一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2017-22(2017-08)

[ポスター講演] 再帰型ニューラルネットワークを用いたセミブラインド音声分離・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 大学院情報学研究科
京都府京都市左京区吉田本町
E-mail: †{wake,bando,mimura,itoyama,yoshii,kawahara}@sap.ist.i.kyoto-u.ac.jp

あらまし　本稿では，ニューラルネットワークを用いたセミブラインド音声強調の手法について述べる．人間とロボットとの対話において，ロボットは自身のマイクロホンに加えて，ロボット自身の発話信号も得ることができるため，ここで扱う音声強調はセミブラインドである．本稿では，セミブラインド音源分離とブラインド残響除去の 2 つのモジュールからなるニューラルネットワークを提案する．この再帰型ニューラルネットワークは，両モジュールに教師信号を用いることでマルチタスク学習を行う．評価実験により，既存のセミブラインド音声強調法と比べて提案手法の有効性を示す．

キーワード　セミブラインド音声強調，セミブラインド音声分離，ブラインド残響除去，再帰型ニューラルネットワーク

[Poster Presentation]Semi-blind speech separation and enhancement using recurrent neural network

Masaya WAKE†, Yoshiaki BANDO†, Masato MIMURA†,

Katsutoshi ITOYAMA†, Kazuyoshi YOSHII†, and Tatsuya KAWAHARA†

† Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501 JAPAN
E-mail: †{wake,bando,mimura,itoyama,yoshii,kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract　This paper describes a semi-blind speech enhancement method using a neural network. In a human-robot speech interaction, the robot inputs not only audio signals recorded by a microphone but also speech signals made by the robot itself, which can be used for semi-blind speech enhancement. We propose a neural network which consists of cascaded two modules: a semi-blind source separation module and a blind dereverberation module. The proposed recurrent neural network is trained in a manner of multi-task learning, i.e., teacher signals are used for both the output of the separation module and the dereverberation module. Experiments are conducted to show the effectiveness of the proposed network.

Key words　Semi-blind speech enhancement, Semi-blind source separation, Blind dereverberation, Recurrent neural network

1. はじめに

ロボットによる対話において，マイクロホンで観測された信号のうち人間の音声のみを強調することは必要不可欠である．ロボットが発話権を正確に同定することはこんなんで，人間とロボットが同時に発話を行うことがあるため，マイクロホンによって観測される音声信号に，人間の音声だけでなくロボット自身の音声も含まれることがある．また，観測された音声信号には，それぞれの音源からの直接音だけでなく，残響音も含まれ，音声認識を困難にする．この問題に対処するために，人間の音声を分離して残響を除去する音声強調が重要となる．

本稿で考える音声強調は 2 つのステップから構成される．第一のステップは，観測された音声信号から人間の発話に由るもののみを分離するセミブラインド音源分離である．対話生成部

