1C-06

移動ロボットのための 時変空間相関行列推定に基づく多チャネル音声強調

藤田 雅彦　　坂東 宜昭　　佐々木 洋子　　大西 正輝
産業総合技術研究所　人工知能研究センター

1. はじめに

展示会場や科学館など混雑環境で働く自律移動ロボット (図 1-(a)) は，来訪者の呼びかけや問いかけに対し適切に応答できることが望ましい．雑踏環境下で頑健に音声を認識するには，不要な雑音を抑圧し認識対象の音声のみを抽出する音声強調が不可欠である [1]．このような環境での音声強調では，深層ニューラルネットワーク (DNN) とビームフォーマを活用する多チャネル音声強調が高い性能を発揮している [2,3]．

自律移動ロボットにおける音声強調では，ロボット自身の移動や歩きながら話しかけてくる話者など，音源の相対的な移動を考慮しなければならない．移動音源の強調では，時変の空間相関行列 (SCM) を用いるビームフォーマが有効である [4]．この手法は，各時間フレームの観測信号を自己注意機構により重み付き平均して時変 SCM を推定する．しかし，本手法は観測信号全体を用いて推論するため，そのままではロボット対話など実時間性が必要なタスクには不向きだった．

本研究では，自律移動ロボットのための実時間深層ビームフォーマを提案する (図 1-(b))．単純に従来の時変ビームフォーマを逐次推論するのみでは，頻繁なビームの更新により計算コストが膨大となってしまう．そこで本研究では，フレーム単位で行っていた時変 SCM 推定の代わりに，ブロックオンライン型の注意機構を用いることで計算コストと性能を両立する．日本科学未来館で実際に自律移動ロボットを用いて収集した雑音信号により，提案法の有効性を確認した．

2. ブロックオンライン深層ビームフォーマ

自律移動ロボットで用いる音声強調は，実時間で移動音源を高精度に抽出できる必要がある．本研究では，移動音源に有効な注意機構を用いた時変 SCM 推定に基づくビームフォーミングをブロックオンライン拡張する．

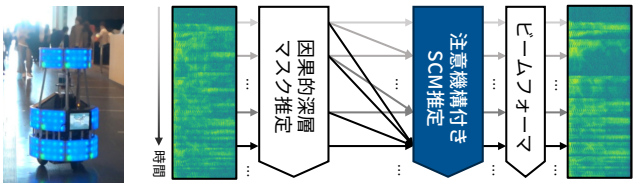
2.1 問題設定

自律移動ロボットが観測した混合音と抽出したい話者の方位を入力して音声を強調する．

入力: M チャネルの観測音 $\mathbf{x}_{ft} \in \mathbb{C}^M$ と，画像認識などで得た話者方向ベクトル $\mathbf{d}_t \in \mathbb{R}^3$ ($\|\mathbf{d}_t\| = 1$).
出力: 方向 $\mathbf{d}_t$ に存在する目的話者の強調音声 $\hat{s}_{ft} \in \mathbb{C}$

ただし，$f = 1, 2, \ldots, F$ および $t = 1, 2, \ldots, T$ はそれぞれ周波数と時間の番号を表す．提案法は，B フレームごとに処理するブロックオンライン型として定式化するため，各ブロックの初めのフレーム番号を $t_i = Bi$ とし，$i = 1, 2, \ldots$ はブロック番号を表す．

Multichannel Speech Enhancement for a Mobile Robot Based on Time-Varying SCM Estimation: *M. Fujita, Y. Bando, Y. Sasaki, M. Onishi.*



(a) Peacock　　(b) 実時間深層ビームフォーマの概要

図 1: 日本科学未来館で自律移動する Peacock (a) と，深層マスク型・実時間ビームフォーマの概要 (b).

2.2 ビームフォーミング

本研究では，観測音 $\mathbf{x}_{ft}$ は目的音声 $\mathbf{s}_{ft} \in \mathbb{C}^M$ とそれ以外の雑音 $\mathbf{n}_{ft} \in \mathbb{C}^M$ の和であると仮定する．

$$\mathbf{x}_{ft} = \mathbf{a}_{i,f} s_{ft} + \mathbf{n}_{ft} \quad (t_i \le t < t_{i+1}) \tag{1}$$

ここで，$\mathbf{a}_{i,f} \in \mathbb{C}^M$ はステアリングベクトルを表す．この仮定のもと，ビームフォーマ $\mathbf{w}_f \in \mathbb{C}^M$ を用いて混合音から目的音声 $\hat{s}_{ft}$ を推定する．

$$\hat{s}_{ft} \leftarrow \mathbf{w}_{i,f}^{\mathsf{H}} \mathbf{x}_{ft} \tag{2}$$

