5R-05

深層学習を用いた複数音声強調処理のアンサンブル手法の検討

藤田 雅彦[1], 糸山 克寿[1], 西田 健次[1], 中臺 一博[1,2]
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

音声は意思疎通を図るうえで最も自然で使いやすい手段の一つである．コンピュータやロボットが実環境で音声を適切に扱うことができるようになれば，人との自然なコミュニケーションの実現に近づくことが期待される．ほとんどの場合，実環境で収録された音声には雑音や残響が含まれており，これらは音声認識などのサービスの性能を低下させてしまう．音声強調は，音声に混入した雑音や残響を取り除き，人間にとって聞きやすく，コンピュータやロボットにとって処理しやすいクリーンな音声を得るための技術である．

音声強調はこれまでに数多く研究されている．例えば，Heymann らはニューラルネットワーク (NN) を用いて時間周波数マスクを推定し，さらにこれをビームフォーミングと組み合わせる手法を提案した [1]．時間周波数マスクは，雑音混じり音声のスペクトログラムと同じサイズの行列として定義され，スペクトログラムとマスクの要素積をとることで目的音声のみを通過させ音源分離や音声強調を実現する手法である．この手法には NN によってマスクを正確に推定できるという利点がある一方で，NN の学習に用いていない未知の環境雑音に対しては性能が劣化してしまうという問題がある．この他にも，独立成分分析 (ICA) [2] や非負値行列因子分解 (NMF) [3] 等を用いた音声強調手法がこれまでに提案されているが，いずれの手法についても収録環境や雑音の種類に得手不得手があるため，あらゆる環境でロバストに動作する音声強調手法が必要とされている．

本稿では，畳み込みニューラルネットワーク (CNN) を用いて複数の音声強調処理をアンサンブルした，環境や雑音の変化に対してロバストな音声強調手法を提案する．アンサンブルは機械学習で用いられる識別器の汎化性能を向上させるために用いられる手法である．複数の音声強調手法から生成された時間周波数マスクを CNN を用いてアンサンブルすることで，それぞれの手法の長所を併せ持つアンサンブル時間周波数マスクを生成し，音声を強調する．

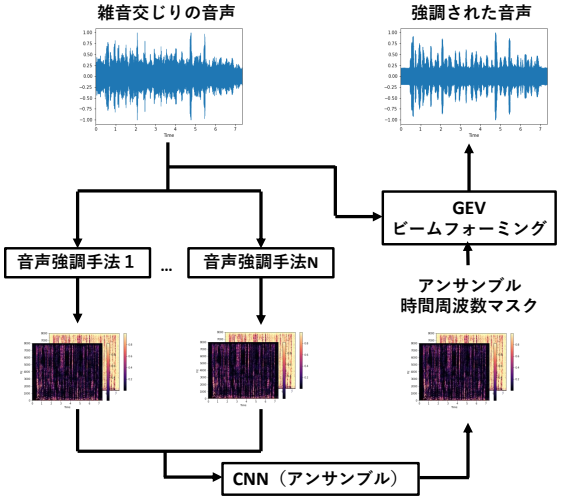

Fig. 1　提案手法のシステム概要

2　提案手法

雑音を含む音声を入力として，全 N 個の音声強調手法を用いてそれぞれで時間周波数マスクを推定する．アンサンブルにより統合されたマスク（アンサンブル時間周波数マスク）を生成し，これをもとにビームフォーミングを行い，目的の音声を強調抽出する．このシステムは藤田らが提案したもの [4] とほぼ同一の構成であるが，マスクのアンサンブルに CNN を用いることで，時間周波数ごとに異なるマスク重みを入力音声に応じて推定することが可能となる．アンサンブル時間周波数マスクを用いた Generalized EigenValue (GEV) ビームフォーミング [1] により，最終的な強調音声を得る．Fig. 1 に提案手法の概要を示す．

CNN の構成を Table 1 に示す．入力は各音声強調手法によって生成された Nch のマスクであり，出力は 1ch のアンサンブルマスクである．全体の構成は，2 次元の畳み込みを行う前半部と逆畳み込みを行う後半部分に分かれる．出力層の手前にソフトマックス関数を導入し，入力との積を取ることで出力としている．これは時間周波数マスクの各時間周波数ビンで次のような入力マスクの加重平均が行われることを期待している．

$$\hat{\boldsymbol{M}}_{ft} = \sum_{n=1}^{N} \alpha_{nft} \boldsymbol{M}_{nft} \tag{1}$$

An Ensemble Method for Multiple Speech Enhancement Using Deep Learning
Masahiko Fujita[1], Katsutoshi Itoyama[1], Kenji Nishida[1], Kazuhiro Nakadai[1,2]
1 Dept. of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology
2 Honda Research Institute Japan Co., Ltd.

