一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-052-5 (12/3)

Mask U-Net を用いた環境音セグメンテーションの提案

Environmental sound segmentation utilizing Mask U-Net

周藤唯[*1]，糸山克寿[*1]，西田健次[*1]，中臺一博[*1,2]
Yui SUDOU，Katsutoshi ITOYAMA，Kenji NISHIDA，Kazuhiro NAKADAI
東京工業大学 工学院 システム制御系[*1]
（株）ホンダ・リサーチ・インスティチュート・ジャパン[*2]
Tokyo Institute of Technology，Honda Research Institute Japan

{sudo，itoyama，nakadai}@ra.sc.e.titech.ac.jp

Abstract

本稿では，Mask U-Net を用いた環境音のセグメンテーションについて提案する．近年，音データを用いた防犯監視システムや高齢者見守りシステムなどの要望が高まってきている．環境音のセグメンテーションは，空間情報を用いる伝統的な音源分離とは異なり，事前学習した音源の性質をもとに，区間検出，分離，識別を同時に行う手法である．このような手法として，画像の Semantic Segmentation 用に提案された U-Net を歌声信号分離に適用した例があるが，限定したクラスを対象とした適用にとどまっていた．本研究では，U-Net を用いたセグメンテーションと CNN を用いた音響イベント検出を組み合わせた Mask U-Net による環境音セグメンテーション法を提案し，より一般的なクラスを対象とした 75 クラスの環境音が含まれる混合音データに適用することによってその有効性を示す．結果として，従来手法と比較し，学習速度，音源分離性能が向上することを確認した．

1 はじめに

近年，音データを用いた防犯監視システムや高齢者の見守りシステムなど，音による環境理解に関する研究が数多くなされている[Peng 12]．例えば，高齢者の見守りシステムであれば，分離された咳の音データを用いた詳細な診断や，製造業であれば，分離された作動音を異常検知や生産データトレサビリティとして用いるといったように音源信号の有無だけではなく，複数の音源からの信号が混在した混合音から分離，識別を伴ったセグメンテーションが必要である.こうした研究として，従来は，音源の空間的な情報を利用して分離抽出を行い，分離抽出信号を識別するといった手法や，時間的なオーバーラップはないと仮定して，区間ごとに支配的な音源を識別する手法が多かった．これに対し，画像の Semantic Segmentation 用に提案された深層学習手法である U-Net[Ronneberger 15]を適用した歌声信号分離が高い性能を有する手法として報告されている[Jansson 17]．しかし，実環境に存在する多クラスを対象にした音源分離に関する研究は少ない．また，画像の Semantic Segmentation では，対象クラスのサイズが大きい，あるいは小さい場合，性能が悪化することが報告されている [Zhao 17]．したがって，少数かつ限定されたドメインのクラスの音源分離に適用された従来手法を多クラスの環境音に適用すると，時間方向にスパースな音データに対して，性能が悪化してしまうことが考えられる．

そこで，本研究では，従来の U-Net を用いた音源分離手法に対し、事前に学習された音響イベント検出モデルを併用した Mask U-Net を提案する．作成した多クラスの環境音データセットにおいて，既存手法と比較することで，その有効性を評価する．

2 従来手法

従来，非負値行列分解(Non-negative Matrix Factorization)による音源分離が提案されており，モノラル音源分離や歌声分離などの分野でその有効性が示されている[Smaragdis 14]．しかし，近年では，深層学習モデルを用いた音源分離手法が高い性能を示すことが報告されており，例えば U-Net を用いた手法が提案されている[Stoller 18]．しかし，ほとんどの研究はスピーチや歌声分離といった，少数のクラスを対象としており，実環境に存在する多数のクラスを対象にした音源分離に関する研究は少なく，従来手法をそのまま多クラスの環境音に適用したとしても，性能が悪化してしまう可能性がある．一方，画像のセグメンテーション手法としては，多数のクラスを対象とした手法が数多く提案されているが，中でも，従来のセグメンテーション手法に物体検出手法である Faster R-CNN[Ren 16]を組み合わせた Mask R-CNN が高いセグメンテーション性能を示すことが報告されている[He 18]．

