音量比特徴量の重み付きクラスタリングによるドラムセット録音時の被り音抑圧 *

☆鈴木慶, 北村大地 (香川高専)

1 はじめに

ドラムセットの録音では，Fig. 1 に示すようにキックドラム（kick drum: KD），スネアドラム（snare drum: SD），ハイハットシンバル（hi-hat cymbal: HH）等の各音源にマイクロホンを近接させ，多チャネル信号を得る．その後，各マイクロホンの観測信号に適切な処理を施してミキシングすることで，ドラムセット全体の音を再構成する．このようなマイキングは，近接する音源（目的音源）のみの信号を得る目的がある．しかし，実際には Fig. 1 のように他の音源の音も混入してしまう．この混入音を被り音と呼び，音楽ライブ演奏や録音・ミキシングでの品質低下を招く大きな原因となる．目的の音源に近接したマイクロホンに混入する被り音を抑圧することができれば音楽ライブ演奏や録音・ミキシングにおける音質向上の他に，ドラム演奏支援 [1] への活用も期待できる．

ドラムセット全体の音を，KD や SD 等の個々の音源に分離する技術はドラム音源分離（drum source separation: DSS）と呼ばれ，深層ニューラルネットワーク（deep neural network: DNN）が用いられる [2, 3]．このアプローチでは，モノラル及びステレオの混合信号から分離音を予測するために，ドラムの各音源の学習用データセットを用いているが，更なる性能改善には学習データや DNN の大規模化を要すると報告されている [2]．

ドラムセット全体の多チャネル録音信号の被り音抑圧では，各マイクロホンにおいて目的音源の音量が被り音の音量よりも大きく観測される．これは，各目的音源に各マイクロホンを近接させているためである．そのため，多チャネル信号を入力とする DNN を学習する方法が合理的である．しかし，ドラムセットの多チャネル信号の大規模なデータセットは現状存在せず，また録音作業が煩雑なため作成も困難である．

少ない学習データで DNN の効率的に学習する工夫の一つは，補助的な情報を一緒に DNN に入力することである．例えば，教師無し手法で大まかに音源分離された信号を新しい補助的な情報として DNN の入力に用いる手法が考えられる．このような補助入力情報の推定精度は高いほど DNN の学習に効果的であるため，本稿ではまず，ドラムセットの被り音の大まかな抑圧が単純な教師無しクラスタリングでどの程度実現できるかについて調査し，今後の DNN 学習における補助入力としての有用性について検討する．

2 ドラムセットの多チャネル録音信号の小規模データセット作成

2.1 多チャネル信号のデータセット

具体的な観測信号の例として，ドラムセットの多チャネル録音信号の小規模なデータセットを作成した．本データセットは，BD，SD，HH，クラッシュシンバル（16 及び 18 インチ），ライドシンバル，及びタム（ハイ，ロー，フロア）の 9 つの音源が含まれる．マイクロホンは一般的なドラムセットの録音を模擬して Fig. 2 のように各音源に近接させて配置している．演奏したドラム楽譜の一部を，Fig. 3 に示す．本データセットでは，全音源を同時に演奏した多チャネル信号と，各音源を個別に演奏した多チャネル信号の両方を録音録音している．そのため，特定の音源の全マイクロホンへの被り音が収録されている．本稿では，このデータセット中の BD，SD，及び HH の 3 つの音源を対象とした被り音抑圧を行う．なお今後は，本データセットの公開を予定している．

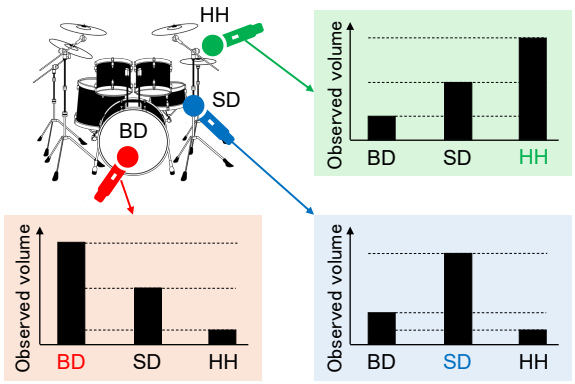


Fig. 1: Relative volume levels of microphones placed in close proximity to each sound source.

