特別研究論文

（査読済み）

研　究　題　目

音量比による事前処理を援用した深層ステレオ音楽分離

提出年月日		2026 年　1 月　28 日
氏	名	加藤　大輝
主	査	北村　大地　准教授
副	査	柿元　健　准教授
副	査	村上　幸一　准教授

香川高等専門学校
専攻科
創造工学専攻

NATIONAL INSTITUTE OF TECHNOLOGY KAGAWA COLLEGE
香川高専

Deep Stereo Music Separation Using Preprocessing Based on Interchannel Amplitude Ratio

Taiki Kato

Advanced Course in Industrial and Systems Engineering
National Institute of Technology, Kagawa College

Abstract

Music separation isolates individual sound sources, such as vocals and instruments, from music signals. It is essential for applications such as automatic music transcription, music recognition, analysis, and genre classification. Recent advances in deep neural networks (DNNs) have significantly improved source separation. End-to-end DNN models that operate directly on waveforms achieve high separation accuracy. Such state-of-the-art methods achieve strong performance at the cost of requiring large training datasets, leading to substantial computational costs. Auxiliary information, such as left–right source placement, is effective for improving separation accuracy while reducing the learning burden. By explicitly providing spatial cues between sources, it facilitates learning and enhances separation performance. However, real-world music signals often employ various stereo production techniques, including stereo effects and reverberation, which makes extracting reliable features for separation difficult. In this paper, we focus on simple spatial cues derived directly from stereo music signals. Specifically, we use directional separation signals based on inter-channel amplitude ratios as auxiliary information for DNN-based music separation. FiLM reduces the complexity of internal representations, leading to accurate and stable music separation. Performance evaluation shows that the proposed method outperforms both DNN models without auxiliary information and conventional auxiliary-based methods. Specifically, it achieves an average source-to-distortion ratio (SDR) improvement of 12.9 dB across all sources, which is higher than the other compared methods. The proposed method also generalizes well to inputs with varying numbers of sources and previously unseen sounds. These results demonstrate that leveraging auxiliary information can enhance both separation accuracy and generalization. Future work will explore combining this approach with more sophisticated spatial cues and improved conditioning mechanisms. The insight gained may also apply to other domains, such as speech recognition and environmental sound analysis.

Key Words: deep neural networks, stereo music separation, feature-wise linear modulation, inter-channel level difference

（和訳）

音楽分離は，音楽信号に含まれるボーカルや楽器音など複数の音源を個別に抽出する技術であり，音楽認識や解析，自動採譜，ジャンル認識など，幅広い応用が期待される重要な研究分野である．近年，深層ニューラルネットワーク（deep neural network: DNN）の発展により，音源分離技術は大きく進展している．特に波形そのものを入出力とする end-to-end モデルの DNN は高精度な分離を実現している．しかし，現在の深層学習を用いた最先端の音源分離手法は，高性能である一方で，膨大な学習データを用いているため，それを処理するための莫大な計算コストを必要とするという課題がある．分離精度を向上させつつ学習負担を軽減する手法として，音源の左右配置などの補助情報を活用する手法が有効である．補助情報により音源の空間的な位置関係をモデルに明示的に与えることで，学習負担の軽減と分離精度の向上が期待できる．しかし，実際の音楽信号ではステレオエフェクトやリバーブなど多様なステレオ化手法が用いられている．そのため，分離のための正確な特徴量の抽出は困難である．本論文では，既知のステレオ音楽信号から直接生成可能な情報のうち，特にステレオ化の最も基本的な処理に基づく空間的手がかりに着目する．具体的には，左右チャネル間の音量比に基づく方位分離信号を DNN の補助情報として利用する．この補助情報を活用し，特徴量的線型変調（feature-wise linear modulation: FiLM）を介して DNN に適応させる新しいステレオ音楽分離手法を提案する．FiLM による動的条件付けにより，DNN が学習すべき内部表現の複雑性を低減し，高精度かつ安定した音楽分離を実現する．性能評価の結果，提案手法は補助情報を用いない DNN モデルおよび既存の補助情報を活用する手法のいずれよりも高い分離精度を達成した．特に，全音源に対する平均 source-to-distortion ratio（SDR）の改善量は 12.9 dB に達し，比較手法を上回る結果となった．
さらに，ステレオ音楽信号内の音源数が変動する入力や未知音源に対しても良好な性能を維持し，高い汎化性能を示した．これらの結果から，補助情報を活用することで，ステレオ音楽分離における分離精度と汎化性能の双方を向上できることが確認された．今後は，より高度な空間表現や改良された条件付け機構と組み合わせることで，さらなる性能向上が期待される．また，本論文で得られた学習に補助的な情報を加えることで性能向上が可能という知見は，音声認識や環境音解析など，ステレオ音楽分離以外の分野にも応用可能である．

目次

第 1 章

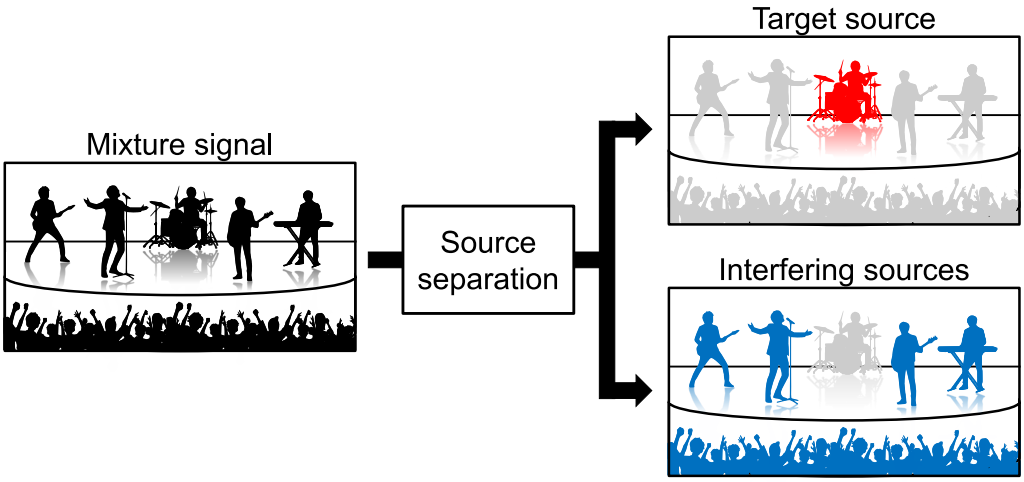
緒言

1.1　本論文の背景

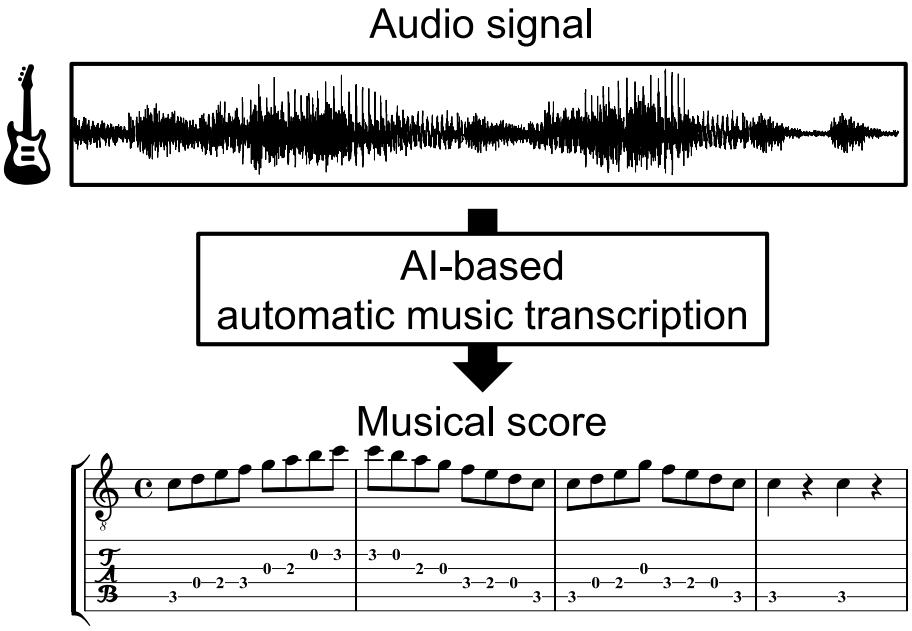
音楽信号は，ボーカルや楽器音など複数の音源が時間・周波数的に重なり合った複雑な信号である．現在，流通している音楽信号の多くはステレオ信号として記録・配布されており，左右チャネルを用いることで音源の空間的な配置が表現されている．このようなステレオ音楽信号に含まれる空間的配置に関する情報は，音源分離における重要な手がかりとなり得る．このような背景のもと，観測されたステレオ音楽信号から，各音源の独立した信号を推定するステレオ音楽分離技術が研究されている．この技術は音響信号処理における重要な研究テーマであり，音声処理をはじめ幅広い応用が期待される．具体的な応用例としては，Fig. 1.1 のような音楽信号から雑音を除去して目的の音源のみを抽出・強調するタスクや，ライブ会場で収録された複雑な混合音から，ドラムなどの特定の楽器音のみを抽出する音源分離タスク，さらには楽器音の自動採譜タスクなどが挙げられる．これらのタスクを満たすには高精度なステレオ音楽分離手法が求められる．

初期のステレオ音楽分離手法では，時間周波数領域における信号の疎性や構造に着目した劣決定音源分離の枠組みが用いられてきた [1, 2]．これらの手法では，左右チャネル間の相関や振幅差といった空間的な手がかりを利用することで，観測数よりも音源数が多い劣決定条件下での音源分離が試みられてきた．その後，音楽信号の持つスペクトル構造を明示的にモデル化できる点から，非負値行列因子分解（nonnegative matrix factorization: NMF）を基盤とする手法が広く研究されるようになった．特に，確率モデルに基づく NMF や，事前に学習した音源モデルを用いる教師あり NMF を導入することで，分離性能の向上が図られてきた [3, 4, 5]．一方で，これらの手法は，精緻な信号モデル設計や多数のパラメータ推定を必要とする場合が多く，計算コストや実環境への適用性の観点で課題が残されている．

これに対して，近年の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では，DNN が分離後の信号や分離マスクを直接推定することで，高精度な音源分離が実現されている [6, 7, 8]．特に，Fig. 1.2 (a) に示すような，波形そのものを入出力とする end-to-end モデルは，複雑な音楽信号に対しても高い分離性能を示している [9]．これらの DNN ベースの手法は，従来の統計的手法では扱いが困難であった音源間の複雑な重な



(a)



(b)

Fig. 1.1. Application examples of audio signal processing: (a) music source separation, and (b) music instrument transcription.

りや非線形な混合を学習によって表現できる点に特徴がある．一方で，高品質な分離性能を得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．DNN による音源分離性能を向上させつつ，学習負担やネットワークの複雑性を抑えるための方法として，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アプローチが注目されている．

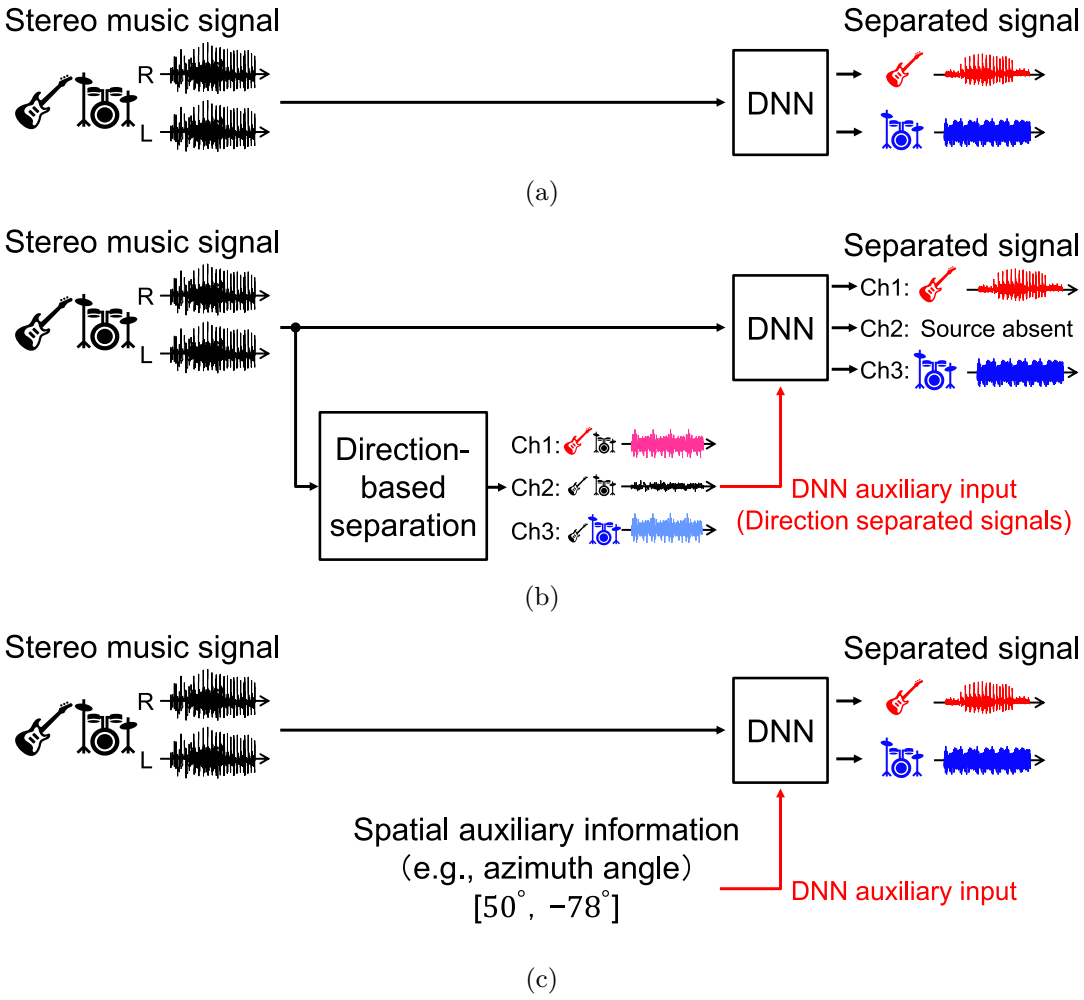


Fig. 1.2. Comparison of DNN-based music source separation architectures: (a) end-to-end model, (b) proposed FiLM-based architecture using stereo-derived directional cues, and (c) model incorporating auxiliary spatial information.

1.2　本論文の目的

本論文の目的は，従来の DNN による音楽音源分離が，大規模な学習データや多大な計算資源に依存しているという課題を克服するため，追加の外部計測情報に依存せず，ステレオ音楽信号のみから生成可能な補助情報を活用して，DNN による音源分離精度を向上させることである．具体的には，ステレオ音楽信号の左右チャネル間音量比に基づき生成可能な方位分離信号に着目し，これを特徴量的線型変調（feature-wise linear modulation: FiLM）[10] を介して DNN に適応させる新しいステレオ音楽分離手法を提案する．具体的なネットワークアーキテクチャを Fig. 1.2 (b) に示す．提案手法は，理想的には DNN が内部で自動的に学習可能である音源の空間的関係を，方位分離という古典的かつ粗い特徴量としてあらかじめ設計し入力

するアプローチと位置付けられる．これは，空間的対応関係をすべて DNN に学習させるのではなく，学習が容易で解釈性の高い形に事前に圧縮することで，ネットワーク構造の単純化および学習コストの低減を図ることを目的としている．従来の end-to-end モデル [7, 9] に補助情報を付与することで，高精度な音源分離と実用性の両立を図る．本手法で生成される分離信号は，従来の音源数に基づく分離とは異なり，与えられたステレオ音楽信号から生成される方位分離信号のチャネル数に応じて出力チャネルが決定される点に特徴がある．すなわち，入力信号中にいくつ音源が存在していても，方位分離信号のチャネル数が N であれば常に N チャネルの出力を生成する．各チャネルには，その方位において最も主要な音源成分が割り当てられ，異なるチャネル間で同一音源が重複して出力されることはない．また，特定の方位に支配的な音源が存在しない場合，そのチャネルは実質的に音成分を持たない出力となる．このように，本手法は空間的な優位性に基づくチャネル依存の分離構造を持ち，従来のソース数依存型の音源分離とは本質的に異なるアプローチである．

なお，ステレオ音楽分離において，空間的特徴量を補助情報として用いる先行研究も報告されている [11]．Fig. 1.2 (c) に示すように，本論文と同様に，空間的手がかりを明示的に DNN に与えるという点で近い思想を持つ．本論文では，より取得が容易で汎用性の高い特徴量として，チャネル間音量比に基づく方位分離信号を採用し，その有効性を実験的に検証する．

1.3　本論文の構成

本論文は以下のように構成されている．2 章では，ステレオ音楽分離の基礎理論と先行研究 [11] について解説する．特に，STFT やメルスペクトログラムといった時間周波数特徴量，DNN における FiLM による条件付け，方位分離や先行研究 [11] での空間情報の活用について述べ，本論文の課題設定の根拠を示す．3 章では，本論文の提案手法である音量比を用いた方位分離の事前処理と FiLM によるステレオ音楽分離手法の全体像を示す．左右チャネル間の音量比に基づく方位分離信号の生成，DNN の入出力やデータセット作成方法，ネットワーク構造，損失関数までを詳細に述べる．4 章では，提案手法の有効性を検証するために，比較手法を用いた性能評価実験を行う．指標を用いて性能改善の程度を評価し，FiLM による条件付けの影響も示す．5 章では，提案手法の汎化性能を評価する実験を行う．学習データに依存せず，未知音源や音源数が変動する入力に対しても高精度な分離が可能かを検証する．最後に 6 章では，本論文で得られた知見を総括し，今後の課題や他分野への応用可能性について述べる．

第 2 章

基礎理論および先行研究

2.1 はじめに

本章では，本論文で提案する音源分離手法の構成要素となる基礎技術および先行研究について述べる．まず，2.2 節では時間周波数解析の基本技術である短時間フーリエ変換（short-time Fourier transform: STFT）について，2.3 節では提案手法に関わる特徴変調の技術を概説する．続く 2.4 節では，本論文で扱う問題に関連する先行研究を整理する．これは後の実験で比較手法としても使用する．最後に，2.5 節で本章をまとめる．

2.2 STFT

STFT は，音響信号の時間的に変化するスペクトルを，時間周波数領域と呼ばれる二次元の特徴量空間で表現するための変換手法である．STFT の概要を Fig. 2.1 に示す．STFT では，音響信号の時間波形を短時間区間に分割し，窓関数を乗じたうえで周波数領域へと変換する．音響信号の時間波形を次式で定義する．

$$\boldsymbol{y} = [y(1), y(2), \cdots, y(l), \cdots, y(L)]^{\mathrm{T}} \in \mathbb{R}^{L} \tag{2.1}$$

ここで，$\cdot^{\mathrm{T}}$ は転置，L は時間信号 $\boldsymbol{y}$ の長さ，$l = 1, 2, \cdots, L$ は時間信号 $\boldsymbol{y}$ の離散時間サンプルをそれぞれ表す．短時間区間長（窓長）および短時間区間のシフト長をそれぞれ Q および τ としたとき，時間領域の信号時間領域の信号 $\boldsymbol{y}$ の j 番目の短時間区間（時間フレーム）の信号 $\tilde{\boldsymbol{y}}^{(j)}$ は次式で表される．

$$\tilde{\boldsymbol{y}}^{(j)} = [y((j-1)\tau + 1), y((j-1)\tau + 2), \cdots, y((j-1)\tau + Q)]^{\mathrm{T}} \tag{2.2}$$

$$= [\tilde{y}^{(j)}(1), \tilde{y}^{(j)}(2), \cdots, \tilde{y}^{(j)}(q), \cdots, \tilde{y}^{(j)}(Q)]^{\mathrm{T}} \in \mathbb{R}^{Q} \tag{2.3}$$

ここで，$j = 1, 2, \cdots, J$ および $q = 1, 2, \cdots, Q$ は，それぞれ時間フレームおよび時間フレーム内のサンプルを示す．また，フレーム数 J は次式によって与えられる．

$$J = \frac{L}{\tau} \tag{2.4}$$

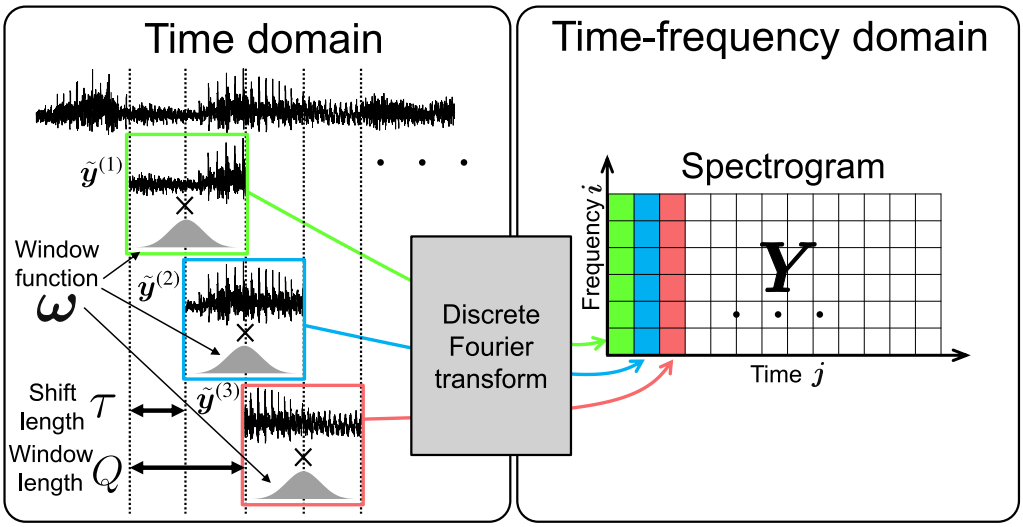


Fig. 2.1. Mechanism of STFT.