本稿では，Mask R-CNN のアプローチを参考に，U-Net を用いた従来手法に，CNN を用いた音響イベント検出手法を組み合わせることによる環境音セグメンテーション法を検討し，Mask U-Net として提案する．

3 提案手法

本稿で提案する Mask U-Net 全体の構造を図 1 に示す．混合音として観測される環境音に対し，短時間フ

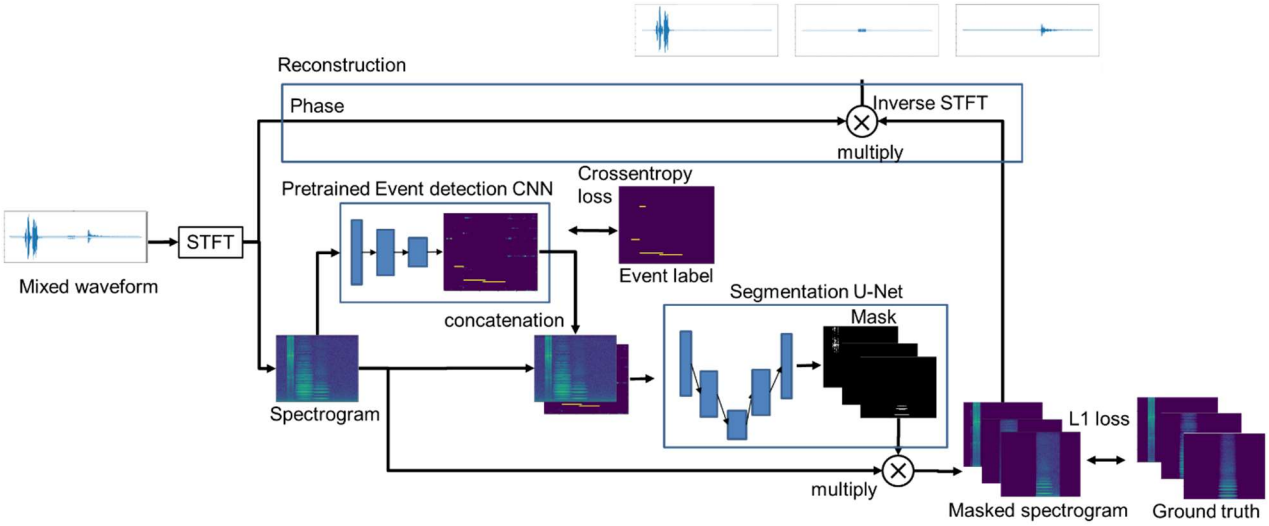


Figure 1 Architecture of Mask U-Net

ーリエ変換(STFT)を適用し，スペクトログラムに変換する．得られたスペクトログラムを画像と見立て，深層学習モデルに入力し，入力されたスペクトログラムから，各クラスを分離するマスクを予測する．マスクにより得られた各環境音のスペクトログラムは振幅情報しか持たないので，元の混合音データを STFT した際の位相を用いて，逆短時間フーリエ変換(Inverse STFT)を行い，時間領域信号に変換する．図 1 の Segmentation U-Net 部が従来手法を表し，提案手法は，その前段に音響イベント検出部を設ける．

本節では，従来手法である U-Net を用いた音源分離，CNN を用いた音響イベント検出手法について説明し，その後，提案する Mask U-Net について説明する．

