球面調和関数展開に基づく近接音抽出を用いた時間-周波数マスク推定による近接／遠方音分離

西口 草太
法政大学大学院 情報科学研究科 情報科学専攻
学生証番号 18t0012
E-mail:sota.nishiguchi.4b@stu.hosei.ac.jp

Abstract

We propose the combination of a physical-model-based and a deep-learning (DL)-based source separation for near- and far-field source separation. The DL-based near- and far-field source separation method uses spherical-harmonic-analysis-based acoustic features. Deep learning is a state-of-the-art technique for source separation. In this approach, a bidirectional long short term memory (BLSTM) is used to predict a time-frequency (T-F) mask. To accurately predict a T-F mask, it is necessary to use acoustic features that have high mutual information with the oracle T-F mask. In this study, low-frequency-band near- and far-field sources are estimated based on spherical harmonic analysis and used as acoustic features. Subsequently, a DNN predicts a T-F mask to separate all frequency bands. Our experimental results show that the proposed method improved the signal-to-distortion-rate by 8–10 dB compared to the harmonic-analysis-based method. IIn addition, the proposed method improved the PESQ and STOI compared to the conventional DL-based T-F mask estimation method.

1 まえがき

音源分離は雑音条件下での音声認識や話者識別のフロントエンド処理として有効である．既存の音源分離手法の多くは方向 [1] やスペクトルの時間周波数構造 [2]，またはその両方 [3] に焦点を当てて目的音と雑音を分離している．スペクトルを用いた手法は音声と非音声雑音の分離や，歌声と楽器音の分離の様に目的音と雑音の音色に明確な差がある場合は有効だが，目的音と雑音がともに音声であるような場合はスペクトル情報のみでの分離は難しい．時間周波数構造に着目した分離手法も研究されているが，音声と音声の混合 (特に同じ性別の話者) の場合に個々の音声の周波数構造が曖昧になり，分離の精度がそれほど上がらないことが課題となっている．方向を用いた分離では２音源の方向が近づくと分離精度が低下し，同方向の音源に関しては分離できない [1]．そこで本研究では，これらの従来の音響特徴が利用できない場面として，同じ方向の近接と遠方にある音源の混合音声を対象とした複数話者音源分離を考え，マイクと各音源の距離の違いに着目した近接音／遠方音分離を目指す．

近接音／遠方音の定義のために信号の波面形状を考える．ある点音源がマイクの無限遠にあると仮定すると，その音源からの受信信号は完全な平面波となる．一方で点音源がマイクの近傍にある場合は，その音源からの受信信号は球面波となる．これより，本論では球面波とみなせる信号を近接音とし，平面波とみなせる信号を遠方音と定義する．これらの境界 r，つまり「どれだけ音源が離れると波面が平面波とみなせるか.」は，信号の波長 λ とマイクのアレイ直径 L によって，$r = 10 \cdot L^2/\lambda$ と近似できることが知られている [4]．一般に長波長の信号ほど，より短い距離で平面波に近似でき，直径 0.1m の球面アレイを用いた場合，500Hz の信号はおよそ 0.15m で平面波となる．到来する波面を利用すれば，0.15m より近くの音源と遠くの音源を分離でき，雑音除去や残響の除去に利用できる．

上記の到来波面を利用した近接音抽出法として，球面調和関数に基づく手法が提案されている [5]．羽田らは中空の球面マイクロホンアレイを用い，球表面の音圧分布から中心音圧を球面調和関数展開により推定し，到来音との差をとることで近接音を分離する方法を提案した [5]．しかし前述したように音源から出た波が平面波に変化する距離は，信号の波長に反比例（周波数に正比例）するため，分離可能な周波数には上限がある．上限周波数は球面アレイの直径とマイク間距離にも依存しており，音声認識でよく用いられるサンプリングレート 16kHz の信号の全帯域での分離を想定しても，マイク数とアレイの大きさの観点から実現は困難である．

物理モデルと異なるアプローチとして機械学習による音源分離手法があり，近年では深層学習を用いた手法が提案されている [6, 7, 8, 9]．この方法では，ウィナーフィルタのような時間周波数 (T-F: time-frequency) マスク [10] をディープニューラルネットワーク (DNN: deep neural network) を利用して推定する．既存研究の多くはスペクトルの時間周波数構造 [6] や方向 [7] に焦点を当て，対数メルスペクトルやビームフォーミングの出力を音響特徴量として使用している．スペクトルを用いた手法では，混合音声の時間周波数構造のスパース性を用い，話者の声質や発話文のつながりからそれぞれの音声を強調する T-F マスクを推定する．しかし，目的音声を指定して抽出することができないため，特定の音源を取りだすという応用には向かない．適応発話を用い特定の話者の音声を強調する手法 [11] もあるが，事前に目的話者を定義できないパブリックな環境下では利用できないと考える．また信号対雑音比 (SNR:Signal to Noise Ratio) が低い場合や，同性の複数話者の混合音声に対しては各音声の周波数構造が曖昧になり，マスクの推定精度が大

Supervisor: Prof. Katunobu Itou

きく下がってしまう [13].