2.2 各音源の音量比

Table 1 は，KD，SD，及び HH の各音源に近接させたマイクロホンの観測信号における被り音のエネルギーを，本データセットの冒頭 10 秒間で求めた結果である．ただし，近接させている音源（目的音源）のエネルギーを 0 dB として正規化しているため，目的音と被り音の相対的なエネルギーを示している．この結果より，被り音の傾向として SD はいずれのマイクロホンに対しても大きなエネルギーとなることが分かる．特に HH の近接マイクロホンに対しては，HH よりも SD の方が大きなエネルギーで観測されている．この原因として，元々SD と HH が空間的に近い配置となっている点と，SD 自体が音源として大きなエネルギーを生じる特性を持つ点が挙げられる．従って，1 章で述べた「被り音の音量は目的音源の音量よりも小さい」という仮定は，HH の近接マイクロホンには成立しない．それでも，各音源の音量の違いは顕著であり，次章で述べる教師無し被り音抑圧の手がかりとして利用できると思われる．

* Bleeding-sound reduction for drums recording using weighted clustering of volume-ratio features. By Kei SUZUKI and Daichi KITAMURA (NIT, Kagawa).

(a) BD

(b) SD

(c) HH

Fig. 2: Example of microphone arrangements for (a) BD, (b) SD, and (c) HH.

3　教師無し手法によるドラム被り音抑圧

3.1　音量比を用いた特徴量空間への変換

Table 1 のような音源間の音量比を用いた教師無しドラム被り音抑圧について検討する．今，観測信号の振幅スペクトログラムの周波数ビン f 及び時間フレーム t の全マイクロホンの成分を $\boldsymbol{a}_{f,t} \in \mathbb{R}^M_{\geq 0}$ と表す．ここで，$f \in \{1, 2, \cdots, F\}$ 及び $t \in \{1, 2, \cdots, T\}$ はそれぞれ周波数ビン及び時間フレームのインデクスである．この $\boldsymbol{a}_{f,t}$ を次式のように極座標で表現する．

$$r_{f,t} = ||\boldsymbol{a}_{f,t}||_2 \tag{1}$$

$$\theta_{f,t,m} = \arctan\left(\frac{a_{f,t,m'+1}}{\sqrt{\sum_{m=1}^{m'} a_{f,t,m}^2}}\right) \tag{2}$$

ここで，$\|\cdot\|_2$ は L_2 ノルム，$m' \in \{1, \cdots, M-1\}$ は（m とは別の）マイクロホンのインデクスである．$r_{f,t}$ は $\boldsymbol{a}_{f,t}$ の長さ，$\theta_{f,t,m'}$ は $m = 1$ 番目から $m = m' + 1$ 番目までのマイクロホンの音量比に対応する角度である．この音量比特徴量は $\boldsymbol{\theta}_{f,t} = [\theta_{f,t,1}, \cdots, \theta_{f,t,M-1}]^{\mathrm{T}} \in [0, \pi/2]^{M-1}$ としてベクトルで表す．ここで，$\cdot^{\mathrm{T}}$ は転置を表す．式 (1) 及び (2) によって，M 次元非負ベクトル $\boldsymbol{a}_{f,t}$ は $(r_{f,t}, \boldsymbol{\theta}_{f,t})$ という特徴量に変換される．

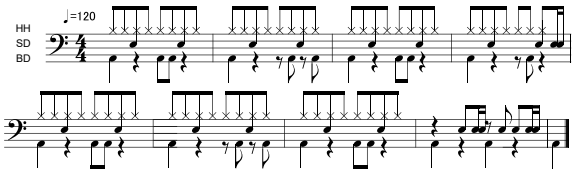


Fig. 3: Example of drums score used in recording.