3.1 U-Net のネットワーク構造

図 2 に本稿で使用した U-Net[Jansson 17]の構造を示す．U-Net は，エンコーダ層とデコーダ層で構成されている．エンコーダ層は，画像サイズを半減させながらチャネル数を 2 倍にする 2 次元畳み込みの繰り返し構造を持つ．すべてのエンコーダ層は 3x3 サイズのカーネルを持ち，ストライドは 2，パディングは 1 とする.各エンコーダ層では，batch normalization と leakness 0.2 の Leaky ReLU[Maas 13]を使用する．デコーダ層は，画像サイズを 2 倍にし，チャネルの数を半分に減らす deconvolution の繰り返し構造を持つ．3x3 サイズのカーネルを持ち，ストライドは 2，パディングは 1 とする．各デコーダ層では，batch normalization および ReLU を使用し，最初の 3 層は，50%のドロップアウトを適用する．さらに，同じ画像サイズを有するエンコーダ層およびデコーダ層はスキップ結合をもつ．これにより，低レベルの情報が高解像度入力から高解像度出力に直接流れる．最終層の活性化関数には Softmax 関数を用いる．学習には ADAM[Kingma 14]を用い，0.001 の学習率で 100 epoch 分学習させる．

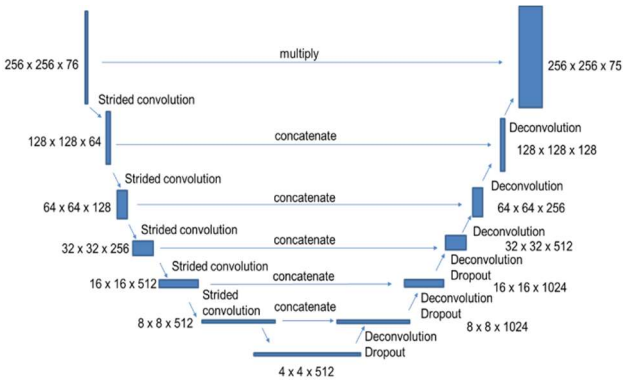


Figure 2 U-Net architecture

計算コストを削減するため，各音源は 16 kHz にダウンサンプリングする．フレーム長 512 サンプル，シフト長 128 サンプル，ハミング窓を窓関数として用いた短時間フーリエ変換(STFT) を行うことで振幅スペクトルを得る．その後，各音データの最大振幅値を用いて正規化を行う．

3.2 損失関数

学習に用いる損失関数を式(1)に表す．X は各環境音が混合された音源分離前のスペクトログラム， Y は各環境音のスペクトログラムの大きさを表す．モデルを学習するために使用される損失関数は，分離後の各環境音スペクトログラムとマスクされた入力スペクトログラムの差の $L_{1,1}$ ノルムを表す．