本研究では物理モデルによる手法 [5] と深層学習による手法 [6] を組み合わせ，球面調和関数展開に基づく音響特徴量を用いた深層学習による近接音／遠方音分離法を提案する．提案手法では事前処理として球面調和関数展開に基づく近接音抽出法により低周波帯域の近接音と遠方音を抽出する．その後，抽出した低域音声と混合音声を特徴量として BLSTM モデルを学習し，高周波帯域を含んだ音源分離 T-F マスクを推定することで，より高音質な近接／遠方音分離を実現する．

2 先行研究

2.1 球面調和関数展開に基づく近接音分離

近接音 $S_{t,f}$ と遠方音 $N_{t,f}$ を $M+1$ 本のマイクロホンで観測し，2 つの音源を分離することを考える．m 番目のマイクロホンで観測される信号 $X_{t,f}^{(m)}$ は次の式で表せる．

$$X_{t,f}^{(m)} = S_{t,f}^{(m)} + N_{t,f}^{(m)} \tag{1}$$

ここで t と f はそれぞれ時間と周波数のインデックスである．また $S_{t,f}^{(m)}$ と $N_{t,f}^{(m)}$ はそれぞれ m 番目のマイクロホンに到来した近接音と遠方音のスペクトログラムである．$S_{t,f}^{(m)}$ と $N_{t,f}^{(m)}$ はそれぞれ音源とマイク間の伝達関数を含むものとする．

羽田らは球面調和関数展開に基づく近接音分離法を提案した [5]．この手法では中空の球面アレイが用いられており，球の中央に 1 つのマイク（$m=0$），球の表面に M 個のマイクが等角度，等間隔に配置される．すべての入射波が平面波であると仮定すると，球面調和関数展開により，球面の中心音圧を球面上の音圧から補間できる．ここで近接音は球面波として到来するため，観測音圧と補間音圧の残差信号として近接音を得られる．

$$\hat{S}_{t,f,\mathcal{D}} = X_{t,f,\mathcal{D}}^{(0)} - \sum_{m=1}^{M} \frac{1}{J_0(kr)} \frac{1}{M} X_{t,f,\mathcal{D}}^{(m)} \tag{2}$$

ここで添字 $\mathcal{D}$ は信号がダウンサンプリングされたことを示す．$J_0(kr)$ は 0 次の球面ベッセル関数，k は波数，r は球の半径である．

球面調和関数展開に基づく音源分離では，分離可能な周波数の上限は球面アレイの半径に依存する．例えば $r=5$ cm のとき，球ベッセル関数のゼロ点が 3400Hz 付近に存在するので，ナイキスト周波数がゼロ点の周波数よりも低くなるように信号をダウンサンプリングする必要がある．また波長の大きな信号ほど短い距離で平面波に近似できるため，近接音源が少しでもマイクから離れると，低周波成分が誤って遠方音とみなされ減衰する．近接音抽出により低周波や高周波成分が欠如または減衰してしまうため，この手法を音声認識などのフロントエンド処理に直接使用することは難しい．

2.2 深層学習による T-F マスク推定

T-F マスク処理は入力音を周波数領域で分離する音源分離技術として用いられてきた．T-F マスクを用いた音源強調では観測信号に T-F マスクを乗じることで，特定の成分を強調した出力信号 $\hat{S}_{t,f}$ が得られる．

$$\hat{S}_{t,f} = G_{t,f} X_{t,f} \tag{3}$$

ここで $G_{t,f}$ は T-F マスクである．T-F マスクの推定には，多チャンネル音源を用いた手法 [14] や，非負値行列因子分解に基づく手法 [2] などがある．

また，深層学習を利用した T-F マスクの推定法も提案されている．典型的な深層学習による手法では，T-F マスク $\boldsymbol{G}_t := (G_{t,1}, ..., G_{t,F})^\top$ を次のように推定する．

$$\hat{\boldsymbol{G}}_t = \mathcal{M}(\boldsymbol{\phi}_t|\Theta) \tag{4}$$

ここで $\mathcal{M}$ は DNN や LSTM などのニューラルネットワークに基づく回帰関数であり，$\boldsymbol{\phi}_t$ は t 番目のフレームでの音響特徴ベクトル，Θ はニューラルネットワークのパラメータ，$\top$ は転置を意味する．T-F マスクを正確に予測するには，T-F マスクとの相互情報量が高い音響特徴量を使用する必要がある [15]．しかし，近接音と遠方音を分離する T-F マスクの推定に有効な音響特徴量は知られていないため，深層学習は近接音／遠方音の分離には利用されていない．

