7N-06

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

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

近年, 深層学習の発展に伴い，音声認識技術の性能も大きく向上している．音声認識の主要な問題の一つとして, 雑音が酷い環境において単語や発話個所の誤検出が増え, 認識精度が劣化することが挙げられる．この問題に対し, 音声強調技術が長年にわたり研究されているが，その中でも, 時間周波数マスクとビームフォーミングを組み合わせた音声強調手法は, 高い強調性能があることが報告されている [1, 2]．一方で, 従来の時間周波数マスクベースの音声強調手法は, 単一のモデルやネットワークから時間周波数マスクを推定するため，十分に入力信号に含まれる音声強調の鍵となる特徴量を活かしきれていないと考えられる．そこで，本稿では異なる音声強調手法から推定される複数の時間周波数マスクをアンサンブルすることで, 処理のロバスト性の向上を図る．アンサンブルはパターン認識分野等で用いられるモデルの汎化性能を向上させる手法である．

2　提案手法

Fig.1 に提案手法のフローを示す．入力音声信号には, 目的音声の他に雑音が混入している．この入力信号を, 短時間フーリエ変換により振幅スペクトログラムに変換し, 得られる振幅スペクトログラムに対して N 個の音声強調手法を適用し, N 個の時間周波数マスクを推定する．これらのマスクをアンサンブルすることで, アンサンブル時間周波数マスクを生成する．さらに, 得られるアンサンブル時間周波数マスクから空間相関行列を推定し, Generalized EigenValue (GEV) ビームフォーミング [1] を行うことで, 目的音声を強調抽出する．

本稿では, アンサンブル方法として, 単純平均と加重平均を提案する．

単純平均では, アンサンブル対象の音声強調手法から 2 つ以上の手法を選び, それらから生成したマスクに対して単純平均を行う．

加重平均では, 学習用のデータを用いて, 事前に最適化問題を解き, 各手法のマスクに対する重みを決定する．具体的には、下式で表すように，理想マスク I とアンサンブル時間周波数マスクの平均二乗誤差を最小化するマスク重み $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_N]^T$ を求める最適化問題として定義した。

$$\boldsymbol{\alpha} = \min_{\boldsymbol{\alpha}} \frac{1}{DFT} \sum_{D,F,T} \left(I_{dft} - \sum_{n=1}^{N} \alpha_n M_{ndft} \right)^2$$

ここで, $I \in \mathbb{R}^{D\times F\times T}$ は理想マスク, $M_n \in \mathbb{R}^{D\times F\times T}$ は強調手法 n $(n = 1, 2, ..., N)$ から生成した目的音声または雑音の時間周波数マスク, α_n は各マスクに対する重みを表す．また, N は用いるアンサンブルに用いる強調手法の個数, D はデータ数, F は周波数, T はフレーム数, d, f, t はこれらに対応するインデックスを示す．制約条件として, $\sum \alpha_n = 1, 0 \leq \alpha_n \leq 1$ を課した．

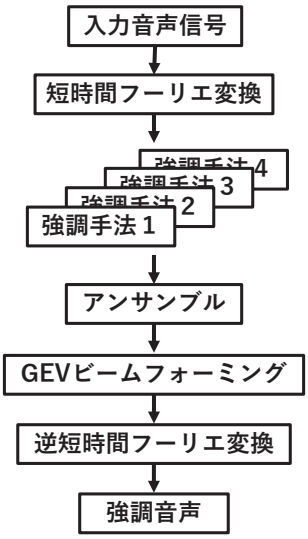

Fig. 1　System Overview of the Proposed Method

3　実験・考察

提案手法の有効性を評価するため, 強調された音声の品質評価を行った．評価指標として人間の聴感と相関がある Short-Time Objective Intelligibility (STOI) [3] と Perceptual Evaluation of Speech Quality (PESQ) [4] を用いた．STOI は 0–1, PESQ は 1–5 のそれぞれ実数で算出され, 値が大きいほど人にとって聴きやすい音声となる。また, 評価のためのデータとして CHiME-3 Challenge [5] のデータセットを用いた．CHiME-3 デー

A Study on a Speech Enhancement Method Using Ensemble Time-Frequency Masking
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.

タセットはWSJ0コーパス [6] と環境雑音を組み合わせて作成され, 実録音と模擬録音で構成されている. 評価には, “on the bus”,“cafe”,“pedestrian area”,“street” の 4 種類の雑音環境で, 4 人の話者の発話を 16 kHz, 6 ch で録音した合計 1320 個の評価用実録音を用いた.

