ゲート付き CNN を用いた深層クラスタリングによる音源分離 *

◎李 莉 [1,2], 亀岡 弘和 [1]
[1] 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
[2] 筑波大学

1 はじめに

本稿では，モノラルの複数音声分離問題を扱う。複数の音声信号が混じる混合信号から各音声信号を分離する音声分離技術は音声認識や補聴器などを高精度化するために重要となる技術である。近年の深層学習の発展により，いくつかニューラルネットワーク（Neural Network: NN）を用いた強力な音声分離手法が提案されており [1–5]，特に教師ありタスクにおいて従来のモデルベースの音声分離手法を凌駕する性能を発揮することが報告されている。深層クラスタリング [4] はその代表例である。

深層クラスタリングでは，混合信号のスペクトログラムの時間周波数点ごとに埋め込みベクトルを考え，同一音源が支配的な時間周波数点の埋め込みベクトルが互いに近接するように時間周波数点特徴から埋め込みベクトルへの写像を NN を用いて学習することで，テスト時に埋め込みベクトルにクラスタリングを行うことにより各音源の時間周波数マスクを推定する方法である。従来，埋め込みベクトルへの写像関数として，時系列データを扱う再帰型 NN（Recurrent NN: RNN）の一種である双方向長･短期記憶（Bidirectional long short-term memory: BLSTM）ネットワークが用いられている。LSTM では入力ゲート，忘却ゲートと出力ゲートと呼ぶ三つの制御ゲートを用いることで，従来の RNN の勾配消失問題を解決し，入力と出力の長期依存関係を学習することが可能になった。その優れた性能は今まで数多くのタスクにおいて示されいた。しかしながら，一般に BLSTM を含む RNN には，多層になるほど過学習が生じやすくなる点や学習に多大な計算コストを要する点などの問題点があることが知られている。

一方，畳み込み NN（Convolutional NN: CNN）は RNN に比べ過学習を起こしにくい点，並行計算に向いている点などの利点を有することから，近年時系列データを扱うモデルとしても注目が高まっている。最近提案された「ゲート付き CNN」[6] は，時系列データの長期依存関係を効率的に捉えることが可能で，入力文章における後続単語を予測する言語モデルとしての能力が LSTM を凌駕することが報告されている。通常の CNN を用いてデータの長期依存関係を学習するためには通常ネットワークの多層化が必要となるが，多層化により勾配消失問題が生じやすくなる点が課題であった。これに対しゲート付き CNN は，LSTM と同様に線形出力を変調させるゲート構造を CNN に導入することにより各層で通過させたい情報の制御を可能にしつつ勾配消失を防ぐことができる特長を有している。本稿では，以上のゲート付き CNN の特長に着目し，埋め込みベクトルへの写像をゲート付き CNN でモデル化した深層クラスタリングを提案する。

2 深層クラスタリング

C 個の音源からなる混合信号の時間周波数表現（スペクトログラム）をベクトル化したものを $\mathbf{x} = [x_1, \ldots, x_n, \ldots, x_N]^T \in \mathbb{R}^N$ とする。ただし，n は時間周波数点 (f, t) に対応するインデックスを表し，N は時間周波数点の総数 $F \times T$ である。深層クラスタリングではまず，スペクトログラムの各点 x_n ごとにノルムが 1 の D 次元埋め込みベクトル $\mathbf{v}_n = [v_{n,1}, \ldots, v_{n,D}]$ を考え，同一音源が支配的な時間周波数点の埋め込みベクトルが互いに接近するように線形写像 $\mathbf{V} = g_{\mathbf{\Theta}}(\mathbf{x})$ を学習することが目標である。ただし，$\mathbf{V} = [\mathbf{v}_1; \ldots; \mathbf{v}_N] \in \mathbb{R}^{N \times D}$ である。従来の深層クラスタリングでは，$g_{\mathbf{\Theta}}$ は BLSTM によりモデル化されており，Θ はそのパラメータを表す。$\mathbf{x}$ の各時間周波数点 n で支配的な音源ラベルを示した one-hot ベクトル（行ベクトル）を $\mathbf{y}_n \in \{0,1\}^{1 \times C}$ とし，$\mathbf{Y} = [\mathbf{y}_1; \ldots; \mathbf{y}_N] \in \{0,1\}^{N \times C}$ とすると，深層クラスタリングでは

$$\mathcal{J}(\mathbf{V}) = ||\mathbf{V}\mathbf{V}^T - \mathbf{Y}\mathbf{Y}^T||_F^2 \qquad (1)$$