本研究では，強調音の歪みが少なく音声認識性能を得やすい最小無歪ビームフォーマ $\mathbf{w}_{i,f}^{\mathrm{mvdr}}$ [5] を用いる．

$$\mathbf{w}_{i,f}^{\mathrm{mvdr}} \triangleq \mathrm{tr}\left(\mathbf{G}_{i,f}^{-1}\mathbf{H}_{i,f}\right)^{-1} \mathbf{G}_{i,f}^{-1}\mathbf{H}_{i,f}\mathbf{u} \tag{3}$$

ただし，$\mathbf{H}_{i,f} = \mathbf{a}_{i,f}(\mathbf{a}_{i,f})^{\mathsf{H}} \in \mathbb{S}_+^{M\times M}$ および $\mathbf{G}_{i,f} \in \mathbb{S}_+^{M\times M}$ は音声と雑音の SCM，$\mathbf{u} \in \mathbb{R}^M$ は 1 番目のみ値を持つ単位ベクトルを表す．式 (3) のビームの構築には SCM の推定およびその逆行列計算が含まれており，ブロックサイズ B を小さくすれば高い性能を得られるが計算量が増大するトレードオフが存在する．

2.3 SCM のブロックオンライン深層予測

計算量を抑制しつつ性能を担保するため，ブロック型の注意機構に基づき時変 SCM を推定する．SCM は音声と雑音で同じ方法により得るため，以降では音声の SCM $\mathbf{H}_{i,f}$ についてのみ述べる．SCM $\mathbf{H}_{i,f}$ は，ブロック i におけるフレーム t の注意係数 $c_{i,t} \in \mathbb{R}_+$ ($1 = \sum_{t=1}^{t_i+B} c_{i,t}$) と深層マスク推定 DNN により得られた時間周波数マスク $m_{ft} \in [0,1]$ を用いて以下のように求める．

$$\mathbf{H}_{i,f} = \sum_{t=1}^{t_i+B} c_{i,t}\left\{m_{ft}\mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathsf{H}}\right\} \tag{4}$$

注意係数 $c_{i,t}$ は，DNN の出力であるブロック i のクエリ $\mathbf{q}_i \in \mathbb{R}^D$ と，フレーム t のキー $\mathbf{k}_t \in \mathbb{R}^D$ を用いて以下のように求める．

$$c_{i,t} \propto \frac{1}{\sqrt{D}}\mathbf{q}_i^{\mathsf{T}}\mathbf{k}_t \tag{5}$$

DNN のパラメータは，時間波形に対する信号対雑音比 (SNR) [4] を目的関数として最適化する．

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

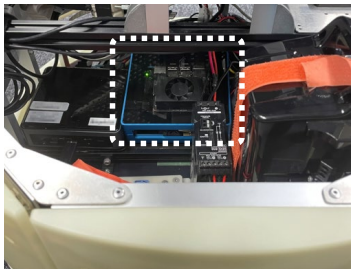


図 2: ロボット上の計算機 (白枠).

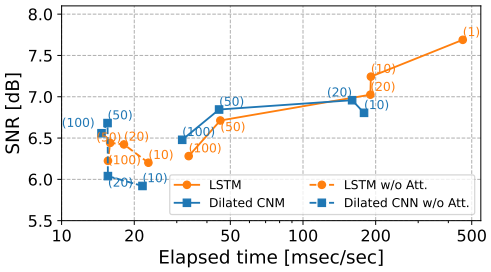


図 3: 推論時間と強調性能. () 内は B を表す.

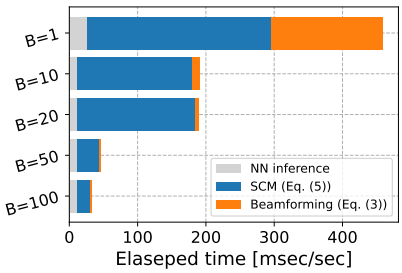


図 4: LSTM の推論時間の内訳.

