6M-01

ロボット対話における深層学習を用いたセミブラインド音声強調

和気 雅弥† 坂東 宜昭‡ 三村 正人‡ 糸山 克寿‡ 吉井 和佳‡ 河原 達也‡
†京都大学 工学部情報学科 ‡京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

自然なロボット対話の実現には，人間同士の自然な対話で起こりうる状況に対応することが重要である．その一つに，一方が発話している最中に他方からの発話が割り込むバージインという状況がある．ロボット対話では，ロボットが発話している最中に人間からの発話が割り込むという状況である．このとき，マイクロホンの観測信号は両発話の混合音であるのでそのままでの音声認識は困難であり，人間の発話を強調する必要がある．

本稿で扱う問題設定はセミブラインド音源分離と呼ばれ，除去の対象となるロボットの音源信号を利用できる．従来手法には，音声の混合モデルを基にして確率的に音声強調を行う手法 [1] や，独立成分分析を用いる Semi-blind ICA (SB-ICA) [2] といった手法が挙げられる．

本稿では，図 1 のような残響のある環境下でのセミブラインド音源分離を考える．観測された混合音からロボットの音源信号を利用して人間の発話を強調するために畳み込みニューラルネットワーク (CNN) を用いたセミブラインド音声強調 (SB-CNN) を提案する．CNN はネットワーク内で局所的な特徴の抽出を学習させる層を用い，その層を重ねることで複雑な特徴抽出を可能とし，特に画像認識の分野で大きな成果を挙げている [3]．音声においても，スペクトルを二次元の画像とみなすことが可能であり，音声の特徴を抽出することで音声の空間モデルを用いずに音源位置などに頑健な音声強調を行う．

2. 畳み込みニューラルネットワーク

本稿では，以下の音声強調の問題を扱う．

入力	1. マイクロホンの観測スペクトル X_{tf}
	2. ロボットの音源スペクトル S^R_{tf}
出力	人間の音源スペクトル S^U_{tf}

上記の t, f はそれぞれ時間フレームと周波数ビンを表す．このときロボット音源，人間音源とマイクロホン間の伝達関数をそれぞれ H_R, H_U とすると，観測スペクトルと各音源スペクトルとの関係は以下の式で表せる．

$$X_{tf} = H_R S^R_{tf} + H_U S^U_{tf}$$

また，本稿で用いるネットワークは複素数を扱わないので，各スペクトルの振幅をネットワークの入出力に用い，時間領域への変換に際してマイクロホン入力の位相を用い，逆短時間フーリエ変換の後に窓関数の影響が及ばないように重みをかけて調整している．

図 2 に提案法で用いる CNN の概形を示す．入力は，マイクロホンの観測とロボットの音源のスペクトルを前後 5 フレーム，各 11 フレームずつとする．前後のフレームも入力に加えることで，時間軸方向での音声の特徴が抽出され，残響除去の効果が期待できる．また，CNN の初めの畳み込み層のフィルタサイズを周波数軸方向では音声の基本周波数より大きくとることで，音声が周波数軸方向に持つ調波構造が考慮されることが期待される．

Semi-blind speech enhancement using deep learning in human-robot interaction : *Masaya Wake, Yoshiaki Bando, Masato Mimura, Katsutoshi Itoyama, Kazuyoshi Yoshii, Tatsuya Kawahara (Kyoto Univ.)*

