A DOMAIN TRANSFER BASED DATA AUGMENTATION METHOD FOR AUTOMATED RESPIRATORY CLASSIFICATION

Zijie Wang *Zhao Wang**

Zhejiang University

ABSTRACT

Automated auscultation has become a hot topic in the medical field for diagnostic and predictive analytic. Automated auscultation aims to improve the classification of respiratory sounds recorded by electronic stethoscope. Researchers have paid great effort on developing intelligent auscultation methods to improve the effectiveness of hearing and assist clinicians, especially deep neural network techniques have been employed in recent years. The performance of deep neural network (DNN) based methods is highly data-dependent. Unfortunately, even the current world's largest publicly available respiratory sound dataset, ICBHI, has only 6898 respiratory cycles with a total length of only 5.5 hours, which become a bottleneck for further improvement of DNN models. Therefore, we propose a data augmentation method for respiratory sounds classification, where the input transformation and migration are implemented. In addition, the classical pipeline which is usually used in the computer vision area is also improved in this work. Experimental results show that the proposed data augmentation methods could improve separation performance than the baseline methods. Especially, the proposed data augmentation could be easily implemented in the existing automated auscultation approaches.

Index Terms— respiratory sound classification, splicing audio, mixup, audio data augmentation, ICBHI dataset

1. INTRODUCTION

Respiratory diseases like chronic obstructive pulmonary disease (COPD), Asthma, Pneumonia, Tuberculosis and Lung cancer are leading causes of death and disability worldwide. Sixty-five million people suffer from COPD with 3 million death each year, now making it the third leading cause of death worldwide [1]. Early diagnosis is critical in preventing the spread of respiratory disease and reducing its adverse effects. Auscultation on the chest using a stethoscope is the standard method for screening and diagnosis of respiratory disease. It provides a low-cost, non-invasive screening method that avoids the risk of radiographic.

Auscultation has its significant drawbacks: it requires a physician with specialized auscultation training to complete the diagnostic process while such subjective method always depends on the physician's auditory perception, their experience, and ability to differentiate lung sounds patterns. In the absence of specialized physicians, such as in remote or backward areas, or in disease pandemics (e.g. COVID-19), such auscultation would be almost impossible to accomplish. The automation of auscultation can effectively assist physicians with relatively low levels of expertise (e.g., rural or community physicians) and greatly increase diagnostic efficiency. If an automated auscultation system can achieve a high accuracy rate, it can even be self-testing, thus detecting the disease early, saving treatment costs and treatment resources, and greatly reducing the likelihood of its deterioration into a serious disease.

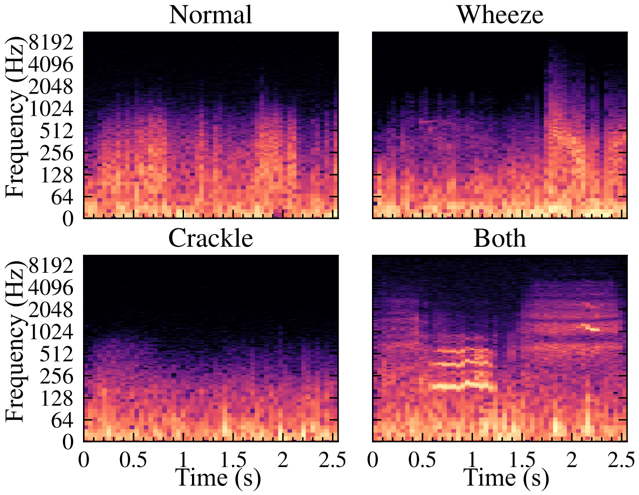

Fig. 1: Examples of different categories of respiratory cycles shown on the log-spectrogram

Since auscultation is the conclusion of the respiratory sounds listening within the body, the most important issue of designing an automatic auscultation system could be considered as the classification of respiratory sounds. Crackle and wheeze are identical and important among these sounds. Wheeze is considered a hallmark of COPD and asthma, manifesting as a continuous segment of high pitch, while a range of serious lung diseases can lead to crackle, which sounds like a discontinuous explosion. As shown in Figure 1, it can be seen that wheeze shows a clear bar trace in the spectrogram,

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

while crackle shows signs of fragmentation.

