[原著論文]

scSE-CRNN と 3 種類の呼吸音変換画像による呼吸音の分類

浅谷　尚希[†]・陸　　慧敏[†]・神谷　　亨[†*]・間普　真吾[††]・木戸　尚治[†††]

[†]九州工業大学　〒804-8550 福岡県北九州市戸畑区仙水町 1-1
[††]山口大学　〒755-8611 山口県宇部市常盤台 2-16-1
[†††]大阪大学　〒565-0871 大阪府吹田市山田丘 2-2
E-mail : asatani.naoki676@mail.kyutech.jp
*責任著者：神谷　　亨
（受理日：2021 年 6 月 23 日，採択日：2021 年 9 月 14 日）

Classification of Respiratory Sounds by scSE-CRNN from Triple Types of Respiratory Sound Images

Naoki ASATANI[†] Huimin LU[†] Tohru KAMIYA[†*] Shingo MABU[††], and Shoji KIDO[†††]

[†]Kyushu Institute of Technology, 1-1 Sensuicho, Tobata-Ku, Kitakyushu, Fukuoka, 804-8550 Japan
[††]Yamaguchi University, 2-16-1 Tokiwadai, Ube, Yamaguchi, 755-8611 Japan
[†††]Osaka University, 2-2 Yamadaoka, Suita, Osaka, 565-0871 Japan
E-mail : asatani.naoki676@mail.kyutech.jp
*Corresponding author : Tohru KAMIYA
（Received on June 23, 2021. In final form on September 14, 2021.）

Abstract : Due to the respiratory diseases such as chronic obstructive pulmonary disease and lower respiratory tract infections nearly 8 million people were died worldwide each year. Reducing the number of deaths from respiratory diseases is a challenge to be solved worldwide. Early detection is the most efficient way to reduce the number of deaths in respiratory illness. As a result, the spread of infection can be suppressed, and the therapeutic effect can be enhanced. Currently, auscultation is performed as a promising method for early detection of respiratory diseases. Auscultation can estimate respiratory diseases by distinguishing abnormal sounds contained in respiratory sounds. However, medical staff need to be trained to perform auscultation with high accuracy. Also, the diagnostic results depend on each staff subjectively, which can lead to inconsistent results. Therefore, in some environments, a shortage of specialized health care workers can lead to the spread of respiratory illness. To solve this problem, an application that analyzes respiratory sounds and outputs diagnostic results is needed. In this paper, we use a newly proposed deep learning model to automatically classify the respiratory sound data from the ICBHI 2017 Challenge Dataset. Short-Time Fourier Transform, Constant-Q Transform, and Continuous Wavelet Transform are applied to the respiratory sound data to convert it into the time-frequency region. Then, the obtained three types of breath sound images are input to CRNN (Convolutional Recurrent Neural Network) having scSE (Spatial and Channel Squeeze & Excitation) Block. The accuracy is improved by weighting the features of each image. As a result, AUC (Area Under the Curve) : (Normal : 0.87, Crackle : 0.88, Wheeze : 0.92, Both : 0.89), Sensitivity : 0.67, Specificity : 0.82, Average Score : 0.75, Harmonic Score : 0.74, Accuracy : 0.75 were obtained.
Keywords : Respiratory Sounds Classification, Computer Aided Diagnosis, Time-Frequency Analysis, Deep Learning

1. 序　論

慢性閉塞性肺疾患，下気道感染症などの呼吸器疾患により，毎年世界中で 800 万人近い人々が死亡している[1]. 呼吸器系の病気は，世界的にも無視することができない病気の一種で，死亡者数削減が重要な課題となっている．呼吸器疾患による死亡者数削減の最も効率的な方法は早期発見であり，それにより感染の拡大を抑えるだけでなく，治療効果を高めることが可能である．

現在呼吸器疾患の早期発見法として，聴診が行われている．罹患者の呼吸音には，気管，気管支，肺などの異常による異常音（副雑音）が含まれることが多い．例えば，主に肺の疾患により断続性ラ音（Crackle）が発生し，気管・気管支の疾患により，連続性ラ音（Wheeze）が発生することが多い．つまり，聴診により，呼吸音に含まれる異常音を聞き分けることにより，疾患を推定することができる．そのため，聴診は簡便で安全・安価な手法として，世界中で広く採用されている[2].

しかし，精度の高い診断を行うためには，長い訓練期間が必要で，聴診者の主観的診断による，診断結果のばらつきの問題がある．さらに，聴診は，画像診断とは異なり，呼吸音を聞いて診断を行うため，診断結果を可視化することが困難であるという問題も存在する．以上のような背景から，近年では，呼吸音を解析して診断結果を出力するアプリケーションの開発が求められている．

先行研究[3, 4]では，書籍[2]の CD に収録された理想的な呼吸音データを対象に，深層学習アプローチによる自動分類を試みている．しかし，将来的な診断支援システムの実現のためには，雑音を多く含む呼吸音が収録されている公開データベースによる，より一般性のある識別法が望まれる．現在，呼吸音解析の領域では，ICBHI（International Conference on Biomedical and Health Informatics）2017

Challenge Respiratory Sound Database を用いた研究が行われている．これは，現在入手可能な最大の呼吸音データセットであり，様々な録音機器や環境下で収録されている[5]．