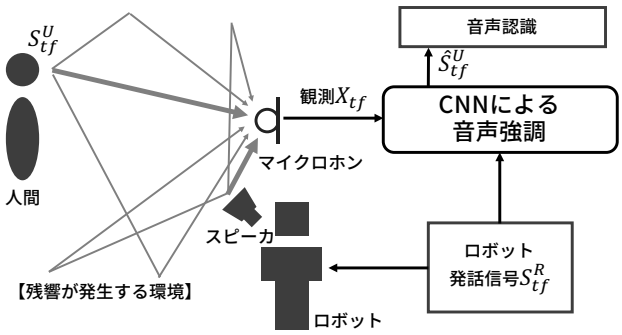

図 1: 状況設定及びデータの流れ

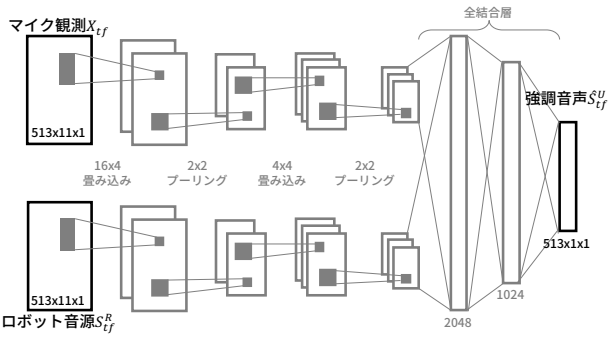

図 2: 用いる CNN の概形

提案法では初めの畳み込み層のフィルタサイズは，時間軸方向は 40 ミリ秒，周波数方向は 250Hz となるよう設定した．

プーリング層や全結合層での活性化関数には，[3] で高速な学習が可能と報告されている Rectified Linear Units (ReLU) を用いる．出力層は振幅スペクトルの各周波数ビンごとの値とし，非負にするため出力層にも ReLU を用いる．

3. 評価実験

CNN を用いた提案法の有効性を示すため，数値実験により評価を行った．

3.1 実験条件

実験のためのデータセットには，新聞記事読み上げコーパス (ASJ-JNAS) [4] の音素バランス 503 文を用い，人間の音声として男性話者，ロボットの音声として女性話者の発話データを利用した．学習データには各 60 話者 3012 発話，評価データには学習データと異なる各 4 話者 200 発話を用い，サンプリングレートは 16kHz，混合時の信号対雑音比 (SNR) は 0dB に統一した．

図 3 に示す部屋の中央にマイクロホンを置き，音源は伝達関数を測定した 40 箇所のうち，学習データにおいては白丸の 36 箇所，評価データにおいては黒丸の 4 箇所の中からランダムに配置する．またマイクロホン，音源の高さはそれぞれ 115cm，110cm である．伝達関数の有効長は 16384 サンプル (1024 ミリ秒) とした．スペクトルの窓幅は 1024 サンプル (64 ミリ秒)，窓のシフト幅は 160 サンプル (10 ミリ秒) とし，初めの畳み込み層の

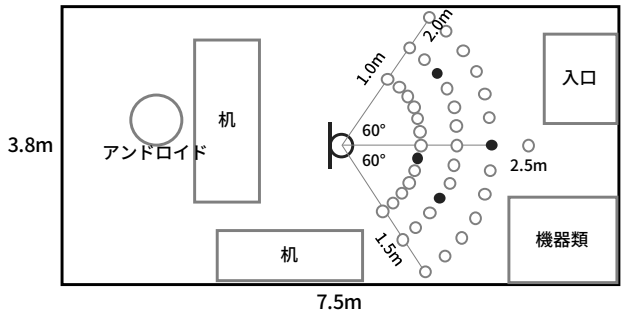

図 3: 実験室の概略図とマイクロホンや音源の配置

フィルタサイズは 16 × 4 とした．重みとバイアスは乱数を用い初期化し，サイズ 256 のミニバッチ学習を 25 エポック行う．損失関数は二乗誤差，学習法は Adagrad ($\mu = 0.005$) とした．

比較のため，音声強調を行わない場合と従来法の Semi-blind ICA に加え，Denoising Autoencoder (DAE) [5], DAE の入力にロボット音源信号を追加したセミブラインド DAE (SB-DAE) による音声強調を用いた．ここで，DAE と SB-DAE のネットワークは 2048 ユニットの隠れ層を 3 層用いた多層パーセプトロンとし，学習時のエポック数や損失関数，学習法は提案法と同じとした．

各手法について学習を行い，結果として出力された評価データ各発話の Source to Distortion Ratio (SDR) の平均値を比較することにより提案法の評価を行う．

3.2 実験結果

表 1 に示すように，提案法は SB-ICA よりも SDR が 0.824dB 改善している．また，DAE より SB-DAE が SDR が 0.602dB 改善しており，ロボット音源信号の利用による性能向上が確認でき，SB-DAE より SB-CNN が SDR が 0.463dB 改善しているため，CNN によるスペクトルの特徴抽出の有効性も確認できる．

