Breathing Sound Segmentation and Detection Using Transfer Learning Techniques on an Attention-Based Encoder-Decoder Architecture

Chiu-Han Hsiao[1], *Member*, *IEEE*, Ting-Wei Lin[2], Chii-Wann Lin[3], Fu-Shun Hsu[4], Frank Yeong-Sung Lin[2], Chung-Wei Chen[4], and Chi-Ming Chung[5]

***Abstract*— This paper focuses on the use of an attention-based encoder-decoder model for the task of breathing sound segmentation and detection. This study aims to accurately segment the inspiration and expiration of patients with pulmonary diseases using the proposed model. Spectrograms of the lung sound signals are used to train the model. The model would first encode the spectrogram and then detect inspiratory or expiratory sounds using the encoded image on an attention-based decoder. With the use of the attention mechanism, physicians would be able to make a more precise diagnosis based on the more interpretable outputs.**

***Clinical relevance*— An attention-based approach is proposed for the task of lung sound segmentation and detection. The results revealed that the most consistent and accurate outputs are achieved by attending at 0.1-second time segments. Future research on the segmentation of adventitious lung sounds based on our proposed model will be able to provide more precise diagnosis recommendations to physicians with more predictability and efficiency.**

I. INTRODUCTION

Respiratory diseases are among the deadliest diseases worldwide. According to the World Health Organization (WHO), chronic obstructive pulmonary disease (COPD) ranks third, and lower respiratory infection ranks fourth in terms of fatality. Some of the most common chronic respiratory diseases are COPD and asthma. COPD caused approximately 3.17 million deaths in 2015, which accounted for 5% of the deaths in that single year. Symptoms of asthma include reduced airflow into and out of the lungs due to the swelling of the bronchial tubes, which narrows the airway. Both these diseases cannot be cured and affect a high proportion of the global population. However, the symptoms can be relieved with appropriate management and the right treatments [1], [2].

Lung sound serves as a crucial indicator of the health and disease of the respiratory system. Physicians widely use auscultation to evaluate pulmonary conditions, which is usually performed using a stethoscope. However, an inexperienced technician may provide a wrong diagnosis. There are two kinds of lung sounds, normal breathing sounds or adventitious breathing sounds. Normal breathing sounds are lung sounds that do not indicate any disorder. An adventitious breathing sound usually indicates the presence of pulmonary disease and can be heard alongside normal breathing sounds. Two kinds of adventitious breathing sounds exist, namely continuous, which includes wheezes, and non-continuous, which provides for crackles [3].

Consequently, to relieve patients of the inconvenience caused by their symptoms, new methods must be developed for identifying respiratory diseases. Incorporating computer analysis and electronic auscultation in the study of lung sounds ensures high accuracy and timely diagnosis. It eliminates the subjectivity of the listener and identifies pathological features that physicians cannot identify. With computer analysis, physicians can provide accurate diagnosis and initiate treatment early, relieving the discomfort of their patients.

Currently, several studies have attempted to incorporate computer algorithms using convolutional neural networks (CNNs) for detecting adventitious lung sounds [3], [4], [5]. However, most of these works focused only on lung sound classification instead of segmentation of the sound signal. This study proposed an attention-based encoder-decoder architecture audio segmentation neural network shown in Fig. 1 to enhance the quality

[1] Chiu-Han Hsiao is with the Research Center for Information Technology Innovation, Academia Sinica, Taiwan, 11529, chiuhanhsiao@citi.sinica.edu.tw.

[2] Ting-Wei Lin is with the Information Management Department, National Taiwan University, Taipei, Taiwan, 10617, r08725007@ntu.edu.tw.

[2] Frank Yeong-Sung Lin is with the Information Management Department, National Taiwan University, Taipei, Taiwan, 10617, yeongsunglin@gmail.com.

[3] Chii-Wann Lin is with the Department of Biomedical Engineering, National Taiwan University, Taipei, Taiwan, 10617, cwlinx@ntu.edu.tw.

[4] Fu-Shun Hsu is with the Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, and Department of Critical Care Medicine, Far Eastern Memorial Hospital, New Taipei, Taiwan, 22060, fushun0607hsu@gmail.com.