Table 1: Relative energy [dB] of bleeding sounds observed by the close microphones

	KD source	SD source	HH source
KD microphone	0	-13.41	-42.43
SD microphone	-25.98	0	-23.41
HH microphone	-6.62	9.24	0

提案手法では，$\boldsymbol{\theta}_{f,t}$ を音量比特徴量空間上のデータ，観測振幅 $r_{f,t}$ を各データの重みと解釈する．KD, SD, 及び HH のマイクロホンを用いた $M = 3$ の例を Fig. 4 に示す．これは $\boldsymbol{\theta}_{f,t}$ を $M-1$ 次元データとし $r_{f,t}$ で重みづけして積み上げたヒストグラムであり，Table 1 のように音量比が明確な場合は，Fig. 4 のように各音源がクラスタを構成する．従って，これらのクラスタを教師無しクラスタリングで分離できれば，被り音の抑圧が可能である．

3.2　重み付きクラスタリング

3.2.1　重み付き k 平均法

各時間周波数で定義される音量比特徴量 $\boldsymbol{\theta}_{f,t}$ に対して，$r_{f,t}$ の重みを考慮したクラスタリングを考える．最も単純な方法は重み付き k 平均法である．いま，音源数はマイクロホン数 M と等しいと仮定する．各クラスタが 1 つの音源に対応すると定義し，K（$= M$）個のクラスタを用意する．クラスタ $k = 1, 2, \cdots, K$ に現在属している音量比特徴量（データ $\boldsymbol{\theta}_{f,t}$）の時間及び周波数インデクスの組の集合を C_k とおくとき，クラスタ k の重心 $\boldsymbol{\mu}_k \in [0, \pi/2]^{M-1}$ を次式で計算する．

$$\boldsymbol{\mu}_k = \frac{\sum_{(f,t)\in C_k} r_{f,t}\boldsymbol{\theta}_{f,t}}{\sum_{(f,t)\in C_k} r_{f,t}} \tag{3}$$

重み $r_{f,t}$ によって，振幅の大きい時間周波数の $\boldsymbol{\theta}_{f,t}$ ほど，$\boldsymbol{\mu}_k$ の計算に強く影響する．次に，式 (3) で更新された $\boldsymbol{\mu}_k$ $\forall k$ を用いて各クラスタの音量比特徴量のインデクス集合 C_k $\forall k$ を更新する．$\boldsymbol{\mu}_k$ と C_k の更新を収束するまで反復することで，クラスタリングの結果が得られる．

3.3 節で述べる被り音抑圧処理のために，クラスタリング結果を用いて各クラスタの分散共分散行列 $\boldsymbol{\Sigma}_k$ 及び寄与率 π_k を次式で求める．

$$\boldsymbol{\Sigma}_k = \frac{\sum_{(f,t)\in C_k} r_{f,t}(\boldsymbol{\theta}_{f,t} - \boldsymbol{\mu}_k)(\boldsymbol{\theta}_{f,t} - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_{(f,t)\in C_k} r_{f,t}} \tag{4}$$

$$\pi_k = \frac{\sum_{(f,t)\in C_k} r_{f,t}}{\sum_f \sum_t r_{f,t}} \tag{5}$$

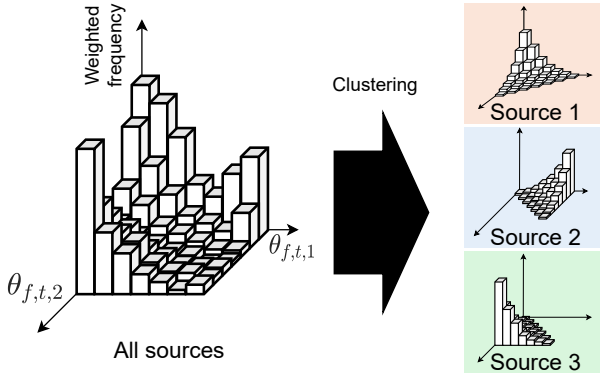


Fig. 4: Weighted histogram of volume-ratio features when $M = 3$.

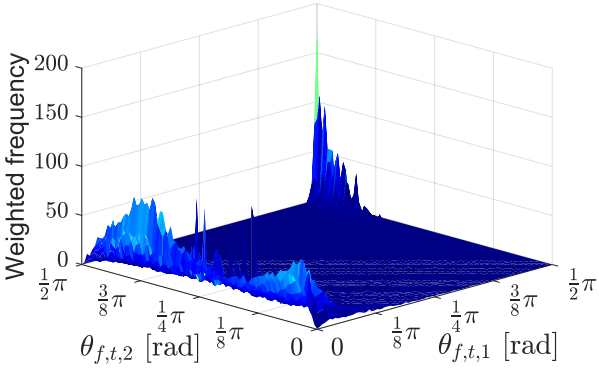


Fig. 5: Weighted histogram of the observed multi-channel signal.

