1R-04

距離ベース時間周波数マスク推定による音声強調手法の検討

石井 遼平 [1]，糸山 克寿 [1,2]，中臺 一博 [1]
1 東京工業大学 工学院 システム制御系　2 (株) ホンダ・リサーチ・インスティチュート・ジャパン

1　はじめに

近年では，マイクが様々な機器に搭載され，音声をいつでも簡単に録音できるようになり，通話やオンライン会議などに広く利用されている．複数の音源が存在する状況では，誰か一人だけの発話を録音しようとして口元にマイクを配置したとしても，録音される信号には他の音源も混入してしまい，円滑なコミュニケーションの妨げとなる．ノイズキャンセリング機能で周囲の雑音を抑圧し，近くの話者の音声のみを強調する機能を備えた製品もあるが，マイクロホンアレイのような特定デバイスの利用を前提とする場合が多く，単一マイクに対するこのような機能は十分に実現されていない．

本稿ではこの問題を解決するために，Fig. 1 のように話者の近くに設置された単一マイクで録音された音声から，近距離音声のみを強調し取り出す，という手法を検討し報告する．マイクは話者のうちいずれかの近くに設置するものとし，設置されたマイクに最も近い話者の音声を近距離音，それ以外の話者の音声を遠距離音とする．単一マイクのみを用いるため，マイクロホンアレイ処理で用いられる位相差などの特徴を手がかりとすることはできないが，音源とマイクとの距離の違いに起因する，直接音成分と残響成分の比率などを手がかりとすることで，近距離音声を強調することを目指す．本稿では，Recurrent Residual Network (RRN) により推定された時間周波数マスクを用いて，モノラル混合音から，近距離話者の音声のみを抽出する音声強調法を提案する．提案手法の有効性を示すため，実環境のインパルス応答から構築されたデータセットを用いて評価実験を行った．

2　関連研究

Patterson [1] らは，距離に基づく音源分離を提案している．畳み込みニューラルネットワーク (Convolutional Neural Network, CNN) と，再帰型ニューラルネットワーク (Recurrent Neural Network, RNN) の一種である Long Short-Term Memory (LSTM) [2] とを組み合わせたモデルにより時間周波数マスクを推定し，音源とマイクロホンとの距離の違いに基づいて混合音を近距離音と遠距離音に分離する．この研究はモノラル音響信号に対しても距離に基づく音源分離が可能であることを示した一方で，性能評価は pyroomacoustics [3] で作成された仮想環境のデータのみが用いられており，実環境での性能評価はなされていなかった．

3　提案手法

近距離音と遠距離音が混在するモノラル混合音のスペクトログラムを入力とし近距離音のスペクトログラムを出力する，本稿で提案する時間周波数マスクを用いた近距離音声強調手法について述べる．その概要を Fig. 2 に示す．時間周波数マスクはスペクトログラムと同じ次元をもち，各要素は $[0, 1]$ の実数の行列である．値が 0 に近ければ混合音中の対応する時間周波数成分を遮断し，1 に近ければ通過させる．時間周波数成分が近距離音を多く含むときに 1 に近く，それ以外を多く含むときに 0 に近くなるように時間周波数マスクを推定することで，混合音から近距離音のみを抽出する．混合音のスペクトログラムを $\boldsymbol{X} \in \mathbb{C}^{T\times F}$，強調近距離音のスペクトログラムを $\boldsymbol{Y} \in \mathbb{C}^{T\times F}$，時間周波数マスクを $\boldsymbol{M} \in [0,1]^{T\times F}$ で表すと，これらの関係は $\boldsymbol{Y} = \boldsymbol{X} \otimes \boldsymbol{M}$ と表される．T は時間フレーム数，F は周波数ビン数を表す．$\otimes$ は行列の要素積を表す演算子である．