これまでに，このデータベースを用いた関連研究として，CNN（Convolutional Neural Network）を用いた手法がいくつか提案されている．呼吸音データを時間周波数領域に変換することにより，二次元画像を生成する．そして，得られた呼吸音変換画像に対し，CNN を用いることで特徴を自動抽出し分類を行う．CNN を用いた手法が，従来の機械学習を用いた手法より高い精度を達成することが確認されており，注目を浴びている．

具体的な手法としては，CNN と機械学習モデルを組み合わせて分類する手法[6]，改良を施した CNN による分類手法[7]，CNN により得られた特徴に対し，近年着目を浴びている Self-Attention 機構を組み合わせた手法[8]，そして HPSS（Harmonic/Percussive Sound Separation）により，異常呼吸音の特徴を強調後，CNN を用いることで分類する手法が提案されている[9]．これらの手法は，短時間フーリエ変換によるスペクトログラムや，MFCC（Mel Frequency Cepstral Coefficients）[10]を CNN の入力として用い，複数特徴量を CNN に同時入力する手法も提案され，精度の向上が確認されている[11]．

しかし，CNN のみによる深層学習モデルでは，異常呼吸音が持つ時系列特徴（持続音や突発音）の考慮が困難である点や，複数特徴量を同時入力する手法では，分類結果が一部の特徴量に影響をうけ，精度低下を引き起こすという問題が存在する．

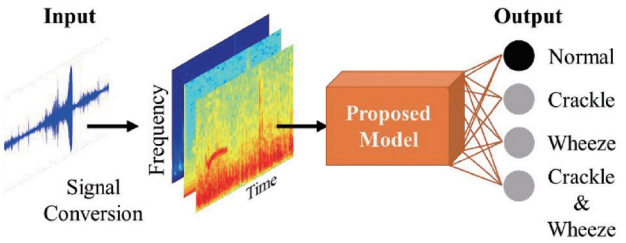


Fig.1　Outline of proposed method

そこで本論文では，複数の時間－周波数変換を用いて生成した呼吸音変換画像を，scSE（Spatial and Channel Squeeze & Excitation）構造をもつ CRNN（Convolutional Recurrent Neural Network）に同時入力することにより，上記問題の解決を図り，精度向上を目指す．

2. 方　法

本論文では，ICBHI 2017 Challenge Respiratory Sound Database に収録されている呼吸音データを対象とした分類手法を提案する．前処理を施した呼吸音データに対し，3 種類の時間－周波数変換を適用し，異なる 3 種類の呼吸音変換画像を生成する．そして，生成した画像を提案する深層学習モデルに同時入力することにより，4 クラス（Normal：正常呼吸音，Crackle：断続性ラ音，Wheeze：連続性ラ音，Both：Crackle &Wheeze）に分類する．提案手法の流れを Fig.1に示し，以下に詳細を示す．

2.1　前処理

ICBHI 2017 Challenge Respiratory Sound Database の呼吸音データは，複数の収録機器，サンプリングレート（44100, 10000，4000 Hz）で収録されている．そのため，4000 Hz にリサンプリングし，音量の正規化により，音量をある程度統一する．そして，呼吸サイクルごとに切り取り，ラベル付け（Normal，Crackle，Wheeze，Both）を行う．

2.2　信号変換

前処理を施した，呼吸音データに対し，3 種類の時間周波数変換を行う．異常呼吸音の特徴は，周波数成分の時間推移に現れるため，横軸を時間，縦軸を周波数とした画像に変換を行うことにより，呼吸音分類に有効な特徴を抽出することができる．また，異常呼吸音がもつ特徴を可視化することができるため，呼吸音の異常領域を視覚的に確認できる診断支援システムが実現でき，従来の聴診における問題の改善につながる．

以下に短時間フーリエ変換，Constant-Q 変換，連続ウェーブレット変換についてそれぞれ説明する．

2.2.1　短時間フーリエ変換

短時間フーリエ変換（STFT：Short-Time Fourier Transform）[3, 12, 13]は，信号の関心領域を，窓関数を用いることで局所的に分離し，周波数解析を行う変換である．窓関数 $\omega(t)$ は有限の時間長 L を持つ信号で，式（1）に示す．

$$\omega(t) = \begin{cases} 1, & (-\frac{L}{2} \le t \le \frac{L}{2}) \\ 0, & (otherwise) \end{cases} \tag{1}$$

解析対象の信号を切り出すため，窓関数を時間軸上に移動させながら信号 $f(t)$ との積をとる．切り出した信号を式（2）に示す．

$$\omega(t-\tau)f(t) \tag{2}$$

ここで，τ は窓関数の位置を表す．切り出された短時間区間信号に対し施すフーリエ変換を，短時間フーリエ変換と呼び，式（3）に示す．

$$F_\omega(\omega,\tau) = \int_{-\infty}^{+\infty} \omega(t-\tau)f(t)e^{-j\omega t}dt \tag{3}$$

そして，短時間フーリエ変換で得られる，パワースペクトルを式（4）に示す．

$$|F_\omega(\omega,\tau)|^2 = \left|\int_{-\infty}^{+\infty} \omega(t-\tau)f(t)e^{-j\omega t}dt\right|^2 \tag{4}$$

時間－周波数平面にパワースペクトルを 3 次元的に表示した画像をスペクトログラムという．本論文では窓関数 $\omega(t)$ として，式（5）に示すハミング窓を利用し，窓幅 L を 40 ms，時間位置 τ を 5 ms ずつずらしながら，短時間フーリエ変換を行い，スペクトログラムを生成する．

$$\omega(t) = \begin{cases} 0.54 + 0.46\cos\left(\frac{2\pi t}{L}\right), & (-\frac{L}{2} \le t \le \frac{L}{2}) \\ 0, & (otherwise) \end{cases} \tag{5}$$