$$L(X,Y;\theta) = ||f(X,\theta)\cdot X - Y||_{1,1} \qquad (1)$$

ここで，$f(X,\ \theta)$は，モデルによって生成されたマスクであるパラメータθを有する入力 X に適用されるネットワークモデルの出力である．

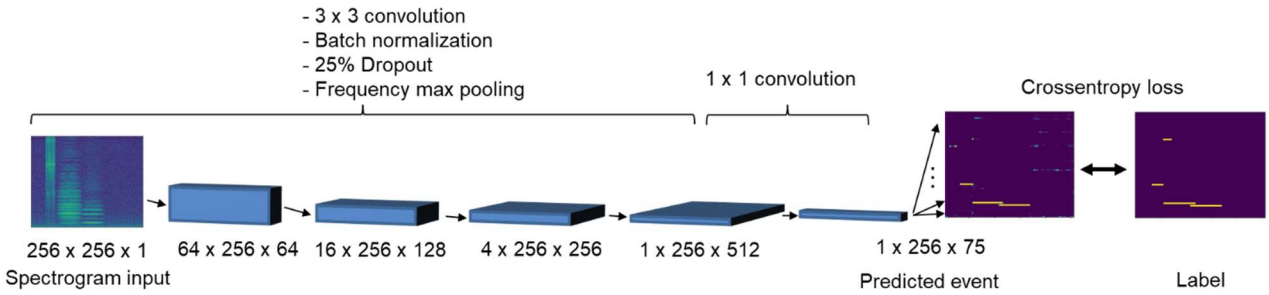


Figure 3 CNN architecture of the sound

3.3 CNN を用いた音響イベント検出

続いて, CNN を用いた音響イベント検出, および, 提案する Mask U-Net への適用方法について述べる. 音響イベント検出の手法については，多くの手法が提案されているが，本稿では，U-Net への結合の容易さを考慮し，スペクトログラムに対し CNN を適用した音響イベント検出を行う[Zhang 15]. CNN の詳細を図 3 に示す．最終層を除くすべての層において，カーネルサイズ 3x3 の畳み込み，ReLU，Max pooling, batch normalization, 25%のドロップアウトを適用する. 時間分解能を保つため，時間方向に対しては，Max pooling を行わず，周波数方向のみに対して，Pooling による次元削減を行う[Cakır 17]．最終層でカーネルサイズ 1x1 の畳み込みを行う.

3.4 音響イベント検出の環境音セグメンテーションへの適用

音響イベント検出部の出力に対し，スペクトログラムと同一次元になるよう後処理を行った後，スペクトログラムと concatenate してセグメンテーション部への入力とする. 具体的には, 音響イベント検出部の出力は，各時刻における各音響イベントの有無を示すベクトルであるのに対し，concatenate されるスペクトログラムは時間-周波数方向に成分を持つ行列であるため，図 4 に示すように，音響イベント検出の出力ベクトルをスペクトログラムの周波数方向の次元と同一になるよう複製することで，スペクトログラムと同一次元の行列にした後 concatenate を行う．また，通常，音響イベント検出では，出力に対して閾値を設け，音響イベントの有無を 2 クラスで表すが, 本手法においては, 音響イベント検出の出力を各環境音の存在確率として捉え，2 値化は行わない.

音響イベント検出の結果を事前情報としてセグメンテーション部に入力することによって，性能向上が可能であると考えられる.

3.4 時間領域の音響信号復元

本稿で使用する深層学習モデルは，各環境音のスペクトログラム、すなわち, 振幅の大きさのみを予測するため，そのままでは時間領域の音源信号を復元することができない．そのため，音源分離前の混合

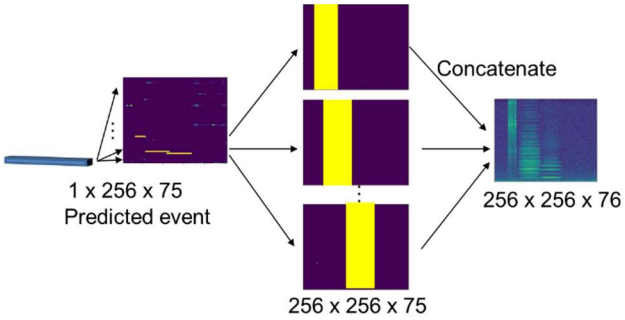


Figure 4 Concatenation of predicted sound events to produce the input of U-Net.

音の位相情報を使用して，時間領域の音響信号復元を行う.

4 評価実験

4.1 データセット

音源分離の学習のためには，混合音とそれに対応する正解音源信号のセットが必要になるため, 表 1 に示すコーパスのうち，ドライソースもしくは他の騒音が少なく，単一クラスのみが含まれた音源を選定し，これらをランダムに合成することで学習データを作成する．コーパスが異なる場合でも類似したクラスの音源はマージし，計 75 クラスの環境音データセットを作成した.計 10,000 個の合成混合音と正解音源信号のセットを作成し，学習データとして使用する. また, 学習データ作成に使用していないドライソースを用いて作成された評価データ 5,000 セットを作成し，以降の性能評価を行った.

