E-002

空間スペクトルを用いた音声強調のための教師あり時間周波数マスク推定
Supervised Time-frequency Mask Estimation Using Spatial Spectrum for Speech Enhancement

市村 匡輝† 武田 龍‡ 駒谷 和範‡
Masaki Ichimura Ryu Takeda Kazunori Komatani

1 はじめに

本研究では，複数のマイクロホンを用いた音声強調の高精度化を目指している．音声強調は雑音を含む入力信号から目的の音声信号を抽出する技術であり，耐雑音性を備えた音声認識には必要不可欠な技術である．一般的には，周囲の環境音や非目的話者の音声が雑音信号となる．本稿では，複数話者の音声が混合した信号を音声強調の対象とする．

深層学習に基づく時間周波数マスク推定とビームフォーマを組み合わせた手法が音声強調技術として有望である [1]．この手法では，入力信号から抽出した特徴量に基づき雑音成分を抑圧するマスクが推定される．マスクを予測するモデルは，事前に教師あり学習で構築される．マスクの推定精度は音声強調の性能を左右するため，予測に用いる特徴量の設計が問題となる．

本研究では，空間情報を用いることで，雑音成分抑圧のためのマスクの推定を高精度化する．空間情報として，方向毎の音の到来可能性を表す空間スペクトルを用いる．これを従来用いられてきたパワースペクトルに基づく特徴量に加えて用い，マスクを高性能に推定する．本稿では，空間スペクトルを始めとする特徴量を有効に活用するため，画像のセグメンテーションタスクで提案された U-Net [2] を用いて，空間スペクトルの有効性を検証する．

2 時間周波数マスクに基づくビームフォーマ

音声強調はビームフォーマで行う．ビームフォーマのフィルタ推定では，雑音の情報を用いる必要がある．雑音成分を抽出するために時間周波数マスクを用いる．

2.1 ビームフォーマ

音源が移動しない場合，時間周波数領域において，フレーム t，周波数ビン f における M チャネルのマイク観測信号 $\boldsymbol{z}_{ft} = [z_{1,ft}, \cdots, z_{M,ft}]^{\mathrm{T}}$ は次式で表現される．

$$\boldsymbol{z}_{ft} = \boldsymbol{a}_f s_{ft} + \boldsymbol{n}_{ft} \tag{1}$$

ここで，s_{ft} は目的信号を表し，$\boldsymbol{n}_{ft} = [n_{1,ft}, \cdots, n_{M,ft}]^{\mathrm{T}}$ は雑音信号を表す．また，$\boldsymbol{a}_f = [a_{1,f}, \cdots, a_{M,f}]^{\mathrm{T}}$ はステアリングベクトルと呼ばれる，伝達関数を表すベクトルである．

ML (Maximum Likelihood) ビームフォーマは，雑音源による空間相関行列 $\boldsymbol{K} = E[\boldsymbol{n}_{ft}\boldsymbol{n}_{ft}^{\mathrm{H}}]$ を用いて雑音源の方向に対して死角を形成することで，目的音方向の音を強調する．このビームフォーマの出力 $\hat{s}_{ft}$ は，重み $\boldsymbol{w}_{\mathrm{ML}\,f}$ を用いて次式で計算される．

$$\hat{s}_{ft} = \boldsymbol{w}_{\mathrm{ML}\,f}^{\mathrm{H}} \boldsymbol{z}_{ft}, \boldsymbol{w}_{\mathrm{ML}\,f} = \frac{\boldsymbol{K}^{-1}\boldsymbol{a}_f}{\boldsymbol{a}_f^{\mathrm{H}}\boldsymbol{K}^{-1}\boldsymbol{a}_f} \tag{2}$$

† Graduate School of Engineering, Osaka University
‡ The Institute of Scientific and Industrial Research, Osaka University

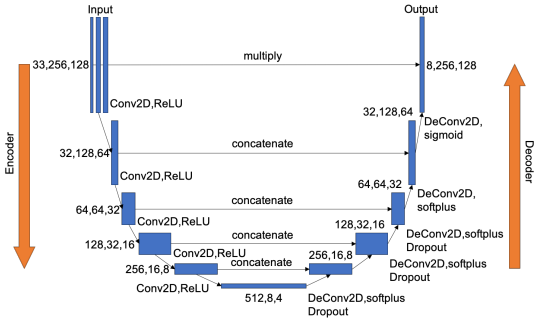

図 1: U-Net の構造とパラメータ設定

ここで，$\{\cdot\}^{\mathrm{H}}$ は複素共役転置を表す記号である．

雑音信号 $\boldsymbol{n}_{ft}$ を直接得ることはできないため，実際には観測信号から推定した雑音信号 $\hat{\boldsymbol{n}}_{ft}$ を空間相関行列の計算に用いる．この雑音信号の推定には，深層学習モデルで予測された時間周波数マスク $\lambda_{l,ft}$ を用いる．l はマイクロホンのチャネルに対応する番号である．このマスクは，スペクトル $z_{l,ft}$ の成分が雑音である確率を表し，雑音成分の可能性が高い場合は 1 に近い値を取る．雑音信号は要素毎にマスクを乗ずることで推定される．