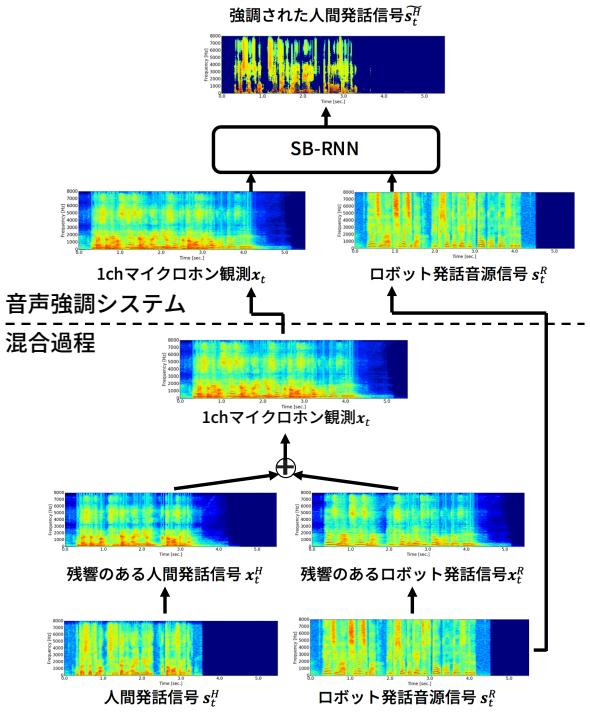

図 1　提案手法の概略図

Fig. 1　An overview of the proposed method

分と音声認識部分が一体的に運用されるとき，システム内でロボットの発話音源を利用することが可能であるため，このステップはセミブラインド音源分離の問題として考えられる．第二のステップは，分離された残響を含む人間発話の音声信号から残響成分を除去した人間発話の音声信号を得るブラインド残響除去である．部屋の大きさや人間とマイクロホンの位置が不明で，それらの情報を用いることなく残響除去を行うため，このステップはブラインド残響除去の問題として考えられる．

本稿では，音声強調を行うセミブラインド再帰型ニューラルネットワーク (SB-RNN) を提案する．このニューラルネットワークは，音源分離を行う音源分離モジュールと残響除去を行う残響除去モジュールの二つのモジュールを接続して一つのネットワークとして構成する．二つのモジュールを接続して一つのネットワークとして学習を行うことで，各モジュールが音源分離や残響除去を行うように学習され，なおかつネットワーク全体が総合的に最適化されるという利点がある．それぞれのモジュールを独立に学習させる場合では，学習時に用いる残響除去モジュールの入力がロボットの音声を一切含まないものとなり，実際の運用の際に音源分離モジュールで分離が不十分だった場合の残響除去の性能が低下すると考えられる．一つのネットワークとして構成させると，同時に音源分離と残響除去を行うため困難になる．提案手法では，音源分離を一つのネットワークで行うとともに，音源分離モジュールと残響除去モジュールの役割を明確にするために，マルチタスク学習 [1] を導入する．

2. 関 連 研 究

ロボット発話の音源信号を用いた音声強調の手法として，これまでにセミブラインド独立成分分析 (Semi-blind ICA) が提案されている [2]．これは音源分離に用いられている独立成分分析 (ICA) [3], [4] をセミブラインドに拡張したものであり，これらの手法は事前学習を必要としないという特徴があるが，音源数と同じかそれを上回るマイクロホンのチャネル数を必要とし，また逐次推定を行うため出力が収束するまでに数秒を要する．

音源分離単独の手法としては，分離を行うマスクを推定する手法が存在する．マスクとなる行列を推定し，入力とマスクの要素積 (アダマール積) を音源分離結果とする．ある時間・周波数ビンの成分がいくつの音源からもたらされているかという仮定により，用いるマスクがハードマスク (バイナリマスクとも呼ばれる) またはソフトマスクのいずれかになる．ハードマスクを用いる手法 [5] は，ある時間・周波数ビンの成分は単一の音源からのみもたらされていると仮定し，マスクのそれぞれの要素の値は 0 か 1 の二値のいずれかのみをとる．ソフトマスクを用いる手法 [6] は，ある時間・周波数ビンの成分が複数の音源からもたらされていると仮定し，マスクのそれぞれの要素の値は 0 から 1 の実数値をとる．

他にも，複数の成分に分割し，それを適宜統合することで音源分離を行う手法も提案されている．これらの手法は音声の低ランク性を仮定し，非負値行列因子分解 (NMF) や確率的潜在要素解析 (PLCA) を用いてマイクロホンの観測信号を複数の成分に分割している [7], [8]．しかし，これらの手法は分割した各成分がどの音源に対応するかの割当てを正確に行う必要があり，混合音源のうち，全てないしは 1 つの音源を除いた全ての音源について事前学習が必要となる．