4.2 音響イベント検出モデルの事前学習結果

事前情報として入力する音響イベント検出部の学習結果について述べる.式(2), (3)を用いて，F 値の算出する．TP，FP，FN はそれぞれ，True positive，False positive，False negative を表し，P，R はそれぞれ，Precision，Recall を表す.

$$\mathrm{P} = \frac{TP}{TP + FP}, \qquad \mathrm{R} = \frac{TP}{TP + FN} \tag{2}$$

$$\mathrm{F} = \frac{2P \cdot R}{P + R} = \frac{TP}{TP + (FN + FP)/2} \tag{3}$$

Table 1 Source domain dataset description

Database set	Contents	# of classes
ATR words	Male, female	2
RWCP	Bell, coin, buzzer, clock, phone, pinpong, whistle, rap, castanet, maracas, alarm, bottle, claps, air pump, book, phone, spray, tear	19
RWC-MusicDatabase	Timpani, cembalo, electric guitar, violin	4
Japan wild bird science	Bird	1
Bird research		
Grasshopper, cricket, grasshopper, singing voice	Insect	1
Sound database	Cat, baby, bird, footstep, frog, bathroom, clock, golf, tennis, trampoline, dog	11
Japanese cicada	Cicada	1
Freesound General-Purpose Audio Tagging Challenge Kaggle	Tearing, shatter, gunshot, fireworks, writing, computer keyboard, scissors, microwave oven, keys jangling, drawer open or close, knock, phone, saxophone, oboe, flute, clarinet, acoustic guitar, tambourine, gong, glockenspiel, snare drum, bass drum, hi-hat, electric piano, harmonica, trumpet, violin, double bass, cello, chime, cough, laughter, applause, finger snapping, fart, burping, cowbell, bark, meow	43
DCASE 2016 Task 2dataset	Clearthroat, cough, doorslam, drawer, keyboard, keysdrop, knock, laughter, pageturn, phone, speech	11
Total (similar classes were merged and target domain related were excluded /16kHz, 16bit)		75

事前学習した音響イベント検出モデルの F 値を算出すると 0.72 であった．また，音響イベント検出結果の例を図 6 に示す. ラベルデータと比較することで，本稿で用いたデータセットに対して，音響イベント検出ができていることが確認できる．

4.3 音源分離の結果

続いて，CNN を用いた音響イベント検出を事前情報として用いた Mask U-Net および既存手法である U-Net のセグメンテーション結果を比較する．式(4)により全体の RMSE(Root Mean Squared Error)，クラスごとの RMSE を算出した結果を表 2 および図 8 に示す．

