左右音量比特徴量を援用した Conv-TasNet によるステレオ音楽分離 *

☆加藤 大輝, 北村 大地 (香川高専), 矢田部 浩平 (農工大)

1 はじめに

音楽信号は，ボーカルや楽器音など複数の音源が時間周波数上で重なり合った複雑な信号である．現在，流通している音楽信号の多くはステレオ信号として流通しており，左右チャネルを用いることで音源の空間配置が表現されている．この各音源の空間配置に関する情報は，音源分離における重要な手がかりとなり得るため，観測ステレオ音楽信号から各音源信号を推定するステレオ音楽分離技術が研究されてきた（例えば [1] 等）．この技術は音楽解析，自動採譜，音楽制作支援など幅広い応用が期待される．

近年，深層ニューラルネットワーク（deep neural network: DNN）を用いた音源分離手法が大きな進展を遂げている．特に Fig. 1 (a) に示すような，波形を直接入出力とする end-to-end モデルは高い分離性能を示している [2]．しかし，高精度な分離を実現するためには大規模な学習データと多大な計算資源を必要とするという課題があり，芸術性を損なわないレベルの音楽分離を安定して実現することは依然として容易ではない．このような背景から，DNN の学習や予測をサポートする目的で，ステレオ音楽信号に含まれる空間的特徴量を補助情報として活用する手法が登場している．例えば文献 [3] では，各音源の方位角を既知として補助情報に用いる DNN が提案されている．しかし，実際の音楽信号にはリバーブ等の複雑なステレオエフェクトが施されており，方位角を定義することさえもまた困難である．

本稿では，一般的なステレオ音楽信号のみから生成可能な空間特徴量を DNN に補助的に入力するステレオ音楽分離を提案する．具体的には， Fig. 1 (b) に示すように，左右チャネル間の音量比に基づく分離信号（方位分離信号）を DNN に基づくステレオ音楽分離に援用する．これにより，DNN が学習すべき内部表現の複雑性を低減し，分離精度が向上することを期待している．

2 提案手法

2.1 提案手法の概要

本稿では，一般的な音楽コンテンツがステレオ形式で流通している点に着目し，左右チャネル間の音量差から得られる粗い空間的特徴量を事前に抽出することで，DNN の学習の負荷を軽減するステレオ音楽分離手法を提案する．左右音量比に基づいて生成される方位分離信号は，ステレオ信号における各方位において支配的な音源成分を強調した中間表現として機能する．従って本手法は，ステレオ音楽信号中の音源と方位の対応関係を，解釈性の高い粗い特徴量として明示的に DNN に与えるアプローチである．

提案手法の全体構成を Fig. 1 (b) に示す．本手法

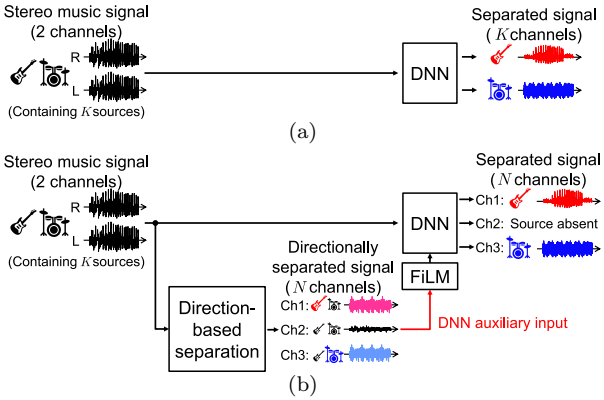

Fig. 1: Comparison of DNN-based music source separation models: (a) conventional end-to-end model and (b) proposed framework.

の特徴の一つは，推定信号のチャネル数が音源数 K ではなく，方位分離信号のチャネル数 N と一致する点である．ステレオ音楽信号中に複数音源が混在していても，各方位にはその方向で支配的な音源成分が割り当てられ，同一音源が複数チャネルに重複や分割して出力されることを抑制できる．従って，出力チャネル数が音源数ではなく方位数 N に依存するため，設計上は音源数を事前に仮定する必要なく適用可能な構成となっている．また本手法では，方位分離信号を単に DNN の入力次元に追加するのではなく，feature-wise linear modulation（FiLM）[4] を用いてネットワーク内部の特徴抽出過程に反映させる．これにより，分離された各音源の出力チャネルが方位分離信号のチャネルと整合することを狙う．以上のように，本手法では，方位に基づくチャネル依存構造と FiLM による条件付けを組み合わせることで，DNN が学習すべき音源と方位の対応関係を構造的に制約でき，限られた学習データでも，高精度なステレオ音楽分離が期待できる．