3.2.2 重み付き混合ガウス分布モデル

k 平均法よりも柔軟なクラスタリングとして，混合ガウス分布モデル（Gaussian mixture model: GMM）が挙げられる．各データに重みを考慮した expectation-maximization（EM）アルゴリズムを用いて GMM を推定し，クラスタリングを行う．重み付きデータに対する EM アルゴリズムは，各パラメータを初期化したうえで，次に示す期待値ステップ（E-step）と最大化ステップ（M-step）を収束するまで反復する．

E-step 各データ点 $\boldsymbol{\theta}_{f,t}$ が k 番目のガウス分布から生成された事後確率（負担率）を，次式で計算する．

$$\gamma_{f,t,k} = r_{f,t} \frac{\pi_k \mathcal{N}(\boldsymbol{\theta}_{f,t}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\boldsymbol{\theta}_{f,t}|\boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})} \tag{6}$$

ここで，$\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ は平均 $\boldsymbol{\mu}$ 及び分散共分散 $\boldsymbol{\Sigma}$ の多変量ガウス分布の確率密度関数を表す．

M-step E-step で求めた $\gamma_{f,t,k}$ を用いて，平均 $\boldsymbol{\mu}_k$，分散共分散行列 $\boldsymbol{\Sigma}_k$，及び混合比 π_k を次式で計算する．

$$\boldsymbol{\mu}_k = \frac{\sum_f \sum_t \gamma_{f,t,k} \boldsymbol{\theta}_{f,t}}{\sum_f \sum_t \gamma_{f,t,k}} \tag{7}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_f \sum_t \gamma_{f,t,k} (\boldsymbol{\theta}_{f,t} - \boldsymbol{\mu}_k)(\boldsymbol{\theta}_{f,t} - \boldsymbol{\mu}_k)^{\mathrm{T}}}{\sum_f \sum_t \gamma_{f,t,k}} \tag{8}$$

$$\pi_k = \frac{\sum_f \sum_t \gamma_{f,t,k}}{\sum_{k'} \sum_f \sum_t \gamma_{f,t,k'}} \tag{9}$$

前項の重み付き k 平均法と同様に重み $r_{f,t}$ が考慮されているため，振幅の大きい時間周波数成分を重視した GMM が推定される．

3.3 マスク生成と被り音抑圧

提案手法では，クラスタリングで得られた $\boldsymbol{\mu}_k$ 及び $\boldsymbol{\Sigma}_k$ から $M-1$ 次元ガウス分布を生成し，被り音抑圧の時間周波数マスクに変換する．まず，各クラスタのガウス分布を $\mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_k, \kappa\boldsymbol{\Sigma}_k)$ と定義する．ここで，$\kappa \geq 1$ は分散共分散行列を拡大する係数であり，大きい値に設定すれば被り音抑圧の時間周波数マスクを緩和できる．次に，各データ $\boldsymbol{\theta}_{f,t}$ に対する事後確率を得る．

$$m_{f,t,k} = \mathcal{N}(\boldsymbol{\theta} = \boldsymbol{\theta}_{f,t}|\boldsymbol{\mu}_k, \kappa\boldsymbol{\Sigma}_k) \tag{10}$$

これによって，クラスタ k（音源 k）に対する時間周波数マスク $\boldsymbol{M}_k \in \mathbb{R}_{\geq 0}^{F \times T}$ が得られる．ここで，$\boldsymbol{M}_k$ は $m_{f,t,k}$ を要素にもつ行列である．この時間周波数マスク $\boldsymbol{M}_k$ を正規化し $\overline{\boldsymbol{M}}_k \in [0,1]^{F \times T}$ に変換した後，目的音源の近接マイクロホンの観測信号の複素スペクトログラムと $\overline{\boldsymbol{M}}_k$ の要素積をとることで被り音が抑圧される．