ただし，信号長 L はフレーム数 J が整数となるように各時間フレームの信号の両端にゼロを挿入する処理 （ゼロパディング） が施されている．このとき時間フレームの信号を全ての j についてまとめた全時間フレームの信号は次式の通り定義できる．

$$\tilde{\boldsymbol{Y}} = [\tilde{\boldsymbol{y}}^{(1)}\ \tilde{\boldsymbol{y}}^{(2)}\ \cdots\ \tilde{\boldsymbol{y}}^{(j)}\ \cdots\ \tilde{\boldsymbol{y}}^{(J)}] \in \mathbb{R}^{Q\times J} \tag{2.5}$$

次に，長さ Q の窓関数を $\boldsymbol{\omega} = [\omega(1), \omega(2), \cdots, \omega(q), \cdots, \omega(Q)]^{\mathrm{T}} \in \mathbb{R}^{Q}$ と定義する．STFT の処理は次式で表される．

$$\boldsymbol{Y} = \mathrm{STFT}_{\boldsymbol{\omega}}(\tilde{\boldsymbol{Y}}) \in \mathbb{C}^{I\times J} \tag{2.6}$$

$$y_{ij} = \sum_{q=1}^{Q} \omega(q) y^{(j)}(q) \exp\left\{\frac{-\iota 2\pi (q-1)(i-1)}{Q}\right\} \tag{2.7}$$

ここで，$\boldsymbol{Y}$ は複素スペクトログラムと呼ばれ，複素数の時間周波数成分を持つ行列である．また，y_{ij} は $\boldsymbol{Y}$ の (i,j) 要素を表す．I は $I = \lfloor \frac{Q}{2} \rfloor + 1$ を満たす整数（$\lfloor\cdot\rfloor$ は床関数），$i = 1, 2, \cdots, I$ は周波数ビンのインデクス，$j = 1, 2, \cdots, J$ は時間フレームのインデクス，ι は虚数単位を示している．このように，時間領域の信号を一定幅 Q の短時間毎に区切って分析窓関数 $\boldsymbol{\omega}$ を乗じて離散フーリエ変換（discrete Fourier transformation: DFT）することで，周波数と時間の 2 次元複素行列であるスペクトログラム $\boldsymbol{Y}$ に変換できる．複素スペクトログラムは各時間周波数の振幅成分と位相成分を持っているが，音源分離等の多くの音響信号処理では，振幅成分のみを取り扱うことが多い．その場合は，複素スペクトログラム $\boldsymbol{Y}$ の各要素に関して絶対値を取った振幅スペクトログラム $|\boldsymbol{Y}| \in \mathbb{R}_{\geq 0}^{I\times J}$ や，絶対値の 2 乗を取ったパワースペクトログラム $|\boldsymbol{Y}|^{\cdot 2} \in \mathbb{R}_{\geq 0}^{I\times J}$ を処理の対象とする．ここで，ベクトルや行列に対する絶対値記号およびドット付き指数乗はそれぞれ要素毎の絶対値および要素毎の指数乗を表す．

本論文では，人間の聴覚特性に基づいた周波数尺度であるメル尺度に変換することで，学習モデルに適した入力特徴量を構築する．具体的には，STFT により得られたパワースペクトロ

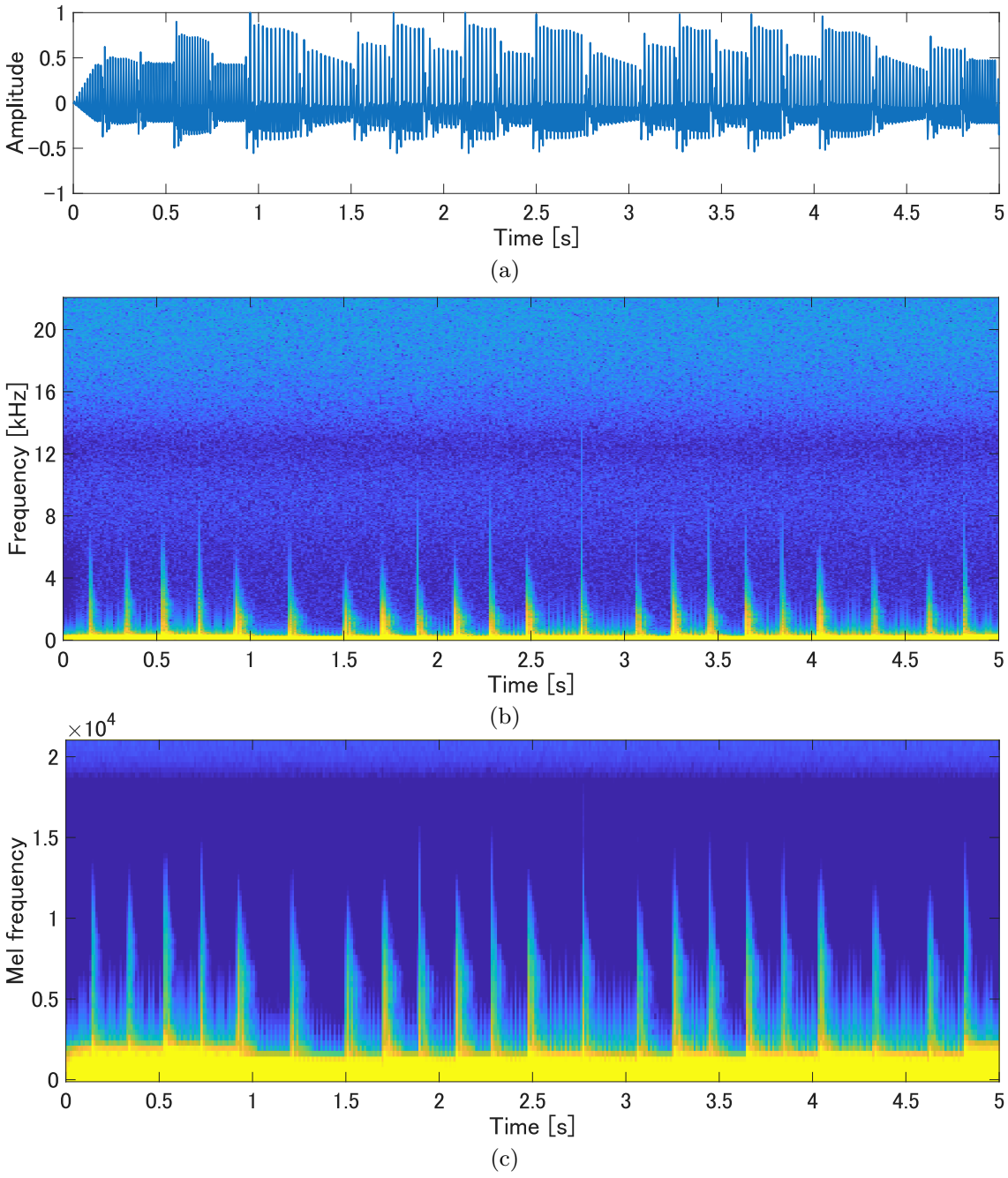


Fig. 2.2. Time-domain and time-frequency representations of an audio signal: (a) waveform, (b) power spectrogram, and (c) mel spectrogram.

グラム $|\boldsymbol{Y}|^{\cdot 2} \in \mathbb{R}_{\geq 0}^{I \times J}$ に対して，メルフィルタバンク行列 $\boldsymbol{W}_{\mathrm{mel}} \in \mathbb{R}_{\geq 0}^{B \times I}$ を適用することで，メルスペクトログラム $\boldsymbol{Y}_{\mathrm{mel}} \in \mathbb{R}_{\geq 0}^{B \times J}$ を得る．

$$\boldsymbol{Y}_{\mathrm{mel}} = \boldsymbol{W}_{\mathrm{mel}} |\boldsymbol{Y}|^{\cdot 2} \tag{2.8}$$

ここで，$\boldsymbol{W}_{\mathrm{mel}}$ は，STFT の周波数軸 I ビンを，聴覚特性に基づく B 個のメル周波数ビンに

次元圧縮する行列であり，第 b 行目が b 番目のメルフィルタに対応する．本論文では，メルスペクトログラム $\boldsymbol{Y}_{\mathrm{mel}}$ を学習モデルへの入力として使用する．時間領域波形，時間周波数表現としてのパワースペクトログラム，およびメルスペクトログラムを Fig. 2.2 に示す．Fig. 2.2 (b) のパワースペクトログラムは STFT の周波数解像度をそのまま保持するため，低周波数成分のパワーが小さい場合はほとんど青色となり視認性が低い．一方，Fig. 2.2 (c) に示すメルスペクトログラムは，人間の聴覚特性を近似するメル尺度に基づき，周波数軸を非線形に再配置した時間周波数表現である．このため，低周波数帯は高周波数帯に比べて相対的に細かい周波数分解能で表現され，知覚的に重要な低周波数成分の構造が視覚的に把握しやすくなっている．

2.3　Feature-wise linear modulation

本節では，補助情報を DNN の内部に組み込む手法の一つである FiLM [10] について述べる．DNN に FiLM を適用した概念図を Fig. 2.3 に示す．一般的な DNN は，入力層，中間層，出力層から構成される．中間層の代表例として畳み込み層（convolution layer: conv 層）があり，conv 層を用いた DNN は畳み込みニューラルネットワーク（convolutional neural network: CNN）[12] と呼ばれる．ほかにも，再帰型ニューラルネットワーク（recurrent neural network: RNN）[13]，RNN の一種である長・短期記憶（long short-term memory: LSTM）[14]，さらに LSTM を双方向に拡張した双方向再帰型ニューラルネットワーク（bidirectional LSTM: BiLSTM）など，様々なアーキテクチャが存在する．本節では，これらの具体的なネットワーク構造に依存しない，一般的な DNN モデルに対する FiLM の適用方法について解説する．

FiLM は，DNN の中間層の出力である特徴量 $\boldsymbol{F} \in \mathbb{R}^{C\times H\times V}$ を，チャネルごとに線形変換する手法である．ここで，$c = 1, 2, \cdots, C$ はチャネル番号，$h = 1, 2, \cdots, H$ は特徴量ビン，および $v = 1, 2, \cdots, V$ は時間フレームを表す．FiLM は，特徴量 $\boldsymbol{F}$ の各チャネル $c = 1, \cdots, C$ に対して，スケーリング $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \cdots, \gamma_c, \cdots, \gamma_C]^{\mathrm{T}}$ とバイアス $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_c, \cdots, \beta_C]^{\mathrm{T}}$ による線形変換を行う．

$$\boldsymbol{F}_c^{\mathrm{out}} = \gamma_c \cdot \boldsymbol{F}_c + \beta_c \tag{2.9}$$

ここで，$\boldsymbol{F}_c \in \mathbb{R}^{H\times V}$ は入力特徴量の c 番チャネル，γ_c および β_c は補助情報に対して線形な変形を与えるパラメータである．単一の FiLM 層の動作の概念図を Fig. 2.4 に示す．γ_c によりチャネルごとの特徴量がスケーリングされ，β_c によりシフトされることがわかる．補助情報から $\boldsymbol{\gamma}$ および $\boldsymbol{\beta}$ を生成するネットワークを FiLM ジェネレータと呼ぶ．補助情報が FiLM ジェネレータ に入力され，特徴量 $\boldsymbol{F}$ に適用される $\boldsymbol{\gamma}$ および $\boldsymbol{\beta}$ が生成されていることが Fig. 2.3 からわかる．

本論文では，補助情報として前節で定義したメルスペクトログラム $\boldsymbol{Y}_{\mathrm{mel}} \in \mathbb{R}^{B\times J}$ を入力とした場合を考える．FiLM ジェネレータは，補助情報を conv 層で処理し，その出力を線形層（以後，Linear 層）に入力することで，$\boldsymbol{\gamma}$ および $\boldsymbol{\beta}$ を生成するネットワークである．一連の

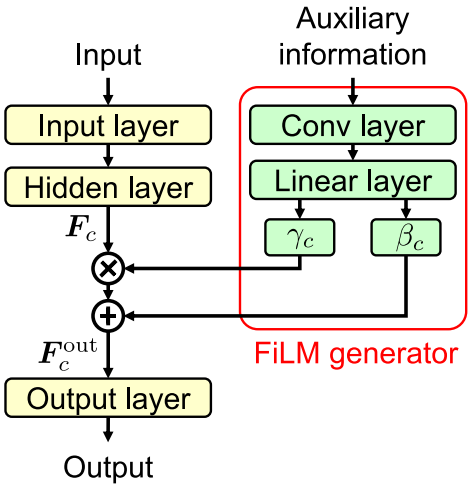


Fig. 2.3. DNN overview with FiLM applied to intermediate features.

処理を $f_\theta(\cdot)$ とすると，$\boldsymbol{\gamma}$ および $\boldsymbol{\beta}$ は次式で表される．

$$(\boldsymbol{\gamma}, \boldsymbol{\beta}) = f_\theta(\boldsymbol{Y}_{\text{mel}}) \tag{2.10}$$

パラメータはチャネルごとに 2 つ (γ_c, β_c) のみで済むため，計算コストも低く，大規模ネットワークや高解像度データでも効率的な FiLM ジェネレータの学習と補助情報の援用が可能となる．FiLM は，視覚質問応答タスクにおける画像と言語を用いた視覚推論 [10] において有効性が示されて以来，画像と音声，音声とテキストといったマルチモーダル学習への応用 [15, 16] や，音声信号やその他の時系列データに対する条件付き変換 [17] など幅広く利用されてきた．さらに，解剖学的画像のノイズ除去タスクにおける空間適応的変調 [18] や，グラフ構造データへの適用 [19] も報告されており，応用範囲は非常に広範である．これらの文献では，FiLM が多様な条件付き変調に有効であることを示しており，本論文において音響特徴量を条件とする FiLM ベース音源分離手法を設計する動機となる．

2.4 先行研究

本節では，補助情報を利用したステレオ音楽分離の関連研究を述べる．特に，空間情報を補助情報として活用する手法である spatially informed network（SpaIn-Net）[11] について解説し，本論文の比較手法として用いる．

2.4.1 SpaIn-Net

SpaIn-Net [11] は，従来のステレオ音楽音源分離の DNN モデルが，ステレオ音楽信号に含まれる空間情報を暗黙的に利用しているのに対し，空間情報を補助情報として明示的かつ能動的に利用することで，分離性能の向上を目指した手法である．特に，同種の楽器が混在するような，スペクトル的・音色的特徴が類似している音源の分離において，空間情報が音源の分離

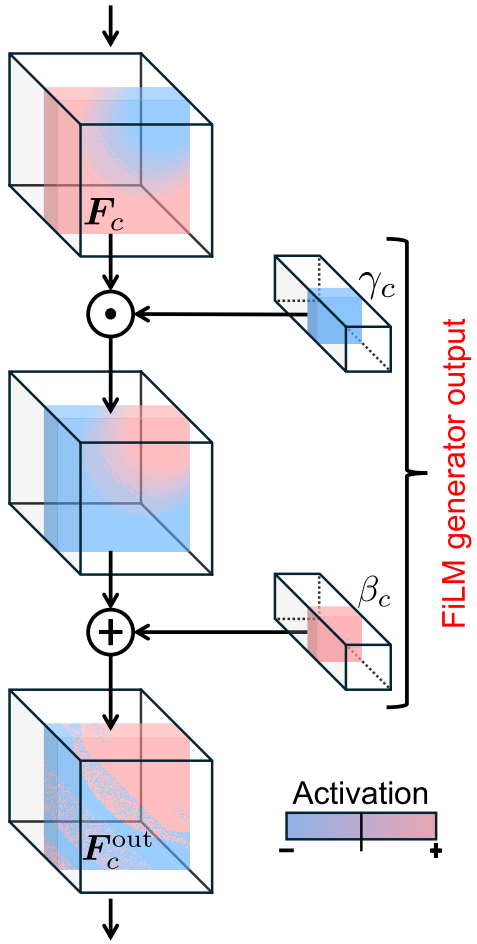


Fig. 2.4. A single FiLM layer for a CNN. ⊙ denotes the Hadamard product, where each feature map is scaled and shifted by $\boldsymbol{\gamma}$ and, $\boldsymbol{\beta}$ respectively.

に役立つことを示している．補助情報として利用可能な空間的特徴量 [20, 21, 22] は多岐にわたるが，文献 [11] ではその一例としてパンニング角 $\boldsymbol{\alpha}$ が用いられている．パンニング角とは，ステレオ音楽信号における左右の音量バランスに基づき，音源が聴取者から見てステレオ空間のどの方向に定位しているかを角度で示したものである．パンニング角の定義と，SpaIn-Net における補助情報としての利用方法を Fig. 2.5 に示す．ステレオ音楽波形とパンニング角 $\boldsymbol{\alpha}$ を SpaIn-Net に同時に入力することで，空間情報を考慮した音源分離が可能となり，各楽器の分離音が出力される．ステレオ音場の正面を 0°，左方向を 90°，右方向を −90° と定義すると，各音源の定位位置は以上の範囲で指定される．ベースが +45°，ドラムが −45° に定位している場合，それぞれ左寄りと右寄りの位置に音源が配置されていることを意味する．このようにパンニング角は，ステレオ音楽信号における空間情報を定量的に表す補助情報として，学習モデルに入力することができる．

ベースラインモデルである SpaIn-Net は，最先端のステレオ音楽音源分離システムである Open-Unmix [23] と CrossNet [24] を統合した XUMX モデル [25] を基盤として構築されて

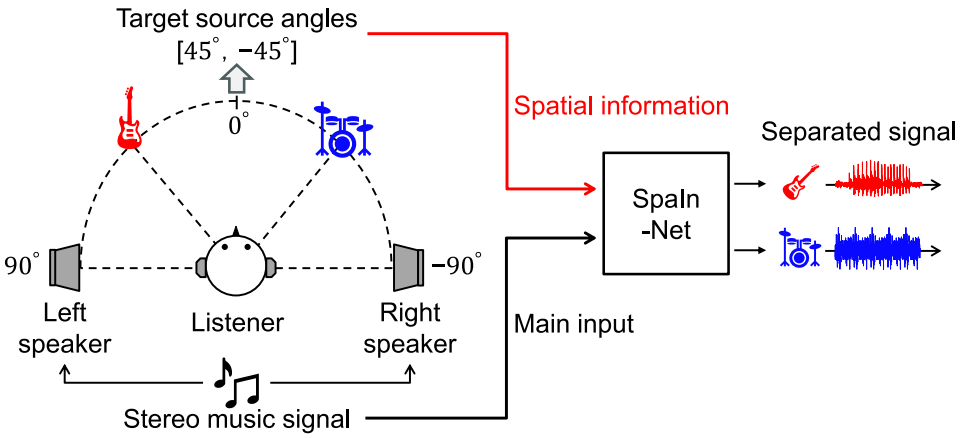


Fig. 2.5. Panning angle definition and its use as auxiliary information in SpaIn-Net.

いる．XUMX と SpaIn-Net のネットワーク構造の概要を Fig. 2.6 に示す．XUMX モデルは，各音源ごとに独立した推論パスを持つ点が特徴である．各サブネットワークは，音源ごとにパラメータが独立した正規化層や BiLSTM 層を有しており，楽器固有の統計的・時間的特徴を個別に学習できるよう設計されている．さらに，CrossNet によって導入された残差接続機構により，平均演算を介した情報交換パスが音源間に追加され，相互補完的に特徴を利用できるようになっている．SpaIn-Net は，以上の XUMX の構造を保持しつつ，位置符号化（positional encoding）によって定位角情報をネットワークへ条件付けすることで，空間情報を直接利用した分離を実現している [26]．補助情報を用いた条件付けの具体的な方法については，次節で詳述する．

2.4.2 角度情報の符号化

前小節で述べたように，SpaIn-Net では各音源 k に対応するパンニング角 $\alpha^{(k)}$ を，補助情報として利用する．パンニング角ベクトルは次式で表される：

$$\boldsymbol{\alpha} = [\alpha^{(1)}, \alpha^{(2)}, \cdots, \alpha^{(K)}] \in [-90, 90]^K \tag{2.11}$$

ここで K はステレオ音楽信号に含まれる音源数である．Fig. 2.6 (b) より，スカラー値の角度情報 $\alpha^{(k)}$ を，ネットワークが処理する高次元のステレオ振幅スペクトログラム $|\boldsymbol{Y}| \in \mathbb{R}_{\geq 0}^{2\times I\times J}$ と統合するため，次元の不一致を解消する必要がある．ここで，各次元はステレオチャネル，周波数ビン，時間フレームを表す．SpaIn-Net では文献 [11] に基づき，各パンニング角度 $\alpha^{(k)}$ を，次式の通り，周波数方向の次元 I に近い次元 D を持つ空間埋め込みベクトルに写像する．

$$\tilde{\boldsymbol{\alpha}}^{(k)} \in [-1, 1]^D \tag{2.12}$$

各次元 $i = 0, 1, \cdots, D-1$ には異なる周波数の sin 関数，cos 関数を適用し，低次元では緩やかな変化，高次元では細かい変化を与えることで，多尺度的に角度を高次元空間へ埋め込む．さらに，ステレオ音場では $-90°$〜$+90°$ の範囲を取るため，負方向の角度を正方向と同じ方

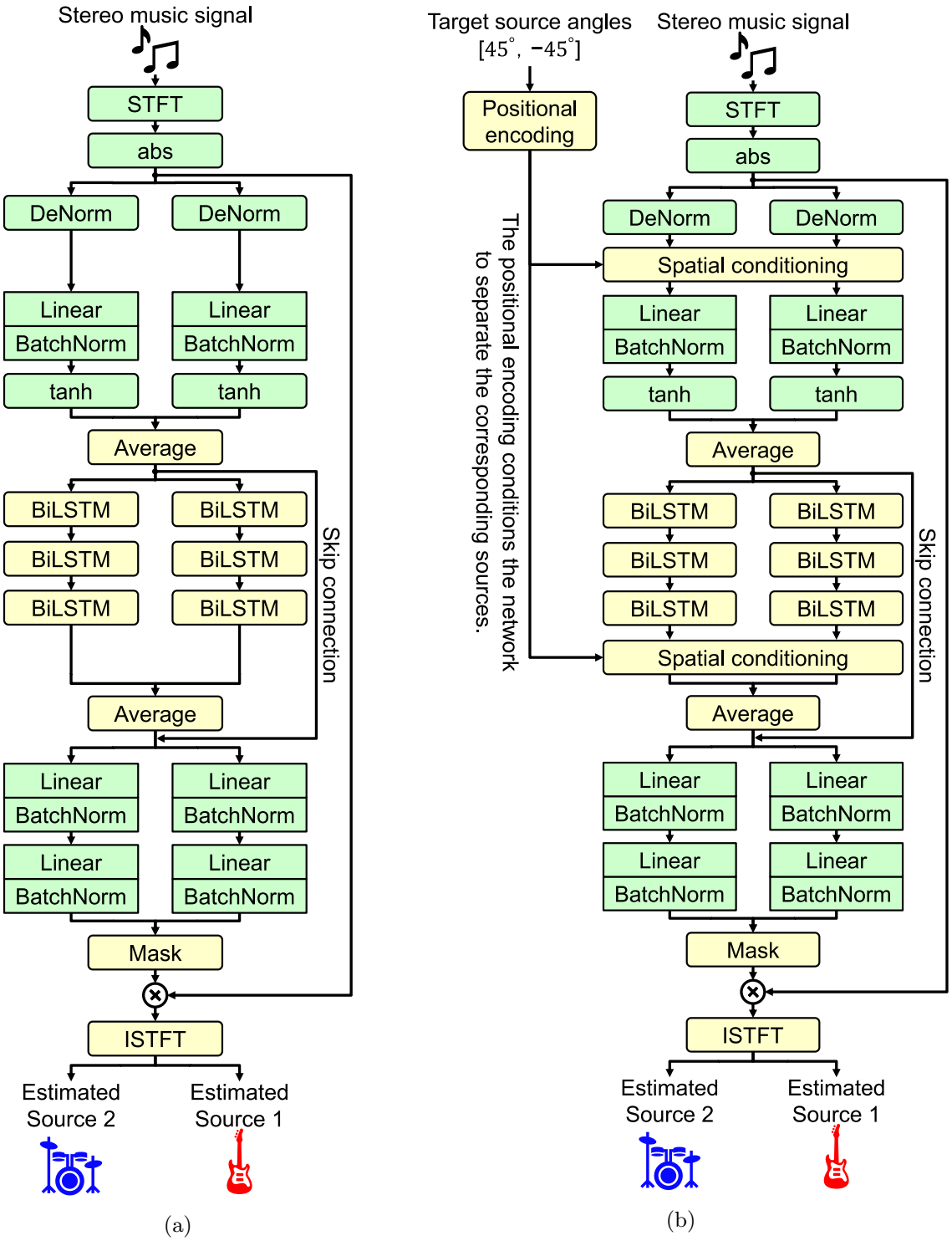


Fig. 2.6. Overview of the baseline XUMX model and the proposed SpaIn-Net: (a) XUMX model, and (b) SpaIn-Net.