残響除去単独の手法も数多く提案されてきており，例えば残響の指数減衰を仮定して，残響成分のスペクトル減算を行う手法や [9]，残響を含まない音声の包絡を復元するフィルタの推定を行う手法 [10]，残響を含まない音声が高い尖度を有することを用いた手法 [11] などが提案されている．これらの手法においては，インパルス応答や残響を含まない音声の性質について何らかの仮定をおいている．

深層学習の隆盛と共に，それによる手法が音源分離や残響除去の手法として提案されてきており [12], [13]，基本的な多層パーセプトロンを用いた手法でも既存の手法より優れた性能を示している [14]．また，音声信号の時間的相関を考慮できる再帰型ニューラルネットワーク (RNN) を用いた音源分離や残響除去の手法が提案され，これらの手法がより良い性能を示している [15]〜[17]．ニューラルネットワークを用いる手法では大量の学習データを必要とするが，その一方であらゆる音声の混合や残響のモデルを表現することができ，また音源分離や残響除去に対する頑健性が期待される．

3. 提 案 手 法

図 1 に SB-RNN を用いた提案手法の概要を示す．このニューラルネットワークは，単チャネルマイクロホンが観測した音

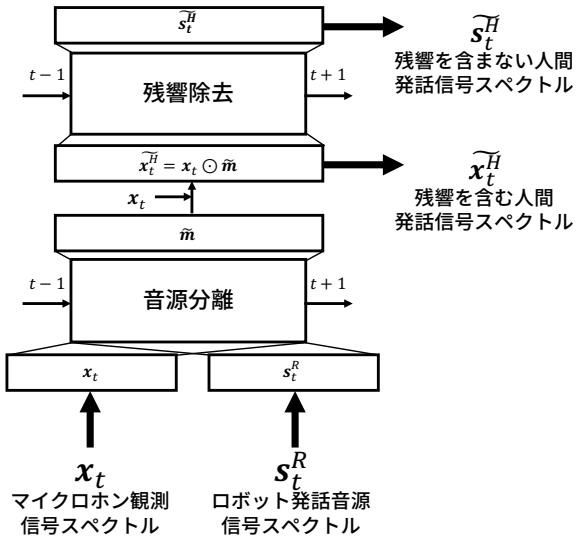

図 2　提案手法のネットワークの構成
Fig. 2　A structure of the proposed network

声信号とロボット発話の音源信号を入力とし，音源分離及び残響除去によって強調された人間の発話信号を出力とする．$\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tF})$ を時間 t におけるマイクロホン観測のスペクトル，及び $\boldsymbol{x}_t^R$ と $\boldsymbol{x}_t^H$ をそれぞれロボット，人間の残響を含む音声信号とすると，ロボット及び人間とマイクロホンとの間のインパルス応答をそれぞれ $\boldsymbol{h}^R, \boldsymbol{h}^H$ とした時，

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{x}_t^R + \boldsymbol{x}_t^H \\ \boldsymbol{x}_t^R &= \boldsymbol{h}^R \odot \boldsymbol{s}_t^R \\ \boldsymbol{x}_t^H &= \boldsymbol{h}^H \odot \boldsymbol{s}_t^H \end{aligned} \tag{1}$$

となる．ここで F は周波数ビンの数であり，$\odot$ は要素積を表す．一般的なニューラルネットワークで扱うことのできる値は実数であるため，本稿で提案する SB-RNN では振幅スペクトルでの処理を行う．すなわち

$$\begin{aligned} |\boldsymbol{x}_t| &= |\boldsymbol{x}_t^R| + |\boldsymbol{x}_t^H| \\ |\boldsymbol{x}_t^R| &= |\boldsymbol{h}^R| \odot |\boldsymbol{s}_t^R| \\ |\boldsymbol{x}_t^H| &= |\boldsymbol{h}^H| \odot |\boldsymbol{s}_t^H| \end{aligned} \tag{2}$$

