 堀池 大樹, シャイブラー ロビン, 若林 佑幸, 小野 順貴. マイクロホンアレイとブリンキーを用いたマルチモーダル複数音声強調. 音響学会秋季研究発表会 講演論文集, pp. 181–182, September 2019.

[4] Daiki Horiike, Robin Scheibler, Yukoh Wakabayashi, and Nobutaka Ono. Blink-former: Light-aided beamforming for multiple targets enhancement. In *In Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 443–448, December 2019.

[5] 堀池 大樹, 石井奏人, シャイブラー ロビン, 若林 佑幸, 小野 順貴. 音光変換デバイスブリンキーを用いた複数音源定位. 音響学会春季研究発表会 講演論文集, pp. 253–254, March 2020.

[6] Daiki Horiike, Robin Scheibler, Yuma Kinoshita, Yukoh Wakabayashi, and Nobutaka Ono. Energy-based multiple source localization with blinkies. In *In Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 443–448, December 2020.

著者紹介

堀池 大樹

1997 年 3 月　神奈川県生まれ
2015 年 3 月　山梨県立吉田高等学校 普通科 卒業
2015 年 4 月　首都大学東京 システムデザイン学部 情報通信システムコース 入学
2019 年 3 月　首都大学東京 システムデザイン学部 情報通信システムコース 卒業
2019 年 4 月　首都大学東京大学院 システムデザイン研究科 情報科学域 博士前期課程 入学
2021 年 3 月　東京都立大学大学院 システムデザイン研究科 情報科学域 博士前期課程 修了見込