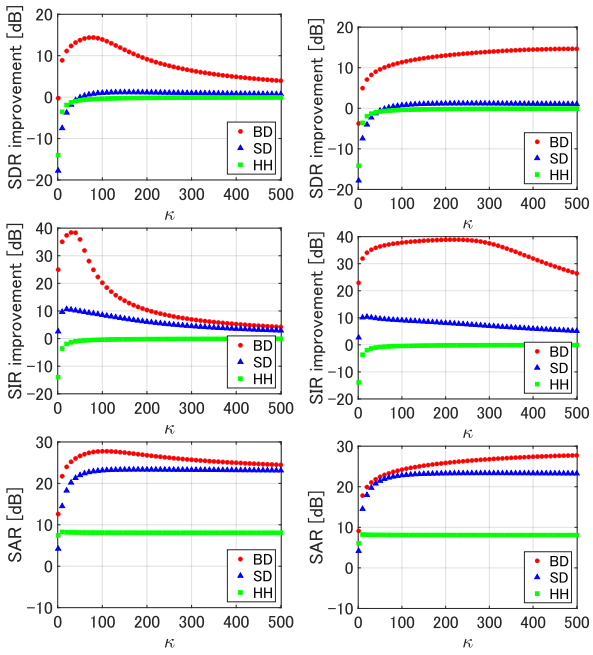
4 実験条件

4.1 実験方法

Fig. 3 の楽譜通りに演奏した BD，SD，及び HH に対して，これらの音源に近接させた 3 個のマイクロホンの観測信号を用意し，前節の教師無しクラスタリングによる被り音抑圧を適用した．ただし，近接マイクロホンは BD，SD，HH の順番で $m = 1, 2, 3$ と定義し，音源毎に演奏した際の多チャネル観測信号をマイクロホン毎に足し合わせることで，混合信号を模擬した．短時間フーリエ変換の窓長は 21.3 ms，シフト長は 10.7 ms，窓関数はブラックマン窓とした．k 平均法と GMM のいずれにおいても，異なる乱数初期値を用いて 20 回実施した結果の中から，被り音抑圧効果が最良のモデルについて報告する．時間周波数マスクを緩和する係数 κ は $[1, 500]$ を等間隔に 50 分割して探索し，分離性能を比較した．評価尺度には，信号対歪み比（source-to-distortion ratio: SDR）及び信号対干渉音比（source-to-interference ratio: SIR）の改善量，及び信号群対歪み比（sources-to-artifact ratio: SAR）を用いた [4]．本実験では，SIR は被り音の抑圧度合い，SAR は抑圧処理で生じる人工歪みの少なさ，SDR は SIR と SAR の両方を加味した被り音抑圧の総合的な品質を示す．

4.2 実験結果

Fig. 5 は，本実験で用いた観測信号に対して，3.1 節の手法で重み付きヒストグラムを求めた結果である．明確なクラスタが観測されており，クラスタリングによる被り音抑圧がある程度可能であることが分かる．しかしながら，一部のクラスタ（具体的には SD と HH）にはある程度重なりが生じていることもわかる．単純なクラスタリングだけでこれらを分離することは困難だが，1 章で述べたように，クラスタリングによる大まかな分離を補助的な入力とする DNN であればこのような信号の被り音抑圧も期待できる．



(a) k-means (b) GMM

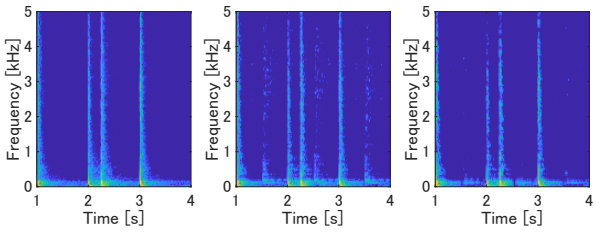
Fig. 6: Performance of bleeding-sound reduction when $\kappa = 100$ for (a) k-means and (b) GMM.