について考えることになるが，以下絶対値記号を省略し，$\boldsymbol{x}_t$ が振幅スペクトルを表すものとする．SB-RNN によって得られた人間の発話信号推定スペクトル $\tilde{\boldsymbol{s}}^H$ を時間領域に復元する際には，マイクロホン観測の位相 $\arg(\boldsymbol{x}_t)$ を用いる．

図 2 に本稿で提案する SB-RNN の構造を示す．SB-RNN は音源分離モジュールと残響除去モジュールからなり，各々以下で説明する．

3.1　音源分離モジュール

音源分離モジュールではマイクロホン観測 $\boldsymbol{x}_t$ を分離するためのマスク $\boldsymbol{m}_t = (m_{t1}, \ldots, m_{tF})$ を推定する．入力はマイクロホン観測 $\boldsymbol{x}_t$ 及びロボット発話の音源信号 $\boldsymbol{s}_t^R$ であり，出力は推定マスク $\tilde{\boldsymbol{m}}_t$ とする．出力されたマスク $\tilde{\boldsymbol{m}}_t$ を用いて，マイクロホン観測 $\boldsymbol{x}_t$ を人間の発話による信号 $\tilde{\boldsymbol{x}}_t^H$ とロボットの発話による信号 $\tilde{\boldsymbol{x}}_t^R$ に以下の式を用いて分離を行う．

$$\begin{aligned} \tilde{\boldsymbol{x}}_t^H &= \tilde{\boldsymbol{m}}_t \odot \boldsymbol{x}_t, \\ \tilde{\boldsymbol{x}}_t^R &= (\boldsymbol{1} - \tilde{\boldsymbol{m}}_t) \odot \boldsymbol{x}_t. \end{aligned} \tag{3}$$

なお，$\boldsymbol{1}$ はすべての要素が 1 である F 次元のベクトルである．

音源分離モジュールは 5 層からなるネットワークである．中間の隠れ層は 3 層で，それぞれのノード数は入力層が $2F$，隠れ層が各 500，出力層が F である．音声は時間方向に強い相関を保つため，隠れ層のうち中央の 2 番目の隠れ層に再帰層を導入することでその相関を活用した音源分離の性能向上が期待できる．活性化関数は入力層と隠れ層には rectrified linear unit (ReLU) を利用するが，出力層はソフトマスクとするために 0 から 1 の実数をとるシグモイド関数を採用する．マスクを用いることにより，発話が存在しない部分に雑音が混入することを防ぎ，音源分離の性能の向上が期待できる．

再帰を行うフレーム数が多いほど長時間の特徴を考慮することができるが，必要とする計算時間やメモリ容量などが増大し，学習が現実的でなくなるため，本手法においては 512 ミリ秒に相当する過去 32 フレームを参照する．

3.2　残響除去モジュール

残響除去モジュールでは音源分離モジュールによって分離された残響を含む人間の発話信号 $\tilde{\boldsymbol{x}}_t^H$ を入力とし，残響を含まない人間の発話信号 $\tilde{\boldsymbol{s}}_t^H$ を直接推定する．残響除去モジュールも音源分離モジュールと同様に 5 層からなるネットワークとし，入力層，3 層の隠れ層，出力層を持つものとして，そのうち 2 番目の隠れ層に再帰層を導入する．音源分離モジュールと異なり，入力のスペクトルが 1 つであるので，ノード数は入力層が F，隠れ層が各 500，出力層が F である．音声の残響は近接フレーム間で非常に強い相関があるため，残響除去にも隠れ層を導入することで残響除去の性能向上が期待できる．活性化関数は入力層と隠れ層は音源分離モジュールと同様に ReLU であるが，音源分離モジュールと異なり出力が振幅スペクトルであるので，非負値を出力する ReLU を出力層にも採用する．