3. 評価実験

日本科学未来館で収録した環境音に，WSJ0 英語読み上げ音声を重畳した混合音を用いて，提案法を評価した．

3.1 データセット

日本科学未来館における実証実験として，自律移動ロボット Peacock [6] に常設展 3 階を巡回させながら雑踏環境音を収録した (図 1-(a))．録音には 16 kHz，24 bit，16 チャネルのマイクアレイ用いた．本稿では，2022 年 11 月 26 日および 27 日に収録した約 15 時間の多チャネル環境録音を用いた．得られた雑音信号に，移動音声音源を数値シミュレーションして 5 秒分を重畳し混合音を生成した．音声信号には WSJ0 英語読み上げ音声を用い，ランダムに生成した移動経路に沿って等速直線運動させた．シミュレーションには鏡像法を用い，残響時間 RT_{60} は 600 ms で 10 m 四方の室内を仮定してインパルス応答を生成し，音源信号に畳み込んだ．この時，音声源の移動速度は 0.0 から 5.0 m/s，SNR は -10 から 5 dB の範囲で乱択した．学習，検証，評価データとして，それぞれ 20000, 1000, 1000 個の混合音を生成した．

3.2 実験設定

異なる特性を持つ以下の 2 種の DNN を評価した．1 つめは，長短期記憶ネットワーク (LSTM) [2] であり，過去のブロックの情報など長期的な依存関係を扱えるが，未来の情報は扱えない．ユニット数は 256，層数は 2 とした．2 つめは，1 次元膨張畳み込み NN (Dilated CNN) [7] であり，並列計算を活用するためブロック内の全フレームを一挙に推論する．ブロック内の未来の情報も活用できるが，LSTM のように長期的な依存関係は扱えない．256 チャネルで 4 層から成るモジュールを 4 つ重ねた．入力特徴量には，対数パワースペクトログラムとチャネル間位相差，音源の正解方位ベクトルを与えた．クエリとキーの埋め込み次元 D は 512 とし，クエリ $\mathbf{q}_i$ はフレームごとに出力した埋め込みのブロック内平均とした．

各 DNN は，学習率 10^{-3} の Adam を用いて学習した．スペクトログラムの生成では窓長 1024，ホップ長 160 の短時間フーリエ変換を用いた．ブロックサイズ B は $\{1, 10, 20, 50, 100\}$ を評価した．学習ではエポック数を 50，バッチサイズは 64 とした．音声強調性能を SNR で評価し，実行時間を NVIDIA Jetson Xavier NX (図 2) を用いて計測した．また，式 (4) の注意機構を用いずにブロック内のフレームを平均する場合と比較した．

3.3 実験結果

図 3 に示すように，注意機構を用いれば，用いない場合 (w/o Att) に比べ計算時間は増加するものの，性能を大きく改善できた．LSTM を用いた場合は，ブロックサイズ B が小さいほど高い強調性能を得られた．一方，Dilated CNN の場合は，B が 10 のとき LSTM に比べ性能が劣化している．Dilated CNN はブロック内の情報のみで推論するため，十分な文脈情報が得られなかったためと考えられる．また，B が 50 以上の場合は，Dilated CNN が LSTM より高い性能を示しており，ブロック内の未来の情報を活用できたと考えられる．全ての場合で入力長より短い時間で処理できているが，実際には他の推論処理が並列実行されるため，より短い処理時間が望ましい．強調性能を優先する場合は $B = 10$ の LSTM を，計算時間を優先する場合は，$B = 50$ の Dilated CNN を用いれば良いことが分かった．

図 4 に示す通り，$B = 1$ では，DNN の推論時間ではなく SCM とビームフォーミングの計算時間がボトルネックであった．ブロックオンライン化により，これらの計算時間を B の増大に伴い効果的に削減できている．$B = 10, 20$ では，SCM 計算がボトルネックとなっており，今後は注意機構の計算時間削減を進める．

4. おわりに

本稿では，ロボット上の計算機で実時間動作できる，移動音源の音声強調法を開発した．実雑音を用いた評価実験を行い，ブロックオンライン処理により，最低限の性能劣化で計算効率を改善できることを確認した．今後は，逐次処理するモジュールとしてロボット上に実装し，音声対話など後段の処理と統合性能を評価する．

謝辞: 本研究の一部は，NEDO の支援を受けた．また，日本科学未来館の協力のもと実施した．

参考文献

[1] J. Barker *et al.* The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes. *CSL*, 46:605–626, 2017.
[2] J. Heymann *et al.* Neural network based spectral mask estimation for acoustic beamforming. In *IEEE ICASSP*, 196–200, 2016.
[3] T. Nakatani *et al.* Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming. In *IEEE ICASSP*, 286–290, 2017.
[4] T. Ochiai *et al.* Mask-based neural beamforming for moving speakers with self-attention-based tracking. *arXiv preprint arXiv:2205.03568*, 2022.
[5] On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Transactions on audio, speech, and language processing*, 18(2):260–276, 2009.
[6] Y. Sasaki *et al.* Long-term demonstration experiment of autonomous mobile robot in a science museum. In *IEEE IRIS*, 304–310, 2017.
[7] Y. Luo *et al.* Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE TASLP*, 27(8):1256–1266, 2019.

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.