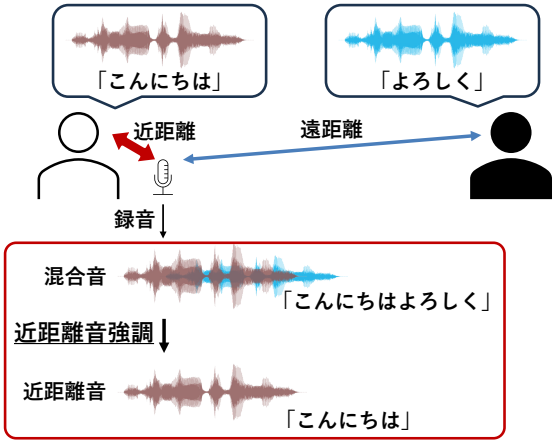

Fig. 1　提案手法が想定する状況

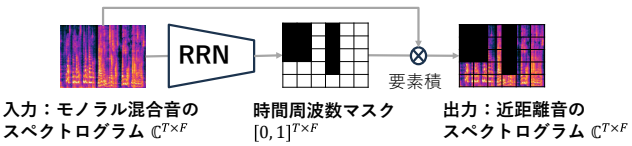

Fig. 2　提案手法の概要

マイクで収録される近距離音と遠距離音は，そのどちらに対しても直接音成分と残響成分が混ざって観測されるものの，それらの比率は距離によって異なり，近距離音では直接音成分が，遠距離音では残響成分が，それぞれ相対的に大きくなる．混合音のスペクトログラムから，深層学習モデルでこのような直接音成分と残響成分の比率を抽出することができれば，それに基づいて近距離音のみを強調する時間周波数マスクを得ることができる．そのためには，直接音・残響成分の比率に関する局所的な特徴を抽出する機能と，音声の時間的な連続性を考慮してマスクに反映させる機能を併せ持つモデルが必要となる．

時間周波数マスクの推定には本稿で提案する Recurrent Residual Network (RRN) を用いる．これは，CNN の一種である残差ネットワーク (Residual Net, ResNet) [4] と RNN の一種である Gated Recurrent Unit (GRU) [5] を組み合わせたネットワークである．Patterson [1] らの手法と比べると，CNN が ResNet に，LSTM が GRU に変更されている．ResNet は一般的な CNN に比べて勾配消失が起こりにくく深

Speech Enhancement by Distance-based Time-Frequency Mask Estimation
Ryohei Ishii[1], Katsutoshi Itoyama[1,2], Kazuhiro Nakadai[1]
1 Dept. of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology
2 Honda Research Institute Japan Co., Ltd.

Table 1　RRN の構成．T は時間フレーム数，F は周波数ビン数を表す．入力はスペクトログラムの実部と虚部をそれぞれ別チャネルとしている．

Input: Spectrogram of mixed sound, $2 \times T \times F$
Conv2D 7×7 @ 64, ReLU
4× Conv2D 3×3 @ 64, ReLU
4× Conv2D 3×3 @ 128, ReLU
4× Conv2D 3×3 @ 256, ReLU
GRU $256 \rightarrow 30$
Linear $30 \rightarrow 16$, ReLU
Linear $16 \rightarrow 1$, ReLU
Output: Time-Frequency mask, $T \times F$

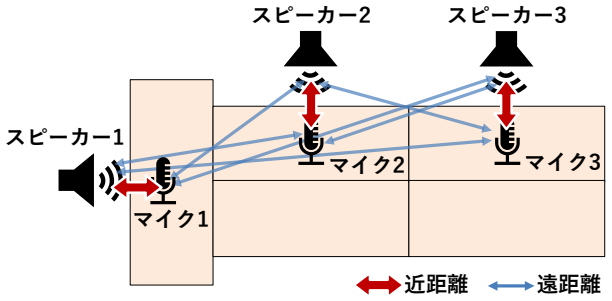

Fig. 3　インパルス応答収録環境の例

い層を構成しやすいため，直接音と残響成分の比率に関する局所的な音響特徴の抽出性能向上をねらい採用した．オリジナルの ResNet では，畳み込み層で抽出された特徴を圧縮するために Max Pooling 層が用いられているが，提案手法では入力スペクトログラムと同じサイズの時間周波数マスクを求める必要があるため，Max Pooling 層を取り除いている．また GRU は，LSTM よりも単純な構造で同等の性能をもつため，ネットワークのパラメータ数を抑制しつつ時間的な連続性を考慮するために採用した．RRN の具体的な構成を Table 1 に示す．