The task of respiratory sound classification was done early on by handcraft feature design and be replaced by popular deep learning methods in recent years. Besides, input features have consequently become MFCC, filterbank, spectrograms, Mel-spectrograms, and so on. Many CNN or RNN based methods have been conducted in recent years, as well as some hybrid models.

These off-the-shelf valid models or backbone models are widely used in many tasks via transfer learning framework with fine-tuning process. In such approaches, the performance of adopted task is highly related with the amount and distribution of training data used for fine-tuning.

The proposed work aims to systematically study the data augmentation techniques for the task of respiratory recognition while such medical related tasks always suffered from data usage and imbalance problems. The contributions of proposed work include:

1. Completed the transfer of a simple structure for computer vision, which achieves outstanding results on audio tasks through a simple structure.

2. Made a fine-grained design for audio data augmentation, which provide significant performance as well as interpretability.

2. PROPOSED MODEL

The proposed pipeline is straightforward: After diverse preprocessing of the audio, the spectrogram is extracted and fed to the model. The inputs are sequentially passed through backbone, pooling layer, and linear layers with dropout, and then passed through the classifier to obtain classification logits as the output of the model.

2.1. Data Utilization

Although ICBHI [2] is already the largest publicly available breath sound training dataset, it still only has a total of 6898 respiratory cycle samples, which is too small for training a larger DNN. Therefore, the major of our work is to utilize this audio data more efficiently, most notably data augmentation. The proposed work focus on splicing and circular padding while traditional methods including adding noise, time stretching and pitch shifting are also employed.

First of all, we take the benefits of traditional audio data augmentation are necessary, including adding noise, time stretching, pitch shifting, and so on. The size of the training set can be effectively expanded with such methods. In order to enhance the model's robustness to the noise occurs in the data, we have used strong noise augmentation in particular, including white noise and some pre-recorded environmental noise audio, and the signal to noise ratio (SNR) was randomly sampled between 1 and 10.

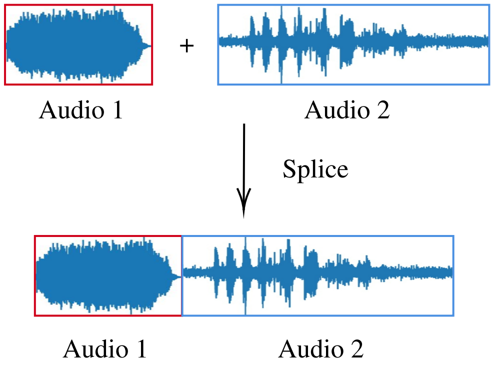

Fig. 2: Splice samples

We noticed that some of the respiratory cycle samples are very short (min 0.2 sec) and some of them are quite long (max 16 sec), while the average length of the respiratory cycle is 2.7 seconds. The model could meet the over-fitting problem due to a small number of special samples. To further improve the training data quality, we employ a zero-fill for such too-short audio (zero padding), truncate the too-long audio, and make all the audio uniform in length. There are two improvements to such a common operation.

Splicing Audio: Experiments found that such treatment is far from sufficient and the model is still easily over-fitted with too short samples. So we splice the shorter samples of the same category as shown in Fig. 2 . This method not only alleviates the over-fitting problem in too-short samples, but also makes the model focus more on details of the respiratory sound content rather than the recording activities, e.g., the process of the beginning and end of breathing. The method also allows the number of samples with different labels to be balanced.

Circular Padding: To make the model focus only on the details of the breathing sound, we discarded the zero padding and replaced it with circular padding. This padding is applied to all audio, which means that the test set is also affected.

With both of these approaches, we want the input audio to be no longer a single respiratory cycle, but a longer period of breathing, allowing the model to determine if there is a respiratory problem. While each sample in the test set has only one respiratory cycle, we also treat it as a partial segment of a long respiratory sound containing many cycles by means of circular padding. Experiments show that the above approach results in a remarkable improvement in the accuracy of the test set.