2.2 左右音量比に基づく方位分離信号

ステレオ音楽信号では，各音源が左右チャネル間の音量差により特定の方向に定位するようミキシングされる傾向にある．この左右音量差は，人間の音像定位における主要な手がかりである音圧レベル差に対応しており，ステレオ音楽信号に内在する重要な空間的特徴量である．

最も単純な方位感の付与方法は，Fig. 2 に示すように各音源に左右チャネルの音量比を与えることである．この操作はパンニングと呼ばれ，音量の強いチャネルの方位に音像が定位する．このような左右音量比の音源毎の違いが，ステレオ音楽信号における重要な空間的特徴量となる．

提案手法では，観測信号から作成する DNN 用の補助情報として，左右音量比の違いに基づく N チャネ

*Stereo music source separation using Conv-TasNet with inter-channel level difference features. By Taiki KATO, Daichi KITAMURA (NIT, Kagawa), and Kohei YATABE (Tokyo University of Agriculture and Technology).

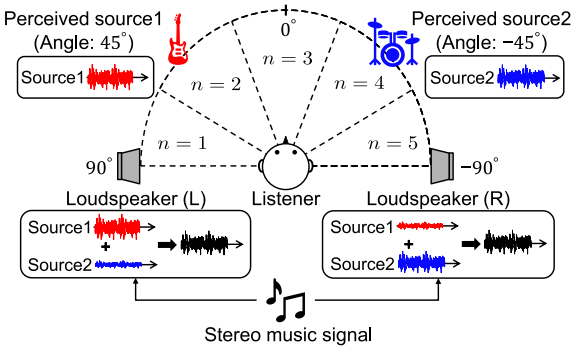

Fig. 2: Sourcewise panning based on inter-channel level difference.

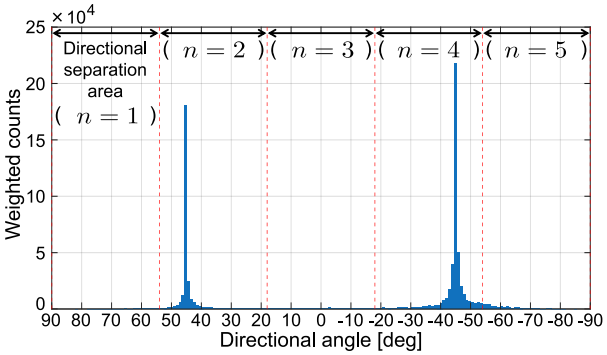

Fig. 3: Amplitude-weighted directional histogram of the stereo mixture depicted in Fig. 2, where $K = 2$ and $N = 5$.

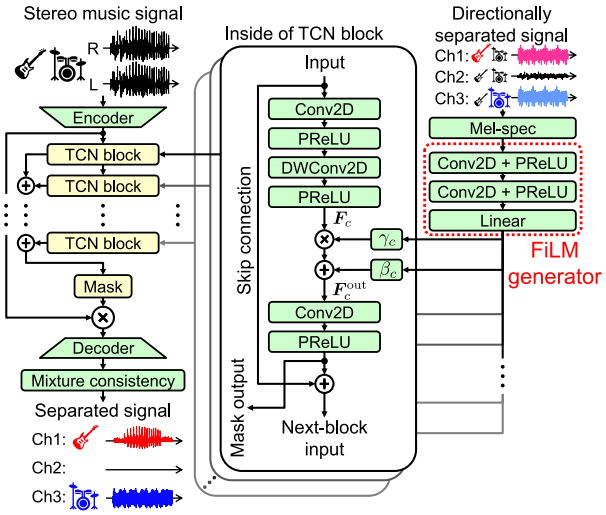

Fig. 4: Overview of the proposed IC Conv-TasNet with FiLM-based conditioning.