法で符号化すると左右対称な表現となる．そこで SpaIn-Net では，正方向と負方向で異なるスケーリング則を用い，左右方向の非対称性を埋め込み空間に反映している．次元を D，インデックスを i（$0 \leq i < D/2$）とすると，偶数次元 $2i$ に sin 関数，奇数次元 $2i+1$ に cos 関数を割り当て，次式で表される．

$$\mathcal{P}(\alpha^{(k)}, 2i) = \sin\left(\frac{|\alpha^{(k)}|}{90^{(D-2i)/D}}\right) \tag{2.13}$$

$$\mathcal{P}(\alpha^{(k)}, 2i+1) = \cos\left(\frac{|\alpha^{(k)}|}{90^{(D-2i)/D}}\right) \tag{2.14}$$

$$\mathcal{M}(\alpha^{(k)}, 2i) = \sin\left(\frac{\alpha^{(k)}}{90^{2i/D}}\right) \tag{2.15}$$

$$\mathcal{M}(\alpha^{(k)}, 2i+1) = \cos\left(\frac{\alpha^{(k)}}{90^{2i/D}}\right) \tag{2.16}$$

これにより，正方向の角度 $\alpha(k) \geq 0$ は，式 (2.13) および (2.14) に示すように，$(D-2i)/D$ によって周期スケールを調整した sin・cos 関数により符号化される．ここで $\mathcal{P}$ は，正弦関数と余弦関数を交互に用いて定義される空間埋め込み関数である．一方，負方向の角度 $\alpha(k) < 0$ は，式 (2.15) および (2.16) に示すように，$2i/D$ を用いた異なる周期スケールで符号化される．ここで $\mathcal{M}$ も同様に，正弦関数と余弦関数を交互に用いて定義されるが，周期スケーリングを反転させることで，左右方向の非対称性が埋め込み空間に反映される．各音源のパンニング角度 $\alpha^{(k)}$ の空間埋め込みベクトル $\tilde{\boldsymbol{\alpha}}^{(k)}$ は次式で表される．

$$\tilde{\boldsymbol{\alpha}}^{(k)} = \begin{cases} [\mathcal{P}(\alpha^{(k)}, 0), \mathcal{P}(\alpha^{(k)}, 1), \cdots, \mathcal{P}(\alpha^{(k)}, D-1)]^{\mathrm{T}}, & \alpha^{(k)} \geq 0 \\ [\mathcal{M}(\alpha^{(k)}, 0), \mathcal{M}(\alpha^{(k)}, 1), \cdots, \mathcal{M}(\alpha^{(k)}, D-1)]^{\mathrm{T}}, & \alpha^{(k)} < 0 \end{cases} \tag{2.17}$$

これらの式によりベクトル $\tilde{\boldsymbol{\alpha}}^{(k)}$ は，左右方向の非対称性が高次元空間に反映される．ベクトル $\tilde{\boldsymbol{\alpha}}^{(k)}$ の一例を Fig. 2.7 に示す．ここで，embedding value は，sin・cos 関数により生成された空間埋め込みベクトルの各要素の値（$[-1, 1]$）を示している．これは音響スペクトルやエネルギー分布を表すものではなく，角度情報を高次元空間に写像した埋め込み表現の可視化である．角度 $\alpha^{(k)}$ が小さい場合，各次元の変化は緩やかであり，低次元成分に主に情報が反映されることがわかる．一方，角度 $\alpha^{(k)}$ が大きい場合，高次元成分にも細かい周期的変化が生じ，より精密な方向情報が符号化される．また，正方向 $\alpha^{(k)} \geq 0$ と負方向 $\alpha^{(k)} < 0$ では，sin・cos のスケーリング則が異なるため，埋め込みベクトルが左右で反転した形状となり，左右方向の非対称性が表現されていることが確認できる．

空間情報とスペクトル情報を組み合わせる手法は文献 [11] で 3 種類提案されている．結合，加算，そして適応インスタンス正規化（adaptive instance normalization: AdaIN）である．まず結合の場合，ステレオ音楽信号から得られた振幅スペクトログラム $|\boldsymbol{Y}| \in \mathbb{R}_{\geq 0}^{2\times I\times J}$ をスタックし，$|\tilde{\boldsymbol{Y}}| \in \mathbb{R}_{\geq 0}^{2I\times J}$ を得る．このとき，周波数ビンの上半分と下半分がそれぞれ左チャネルおよび右チャネルの振幅スペクトログラムに対応する．空間埋め込みベクトル $\tilde{\boldsymbol{\alpha}}^{(k)}$ を周波数方向 I に繰り返し結合することで，空間埋め込み行列 $\boldsymbol{A}^{(k)}$ が構成される．

$$\boldsymbol{A}^{(k)} = [\tilde{\boldsymbol{\alpha}}^{(k)}, \cdots, \tilde{\boldsymbol{\alpha}}^{(k)}]^{\mathrm{T}} \in \mathbb{R}^{D\times J} \tag{2.18}$$

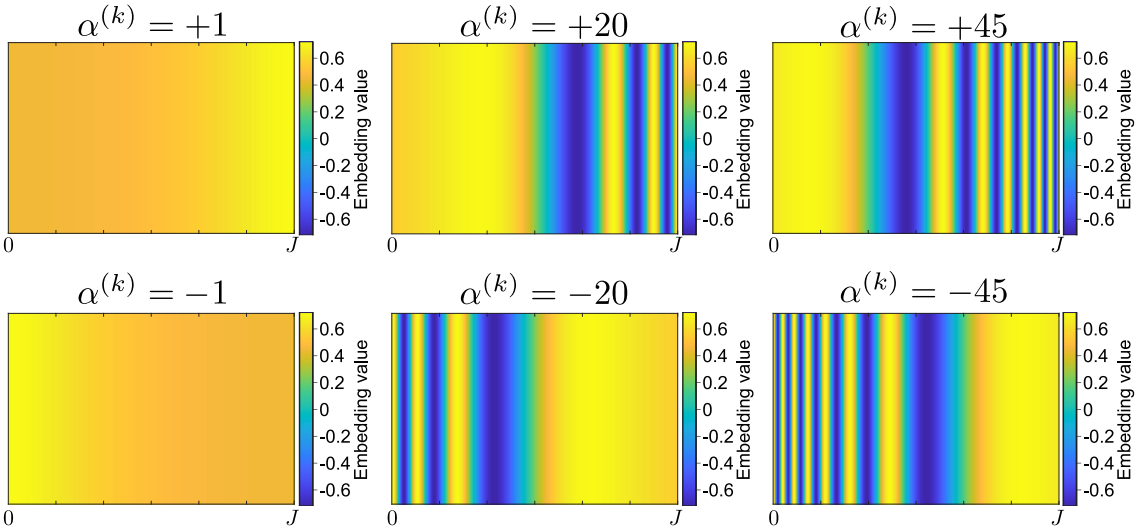


Fig. 2.7. Examples of spatial embedding vectors for different panning angles.

これにより，k 番目の条件付け入力ベクトルは次式となり，k 番目の音源に条件付けが可能となる．

$$\left[|\tilde{\boldsymbol{Y}}|^{\mathrm{T}} \ \boldsymbol{A}^{(k)\mathrm{T}} \right]^{\mathrm{T}} \in \mathbb{R}^{(2I+D)\times J} \tag{2.19}$$

次に加算の場合，次元の整合をとるために時間・周波数次元が一致するように $D = 2I$ を満たす必要がある．

$$|\tilde{\boldsymbol{Y}}| + \boldsymbol{A}^{(k)} \in \mathbb{R}^{2I\times J} \tag{2.20}$$

最後に，AdaIN について述べる．AdaIN は元々画像のスタイル転送で提案 [27] され，あるコンテンツ特徴を目標とするスタイル特徴に統計的に整合させることを目的とする．文献 [11] では，空間情報をコンテンツ特徴 $\tilde{\boldsymbol{\alpha}}^{(k)}$，スペクトル情報をスタイル特徴 $|\tilde{\boldsymbol{Y}}|$ と見なす．AdaIN の目的は，次式のように平均と標準偏差を揃えることである．

$$\mathrm{AdaIN}(|\tilde{\boldsymbol{Y}}|, \boldsymbol{A}^{(k)}) = \sigma(\boldsymbol{A}^{(k)}) \odot \frac{|\tilde{\boldsymbol{Y}}| - \mu(|\tilde{\boldsymbol{Y}}|)}{\sigma(|\tilde{\boldsymbol{Y}}|)} + \mu(\boldsymbol{A}^{(k)}), \tag{2.21}$$

ここで各フレーム j において，位置埋め込み $\tilde{\boldsymbol{\alpha}}^{(k)}$ の平均と分散を振幅スペクトログラムの平均と分散に揃える．

2.4.3　学習時の損失関数

本項では，学習時に使用する損失関数について述べる．時間領域および周波数領域の双方において損失を計算する multi-domain loss (MDL) を導入する．損失関数の計算において信号はモノラルとして扱うため，定義もモノラルとする．モノラル音楽信号の時間波形 $\boldsymbol{x} \in \mathbb{R}^{L}$ は

次式で定義する．

$$\boldsymbol{x} = [x(1),\ x(2), \cdots,\ x(l), \cdots,\ x(L)]^{\mathrm{T}} \in \mathbb{R}^{L} \tag{2.22}$$

モノラル音楽信号は K 個の音源信号 $\boldsymbol{s}_k \in \mathbb{R}^{L}$ から構成される．各音源の信号を次式で定義する．

$$\boldsymbol{s}_k = [s_k(1),\ s_k(2), \cdots,\ s_k(l), \cdots,\ s_k(L)]^{\mathrm{T}} \in \mathbb{R}^{L} \tag{2.23}$$

ここで，$k = 1, 2, \cdots, K$ は音源のインデックスを示す．ネットワークにより推定される信号 $\hat{\boldsymbol{s}}_k \in \mathbb{R}^{L}$ も同様に次式で定義する．

$$\hat{\boldsymbol{s}}_k = [\hat{s}_k(1),\ \hat{s}_k(2), \cdots,\ \hat{s}_k(l), \cdots,\ \hat{s}_k(L)]^{\mathrm{T}} \in \mathbb{R}^{L} \tag{2.24}$$

音源信号 $\boldsymbol{s}_k$ および推定信号 $\hat{\boldsymbol{s}}_k$ の振幅スペクトログラムは $\boldsymbol{S}_k^{(\mathrm{TF})} \in \mathbb{R}_{\geq 0}^{I \times J}$ および $\hat{\boldsymbol{S}}_k^{(\mathrm{TF})} \in \mathbb{R}_{\geq 0}^{I \times J}$ と定義する．MDL は時間周波数領域の平均二乗誤差 (MSE) と，時間領域の加重 SDR (weighted signal-to-distortion ratio: wSDR) を組み合わせた損失であり，次式で定義される．

$$\mathcal{L}_{\mathrm{MDL}} = \mathcal{L}_{\mathrm{MSE}} + r\mathcal{L}_{\mathrm{wSDR}} \tag{2.25}$$

ここで r は 2 種類の損失を混合するための重み係数である．

時間周波数領域の損失関数について述べる．振幅スペクトログラム $\boldsymbol{S}_k^{(\mathrm{TF})}$ およびその推定値 $\hat{\boldsymbol{S}}_k^{(\mathrm{TF})}$ の成分 $s_{ijk}^{(\mathrm{TF})}$ および $\hat{s}_{ijk}^{(\mathrm{TF})}$ を用いて，時間周波数領域の平均二乗誤差（mean squared error: MSE）を次式で定義する．

$$\mathcal{L}_{\mathrm{MSE}} = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{i=1}^{I} \left(s_{ijk}^{(\mathrm{TF})} - \hat{s}_{ijk}^{(\mathrm{TF})} \right)^2 \tag{2.26}$$

時間領域の損失関数について述べる．音源信号 $\boldsymbol{s}_k$，推定信号 $\hat{\boldsymbol{s}}_k$，および混合信号 $\boldsymbol{x}$ を用いて wSDR は次式で与えられる．

$$\mathcal{L}_{\mathrm{wSDR}} = \sum_{k=1}^{K} \left\{ -\rho_k \frac{\langle \boldsymbol{s}_k, \hat{\boldsymbol{s}}_k \rangle}{\|\boldsymbol{s}_k\| \, \|\hat{\boldsymbol{s}}_k\|} - (1 - \rho_k) \frac{\langle \boldsymbol{x} - \boldsymbol{s}_k,\ \boldsymbol{x} - \hat{\boldsymbol{s}}_k \rangle}{\|\boldsymbol{x} - \boldsymbol{s}_k\| \, \|\boldsymbol{x} - \hat{\boldsymbol{s}}_k\|} \right\} + 1 \tag{2.27}$$

wSDR で用いる重み ρ_k は，音源信号 $\boldsymbol{s}_k$ のエネルギーと正解以外の成分 $\boldsymbol{x} - \boldsymbol{s}_k$ のエネルギー比に基づき次式で定義される．

$$\rho_k = \frac{\|\boldsymbol{s}_k\|^2}{\|\boldsymbol{s}_k\|^2 + \|\boldsymbol{x} - \boldsymbol{s}_k\|^2} \in [0, 1] \tag{2.28}$$

音源信号のエネルギーが支配的な区間では ρ_k が大きくなり，推定信号と音源信号の相関をより重視する．正解以外の成分が多い区間では ρ_k が小さくなり，残差の影響を抑制する効果が強まる．wSDR 損失 $\mathcal{L}_{\mathrm{wSDR}}$ は，信号間の相関係数に基づく疑似的な SDR 損失として機能する．第一項は推定信号 $\hat{\boldsymbol{s}}_k$ と音源信号 $\boldsymbol{s}_k$ との相関を評価しており，推定が正解に近いほど値が小さくなる．第二項は推定信号と正解以外の成分，すなわちモノラル音楽信号 $\boldsymbol{x}$ から音源信号 $\boldsymbol{s}_k$ を引いた残余信号との相関を評価しており，正解以外の成分が推定信号に混ざらないように抑制する役割を持つ．このようにして，wSDR は時間的に変化する信号の特性に応じて，音源信号と残差の重要度を柔軟に調整することができる．

2.5　本章のまとめ

本章では，本論文で提案する音源分離手法に関連する基礎技術および先行研究について述べた．まず，時間周波数解析の基本技術である STFT について述べた．次に，提案手法に用いる FiLM の概要を紹介した．さらに，本論文で扱う問題に関連する先行研究を整理し，後の実験での比較手法として活用する知見をまとめた．これらの技術と知見を基礎として，次章では提案手法の構成と具体的なアプローチについて詳述する．

第 3 章

提案手法

3.1　はじめに

前章では，提案する音源分離手法の基礎技術および関連研究について述べた．本章では，本論文で提案するステレオ音楽音源分離手法の詳細について説明する．まず，3.2 節で提案手法の全体像を示す．続く 3.3 節では，提案手法における補助情報となる方位分離の手法について述べる．3.4 節では，DNN モデルの入力・出力形式および学習用データセットの生成方法を解説する．3.5 節では，DNN に FiLM を適用したモデル構造について詳述する．3.6 節では，学習に用いる損失関数を導出する．最後に，3.7 節で本章の内容をまとめる．

3.2　提案手法の全体像

本論文では，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推定される粗い空間手がかりを事前に抽出することで，DNN が直接扱うべき学習負荷を大幅に軽減することを目的としている．この事前処理により，入力信号中に含まれる主要音源のおおまかな方向性があらかじめ分離され，DNN はその空間的手がかりを利用することで，より効率的かつ安定した音源分離を実現できる．ここで述べる方位情報とは，ステレオ音楽信号の左右音量差に基づいて生成される方位分離信号を指す．これらは，各方向に優勢な音源成分を強調した中間表現として機能する．本節ではその概略のみについて述べ，方位分離信号の具体的な生成手法については次節で詳述する．本手法では，方位分離で得られた方位分離信号を単に DNN の入力として与えるのではなく，ネットワーク内部の特徴抽出過程に反映させることで，より方向性に整合した分離を達成する．Fig. 3.1 に示すとおり，方位分離によって得られた方位分離信号を DNN 内部で利用するために，FiLM による条件付け機構を導入している．FiLM は，入力特徴に対してチャネルごとのスケーリングおよびシフトを付与することで，方位情報に応じた特徴強調を実現する手法である．

本手法の特徴として，分離の出力チャネル数が音源数ではなく，方位分離によって生成されるチャネル数により決定される点が挙げられる．すなわち，入力中に複数音源が混在していて

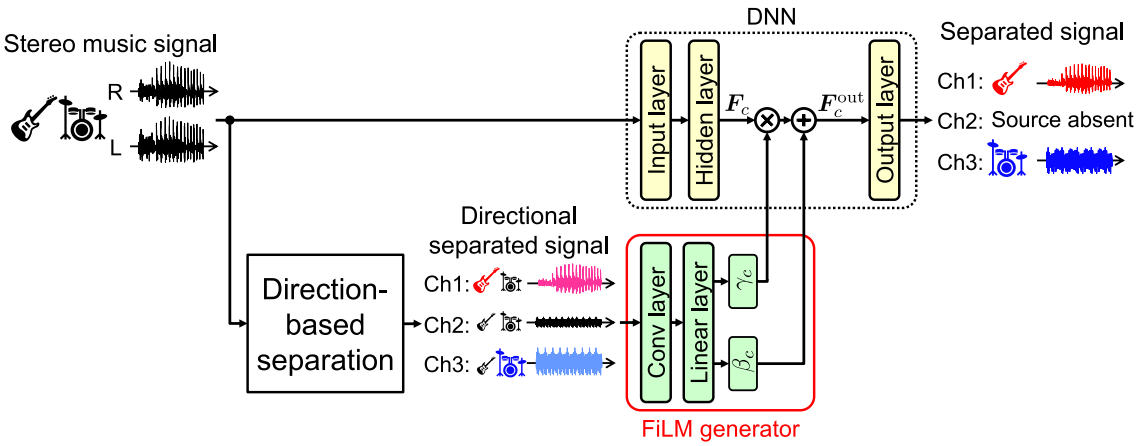


Fig. 3.1. Overview of the proposed method.

も，方位ごとに定義されたチャネル数が N であれば，常に N 個の出力が生成され，各チャネルにはその方向で支配的となる音源成分が配置される．ある方位に優勢な音源が存在しない場合には，対応するチャネルはほとんど音成分を含まない出力となる．この構造により，異なる方位チャネル間で同一音源が重複して出力されることを避けることができる．さらに，本手法はステレオ音楽信号内に実際にいくつ音源が存在するかに依存せずに学習や推論を行える点で，従来手法とは異なる特徴を有する．従来の音源分離モデルの多くは，音源数をあらかじめ固定し，その数に合わせてネットワークの出力次元を設計する必要があった．そのため，想定より多い音源が入力された場合には分離精度が低下し，音源数が少ない場合には空チャネルが生じるなど，柔軟性に欠けるという問題があった．提案手法では方位に基づくチャネル設計を出力の基準としているため，学習時に音源数を固定する必要がなく，実際の音源数が変動する一般的な音楽コンテンツに対しても安定した処理が可能である．もちろん，方位チャネル数 N が混合音源数 K より小さい場合には，同一方向に位置する複数音源が同一チャネルに混在する可能性はある．しかし，現代の商用音楽作品で一般的に混在するパート数を十分に上回るように方位分割数を設定することで，この問題は実質的に緩和され，ほとんどの実用シナリオで有効に機能する．

以上のように，方位に基づくチャネル依存構造を採用することで，従来の音源数依存型手法にはない柔軟性を実現するとともに，事前処理段階で空間的手がかりを付与することで DNN に求められる表現学習の負担を軽減し，限られた学習データでも高い分離性能を達成できる点が本手法の利点である．

3.3　方位分離

人間は，自然環境における音源の方向を，左右耳に到達する音の差異に基づいて推定する．特に，頭部や耳介の影響によって生じる時間差および音量差（interaural level difference: ILD）が主要な手がかりとなり，知覚される音像の方向が決定される．このうち，高域成分を

多く含む音源や人工的にミックスされた音源では，ILD が定位に強く寄与することが知られている．ステレオ音楽信号においては，これらの聴覚的手がかりを人工的に再現するため，左右チャネルの音量比を操作することで音像の方位を制御する．一般的な楽曲制作においては，モノラル音楽信号に対して Fig. 3.2 に示すように左右チャネルの音量比を調整し，左前方へ定位させたい音源には左チャネルを大きく，右チャネルを小さく設定する．右前方への定位ではこの関係が反転する．この操作はパンニングと呼ばれ，サイン則やタンジェント則などの法則に基づいて左右チャネルの音量比が決定される [28]．本論文ではデータ作成の観点から，最も一般的に用いられるサイン則に基づくモデルを採用する．

モノラル音源信号に対してパンニング係数を適用することでステレオ音楽信号を生成する手法について述べる．各音源 k に対して左右チャネルのゲインをサイン則に基づき定義し，そのゲインをモノラル音楽信号に適用することでステレオ波形を構成する．サイン則に基づく左右チャネルのゲインは次式のように定義する．

$$g_k^{(\mathrm{L})} = \cos\left(\frac{\pi}{4}\left(1+\frac{\theta_k}{90}\right)\right) \tag{3.1}$$

$$g_k^{(\mathrm{R})} = \sin\left(\frac{\pi}{4}\left(1+\frac{\theta_k}{90}\right)\right) \tag{3.2}$$

ここで，$\theta_k \in [-90, 90]$ は音源 k の定位方向を表すパンニング角であり，$k = 1, 2, \cdots, K$ は音源のインデックスを示す．次に，これらのゲインをモノラル音楽信号に適用することで，左右チャネルからなるステレオ音楽信号を構成する．音源 k のモノラル音楽信号を $\boldsymbol{x}_k \in \mathbb{R}^L$ とすると，ステレオ音楽信号 $\boldsymbol{x}_k^{(\mathrm{L})}$ および $\boldsymbol{x}_k^{(\mathrm{R})}$ は次式で与えられる．

$$\boldsymbol{x}_k^{(\mathrm{L})} = g_k^{(\mathrm{L})}\,\boldsymbol{x}_k \tag{3.3}$$

$$\boldsymbol{x}_k^{(\mathrm{R})} = g_k^{(\mathrm{R})}\,\boldsymbol{x}_k \tag{3.4}$$

ここで，L は信号長を表し，$\boldsymbol{x}_k^{(\mathrm{L})}$ および $\boldsymbol{x}_k^{(\mathrm{R})}$ はそれぞれ左チャネルおよび右チャネルの波形である．このとき，音源を左方向に定位させたい場合には $g_k^{(\mathrm{L})}$ が大きく，右方向では $g_k^{(\mathrm{R})}$ が大きくなる．ステレオ音楽信号に含まれる各音源成分は，左右チャネル間の相対的な音圧レベルに応じて，特定方向に定位知覚される構造を備えている．ステレオ音楽信号の左右チャネルの信号 $\boldsymbol{x}^{(\mathrm{L})}$ および $\boldsymbol{x}^{(\mathrm{R})}$ は次式で定義される．

$$\boldsymbol{x}^{(\mathrm{L})} = \sum_{k=1}^{K} \boldsymbol{x}_k^{(\mathrm{L})} \tag{3.5}$$

$$\boldsymbol{x}^{(\mathrm{R})} = \sum_{k=1}^{K} \boldsymbol{x}_k^{(\mathrm{R})} \tag{3.6}$$

ステレオ音楽信号 $\boldsymbol{X} \in \mathbb{R}^{L\times 2}$ は，左右チャネルを列ベクトルとして並べた行列として次式で定義される．

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}^{(\mathrm{L})} & \boldsymbol{x}^{(\mathrm{R})} \end{bmatrix} \in \mathbb{R}^{L\times 2} \tag{3.7}$$

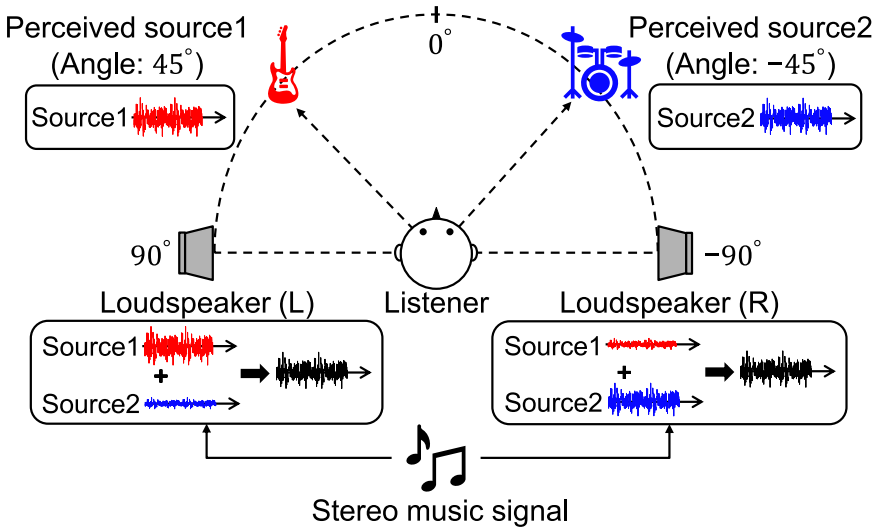


Fig. 3.2. Conceptual diagram of sound source localization based on ILD.

次に，ステレオ音楽信号 $\boldsymbol{X}$ を用いて，各方位に対応する方位分離信号を生成する処理について述べる．方位分離の基本的な考え方は，左右チャネル間の音量比に基づき，各時間周波数成分の方向角 $\phi_{ij} \in [-90°, 90°]$ を推定し，特定の方向に対応する成分のみを抽出することである．まず，ステレオ音楽信号 $\boldsymbol{X}$ に対して短時間フーリエ変換を適用し，複素スペクトログラム $\tilde{\boldsymbol{X}} \in \mathbb{C}^{I \times J \times 2}$ を得る．$\tilde{\boldsymbol{X}}$ の左右チャネルにおける時間周波数成分を $\tilde{x}_{ij}^{(\mathrm{L})}$ および $\tilde{x}_{ij}^{(\mathrm{R})}$ とすると，ϕ_{ij} は，左右チャネルの ILD に基づき，サイン則の逆写像として次式により推定する．

$$\phi_{ij} = \arcsin \frac{\left|\tilde{x}_{ij}^{(\mathrm{L})}\right| - \left|\tilde{x}_{ij}^{(\mathrm{R})}\right|}{\left|\tilde{x}_{ij}^{(\mathrm{L})}\right| + \left|\tilde{x}_{ij}^{(\mathrm{R})}\right|} \tag{3.8}$$

ここで $\left|\tilde{x}_{ij}^{(\mathrm{L})}\right|$ および $\left|\tilde{x}_{ij}^{(\mathrm{R})}\right|$ は，それぞれ周波数ビン i，時間フレーム j における左右チャネルの振幅スペクトログラムを表す．式 (3.8) は左右振幅差を $\left|\tilde{x}_{ij}^{(\mathrm{L})}\right| + \left|\tilde{x}_{ij}^{(\mathrm{R})}\right|$ で正規化することにより，ILD を $[-1, 1]$ に写像し，arcsin を適用することでパンニング角 $[-90°, 90°]$ を一意に定める．

推定された方向角 ϕ_{ij} を全時間周波数成分について集計することで，入力信号に含まれる音源の方位分布を表す重み付きヒストグラムを作成できる．このヒストグラムは主に音源の空間配置を視覚的に確認するためのものであり，後述のマスク生成そのものは，方向角 ϕ_{ij} の統計分布をガウス混合モデル（gaussian mixture model: GMM）[29] によって推定することで行う．Fig. 3.3 に，Fig. 3.2 の音源配置を重み付きヒストグラムとして可視化した例を示す．ここでは方位分割数を $N = 5$ とした．このヒストグラム上の特定方向に対応する時間周波数領域の成分を抽出することで，方位ごとの分離信号を生成する．具体的には，各方向領域に対応するマスクを適用し，その領域のエネルギー成分のみを残す．逆短時間フーリエ変換を行うことで，時間領域のステレオ形式の方位分離信号が得られる．あらかじめ指定した N 個の分

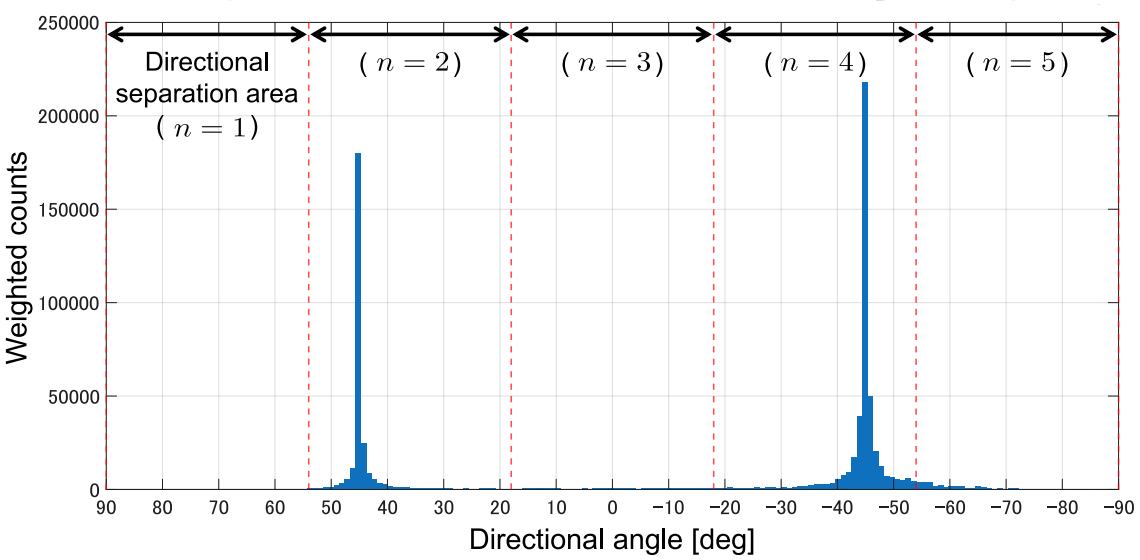


Fig. 3.3. Weighted histogram of direction angles ϕ_{ij} estimated from the stereo mixture.