$$\hat{\boldsymbol{n}}_{ft} = [\lambda_{1,ft} z_{M,ft}, \cdots, \lambda_{M,ft} z_{M,ft}]^{\mathrm{T}} \tag{3}$$

2.2 U-Net を用いた時間周波数マスク推定

時間周波数マスクを予測するモデルは予め教師あり学習で構築される．モデルの学習には，入力信号と教師信号のペアが必要となる．教師信号となるマスクとして，IRM (Ideal Ratio Mask)[3] を与える．目的信号 s_{ft} と雑音信号 $n_{l,ft}$ を用いて，IRM $\lambda_{l,ft}$ は次のように表される．

$$\lambda_{l,ft} = \left(\frac{|s_{ft}|^2}{|s_{ft}|^2 + |n_{l,ft}|^2} \right) \tag{4}$$

U-Net は，観測信号のパワースペクトル $|z_{l,ft}|$ を二次元画像と見なすことで，音声抽出タスクにおいても高い分離性能を発揮している [4]．図 1 に [4] に基づく U-Net の構造を示す．図の数字は，各層におけるチャネルと二次元データのサイズを表している．基本的には，Encoder 層と Decoder 層を組み合わせた U 型の畳み込みネットワークで構成される．Encoder 層では，チャネル数を 2 倍にしつつ，サイズを $\frac{1}{2}$ にする convolution を行う．Decoder 層では，Encoder 層における同じサイズの情報を concatenate することで活用し，チャネル数を $\frac{1}{2}$ にしつつ，サイズを 2 倍にする deconvolution を行う．出力層では，活性化関数に sigmoid 関数を用いる．

3 提案手法

本研究では，目的となる音源の到来方向は既知であると仮定する．これは，抽出すべき音声とその他の音声を区別するためである．また，音源は移動しないとする．

3.1 目的音方向および音源方向特徴量

音声信号における空間スペクトルとは，音源の属性と音の到来方向との関係を表した値である．到来方向 θ，音源数を N とした場合の MUSIC (MUltiple Signal Classification) 法による空間スペクトルを $P_{\theta,f}$ で表す．

$$P_{\theta,f} = \|\boldsymbol{a}_{\theta,f}\|^2 / \sum_{i=N+1}^{M} \left|\boldsymbol{a}_{\theta,f}^{\mathrm{H}} \boldsymbol{e}_i\right|^2 \quad (5)$$

ここで，$\boldsymbol{a}_{\theta,f}$ は方向 θ におけるステアリングベクトルを表す．また，$\boldsymbol{e}_i$ は，空間相関行列 $\boldsymbol{R} = E[\boldsymbol{z}_{ft}\boldsymbol{z}_{ft}^{\mathrm{H}}]$ の i 番目に大きい固有値に対応する固有ベクトルを表す．

空間スペクトルを基に，目的音の到来方向を基準とした音源方向特徴量を算出する．マイクロホンの位置を基にした絶対的な方向情報を用いた場合には，目的音方向と雑音方向の区別がつかないためである．まず，ある設定した長さのフレーム毎に空間スペクトルを算出し，その間の値を一定とした $P_{\theta,ft}$ を計算する．次に，θ は，目的音の到来方向 α から時計回りにある角度 ϕ ずつ，2π 分を設定する．音源方向特徴量を $\boldsymbol{P}_{ft} = [P_{\alpha,ft}, P_{\alpha+\phi,ft}, P_{\alpha+2\phi,ft}, \cdots, P_{\alpha+2\pi,ft}]^{\mathrm{T}}$ と表す．

3.2 学習モデルへの反映

観測信号のパワースペクトル $|z_{ft}|$ や $\boldsymbol{P}_{ft}$ といった，時間周波数上の情報をチャネルの軸方向に concatenate してまとめ，U-Net のチャネル部分に拡張して入力する．各層では 2 次元の畳込み演算が行われ，各入力と各出力のチャネル毎に別の重みフィルタが用意される．これにより，チャネル間での関係性を考慮した学習を行うことが見込まれ，結果として特徴量間の情報を有効に活用されることが期待できる．