ルの（すなわち，N 種の方位の）方位分離信号を事前に得る．具体的には，観測ステレオ信号の各時間周波数成分の左右チャネルの振幅比から時間周波数毎の方位角を算出し，その時間周波数の振幅値で重み付けした方位ヒストグラムを求める．Fig. 2 のステレオ音楽信号（音源数は $K = 2$）を例として求めた重み付け方位ヒストグラムを Fig. 3 に示す．ステレオ音楽信号中の各音源が占める方向角の分布が，特定の方位に集中していることが確認できる．提案手法では，このヒストグラムを N 個（Fig. 3 の例では $N = 5$）の方位領域にハードに分割することで，N チャネルの方位分離信号を得る．これにより，観測ステレオ音楽信号中の各音源が優勢な方位を粗く反映した中間表現が得られ，後段の DNN に対する補助情報として利用できる．注意点として，この方位分離は，全ての音源を完全に分離することを目的とするものではない．観測信号のみから直接得ることができる粗い分離信号でよく，それ以上の音源分離は DNN に期待されている．従って，同一方位に複数の音源が存在する場合には，それらは分離されず，同一の方位チャネルに統合された信号として出力される．本手法では，DNN の最終出力もまた音源単位ではなく方位単位で定義されており，同一方向に定位する複数の音源をまとめて 1 つの方位成分として推定することを正解とする．

本稿では，各音源にパンニングが施された楽曲を想定し，各方位領域には最大 1 つの音源が存在することを仮定する．従って，同一音源が複数の方位チャネルに重複して割り当てられることはない．より現実的な，同一方位に複数の音源が存在する信号や，ステレオエフェクトなどが適用された信号への効果検証及び改良は，今後の課題とする．

2.3 補助情報を活用した Conv-TasNet 拡張

提案手法では，end-to-end 型の DNN として，Conv-TasNet [5] の拡張版である inter-channel Conv-TasNet（IC Conv-TasNet）[6] を基本構造として用いる．まず，提案手法のモデルを Fig. 4 に示す．ステレオ音楽信号をメイン入力として受け取り，事前に観測信号から得た方位分離信号を補助情報として FiLM 層 [4] に入力することで，各 TCN ブロックの特徴量を条件付けしている．さらに，出力分離信号の総和が入力混合信号と一致するように，mixture consistency 層を最終層に用いている．

次に，ネットワーク内部について述べる．まず観測ステレオ音楽信号をエンコーダで特徴空間に変換した後，複数の時系列畳み込みネットワーク（TCN ブロック）を通じて時間的依存関係を学習する．また，補助情報として与えられた方位分離信号はメルスペクトログラムに変換した後，いくつかの層（FiLM ジェネレータ）を通して各 TCN ブロックの特徴マップ $\boldsymbol{F}_c \in \mathbb{R}^{C\times H\times D}$ に対するスケーリング係数 $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \gamma_c, \cdots, \gamma_C]^{\mathrm{T}}$ 及びバイアス $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_c, \beta_C]^{\mathrm{T}}$ を生成する．ここで，$c = 1, 2, \cdots, C$ は TCN ブロックのインデックス，H 及び D は特徴マップの次元を表す．各 TCN ブロックにおいて係数 γ_c 及び β_c は次式のように適用され，特徴毎に条件付けが行われる．

$$\boldsymbol{F}_c^{\mathrm{out}} = \gamma_c \odot \boldsymbol{F}_c + \beta_c \tag{1}$$

最後に mixture consistency 層について述べる．推定信号 $\hat{\boldsymbol{s}}_n = [\hat{s}_n(1), \hat{s}_n(2), \cdots, \hat{s}_n(L)]^{\mathrm{T}} \in \mathbb{R}^L$ を，各方位領域に対応する N 個の（方位分離された）モノラル信号と定義し，その正解信号 $\boldsymbol{s}_n = [s_n(1), s_n(2), \cdots, s_n(L)]^{\mathrm{T}} \in \mathbb{R}^L$ と比較して学習する．ここで L は時間信号の信号長を表す．また，観測ステレオ音楽信号をモノラル化した信号を $\boldsymbol{x} = [x(1), x(2), \cdots, x(L)]^{\mathrm{T}} \in \mathbb{R}^L$ とおくとき，推定信号 $\hat{\boldsymbol{s}}_n$ の n に関する総和が $\boldsymbol{x}$ に一致することが