[4] Chung-Wei Chen is with the Department of Critical Care Medicine, Far Eastern Memorial Hospital, New Taipei, Taiwan, 22060, femhcs@gmail.com.

[5] Chi-Ming Chung is with the Computer Science & Information Engineering Department, National Taiwan University, Taipei, Taiwan, 10617, b05902026@ntu.edu.tw.

of computer-assisted lung sound analysis, focusing on the segmentation of the inspiration phase and expiration phase of human breathing.

II. LITERATURE REVIEW

A. Inspiration/Expiration and Adventitious Lung Sound

Inspiratory–expiratory ($I:E$) ratio is the ratio of the duration of inspiration and expiration time shown in (1) [6]. The duration for each phase is dependent on this ratio and the overall respiratory rate (1) [7]. A typical breathing sound for most situations typically has an $I:E$ ratio of 1:2 for adults, whereas children tend to have an $I:E$ ratio of 1:3.

$$I:E \text{ ratio} = t_{inspiration} : t_{expiration} \tag{1}$$

Adventitious lung sounds refer to lung sounds that are heard in addition to normal breathing sounds. The first characteristic observed to classify adventitious lung sounds is whether it is continuous or non-continuous. Non-continuous sounds include crackle, which typically lasts for less than 25 ms. Wheezes and rhonchi are constant sounds and last for more than 250 ms. A relationship exists between adventitious lung sounds and the inspiration/expiration phase. For instance, crackles are generated due to small airways and short openings during inspiration. Thus, crackles are typically inspiratory. Wheezes and crackles occur during the expiratory phase or both inspiratory and expiratory phases. However, they are not observed in the inspiratory phase alone [8].

B. Neural Networks for Lung Sound Classification

Several studies have focused on the classification of lung sounds; they have used recordings of lung sounds for classification tasks [3], [4], [5].

These tasks [3], [4], [5] typically include the classification of lung sounds (normal sounds, crackles, wheezes, stridors, and squawks). Other classification tasks are aimed toward the detection of five diseases (asthma, bronchitis, pneumonia, pneumoconiosis, and COPD) and whether a patient is healthy [4]. These studies used the architecture of CNN. Due to the introduction of CNN, capabilities of computer vision tasks, including object detection [9], and object classification [10] have drastically improved. Moreover, CNN architecture has been used in several cases regarding the domain of medical sound detection. A study [11] presented a cough sound detection system on a wearable device. Heartbeat detection [12] and heart sound classification [13] have been applied in a CNN for their respective tasks.

C. Encoder-Decoder Architecture

Despite being a powerful machine learning model, deep neural networks (DNNs) can be applied only to problems that have a fixed length of inputs and targets. They cannot be used for tasks such as audio segmentation because most of these problems have arbitrary sequence lengths [14]. An encoder-decoder model takes input in a fixed form using an encoder and decodes it using a decoder. With this mechanism, encoder-decoder models can overcome challenges faced by DNNs by having different input and output lengths. Therefore, this kind of architecture has been popular for sequence-to-sequence tasks, which include image captioning, speech recognition [15], and machine translation [14]. As the aim was to apply audio segmentation to respiratory sounds, a typical sequence-to-sequence task, the encoder-decoder is a perfect model for this goal.

D. Attention Mechanism

The attention mechanism was introduced as a solution to the issue of recurrent neural network sequence-to-sequence models, where the standard version of the sequence-to-sequence model is unable to process long input sequences because only the latest hidden state is used as the context in the decoder [16]. The attention mechanism surmounts this challenge by mapping between the decoder output and the encoder hidden states. It provides the decoder with the capability to access the entire sequence and to focus on specific regions to produce the output. Thus, "attention" on the relevant area of the input sequence is formed.

E. The Attention Model for Image Captioning

In [17], an attention-based encoder-decoder automatic image captioning model was introduced. Exploiting the power of the attention mechanism, the model is able to output a sentence describing an image. Not only does it produce state-of-the-art results, but the attention mechanism also makes the output more interpretable by visualizing where the model focuses on at each timestamp.