4 評価実験

4.1 実験設定

目的音と雑音のベースとなるモノラル音声は，J-NAS (Japanese Newspaper Article Speech) の新聞記事読み上げコーパスを使用した．サンプリング周波数は 16,000[Hz] である．男女 10 人ずつ，それぞれ 100 発話を用いて，SNR (Signal-to-Noise Ratio) を 10〜30[dB] でランダムに設定し，無響室環境で 8ch のマイクロホンによって収録された方向解像度が 1° のインパルス応答を用いて，最低でも 20° 異なるランダムな到来方向で，2000 の混合音を合成した．そのうち，80%を学習用データとし，10%ずつ検証用データとテスト用データに割り当てた．

混合音毎にパワースペクトル，空間スペクトルを算出し，対数を取った上で最小値 0，最大値 1 に正規化する．空間スペクトルの方向解像度は 20° とし，120 フレーム毎に算出した．また，パワースペクトルでは失われている位相情報を加えるため，IPD (Intermicrophone Phase Difference) を算出し，同じく正規化を行う．パワースペクトルに対し，空間スペクトルと IPD をそれぞれ concatenate するか否かを組み合わせたものを入力特徴量とする．ネットワークには，1 フレームをシフトさせながら，128 フレームずつを切り出して入力する．短時間周波数変換のパラメータには，フレーム幅 512[pt]，シフト幅 128[pt]，窓関数にハミング窓を使用した．

U-Net のパラメータは図 1 のように設定した．学習におけるミニバッチサイズは 128，誤差関数には L1 ノルムを使用した．Adam optimizer を用いて学習を行い，そのパラメータは推奨値を用いた [5]．epoch 数は 150 とした．検証用データで誤差関数値が最小となったパラメータセットを性能評価に用いた．

表 1: 実験結果

特徴量		SDR[dB]		マスク
空間スペクトル	IPD	測定値	改善量	RMSE
✓	✓	**8.46**	**2.85**	0.242
✓		8.44	2.83	**0.232**
	✓	8.40	2.79	0.250
		8.32	2.71	0.239

ビームフォーマの出力音声信号と推定したマスクの評価を行う．前者は，信号の歪みの大きさを評価する尺度の SDR (Signal-to-Distortion Ratio)，後者は，IRM との誤差を評価する RMSE (Root Mean Square Error) を用いる．SDR は混合音毎に次式で表される．

$$SDR = 10\log_{10}\left(\sum_{ft}|s_{ft}|^2 / \sum_{ft}\left\{|s_{ft}| - c|\hat{s}_{ft}|\right\}^2\right) \quad (6)$$

$$c = \sqrt{\sum_{ft}|s_{ft}|^2 / \sum_{ft}|\hat{s}_{ft}|^2} \quad (7)$$

4.2 実験結果

実験結果を表 1 に示す．改善量は，混合信号に対する SDR と音声強調後の SDR との差を表している．それぞれの値は各混合音における値の平均値を示している．

マスク推定性能 (RMSE) は，空間スペクトルを加えた場合が最も良い結果となった．空間スペクトルと IPD を加えた場合と IPD を加えた場合，空間スペクトルを加えた場合とパワースペクトル単体とした場合をそれぞれ比較すると，前者では 0.008，後者では 0.007 改善した．

音声強調性能 (SDR) は，空間スペクトルと IPD の両方を加えた場合が最も良い結果となった．空間スペクトルと IPD を加えた場合と IPD を加えた場合，空間スペクトルを加えた場合とパワースペクトル単体とした場合をそれぞれ比較すると，前者では 0.06，後者では 0.12 改善した．

5 おわりに

時間周波数マスク推定に基づくビームフォーマについて，空間スペクトルの有無による性能差を評価した．今後は，残響を含むような実環境に近い音声に対する空間スペクトルの有効性の検証を行う．

参考文献

[1] J. Heymann et al., “Neural network based spectral mask estimation for acoustic beamforming,” in *proc. of ICASSP*, 2016, pp. 196–200.

[2] O. Ronneberger et al., “U-net: Convolutional networks for biomedical image segmentation,” in *proc. of MICCAI*, 2015, vol. 9351 of *Lecture Notes in Computer Science*, pp. 234–241.

[3] Y. Wang et al., “On training targets for supervised speech separation,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[4] A. Jansson et al., “Singing voice separation with deep u-net convolutional networks,” in *proc. of ISMIR*, 2017, pp. 745–751.

[5] D. P. Kingma et al., “Adam: A method for stochastic optimization,” in *proc. of ICLR*, 2015.