望ましい．そこで，補正前の推定信号 $\tilde{\boldsymbol{s}}_n$ に対して次式の射影を適用する [7].

$$\hat{\boldsymbol{s}}_n = \tilde{\boldsymbol{s}}_n + \frac{1}{N}\left(\boldsymbol{x} - \sum_{n=1}^{N} \tilde{\boldsymbol{s}}_n\right) \tag{2}$$

この射影により，DNN 出力の総和が入力信号と整合するような一貫性のある音源分離 DNN の学習が可能となる．

2.4 DNN 学習時の損失関数

DNN 学習時に用いる損失関数について述べる．提案手法の DNN モデルは Fig. 1 (b) のように，優勢な音源が存在する方位に対応するチャネル（有音チャネル）にはその音源の分離信号を出力し，優勢な音源がほぼ存在しない方位に対応するチャネル（無音チャネル）には無音信号を出力することを期待している．そこで，有音チャネルに対しては真の分離信号と推定信号間の threshold scale-invariant signal-to-noise ratio（SI-SNR）[8] を用い，また無音チャネルに対しては L_1 ノルム損失を用いる．すなわち，無音になるべきチャネルの推定信号の無音化を L_1 ノルムの最小化として実装している．また，有音チャネルのインデクス集合 $\mathcal{C}_{\text{act}}$ と無音チャネルのインデクス集合 $\mathcal{C}_{\text{sil}}$ は，各チャネルの真の分離信号 $\boldsymbol{s}_n$ のエネルギーに基づき次式のように判定して定める．

$$\mathcal{C}_{\text{act}} = \{\, n \mid \|\boldsymbol{s}_n\|_2^2 \geq \varepsilon_s \,\} \tag{3}$$

$$\mathcal{C}_{\text{sil}} = \{\, n \mid \|\boldsymbol{s}_n\|_2^2 < \varepsilon_s \,\} \tag{4}$$

ここで ε_s は無音判定の閾値である．

有音チャネルに対して用いる threshold SI-SNR は，真の分離信号 $\boldsymbol{s}_n$ と推定信号 $\hat{\boldsymbol{s}}_n$ から次式で計算する．

$$\mathcal{L}_{\text{SI-SNR}}(\boldsymbol{s}_n, \hat{\boldsymbol{s}}_n) = -10\log_{10}\frac{\|\boldsymbol{t}_n\|_2^2}{\|\boldsymbol{e}_n\|_2^2 + \tau\|\boldsymbol{t}_n\|_2^2 + \varepsilon} \tag{5}$$

ここで，射影係数 δ，ターゲット成分 $\boldsymbol{t}_n$，及び残差成分 $\boldsymbol{e}_n$ はそれぞれ次のように定義される．

$$\delta = \frac{\langle \hat{\boldsymbol{s}}_n, \boldsymbol{s}_n \rangle}{\|\boldsymbol{s}_n\|_2^2 + \varepsilon}, \quad \boldsymbol{t}_n = \delta\, \boldsymbol{s}_n, \quad \boldsymbol{e}_n = \hat{\boldsymbol{s}}_n - \boldsymbol{t}_n \tag{6}$$

さらに，$\varepsilon > 0$ は数値安定化のための微小定数，τ は soft-threshold パラメータで，残差が小さい場合に SI-SNR が過剰に大きくなることを抑制する．

最終的に，全体の損失関数は次式となる．

$$\mathcal{L} = \frac{1}{N}\left(\lambda \sum_{n\in\mathcal{C}_{\text{sil}}} \|\hat{\boldsymbol{s}}_n\|_1 + \sum_{n\in\mathcal{C}_{\text{act}}} \mathcal{L}_{\text{SI-SNR}}(\boldsymbol{s}_n, \hat{\boldsymbol{s}}_n)\right) \tag{7}$$

ここで，λ は 2 つの損失関数のバランスを決めるパラメータである．

3 提案手法と比較手法の性能評価実験

3.1 実験条件

提案手法の有効性を検証するため，ステレオ音楽信号を用いた音源分離実験を行う．本実験では，IC Conv-TasNet [6] 及び SpaIn-Net [3] を比較手法として用い，提案手法の有効性を評価した．SpaIn-Net には各音源の理想的な定位角度を補助情報として与え，上限性能の評価を行った．