2.2.2　Constant-Q 変換

短時間フーリエ変換は，窓関数の窓幅が固定であるため，高周波数では十分な周波数分解能であるにも関わらず，低周波数での分解能は不十分となる問題がある．Constant-Q 変換（CQT：Constant-Q Transform）[3, 14]は，短時間フーリエ変換と同様に窓関数を用いた変換であるが，各周波数帯域で窓幅を変化させることで上記問題を改善した変換法である．Constant-Q 変換を式（6）に示す．

$$X[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} W[k,n]x[n]\exp\left(-j\frac{2\pi Q}{N[k]}n\right) \tag{6}$$

ここで，$W[n]$は窓関数，$x[n]$は元信号，Q は定数である．また，$N[k]$は式(7)で算出される．

$$N[k] = \frac{f_s}{f_k} Q \tag{7}$$

f_s はサンプリング周波数(4000 Hz)，f_k は各周波数帯域における中心周波数である．そして，定数 Q は式(8)で算出される．

$$Q = (2^{\frac{1}{b}} - 1)^{-1} \tag{8}$$

ここで，b は 1 オクターブの分割数で，本論文では，1 オクターブを 48 分割で計算している．窓関数には式(9)に示すハニング窓を採用し，式(6)の Constant-Q 変換により得られた結果から，短時間フーリエ変換と同様にスペクトログラムを生成する．

$$\omega(t) = \begin{cases} 0.5 + 0.5\cos\left(\frac{2\pi t}{L}\right), & \left(-\frac{L}{2} \le t \le \frac{L}{2}\right) \\ 0, & (otherwise) \end{cases} \tag{9}$$

2.2.3 連続ウェーブレット変換

連続ウェーブレット変換(CWT: Continuous Wavelet Transform)[3, 13, 15]は，Constant-Q 変換同様，短時間フーリエ変換の分解能問題を改善した変換手法である．しかし，窓関数により信号を切り出し，フーリエ変換を行う短時間フーリエ変換，Constant-Q 変換とは異なり，連続ウェーブレット変換は，ウェーブレット関数を拡大縮小することで，対象とする信号の時間周波数解析を行う点が特徴である．

連続ウェーブレット変換は，ウェーブレット関数 $\psi(t)$ と，対象とする信号との内積で式(10)のように定義される．

$$W_\psi(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} f(t)\psi\left(\frac{t-b}{a}\right)^* dt \tag{10}$$

ここで，a はスケールの比率でウェーブレット関数の拡がりを示し，b は時間位置を示す．式(10)は解析対象の信号とスケール変換したウェーブレット関数との相関を表す．

また，式(10)の 2 乗関数は，時間－スケール平面のエネルギー分布を示し，得られる二次元平面をスカログラムという．

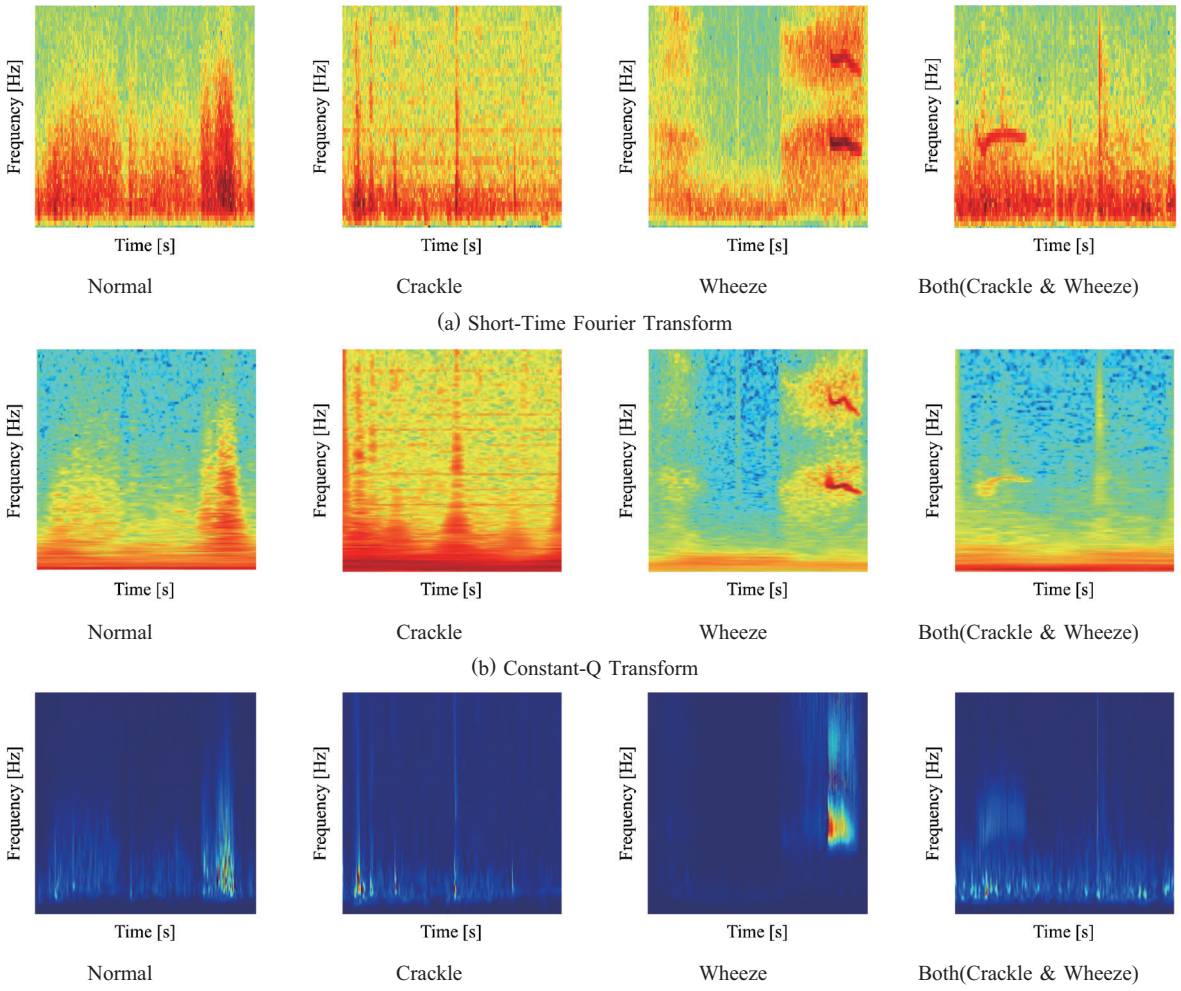
本論文では，ウェーブレット関数 $\psi(t)$ に，式(11)に示す，ガボールウェーブレットを用いてスカログラムの生成を行う．