アンサンブルには, [1, 7, 8, 9] の 4 つの音声強調手法を用いた. [1] は双方向 LSTM ネットワーク (BLSTM) によって時間周波数マスクを推定する. [7] は多チャネル非負値行列因子分解 (MNMF) と DNN を組み合わせてマスクを推定する. [8] では 2 種類の LSTM ネットワークによる 2 つの信号変換 (Dual-signal Transformation LSTM Network: DTLN) によって音声を強調する. [9] は敵対性生成ネットワークを音声強調に用いた手法 (Speech Enhancement Generative Adversarial Network: SEGAN) である. アンサンブルの対象として 4 つの音声強調手法を用いたので, 全部で 11 通りの単純平均アンサンブルを行った.

評価実験を行う前に, 先述した加重平均アンサンブルの重み決定アルゴリズムを用いて, 4 つの手法から生成したマスクに対する重みを推定した. 重み推定には, 84 人の発話と 4 種類の雑音から作成された合計 7138 個で構成される CHiME-3 データセットの学習用模擬録音を用いた. 各強調手法 (BLSTM, MNMF, DTLN, SEGAN) から生成された目的音声と雑音の時間周波数マスクに対する重み $\boldsymbol{\alpha}_X, \boldsymbol{\alpha}_N$ をそれぞれ示す.

$$\begin{aligned} \boldsymbol{\alpha}_X &= [0.22, 4.26 \times 10^{-6}, 0.63, 0.15]^T \\ \boldsymbol{\alpha}_N &= [0.28, 0.40, 0.26, 0.06]^T \end{aligned}$$

BLSTM や DTLN に対する重みが大きいのは, LSTM ネットワークによって正確にマスクを推定することができているためであると考察する. また, MNMF の雑音のマスクに対する重みが大きいのは, 雑音のモデル化が正確に行われているためだと考察する.

各手法によって強調した音声を STOI, PESQ で評価し, 平均を取ったものを Table 1 に示す. いずれの指標においても, BLSTM と DTLN の単純平均がアンサンブルをしていない場合を上回り, 全体で最も良い結果となった. 加重平均アンサンブルは, 最も良いとはならなかったが, 最も良い組合せと同等の結果が得られた. 加重平均アンサンブルが最も良くならなかった理由として、理想マスクとの誤差の最小化が必ずしも, STOI や PESQ での評価においては最適でなかったからだと考察する.

4 おわりに

本稿では, 複数の時間周波数マスクから生成したアンサンブル時間周波数マスクを用いた音声強調手法を提案した. 強調された音声の品質評価を行い, アンサンブルの有効性を示した. 加重平均アンサンブルで推定した重みや設定した目的関数の妥当性の検証, 音声認識を用いた評価を行うことが今後の課題である.

Table 1 Results of STOI and PESQ evaluation of speech enhanced by each method.

強調手法	STOI	PESQ
強調なし	0.413	1.849
BLSTM	0.521	2.306
MNMF	0.481	2.060
DTLN	0.521	2.323
SEGAN	0.486	2.099
BLSTM/MNMF	0.519	2.273
BLSTM/DTLN	**0.524**	**2.332**
BLSTM/SEGAN	0.518	2.280
MNMF/DTLN	0.519	2.282
MNMF/SEGAN	0.486	2.084
DTLN/SEGAN	0.518	2.289
BLSTM/MNMF/DTLN	0.522	2.295
BLSTM/MNMF/SEGAN	0.517	2.262
BLSTM/DTLN/SEGAN	0.522	2.313
MNMF/DTLN/SEGAN	0.517	2.267
BLSTM/MNMF/DTLN/SEGAN	0.521	2.289
加重平均アンサンブル	0.522	2.294

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 *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[2] Xiong Xiao et al. On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3246–3250. IEEE, 2017.

[3] Cees H Taal et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

[4] 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.

[5] Jon Barker et al. The third ‘chime’ speech separation and recognition challenge: Dataset, task and baselines. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 504–511. IEEE, 2015.

[6] John Garofalo et al. Csr-i (wsj0) complete. *Linguistic Data Consortium, Philadelphia*, 2007.

[7] Kouhei Sekiguchi et al. Semi-supervised multichannel speech enhancement with a deep speech prior. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2197–2212, 2019.

[8] Nils L Westhausen et al. Dual-signal transformation lstm network for real-time noise suppression. *arXiv preprint arXiv:2005.07551*, 2020.

[9] Santiago Pascual et al. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.