3 提案手法

先行研究 [16] では，低域音声とマスクとの対応関係に着目し，従来手法により分離した近接音と遠方音の対数メルスペクトルを特徴量とした DNN モデルにより，音声の高域を含む近接音強調を実現した．既存手法と比べ，抽出音の信号対歪率 (SDR: signal-to-distortion rate) が大きく改善したものの，PESQ(perceptual evaluation of speech quality) と STOI(short-time objective intelligibility measure)[17] がやや低下し課題の残る結果となった．

[16] では 4 層の DNN とコンテキスト処理を用いたモデルにより T-F マスクを推定していた．コンテキスト処理は約 0.2 秒間であり，トライフォンレベルの依存関係をふまえてマスクを推定するモデルを想定した．大きく抑揚のついた発話や文頭・文末の様に前後の情報が無い箇所では，分離音声の音質が低下することが確認され，PESQ や STOI の低下につながったと考える．

ここで，目的音・雑音がともに非定常な音声信号であることに着目し，重畳がない区間から重畳区間のマスクを推定できるようなモデルを考える．より長い時間の依存関係をふまえた学習が必要となる一方で，局所的な音韻の変化にもロバストなモデルが必要となる．単純にコンテキスト処理の区間を長くすると，局所的な情報の重みが小さくなってしまう可能性があるため，DNN によるマスク推定では音声の長時間の依存関係を利用しづらい．BLSTM は時系列データに対して前後の時刻の出力を再帰的に入力として利用することで，長期的な時間依存をふまえた学習が可能である．また再帰的入力に対する忘却率を学習することで，時間依存がある部分とそうでない部分で特徴量の取捨選択ができる．BLSTM とコンテキスト処理により前後の単語や文節レベルの依存関係を踏まえたマスク推定が可能である [19]．本論では BLSTM に畳み込みニューラルネットワーク (CNN) を組み合わせた CNN-BLSTM モデルによるマスク推定を考える．

また先行研究では音源位置の変化により事前分離した近接音がひずむと，マスク推定モデルの学習が難しくなり音質が悪化した．これは学習データの作成時に，音源の位置を固定してシミュレーションを行っていたことが原因である．そこで音源位置を移動させてシミュレーションを行い，音源位置や空間の変化に頑健なマスク推定モデルの学習について実験・考察する．

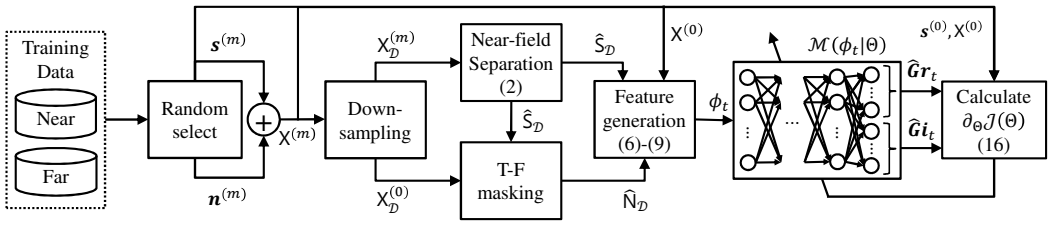

図 1. 提案手法の学習手順

3.1 音響特徴量

提案手法の音響特徴量を定義する．$\hat{S}_{t,f,\mathcal{D}}$ には事前分離で分離しきれなかった雑音成分が含まれる可能性がある．また $\hat{N}_{t,f,\mathcal{D}}$ には目的音成分が含まれる可能性がある．そこで目的音の推定値に加えて雑音の推定値も特徴量に含める．低周波帯域の雑音成分は次のように求まる．

$$\hat{N}_{t,f,\mathcal{D}} = \frac{|X^{(0)}_{t,f,\mathcal{D}}| - |\hat{S}_{t,f,\mathcal{D}}|}{|X^{(0)}_{t,f,\mathcal{D}}|} \cdot X^{(0)}_{t,f,\mathcal{D}} \tag{5}$$

目的音と雑音，混合音の対数振幅スペクトログラムを用いて，次の特徴量ベクトルを定義する．

$$\begin{aligned}\boldsymbol{\phi}_t := (&\hat{\boldsymbol{s}}_{t-C,\mathcal{D}}, \hat{\boldsymbol{n}}_{t-C,\mathcal{D}}, \boldsymbol{x}_{t-C}, ..., \\ &\hat{\boldsymbol{s}}_{t+C,\mathcal{D}}, \hat{\boldsymbol{n}}_{t+C,\mathcal{D}}, \boldsymbol{x}_{t+C})^\top\end{aligned} \tag{6}$$

$$\hat{\boldsymbol{s}}_{t,\mathcal{D}} := \ln\left(\mathrm{Abs}\left[\left(\hat{S}_{t,1,\mathcal{D}}, \hat{S}_{t,2,\mathcal{D}}, ..., \hat{S}_{t,F_d,\mathcal{D}}\right)\right]\right) \tag{7}$$

$$\hat{\boldsymbol{n}}_{t,\mathcal{D}} := \ln\left(\mathrm{Abs}\left[\left(\hat{N}_{t,1,\mathcal{D}}, \hat{N}_{t,2,\mathcal{D}}, ..., \hat{N}_{t,F_d,\mathcal{D}}\right)\right]\right) \tag{8}$$

$$\boldsymbol{x}_t := \ln\left(\mathrm{Abs}\left[\left(X^{(0)}_{t,1}, X^{(0)}_{t,2}, ..., X^{(0)}_{t,F}\right)\right]\right) \tag{9}$$