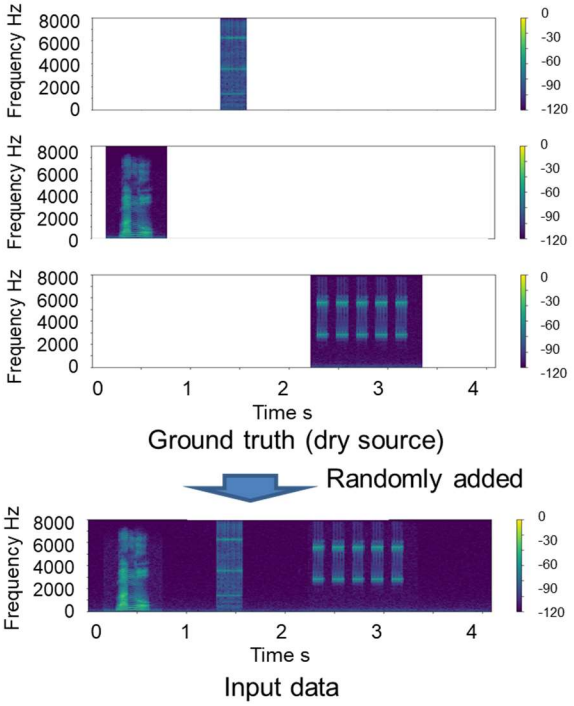


Figure 5 Mixed sound data

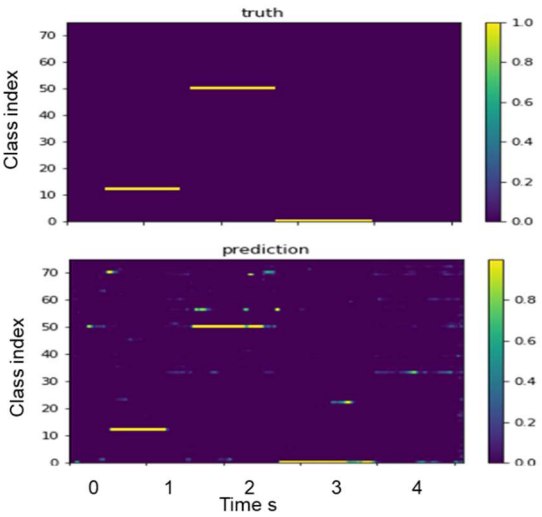


Figure 6 Results of predicted sound event detection

$$\mathrm{RMSE} = \sqrt{1/N(\boldsymbol{Y}_{true} - \boldsymbol{Y}_{pred})^2} \qquad (4)$$

表 2 に示す通り，従来手法と比較し，提案手法の RMSE が小さいことから，提案した Mask U-Net によって高精度に音源分離ができていることがわかる．図 8 も同様に，どのクラスも全体的に提案手法の方が，RMSE が小さいことがわかる．また，図 7 に各クラスに分離したスペクトログラム画像の例を示す．従来手法では，局所的に誤ったクラスに分類されている場合や，単一クラスの環境音に対し，複数のクラスが予測されている場合があることがわかる．また，オーバーラップ部分においては，サイズの大きい方