離方向に対して，時間周波数マスク $\boldsymbol{M}_n \in [0,1]^{I\times J}$ を生成する．$n = 1, 2, \cdots, N$ は方位分離信号のインデックスを表す．得られた方向角 ϕ_{ij} の分布を N 個の方向クラスタに分割するため，GMM により方向角分布を統計的にモデリングする．GMM は各方向クラスタ n を平均 μ_n，分散 σ_n^2 をもつガウス分布として表し，混合比 π_n により各クラスタの寄与度を表す．GMM のパラメータ $\{\pi_n, \mu_n, \sigma_n^2\}$ は方向角 ϕ_{ij} に対して EM アルゴリズム [29] により推定される．GMM に基づき，マスク値は次式で与えられる．

$$m_{ijn} = \frac{\pi_n \mathcal{N}(\phi_{ij}; \mu_n, \sigma_n^2)}{\sum_{n'=1}^{N} \pi_{n'} \mathcal{N}(\phi_{ij}; \mu_{n'}, \sigma_{n'}^2)} \tag{3.9}$$

ここで $\mathcal{N}(\phi_{ij}; \mu_n, \sigma_n^2)$ は平均 μ_n，分散 σ_n^2 のガウス分布の値を表す．m_{ijn} は，各方向クラスタ n に属するソフトマスクであり，全クラスタにわたる総和が 1 になるという特徴をもつ．この性質から，任意の時間周波数成分に対して，次式が常に成り立つ．

$$\sum_{n=1}^{N} m_{ijn} = 1 \tag{3.10}$$

各方向 n の分離信号のスペクトログラム $\hat{\boldsymbol{X}}_n \in \mathbb{C}^{I\times J\times 2}$ は次式で表される．

$$\hat{\boldsymbol{X}}_n = \boldsymbol{M}_n \odot \tilde{\boldsymbol{X}} \tag{3.11}$$

ここで $\odot$ は要素ごとの積を表す．マスクにより分離されたスペクトログラム $\hat{\boldsymbol{X}}_n$ に対して逆短時間フーリエ変換を適用することで，各方向に対応するステレオ音楽信号が得られる．さらに，このステレオ音楽信号から左右チャネルを統合してモノラル化した波形 $\hat{\boldsymbol{x}}_n \in \mathbb{R}^L$ を方位分離信号として定義する．本論文では，この方位分離信号を後段のネットワークに与える補助情報として利用する。この処理により，各音源の定位方向に対応する時間周波数成分が相対的に強調された方位分離信号が得られる．

3.4　DNN の入出力とデータセット生成

本節では，提案手法で用いる DNN の入出力構造と，学習に必要なデータセットの一般的な生成方法について述べる．Fig. 3.1 に示すように，本手法の入力はステレオ音楽信号 $\boldsymbol{X}$ および $\boldsymbol{X}$ に基づいて得られるモノラル化した方位分離信号 $\hat{\boldsymbol{x}}_n$ である．DNN はこれらを入力として受け取り，各方位に対応する推定信号 $\hat{\boldsymbol{s}}_n \in \mathbb{R}^L$ を推定する．DNN の出力である推定信号 $\hat{\boldsymbol{s}}_n$ は，対応する正解信号 $\boldsymbol{s}_n \in \mathbb{R}^L$ と比較することで学習される．正解信号および推定信号はそれぞれ次式で定義する．

$$\boldsymbol{s}_n = [s_n(1),\ s_n(2),\ \cdots,\ s_n(L)]^{\mathrm{T}} \in \mathbb{R}^L \tag{3.12}$$

$$\hat{\boldsymbol{s}}_n = [\hat{s}_n(1),\ \hat{s}_n(2),\ \cdots,\ \hat{s}_n(L)]^{\mathrm{T}} \in \mathbb{R}^L \tag{3.13}$$

学習に用いるステレオ音楽信号は，複数の単一音源から構成される公開音楽データセットを用いて生成することを想定する．一般に，音楽データセットに含まれる各音源はトラックごとに信号長が異なるため，学習に適した一定長のサンプルを得る目的で，時間方向に固定長で切り出す前処理を行う．この切り出し処理は，Fig. 3.4 に示すように，入力波形を時間方向に一定長で抽出する操作である．一定のシフト幅で移動させながら抽出するため，隣接するサンプル間には時間的な重なりが生じる．切り出した各単一音源に対しては，所定の方位分割数 N を仮定し，各音源をいずれかの方位領域に割り当てることで空間配置を模擬する．その概念図を Fig. 3.5 に示す．方位角の取り得る範囲や分割方法は想定する収録条件や応用に応じて柔軟に設定可能であり，本手法は特定の角度範囲に依存しない．割り当てられた方位領域内において，定位角度は所定の分布に従って決定され，対応する左右チャネルのゲインはサイン則に基づくパンニングにより算出される．学習データにおける方位と音源の対応関係を明確にするため，1 つの方位領域には高々 1 種類の音源のみが配置されるよう制約を設ける．以上の制約を満たすため，方位分割数 N は音源数 K 以上とし，$N \geq K$ が成り立つように設定する．この制約により，各方位に対応する正解信号が一意に定まり，DNN が方位ごとの音源分離を学習しやすくなる．

以上の処理を各音源に対して行い，時間波形上で加算することで，複数の音源が異なる方位から到来する状況を模擬したステレオ音楽信号 $\boldsymbol{X}$ を生成する．生成されたステレオ音楽信号に対して方位分離処理を適用することで，N 個の方位分離信号 $\hat{\boldsymbol{x}}_n$ が得られる．DNN が出力すべき正解信号 $\boldsymbol{s}_n$ は，各方位領域に配置された単一音源の混合前モノラル信号とする．該当する方位に音源が存在しない場合には，全サンプル値が 0 の無音信号を正解として割り当てる．

Fig. 3.6 は，上述したデータセット生成手順を特定の方位分割数および音源配置条件のもとで適用した一例を示している．本図は処理の流れを視覚的に理解することを目的としたものであり，方位角の範囲，分割数 N，および音源数は応用や実験設定に応じて変更可能である．

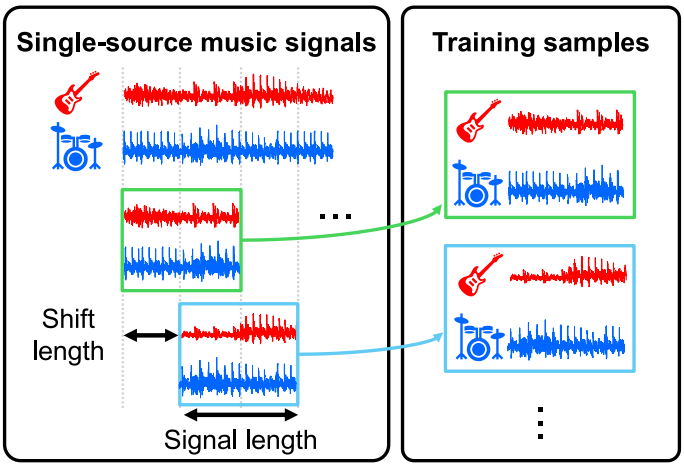


Fig. 3.4. Segmentation of single-source audio into fixed-length training samples.

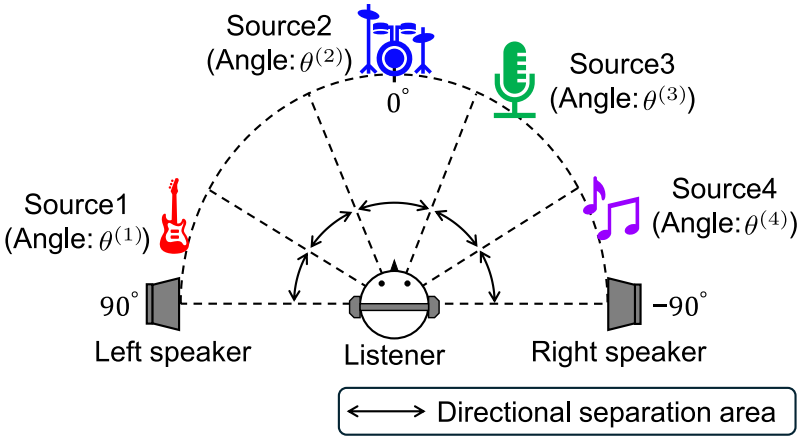


Fig. 3.5. Division of the azimuth range into N direction sectors and the random assignment of each source to a sector.

3.5　DNN の構造

本論文では，波形を直接入力し波形を出力する end-to-end 型の DNN を用いる．基盤モデルとして，モノラル入力に対して高い分離性能を示す完全畳み込み時間領域音声分離ネットワーク（fully-convolutional time-domain audio separation network: Conv-TasNet）[7] を採用する．さらに，本論文ではその拡張版であり，ステレオ入力を扱うためにチャネル間の相関を利用可能とした inter-channel Conv-TasNet（IC Conv-TasNet）[30] を基本構造として用いる．

本手法では，メイン入力としてステレオ音楽信号を IC Conv-TasNet に与えるとともに，事前処理によって得られた方位分離信号 $\hat{\boldsymbol{x}}_n$ を補助情報としてネットワーク内部に組み込む．補助情報は，ネットワーク中間層に追加した FiLM ジェネレータを介してメイン入力の特徴量を

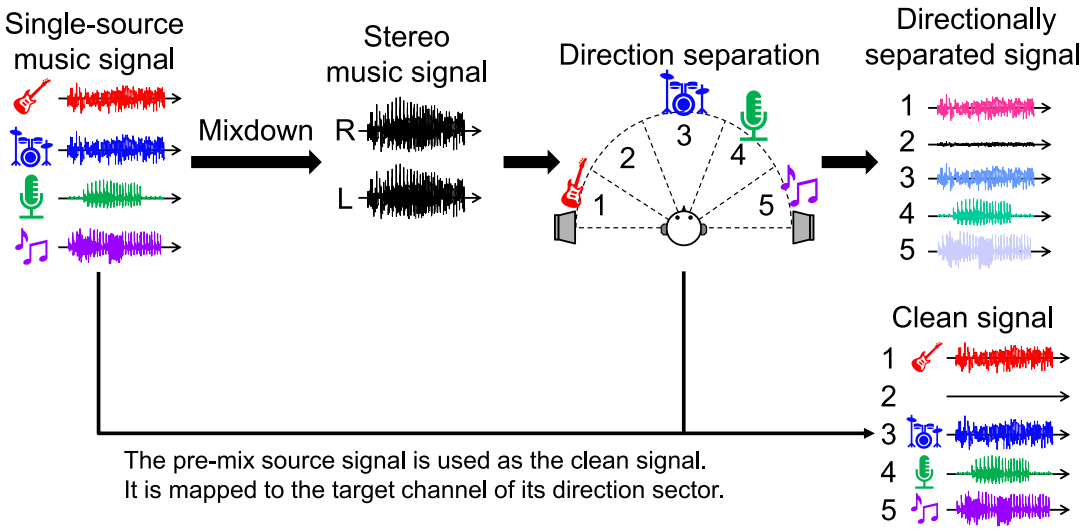


Fig. 3.6. Summary diagram of stereo music signal, direction separated signals, and clean signals.

条件付けする形で利用される．これにより，DNN は方位分離信号に基づく空間的な条件付けを通じて，方位ごとの音源分離を効率的に学習可能となり，高精度な音源分離が期待される．本手法で用いるネットワーク構造の概要を Fig. 3.7 に示す．エンコーダ，時系列畳み込みネットワーク（temporal convolutional network: TCN）ブロック，およびデコーダから構成される IC Conv-TasNet の基本的な処理流れに加え，補助情報を注入する FiLM ジェネレータの動作を示している．FiLM ジェネレータが生成する $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ は，各 TCN ブロックに対して個別の調整量として設定される．

本論文のモデルでは，N 個の方位領域に対応する波形推定を行い，出力 $\hat{\boldsymbol{s}}_n$ を最終的な推定信号とする．本論文ではデータセット生成時に「1 方位領域には高々 1 種類の音源のみが存在する」という制約を課しているため，推定信号 $\hat{\boldsymbol{s}}_n$ は，対応する方位に配置された単一音源の推定結果，あるいは音源が存在しない場合には無音信号を表す．本手法で扱う音源分離問題においては，推定信号 $\hat{\boldsymbol{s}}_n$ の和が，ステレオ音楽信号 $\boldsymbol{X}$ をモノラル化したモノラル音楽信号 $\boldsymbol{x}$ と一致することが望ましい．この条件は混合整合性（mixture consistency）と呼ばれる．音源が存在しない方位領域においても，推定信号 $\hat{\boldsymbol{s}}_n$ が厳密にゼロとなる保証はないため，混合整合性は全方位領域に対応する推定信号の和として定義される．以上を踏まえ，次式を満たすことが期待される．

$$\sum_{n=1}^{N} \hat{\boldsymbol{s}}_n = \boldsymbol{x} \tag{3.14}$$

本論文では，最終的な推定信号 $\hat{\boldsymbol{s}}_n$ が混合整合性を満たすことをモデルの要件とする．しかし，DNN が直接生成する波形は，推定誤差やスケール不変損失の影響により，式 (3.14) を厳密には満たさない場合がある．そこで，混合整合性を満たす最終出力 $\hat{\boldsymbol{s}}_n$ を得るために，DNN の

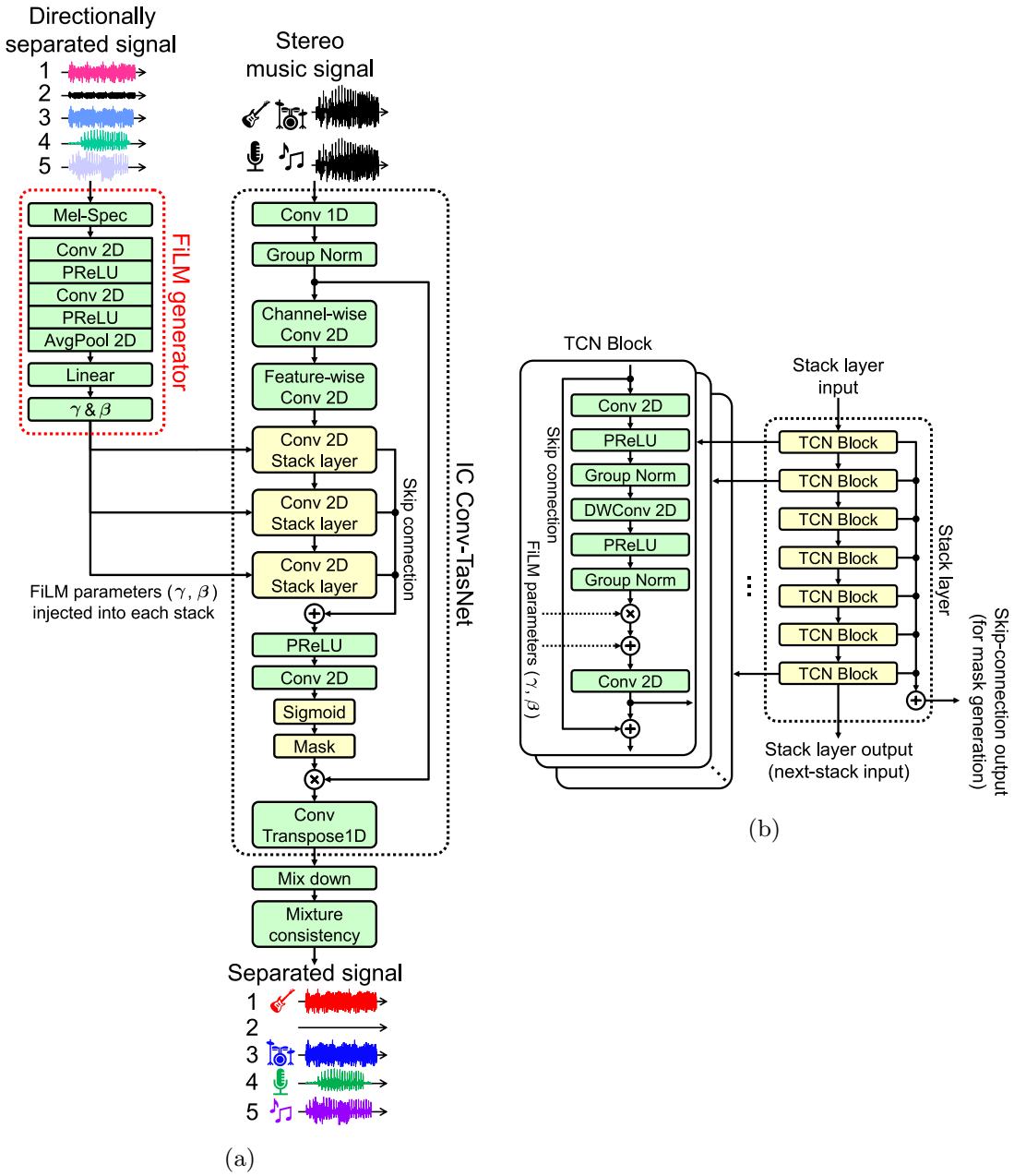


Fig. 3.7. Overview of the proposed network: (a) FiLM-integrated IC Conv-TasNet with stereo input and (b) expanded view of the TCN stack and its constituent TCN blocks.

内部出力として波形信号 $\tilde{\boldsymbol{s}}_n \in \mathbb{R}^L$ を導入し，これに対して混合整合性射影に基づく処理 [31] を適用する．具体的には，$\tilde{\boldsymbol{s}}_n$ に修正量 $\Delta\boldsymbol{s}_n$ を加えることで $\hat{\boldsymbol{s}}_n = \tilde{\boldsymbol{s}}_n + \Delta\boldsymbol{s}_n$ が混合整合性を満たすよう，線形射影処理を行う．まず推定信号の和を次式とする．

$$\boldsymbol{z} = \sum_{n=1}^{N} \tilde{\boldsymbol{s}}_n \tag{3.15}$$

修正量は次式で与えられる．

$$\Delta \boldsymbol{s}_n = \frac{1}{N}\left(\boldsymbol{x} - \boldsymbol{z}\right), \tag{3.16}$$

最終的な出力信号は次式となる．

$$\hat{\boldsymbol{s}}_n = \tilde{\boldsymbol{s}}_n + \frac{1}{N}\left(\boldsymbol{x} - \sum_{n=1}^{N} \tilde{\boldsymbol{s}}_n\right) \tag{3.17}$$

この処理により，最終出力の推定信号 $\hat{\boldsymbol{s}}_n$ の和は，モノラル音楽信号 $\boldsymbol{x}$ と一致するように補正される．

3.6　DNN 学習時の損失関数

本節では，時間領域で用いる評価指標として閾値付き尺度不変信号対雑音比（threshold scale-invariant signal-to-noise ratio: threshold SI-SNR）[32, 33] を定義する．さらに，L_1 ノルム損失を組み合わせ，全体損失 $\mathcal{L}$ を導入する．

Threshold SI-SNR の定義には，推定信号 $\hat{\boldsymbol{s}}_n$ と正解信号 $\boldsymbol{s}_n$ の直交分解に基づくターゲット成分と残差成分を用いる．推定信号 $\hat{\boldsymbol{s}}_n$ が正解信号 $\boldsymbol{s}_n$ をどの程度再現できているかを評価するため，まず $\hat{\boldsymbol{s}}_n$ を $\boldsymbol{s}_n$ 方向の成分とそれ以外の直交成分に分解する．正解信号方向への射影係数 δ は，次式で定義されるスカラー値として表される．

$$\delta = \frac{\langle \hat{\boldsymbol{s}}_n, \boldsymbol{s}_n \rangle}{\|\boldsymbol{s}_n\|_2^2 + \varepsilon} \tag{3.18}$$

ここで，$\|\cdot\|_2$ はベクトルの L_2 ノルムである．分母には数値的安定性を確保するため，小さな正定数 $\varepsilon > 0$ を加えている．射影係数 δ を導入することで，$\hat{\boldsymbol{s}}_n$ に含まれる $\boldsymbol{s}_n$ と同じ方向の成分を厳密に抽出でき，残差成分 $\boldsymbol{s}_{\text{noise}}$ は $\boldsymbol{s}_n$ に直交し，推定誤差のみを表すという利点が得られる．射影係数を用いて，ターゲット成分 $\boldsymbol{s}_{\text{target}}$ および残差成分 $\boldsymbol{s}_{\text{noise}}$ を次式で定義する．

$$\boldsymbol{s}_{\text{target}} = \delta\, \boldsymbol{s}_n \tag{3.19}$$

$$\boldsymbol{s}_{\text{noise}} = \hat{\boldsymbol{s}}_n - \boldsymbol{s}_{\text{target}} \tag{3.20}$$

Threshold SI-SNR では，残差成分が小さい場合に SI-SNR が過剰に大きくなるのを抑制するため，soft-threshold パラメータ τ を導入する．具体的には，デシベル値 a_{th} から次式で定義される．

$$\tau = 10^{a_{\text{th}}/10} \tag{3.21}$$

これらの定義をもとに，threshold SI-SNR を次式で定義する．

$$\mathcal{L}_{\text{SI-SNR}_\tau}(\boldsymbol{s}_n, \hat{\boldsymbol{s}}_n) = -10 \log_{10} \frac{\|\boldsymbol{s}_{\text{target}}\|_2^2}{\|\boldsymbol{s}_{\text{noise}}\|_2^2 + \tau \|\boldsymbol{s}_{\text{target}}\|_2^2 + \varepsilon} \tag{3.22}$$

しかし，提案手法の出力には無音チャネルが含まれることがあり，この場合，正解信号 $\boldsymbol{s}_n$ がゼロとなるため SI-SNR が発散し，学習が不安定になる．そこで，本論文では無音チャネルに対しては SI-SNR に代えて，次式に示す損失関数を用いる．

$$\mathcal{L}_{L_1}(\hat{\boldsymbol{s}}_n) = \|\hat{\boldsymbol{s}}_n\|_1 \tag{3.23}$$

ここで，$\|\cdot\|_1$ はベクトルの L_1 ノルムである．無音チャネルと有音チャネルで損失関数を使い分けるため，各音源 $\boldsymbol{s}_n$ のエネルギーを用いて，無音チャネルと有音チャネルを分類する．無音チャネル集合を $\mathcal{C}_{\text{sil}}$，有音チャネル集合を $\mathcal{C}_{\text{act}}$ とする．

$$\mathcal{C}_{\text{sil}} = \{\, n \in \{1, \cdots, N\} \mid \|\boldsymbol{s}_n\|_2^2 < \varepsilon_s \,\} \tag{3.24}$$

$$\mathcal{C}_{\text{act}} = \{\, n \in \{1, \cdots, N\} \mid \|\boldsymbol{s}_n\|_2^2 \geq \varepsilon_s \,\} \tag{3.25}$$

ここで $\varepsilon_s > 0$ は無音判定のための閾値である．

以上を用いて，全体損失 $\mathcal{L}$ は次式で定義される．

$$\mathcal{L} = \frac{1}{N}\left(\lambda \sum_{n \in \mathcal{C}_{\text{sil}}} \mathcal{L}_{L_1}(\hat{\boldsymbol{s}}_n) + \sum_{n \in \mathcal{C}_{\text{act}}} \mathcal{L}_{\text{SI-SNR}_\tau}(\boldsymbol{s}_n, \hat{\boldsymbol{s}}_n)\right) \tag{3.26}$$