ここで C はコンテキストウィンドウのサイズであり，Abs[·] は要素ごとの絶対値を表す．F_d はダウンサンプリングした音声のナイキスト周波数に対応するインデックスである．コンテキスト処理を施した一定時間のスペクトルを特徴量に用いることで，先行音韻または後続音韻の影響を考慮したマスク推定モデルとなることを期待する．

短時間フーリエ変換で得られたスペクトログラムの各時刻での特徴量を計算し，近接音を強調する複素振幅マスクの実部 $\hat{\boldsymbol{G}}_{r,t}$ と虚部 $\hat{\boldsymbol{G}}_{i,t}$ をそれぞれ推定する．

$$\boldsymbol{H}_t = \mathcal{M}(\boldsymbol{\phi}_t|\Theta) \tag{10}$$

$$\hat{\boldsymbol{G}}_{r,t} = (H_{t,1}, H_{t,2}, \ldots, H_{t,F}) \tag{11}$$

$$\hat{\boldsymbol{G}}_{i,t} = (H_{t,F+1}, H_{t,F+2}, \ldots, H_{t,2F}) \tag{12}$$

DNN の出力次元はスペクトログラムの周波数ビンの倍に設定し，前半部を実部マスク，後半部を虚部マスクとして利用する．推定されたマスクと混合音 $\boldsymbol{X}^{(0)}_t := (X^{(0)}_{t,1}, \ldots, X^{(0)}_{t,F})^\top$ を用いて，高サンプリングレートの近接音を抽出する．

$$\hat{\boldsymbol{S}}_{r,t} = \hat{\boldsymbol{G}}_{r,t} \odot \boldsymbol{X}^{(0)}_{r,t} - \hat{\boldsymbol{G}}_{i,t} \odot \boldsymbol{X}^{(0)}_{i,t} \tag{13}$$

$$\hat{\boldsymbol{S}}_{i,t} = \hat{\boldsymbol{G}}_{r,t} \odot \boldsymbol{X}^{(0)}_{i,t} + \hat{\boldsymbol{G}}_{i,t} \odot \boldsymbol{X}^{(0)}_{r,t} \tag{14}$$

$$\hat{\boldsymbol{S}}_t = \hat{\boldsymbol{S}}_{r,t} + i\hat{\boldsymbol{S}}_{i,t} \tag{15}$$

ここで $\odot$ は要素ごとの積であり，$\boldsymbol{X}^{(0)}_{r,t}, \boldsymbol{X}^{(0)}_{i,t}$ はそれぞれ $\boldsymbol{X}^{(0)}_t$ の実部と虚部である．

3.2 目的関数

BLSTM のパラメータ Θ の学習には目的音・雑音波形の平均絶対誤差とコサイン類似度を用いた次の目的関数 $\mathcal{J}(\Theta)$ を用いた．

$$\hat{\boldsymbol{s}} = \mathrm{ISTFT}\left[\mathcal{M}(\boldsymbol{\Phi}|\Theta) \odot \mathsf{X}^{(0)}\right] \tag{16}$$

$$\hat{\boldsymbol{n}} = \boldsymbol{x} - \hat{\boldsymbol{s}} \tag{17}$$

$$\mathcal{J}_1(\Theta) = \frac{1}{K}\left\|\boldsymbol{s} - \hat{\boldsymbol{s}}\right\|_1 + \frac{1}{K}\left\|\boldsymbol{n} - \hat{\boldsymbol{n}}\right\|_1 \tag{18}$$

$$\mathcal{J}_2(\Theta) = \frac{1}{K}\left\|\alpha \cdot \cos(\boldsymbol{s}, \hat{\boldsymbol{s}}) + (1-\alpha) \cdot \cos(\boldsymbol{n}, \hat{\boldsymbol{n}})\right\|_1 \tag{19}$$

$$\mathcal{J}(\Theta) = \mathcal{J}_1(\Theta) - \mathcal{J}_2(\Theta) \tag{20}$$

$\|\cdot\|_1$ は L1 ノルム，cos(·) はコサイン類似度である．また α はフレームごとの $\mathbf{x}$ に対する目的音のパワー比である．上記の目的関数を最小化するようにパラメータを学習することで，目的音の波形の絶対誤差を小さく，かつ相関を大きくするようなマスクが推定できる．マスク処理したスペクトログラムを逆フーリエ変換し，オーバーラップ加算後の波形を見ることで，フレーム間の位相ズレによるノイズやミュージカルノイズを抑える効果を期待する．