While the above approach allows a degree of greater focus on the details of the sound, the processes of the breathing cycle, such as the beginning and end, may still be evident. To address such issues, we further enhance the data by introducing the mixup [3]. The Mixup method achieves data enhancement by fusing the spectrograms and labels of different samples. Specifically, for each sample x_a with one-hot label

y_a, mixup will first randomly sample from the beta function to obtain λ. And randomly draw another sample x_b with one-hot label y_b from the dataset. The generated result $\tilde{x}$ and $\tilde{y}$ is shown in the following equation.

$$\begin{aligned}\tilde{x} &= \lambda x_i + (1-\lambda)x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j\end{aligned} \tag{1}$$

Mixup is widely used in the field of computer vision. Numerous experiments have demonstrated that mixup encourages the model to have a linear understanding of the samples, which can effectively reduce the interference of noisy samples and make the decision boundary smoother, thus significantly improving the generalization ability of the model. In our experiments, mixup greatly enhances the generalization of the model and boosts the test set accuracy noticeably.

We found that when training for unbalanced categories, the performance will be slightly improved by using weighted classification losses. Specifically, the loss weight of categories with a large number of samples is reduced, while the loss weight of categories with a small number of samples is increased.

2.2. NetWork Structure

We use ResNeSt [4] as backbone model, which is based from ResNet [5], ResNeXt [6], SK-Net [7] and SE-Net [8]. The ResNeSt structure introduces the split-attention module, rethinking the Multi-path mechanism of GoogleNet, referencing ResNeXt's group convolution in ResNet bottle, and borrowing from SE-Net's channel-attention mechanism by adaptive re-calibrating channel feature responses, and SK-net's feature-map attention by introducing two network branches. The IBN module proposed in IBN-Net [9] also has a significant improvement in the generalizability of the model in the audio domain, which can also be noticed in some audio-related tasks as [10].

We started with a pre-trained ResNeSt model based on ImageNet. Besides, it has noticed that the pre-trained model performs better compared to the random initialization, although the image domain is different from acoustic spectrograms. A spatial aggregation module is designed to compress the larger feature maps output by backbone into vectors. The two most common spatial aggregation modules are average pooling and maximum pooling. But both of these methods are very corrupt to the original spatial information. For this reason, we use Generalized Mean Pooling (GeMPool) [11] to replace the original spatial aggregation module. It learns the spatial properties by a learnable parameter p, and thus can utilize the spatial information to a greater extent. It is worth mentioning that it is equivalent to average pooling when $p = 1$ and to maximum pooling when $p \to \infty$. For the feature space $\mathcal{X} \in \mathbb{R}^{C \times F \times T}$, the formula of GeMPool is shown below:

$$\mathcal{F} = [f_1 \dots f_k \dots f_K^\top], f_k = \left(\frac{1}{|\mathcal{X}_k|} \sum_{x \in \mathcal{X}_k} x^p \right)^{\frac{1}{p}} \tag{2}$$

In many speech audio tasks, it is common practice to use dropout [12] to resist overfitting, which usually results in significant model improvement. [13] states that the dropout layer should appear after all batch-normalization layers, so we designed multiple linear layers for sorting and extracting information from the backbone output. And the dropout module is included in it. Our additional linear layer with dropout of the model is efficient on anti-over-fitting, which is a dropout-linear-ReLU-dropout-classifier structure. The overall model structure is (OutDim 2048) → GemPool(on frequency dimension) → GemPool(on time dimension) → BatchNorm → Dropout → Linear (2048 → 512) → ReLU → Dropout → Linear (512 → 512) → ReLU → Dropout → Linear (512 → 4).

3. EXPERIMENTAL RESULTS

3.1. Dataset

To evaluate the performance of our model, we conducted several experiments on the largest publicly available respiratory sound benchmark datasets from International Conference on Biomedical and Health Informatics (ICBHI) scientific challenge respiratory sound database by [2].