Table 1: Experimental conditions

Input signal length	10 s
Sampling frequency	16 kHz
Number of directional channels	$N = 5$
Number of sources	$K = 4$
Batch size	4
Gradient accumulation steps	4
Training epochs	200
Early stopping (patience)	30
Loss function threshold	$\tau = -30$ dB

本実験における実験条件を Table 1 に示す．使用する音楽データのドライソースは MUSDB18 [9] に含まれるモノラル音源信号（vocal，bass，drum，及び other）とした．各楽曲において 10 秒の長さの区間をランダムに定め，その区間の各音源信号を切り出し，パンニングにより観測ステレオ音楽信号を生成して学習及び評価に用いた．但し，学習データと評価データでは異なる楽曲を用いている．本条件におけるデータセットは，学習データが 32,720 秒，検証データが 3,130 秒，評価データが 3,650 秒で構成した．評価指標には推定信号と正解信号の全体的な類似度を示す source-to-distortion ratio (SDR) [10] を用いた．

3.2 実験結果

各手法における音源毎の SDR をバイオリンプロットとして Fig. 5 に示す．また，Table 2 には，各音源に対する平均 SDR と，各手法による SDR 改善量（ΔSDR）を示す．混合信号は全音源で SDR が低く，他音源からの強い干渉が存在することが分かる．方位分離信号は粗い空間特徴量として推定された不完全な分離信号であるが，すべての音源で SDR が向上しており，DNN に与える補助情報として一定の有益な情報となっていることが確認できる IC Conv-TasNet では，すべての音源において中央値の SDR 改善は確認できるものの，分布の広がりが比較的大きく，楽曲毎の分離性能にばらつきが見られる．全音源平均の ΔSDR は 9.14 dB 程度であり，音源によっては十分な改善が得られない場合も存在する．SpaIn-Net は，全音源において IC Conv-TasNet を上回る SDR 改善を示しており，中央値の上昇とともに分布の幅も縮小している．全音源平均では 11.43 dB の改善が得られており，補助情報として定位角度を導入することで，分離性能及びその安定性が向上していることが分かる．これは，補助情報として理想的な定位角度を与えることで，空間情報を直接モデルに反映できているためと考えられる．提案手法は，bass，drum，other，及び vocal のすべての音源において最も高い SDR 改善を示している．全音源平均の ΔSDR は 12.92 dB に達しており，既存手法を上回る結果となった．バイオリンプロットからも，分布全体が高い SDR 改善量側へシフトしていることが確認でき，中央値において最も良好である．以上より，提案手法は従来の IC

Table 2: Average scores [dB] of mixture SDR and the SDR improvement from the mixture for each source

Source		bass	drum	other	vocal	Avg
Mixture	SDR	-4.97	-4.29	-5.86	-6.61	-5.37
Directionally separated signal	ΔSDR	5.12	5.92	6.58	8.04	6.41
IC Conv-TasNet		9.69	8.07	8.15	10.66	9.14
SpaIn-Net (Oracle direction)		11.34	10.79	10.10	13.48	11.43
Proposed method		**12.78**	**11.64**	**12.13**	**15.14**	**12.92**

Conv-TasNet 及び SpaIn-Net を明確に上回ることが示され，方位分離によって得られる粗い音源分離信号がステレオ音楽分離 DNN の性能を押し上げる重要な補助情報となりえることを確認した．

4 おわりに

本稿では，左右チャネル間の音量比に基づく方位分離信号を DNN の補助情報として活用し，FiLM を介して IC Conv-TasNet に統合するステレオ音楽分離手法を提案した．評価の結果，提案手法は従来の IC Conv-TasNet 及び SpaIn-Net と比較して分離精度が向上することを確認した．この結果は，方位分離信号による条件付けが，ステレオ音楽分離性能の向上に有効であることを示している．

謝辞 本研究の一部は Google Gift 2024 の助成を受けたものである．また，本研究を進めるにあたり，Google DeepMind の小泉悠馬氏から有益なご助言をいただいたため，ここに感謝の意を表する．

参考文献

[1] D. Kitamura, H. Saruwatari, H. Kameoka, Y. Takahashi, K. Kondo, and S. Nakamura, "Multichannel signal separation combining directional clustering and nonnegative matrix factorization with spectrogram restoration," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 4, pp. 654–669, 2015.