$$\psi(t) = \frac{1}{2\sqrt{\pi}\sigma} e^{-\frac{t^2}{\sigma^2}} e^{-it} \tag{11}$$

以上の各変換法により得られた変換画像の例を，それぞれ Fig.2に示す．

2.3 識別器による分類

本論文では，改良を施した CRNN に scSE Block を組み込み，生成した 3 種類の呼吸音変換画像を同時入力することにより，呼吸音の自動分類を行う．以下に提案手法の詳細を示す．



(a) Short-Time Fourier Transform

(b) Constant-Q Transform

(c) Continuous Wavelet Transform

Fig.2 Images generated by each conversion method

2.3.1 CRNN

CRNN は CNN と RNN(Recurrent Neural Network)を組み合わせた深層学習モデルである．CNN により画像特徴量を抽出し，各時刻における特徴量の変化を，RNN を用いることで表現することができる．このモデルにより，CNN のみでは困難であった異常呼吸音がもつ時系列特徴を考慮することができる．実際に，楽曲分類の分野[16]や，鳥の鳴き声検出[17]の分野で利用され，高い精度を達成している．

そこで本論文では，呼吸音の分類精度向上を図るため，CRNN を基本構造として採用する．ただし，さらなる精度向上のため，RNN モデルで採用した LSTM(Long Short-Term Memory)を，双方向の系列を学習できる Bi-LSTM (Bidirectional LSTM)[18]に変更する．

LSTM は，時間経過とともにその状態を維持するメモリセルに対し，入力ゲート，忘却ゲート，出力ゲートを施し情報の出入りを制御することで，長期の系列学習を可能とした RNN モデルの一種である[19]．また，Bi-LSTM は順方向 LSTM の出力と逆方向 LSTM の出力を結合することで，文脈の前後の特徴を考慮した学習を可能とするモデルである．

異常呼吸音には，突発異常音と，持続異常音が存在するため，異常音発生時の前後の特徴を考慮することにより，それらの分類精度の向上を図る．本論文で提案する CRNN モデルを Fig.3，Table 1 に示す．

また，本提案モデルと類似したモデルに Bi-CLSTM (Bidirectional-Convolutional LSTM)が提案されている[20]．このモデルは，動画像処理の分野で主に扱われる CLSTM (Convolutional LSTM)[21]を，時間軸方向に対し双方向に学習するモデルである．LSTM は通常 1 次元の特徴量情報を入力として扱うが，CLSTM は LSTM に畳み込み構造を持たせることで，入力する特徴量情報を 2 次元に拡張したモデルである．そのため，本提案モデルの，CNN により得られた 1 次元の特徴量を Bi-LSTM に入力し，時系列特徴の学習を行う点が Bi-CLSTM と異なる．

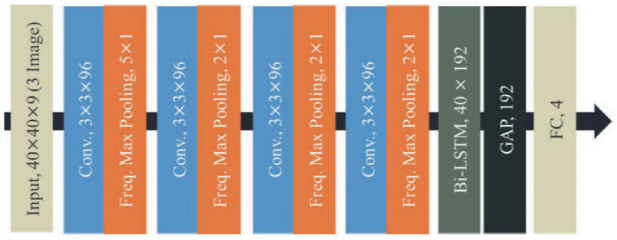


Fig.3 Architecture of improved CRNN
Freq. Max Pooling : Frequency Max Pooling,
GAP : Global Average Pooling

Table 1 Details of improved CRNN
BN : Batch Normalization, ReLU : Rectified Linear Unit