The ICBHI database is widely used in respiratory sound related works, which contains 920 respiratory recordings from 126 different patients. Totally 6898 respiratory cycles included in recordings are annotated as one of four classes: normal, crackles, wheezes, and both (crackles and wheezes), see Table 1 for details. The respiratory cycle duration ranges from 0.2s to 16s and the average cycle duration is 2.7s, and it has 5.5 hours in total. Fig. 1 illustrates how these categories differ in terms of spectrogram. Each recording was tagged with one of 7 different chest recording positions, and one of 4 different recording devices. And each recording has annotated plenty of respiratory cycles. As mentioned above, the database gives a wealth of additional information which is useful for us to analyze and model the problem. Besides, a considerable part of the recordings in the database is noisy, which makes the problem complex, challenging, and close to reality.

Some existing methods conducted on ICBHI are evaluated based on 5-fold [14] or random splitting methods [15]. In order to conduct a fair comparison, We evaluate our proposed data augmentation model with official ICBHI challenge criteria, where training and test sets are split in a 60:40 ratio and the training and test sets are divided comparably to completely different patients.

Due to the original paper of the database [2, 16], the performances of classifiers were evaluated using officially pro-

Table 1: Number of different categories of respiratory cycles

	Normal	Crackles	Wheezes	Both	Total
No.	3642	1600	1150	506	6898

Table 2: The ablation study table. Baseline is a ResNeSt-IBN based network, which is only has a backbone, average pooling, and a linear classifier layer. The improvements are aggregated in turn.

	Sp(%)	Se(%)	Sc(%)
Baseline	68.6	28.7	49.6
+ Circular padding	66.0	34.4	50.3
+ GeMPool	68.1	33.3	50.7
+ Dropout	67.8	37.1	52.2
+ Mixup	**71.9**	35.6	53.8
+ Splice	70.4	**40.2**	**55.3**

posed scores i.e. sensitivity (Se), specificity (Sp), and overall score (Sc), the specific formula is follows [2].

$$Se = \frac{C_c + C_w + C_b}{T_c + T_w + T_b}, Sp = \frac{C_n}{T_n}, Sc = \frac{Se + Sp}{2} \quad (3)$$

Where notation C_i, T_i $(i = n, c, w, b)$ indicates the number of correct classifications and the total number of classifications, respectively. And n, c, w, b denote normal, crackles, wheezes, and both (crackles and wheezes), respectively.

3.2. Evaluation settings

All the audios are re-sampled uniformly to 8000Hz and limited to about 8 sec. The STFT with 512 samples size of FFT, and 256 sample hop lengths is applied to the audios. The model is trained using Adam optimizer while learning rate is 0.0006, batch size is 64, and weight decay is 0.001. Warm-up strategy and cross-entropy loss are also used. All dropout rates are set to 0.5. The Mix-up α is set to 0.2. Label-smooth [17] ϵ is set to 0.01. Each category is assigned the same number of spliced samples. After the training with augmentation is finished, it is also necessary to fine-tune the model back to the original data using a smaller learning rate.

3.3. Results

To demonstrate the effectiveness of our individual modules and enhancement methods, an ablation study has been taken and the result is shown in Table 2. A ResNeSt-IBN based network is chosen as baseline, where microphone information is employed as ResNeSt's top metrics. Data from different microphones is itself somewhat skewed and is excluded from the comparison table. Except for this, some enhancement techniques are used, such as smart padding to detect whether the respiratory cycle is normal before and after each abnormal respiratory cycle.

Table 3: Performance on ICBHI dataset

Model	Sp(%)	Se(%)	Sc(%)
GRU	46.7	29.4	38.2
HMM [16]	-	-	39.5
GRU + augmentation	60.5	29.5	45.0
ARNN [18]	81.3	17.8	49.6
Baseline	68.6	28.7	49.6
LungBRN [19]	69.2	31.1	50.2
LungBRN+NL [20]	63.2	41.1	52.3
Respirenet-w/o tricks [21]	71.4	39.0	55.2
Respirenet-w/o RL [21]	71.8	39.6	**55.7**
Baseline + augmentation	70.4	**40.2**	55.3

To further demonstrate the effectiveness of proposed model, a simple single-layer 128-dimensional GRU model is conducted to show the performance of our data enhancement approach and modular design. It could see that the performance of the simple GRU model has been significantly improved while the baseline model is able to reach competitive performance with our proposed data augmentation method in Table 3. Besides, we found in our experiments that the addition of mixup makes the model much more difficult to converge compared to other augmentations. The performance on the validation set is also relatively more stable.