3.3　マルチタスク学習

提案手法では，音源分離モジュールによる分離結果の教師信号として $\boldsymbol{x}_t^H$，残響除去モジュールによる残響除去結果の教師信号として $\boldsymbol{s}_t^H$ の二つを用いる．これにより単一のネットワークが 2 つの出力を行うマルチタスク学習 [1] となり，音源分離モジュールが主に音源分離を，残響除去モジュールが主に残響除去を行うように学習が行われる．学習時に用いる損失関数は二乗誤差とする．すなわち，損失を J としたとき以下のように表される．

$$\begin{aligned} J &= J_S + J_D, \\ J_S &= \|\boldsymbol{x}_t^H - \tilde{\boldsymbol{x}}_t^H\|_2^2, \\ J_D &= \|\boldsymbol{s}_t^H - \tilde{\boldsymbol{s}}_t^H\|_2^2. \end{aligned} \tag{4}$$

ネットワークの最適化には Adam [18] を用い，また各モジュールの出力層を除く各層には初期値依存性を軽減できるバッチ正規化 [19] を適用する．

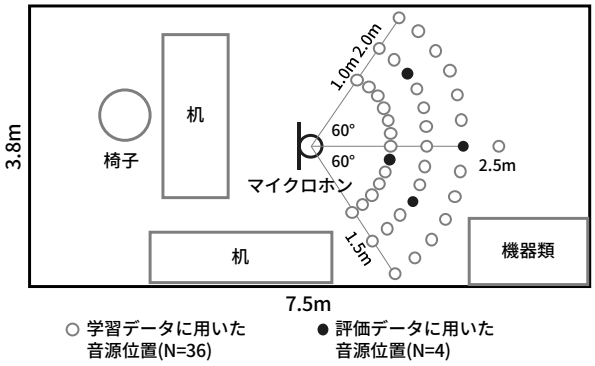

図 3　インパルス応答を測定した部屋の概略図

Fig. 3　A schematic of the room where the impluse responces are recorded

4. 評 価 実 験

提案手法の音声強調性能を評価するために，評価実験を行なった．本稿では評価尺度として source to distortion ratio (SDR) [20] を用いた．

4.1 実 験 環 境

学習及び評価のデータセットとして，新聞記事読み上げコーパス (ASJ-JNAS) [21] の音素バランス 503 文を用いた．男性及び女性の発話をそれぞれ学習用に 3012 発話，評価用に 200 発話用い，男性の音声を人間の音声，女性の音声をロボットの音声とした．なお，学習用データと評価用データに用いた話者に重複はないが，発話している文章には重複が存在する．マイクロホン観測は図 3 に示した室内で事前に計測したインパルス応答を用いてシミュレートしている．40 点で計測されたインパルス応答のうち，36 点を学習時に用いる音源位置，4 点を評価時に用いる音源位置と分けている．音声混合時の signal to noise ratio (SNR) は −6.0dB，−3.0dB，0.0dB，3.0dB，6.0dB，9.0dB の 6 つとし，各 SNR の音声の数が均等になるように配分した．

提案手法 (SB-RNN) の有効性を示すために，他の手法との比較を行なった．処理なしはマイクロホン観測信号に処理を加えず，そのままの信号を用いるものである．また，Semi-blind ICA はロボット聴覚ソフトウエア HARK [22] に搭載されているものを用いた．

残りの手法はニューラルネットワークを用いた手法であり，それぞれのネットワークの構造を図 4 に示す．“ブラインド音声強調” (図 4-A) の手法は，入力のうちのロボットの発話音源 $\boldsymbol{s}_t^R$ を用いず，マイクロホン観測信号 $\boldsymbol{x}_t$ のみから音声強調を行うものである．“再帰なし” (図 4-B) の手法は，音源分離モジュール及び残響除去モジュールの再帰層から再帰を取り除いた，単純な多層パーセプトロンである．“シングルタスク学習” (図 4-C) の手法は，ネットワークの構成は提案手法と同じであるが，音源分離モジュールによる音源分離結果の $\tilde{\boldsymbol{x}}_t^H$ に関する教師信号は与えず，ネットワークの最終的な出力のみを用いてネットワークを学習させる．“個別モジュール” (図 4-D) の手法は，音源分離モジュールと残響除去モジュールを別のネットワークとして独立に学習させ，評価時には音源分離モジュールの出力を残響除去モジュールの入力とする．