III. MODEL AND DATA ACQUISITION

The proposed model can segment lung sounds by taking recordings of lung sounds as input and returning an analysis indicating the start time and end time of each inspiration or expiration.

A. Data Acquisition and Data Preprocessing

The data collection is approved by the review of the institutional review board at the Far East Memorial Hospital, IRB No. 107052-F. The patient informed consent was done by the approved inform consent form. The data was encoded to de-linkage patient identity. 22 Patients were enrolled in this study at the gender ratio of Male: Female 1.2: 1. Digital stethoscope (Littmann 3200, 3M corp: sampling frequency 2000Hz) or anti-noising microphone set (Accursound, Heroic Faith Medical Science Co., Ltd: sampling frequency 4000Hz) were used for acoustic data acquisition. The acoustic data are 15 seconds recordings in .wav format.

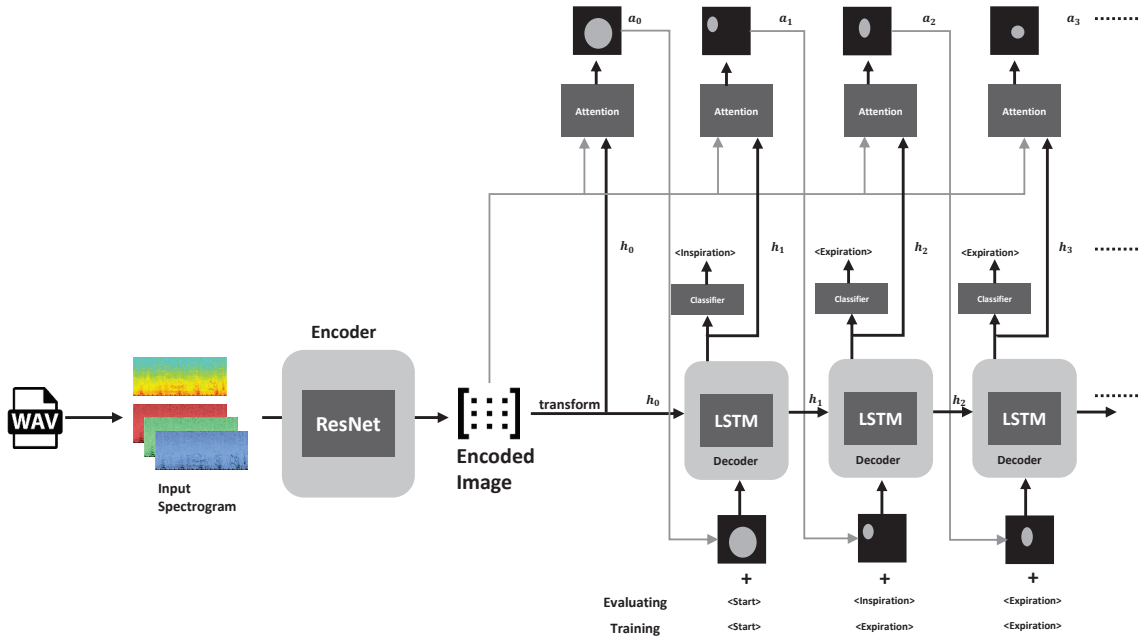

Fig. 1: System Architecture.

TABLE I: PERFORMANCE ON VARIOUS PARAMETER SETTINGS

Seconds per Segment	Encoder Model	Performance Metrics						
		Accuracy	Inspiration Recall	Expiration Recall	Inspiration Precision	Expiration Precision	Inspiration F1-Score	Expiration F1-Score
0.1	ResNet50	0.91908	0.90476	0.9322	0.90476	0.9322	0.90476	0.9322
0.1	ResNet101	0.91751	0.857143	0.9	0.85714	0.9	0.85714	0.9
0.1	ResNet152	0.91605	0.90476	0.9322	0.90476	0.9322	0.90476	0.9322
0.5	ResNet50	0.87638	0.46667	0.61538	0.875	0.88889	0.6087	0.72727
0.5	ResNet101	0.92006	0.28571	0.5	0.5	0.5	0.36364	0.5
0.5	ResNet152	0.90153	0.6	0.76923	0.81818	1	0.69231	0.86957
1	ResNet50	0.88123	1	1	0.16667	1	0.28571	1
1	ResNet101	0.9113	0.28571	0.85714	1	0.85714	0.44444	0.85714
1	ResNet152	0.91359	0.14286	1	0.33333	0.77778	0.2	0.875