4 評価実験

4.1 実験条件

近接音抽出法の出力音を利用した T-F マスク推定手法によって，近接音源と遠方音源の高音質な分離ができるかを確認する．評価尺度には SDR, PESQ, STOI を用いて，提案手法と従来の近接音抽出法との比較を行う．またマスク推定への事前分離音の貢献を示すために，提案手法と同じトポロジーの BLSTM に混合音声のみを特徴量として与えたモデルとの比較も行った．

4.1.1 学習データセット

学習用データの作成には JNAS 音声コーパスを使用した．コーパスに含まれる男女各 153 話者のデータを学習用の 148 話者と評価用の 5 話者にそれぞれ分け，学習用に割り当てた話者の音素バランス文発話を用いる．男女各 148 名による 14800 個の音声からランダムに目的音源と雑音音源を 15000 組選択し，これらに鏡像法 によって生成した近接と遠方の 2 パターンのインパルス応答を畳み込むことで，同じ方向の近接と遠方にある音源を作成した．鏡像法のパラメータを表 1 に，マイクと音源の位置を図 5 に示す．球面アレイは，半径 5cm で $M+1 = 33$ 個のマイク素子を持つ球面中空アレイを想定した．$m = 1, ..., 32$ 番目のマイクロホンは接頂二十面体の各面の中央にそれぞれ配置し，$m = 0$ 番目のマイクは球の中心に配置した．もとの音声のサンプリングレートは 16kHz とし，近接音抽出 [5] の前処理として 6kHz にダウンサンプリングした．

4.1.2 CNN-BLSTM の構造と設定

今回は CNN2 層と BLSTM2 層を組み合わせた全 4 層のマスク推定モデルを利用する．CNN の 1 層目ではスペクトログラムに 11x15 の 30ch フィルタをかける．これにより各時刻，各周波数ビンに対し，時間については前 5 フレームと後ろ 5 フ

表 1. 鏡像法シミュレーターの条件

パラメータ	設定値
空間の大きさ	2x2x2 m^3
残響時間 (RT_{60})	0.07 s
音速	340 m/s

オブジェクト	座標 (m)		
	x	y	z
マイクロホン	1	0.5	1
近接音	1	0.6	1
遠方音	1	1.8	1

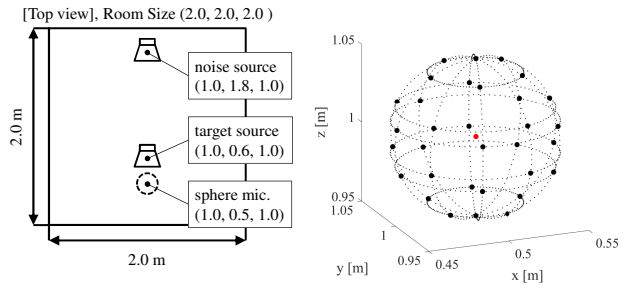

図 2. マイクロホンと音源の配置
各座標 (x, y, z)[m] はマイクと音源の位置を示す．

レーム，周波数については上 7 ビンと下 7 ビンをまとめた 30 次の特徴量ができる．2 層目は 11x15 の 2ch のフィルタにより 1 層目の出力を 60ch に変換する．それらをプーリング層により 1ch に圧縮し，BLSTM 層の入力とする．BLSTM は共にノード数 400 点の完全接続 BLSTM を用いた．出力層（T-F マスク）と隠れ層の活性化関数にはそれぞれ恒等関数とランプ関数（ReLU: rectified linear unit）を用いた．入力ベクトルと BLSTM の出力は短時間フーリエ変換（STFT: short-time Fourier transform）により変換した対数振幅スペクトログラムとした．SFTF のフレームサイズは 512 点，シフト幅は 256 点である．

4.2 評価結果

評価用データの作成には JNAS の新聞読み上げ文を用いた．ソース音源の発話話者は学習データに含まれない男性 5 名，女性 5 名で発話文は 100 種類である．ソース音源を目的音と雑音にランダムに分け，表 1 の条件でインパルス応答を畳み込み，−5，0，5dB の 3 種類の SNR で混合音を 300 サンプル作成した．目的音と雑音はそれぞれ近接音と遠方音とし，SDR，STOI，PESQ の 3 つの客観的手法を用いて従来手法 [5] と提案手法を比較した．評価結果を図 6 に示す．いずれの SNR についても従来の近接音抽出法よりも評点が向上しており，高域成分を含む分離が為されたことで音質が向上した．混合音のみを特徴量とした T-F マスク推定モデルと比較すると，より雑音が大きい SNR−5dB の条件下で音質の改善が顕著だった．このことから従来手法により分離した低域音声が，より SNR の低い厳しい条件下でのマスク推定に大きく貢献したことが分かる．目的音のスペクトルが雑音に大きく埋もれた場合，時間周波数構造のみによる目的音声の判別は困難となる．しかし近接音抽出法により分離した低域音声を用いることで，マスクの推定が容易となり分離精度の向上につながったと考える．