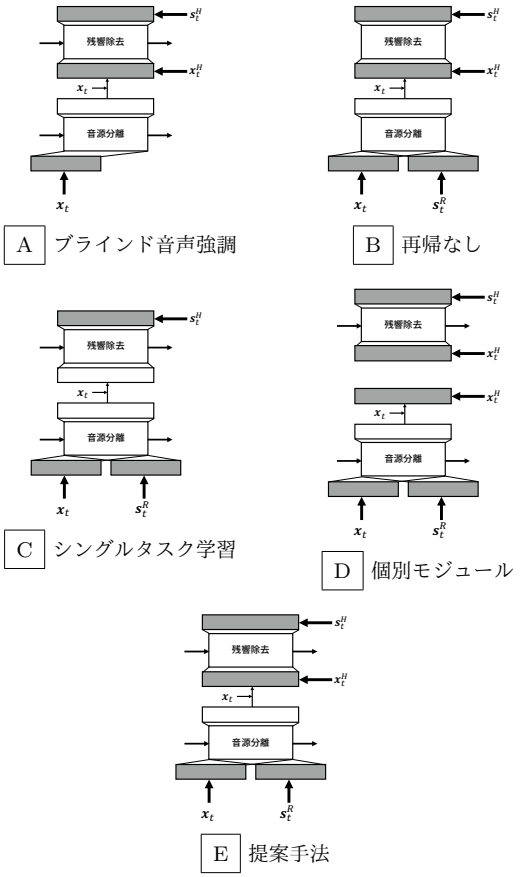

図 4　提案手法との比較を行ったネットワーク

Fig. 4　Networks compared with the proposed SB-RNN

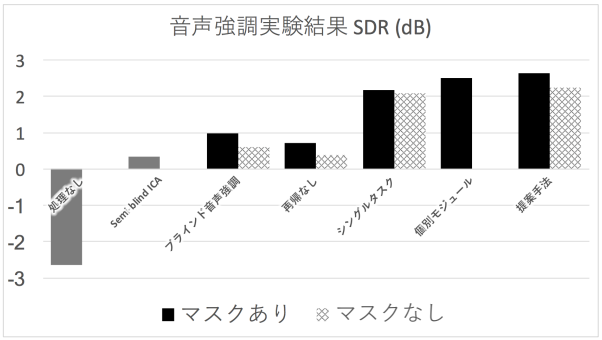

図 5　実験結果の平均 SDR (dB)

Fig. 5　Results of the experiments, the average of SDR (dB).

また，これらの比較手法及び提案手法 (図 4-E) のうち，モジュールが接続されている手法に関しては，それぞれ音源分離モジュールの出力をマスク $\tilde{\boldsymbol{m}}_t$ ではなく，振幅スペクトル $\tilde{\boldsymbol{x}}_t^H$ そのものとした場合についても実験を行なった．

4.2 実 験 結 果

図 5 に実験結果を示す．提案手法による音声強調結果の SDR は既存手法である SB-ICA と比べて 2.3dB 改善している．学習と評価時で用いている音源位置が異なっていることから，提案手法は音源位置に対する頑健性を実現いると考えられる．ま