4. CONCLUSION AND FUTURE WORK

To sum up, respiration sound classification work suffered from data usage and imbalance problems. Although many audio data augmentation techniques have been developed, few work have been reported on such lung related tasks. We have conducted a systematically study on respiratory data augmentation to tackle this issue. Experimental results have shown that the proposed model could significantly improve classification performance. A further comprehensive study would be taken once the SOTA work's source code is published. The proposed data augmentation tool will be published online soon.

Acknowledgement

This work was supported by the National Key Research and Development Project of China (No.2021ZD0110500) and Ningbo Natural Science Foundation (2021J167).

5. REFERENCES

[1] Peter GJ Burney, Jaymini Patel, Roger Newson, Cosetta Minelli, and Mohsen Naghavi, "Global and regional trends in copd mortality, 1990–2010," *European Respiratory Journal*, vol. 45, no. 5, pp. 1239–1247, 2015.

[2] BM Rocha, Dimitris Filos, L Mendes, I Vogiatzis, E Perantoni, E Kaimakamis, P Natsiavas, Ana Oliveira, C Jácome, A Marques, et al., "A respiratory sound database for the development of automated classification," in *International Conference on Biomedical and Health Informatics*. Springer, 2017, pp. 33–37.

[3] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[4] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al., "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[7] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang, "Selective kernel networks," *CoRR*, vol. abs/1903.06586, 2019.

[8] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *CoRR*, vol. abs/1709.01507, 2017.

[9] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 464–479.

[10] Xingjian Du, Zhesong Yu, Bilei Zhu, Xiaoou Chen, and Zejun Ma, "Bytecover: Cover song identification via multi-loss training," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 551–555.

[11] Filip Radenović, Giorgos Tolias, and Ondřej Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[13] Xiang Li, Shuo Chen, Xiaolin Hu, and Jian Yang, "Understanding the disharmony between dropout and batch normalization by variance shift," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2682–2690.

[14] Kirill Kochetov, Evgeny Putin, Maksim Balashov, Andrey Filchenkov, and Anatoly Shalyto, "Noise masking recurrent neural network for respiratory sound classification," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 208–217.

[15] Truc Nguyen and Franz Pernkopf, "Lung sound classification using snapshot ensemble of convolutional neural networks," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 760–763.

[16] Nikša Jakovljević and Tatjana Lončar-Turukalo, "Hidden markov model based respiratory sound classification," in *International Conference on Biomedical and Health Informatics*. Springer, 2017, pp. 39–43.

[17] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton, "Regularizing neural networks by penalizing confident output distributions," *CoRR*, vol. abs/1701.06548, 2017.

[18] Zijiang Yang, Shuo Liu, Meishu Song, Emilia Parada-Cabaleiro, and Björn W Schuller, "Adventitious respiratory classification using attentive residual neural networks.," in *INTERSPEECH*, 2020, pp. 2912–2916.

[19] Yi Ma, Xinzi Xu, Qing Yu, Yuhang Zhang, Yongfu Li, Jian Zhao, and Guoxing Wang, "Lungbrn: A smart digital stethoscope for detecting respiratory disease using bi-resnet deep learning algorithm," in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2019, pp. 1–4.

[20] Yi Ma, Xinzi Xu, and Yongfu Li, "Lungrn+ nl: An improved adventitious lung sound classification using nonlocal block resnet neural network with mixup data augmentation.," in *Interspeech*, 2020, pp. 2902–2906.

[21] Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain, "Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting," *arXiv preprint arXiv:2011.00196*, 2020.