Layer	Size/Stride	Output	Activation
Input	-/-	40×40×9	-
Conv.1	3×3/1	40×40×96	BN, ReLU
Max Pooling	5×1/1	8×40×96	-
Conv.2	3×3/1	8×40×96	BN, ReLU
Max Pooling	2×1/1	4×40×96	-
Conv.3	3×3/1	4×40×96	BN, ReLU
Max Pooling	2×1/1	2×40×96	-
Conv.4	3×3/1	2×40×96	BN, ReLU
Max Pooling	2×1/1	1×40×96	-
Reshape	-/-	96×40	-
Bi-LSTM	-/-	192×40	BN
GAP 1D	-/-	192	-
FC	-/-	4	Softmax

2.3.2 scSE Block

本提案手法では，3 種類の呼吸音変換画像を同時入力することで分類精度の向上を図る．しかし，従来手法において複数画像を同時入力することで，分類結果が一部の特徴量に影響を及ぼされ，精度低下を引き起こす問題が確認された[3, 11]．この問題を改善するため，本論文では，Fig.3 に示す改良型 CRNN に，scSE Block[22]を組み込んだ新規深層学習モデル(Fig.4)を用いて，呼吸音の分類を行う．

scSE Block は，画像の特定空間に注目する sSE(Channel Squeeze and Spatial Excitation)(Fig.5の上部)と，CNN で得られた特徴マップの中，特定の特徴に注目する cSE(Spatial Squeeze and Channel Excitation)(Fig.5の下部)を組み合わせたモジュールである．CRNN に scSE Block を組み込み，画素間，チャンネル間に Attention をとることで，空間的にもチャンネル的にも関連性があり，より意味のある特徴マップを学習することが可能となる．

その結果，各画像の重み付けや，Crackle と Wheeze が混在するクラスでも，異常音の空間的位置を特徴付けることが可能であると考え，採用した．詳細な sSE Block と cSE Block のモデル構造を Table 2，Table 3 に示す．

Fig.4 Incorporation of scSE block

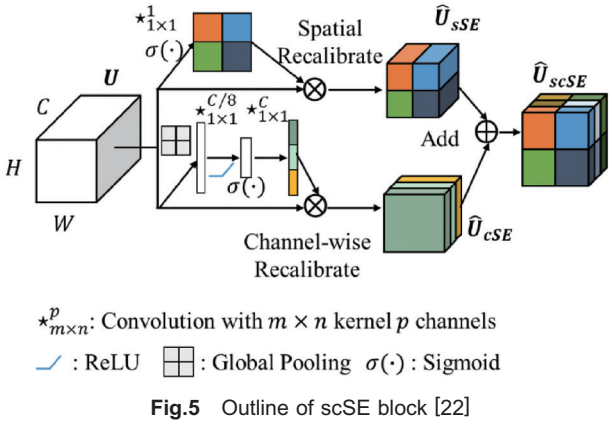


Fig.5 Outline of scSE block [22]

Table 2 Details of sSE block

Layer	Size/Stride	Output	Activation
Input	-/-	40×40×96	-
Conv.1	1×1/1	40×40×1	Sigmoid
Multiply	-/-	40×40×96	-

Table 3 Details of cSE block

Layer	Size/Stride	Output	Activation
Input	-/-	40×40×96	-
GAP 2D	-/-	96	-
FC1	-/-	12	ReLU
FC2	-/-	96	Sigmoid
Multiply	-/-	40×40×96	-

2.3.3 モデルの学習

モデルの最適化には，Keras[23]で実装されている Adam（Adaptive moment estimation）を用いて行う．

また，ICBHI 2017 Challenge Respiratory Sound Database に収録されている呼吸音データは，各クラスでデータ数に偏りがある．そのため，損失関数を式(12)に示す weighted categorical cross-entropy loss[24]を利用する．

$$loss = -\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M} w_k \times y_m^k \times \log(h_\theta(x_m, k)) \qquad (12)$$

ここで，M は学習データ数，K は分類クラス数，w_k はクラス k における重み，y_m^k はクラス k に属する学習データ m の正解ラベル，x_m は入力データ，h_θ はネットワークモデルを表す．この損失関数は，分類クラスに対し重みを付けることで，不均衡データによる精度の低下を防ぐことが可能である．各クラスの重みは，それぞれ Normal を 0.2, Crackle を 0.5，Wheeze を 1.0，Both（Crackle & Wheeze）を 1.0 と各データ数を考慮して経験的に決定した．

最後に，ハイパーパラメータは，バッチサイズを 128，最大エポック数を 100，学習率を 0.005 として学習を行った．

3. 実験と結果

3.1 呼吸音データの詳細

本論文では，ICBHI 2017 Challenge Respiratory Sound Database を対象とした実験を行う．このデータセットには，126 人の患者から，電子聴診器を用いて聴取した 920 個の音声ファイルで構成される．音声ファイルは，呼吸サイクルごとに区切られ，Normal，Crackle，Wheeze，Both（Crackle & Wheeze）のラベルが付与されている．各クラスの呼吸サイクル数を Table 4 に示す．

Table 4 Number of data points for each respiratory sound

Class	Number
Normal	3642
Crackle	1864
Wheeze	886
Both（Crackle & Wheeze）	506
Total	6898