を最小化するように Θ を学習する。ただし，$||\cdot||_F^2$ は Frobenius ノルムを表す。$\mathbf{Y}\mathbf{Y}^T$ は，n 行 n' 列目の要素が時間周波数点 n と n' において同一音源が支配的のときに 1，そうでないときに 0 であるような $N \times N$ のバイナリ行列で，類似度行列と呼ぶ。パラメータ Θ の学習完了後，入力信号のスペクトログラム $\mathbf{x}$ に対し $\mathbf{V}$ を算出し，$\mathbf{V}$ の各行ベクトルをデータベクトルとしてクラスタリング（k 平均クラスタリングなど）を行うことで，同一音源が支配的な時間周波数点の集合を得ることができる。これにより，音源分離を行うための各音源の時間周波数マスクを構成することができる。

*Multi-speaker separation using deep clustering with gated CNN, Li Li (NTT Communication Science Laboratories/University of Tsukuba), Hirokazu Kameoka (NTT Communication Science Laboratories)

従来，NN を適用した音声分離法としては入力スペクトログラムに対し各時間周波数点の音源ラベルを直接予測するアプローチ（例えば [1–3]）が主流であったが，このアプローチでは，教師データを用意する際，各スペクトログラム間で音源ラベルが一貫していない場合，性能低下に直結する点に課題があった。例えば，音源 A と音源 B からなる混合信号のスペクトログラム A とスペクトログラム B を学習データとして，スペクトログラム A では音源 A にラベル $[1,0]$，音源 B にラベル $[0,1]$ が付与され，逆にスペクトログラム B では音源 A にラベル $[0,1]$，音源 B にラベル $[1,0]$ が付与されていたとすると，このような学習データで学習された識別器は，テストデータのスペクトログラムの各点で音源 A と音源 B を識別する能力をもちえない。このため，教師データを準備する際は，各スペクトログラム間で一貫したラベルを慎重に付与する必要があり，利用場面によっては難点となりえる。これに対し深層クラスタリングは教師データとして，各時間周波数点に付与される音源ラベル $\mathbf{Y}$ は必要とせず，代わりに，各スペクトログラムの時間周波数点のペアごとに支配的な音源が同一かどうかを示す類似度行列 $\mathbf{YY}^T$ を用いる手法となっている。このようなラベルの付与にかかる労力は，全データ間で一貫した音源ラベルを付与する労力に比べて小さく済むため，実用上のメリットが大きい。

3　提案手法

従来の深層クラスタリング法では，埋め込みベクトルへの写像関数 $g_{\mathbf{\Theta}}$ として RNN の一種である BLSTM ネットワークが用いられていたが，RNN ベースのネットワークは多層になると学習が安定しない，学習に時間がかかる，過学習しやすい，などの問題が生じることが知られている。そこで本稿では，CNN の特長に着目し，$g_{\mathbf{\Theta}}$ を CNN でモデル化した深層クラスタリングを提案する。更に，複数音声分離タスクに対する適切なネットワークアーキテクチャについて検討する。具体的には，(1) 1 次元 CNN or 2 次元 CNN ; (2) ゲート付き CNN ; (3) Dilated CNN ; (4) Strided CNN ; (5) スキップアーキテクチャを組み合わせたネットワークアーキテクチャを用いる。

(1) 1 次元 CNN or 2 次元 CNN

1 次元 CNN は，入力 $\mathbf{x}$ を F チャンネルのサイズが $1 \times T$ の画像と見なし，出力 $\mathbf{V}$ を $F \times D$ チャンネルのサイズが $1 \times T$ の画像と見なす場合，2 次元 CNN は，入力 $\mathbf{x}$ を 1 チャネルのサイズが $F \times T$ の画像と見なし，出力 $\mathbf{V}$ を D チャネルのサイズが $F \times T$ の画像と見なす場合にそれぞれ相当するものとする。

