4V-04

球面調和関数に基づく距離別分離音を用いた深層学習による近接音声分離 *

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，羽田らの提案した手法 [5] の課題である高周波数成分の分離のために，既存手法で分離した低域成分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．その結果，抽出音声の信号対歪率 (SDR: signal-to-distortion rate) が大きく向上した．

本論文では上記の特徴量に混合音を加えてマスク推定を行う．またマスク処理後の波形の絶対誤差を目的関数に用いることで，抽出音の音質向上を図る．

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 [5].

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

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．

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．本論では $r = 5$ cm の場合を想定する．このときの上限周波数はおよそ 3.4kHz となり，この手法を音声認識のようなフロントエンド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] 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．既存手法と比べ，抽出音の SDR が大きく改善したものの，PESQ(perceptual evaluation of speech quality) と STOI(short-time objective intelligibility measure)[6] がやや低下し課題の残る結果となった．

そこで本論では DNN マスク推定モデルに 3 つの変更を加える．1 つ目の変更として特徴量に混合音のスペクトル情報を追加する．先行研究ではメルフィルタバンク領域でのマスク推定を行い，低域情報から全域のマスクを推定した．混合音のスペクトルを追加しスペクトル領域でのマスク推定を行うことで，高域成分をより厳密に抽出することができると考える．2 つ目に目的関数をスペクトル領域での二乗誤差から，マスク処理後の波形領域での絶対誤差に変更する．マスク処理したスペクトログラムの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．3 つ目に DNN マスク処理の出力を高域のみに限定し，既存手法の低域音声に時間領域で合成する．低域に関しては既存手法のほうが分離性能が優れていたため，高域のみをマスク処理することで音質改善を図る．

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

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

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

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

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

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．

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

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

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

---

*: Near-speech Separation by Deep-Learning using Separated Speeches based on Spherical-harmonic-analysis, *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

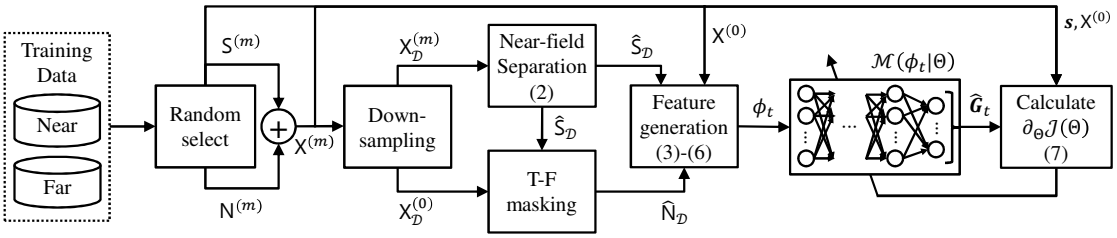

図 1. 提案手法の学習手順

来手法の比較を行った．

3.1 実験条件

3.1.1 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音声を使用した．男性 11 人と女性 11 人による 2200 発話の半分を目的音源，残りの半分をノイズ音源にランダムに分割した．これらに “RIR generator” [7] を用いて生成した 2 パターンのインパルス応答を畳み込み，さらに $-6, 0, 6$dB の 3 つの SNR で目的音とノイズを混合して学習データを作成した．狭くて残響の少ない “*Room A*”（$RT_{60} = 0.07$s）と，広くて残響の長い “*Room B*”（$RT_{60} = 0.5$s）を想定し，RIR generator のパラメータを設定した．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 “*Room A*” では 1.3m，“*Room B*” では 5.5m マイクから離して配置した．テストデータには男性 3 名，女性 3 名の 300 発話からなる日本語音声データベースを用いた．これらの発話音声を目的音とノイズ音にランダムに分け，学習データと同様にインパルス応答を畳み込んだ．

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た．もとの音声のサンプリングレートは 16kHz とし，(2) の前処理として 6kHz にダウンサンプリングした．

3.1.2 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 $C = 5$ 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．

3.2 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．目的音とノイズはそれぞれ近接音と遠方音とし，3 種類の SNR 条件と 2 種類の空間条件で作成したテストデータを用いて評価を行った．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．提案手法 5. は先行研究と比べ SDR と STOI が改善された．目的関数を波形の絶対誤差に変更したことで，フレーム間の位相ずれやミュージカルノイズが低減されたことが分かる．

提案手法 6. では PESQ と STOI が大幅に改善された．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，従来法と提案法の境界周波数を下げることで SDR の低下は防ぐことができると考える．また IBM Text-to-Speech を用いて抽出音声の音声認識実験を行った．表 1 の誤り率は認識結果の単語誤り率である．提案手法 6. が最も誤り率が低く，先行研究の認識誤りを約 30% 削減した．

表 1. 各手法における音質評価尺度と音声認識誤り率

手法	出力音声 低域	出力音声 高域	音質評価尺度 SDR	PESQ	STOI	誤り率 (%)
1.	混合音		3.04	1.799	0.810	84.6
2.	従来法	なし	8.53	2.600	0.920	41.9
3.	従来法	混合音	8.74	2.605	0.938	25.6
4.	先行研究 [4]		9.56	2.599	0.907	24.9
5.	提案法		10.60	2.584	0.917	28.5
6.	従来法	提案法	8.46	2.650	0.939	16.2
ドライソース			-	-	-	4.5

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．先行研究の課題であったミュージカルノイズの低減と音質改善を目的とし，DNN の特徴量と目的関数に変更を加えた．

実験の結果，先行研究と比べて SDR が約 1dB 向上し，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，PESQ と STOI が大幅に改善され，音声認識誤り率を先行研究から約 30% 改善した．今後の課題として，音声認識率をさらに向上させるために，スペクトル包絡の誤差を最小化するような DNN の学習を検討する．

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

[6] C. H. Taal, et al., IEEE Trans. Audio, Speech and Language Processing, pp.2125–2136, 2011.

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