5Q-04

円形マイクアレーを想定した 球面調和関数展開に基づく近接／遠方音分離 T-F マスク推定 *

西口 草太 †, 小泉 悠馬 ‡, 原田 登 ‡, 伊藤 克亘 §

1 序論

音源分離は音声認識や異常音検知などのフロントエンド処理として研究されてきた．ほとんどの音源分離手法は方向 [1] やスペクトル [2]，またはその両方 [3] に焦点を当てて目的音とノイズを分離している．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．

先行研究 [4] では，羽田らの提案した近接音抽出手法 [6] の課題である高域での音源分離のために，既存手法で分離した低域音声を音響特徴量としたディープニューラルネットワーク (DNN) を利用し，高域を含んだ音源分離マスクを推定することを考えた．鏡像法シミュレーションにより 32 個のマイク素子を持つ球面マイクロホンアレー信号をシミュレートし，既存手法で分離可能な上限周波数は 3kHz として実験を行った．

本論では実機を想定した 7ch の円形マイクアレーを用いたシミュレーション実験を行い，特徴量として利用できる低域音声の上限がより低い場合の近接音／遠方音分離に取り組む．

2 提案手法

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 $M+1$ 個のマイク素子を搭載した球面アレイで観測し，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である．

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．近接音は次の式で得られる [6].

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

ここで添え字 $\mathcal{D}$ 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．この手法では中空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．円形マイクアレーを用いた場合，円周上に M 個のマイクと中心に 1 つのマイクが配置され，音源が円の中心を通る垂直線上にあれば，式 (2) により近接音が得られる．

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アレー半径とマイク数，球面調和関数の最大展開次数に依存する．本論では $r = 4\text{cm}$，$M = 6$ の場合を想定する．この場合，球面調和関数展開の最大展開次数は $N = 1$(N は $(N+1)^2 < M$ を満たす最大の整数) である [6]．このとき球面調和関数展開に基づく近接音抽出法の上限周波数はおよそ 1200Hz となり，この手法によって分離した音声を音声認識や話者識別などの信号処理に直接使用することは難しい．

2.2 DNN による近接音分離マスク推定

先行研究 [4],[5] では，2.1 章の低サンプリングレート音声を特徴量とした DNN により，近接音の全帯域 T-F マスク (時間周波数マスク) を推定することで高域を含めた近接音分離を実現した．本論では低サンプリングレート音声の上限周波数が低いため，マスク推定モデルの構造についても検討する必要がある．

まず音響特徴量 $\boldsymbol{\phi}_t$ を次のように定義する．

$$\boldsymbol{\phi}_t := (\hat{\boldsymbol{s}}_{t-C,\mathcal{D}}, \hat{\boldsymbol{n}}_{t-C,\mathcal{D}}, \boldsymbol{x}_{t-C}, ..., \hat{\boldsymbol{s}}_{t+C,\mathcal{D}}, \hat{\boldsymbol{n}}_{t+C,\mathcal{D}}, \boldsymbol{x}_{t+C})^\top \tag{3}$$

$$\hat{\boldsymbol{s}}_{t,\mathcal{D}} := \ln\left(\text{Abs}\left[\left(\hat{S}_{t,1,\mathcal{D}}, \hat{S}_{t,2,\mathcal{D}}, ..., \hat{S}_{t,F_C,\mathcal{D}}\right)\right]\right) \tag{4}$$

$$\hat{\boldsymbol{n}}_{t,\mathcal{D}} := \ln\left(\text{Abs}\left[\left(\hat{N}_{t,1,\mathcal{D}}, \hat{N}_{t,2,\mathcal{D}}, ..., \hat{N}_{t,F_C,\mathcal{D}}\right)\right]\right) \tag{5}$$

$$\boldsymbol{x}_t := \ln\left(\text{Abs}\left[\left(X_{t,1}^{(0)}, X_{t,2}^{(0)}, ..., X_{t,F}^{(0)}\right)\right]\right) \tag{6}$$

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．$\hat{N}_{t,f,\mathcal{D}}$ は低周波帯域のノイズ成分であり，$\hat{S}_{t,f,\mathcal{D}}$ を用いたフィルタ処理により得られる．また，F_C は特徴量として利用する周波数の上限である．

DNN のパラメータ Θ は次の平均絶対誤差（MAE）を最小化するように学習される．

$$\mathcal{J}(\Theta) = \frac{1}{K} \left\| \boldsymbol{s} - \text{ISTFT}\left[\mathcal{M}(\boldsymbol{\Phi}|\Theta) \odot \mathrm{X}^{(0)}\right] \right\|_1 \tag{7}$$

$$\mathrm{X}^{(0)} := \{\mathbf{X}_1^{(0)}, \cdots, \mathbf{X}_T^{(0)}\},\ \mathbf{X}_t^{(0)} := \left(X_{t,1}^{(0)}, ..., X_{t,F}^{(0)}\right)^\top \tag{8}$$

ここで $\odot$ は要素ごとの積であり，$\|\cdot\|_p$ は L_p ノルム，$\boldsymbol{s} \in \mathbb{R}^K$ は時間領域の目的音，$\boldsymbol{\Phi} := \{\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_T\}$ である．また ISTFT[·] は逆短時間フーリエ変換である．