(2) ゲート付き CNN

ゲート付き CNN は，元々単語列の予測モデルとして最初に導入された CNN の一種で，同条件の実験で

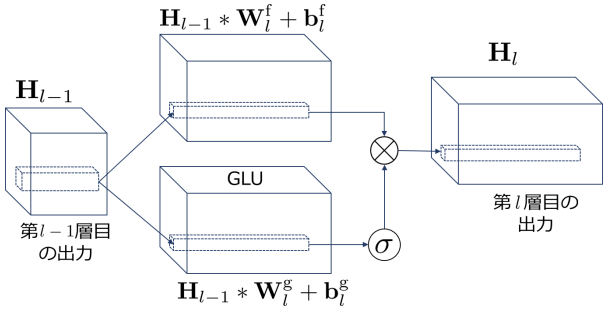

Fig. 1　ゲート付き CNN のアーキテクチャ

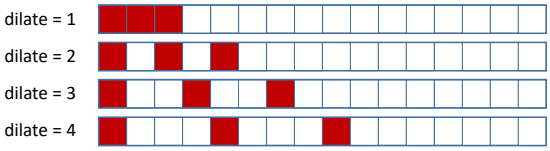

Fig. 2　異なる dilate 数における 1 次元 dilated CNN。

LSTM を超える単語予測性能を発揮することが報告されている。図 1 にはゲート付き CNN のアーキテクチャを示す。第 $l-1$ 層目の出力を $\mathbf{H}_{l-1}$ で表すものとし，ゲート付き CNN では，第 l 層目の出力 $\mathbf{H}_l$ は

$$\mathbf{H}_l = (\mathbf{H}_{l-1} * \mathbf{W}_l^{\mathrm{f}} + b_l^{\mathrm{f}}) \otimes \sigma(\mathbf{H}_{l-1} * \mathbf{W}_l^{\mathrm{g}} + b_l^{\mathrm{g}}) \quad (2)$$

で与えられる。ただし，$\otimes$ と σ は要素ごとの積とシグモイド関数を表し，$\mathbf{W}_l^{\mathrm{f}} \in \mathbb{R}^{D_l \times D_{l-1} \times \tilde{N}_l \times \tilde{M}_l}$，$\mathbf{W}_l^{\mathrm{g}} \in \mathbb{R}^{D_l \times D_{l-1} \times \tilde{N}_l \times \tilde{M}_l}$，$b_l^{\mathrm{f}} \in \mathbb{R}^{D_l \times N_l \times M_l}$ と $b_l^{\mathrm{g}} \in \mathbb{R}^{D_l \times N_l \times M_l}$ が推定すべきパラメータとなる。式 (2) を要素ごとに表記すると

$$H_{l,d,n,m} = \left(\sum_{d'=0}^{D_{l-1}-1} \sum_{n'=0}^{\tilde{N}_l-1} \sum_{m'=0}^{\tilde{M}_l-1} w_{l,d,d',n-n',m-m'}^{\mathrm{f}} H_{l-1,d',n',m'} + b_{l,d,n,m}^{\mathrm{f}} \right) \times \sigma\left(\sum_{d'=0}^{D_{l-1}-1} \sum_{n'=0}^{\tilde{N}_l-1} \sum_{m'=0}^{\tilde{M}_l-1} w_{l,d,d',n-n',m-m'}^{\mathrm{g}} H_{l-1,d',n',m'} + b_{l,d,n,m}^{\mathrm{g}} \right)$$

となる。第二項のような非線形活性化関数を Gated Linear Unit (GLU) と呼ぶ。

(3) Dilated CNN

Dilated CNN は，畳み込みする際にフィルタの隣接する点の間に 0 を埋めることによりパラメータ数を増えずに受容野の範囲を広げる CNN である。図 2 には異なる dilate 数における 1 次元 dilated CNN を示す。dilate 数が 1 の時には通常の CNN で，dilate 数が増えれば増えるほど大きなフィルタが作成され，受容野を広げることができる。

(4) Strided CNN

Strided CNN は，フィルタの畳み込みの適用間隔を 1 以外にすることを許容した CNN である。ストライド幅が 2 のときは畳み込みの出力のサイズは 1/2 になる。

Table 1　実験で採用した CNN のアーキテクチャ。“1D”と“2D”は 1 次元 CNN と 2 次元 CNN，“B”はボトルネックアーキテクチャ，“DC”は dilated CNN，“w/o skip”と“w/ skip”はスキップアーキテクチャのありなしを表す。表中の表記“$\tilde{N}_l \times \tilde{M}_l, D_l, \alpha, \beta$”は各層のフィルタサイズ’$\tilde{N}_l \times \tilde{M}_l$，出力チャンネル数 D_l，stride，dilation を表す。ただし，”64/128” はそれぞれ 5.5 時間および 30 時間の学習データのときに設定したチャネル数を表す。また，全層においてバッチ正規化層を含めた。