図 4 に音声強調の結果の一例を示す．人間が発話していない区間では CNN による強調結果でもほぼ無音になっており，また低周波数では調波構造が高精度に再現されている．一方，高周波数を中心に調波構造が曖昧な部分や，時間フレームに関わらず振幅が 0 として出力されている周波数ビンがある．これは損失関数に値のスケールに影響を受ける平均二乗誤差を用いており，比較的振幅の大きい低周波数ではよく学習が行えているものの，振幅が比較的小さい高周波数では学習が十分でないためだと考えられる．

3.3 考察

提案法の CNN は畳み込み層が 2 層で，複雑な特徴抽出が困難である．解決のために Deep Residual Learning [6] を用い，音声の複雑な特徴抽出が可能な深いネットワークを用いる必要があると考えられる．また，損失関数に一般化 Kullback-Leibler divergence や Itakura-Saito divergence などの振幅比を含んだ損失関数を用いることで，振幅が比較的小さい高周波数の学習が十分に行われ，調波構造の再現が期待される．

更に，提案手法は位相情報を用いず時間軸領域への変換にマイクロホン観測の位相を使用しているため，出力の音声に歪みが生じ，SDR 値の改善に限界があると考えられる．この問題は，多チャンネルのマイクロホン観測の利用や，複素数を扱えるニューラルネットワークを用いることによる改善が期待される．その他，自然な音声を出力するために Deep Convolutional Generative Adversarial Networks [7] の手法を導入することも考えられる．

表 1: 各手法による評価データの SDR (dB)

手法	SDR(dB)
強調なし	-2.970
DAE	0.547
SB-ICA	0.788
SB-DAE	1.149
SB-CNN(提案法)	1.612

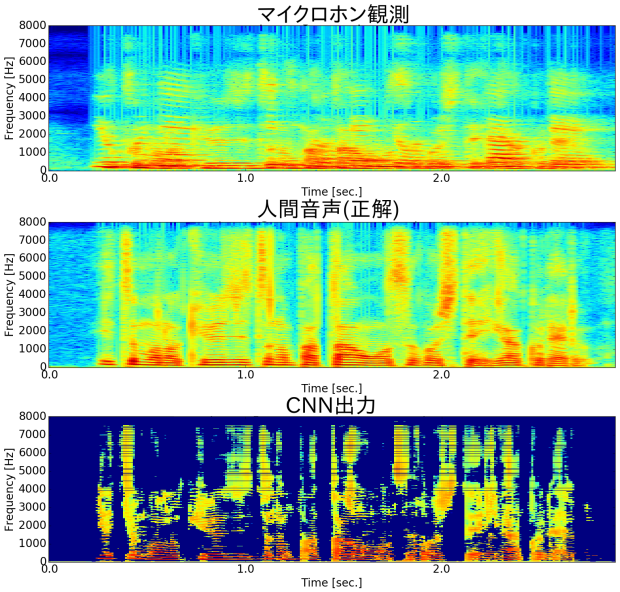

図 4: CNN の出力スペクトルの例

4. おわりに

本稿では，ロボットの発話と人間の発話が重なる状況において，マイクロホンの観測スペクトルと既知であるロボットの音源スペクトルから CNN を用いて人間の発話を得るための手法について述べた．実験では，提案法により強調された音声の SDR が SB-ICA より 0.824dB, SB-DAE より 0.463dB 改善したことを確認した．今後は，考察で述べた手法の導入による音声強調性能の比較や，実環境でのデータを用いた評価を行う．

参考文献

[1] 前澤 陽 *et al.* パワースペクトル上の無限オーダー畳込み型室内伝達モデルに基づく実時間セミブラインド音声強調. *2013* **年 日本音響学会 秋季研究発表会**, 2–1–3.

[2] R. Takeda *et al.* Barge-in-able robot audition based on ICA and missing feature theory under semi-blind situation. *IEEE/RSJ IROS 2008*, 1718–1723.

[3] A. Krizhevsky *et al.* Imagenet classification with deep convolutional neural networks. *NIPS 2012*, 1097–1105.

[4] K. Itou *et al.* Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research. *J. of ASJ (E)*, 20(3):199–206, 1999.

[5] P. Vincent *et al.* Extracting and composing robust features with denoising autoencoders. *ACM ICML 2008*, 1096–1103.

[6] K. He *et al.* Deep residual learning for image recognition. *IEEE CVPR 2016*, 770–778.

[7] A. Radford *et al.* Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.