Meanwhile, the research associate will record the pre-annotated adventitious sound and the demographics of the patient into the metadata chart. All of the acoustic data were then annotated by experienced respiratory therapists or medical doctors to indicate the period of inspiration, expiration, and adventitious sound at the resolution of a sub-millisecond range. The segmentation of inspiration and expiration episodes were modeled using the spectrogram of the acoustic signal and acoustic features extracted with ResNet encoder.

B. Model

Next, a neural network must be defined for lung sound audio segmentation; an attention-based encoder-decoder architecture (inspired by image captioning systems) was used. The encoder uses a ResNet, which converts our spectrogram into a fixed form. The decoder consists of a long short-term memory (LSTM) block for sequence analysis and an attention mechanism for creating a weighted image so that the model can focus on specific part of the spectrogram at each step of time.

The input spectrogram first passes through a ResNet model, which acts as an encoder, encoding the spectrogram into a fixed encoded image (E). A transformation must first be applied to the encoded image, outputting a hidden state of the encoder (h_0). Then, h_0 must be used in an attention mechanism along with the previously encoded image to compute a weighted image (a_0) for timestamp t_1.

At timestamp t_t, the weighted image (a_{t-1}) and the encoded image (h_{t-1}) serve as the inputs for the LSTM block. The output of the LSTM block (h_t) is used in a classifier to determine the result of t_t. Then, h_t is used in the attention mechanism with the encoded image (E), creating the weighted image (a_t), which indicates the area of focus for the next timestamp (t_{t+1}).

IV. EXPERIMENT AND RESULTS

In this section, the methodology proposed and the experiment environment deployed in are described. Several variations of ResNet (ResNet50, ResNet101, ResNet152) models were adopted to encode the input spectrogram

into encoded images. Three different lengths of time segments (0.1 sec, 0.5 sec, 1 sec) with their respective labels as the inputs were experimented to see which yields the best output. The outputs have to be at the granularity given for accuracy check.

The original dataset contains 489 recordings of lung sounds. Of all the recordings, 440 recordings were used for training, while the residual 49 recordings were used as evaluation data. All the data have been converted to spectrograms prior to the training stage.

When evaluating the performance of our models, several performance metrics are used. The performance metrics are segment-based, as were introduced in [18]. The following performance metrics are adopted, accuracy as well as the recall, precision, and F1-Score of inspiration and expiration, respectively.

The models were trained for 150 epochs on computers with either of two GPU cards, correspondingly (NVIDIA GTX1080Ti and NVIDIA GTX1080). Table 1 shows the performance of the model under each parameter combination. The data are obtained at the 150th epoch.

The ten-fold cross-validation is performed on the model using the 0.1 seconds per time segment and ResNet101-as-encoder setting for validating the consistency of our model. The accuracies of each fold of the cross-validation are presented in Table 2.

TABLE II: TEN-FOLD CROSS-VALIDATION ACCURACY

Fold	**Accuracy (%)**
1	91.72085212
2	91.71504896
3	91.52566394
4	91.78356289
5	91.4852517
6	91.91226181
7	91.80739176
8	91.31752191
9	91.95257249
10	91.76185349

The learning curves of accuracy for each time segment length (0.1 sec,0.5 sec, and 1 sec), using ResNet101 as encoder, are shown in Fig. 2, 3, 4, respectively. Learning curves can help us further understand how fast the models converge under their given parameters.

V. DISCUSSION

In Table 1, it is observed that the parameter settings of 0.5 seconds time segment with ResNet101 as the encoder has the best accuracy score. However, the performance of using 0.5 seconds segments is somewhat inconsistent, and the performance in F1-Score is rather poor while applying the 0.1 seconds per time segment granularity can produce a consistent 0.91-0.92 accuracy. Not only is using 0.1 seconds granularity more consistent, but it is also observed that this model yields better performance in other metrics.