[2] S. Araki, N. Ito, R. Haeb-Umbach, G. Wichern, Z.-Q. Wang, and Y. Mitsufuji, "30+ years of source separation research: Achievements and future challenges," *in Proc. ICASSP*, 2025.

[3] D. Petermann and M. Kim, "SpaIn-Net: Spatially informed stereophonic music source separation," *in Proc. ICASSP*, pp. 106–110, 2022.

[4] P. Ethan, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," *in Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[6] D. Lee, S. Kim, and J. W. Choi, "Inter-channel Conv-TasNet for multichannel speech enhancement," *arXiv preprint arXiv:2111.04312*, 2021.

[7] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," *in Proc. ICASSP*, pp. 900–904, 2019.

[8] Y. Koizumi, S. Karita, S. Wisdom, H. Erdogan, J. R. Hershey, L. Jones, and M. Bacchiani, "DF-Conformer: Integrated architecture of Conv-TasNet and conformer using linear complexity self-attention for speech enhancement," *in Proc. WASPAA*, pp. 161–165, 2021.

[9] Z. Rafii, A. Liutkus, F. R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," *in Proc. ISMIR*, 2017.

[10] 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.

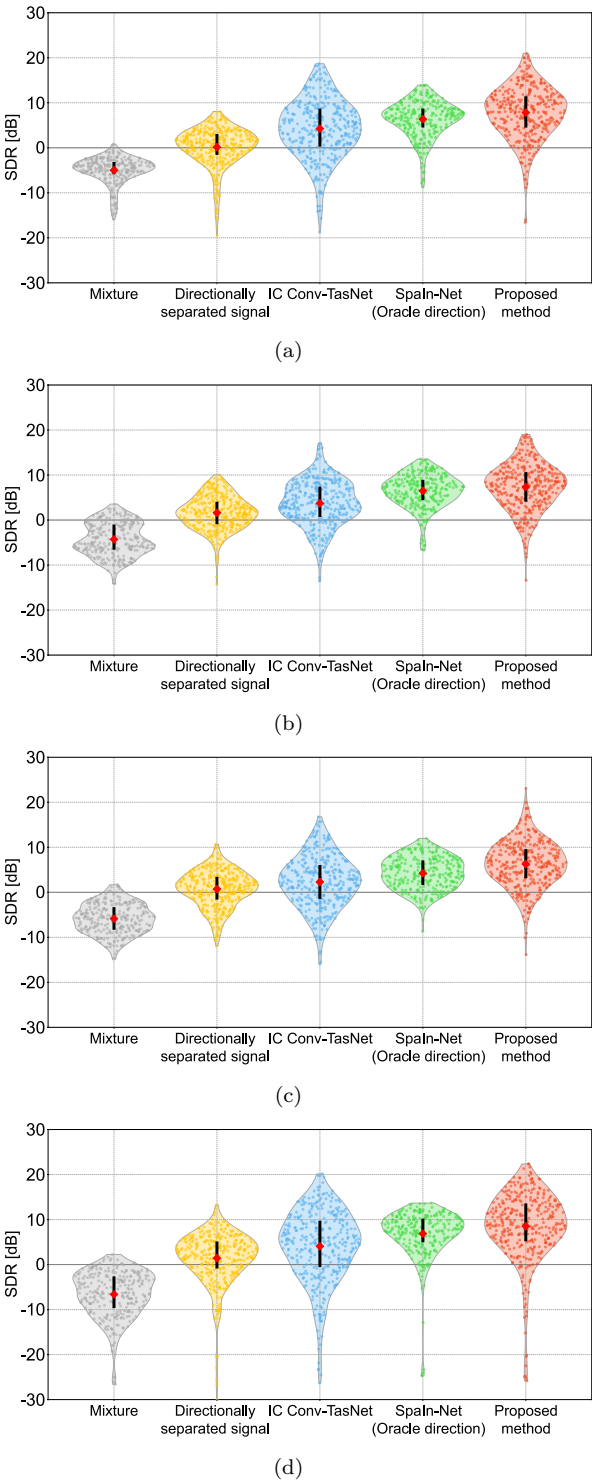

Fig. 5: Violin plots of SDR for each method: (a) bass, (b) drum, (c) other, and (d) vocal.