3 評価実験

提案手法の性能を客観評価により検討した．評価尺度には SDR，PESQ，STOI を用いて，提案手法と従来手法の比較を行った．

---

*: Near- and Far-speech Separation by T-F mask Estimation using Separated Speeches based on Spherical-harmonic-analysis with a Circular Microphone Array, *Sota Nishiguchi (Hosei Univ.) et al.*

† 法政大学大学院 情報科学研究科
‡ NTT
§ 法政大学 情報科学部

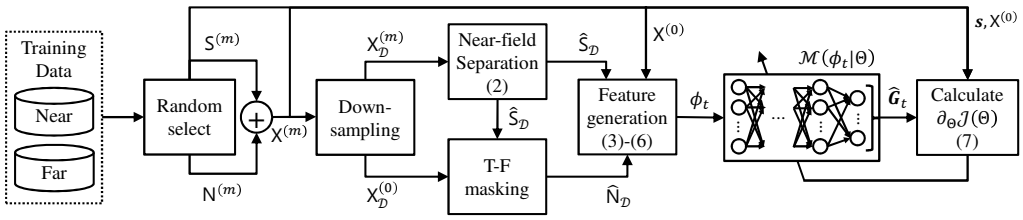

図 1. 提案手法の学習手順

3.1 学習データセット

目的音源とノイズ音源には ATR 日本語音声データベースの A セットから J セットの音声を使用した．男性 22 人と女性 22 人による 6640 発話の半分を目的音源，残りの半分を雑音音源にランダムに割り当てた．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．残響時間（RT_{60}）は 0.07s として，目的音源をマイクから 0.1m 離れた位置に配置し，ノイズ音源は 1.3m に配置した． 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，近接音と遠方音の混合音声を作成した．球面アレイは，半径 4cm で $M+1=7$ 個のマイク素子を持つ円形アレイを想定した．$m=1,...,6$ 番目のマイクロホンは円周上に等間隔で配置し，$m=0$ 番目のマイクは円の中心に配置した．もとの音声のサンプリングレートは 16kHz とし，近接音抽出 [6] の前処理として 2kHz にダウンサンプリングした．

3.2 DNN の構造と設定

提案手法では，ノード数 512 点の隠れ層 4 層で構成された完全接続の DNN を使用した．出力層（T-F マスク）と隠れ層の活性化関数にはそれぞれシグモイド関数とランプ関数（ReLU: rectified linear unit）を用いた．またコンテキストサイズは $C=10$ とした．STFT のフレームサイズは 512 点，シフト幅は 256 点とした．

3.3 客観評価

SDR，STOI，PESQ の 3 つの客観的手法を用いて先行研究 [4] と提案手法を比較した．評価用データの作成には JNAS の音素バランス文を用いた．ソース音源には評価用の男女各 5 名の話者による 100 種類の発話文を用いた．これらの発話音声から目的音と雑音をランダムに選択し，学習用データと同様に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．評価結果を表 1 に示す．数値は全てのテストデータに対する評点の平均である．

提案法では従来の近接音抽出法よりも PESQ と STOI が向上し，SDR は従来法に及ばない結果となったものの混合音からの改善値で見ると 32ch 球面アレーを用いた先行研究 [4] と同等以上であった．SDR は波形の歪みに起因する評点であり低域成分に強く依存するために従来法でも高い値となった．また従来法では上限周波数付近に推定誤差によるノイズ成分が生じるため，従来法の境界周波数に対して余裕をもってダウンサンプリングしたことでノイズ成分が小さくなり，抽出音の SDR の低下が抑えられたことが要因であると考える．

表 1. 音質評価尺度の評点

出力音声		音質評価尺度		
低域	高域	SDR	PESQ	STOI
混合音		1.64	1.26	0.81
従来法		12.23	1.68	0.78
従来法	混合音	11.80	1.59	0.90
提案法		10.44	1.90	0.90
従来法	提案法	13.54	2.06	0.92
cIRM		21.90	3.55	0.985

[4] の結果と比べ，従来法，提案法のいずれも PESQ が低下している．分離音を音声認識などに利用することを考えると，PESQ や STOI のさらなる向上が必要となる．提案法では音韻レベルのコンテキストを踏まえた特徴量を用いて，話者の声質や各音韻に対して低域音声から全帯域のマスクを推定するモデルを想定した．この場合，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た．スペクトルの時間周波数構造について，長短期記憶ユニット (LSTM) のような長期的な依存を学習するモデルを考えることで，単語や文節レベルのコンテキストを考慮したマスク推定モデルが学習できる [7]．本提案手法と同様に，LSTM による T-F マスク推定モデルの特徴量に事前分離音を利用することでさらなる音質の向上が期待できる．

4 結論

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．

実験の結果，先行研究よりも厳しい上限周波数でも音源分離により SDR が向上し，PESQ，STOI の改善も見られた．また，提案法の低域を従来手法の抽出音に差し替えることで，音質が大幅に改善された．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．

参考文献

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

[2] P. Smaragdis et al., Proc. WASPAA, pp.177–180, 2003.

[3] D. Kitamura, et al, IEEE/ACM Trans. Audio, Speech and Language Processing, pp.1626–1641, 2016.

[4] S. Nishiguchi, et al., IWAENC, pp.510–514, 2018.

[5] S. Nishiguchi, et al., 情報処理学会全国大会講演論文集, pp.557–558, 2019.

[6] Y. Haneda, et al., Proc. ICASSP, pp.604–608, 2014.

[7] H. Erdogan, et al., Interspeech 2018, pp.3499–3503, 2018.