l	2D, B, w/o skip	2D, B, w/ skip	2D, DC, 5L	2D, DC, 8L	1D, DC
1	5×5, 64/128, 1, 1	5×5, 64/128, 1, 1	3×3, 64/128, 1, 1	3×3, 64, 1, 1	1×3, 512, 1, 1
2	4×4, 64/128, ↓ 2, 1	4×4, 64/128, ↓ 2, 1	3×3, 64/128, 1, 2	3×3, 128, 1, 1	1×3, 1024, 1, 2
3	3×3, 64/128, 1, 1	3×3 ,64/128, 1, 1	3×3, 64/128, 1, 3	3×3, 256, 1, 2	1×3, 2048, 1, 3
4	4×4, 64/128, ↓ 2, 1	4×4, 64/128, ↓ 2, 1	3×3, 64/128, 1, 4	3×3, 256, 1, 4	1×3, 4096, 1, 4
5	3×3, 64/128, 1, 1	3×3, 64/128, 1, 1	3×3, D, 1, 5	3×3, 256, 1, 8	1×3, 4096, 1, 4
6	4×4, 64/128, ↑ 2, 1	4×4, D, ↑ 2, 1		3×3, 256, 1, 16	1×3, 2048, 1, 4
7	4×4, D, ↑ 2, 1	4×4, D, ↑ 2, 1		3×3, 128, 1, 1	1×3, FD, 1, 4
8				3×3, D, 1, 1	

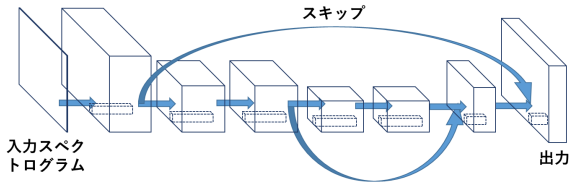

Fig. 3　スキップアーキテクチャを用いた strided CNN。

(5) スキップアーキテクチャ

スキップアーキテクチャは，入力または第 l 層の出力を第 $l+1$ 層以外にも第 $l+l'$ 層 $(l'>1)$ の入力とする NN のアーキテクチャをさす。それにより，strided CNN のストライド幅が 1 以上の場合に失った入力や低次特徴量の情報を深い層での高次特徴量を計算するときに用いることが可能になる。例として，図 3 はスキップアーキテクチャと strided CNN を用いた提案手法のアーキテクチャを示している。

4　モノラル音声分離実験

4.1　実験条件

提案手法の有効性を確認するため，Wall Street Journal（WSJ0）コーパスを用いて従来の BLSTM を用いた深層クラスタリングとの比較実験を行った。評価は処理前と処理後の信号対歪み比 (signal-to-distortion ratios: SDR)[10] の改善値を用いた。学習データ、検証データとテストデータは従来の研究 [4, 8] に従って [9] のプログラムで作成した。NN を学習するための学習データ 30 時間と検証データ 10 時間は WSJ0 データベースの `si_tr_s` フォルダーから任意の話者二人の音声を $[0, 10]$dB の範囲内で任意に選んだ信号対雑音比で重畳させて作成した。テストデータは `si_dt_05` と `si_et_05` フォルダーから任意の話者の音声を重畳させ，2 話者と 3 話者の混合信号を 5 時間ずつ作成した。ただし，学習データとテストデータには同一の話者の音声を含まないようにした。提案手法が少量の学習データでも過学習を起こさず動作するかどうかを確認するため，小規模の学習データ（約 5.5 時間分）と検証データ（約 0.5 時間分）を作成した。

すべての音声信号は 8kHz にダウンサンプリングし，フレーム長 254 サンプル点，フレーム間隔 127 点のハニング窓を用いて Fourier 変換を行い，周波数ビン数は $F = 128$ の対数振幅スペクトログラムを算出した。提案手法に採用した CNN アーキテクチャは任意の長さの入力に対応できるようにしたが，誤差逆伝播の勾配を計算する時間を節約するため，学習時においては，入力対数振幅スペクトログラムを $T = 128$ フレームごとにセグメンテーションを行い，その中の任意の 4 セグメントを用いて勾配を計算した。また，エネルギーが小さい時間周波数点（無音区間など）はどれの音源においても支配的ではないため，学習を不安定に導く可能性がある。この問題を防ぐため，従来研究と同様に誤差逆伝播の勾配を計算する際には，振幅が-40dB 以下の時間周波数点を考慮しないことにした。埋め込みベクトルの次元は 20 または 40 にした。NN のパラメータ学習は Adam optimizer を用いて行った。表 1 に本実験で提案法として採用した CNN の具体的なアーキテクチャを示す。全てのアーキテクチャにおいて全層でバッチ正規化を行い，GLU を活性化関数として用いった。