In the case of 0.1 second time segments, the model applied to the encoder does not have a significant impact on the performance. Yet, the achievements of other granularities yield unstable returns, with the ResNet50 models generally producing the worst outcomes.

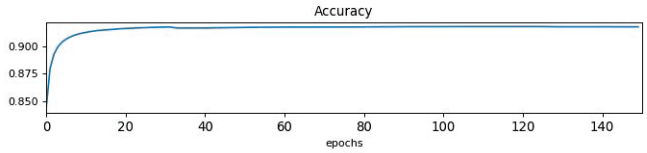

Fig. 2: Learning Curve in Case of 0.1 Seconds per Time Segment with ResNet101

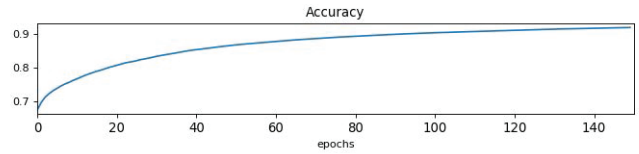

Fig. 3: Learning Curve in Case of 0.5 Seconds per Time Segment with ResNet101

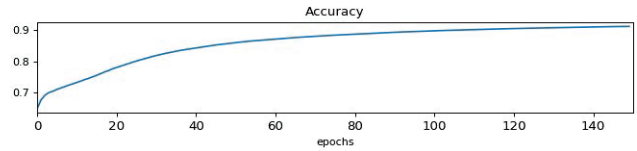

Fig. 4: Learning Curve in Case of 1 Second per Time Segment with ResNet101

Typically, the training time with the cases of 0.1, 0.5, 1 second time segment models take about 3, 5, 12 hours of training time for 150 epochs, respectively. Despite having the longest training time for 150 epochs, from Fig. 2, 3, 4, the results show that the performance of 0.1-second segment models converges at the optimal solution at around 20 epochs, which is approximately 1.6 hours. It implies that not only does shorter time segments achieve better prediction outcomes, it also takes a shorter time to train as well.

The profound differences in the recall rate of various time interval may imply a future research target to improve computer-assisted lung sound and respiratory phases classification. Performance of the classification system will be enhanced by time segmentation range optimization. Our result indicated a potential benefit for

higher sampling frequency on recall rate. However, the sampling frequency is directly related to the computing complexity and the scalability of the inference system. Besides, the higher sampling frequency could introduce a bias to the recall rate or increase the likelihood of overfitting the model.

An ordinary respiratory episode consists of an inspiratory phase and an expiratory phase. Between the inspiration and expiration episode, the resistive pressure of the chest chamber leads to a temporary pause of airflow, which results in a long silence period of the acoustic signal. Therefore, the silenced non-respiratory episode is predominant instead of the inspiration and expiration episodes. In our study dataset, each of the inspiratory epoch ranged from 0.599 to 1.037 seconds (first and third quantiles), and each of the expiratory epoch ranged from 0.395 to 1.080 seconds (first and third quantiles). Most of the 15 seconds recordings contained 3 to 5 times of respiration, which leads to a slight imbalance of inspiration/expiration and silence episode. Although, the higher sampling frequency (i.e., the 0.1 seconds sampling timestamp) produced a better recall rate, it could be a bias, caused by the increased minority subsets by oversampling. Further examination could be done to delineate insight of higher sampling frequency.

It is also worth noting that an attention-based model has the advantage of higher interpretability than others. In Fig. 5, each sub-figure of the upper row reveals the output of each timestamp, with the lighted area indicating the area of the spectrogram that the model is "attending" to. The bottom row shows the ground truth for each timestamp. It can be seen that for each timestamp, the model attends to a different area of the spectrogram for events. Therefore, it can be inferred that the model has learned to attend to the relevant field in each timestamp throughout the time sequence.