3.2 評価方法

本論文では，5 分割交差検証による性能評価を行う．各クラスにおける，ROC（Receiver Operating Characteristic）曲線に基づく AUC（Area Under the Curve）および，ICBHI 2017 Challenge Respiratory Sound Database の評価指標に基づく，AS（Average Score）と HS（Harmonic Score），そして全体正解率（Accuracy）により評価する．AUC は分類する 4 つのクラスのうち，対象とする 1 つのクラスを陽性，その他のクラスを陰性とし，各クラス繰り返すことで算出する．また，AS は SE（Sensitivity）と SP（Specificity）の平均，HS は SE と SP の調和平均を表す．それぞれの指標は Table 5 に示す混同行列に対し，以下の式で算出する．

$$\mathrm{SE} = \frac{C_c + W_w + B_b}{C + W + B} \qquad (13)$$

$$\mathrm{SP} = \frac{N_n}{N} \qquad (14)$$

$$\mathrm{AS} = \frac{\mathrm{SE} + \mathrm{SP}}{2} \qquad (15)$$

$$\mathrm{HS} = \frac{2 \times \mathrm{SE} \times \mathrm{SP}}{\mathrm{SE} + \mathrm{SP}} \qquad (16)$$

ここで，C，W，B，N は各ラベルのデータ総数，C_c，W_w，B_b，N_n などの下付き文字は予測ラベルを表す．例えば N_c は予測ラベルが Crackle で，正解ラベルが Normal であったデータ数を表す．また，提案モデルと他のモデルの Accuracy に関し，2 つのモデルの Accuracy が等しいという帰無仮説のもと，有意水準を 0.05 に設定した両側 t 検定を行う．

3.3 実験結果

Table 6 に AUC による性能評価，Table 7 に ICBHI 2017 Challenge Respiratory Sound Database の性能評価を用いた結果と，両側 t 検定における p 値，そして Table 8 に scSE

Table 5 Confusion matrix

		Prediction label			
		Crackle	Wheeze	Both	Normal
True label	Crackle	C_c	C_w	C_b	C_n
	Wheeze	W_c	W_w	W_b	W_n
	Both	B_c	B_w	B_b	B_n
	Normal	N_c	N_w	N_b	N_n

Table 6 Comparison of AUC results

Model	AUC			
	Normal	Crackle	Wheeze	Both
STFT + Original CRNN	0.83	0.84	0.88	0.84
CQT + Original CRNN	0.79	0.79	0.84	0.82
CWT + Original CRNN	0.79	0.82	0.83	0.81
STFT&CQT + Original CRNN	0.83	0.83	0.87	0.84
STFT&CWT + Original CRNN	0.84	0.84	0.88	0.85
CQT&CWT + Original CRNN	0.83	0.83	0.86	0.83
3 Images + Original CRNN	0.84	0.84	0.88	0.86
3 Images + CRNN（Bi-LSTM）	0.87	0.88	0.90	0.89
Proposed model : 3 Images + scSE-CRNN（Bi-LSTM）	0.87	0.88	0.92	0.89

Table 7 Comparison of results by performance evaluation of ICBHI 2017 Challenge Respiratory Sound Database

Model	SE	SP	AS	HS	Accuracy	p 値
STFT + Original CRNN	0.57	0.81	0.69	0.67	0.69	0
CQT + Original CRNN	0.52	0.78	0.65	0.62	0.65	0
CWT + Original CRNN	0.58	0.71	0.64	0.64	0.65	0
STFT&CQT + Original CRNN	0.58	0.80	0.69	0.67	0.70	0
STFT&CWT + Original CRNN	0.60	0.78	0.69	0.68	0.69	0
CQT&CWT + Original CRNN	0.57	0.78	0.68	0.66	0.68	0
3 Images + Original CRNN	0.60	0.82	0.71	0.69	0.71	0
3 Images + CRNN（Bi-LSTM）	0.64	0.83	0.74	0.72	0.74	0.04
Proposed model : 3 Images + scSE-CRNN（Bi-LSTM）	0.67	0.82	0.75	0.74	0.75	-
3 Images + VGG 16&LSTM [3]	0.54	0.79	0.67	0.64	0.67	-
STFT&CWT + VGG 16 [4]	0.53	0.77	0.65	0.63	0.66	-
CNN + SVM [6]	0.51	0.78	0.65	0.62	0.66	-
CNN + LDA-RSE [7]	0.58	0.83	0.71	0.68	0.71	-
CNN + Attention [8]	-	-	-	-	0.67	-
HPSS + VGG 16&SVM[9]	0.51	0.83	0.67	0.63	-	-
Multi Image + CNN [11]	0.47	0.67	0.57	0.55	-	-
RNN [25]	0.58	0.73	0.66	0.65	-	-

Block 組み込み前後における，混同行列の比較結果を示す．提案手法の性能評価には，短時間フーリエ変換によるスペクトログラム，Constant-Q 変換によるスペクトログラム，連続ウェーブレット変換によるスカログラムをそれぞれ単体でオリジナルの CRNN に入力した場合，3 種類画像の内，2 種類を同時にオリジナルの CRNN に同時入力した場合，3 種類の画像をオリジナルの CRNN に同時入力した場合，3 種類の画像を改良型 CRNN（scSE Block 無し）に同時入力した場合，そして 3 種類の画像を改良型 CRNN に scSE Block を組み込んだ提案モデルに入力した場合（提案手法）との比較を行う．

Table 6，Table 7 より，提案手法において，AUC は Normal : 0.87，Crackle : 0.88，Wheeze : 0.92，Both : 0.89，そして，AS : 0.75，HS : 0.74，Accuracy : 0.75 の精度を得た．また，他の全ての手法に対し，p 値が有意水準 0.05 を下回っており，本提案手法の有意性が確認された．