各集合内でチャネルごとの損失を平均化した後，無音チャネルでは平均 L_1 ノルム損失，有音チャネルでは平均 threshold SI-SNR 損失が計算される．各チャネルの損失を平均することで，無音チャネルと有音チャネルの損失を適切に統合しつつ、チャネルごとの寄与比率の偏りを防ぐ．さらに，L_1 ノルム 損失と threshold SI-SNR 損失間のスケール差を補正する目的で，スケーリング係数 λ を導入する．

3.7 本章のまとめ

本章では，本論文で提案するステレオ音楽音源分離手法について述べた．左右チャネル間の音量差に基づく方位分離を事前処理として導入し，その結果を FiLM により DNN 内部に条件付けすることで，方位情報を活用した音源分離モデルを設計した．また，方位チャネル数に基づく出力設計，データセット生成方法，および無音チャネルを考慮した損失関数を示した．次章では，データセットを用いた実験を行い，提案手法の有効性を検証する．

第 4 章

提案手法と比較手法の性能評価実験

4.1　はじめに

前章では，ステレオ混合信号と方位分離信号を併用することで，高精度な音源分離を実現する提案手法の構成について述べた．本章では，提案手法の有効性を検証するため，各種モデルを用いた比較実験を行う．まず，補助情報の有無や種類の違いが分離性能に与える影響を調べるため，複数のモデル構成を設計し，精度・安定性・実用性の観点から評価を行う．続いて，実験条件および評価指標を示し，最後に得られた結果について詳細な分析を行う．

4.2　実験の目的

本章では，提案手法の有効性を検証することを目的として，補助情報の有無や種類が音源分離性能に与える影響を比較する．特に，本論文で提案する FiLM を介した方位分離信号の活用が，既存手法およびベースラインと比較してどの程度の性能向上をもたらすかを明らかにする．

比較対象とするモデル構成は以下の 3 種類である．まず，補助情報を一切用いず，ステレオ音楽信号のみを入力とする IC Conv-TasNet をベースラインとする．提案手法における空間的特徴量の導入が，分離性能向上に本質的に寄与しているかを検証する基準とする．次に，空間情報となる定位角度を特徴量として利用する SpaIn-Net を比較手法と位置付ける．SpaIn-Net は空間的特徴量（定位角）を補助情報として活用する点で提案手法と近しい手法であり，提案手法においてステレオ音楽信号のみから生成可能な空間特徴量が，定位角度の推定に高度な前処理を要し実運用での適用性に課題が残る既存手法と比較して，分離精度の向上に寄与するかを検証する基準とする．本論文では，定位角度はあらかじめ分かっているものとし，理想的な補助情報を与えた際の性能上限を示すものとして位置づける．最後に，提案手法では，ステレオ音楽信号の左右チャネルの音量比から得られる方位分離信号を補助情報とし，FiLM を介して IC Conv-TasNet 内部に適応的に反映させる．この特徴量は取得が容易であるにもかかわらず，分離性能の向上に寄与し得る点に特徴がある．

4.3 実験条件

本節では，提案手法および比較手法の性能評価を行うために用いたデータセット，モデル構成，学習条件，および評価指標について述べる．データセットの作成方法については 3.4 節で詳述したため，本節では学習に用いた音源数構成と学習条件を中心に説明する．

4.3.1 データセット

本実験におけるデータセットおよび実験条件を Table 4.1 に示す．使用する音楽データのドライソースは MUSDB18 [34] に含まれるモノラル音源信号（vocal，bass，drum，および other）とした．各楽曲において，各音源信号を時間方向に 10 秒の固定長で切り出し，パンニング処理を施すことで観測ステレオ音楽信号を生成し，学習，検証，および評価に用いた．この切り出しは，一定のシフト幅で移動させながら行うため，隣接するサンプル間には時間的な重なりが生じる．学習，検証，および評価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．なお，学習データには 100 曲，検証および評価データにはそれぞれ 10 曲の楽曲を用いた．学習・検証・評価データの時間配分については，検証および評価において同程度の信頼性で性能を確認することを目的として，両者をほぼ同規模とし，学習データについては十分な学習量を確保するため，それらのおよそ 10 倍の時間を割り当てた．また，各音源信号の切り出し長を 10 秒とした理由については，音楽信号に含まれるフレーズや音響的変化を十分に含みつつ，計算資源や学習の安定性とのバランスが取りやすい長さとして，経験的に妥当であると判断したためである．

各方位領域には最大 1 つの音源が存在することを仮定しているため，方位分割数 N は音源数 K 以上に設定した．提案手法では，出力チャネル数が音源数ではなく方位分割数 N に依存するため，本来は音源数を固定せずに利用可能である．しかし，本実験では比較条件を統一するために音源数を K に固定した設定で評価を行う．

4.3.2 モデル構成および比較手法

提案手法およびベースラインは，いずれも IC Conv-TasNet を基本骨格として構築した．エンコーダのフィルタ長，TCN ブロック数，チャネル数などのネットワークパラメータは両手法で統一し，モデル容量を揃えること公平な比較を行った．

比較手法である SpaIn-Net は定位角度を外部特徴量として利用する既存手法であり，提案手法とはネットワーク構造が異なる．SpaIn-Net では定位角度の正確な推定を前提とするため，本実験ではデータセット作成時に得られる定位角度の正解値を入力として与え，理想的条件下での上限性能を評価した．一方，提案手法およびベースラインでは定位角度を使用せず，必要に応じて左右チャネルの音量比に基づく補助特徴量のみを利用した．

4.3.3　学習・最適化条件

DNN の最適化には Adam [35] を用い，学習の高速化とメモリ効率向上のために混合精度学習 [36] および累積勾配を導入した．混合精度学習では，推論および勾配計算の一部を 16-bit 浮動小数点数で実行しつつ，学習に直接影響する重みパラメータは 32-bit 精度で保持する方式を採用した．累積勾配では，バッチサイズ 4 の勾配を 4 回分蓄積し，合計 16 バッチ相当の勾配を用いて 1 回の重み更新を行った．これにより，大きなバッチサイズで学習した場合と同等の安定した最適化効果を得つつ，GPU メモリ使用量を抑制できる．損失関数には threshold SI-SNR を採用し，閾値は -30 dB に設定した．学習は最大 200 epochs とし，検証損失が 30 epochs 改善しない場合には早期終了を適用した．音源分離性能の指標として，出力信号と参照信号の全体的な類似度を表す source-to-distortion ratio (SDR) [37] を用いた．提案手法は左右音量比のみを補助情報として用いるため，追加の前処理負荷が極めて小さい．

4.4　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 Figs. 4.1〜4.4 に示す．また，各手法におけるステレオ音楽信号の各音源に対する平均 SDR と，各手法による SDR 改善量（ΔSDR）を Table 4.2 に示す．

ステレオ音楽信号のは全音源で平均 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，全音源平均の ΔSDR 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 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．全音源平均の ΔSDR は 11.43 dB であり，補助情報として定位角度を導入することで，分離性能およ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，および 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．以上の結果から，提案手法はベースラインおよび比較手法の双方に対して一貫して優れた性能を示し，既存手法に対する明確な優位性を有することが確認された．

以上より，提案手法は従来の IC Conv-TasNet および 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．

Table 4.1. Dataset and training parameters for chapter 4 experiments

Parameter	Value
Dataset	
Total train audio duration [s]	32,720
Total validation audio duration [s]	3,130
Total test audio duration [s]	3,650
Signal duration per sample [s]	10
Sampling rate [Hz]	16,000
Model configuration	
Number of sources in stereo mixture	$K = 4$
Number of directionally separated channels	$N = 5$
Number of TCN blocks per stack	8
Number of stacks	3
Mixed-precision training	Enabled (FP16 + FP32)
Training hyperparameters	
Learning rate	1×10^{-4}
Batch size (per update)	4
Gradient accumulation steps	4
Early stopping patience [epochs]	30
Number of training epochs [epochs]	200
Loss function threshold (SI-SNR) [dB]	$a_{\text{th}} = -30$
Numerical stability constant for SI-SNR	$\varepsilon = 1 \times 10^{-8}$
Silence threshold	$\varepsilon_s = 1 \times 10^{-4}$

4.5 本章のまとめ

本章では，方位分離信号を補助情報として用いた音源分離手法の有効性を検証するため，IC Conv-TasNet および SpaIn-Net との比較実験を行った．

全音源平均の ΔSDR に着目すると，IC Conv-TasNet は，粗い分離結果である方位分離信号単体と比較して高い分離性能を示している．この結果は，学習モデルを介することで，方位分離信号に含まれる不完全な空間情報や時間的構造がより効果的に利用されていることを示唆している．比較手法である SpaIn-Net は，理想的な定位角度を補助情報として与えることで，全音源において IC Conv-TasNet を一貫して上回る性能を示している．これに対し，提案手法は，定位角度のような精緻な空間情報を用いることなく，ステレオ音楽信号から容易に得られる方位分離信号を FiLM を介して活用することで，全音源において最も高い ΔSDR を達成

Table 4.2. Average SDR [dB] and SDR improvement (ΔSDR) [dB] 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 (Baseline)		9.69	8.07	8.15	10.66	9.14
SpaIn-Net with oracle angles		11.34	10.79	10.10	13.48	11.43
IC Conv-TasNet+FiLM (Proposed method)		**12.78**	**11.64**	**12.13**	**15.14**	**12.92**

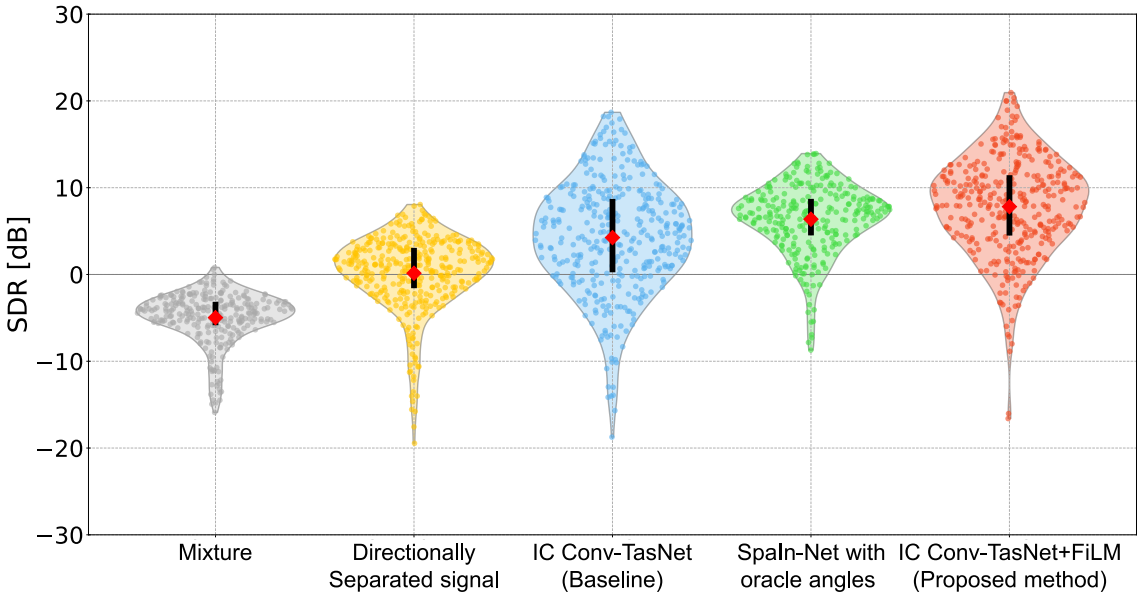


Fig. 4.1. Violin plot of SDR for bass across test data.

している．全音源平均の ΔSDR は 12.92 dB に達しており，比較手法を含む既存手法を上回る結果となった．

以上の結果より，空間情報を明示的な定位角度として与えなくとも，粗い方位分離信号を適切にモデル内部へ統合することで，高精度かつ安定した音源分離が可能であることが示された．特に，FiLM による特徴チャネル方向の条件付けは，空間的手がかりを柔軟に反映する有効な手段であり，追加の前処理を必要としない軽量な構成でありながら，実運用に適した高い性能を実現できる点で有望なアプローチであると結論付けられる．

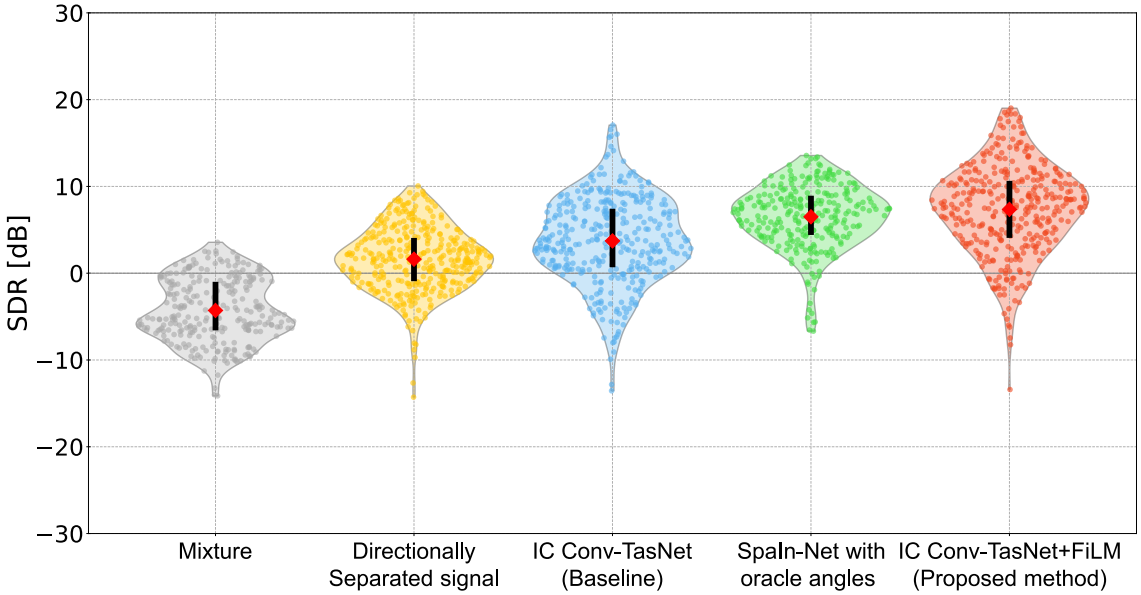


Fig. 4.2. Violin plot of SDR for drum across test data.

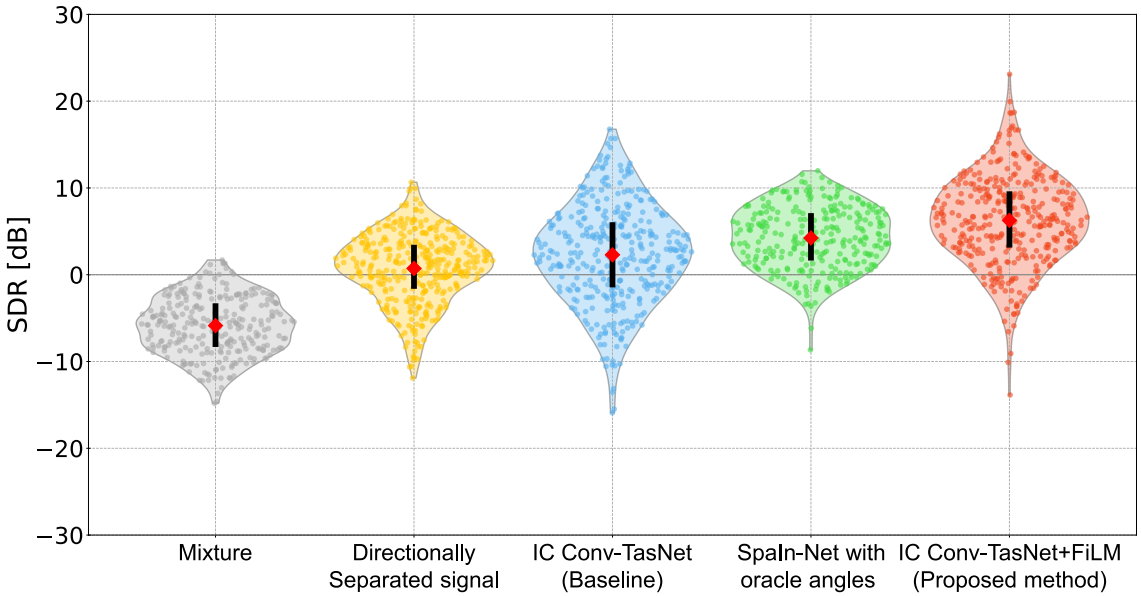


Fig. 4.3. Violin plot of SDR for other across test data.

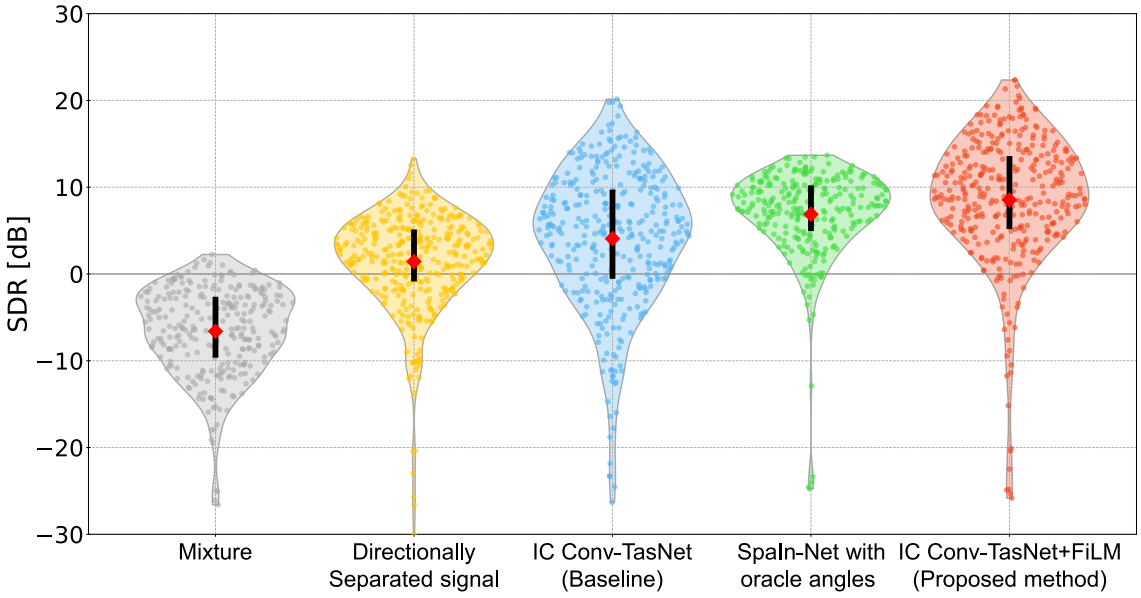


Fig. 4.4. Violin plot of SDR for vocal across test data.

第 5 章

汎化性能の評価実験

5.1　はじめに

前章では，ステレオ混合信号と方位分離信号を併用することで高精度な音源分離を実現する提案手法の構成と性能を示した．しかし，前章で扱った実験は，いずれも学習時と同一の音源数を前提としたデータセットを用いており，実運用で想定されるような音源数が変動する状況に対する汎化性能は検証されていない．実際の楽曲では，曲ごとに編成や音源数が大きく異なるため，学習データに存在しない音源構成に対してもモデルが頑健に動作することが求められる．この点を明らかにするため，本章では学習段階で音源数を変化させる設定を導入し，提案手法が未知の音源構成に対してどの程度適応できるかを検証する．本章では，音源数の異なるデータを学習に取り入れることが提案手法の汎化性能向上にどの程度寄与するか，そして補助情報の有効性が未知条件においても維持されるかを明らかにする．

5.2　実験の目的

本章の目的は，提案手法およびベースラインが，学習時とは異なる音源数や音響条件に対してどの程度汎化できるかを検証することである．そのために，音源数の異なる複数のデータセットを用いてモデルを学習し，学習時に含まれない音源数を持つデータに対して分離性能を評価する．また，音源数の変動に対応できない SpaIn-Net を除外し，固定音源数で学習された前章のモデルを基準として比較することで，音源数変化に対応した学習が性能に与える効果を明確にする．以上により，提案手法が実観測環境のような音源数が変動する条件に対してどの程度適応可能か，さらに補助情報を利用する利点が未知条件においても保持されるかを総合的に評価する．

5.3　実験条件

本節では，提案手法および比較手法の性能を評価するために用いたデータセット構成，モデル設定，学習条件，および評価指標について述べる．データセットの生成方法そのものは 3.4

節で詳述したため，本節では学習に用いた音源数構成と学習条件を中心に説明する．

5.3.1　データセット

本実験におけるデータセットおよび実験条件を Table 5.1 に示す．使用する音楽データのドライソースは MUSDB18 [34] に含まれるモノラル音源信号（vocal，bass，drum，および other）とした．各楽曲において，各音源信号を時間方向に 10 秒の固定長で切り出し，パンニング処理を施すことで観測ステレオ音楽信号を生成し，学習，検証，および評価に用いた．この切り出しは，一定のシフト幅で移動させながら行うため，隣接するサンプル間には時間的な重なりが生じる．学習，検証，および評価では異なる楽曲を用いている．

本章では，学習段階において音源数を変化させた場合の汎化性能を検証することを目的とし，2〜4 音源で構成された複数条件のデータセットを構成した．基本となるモノラル音源信号は共通とし，データセット作成時に一部の音源を除外することで，音源数が異なる条件を生成している．このとき，方位分割数やパンニング条件などの基本的な設定は 4 章と同一である．学習用データには，4 音源条件を 2,893 サンプル，3 音源条件を 2,642 サンプル，2 音源条件を 2,561 サンプル含め，総音源時間は 80,960 秒とした．検証用データおよび評価用データについても，同様に音源数条件を変化させて構成しており，それぞれの総時間を Table 5.1 に示す．なお，学習データには 100 曲，検証および評価データにはそれぞれ 10 曲の楽曲を用いた．学習・検証・評価データの時間配分については，検証および評価において同程度の信頼性で性能を確認することを目的として，両者をほぼ同規模とし，学習データについては十分な学習量を確保するため，それらのおよそ 10 倍の時間を割り当てた．また，各音源信号の切り出し長を 10 秒とした理由については，音楽信号に含まれるフレーズや音響的変化を十分に含みつつ，計算資源や学習の安定性とのバランスが取りやすい長さとして，経験的に妥当であると判断したためである．

本設定では，音源数の減少に伴い無音方向の数が増加する．例えば，4 音源条件では無音方向は 1 方向であるのに対し，3 音源条件では 2 方向，2 音源条件ではさらに多くの無音方向が存在する．このため，音源数が少ないほど分離が容易になる一般的な音源分離問題とは異なり，無音方向の出力抑制や誤検出の防止が重要な課題となる．以上により，本章では音源数の変動に対する汎化性能を，より厳密かつ現実的な条件下で評価する．

5.3.2　モデル構成および比較手法

提案手法およびベースモデルはいずれも，IC Conv-TasNet を基本骨格として構築した．エンコーダのフィルタ長，TCN ブロック数，チャネル数などの主要なハイパーパラメータは両手法で統一し，モデル容量を揃えることで公平な比較を行った．

比較手法として 4 章で用いた SpaIn-Net は，音源数に対応した固定出力構造を前提として設計されており，音源数が変動する条件への直接的な適用は想定されていない．そのため，本章で扱う可変音源数条件における評価対象からは除外した．

Table 5.1. Dataset and training parameters for chapter 5 experiments

Parameter	Value
Dataset	
Total train audio duration [s]	80,960
Total validation audio duration [s]	8,220
Total test audio duration [s]	8,570
Signal duration per sample [s]	10
Sampling rate [Hz]	16,000
Model configuration	
Number of directionally separated channels	$N = 5$
Number of TCN blocks per stack	8
Number of stacks	3
Mixed-precision training	Enabled (FP16 + FP32)
Training hyperparameters	
Learning rate	1×10^{-4}
Batch size (per update)	4
Gradient accumulation steps	4
Early stopping patience [epochs]	30
Number of training epochs [epochs]	200
Loss function threshold (SI-SNR) [dB]	$a_{\mathrm{th}} = -30$
Numerical stability constant for SI-SNR	$\varepsilon = 1 \times 10^{-8}$
Silence threshold	$\varepsilon_s = 1 \times 10^{-4}$

代替の比較として，4 章で用いた「4 音源条件のみのデータセットを用いて学習した提案手法モデル」を比較対象として用いる．本モデルは，学習段階では音源数の変動を考慮していないため，音源数可変条件に対する汎化性能を評価するための適切な比較対象として位置づけられる．

5.3.3 学習・最適化条件

本章における学習および最適化条件は，4 章（4.3.3 節）と同一とした．DNN の最適化には Adam [35] を用い，混合精度学習 [36] および累積勾配を導入した．混合精度学習では，一部の演算を 16-bit 浮動小数点数で実行しつつ，学習に直接影響する重みパラメータは 32-bit 精度で保持した．また，バッチサイズ 4 の勾配を 4 回分蓄積し，合計 16 バッチ相当の勾配を用いて 1 回の重み更新を行った．損失関数には threshold SI-SNR を採用し，閾値は -30 dB に設定した．学習は最大 200 epochs とし，検証損失に基づく早期終了を適用した．音源分離