同様に 2.5kHz 以下の低域のみについて音質評価実験を行った．目的音，分離音のいずれにも 2.5kHz のローパスフィルタをかけて SDR，STOI，PESQ を算出した．評価結果を図 7 に示す．いずれの SNR についても従来の近接音抽出法と比べ，提案法の評点が大きく下回った．提案法の低域での分離精度が従来法に劣った原因については次のように考える．従来法は 32 + 1 チャンネルの多チャンネル信号を入力とした音源分離であるのに対して，提案法はモノラル信号を入力とするブラインド音源分離であり，直接多チャンネル信号を入力としていない．そのため低域成分を従来法と遜色なく分離するためには多チャンネル信号の全てを特徴量として利用し，従来の近接音抽出の機構を含むすべての処理を深層学習でモデル化する必要があると考える．

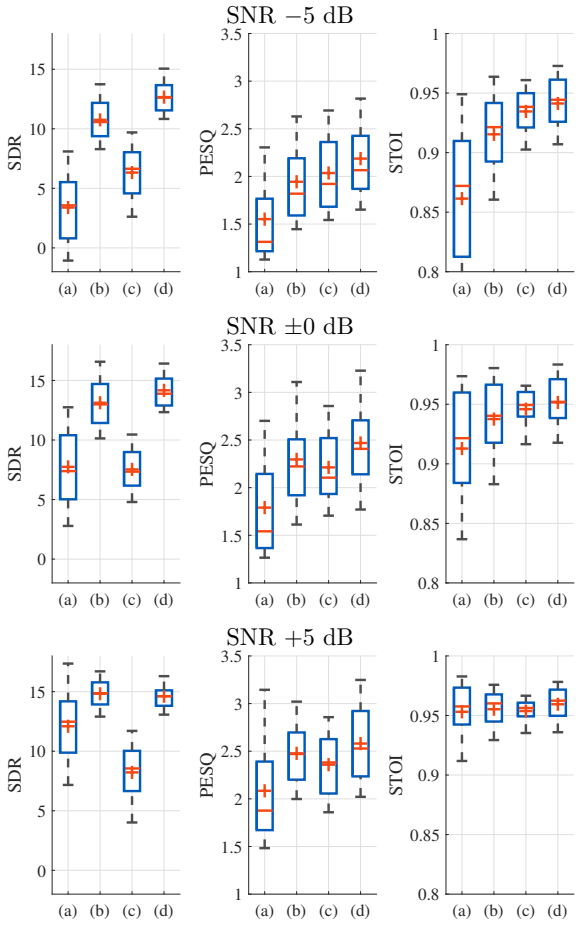

図 3. SNR−5, 0, 5dB における客観評価結果．各箱ひげ図は (a) 観測音，(b) 観測音のみを特徴量としたマスク推定モデルによる出力音，(c) 従来法の出力音 [5]，(d) 提案法の出力音についての評価値である．

4.3 考察

4.2 章では音源の配置が不変な環境を想定していたが，新たに音源の距離について可変な環境を想定した学習データを作成し，近接/遠方音分離モデルを BLSTM により学習した．

4.3.1 学習データセット

目的音源と雑音音源には JNAS 日本語新聞読み上げコーパスの音素バランス 503 文の音声を使用した．男性 148 人と女性 148 人による 14800 発話からランダムに目的音源と雑音音源を 15000 組選択し，これらに “RIR generator” [12] を用いて生成した近接と遠方の 2 パターンのインパルス応答を畳み込み，SNR を −5dB から +5dB の間の一様乱数として 2 つの音声を混合した．マイクと音源の位置を図 5(左) に示す．近接音はマイクとの距離が 0.1m から 0.5m となる位置にランダムに配置するため，0.01m 単位で作成した 41 個のインパルス応答をソース音源に畳み込むことで実装した．遠方音については 0.5m から 1.5m を 0.01m 単位で 101 個のインパルス応答を作成した．上記の工程により作成した 15000 組のデータセットを