4. 考　察

Table 6，Table 7 より各呼吸音変換画像 1 種類を単体で識別器に入力する場合，3 種類の画像の内，2 種類の画像を識別器に同時入力する場合より，3 種類の画像を同時入力する手法が，各評価指標において精度の向上を達成した．また，先行手法[3, 4]を ICBHI 2017 Challenge Respiratory Sound Database に適用した結果との比較により，VGG 16 より本論文で採用した CNN 構造が，呼吸音変換画像からの特徴抽出に有効であることが確認された．これは，オリジナルの CRNN における CNN 構造に対し，VGG 16 は約 2 倍の層数であることから，識別器の学習において過学習が生じたことが原因として考えられる．

次に，CRNN における LSTM を双方向の Bi-LSTM に変更した改良型 CRNN（scSE Block 無し）による分類結果について考察を行う．Bi-LSTM に変更前の結果と比較し，各評価指標において精度が上回った．感度（SE）に 4% の精度向上を示したことから，Bi-LSTM による，異常呼吸音発生前後の時系列特徴を考慮した分類の有効性が確認された．

最後に，本論文提案手法である，scSE-CRNN による分類結果に，他の比較手法との両側 t 検定による統計的有意差を確認した．また，他論文の各 CNN を用いた手法や，RNN を単体で用いた手法と比較し，SE，AS，HS，Accuracy において最も高いスコアを得ることができた．さらに，Table 8 の混同行列を比較すると，scSE Block の組み込みにより，チャンネル間，画素間に重みをつけることで，すべての異常呼吸音クラスにおける分類性能が向上した．つまり，複数画像を識別器に同時入力する際，scSE Block の組み込みが有効であることが確認された．また，Both クラスの分類性能の向上から，scSE Block を用いた Crackle と Wheeze の空間的位置の特徴付けが，異常呼吸音が混在するデータの識別精度向上に寄与することが示された．

本論文で用いた ICBHI 2017 Challenge Respiratory Sound Database に収録されている呼吸音データには，Fig.6に示すような，話声等のノイズを含むデータが多く存在する．例えば，話声は時間軸方向に持続する音となるため，Wheeze の特徴に近い画像が生成される．提案手法による誤分類で特に多かったのが，これらのノイズを含む呼吸音データであり，さらなる精度向上のため，ノイズに頑健な識別器の構築が課題として挙げられる．

解決策として，話声を含むデータに対する，話声と呼吸音の分離が考えられる．音声データのノイズ除去として，デノイジングオートエンコーダを用いた研究[26]が行われ

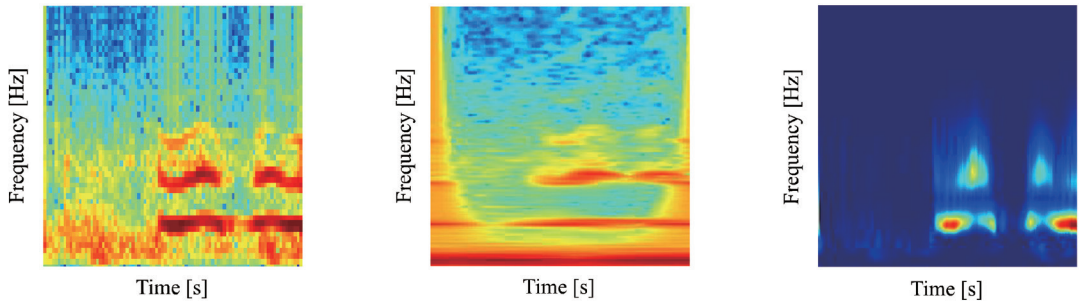
Table 8 Confusion matrix with and without scSE Block

(a) 3 Images + scSE-CRNN（Bi-LSTM）

		Prediction label			
		Crackle	Wheeze	Both	Normal
True label	Crackle	1351	24	40	449
	Wheeze	44	558	92	192
	Both	90	100	265	51
	Normal	492	130	41	2979

(b) 3 Images + CRNN（Bi-LSTM）

		Prediction label			
		Crackle	Wheeze	Both	Normal
True label	Crackle	1315	32	47	470
	Wheeze	47	541	88	210
	Both	90	127	228	61
	Normal	442	146	40	3014



(a) Short-Time Fourier Transform　(b) Constant-Q Transform　(c) Continuous Wavelet Transform

Fig.6 Normal respiratory sounds including speech

ており，本研究においても，意図的に付与した話声を除去するよう，オートエンコーダを学習することで，呼吸音データから話声を除去できると考えられる．

また，呼吸音に，話声等のノイズを付与するデータの水増しを行うことで，テスト時のノイズの影響を低減する，ロバストな識別器の構築が可能になると考える．そのため，今後は上記手法を本研究に導入し，精度の比較検討を行う予定である．

5. 結　論

本論文では，臨床現場において採取された，大規模な呼吸音データセット[5]に収録されている呼吸音を，Normal，Crackle，Wheeze，Both の 4 クラスに自動で分類する新しい深層学習手法を提案した．短時間フーリエ変換，Constant-Q 変換，連続ウェーブレット変換を用いて 3 種類の画像を生成し，scSE Block をもつ改良型 CRNN を用いることにより，呼吸音の分類を行った結果，AUC（Normal : 0.87，Crackle : 0.88，Wheeze : 0.92，Both : 0.89），Sensitivity : 0.67，Specificity : 0.82，Average Score : 0.75，Harmonic Score : 0.74，Accuracy : 0.75 を得た．他手法と比較し，AUC，Sensitivity，Average Score，Harmonic Score，Accuracy において最も高いスコアを得ることができ，ICBHI 2017 Challenge Respiratory Sound Database における本提案手法の優位性が示された．

さらなる分類精度の向上のため，データセットに含まれる，話声などのノイズに対する処理を検討する必要があり，これらは今後の課題である．

謝　辞

本研究は，文部科学省科学研究費補助金（21 H 03840）の補助を受けている．

本研究では，ICBHI 2017 Challenge Dataset（https : // bhichallenge.med.auth.gr/ICBHI_2017_Challenge）を利用している．

文　献

[1] World Health Organization, The top 10 causes of death : https : //www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death.