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

Table 1　CNN の構成

Input: N-ch time-frequency masks, $800 \times 512 \times N$
Conv2D 3×3 @ 32, ReLU
MaxPooling2D 2×2, stride $= 2$
Conv2D 3×3 @ 16, ReLU
MaxPooling2D 2×2, stride $= 2$
Conv2D 3×3 @ 8, ReLU
MaxPooling2D 2×2, stride $= 2$
Conv2D 3×3 @ 8, ReLU
TransConv2D 2×2, stride $= 2$ @ 8, ReLU
TransConv2D 2×2, stride $= 2$ @ 16, ReLU
TransConv2D 2×2, stride $= 2$ @ 32, ReLU
TransConv2D 2×2, stride $= 2$ @ N, ReLU
Softmax
Multiply with input and sum
Output: 1 ch time-frequency mask, 800×512

ここで, $\hat{\boldsymbol{M}} \in \mathbb{R}^{F \times T}$ はアンサンブルマスク, $\boldsymbol{M}_n \in \mathbb{R}^{F \times T}$ は各手法によって生成されたマスク, $\alpha_n \in \mathbb{R}^{F \times T}$ は各マスクに対応する重み行列である.

3　実験・考察

提案手法の有効性を評価するために, シミュレーションによる実験を行った. 評価指標として Perceptual Evaluation of Speech Quality (PESQ) [5] を用いた. PESQ は −0.5 から 4.5 までの実数値で表され, 値が大きいほど人にとって聴きやすい音声であることを示す.

評価のためのデータセットをシミュレーションにより作成した. クリーンな音声として CHiME3 dataset [6] 内の 4 話者の計 330 発話 (16 kHz, 16 bit, 1 ch) を用いた. 雑音として, Freesound[*1]からダウンロードしたヘリコプターの雑音 (Heli) とレストランの室内の雑音 (Babble) の 2 種類を用いた. これらの雑音は 44.1 kHz で提供されているため 16 kHz にダウンサンプリングした. SN 比が 0 dB となるように音量を調節し, Fig. 2 に示すような環境で混合音を作成した. CNN の学習におけるオプティマイザには Adam を用い, 学習率は 0.001 とした. 損失関数として平均二乗誤差を用いた.

アンサンブルの対象の音声強調手法として次の 2 手法を用いた. ひとつは NN でマスクを推定する手法 (NN) [1] である. NN は CHiME3 dataset の学習用の模擬録音データによって学習されており, bus, cafe, pedestrian, street の 4 種類の雑音を既知としている. もうひとつの手法として ILRMA [7] を用いた. ILRMA は低ランクな雑音の高精度な除去が可能である一方で, 低ランクでない雑音を十分に取り除けない場合がある.

各音声強調手法で強調した音声を PESQ で評価した

[*1] https://freesound.org/

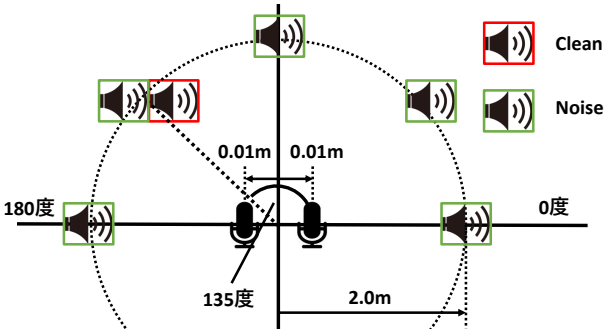

Fig. 2　シミュレーション環境

Table 2　各手法により強調された音声の PESQ スコア

	Noisy	NN	ILRMA	Proposed
Heli	1.12	2.08	2.72	2.27
Babble	1.13	2.32	2.58	**2.64**

結果を Table 2 に示す. Noisy は強調前の観測音声, NN, ILRMA, Proposed はそれぞれの手法による強調音声を表す. ILRMA は低ランクな Heli に対する PESQ スコアが高く, NN は既知である Babble に対する PESQ スコアが高いことが分かる. また, 提案手法は, Babble に対して NN と ILRMA の強調性能を上回った. この結果により, 提案手法は 2 手法の長所を併せ持つマスクを生成できることが示された.

4　おわりに

CNN を用いた時間周波数マスクのアンサンブルを行う音声強調手法を提案した. 評価実験によって CNN によるアンサンブルの有効性を示した.

謝辞　本研究は JSPS 科研費 JP19K12017, JP19KK0260 および JP20H00475 の助成を受けた.

参考文献

[1] J. Heymann et al. Neural network based spectral mask estimation for acoustic beamforming. In *ICASSP*, pages 196–200, 2016.
[2] P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.
[3] K. W. Wilson et al. Speech denoising using nonnegative matrix factorization with priors. In *ICASSP*, pages 4029–4032, 2008.
[4] 藤田 et al. アンサンブル時間周波数マスクによる音声強調手法の検討. In **情処全大**, 2021. 7N-6.
[5] 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.
[6] J. Barker et al. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In *ASRU*, pages 504–511, 2015.
[7] D. Kitamura et al. Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM TASLP*, 24(9):1626–1641, 2016.

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.