4　評価実験

提案手法の有効性を示すために，さまざまな実環境を模して作成した近距離音・遠距離音・背景音からなる混合音に対する近距離音強調性能を評価する実験を行った．近距離音と遠距離音は実環境で収録したインパルス応答を音声コーパス Libri-Light [6] から抽出した音声に畳み込んで作成した．背景音は CHiME3 dataset [7] に含まれる 4 種の雑音からランダムに選択した．評価指標として，強調音声の明瞭度を表す指標である PESQ [8] と STOI [9] を用いた．

インパルス応答の収録は，広さや用途がそれぞれ異なる 10 ヶ所の環境で行った．一例として，5 台の机が置かれた小規模な会議室で話者 3 名の発話を模した環境を Fig. 3 に示す．話者の位置にスピーカーを，それぞれのすぐ前方の卓上にマイクを配置し，各スピーカーから再生した Time Stretched Pulse (TSP) を各マイクロホンで録音し，逆 TSP を畳み込んでインパルス応答を得る．スピーカーとマイクロホンのすべての組み合わせを考え，スピーカーとマイクロホンが近接している場合 (Fig. 3 では赤い矢印) を近距離，それ以外の場合（青い矢印）を遠距離とした．収録環境をその空間の規模に応じて，数名から 10 名程度で利用する小規模の環境 (Room)，数十名程度で利用する中規模の環境 (Office)，100 名以上で利用する大規模の環境 (Hall) の 3 種類に分類し，種類ごとにデータセットを構築した．データセットは近距離音・遠距離音・背景音の混合音（入力）と近距離音のみ（出力）の組からなる．構築したデータセットはおよそ 100 時間分であり，うち 75% を訓練データ，25% を評価データに用いた．

Table 2　PESQ による評価結果

環境	Room	Office	Hall
混合音	1.33	1.31	1.54
近距離強調音	1.85	1.89	1.51

Table 3　STOI による評価結果

環境	Room	Office	Hall
混合音	0.68	0.76	0.80
近距離強調音	0.72	0.89	0.68

ネットワークの学習におけるオプティマイザには Adam [10] を用い，学習率は 0.001 とした．損失関数として，正解の近距離音スペクトログラムと強調近距離音スペクトログラムの平均二乗誤差を用いた．

訓練したモデルを用いて強調した音声を PESQ で評価した結果を Table 2 に，STOI で評価した結果を Table 3 に示す．Room と Office では提案法により PESQ と STOI のいずれもが増加しており，提案手法による強調音声は明瞭度が向上していることが示された．一方で Hall では，PESQ と STOI がいずれも低下した．Hall は大規模な環境であるために反響成分が Room と Office よりも相対的に大きいことが原因の一つであると考えられるが，さらなる分析による原因究明が必要である．

5　おわりに

本稿では，RRN による時間周波数マスクに基づくモノラル混合音に対する近距離音声強調手法を提案した．実環境を模したデータセットを用いた評価実験により，小から中規模の環境では，提案手法による強調音声の明瞭度向上が示された．今後は，大規模環境での明瞭度低下の原因解明，音声強調と音声認識を統合したシステムの開発と評価などに取り組む予定である．

参考文献

[1] K. Patterson et al. Distance-based sound separation. In *INTERSPEECH 2022*, pages 901–905, 2022.
[2] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
[3] S. Robin et al. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *ICASSP 2018*, pages 351–355, 2018.
[4] K. He et al. Deep residual learning for image recognition. In *CVPR 2016*, pages 1–12, 2016.
[5] K. Cho et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP 2014*, pages 1724–1734, 2014.
[6] J. Kahn et al. Libri-Light: A benchmark for ASR with limited or no supervision. In *ICASSP 2020*, pages 7669–7673, 2020.
[7] J. Barker et al. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *ASRU*, pages 504–511, 2015.
[8] ITU-T Recommendation. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862*, 2001.
[9] C. H. Taal et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans. ASLP*, 19(7):2125–2136, 2011.
[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR 2015*, 2015.