性能の評価には，4 章と同様に source-to-distortion ratio (SDR) [37] を用いた．

5.4　実験結果

本章では，前章で学習した固定音源数学習モデルと，音源数を変化させて学習した可変音源数学習モデルを対象とし，学習時とは異なる音源数条件下における音源分離性能を比較・検証する．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 Figs. 5.1〜5.4 に示す．また，各手法におけるステレオ音楽信号の各音源に対する平均 SDR と，各手法による SDR 改善量（ΔSDR）を Table 5.2 に示す．

ステレオ音楽信号のは全音源で平均 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，全音源平均の ΔSDR 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 8.45 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．しかし，前章の固定音源数条件と比較すると性能は低下している．これは，音源数の変動により，モデル内部で想定していない無音方向が増加し，誤検出やリークが発生しやすくなったことが一因と考えられる．ここでリークとは，本来他の音源に属する成分が，分離出力間，あるいは本来無音であるべき方位成分に混入する現象を指す．固定音源数学習モデルは，すべての音源において IC Conv-TasNet を大きく上回る SDR 改善を示し，未知の音源数条件に対しても一定の分離性能を維持できていることが分かる．全音源平均の ΔSDR は 13.24 dB であり，分離性能は向上しているが，本モデルは常に 4 方向の音源存在を仮定して出力する構造であるため，実際には無音となる方向に対しても推定が行われ，性能の上限が制限されている可能性が示唆される．可変音源数学習モデルは，bass，drum，other，および vocal のすべての音源において最も高い SDR 改善を示しており，全音源平均の ΔSDR は 16.03 dB に達した．これは，学習段階で音源数の変動を考慮したデータを与えることで，無音方向の抑制および実音源方向の強調が適切に学習されたためであると考えられる．

固定音源数学習モデルと可変音源数学習モデルの SDR 分布を比較すると，音源数が少ない条件では分離が相対的に容易であるにもかかわらず，固定音源数学習モデルでは高 SDR 領域に属するサンプルの割合が可変音源数学習モデルと比べて明確に少ないことが確認できる．これは，本モデルが常に 4 方向に音源が存在することを仮定して出力を行う構造を採用しているため，実際には音源が存在しない無音方向に対しても推定が行われ，出力エネルギーが分散してしまうことに起因すると考えられる．その結果，本来音源が存在する方向への推定が相対的に弱まり，分離が容易な条件においても SDR の上昇が抑制された可能性が示唆される．

以上の結果より，音源数が変動する未知条件下においては，固定音源数学習モデルでは性能の上限が制約される一方，音源数変動を考慮して学習した可変音源数学習モデルは，高い分離性能を安定して維持できることが明らかとなった．

Table 5.2. Average SDR [dB] and SDR improvement (ΔSDR) for each source under generalized training conditions

Source		bass	drum	other	vocal	Avg
Mixture	SDR	-3.16	-2.60	-4.07	-4.73	-3.64
Directionally separated signal	ΔSDR	6.31	7.59	8.03	9.78	7.93
IC Conv-TasNet (Baseline)		9.07	7.60	7.54	9.59	8.45
Proposed method (Fixed-source training)		13.24	11.58	12.90	15.25	13.24
Proposed method (Generalized training)		**15.68**	**14.70**	**15.73**	**18.00**	**16.03**

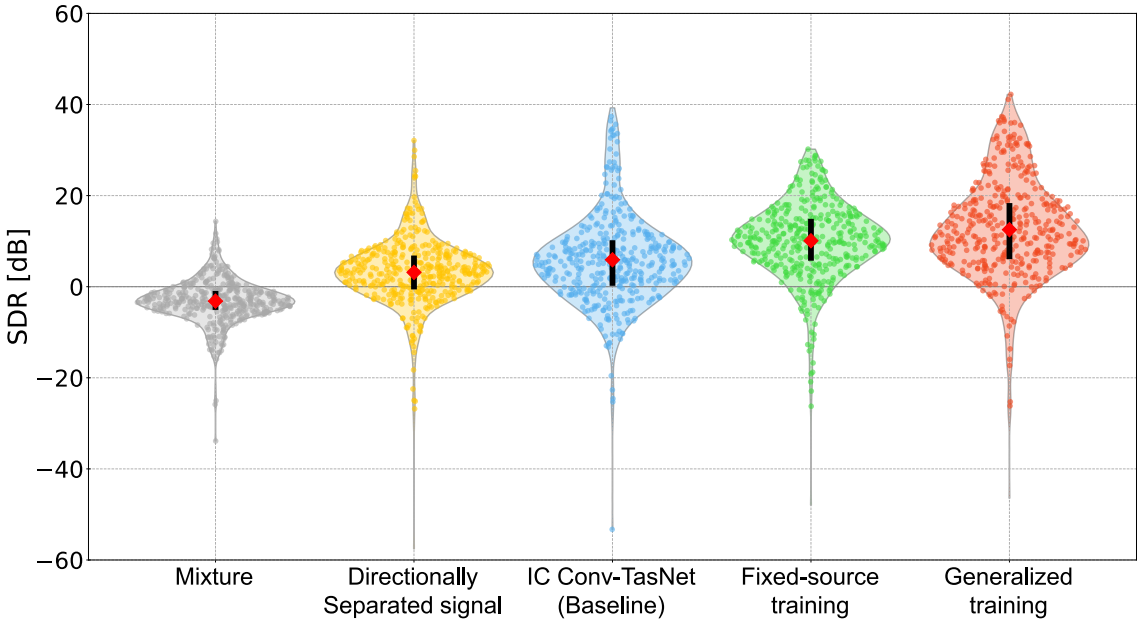


Fig. 5.1. Violin plot of SDR for bass under variable-source conditions.

5.5 本章のまとめ

本章では，音源数が変動する条件下における提案手法の汎化性能を検証するため，学習段階で音源数を変化させた可変音源数学習モデルを構築し，IC Conv-TasNet および固定音源数条件で学習したモデルとの比較実験を行った．

全音源平均の ΔSDR に着目すると，固定音源数学習モデルは，学習時とは異なる音源数条件に対してもベースラインモデルを上回る分離性能を維持していることが確認できる．一方で，音源数の減少に伴い無音方向が増加する条件下では，誤検出やリークの影響により性能の上限が制約される傾向が見られた．これに対し，可変音源数学習モデルは，すべての音源において最も高い ΔSDR を達成しており，未知の音源数条件に対しても安定して高い分離性能を示した．特に，全音源平均の ΔSDR は 16.03 dB に達しており，音源数変動を考慮しない学

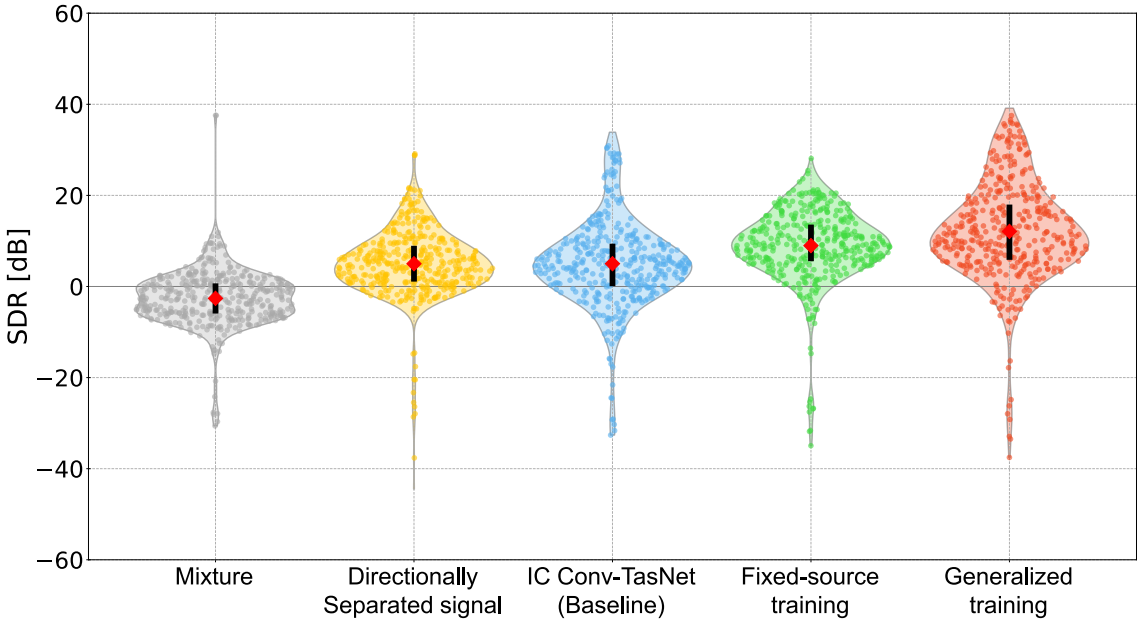


Fig. 5.2. Violin plot of SDR for drum under variable-source conditions.

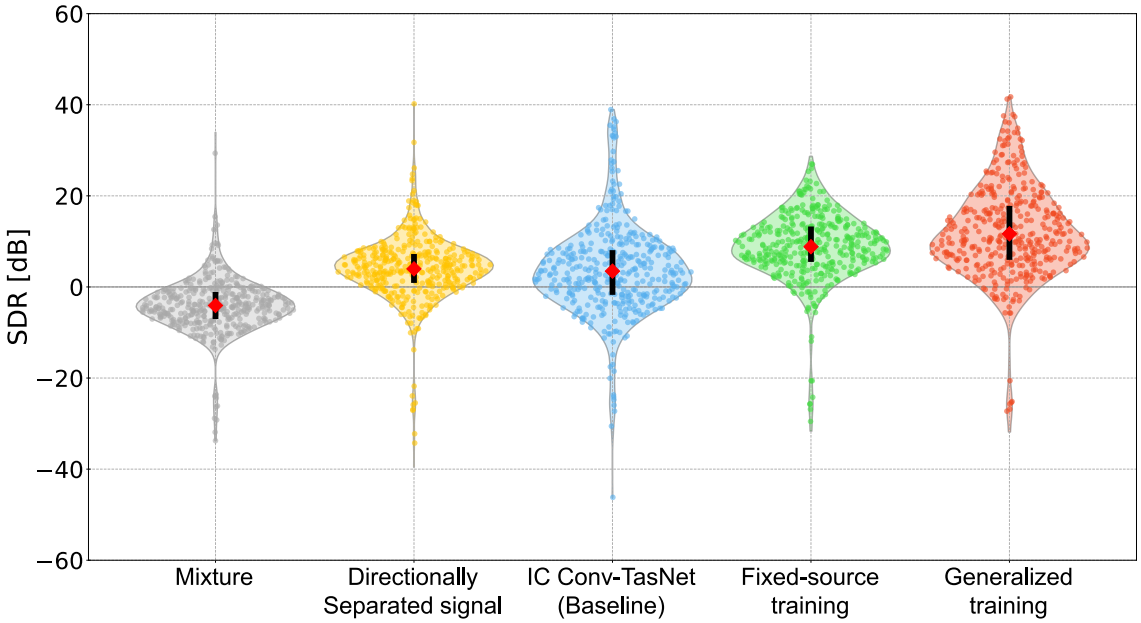


Fig. 5.3. Violin plot of SDR for other under variable-source conditions.

習条件と比較して，顕著な性能向上が確認された．さらに，SDR 分布の解析から，可変音源数学習モデルは平均性能の向上に加え，高 SDR 領域におけるサンプル数の増加や分布全体の高性能側へのシフトを実現しており，楽曲間の性能ばらつきを抑制しつつ，一貫した分離性能を達成できることが明らかとなった．

以上の結果より，学習段階において音源数の変動を明示的に考慮することは，無音方向の抑

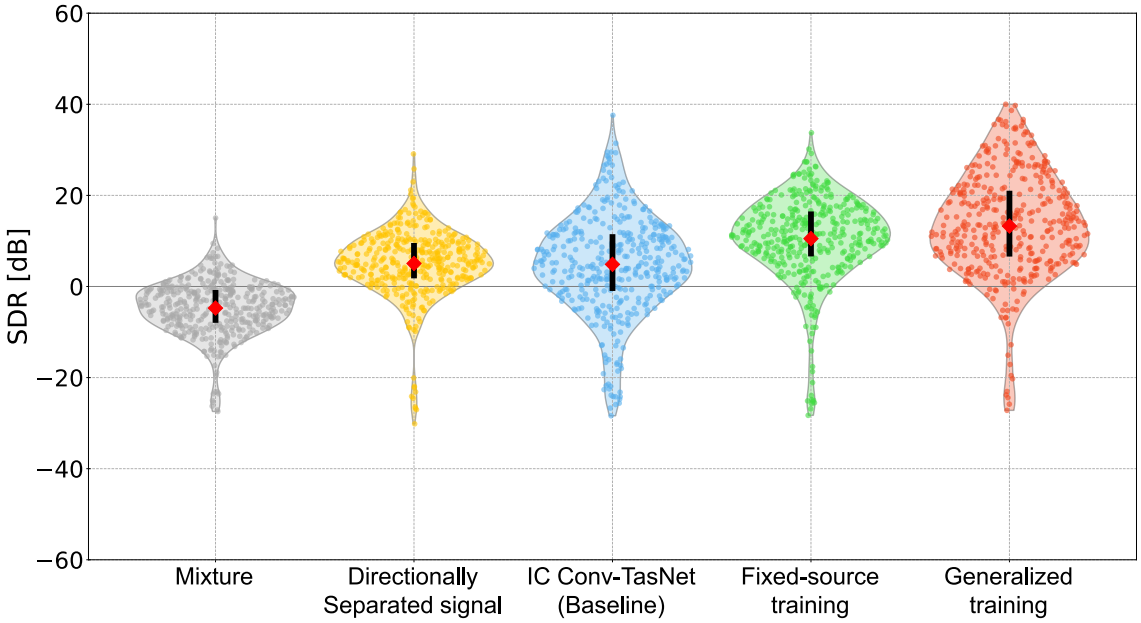


Fig. 5.4. Violin plot of SDR for vocal under variable-source conditions.

制および実音源方向の強調を適切に学習させる上で有効であり，実環境における編成の多様性に対応可能な，頑健な音源分離モデルの構築に有効であることが示された．

第 6 章

結言

本論文では，ステレオ音楽信号から生成可能な方位情報を補助情報として利用し，DNN による音楽音源分離の精度を向上させる手法について検討した．特に，左右チャネル間の音量比に基づいて生成される方位分離信号を FiLM によりモデル内部に適応的に反映させる新たな構成を提案した．

1 章では，ステレオ音楽信号分離に関する背景と課題を整理し，既存の DNN 手法の音楽音源分離が高精度である一方，大規模データや多大な計算資源が必要という問題点を指摘した．これらを踏まえ，追加の計測や外部情報に依存せず，大規模データや計算資源への依存を克服することを目標とし，その手段としてステレオ音楽信号のみから得られる方位分離信号（補助情報）の活用を研究目的として設定した．

2 章では，研究に用いる要素技術の解説および研究目標で述べた補助情報を活用する手法について述べた

3 章では，ステレオ音楽信号の左右チャネル間音量比を用いた方位分離信号の生成手法と，FiLM による条件付け機構を組み込んだ DNN モデルの構造について詳述した．提案手法は，音源数に依存せず，方位ごとに信号を分離するチャネル依存型の構造を採用しており，従来の音源数固定型モデルとは異なる視点から音源分離を実現する点に特徴を有する．

4 章では，SpaIn-Net および IC Conv-TasNet との性能比較実験を行い，提案手法が SpaIn-Net および IC Conv-TasNet と比べて高い SDR 改善を達成できることを示した．特に, FiLM による補助情報の導入が，分離性能の安定化および高精度化に寄与することを実験的に明らかにした．

5 章では，音源数が変動する条件に対する汎化性能を評価した．その結果，音源数固定条件で学習したモデルは未知条件下で性能低下が見られたのに対し，音源数を変化させて学習した可変音源数学習モデルは，高い分離性能を維持できることを確認した．また，分布解析の結果から，提案手法は平均性能だけでなく楽曲間のばらつきを抑制し，安定した分離性能を示すことが明らかとなった．これらの結果は，方位情報と FiLM を組み合わせた本手法が，実環境に近い未知条件に対しても有効であることを示している．

最後に今後の課題について述べる．本論文では左右音量比に基づく方位情報を用いたが，反

射音や残響を含む実環境下では定位情報の推定精度が低下する可能性がある．そのため，位相差情報や時間遅延情報など，他の空間手がかりを組み合わせた拡張が今後の重要な課題となる．また，本手法は従来のパンニングが明確に施された楽曲に対しては有効に機能する一方で，近年の楽曲に多い，音源が中央付近に集中したミキシングが施された音源に対しては課題が残る．具体的には，方位を等間隔に分割する現在の設計では，エネルギーが一部の方位チャネルに過度に集中し，方位分離信号が十分な分離の手がかりとならない場合がある．この問題に対しては，等角度分割の代わりに等エネルギー分割を導入する手法や，方位分布に基づくクラスタリングを用いて適応的に分割数・分割位置を最適化する手法が有効であると考えられる．さらに，本手法では無音方向の抑制をネットワークにより暗黙的に学習させているが，出力チャネル数の自動最適化や動的な音源数推定と統合することで，より柔軟かつ実用的な音源分離システムの構築が可能になると期待される．これらの課題を克服することで，本論文では主に音楽信号を対象としたが，本手法の枠組みは，音声強調や環境音分離，監視音響解析など，他の音響信号処理タスクへの応用可能性を有すると考えられる．以上より，本論文で提案した方位情報と FiLM に基づく音源分離手法は，追加の外部計測情報に依存することなく，高精度かつ高い汎化性能を有する音源分離を実現可能であることを示した．

謝辞

本論文は，香川高等専門学校 電気情報工学科 北村研究室において行った研究をまとめたものである．本研究を進めるにあたり，終始ご指導いただいた北村大地准教授に，深く感謝の意を表します．研究の技術的側面に関する指導に加え，研究課題への向き合い方や思考の整理の仕方など，研究を進めるうえで重要となる多くの示唆を与えていただきました．特に，研究の方向性に迷った際や，思うように結果が得られなかった場面においても，常に丁寧に議論の時間を設けていただき，的確な助言と励ましをいただきました．これらのご指導があったからこそ，本研究を最後までやり遂げることができたと感じています．また，進路選択の際には，より高い目標に挑戦するようご助言と励ましを賜り，大学院進学の決断を後押ししていただきましたことに，心より感謝申し上げます．

本論の副査である柿元健准教授ならびに村上幸一准教授には，論文の構成や記述に関して大変有益な助言を頂き，大変お世話になりました．ここに厚く御礼申し上げます．

また，Google DeepMind の小泉悠馬博士，東京農工大学の矢田部浩平准教授には，研究ミーティングを通じて多角的な視点から多くの貴重なご意見をいただきました．専門的かつ建設的な議論を重ねる中で，自身の研究を客観的に見直す機会を得ることができ，本研究の方向性や位置付けを改めて整理するうえで大きな助けとなりました．ここに心より感謝申し上げます．

北村研究室の同期である鈴木氏，和気氏には，研究室での 3 年間を通して多くの支えをいただきました．研究に関する議論だけでなく，何気ない雑談や日常のやり取りを通して，研究に取り組むうえでの気持ちを保つことができました．また，院試に向けて共に悩み，励まし合いながら過ごした経験は，研究生活を支える大きな力となり，今振り返っても非常に大切な思い出です．皆様と同じ環境で研究に取り組むことができたことに，心より感謝しています．

また，研究室の後輩である小川氏，谷野宮氏，大喜多氏，片山氏，森末氏には，日々の研究室生活を明るく支えていただきました．日常の会話で研究室の雰囲気を明るくしてくれたことに加え，輪講や研究活動に真摯に取り組む姿勢から，多くの刺激を受けました．皆様と共に過ごした日々が，研究に集中できる良い研究環境を築く一助となっていたことを，改めて実感しています．ここに深く感謝の意を表します．

最後に，これまで一貫して支え続けてくれた家族に，心より感謝します．日々の生活を支え，常に見守りながら応援してくれたことが，研究に専念できる環境につながりました．本論文は，多くの方々の支えのもとに完成したものであり，この場を借りて改めて感謝の意を表します．

参考文献

[1] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

[2] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Processing*, vol. 87, no. 8, pp. 1833–1847, 2007.

[3] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[4] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Trans. Signal Process.*, vol. 59, no. 7, pp. 3155–3167, 2011.

[5] 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.

[6] V. S. Kadandale, J. F. Montesinos, G. Haro, and E. Gómez, "Multi-channel U-net for music source separation," *in Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2020, pp. 1–6.

[7] 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.

[8] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3221–3236, 2023.

[9] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2025.

[10] 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.

[11] D. Petermann and M. Kim, "SpaIn-Net: Spatially informed stereophonic music source separation," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 106–110, 2022 .

[12] A. Ajit, K. Acharya, and A. Samanta, "A review of convolutional neural networks," *in Proc. International Conference on Emerging Trends in Information Technology and Engineering (IC-ETITE)*, pp. 1–5, 2020.

[13] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent neural networks: A comprehensive review of architectures, variants, and applications," *Information*, vol. 15, no. 9, p. 517, 2024.

[14] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5929–5955, 2020.

[15] J. Birnbaum, D. Sawyer, and S. Zhang, "Temporal FiLM: Capturing long-range sequence dependencies with feature-wise modulations," *in Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[16] H. Sato, T. Moriya, M. Mimura, S. Horiguchi, T. Ochiai, T. Ashihara, A. Ando, K. Shinayama, and M. Delcroix, "SpeakerBeam-SS: Real-time target speaker extraction with lightweight Conv-TasNet and state space modeling," *in Proc. Interspeech*, pp. 5033–5037, 2024.

[17] D. A. M. G. Wisnu, K. Huang, Y. Zhang, and S. Watanabe, "STSM-FiLM: A FiLM-conditioned neural architecture for time-scale modification of speech," *arXiv preprint arXiv:2510.02672*, 2025.

[18] S. Alfattama and A. Vaish, "Anatomically adaptive feature-wise linear modulation for deep learning-based low-dose CT denoising," *Available at SSRN: 5593009*, 2023, doi:10.2139/ssrn.5593009.

[19] M. Brockschmidt, "GNN-FiLM: Graph neural networks with feature-wise linear modulation," *in Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 119, pp. 1144–1153, 2020.

[20] J. H. Lee, H. S. Choi, and K. Lee, "Audio query-based music source separation," *in Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2019.

[21] P. Seetharaman, G. Wichern, S. Venkataramani, and J. Le Roux, "Class-conditional embeddings for music source separation," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 301–305, 2019.

[22] G. Meseguer-Brocal and G. Peeters, "Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations," *in Proc. Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2019.

[23] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "OpenUnmix: A reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[24] V. A. Kalkhorani and D. Wang, "CrossNet: Leveraging global, cross-band, narrow-band, and positional encoding for single-and multi-channel speaker separation," *arXiv preprint arXiv:2403.03411*, 2024.

[25] R. Sawata, S. Uhlich, S. Takahashi, and Y. Mitsufuji, "All for one and one for all: Improving music separation by bridging networks," *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *in Proc. Adv. Neural Inf. Process. Syst.*, 2017.

[27] X. Huang and S. J. Belongie, "Arbitrary style transfer in realtime with adaptive instance normalization," *in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017.

[28] 安藤彰男, "高臨場感音響技術とその理論," **電子情報通信学会**, vol. 3, no. 4, pp. 33–46, 2010.

[29] V. Melnykov and I. Melnykov, "Initializing the EM algorithm in gaussian mixture models with an unknown number of components," *Comput. Stat. Data Anal.*, vol. 56, no. 6, pp. 1381–1395, 2012.

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

[31] 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. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 900–904, 2019.

[32] 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. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, pp. 161–165, 2021.

[33] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, "Unsupervised sound separation using mixture invariant training," *in Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 3846–3857, 2020.

[34] Z. Rafii, A. Liutkus, F. R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," *in Proc. Int. Soc. Music Inf. Retr. Conf. ISMIR*, 2017.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu, "Mixed precision training," *arXiv preprint arXiv:1710.03740*, 2017.

[37] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind au-

dio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

発表文献一覧

国内学会

1. **加藤大輝**, 北村大地, “単一話者発話区間情報を援用したブラインド音源分離,” 第 27 回日本音響学会 関西支部 若手研究者交流研究発表会, pp. 11, 2024.
2. **加藤大輝**, 北村大地, 矢田部浩平, “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,” 日本音響学会 2026 年春季研究発表会講演論文集, 2026 (in press).