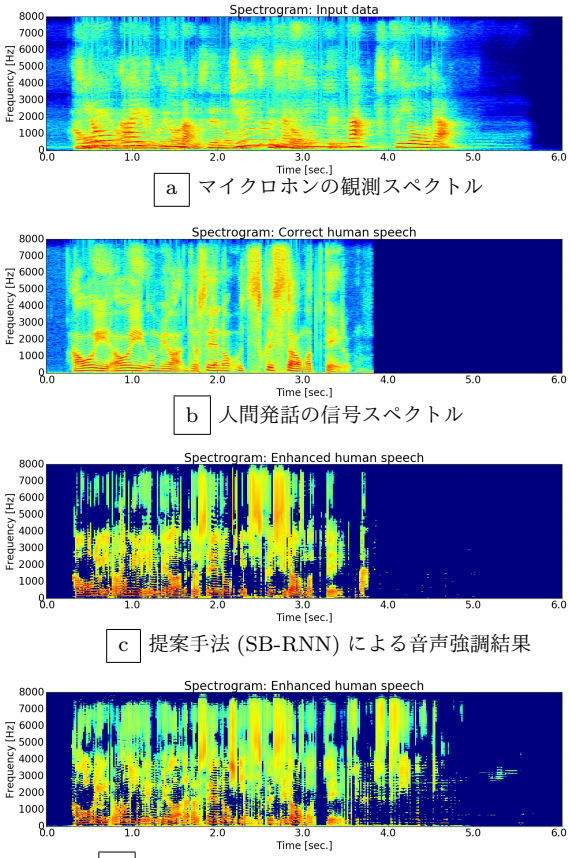

図 6　提案手法による音声強調結果の一例

Fig. 6　An example of the results of speech enhancement by the proposed method.

た，ニューラルネットワークを用いた他の手法と比較して，提案手法がより高い音声強調性能を示していることから，提案手法の優位性を示せた．

ニューラルネットワークを用いた他の手法も Semi-blind ICA と比較して良い性能を示しているが，提案手法ほどの性能は得られなかった．“ブラインド音声強調” と比較することで，提案手法のロボット音源信号を用いた効果を，“再帰なし” と比較することで，時間的相関を考慮できる再帰層の導入の効果を，また，“シングルタスク学習” と比較することで，各モジュールの役割を明確にさせるマルチタスク学習を導入した効果を確認することができた．

また，音源分離モジュールにマスクを用いた場合がマスクを用いていない場合より高い音声強調の性能を示した．これにより，雑音混入を防ぐことができるマスクを導入することの効果を確認することができた．

図 6-c 及び 6-d に提案手法と “個別モジュール” の手法による同一入力からの音声強調結果のスペクトルを示す．スペクトルのおおよそ右半分の区間は，図 6-b に示すように人間が発話を行なっていない区間であるが，“個別モジュール” ではこの区間に雑音が発生している一方で，提案手法ではその雑音が発生していない．これは，学習の際に残響除去モジュールには必ずしも完璧に分離された音声信号ではなく，ロボットの音声が残存している音声信号が入力として与えられることで残響除去モジュールが残響除去のみならず雑音除去の能力も得たものと考えられる．その一方，“個別モジュール” では各モジュールを別々に学習させ，残響除去モジュールでは残響以外の雑音が混入した音声信号を想定していないために，残響以外の雑音が入力として与えられた場合の対処ができない．

5. おわりに

本稿では，音声強調の手法として再帰型ニューラルネットワークを用いた SB-RNN を提案した．再帰層を導入することで音声の時間軸方向への相関を考慮し，マルチタスク学習によってネットワーク内での音源分離と残響除去の役割を明確化した．評価実験の結果，既存の SB-ICA と比較して SDR 値が 2.3dB 改善したことを確認し，また他のネットワークと比較することにより提案手法の有効性を示した．

今後は，現時点では考慮していないロボットや人間の音声に関係ない雑音の除去に対応を行うため，提案手法のネットワークを拡張および改善を行う．また，音声認識率などによる評価も行う．それに加えて，このネットワークを発展させ，音声強調と音声認識を統合したネットワークを構築することも考える．また，提案手法では位相の情報を利用していないため，強調された音声を時間領域に戻すときに人工的な歪みが生じる [23] という問題がある．この問題の解決のために，多チャネルのマイクロホンを用いる方法や複素数を扱えるネットワーク [24] を用いる方法でネットワーク内で位相を取り扱う．