4.2　実験結果

比較実験のため従来の BLSTM ベースの深層クラスタリング法も実装したが，[4, 8] に報告された性能と同等の性能を達成できなかったので，下記，公平を期するため文献中の類似条件での結果を比較対象とする（我々の実装の結果も参考として一部併記する）。提案

Table 2　埋め込みベクトル次元が $D = 20$ のときの，30 時間および 5.5 時間の学習データの下での従来手法と提案手法による平均 SDR 改善値 [dB]。

モデル		学習データ 5.5h	30h
提案手法	2D, B w/o skip	3.90	5.49
	2D, B w/ skip	3.78	5.23
	2D, DC, 5L	**5.78**	**6.78**
	1D, DC	3.94	6.36
従来手法	BLSTM（我々の実装）	1.57	2.46
	BLSTM [4]	-	5.7

Table 3　埋め込みベクトル次元が $D = 40$ のときの，30 時間および 5.5 時間の学習データの下での従来手法と提案手法による平均 SDR 改善値 [dB]。

モデル		学習データ 5.5h	30h
提案手法	2D, DC, 5L	-	6.71
	2D, DC, 8L	6.77	8.32
	1D, DC	-	6.39
従来手法	BLSTM[4]	-	6.0
	BLSTM[8]	-	**9.4**

Table 4　3 音源の音声分離実験における従来手法と提案手法による平均 SDR 改善値 [dB]。

モデル		SDRi [dB]
提案手法	2D, DC, 5L	**3.14**
	2D, DC, 8L	2.43
	1D, DC	2.48
従来手法	BLSTM [4]	2.2

手法の各アーキテクチャが 2 音源分離タスクで埋め込みベクトル次元が 20 の場合に得られた SDR 改善量平均値を表 2 に示す。提案手法においてすべて 30 時間の学習データの下で，すべてのアーキテクチャで従来手法 [4] と同等以上の性能が得られた。2 次元/Dilated/ゲート付き CNN を用いた場合には BLSTM を用いた場合をはるかに超える性能が得られることを確認した。その上，少量の学習データでも過学習を起こさず動作できることを確認した。埋め込みベクトルの次元を 40 に増やした実験結果は表 3 のとおりである。表 2 と比べ，2 音源分離問題に対して埋め込みベクトルの次元数を 20 から 40 に増やすことは大きな性能改善をもたらせないことが分かった。2D/Dilated/ゲート付き CNN を用いたアーキテクチャのの層数を増やしても少量のデータに対して過学習を起こさず 1.6 dB 程度の性能改善が得られた。多層化にすることにより更なる性能向上が期待できる。更に表 4 に 3 音源の音声分離結果を示す。いずれのアーキテクチャも従来手法より高い分離性能が得られた。

5　まとめ

本稿では，CNN が有する利点に着目し，深層クラスタリングにおける時間周波数点から埋め込みベクトルへの写像をゲート付き CNN でモデルする手法を提案した。複数音声分離タスクにおいて適切な CNN アーキテクチャについて検討し，実験により提案法が BLSTM を用いた従来法より高性能であることが確認できたとともに 2 次元/Dilated CNN/ゲート付き CNN を用いた場合に少量の学習データでも過学習を起こさず従来法と相当する性能が得られることを確認した。

参考文献

[1] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Deep learning for monaural speech separation.”, In Proc. ICASSP, pp. 1562–1566, 2014.

[2] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation.”, IEEE/ACM Trans. ASLP, vol. 22, no. 12, pp. 1849–1858, 2014.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks.”, IEEE Signal Processing Letters, vol. 21, no. 1, pp. 65–68, 2014.

[4] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation.”, In Proc. ICASSP, pp. 31–35, 2016.

[5] M. Kolbak, D. Yu, Z. H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks.”, IEEE/ACM Trans. ASLP, vol. 25, no. 10, pp. 1901–1913, 2017.

[6] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” arXiv:1612.08083, 2016.

[7] F. Yu, and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” arXiv:1511.07122, 2015.

[8] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-channel multispeaker separation using deep clustering,” In Proc. Interspeech pp. 545-549, 2016.

[9] Avaliable online: http://www.merl.com/demos/deepclustering

[10] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” IEEE Trans. ASLP, vol. 14, no. 4, pp. 1462-1469, 2006.