Fig. 6 は各クラスタリング（k-means 及び GMM）に対して κ を変化させたときの被り音抑圧の各評価値である．k-means 及び GMM のいずれの結果も，被り音抑圧の性能は音源毎に大きく異なっていることが確認できる．特に SDR 改善量及び SIR 改善量に関して，BD は突出して性能が高く，SD 及び HH は BD と比べて評価値は低く同程度の評価が得られている．これは，Fig. 5 で示した特徴量空間において，「BD のクラスタは明確に孤立していること」及び「SD と HH のクラスタはやや重なりが生じていること」の 2 点が原因と考えられる．κ が小さい場合，時間周波数マスクが被り音の存在する時間周波数スロットを過剰に抑圧し，逆に κ が大きい場合，時間周波数マスクが被り音を抑圧しきれないことがわかる．GMM は k-means と比べて広範囲の κ で高い性能値を示している．これは，k-means より GMM の方がクラスタ構造を適切に捉えているためと考えられる．

Figs. 7–9 に，Fig. 3 の演奏音の 1–4 s におけるリファレンス信号及び各クラスタリングを用いて被り音抑圧を行った際の推定信号のパワースペクトログラムを音源毎に示す．また，これらの推定信号は $\kappa = 100$ の条件での結果の例である．k-means 及び GMM 共に，KD についてはリファレンス信号に近いものが推定されていることがわかるが，SD 及び HH については依然として抑圧しきれていない被り音が残留していることが確認できる．しかし，DNN への補助入力を見据えた大まかな被り音抑圧として考えると，DNN の効果的な学習に有用であることが期待される．

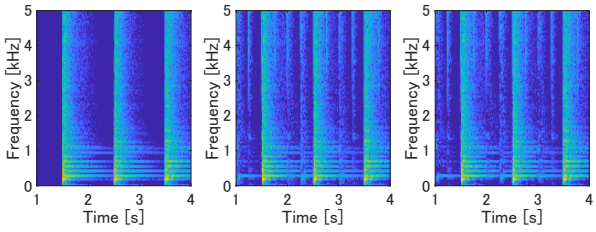
5 おわりに

本稿では，音量比特徴量を用いた教師無し重み付きクラスタリングによる教師無しのドラム被り音抑圧手法を提案した．実験の結果，KD では高い抑圧性能を示したが，HH に近接させたマイクロホンにおける SD の被り音のように目的音源よりも大きな音量を持つ被り音の抑圧には課題が残った．しかしながら，本実験で得られた大まかな分離結果は，データセットが不足しがちな DSS タスクにおける DNN の学習において，有効な補助入力となることが期待できる．



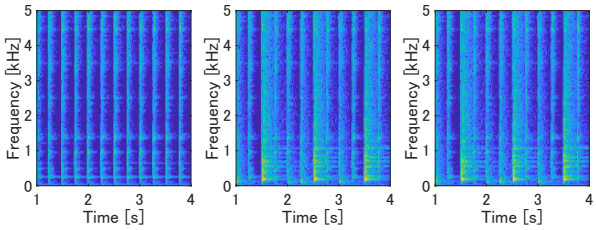
(a) Reference (b) k-means (c) GMM

Fig. 7: Spectrograms of KD source: (a) reference signal, (b) estimated signal obtained by k-means, and (c) estimated signal obtained by GMM.



(a) Reference (b) k-means (c) GMM

Fig. 8: Spectrograms of SD source: (a) reference signal, (b) estimated signal obtained by k-means, and (c) estimated signal obtained by GMM.



(a) Reference (b) k-means (c) GMM

Fig. 9: Spectrograms of HH source: (a) reference signal, (b) estimated signal obtained by k-means, and (c) estimated signal obtained by GMM.

謝辞 本研究の一部は科研費 23K24908 の助成を受けた．

参考文献

[1] M. Hosoya, M. Morise, S. Nakamura, and K. Yoshii, “A real-time drum-wise volume visualization system for learning volume-balanced drum performance,” in *Proc. IFIP Int. Conf. Entertain. Comput.*, pp. 154–166, 2021.

[2] A. I. Mezza, R. Giampiccolo, A. Bernardini, and A. Sarti, “Toward deep drum dource deparation,” *Pattern Recogn. Lett.*, vol. 183, pp. 86–91, 2024.

[3] A. I. Mezza, R. Giampiccolo, A. Bernardini, and A. Sarti, “Benchmarking music demixing models for deep drum source separation,” in *Proc. IEEE Int. Symp. Internet Sounds*, 2024.

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