文　献

[1] R. Caruana, “Multitask learning,” Learning to learn, pp.95–133, Springer, 1998.
[2] R. Takeda, K. Nakadai, K. Komatani, T. Ogata, and H.G. Okuno, “Barge-in-able robot audition based on ICA and missing feature theory under semi-blind situation,” IROS 2008, pp.1718–1723, 2008.
[3] P. Comon, “Independent component analysis, a new concept?,” Signal Processing, vol.36, no.3, pp.287–314, 1994.
[4] N. Murata, S. Ikeda, and A. Ziehe, “An approach to blind source separation based on temporal structure of speech signals,” Neurocomputing, vol.41, no.1, pp.1–24, 2001.
[5] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” IEEE Transactions on Signal Processing, vol.52, no.7, pp.1830–1847, 2004.
[6] A.M. Reddy and B. Raj, “Soft mask methods for single-channel speaker separation,” IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.6, pp.1766–1776, 2007.
[7] A. Cichocki, R. Zdunek, and S.-i. Amari, “New algorithms for non-negative matrix factorization in applications to blind source separation,” ICASSP 2006, vol.5, pp.621–624, 2006.
[8] T. Virtanen, “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria,” IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.3, pp.1066–1074, 2007.
[9] K. Lebart, J.-M. Boucher, and P.N. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” Acta Acustica united with Acustica, vol.87, no.3,

pp.359–366, 2001.
[10] C. Avendano and H. Hermansky, “Study on the dereverberation of speech based on temporal envelope filtering,” ICSLP 1996, vol.2, pp.889–892, 1996.
[11] B.W. Gillespie, H.S. Malvar, and D.A. Florêncio, “Speech dereverberation via maximum-kurtosis subband adaptive filtering,” ICASSP 2001, vol.6, pp.3701–3704, 2001.
[12] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo, “A class of neural networks for independent component analysis,” IEEE Transactions on Neural Networks, vol.8, no.3, pp.486–504, 1997.
[13] Y. Tan, J. Wang, and J.M. Zurada, “Nonlinear blind source separation using a radial basis function network,” IEEE Transactions on Neural Networks, vol.12, no.1, pp.124–134, 2001.
[14] K. Han, Y. Wang, D. Wang, W.S. Woods, I. Merks, and T. Zhang, “Learning spectral mapping for speech dereverberation and denoising,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.6, pp.982–992, 2015.
[15] F. Weninger, F. Eyben, and B. Schuller, “Single-channel speech separation with memory-enhanced recurrent neural networks,” ICASSP 2014, pp.3709–3713, 2014.
[16] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Joint optimization of masks and deep recurrent neural networks for monaural source separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.12, pp.2136–2147, 2015.
[17] A.L. Maas, Q.V. Le, T.M. O’Neil, O. Vinyals, P. Nguyen, and A.Y. Ng, “Recurrent neural networks for noise reduction in robust ASR,” Interspeech 2012, pp.22–25, 2012.
[18] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, pp.1–15, 2014.
[19] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167, pp.1–11, 2015.
[20] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” IEEE Transactions on Audio, Speech, and Language Processing, vol.14, no.4, pp.1462–1469, 2006.
[21] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research.,” Journal of the Acoustical Society of Japan (E), vol.20, no.3, pp.199–206, 1999.
[22] N. Kazuhiro, T. Toru, O.G. Hiroshi, N. Hirofumi, H. Yuji, and T. Hiroshi, “Design and implementation of robot audition system HARK – open source software for listening to three simultaneous speakers,” Advanced Robotics, vol.24, no.5-6, pp.739–761, 2010.
[23] J. Le Roux, N. Ono, and S. Sagayama, “Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction,” Interspeech 2008, pp.23–28, 2008.
[24] M.F. Amin and K. Murase, “Single-layered complex-valued neural network for real-valued classification problems,” Neurocomputing, vol.72, no.4, pp.945–955, 2009.