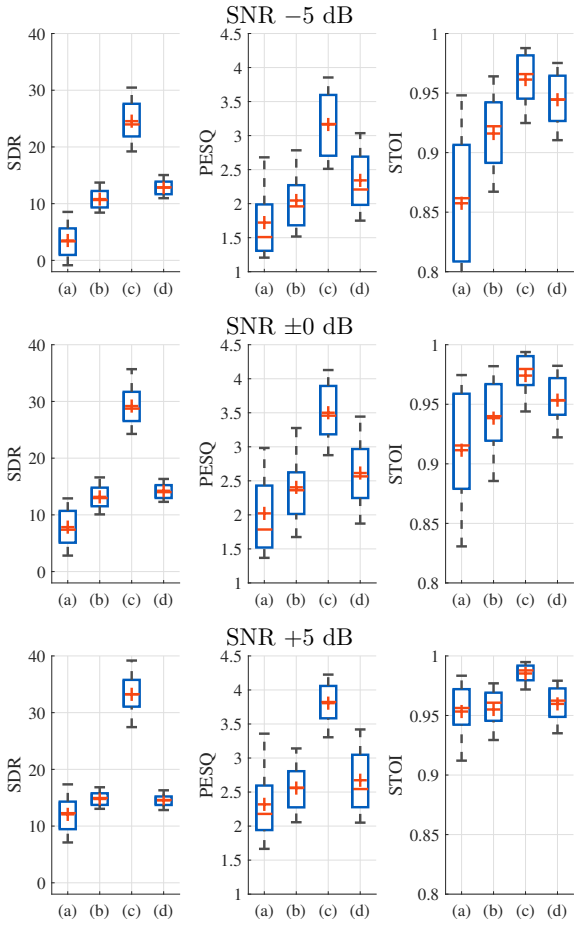

図 4. SNR$-5, 0, 5$dB における 2.5kHz 以下の成分についての客観評価結果．各箱ひげ図は (a) 観測音，(b) 観測音のみを特徴量としたマスク推定モデルによる出力音，(c) 従来法の出力音 [5]，(d) 提案法の出力音についての評価値である．

学習データとする．もとの音声のサンプリングレートは 16kHz とし，[5] の前処理として 6kHz にダウンサンプリングした．

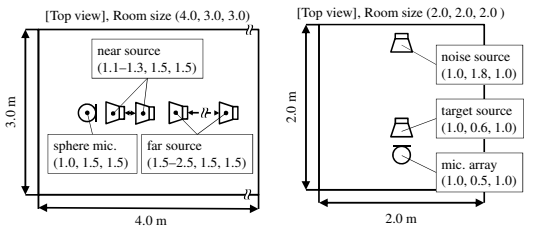

図 5. マイクロホンと音源の配置
各座標 (x, y, z)[m] はマイクと音源の位置を示す．

4.3.2 距離別音質客観評価

評価用のソースには ATR の新聞読み上げ文を用いた．ソース音源の発話話者は男性 5 名，女性 5 名で発話文は 100 種類である．これらの発話音声を目的音と雑音にランダムに分け，学習用データと同じ条件でインパルス応答を畳み込み評価用データとした．0.5m から 1.5m の間を 0.1m 単位で作成した 10 個のインパルス応答を用い，各距離条件に付き 300 個のサンプルを作成した．目的音と雑音はそれぞれ近接音と遠方音とし，SDR，STOI，PESQ の 3 つの客観的手法を用いて従来手法 [5] と提案手法を比較した．また観測音のみを特徴量とした従来のマスク推定モデル [13] との比較も行った．評価結果を図 6 に示す．

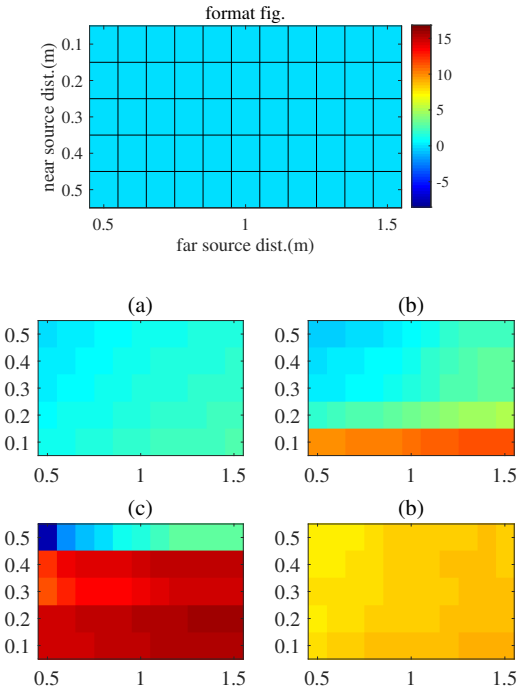

図 6. 近接音/遠方音の各距離条件における抽出近接音の平均 SDR．(a) 観測音，(b) 従来法の出力音 [5]，(c) 提案法，(d)PIT-CNN-BLSTM[13] の出力音についての評価値である．

いずれの手法についても遠方音がマイクに近づくほど音質が悪化した．従来法では 0.7m 付近からグラフの傾きが急になっており，音源固定（近接音 0.1m，遠方音 1.3m のみ）で学習した音源不変モデルでも同様の特徴がみられた．一方で新たに学習した音源可変モデルは，上記の手法と比べて 0.7m 以下での悪化が緩やかになっており，学習データに含まれる範囲であれば，従来法の事前分離音の音質の悪化に対応できることが分かった．ただし，遠方音が 0.5m まで近づくと分離前の混合音よりも SDR，STOI が悪化してしまうため，それよりも遠方音が近い場合は有効な特徴量として機能しないことが予想される．