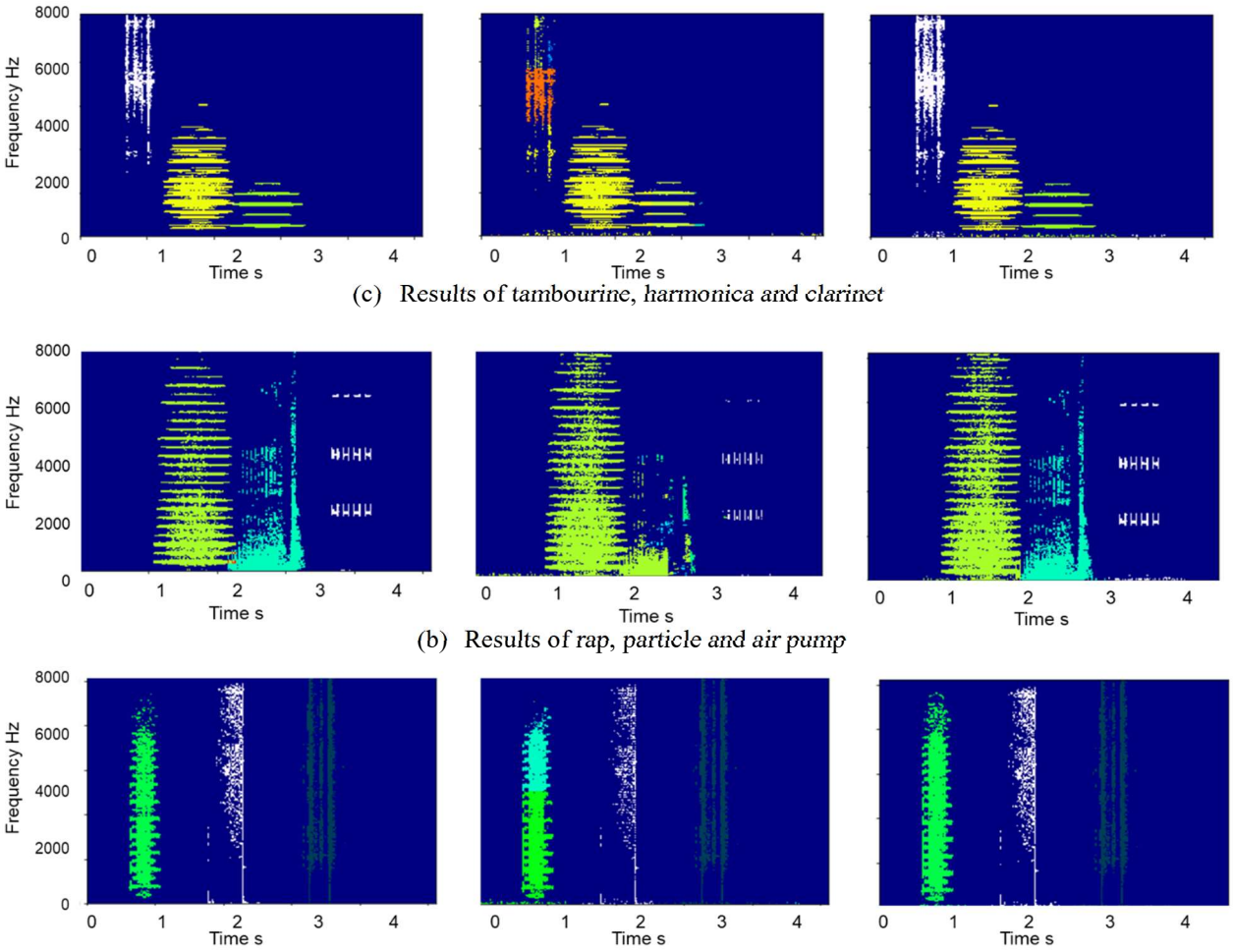


Figure 7 Results sound event separation. Left images show ground truth, center images show the results of U-Net and right images show the results of Mask U-Net

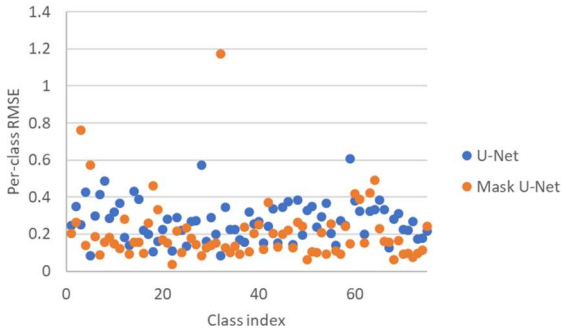
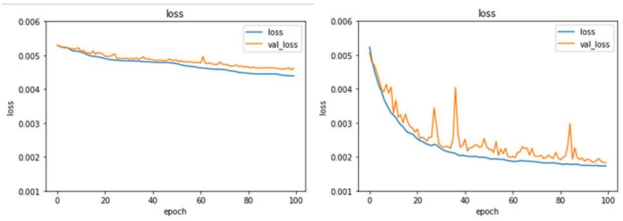


Figure 8 Per-class RMSE

Table 2 Evaluation of RMSE

RMSE	U-Net	Mask U-Net
Total	0.268	**0.201**



(a) U-Net (b) Mask U-Net

Figure 9 Comparison of each loss

のクラスに偏って分離されているのに対し，提案手法は，どのクラスも正しく分離することができていることがわかる．これは，事前に学習した音響イベント検出モデルにより，比較的高い精度で発生している環境音のクラス分類および発生区間が検出できており，その事前情報を U-Net を用いた音源分離モデルに入力しているため,クラス間の境界を明瞭に分離することができていると考えられる．さらに，図 9 に 100 epoch 学習時の損失関数の推移を示す．従来手法と比較し，提案手法はより早く学習が収束していることがわかる．事前情報によって学習速度の向上があることも，性能向上の要因として考えられる．

4.4 分離されたスペクトログラムからの音源再現

図 10 に，分離されたスペクトログラムおよび入力データの位相情報から，時系列データの復元を行っ

た結果を示す. 従来手法では, オーバーラップ部の境界がうまく学習できていないため, 再現された音源にも違うクラスの環境音が混ざってしまっているのに対し, 提案手法では, どのクラスも正しく音源を復元することができているといえる.