付録 A

提案手法と比較手法の性能評価実験における振幅スペクトログラムの比較

A.1　各手法における振幅スペクトログラムの比較

4.4 節において評価した提案手法の性能について，各手法の出力信号に対する振幅スペクトログラムを用いて視覚的に比較した結果を示す．本実験では，音源の配置条件として，Fig. A.1 に示す空間配置の一例を用いている．結果は代表的な配置条件における例であり，すべての条件を網羅するものではないが，各手法の特性を定性的に比較する目的には十分な情報を提供している．Fig. A.2 にはステレオ音楽信号の振幅スペクトログラムを可視化し，Figs. A.3〜A.7 には，当該方位領域 n に対する正解信号，IC Conv-TasNet の出力信号，比較手法である SpaIn-Net の出力信号，および IC Conv-TasNet と FiLM の組み合わせである提案手法の出力信号の振幅スペクトログラムを示し，それらを比較した結果を報告する．なお，SpaIn-Net は音源信号のみを出力するため，無音方向に対応する結果は省略した．

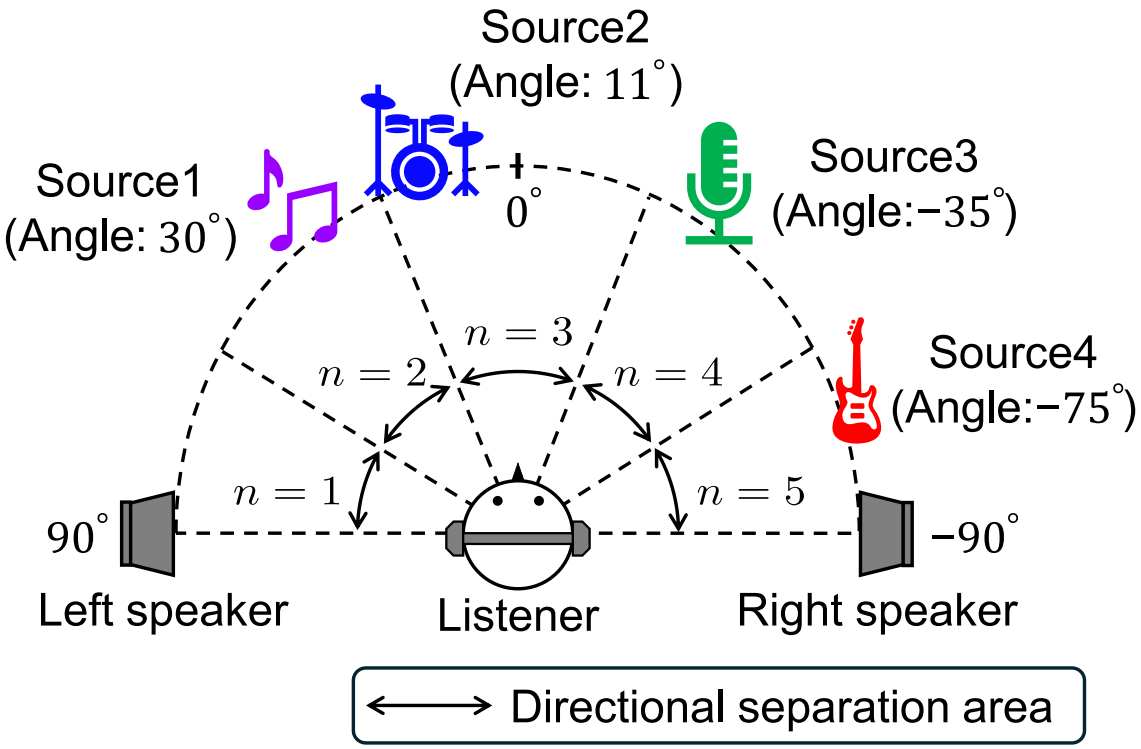


Fig. A.1. Source layout of the experiment.

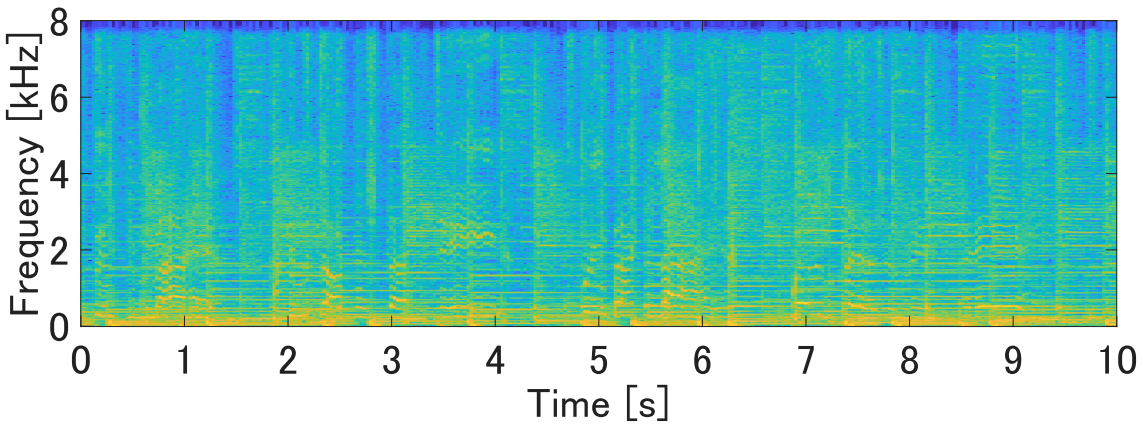


Fig. A.2. Amplitude spectrogram of a stereo music signal. The figure shows the time-frequency representation of the signal used in the experiment.

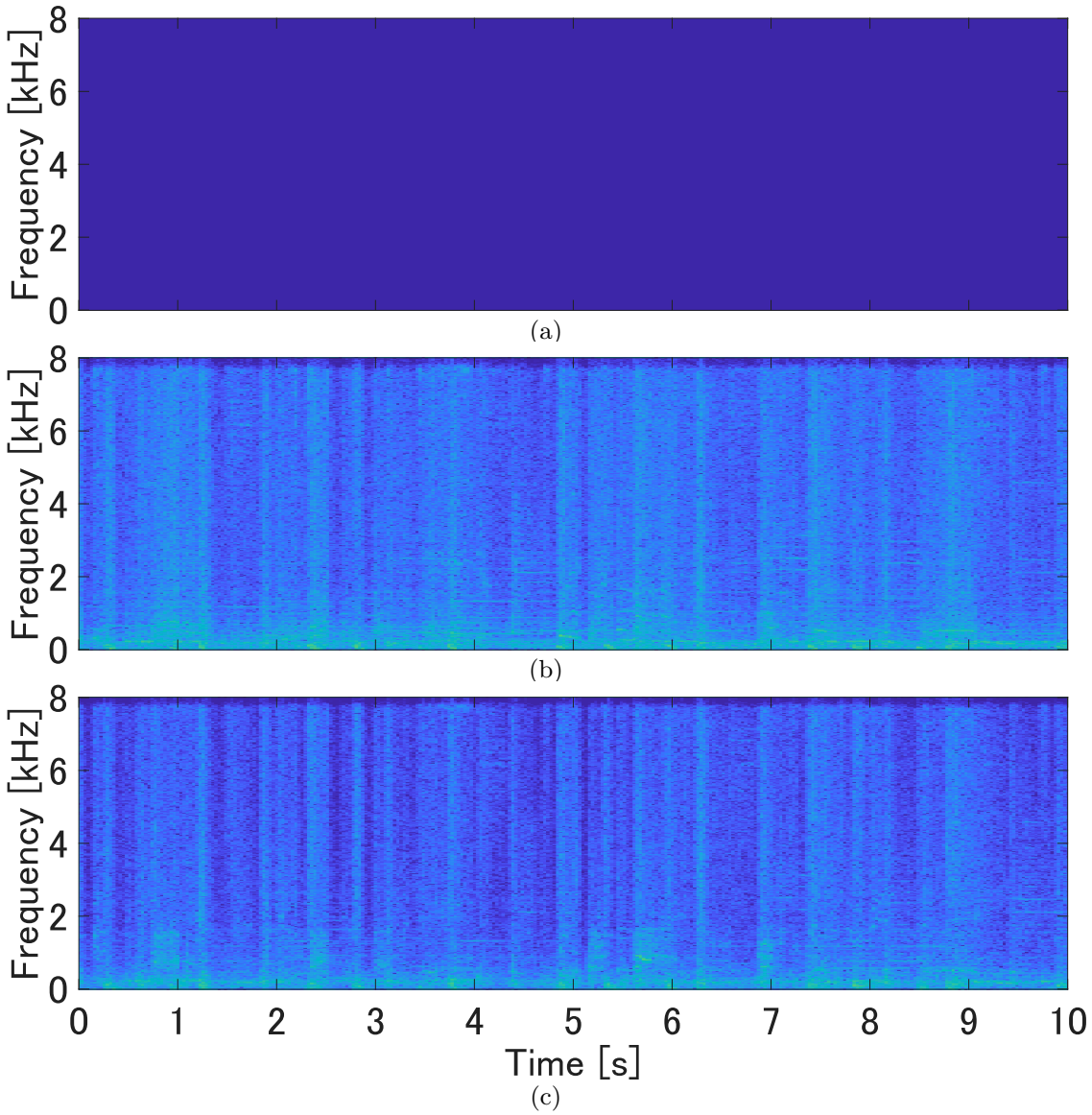


Fig. A.3. Amplitude spectrograms of directionally separated sources for $n = 1$: (a) clean signal, (b) IC Conv-TasNet, and (c) proposed method.

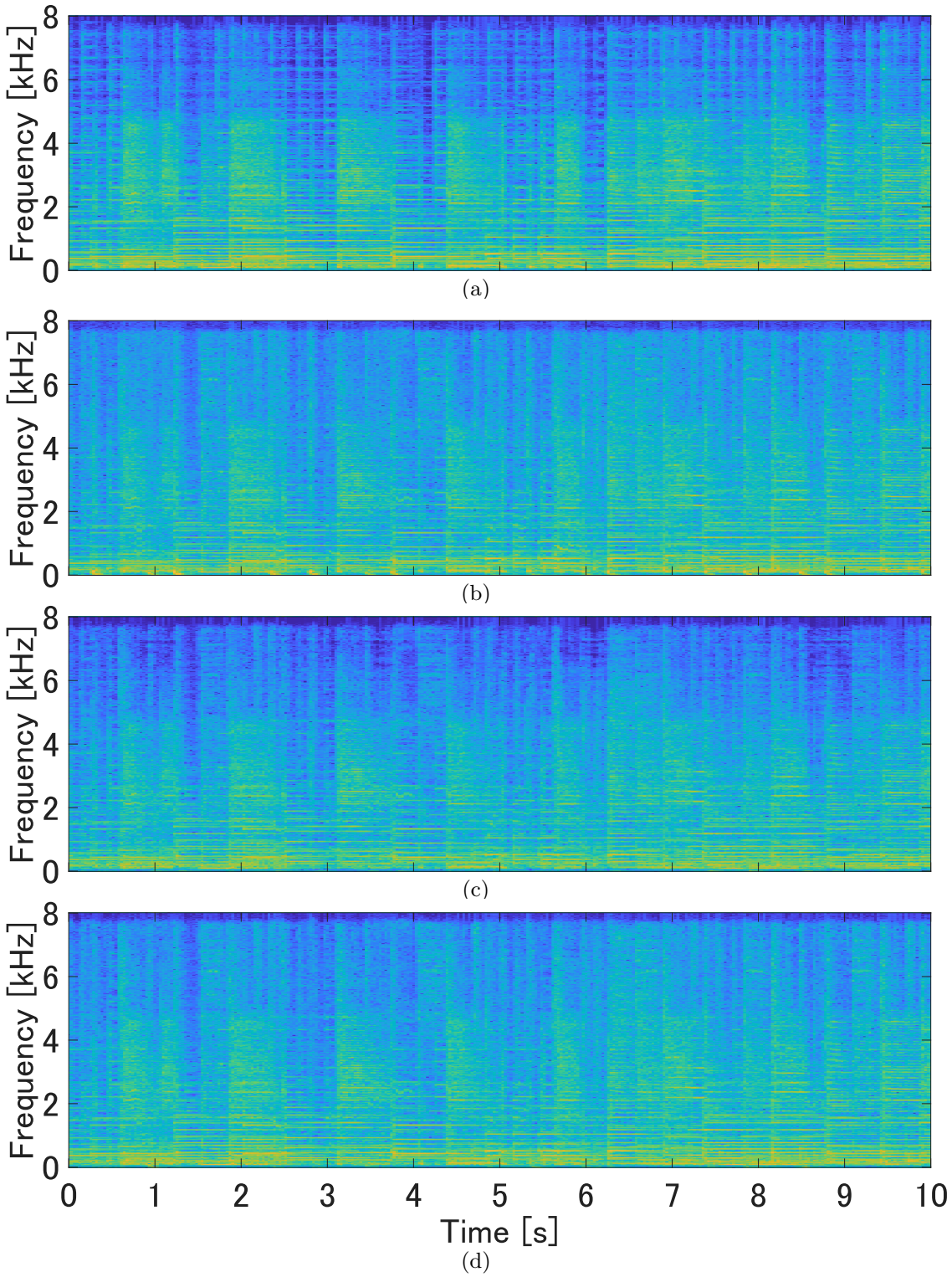


Fig. A.4. Amplitude spectrograms of directionally separated sources for $n = 2$: (a) clean signal, (b) IC Conv-TasNet, (c) SpaIn-Net, and (d) proposed method.

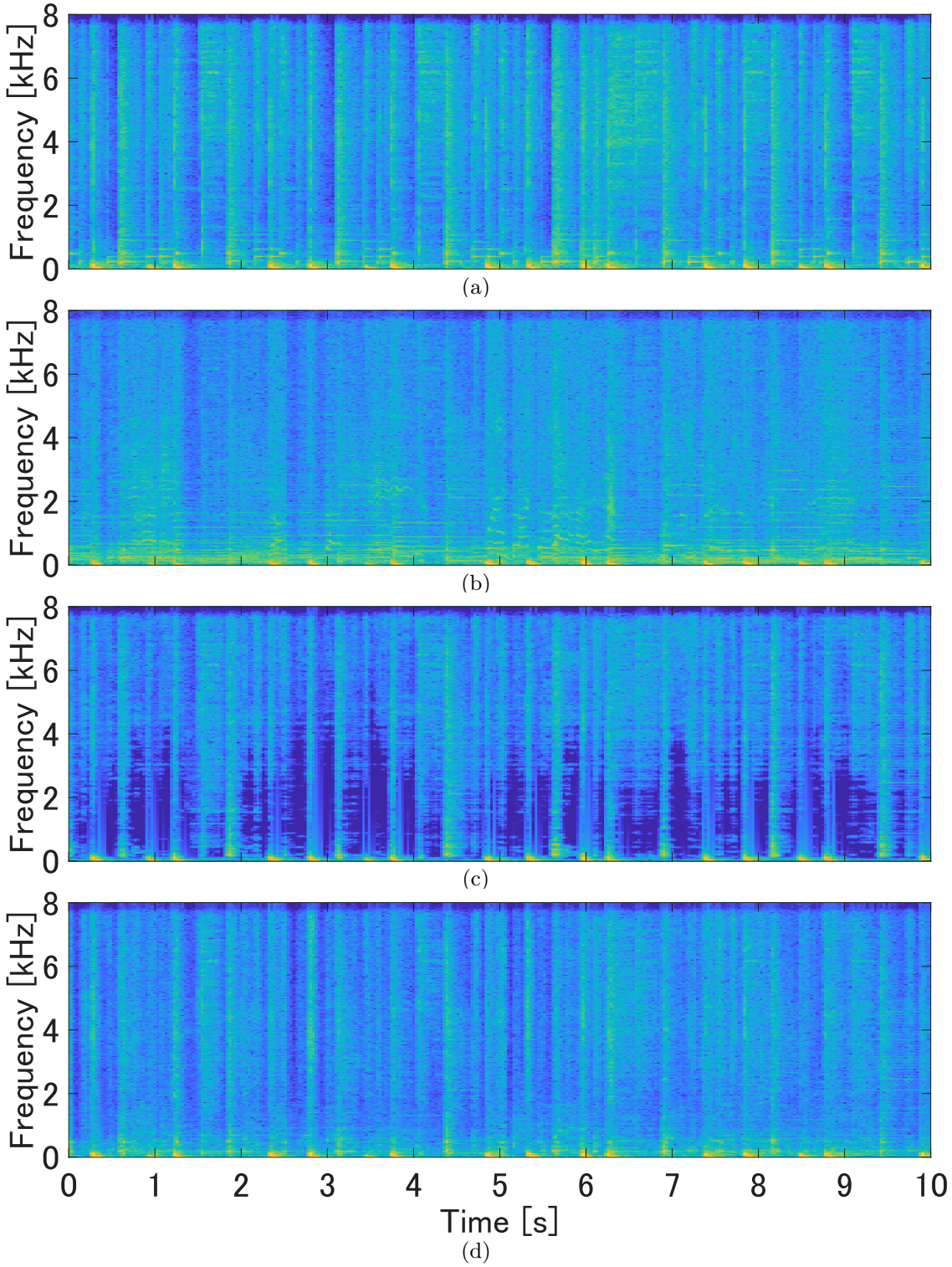


Fig. A.5. Amplitude spectrograms of directionally separated sources for $n = 3$: (a) clean signal, (b) IC Conv-TasNet, (c) SpaIn-Net, and (d) proposed method.

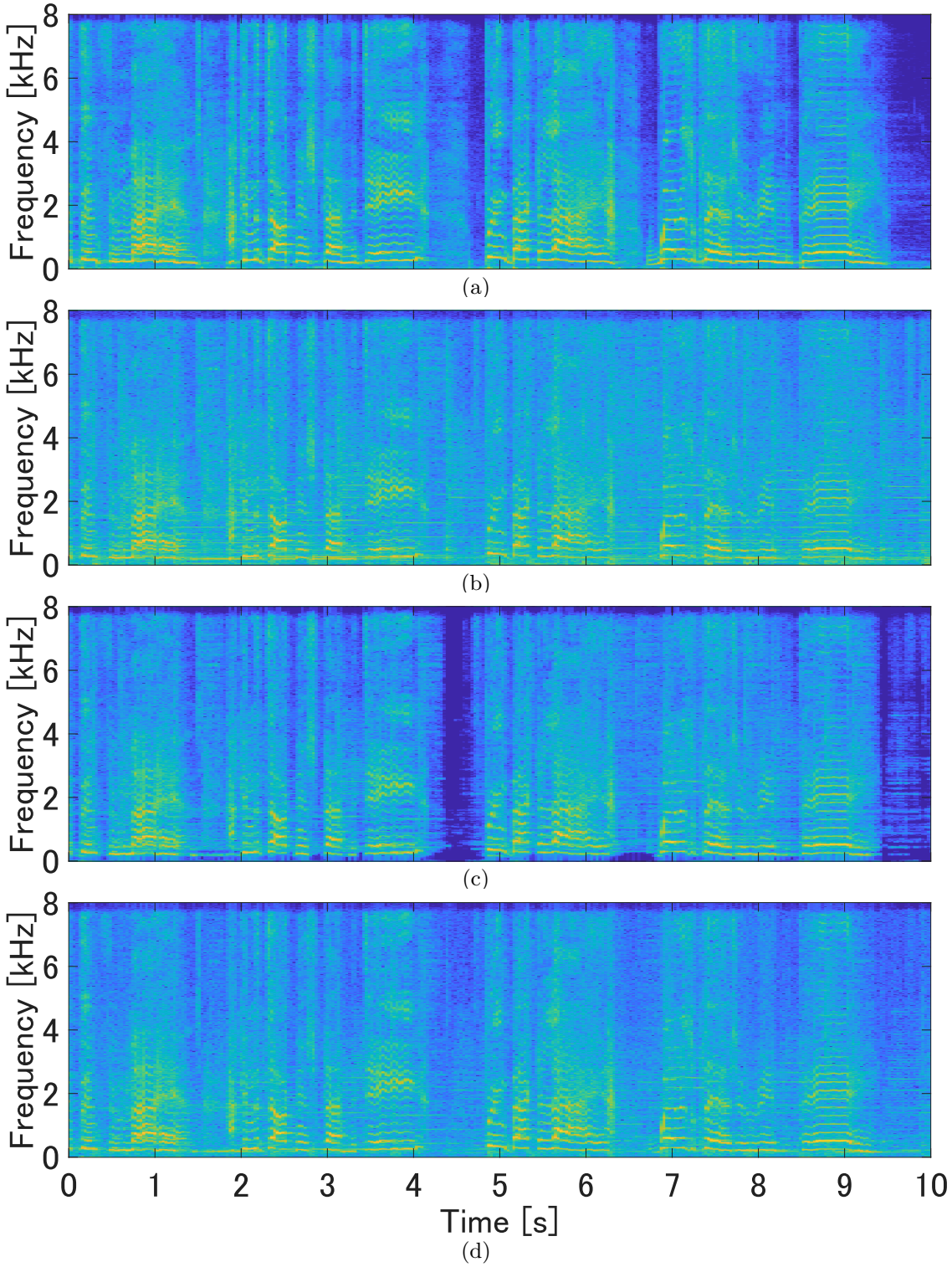


Fig. A.6. Amplitude spectrograms of directionally separated sources for $n = 4$: (a) clean signal, (b) IC Conv-TasNet, (c) SpaIn-Net, and (d) proposed method.

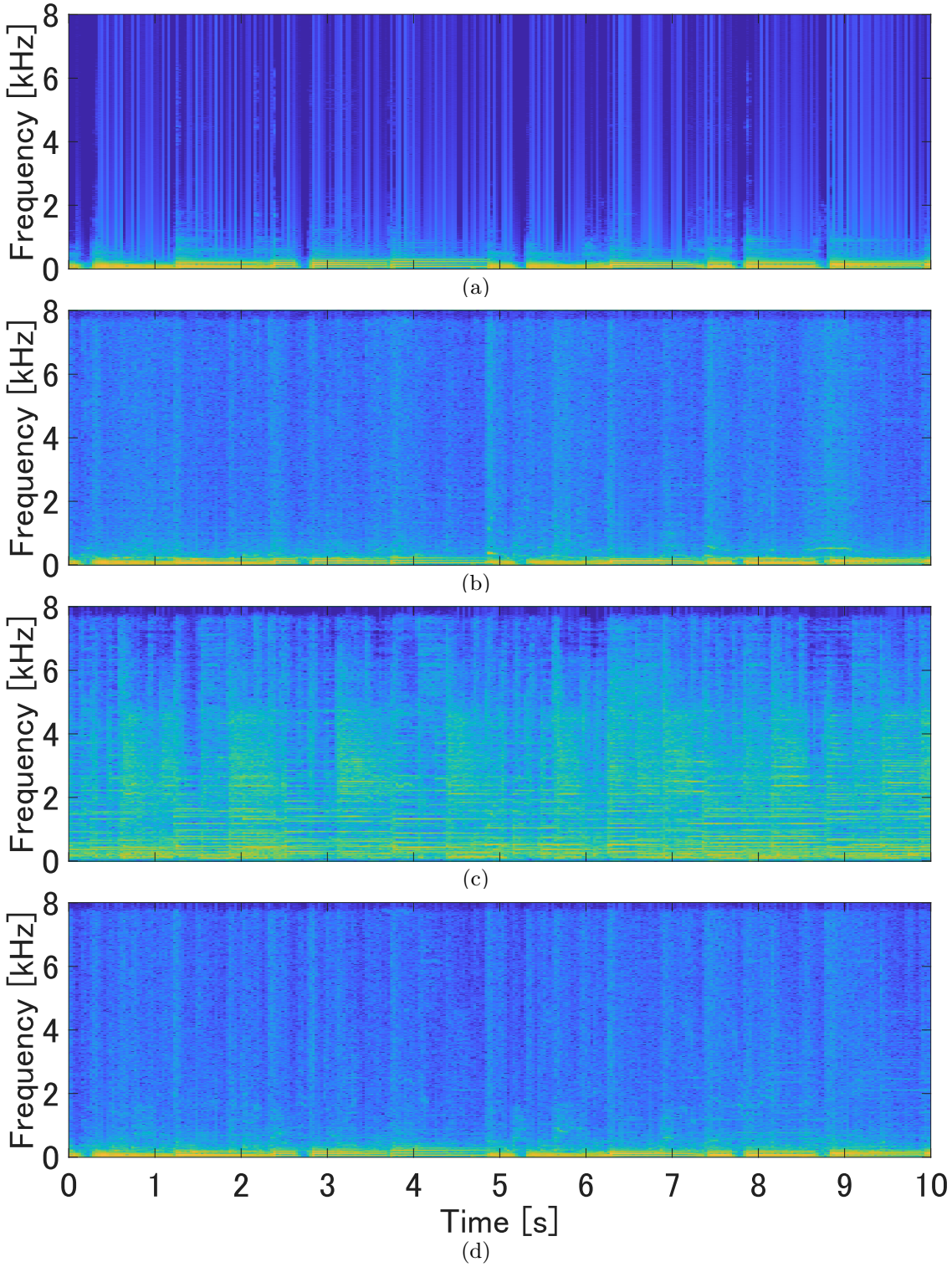


Fig. A.7. Amplitude spectrograms of directionally separated sources for $n = 5$: (a) clean signal, (b) IC Conv-TasNet, (c) SpaIn-Net, and (d) proposed method.