次に空間条件の異なる環境での分離性能を評価した．想定した空間条件，マイクと音源の位置を図 5(右) に示す．SNR やソースについては前述の評価データと同様に作成した．評価結果を図 7 に示す．音源位置固定モデルでは PESQ の平均値が従来法以下まで大きく低下しているのに対して，音源位置をランダムに選択したモデルでは従来法を上回る評点となった．また，SDR と STOI の平均もわずかに向上しており，特に SNR が低い条件下では評価データと同じ空間で学習したモデルと同等の SDR まで向上した．これらから，学習データの作成時にサンプルごとに距離を変化させることで空間の違いによる分離精度の低下を抑制できることが分かった．特に SNR が低い（遠方音が強い）条件下で音質の改善がみられ，これは遠方音を移動させたことで雑音音源の空間特性の変化への頑健性が高まった結果といえる．

5 あとがき

球面調和関数展開による音源分離手法と深層学習による音源分離手法を組み合わせた近接音抽出を提案した．抽出音声の音質改善を目的とし，BLSTM の特徴量と目的関数を検証した．

実験の結果，従来の近接音抽出法や T-F マスク推定法と比べ

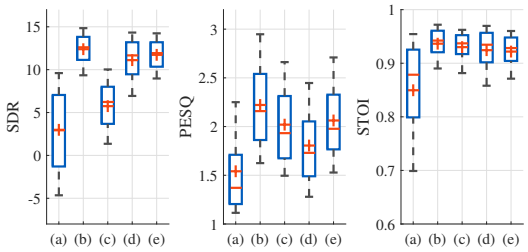

図 7. 異なる空間条件の評価データに対する平均音質評点．(a) 観測音，(b) 評価データと同じ空間で学習したモデル，(c) 従来法の出力音 [5]，(d) 提案法（音源不変モデル），(e) 提案法（音源可変モデル）の出力音についての評価値である．

て SDR，PESQ，STOI いずれの評点においても大きな改善が見られた．今後の課題として，LSTM を用いたマスク推定モデルへの本手法の応用と，多チャンネル信号を入力とする近接音抽出とマスク推定の処理を同時に行う深層学習モデルを検討する．また実環境への応用に向けて実機のマイク数やアレイ半径での近接音抽出法のシミュレーションと，それにより得られた低域音声を用いたマスク推定を実施する必要がある．

参考文献

[1] M. Brandstein et al., “Microphone Arrays,” Springer, 2001.

[2] P. Smaragdis et al., “Non-negative matrix factorization for polyphonic music transcription,” in Proc. WASPAA, 2003.

[3] D. Kitamura, et al, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization” IEEE/ACM Trans. Audio, Speech and Language Processing, pp.1626–1641, 2016.

[4] Rodney A. Kennedy, Thushara D. Abhayapala, and Darren B. Ward, “Broadband Nearfield Beamforming Using a Radial Beampattern Transformation,” in IEEE Trans. Signal Processing, 1998.

[5] Y. Haneda, et al., “Cloase-talking spherical microphone array using sound pressure interpolation based on spherical harmonic expansion,” in Proc of ICASSP, 2014.

[6] H. Erdogan, et al., “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in Proc. ICASSP, 2015.

[7] K. Niwa, et al., “Pinpoint extraction of distant sound source based on DNN mapping from multiple beamforming outputs to prior SNR” in Proc. ICASSP, 2016.

[8] Y. Koizumi, et al., “DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements,” in Proc. ICASSP, 2017.

[9] Y. Koizumi, et al., “DNN-based source enhancement to increase objective sound quality assessment score,’’ IEEE Trans. ASLP, 2018.

[10] Y. Ephraim et al., “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Trans. Audio, Speech and Language Processing, pp.1109–1121, 1984.

[11] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, T. Nakatani, “Single Channel Target Speaker Extraction and Recognition with Speaker Beam,” in Proc. ICASSP, pp.5554–5558, 2018.

[12] E. A. P. Habets, “Room impulse response generator,” `https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator/`.

[13] Morten Kolbk and Dong Yu, “Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks,” IEEE/ACM Transactions on Audio, Speech and Language Processing, pp.1901–1913, 2017.

[14] Y. Hioka, et al., “Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain,” IEEE Trans. Audio, Speech and Language Processing, pp.1240–1250, 2013.

[15] Y. Koizumi, et al., “Informative acoustic feature selection to maximize mutual information for collecting target sources,” IEEE/ACM Trans. Audio, Speech and Language Processing, pp.768–779, 2017.

[16] S. Nishiguchi, et al., “DNN-based Near- and Far-field Source Separation Using Spherical-harmonic-analysis-based Acoustic Features,” IWAENC, pp.510–514, 2018.

[17] C. H. Taal, et al., “An algorithm for intelligibility prediction of time-frequency weighted noisy speech,” IEEE Transactions on Audio, Speech and Language Processing, pp.2125–2136, 2011.

[18] ITU-T “P.862 : Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”

[19] Hakan Erdogan and Takuya Yoshioka, “Investigations on Data Augmentation and Loss Functions for Deep Learning Based Speech-Background Separation,” Interspeech 2018, pp.3499–3503, 2018.