5 おわりに

本稿では, CNN を用いた音響イベント検出と U-Net を用いたセグメンテーション手法を統合した Mask U-Net を用いた, 多クラスの環境音セグメンテーション手法を提案した.

作成した環境音データセットを用いて提案手法を評価したところ, 提案手法は比較的高い精度で検出されたクラスごとの音響イベント区間情報を用いるため, 従来手法に比べ学習速度も速く, より高い音源分離性能を示した.

ただし, 本稿では, 雑音の少ない静かな環境を想定したため, 今後は雑音の大きな環境に対してもロバストな手法に取り組む予定である.

参考文献

[Peng 12] Peng, Y., Lin, C., Sun, M. and Tsai, K.: Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models, Proc. ICME, pp.1281-1221(2012)

[Ronneberger 15] Ronneberger, O.,Fisher, P. and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, In: MICCAI. LNCS, vol. 9351, pp.234-241. Springer(2015)

[Jansson 17] Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A. and Weyde, T.: Singing voice separation with deep U-Net convolutional networks, In Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), pages 323-332(2017)

[Zhao 17] Zhao, J., Shi, J., Qi, X., Wang, X. and Jia, J.: Pyramid Scene Parsing Network,arXiv:1612.01105(2017)

[Smaragdis 14] Smaragdis, P., Fevotte, C., Mysore, G., Mohammadiha, N. and Hoffman, M.: Static and dynamic source separation using nonnegative factorizations: A unified view. IEEE Signal Processing Magazine, 31 (3):66–75(2014)

[Stoller18] Stoller, D., Ewert, S. and Dixon, S.: Wave-U-Net: A multi-scale neural network for end-to-end audio source separation, 19th International Society for Music Information Retrieval Conference(2018)

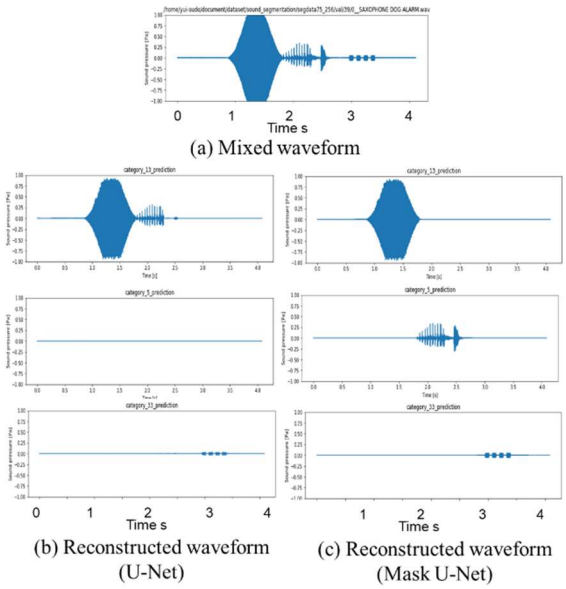


Figure 10 Reconstructed waveform (Saxophone, dog and alarm)

[Ren 17] Ren, S., He, K., Girshick, R. and Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks, in Proc. of NIPS(2016)

[He 17] He, K., Gkioxari, G., Dollar, P. and Girshick, R.: Mask R-CNN, in Proc. of ICCV(2016)

[Maas 13] Maas, A., Hannun, A. and Ng, A.: Rectifier nonlinearities improve neural network acoustic models, in Proc. of ICML (2013)

[Kingma 14] Kingma, D and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv: 1412.6980 (2014)

[Zhang 15] Zhang, H., McLoughlin, I. and Song, Y.: Robust sound event recognition using convolutional neural networks, in Proc. of ICASSP(2015)

[Cakır17] Cakır, E., Parascandolo, G., Heittola, T., Huttunen, H. and Virtanen, T.: Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection, arXiv:1702.06286v1 (2017)