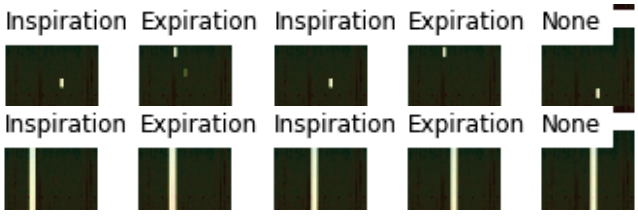

Fig. 5: Attending to Different Parts of the Spectrogram throughout the Time Sequence

VI. CONCLUSIONS

While previous studies have focused on the classification of lung sounds as a whole, An attention-based approach is proposed to the task of lung sound segmentation. By attending at 0.1 seconds time segment, the most consistent and accurate outputs are yielded. The results of this paper will be able to fuel future research on the segmentation of adventitious lung sounds. With the use of the attention-based model, more accurate segmentation outputs will be obtained. Physicians would be able to make more precise diagnosis with the results that have more interpretability.

ACKNOWLEDGMENT

The authors gratefully acknowledge the free support of anti-noising microphone sets from Heroic Faith Medical Science Co., Ltd. for the acoustic data acquisition.

This work was supported in part by Ministry of Science and Technology (MOST), Taiwan, under Grant Number MOST 108-2622-E-002 -012 -CC3.

References

[1] The WHO website. (August 2017). [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/asthma.

[2] The WHO website. (December 2017). [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd).

[3] D. Bardou, K. Zhang, and S. M. Ahmad, "Lung Sound Classification using Convolutional Neural Networks," *Artificial Intelligence in Medicine*, Vol. 88, pp. 58-69, June 2018.

[4] V. Vaityshyn, H. Porieva, and A. Makarenkova, "Pre-trained Convolutional Neural Networks for the Lung Sounds Classification," *Proc. of the IEEE 39th International Conference on Electronics and Nanotechnology (ELNANO)*, Kyiv, Ukraine, April 2019, pp. 522-525.

[5] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of Lung Sounds using Convolutional Neural Networks," *EURASIP Journal on Image and Video Processing*, Vol. 2017, No. 65, pp. 1-9, September 2017.

[6] Scottish Intensive Care Society website. [Online]. Available: https://www.scottishintensivecare.org.uk/training-education/sics-induction-modules/ventilation-i-e-ratio/.

[7] E. Sembroski and A. Bhardwaj. (December 2019). Inverse Ratio Ventilation. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK535395/.

[8] B. Zimmerman and D. Williams. (September 2019). Lung Sounds. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK537253/.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, June 2015, pp. 1-9.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Proc. of the 25th International Conference on Neural Information Processing Systems (NIPS)*, Lake Tahoe, NEVADA, December 2012, pp. 1097-1105.

[11] J. Amoh and K. Odame, "DeepCough: A Deep Convolutional Neural Network in a Wearable Cough Detection System," *Proc. of the IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Atlanta, GEORGIA, October 2015, pp. 1-4.

[12] G. Vrbančič, I. J. Fister, and V. Podgorelec, "Automatic Detection of Heartbeats in Heart Sound Signals Using Deep Convolutional Neural Networks," *Elektronika ir Elektrotechnika*, Vol. 25, No. 3, pp. 71-76, June 2019.

[13] G. D. Clifford, C. Liu, B. Moody, D. Springer, I. Silva, Q. Li, and R. G. Mark, "Classification of Normal/Abnormal Heart Sound Recordings: The PhysioNet/Computing in Cardiology Challenge 2016," *Proc. of the Computing in Cardiology Conference (CinC)*, Vancouver, BC, September 2016, pp. 609-612.

[14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," *Proc. of the Twenty-eighth Conference on Neural Information Processing Systems (NIPS)*, Montréal, CANADA, December 2014, pp. 1-9.

[15] C. C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-Art Speech Recognition with Sequence-to-Sequence Models," *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, April 2018, pp. 4774-4778.

[16] A. Bellur and M. Elhilali, "Audio Object Classification using Distributed Beliefs and Attention," *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (Early Access), pp. 1-1, January 2020. doi: 10.1109/TASLP.2020.2966867.

[17] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML)*, Lille, France, July 2015, pp. 2048-2057.

[18] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for Polyphonic Sound Event Detection," *Applied Sciences*, Vol. 6, No. 6, pp. 1-17, May 2016.