[2] 川城丈夫，阿部直，菊池功次他, CD による聴診トレーニング呼吸音編改訂第 2 版，南江堂，pp.1-6, pp.25-64, 2011.

[3] N. Asatani, T. Kamiya, S. Mabu et al., “Automatic Classification of Respiratory Sounds Considering Time Series Information Based on VGG 16 with LSTM”, 20th International Conference on Control, Automation and Systems, pp. 423-426, 2020.

[4] 南弘毅，陸慧敏，金亨燮他，“時間－周波数解析と畳み込みニューラルネットワークを用いた呼吸音の自動分類”，Medical Imaging Technology, Vol. 38 No. 1, pp. 40-47, 2020.

[5] B. M. Rocha, D. Filos, L. Mendes et al., “An Open Access Database for the Evaluation of Respiratory Sound Classification Algorithms”, Physiological Measurement, Vol. 40, No. 3, 035001, 2019.

[6] F. Demir, A. Sengur, V. Bajaj, “Convolutional Neural Networks Based on Efficient Approach for Classification of Lung Diseases”, Health Information Science and Systems, Vol. 8, No. 1, pp. 1-8, 2020.

[7] F. Demir, A. M. Ismael, A. Sengur, “Classification of Lung Sounds with CNN Model Using Parallel Pooling Structure”, IEEE Access, Vol. 8, pp. 105376-105383, 2020

[8] C. Li, H. Du, B. Zhu, “Classification of Lung Sounds Using CNN-Attention”, EasyChair Preprint no. 4356, 2020.

[9] 丸橋優生，浅谷尚希，陸慧敏他，“HPSS を用いた呼吸音の自動分類”，医用画像情報学会雑誌，Vol. 38, No. 2, pp. 95-100, 2021.

[10] H. Purwins, B. Li, T.Virtanen et al., “Deep Learning for Audio Signal Processing”, Journal of Selected Topics of Signal Proccesing, Vol. 13, No. 2, pp. 206-219, 2019.

[11] K. Minami H. Lu, T. Kamiya et al., “Automatic Classification of Respiratory Sounds Based on Convolutional Neural Network with Multi Images”, 2020 5th International Conference on Biomedical Imaging Signal Processing, pp. 17-21, 2020.

[12] J. Allen, “Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform”, IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 25, No. 3, pp.235-238, 1977.

[13] 和田成夫，よくわかる信号処理－フーリエ解析からウェーブレット変換まで，森北出版，pp.14-84, 2009.

[14] J. Brown, “Calculation of a Constant Q Spectral Transform”, Journal of the Acoustical Society of America, Vol. 89, No. 1, pp. 425-434, 1991.

[15] O. Rioul, M. Vetterli, “Wavelets and Signal Processing”, IEEE Signal Processing Magazine, Vol. 8, No. 4, pp.14-38, 1991.

[16] K. Choi, G. Fazekas, M. Sandler et al., “Convolutional Recurrent Neural Network for Music Classification”,

IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2392-2396, 2017.
[17] E. Cakir, S. Adavanne, G. Parascandolo et al., “Convolutional Recurrent Neural Networks for Bird Audio Detection”, 25th European Signal Processing Conference, pp. 1744-1748, 2017.
[18] A. Graves, J. Schmidhuber, “Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures”, Neural Networks, Vol. 18, pp. 602-610, 2005.
[19] K. Greff, R. K. Srivastava, J. Koutnik et al., “LSTM : A Search Space Odessey”, IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, No. 10, pp. 2222-2232, 2017.
[20] Q. Liu, F. Zhou, R. Hang et al., “Bidirectional-Convolutional LSTM Based Spectral-Spatial Feature Learning for Hyperspectral Image Classification”, arXiv preprint arXiv : 1703.07910, 2017.
[21] 増田正人，中林靖，田村善昭，“Convolutional LSTM を用いた数値流体解析結果の予測”，日本計算工学会論文集, Vol. 2020, No. 1, p. 20201006 2020.
[22] A. G. Roy, N. Navab, C. Wachinger, “Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks”, arXiv preprint arXiv : 1803.02579, 2018.
[23] Keras Documentation : https : //keras.io/ja/.
[24] Y. Ho, S. Wookey, “The Real-World-Weight Cross-Entropy Loss Function : Modeling the Costs of Mislabeling”, arXiv preprint arXiv : 2001.00570, 2020.
[25] K. Kochetov, E. Putin, M. Balashov et al., “Noise Masking Recurrent Neural Network for Respiratory Sound Classification”, International Conference on Artificial Neural Networks, pp. 208-217, 2018.
[26] C. Yu, R. E. Zezario, S. S. Wang et al., “Speech Enhancement Based on Denoising Autoencoder with Multi-branched Encoders”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 28, pp. 2756-2769, 2020.