付録 B

汎化性能の評価実験における振幅スペクトログラムの比較

B.1　各手法における振幅スペクトログラムの比較

5.4 節において評価した提案手法の性能について，各手法の出力信号に対する振幅スペクトログラムを用いて視覚的に比較した結果を示す．本実験では，音源の配置条件として，Fig. B.1 に示す空間配置の一例を用いている．結果は代表的な配置条件における例であり，すべての条件を網羅するものではないが，各手法の特性を定性的に比較する目的には十分な情報を提供している．Fig. B.2 にはステレオ音楽信号の振幅スペクトログラムを可視化し，Figs. B.3〜B.7 には，当該方位領域 n に対する正解信号，IC Conv-TasNet の出力信号，固定音源条件で学習した提案モデルの出力信号，および可変音源条件で学習した提案モデルの出力信号の振幅スペクトログラムを示し，それらを比較した結果を報告する．

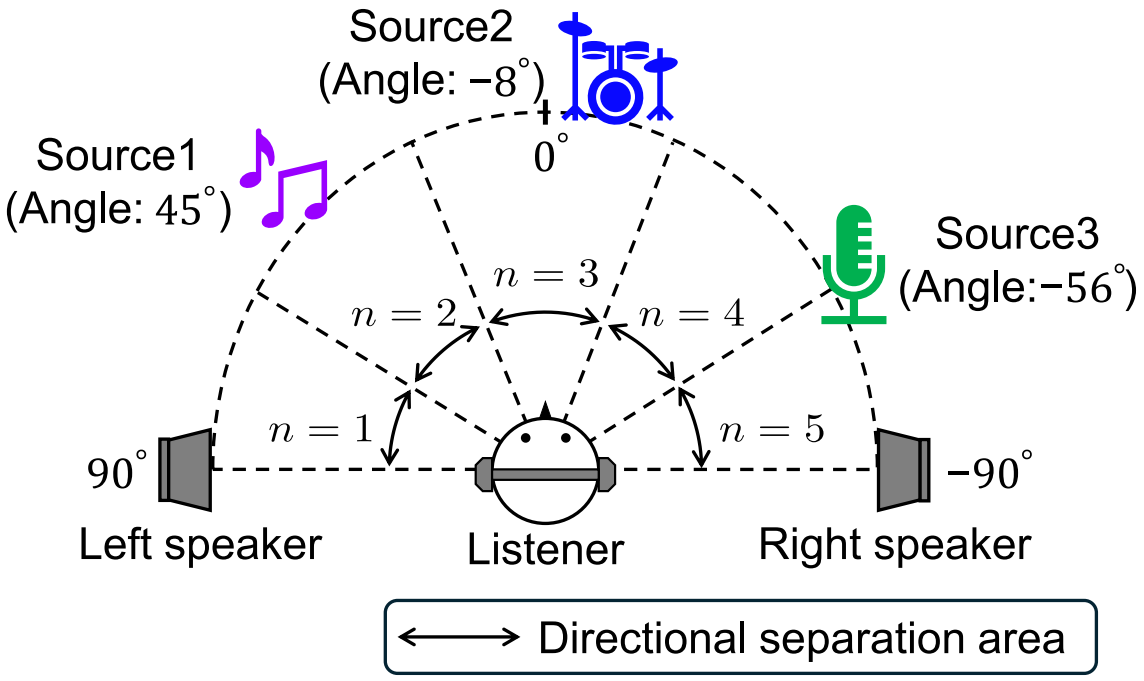


Fig. B.1. Source layout of the experiment.

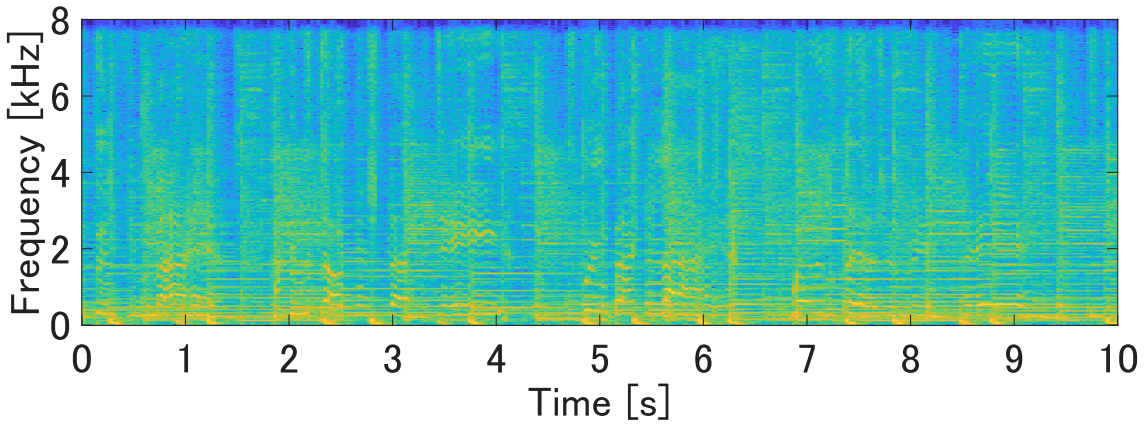


Fig. B.2. Amplitude spectrogram of a stereo music signal. The figure shows the time-frequency representation of the signal used in the experiment.

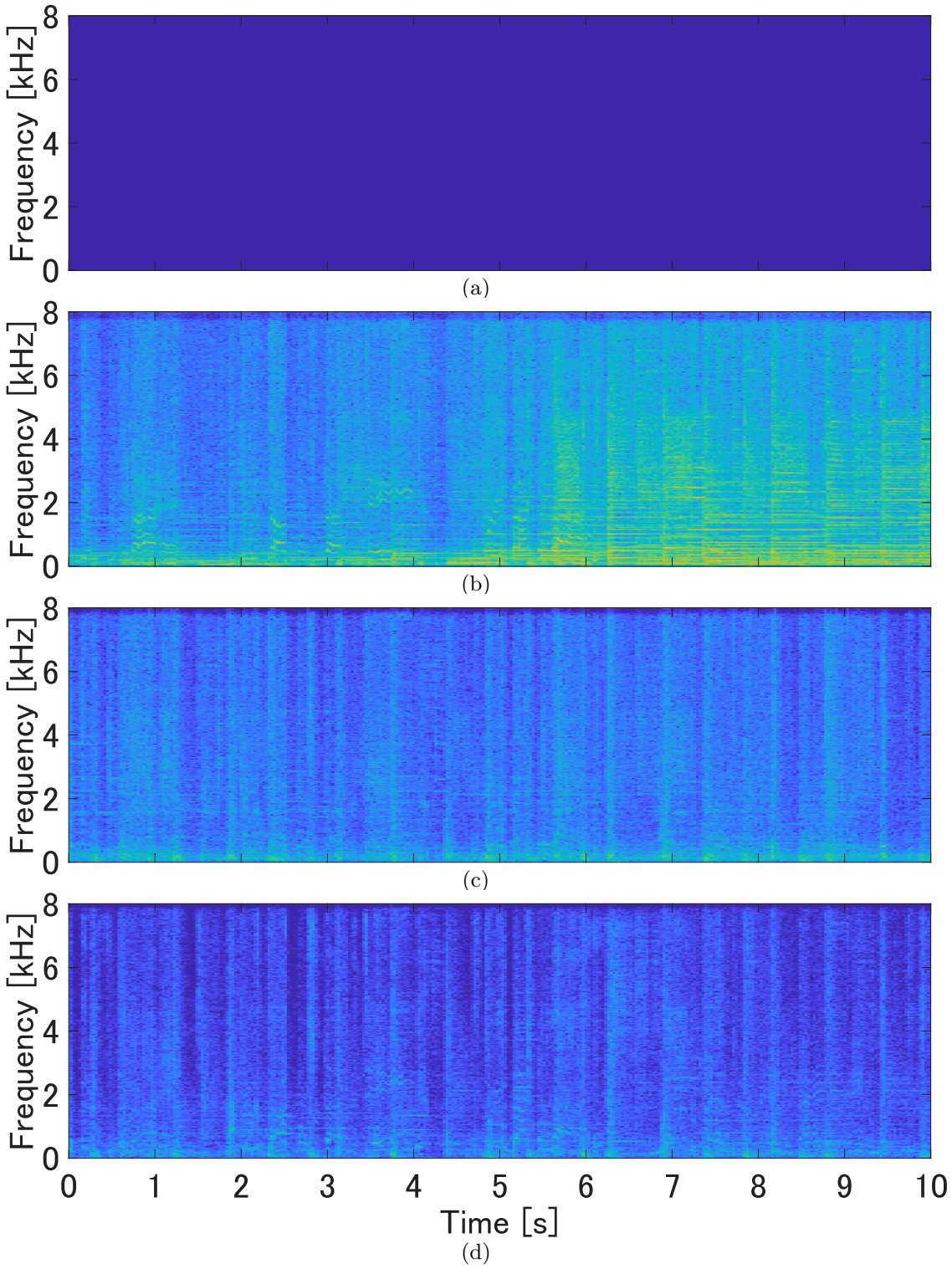


Fig. B.3. Amplitude spectrograms of directionally separated sources for $n = 1$: (a) clean signal, (b) IC Conv-TasNet, (c) proposed model (fixed-source), and (d) proposed model (generalized).

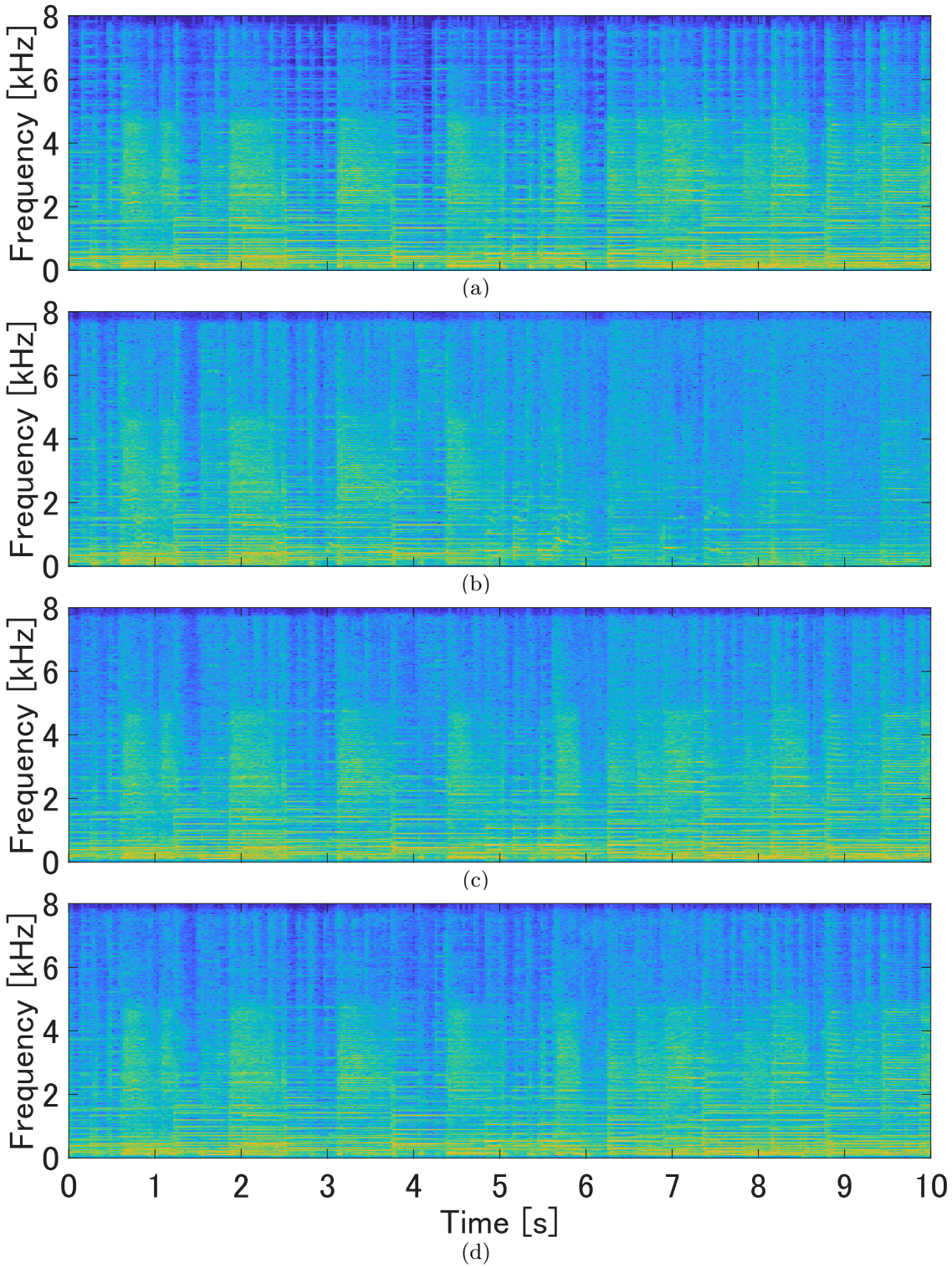


Fig. B.4. Amplitude spectrograms of directionally separated sources for $n = 2$: (a) clean signal, (b) IC Conv-TasNet, (c) proposed model (fixed-source), and (d) proposed model (generalized).

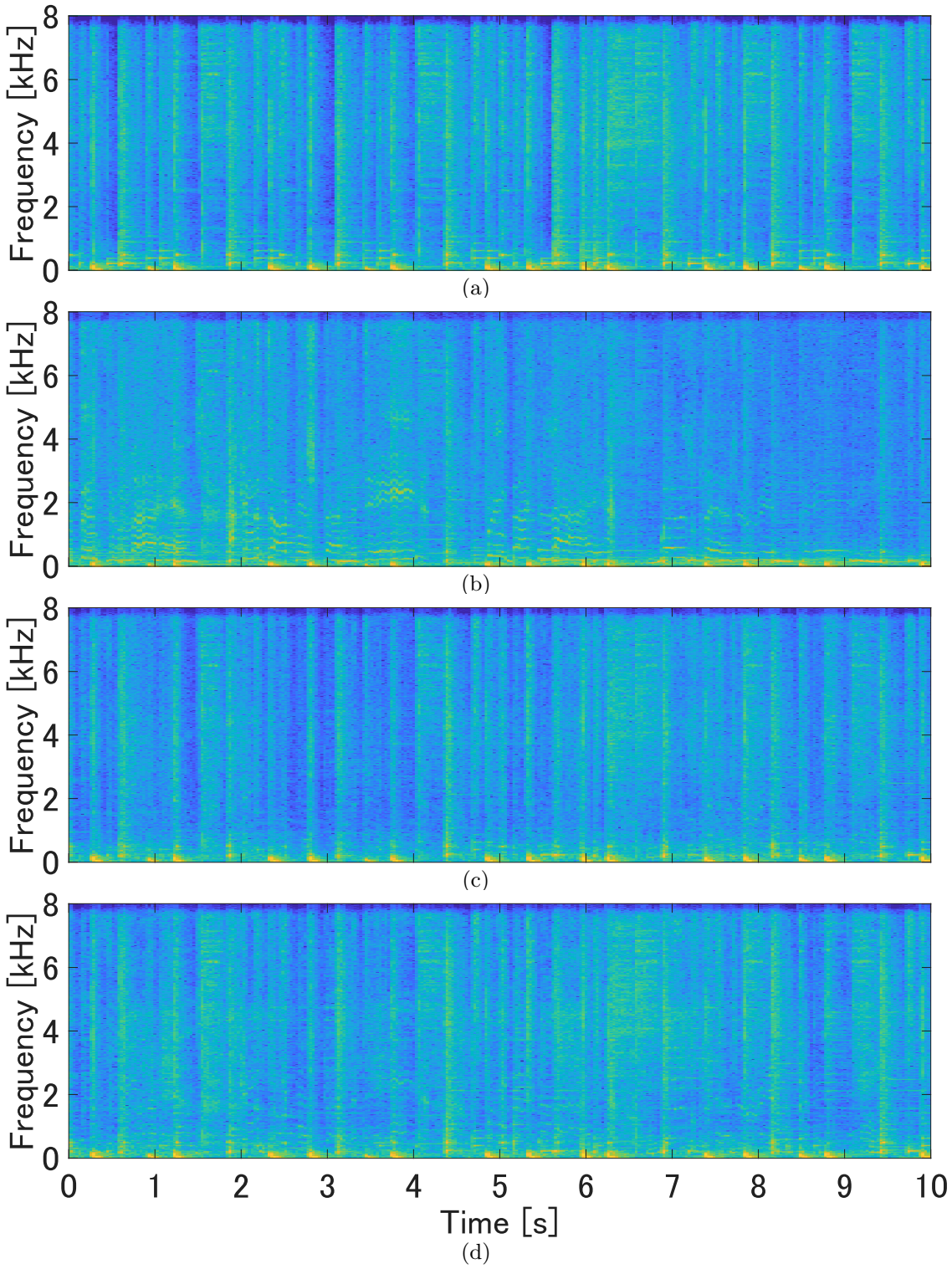


Fig. B.5. Amplitude spectrograms of directionally separated sources for $n = 3$: (a) clean signal, (b) IC Conv-TasNet, (c) proposed model (fixed-source), and (d) proposed model (generalized).

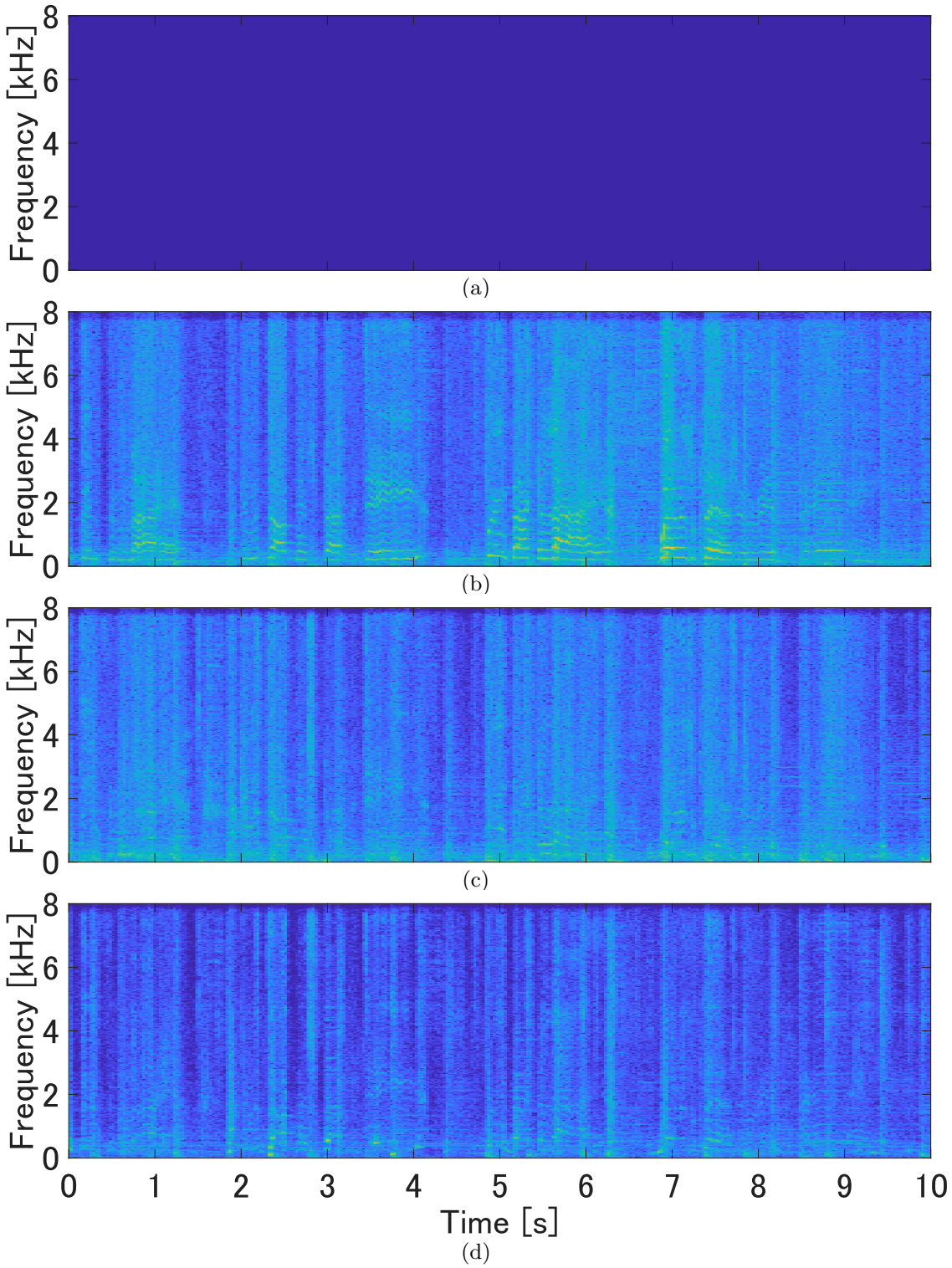


Fig. B.6. Amplitude spectrograms of directionally separated sources for $n = 4$: (a) clean signal, (b) IC Conv-TasNet, (c) proposed model (fixed-source), and (d) proposed model (generalized).

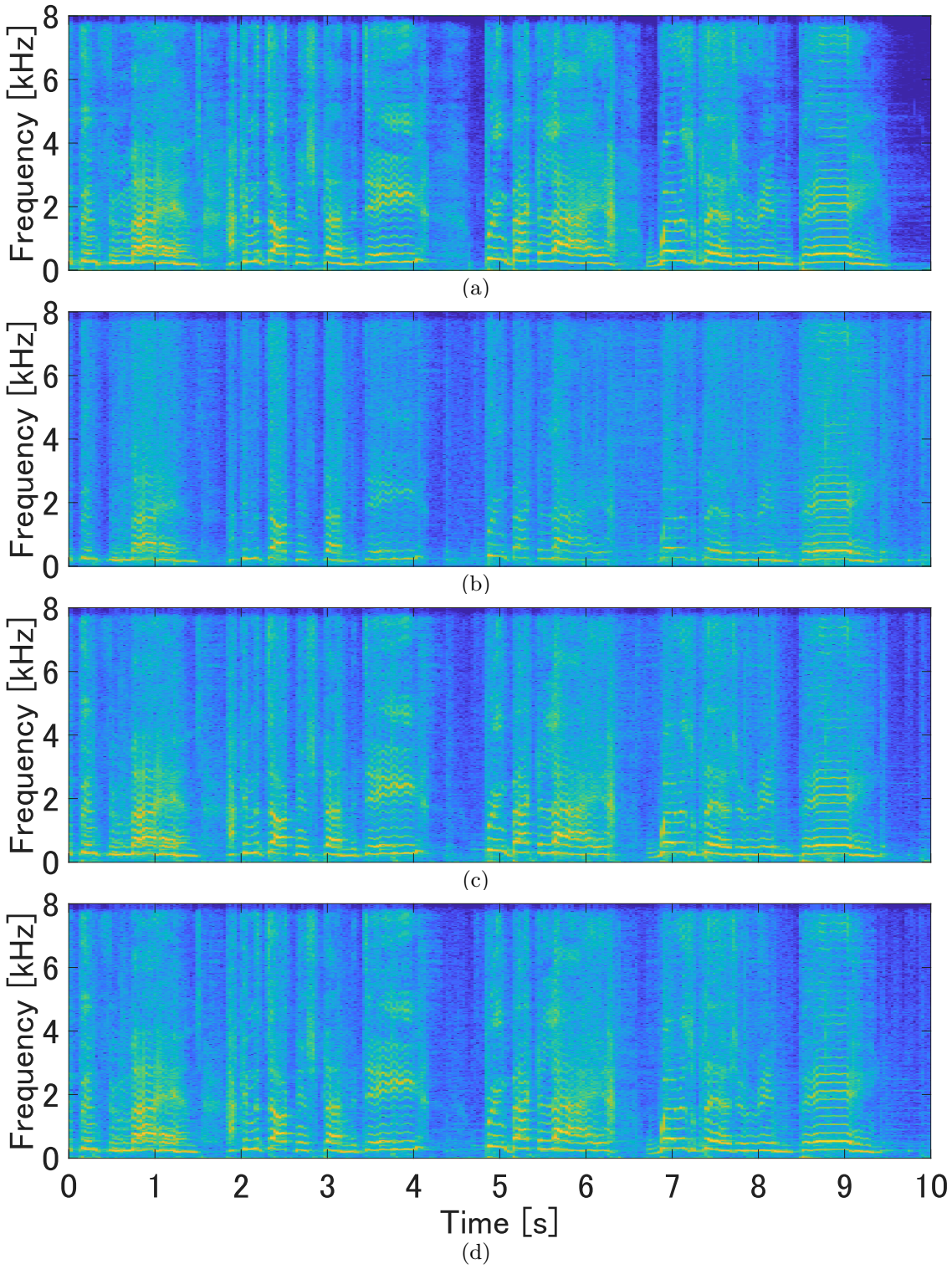


Fig. B.7. Amplitude spectrograms of directionally separated sources for $n = 5$: (a) clean signal, (b) IC Conv-TasNet, (c) proposed model (fixed-source), and (d) proposed model (generalized).