Performance Requirements for Cough Classifiers in Real-World Applications

A.C. den Brinker
Data Science
Philips Research
Eindhoven, NL
bert.den.brinker@philips.com

M. Coman
ICT & SW Engineering
Fontys University
Eindhoven, NL
m.marairina@student.fontys.nl

O. Ouweltjes
Data Science
Philips Research
Eindhoven, NL
okke.ouweltjes@philips.com

M.G. Crooks
Dept. Acad. Respir. Medicine
Hull York Medical School
Hull, UK
ORCID 0000-0001-6876-0258

S. Thackray-Nocera
Dept. Acad. Respir. Medicine
Hull York Medical School
Hull, UK
ORCID 0000-0002-9057-3608

A.H. Morice
Dept. Acad. Respir. Medicine
Hull York Medical School
Hull, UK
ORCID 0000-0002-6135-9610

***Abstract*—In the context of monitoring respiratory diseases, an unobtrusive cough monitor is an attractive tool. Preferably, such tool requires little or no customization. We address the question of the feasibility of such a device. A large database of sounds including coughs and other events was available. Using deep learning, a general cough classifier was constructed. The plug-and-play feasibility of such cough classifier is addressed by a leave-one-patient-out procedure. For a large part of the cohort (80%), the performance of the classifier is excellent meaning an area under the curve (AUC) of larger than 0.9. On top of that, estimates are derived for its success in practical scenarios by considering the prevalence of cough and the required specificity. It is shown that the acoustic environment can be harsh, requiring very high specificities. From the results, we argue that for real-world applications customization will be required. For part of the population, it suffices to set a patient-specific operation point in generic cough classifier, but for some part a personalized cough classifier will be needed.**

***Index Terms*—Respiratory diseases, COPD, cough, machine learning, deep learning**

I. Introduction

Cough is a characteristic associated with many respiratory diseases. For chronic respiratory diseases like chronic obstructive pulmonary disease (COPD) and asthma, it is one of the symptoms present in the associated questionnaires: the COPD assessment test (CAT) and Asthma Control Test (ACT), respectively. From a patient perspective, an unobtrusive measurement of health status is typically preferred over obtrusive measurements and, presumably, over questionnaires as well. From a clinician perspective, objective measurements are considered very valuable and largely lacking in current monitoring programs. Dyspnoea and cough were the most reported symptoms at the onset of acute exacerbation of COPD (AE-COPD) [1]–[3]. Cough in particular is a phenomenon that can be captured in an unobtrusive way yielding an objectively measured symptom [4].

Monitoring of patients at home would preferably be done in a plug-and-play fashion: a device supplied to the patient, easy to install and almost directly usable. This would call for a generic cough classifier, preferably one without the need of customization, otherwise one with easy means of customization. In this context, this paper studies the possibility of constructing such a cough classifier for a particular cohort of patients.

In a recent trial, data of COPD patients in their home environment were collected. The data consist of one second audio snippets of night-time recordings in the vicinity of COPD patients, where the monitoring period ran over a period of 90 days. Part of this data has been annotated and is used in combination with a deep learning algorithm for classification of coughs.

Several studies have reported on general cough classifiers, e.g. [5]–[10]. These classifier were developed on other data sets (patients, acoustic environments, sampling rates, features) and are typically not publicly available. We therefore started from scratch mainly to attain a model tailored to the current data (patients and acoustic environments). Rather than going for a competing cough classifier, the aim of this paper is to obtain insight into requirements for cough monitors when used in real home environments and, in particular, those of COPD patients. Therefore, estimates of the prevalence of coughs have been made as well, in order to define requirements for the cough classifier. The results of this study are a stepping stone in realizing real-world cough monitoring applications.

The outline of the paper is as follows. In Section II, we discuss the data and its annotation. Next, we consider the developed convolutional neural network used for classification and its performance. Specifically, the distribution of the performance over the available cohort is discussed in Section IV. In a similar way, the acoustic environment is charted in Section V.

II. Data

We conducted a prospective longitudinal study of continual cough monitoring in COPD patients experienced in telemon-

itoring. Participants underwent domiciliary cough monitoring and completed a daily questionnaire for 90 days, scheduled to be filled out in the morning. The study was reviewed and approved by the North East-York Research Ethics Committee (REC Ref: 15/N/0291), the United Kingdom Health Research Authority and the Internal Committee Biomedical Experiments of Philips Research. We will not discuss the protocol at length but only in as far as it is relevant for the current paper.

The cough monitor consisted of a stationary microphone paired wirelessly with a laptop computer. Based on previous data revealing a high correlation between day and night time cough frequencies [11], we chose the sleeping area in which to position the microphone (typically the bedside table). The cough monitoring system requires no user input following initial installation.

The cough monitor analyzed the audio for features, and stored the features at moments of non-stationary acoustic scenes, see Section V. On top of that, a limited number of one second long sound snippets were stored. These snippets enabled annotation while ensuring privacy as it is impossible to overhear any conversation. Data of a total of 28 patients was considered useful as material for the development of a deep learning algorithm. Only these snippets are used for the development of a generic cough classifier for COPD patients and not any of the stored audio features.

From the 28 patients, there were 20 having a partner and 8 were single. In almost all cases, coughs and non-coughs were annotated, except for four cases where the annotators thought it possible to distinguish between the patient and the partner. Coughs from the partner are not part of either class.

The applicability of the developed algorithm depends on a large degree on the annotation. Care was taken to have proper annotations. A tool was developed to support the annotation. During annotation, the audio snippets were presented as an audio file and its waveform was shown on the screen. Particulars of visuals of a common cough sound (explosive phase, intermediate phase, voiced phase) were known to the annotators. Given the combined audio and visuals, the annotator's task was to decide whether or not there was a cough starting with an explosive phase in the first half of the second.

The total number of coughs and non-coughs is highly variable over the patients. The statistics of the annotations and coughs over the patients is provided in Table I in terms of minimum, maximum and median. In order to balance this variation, patients with either very high amount of cough and/or non-cough annotations were separated and only part of their data was used. In this way at least a partial balance was created without scaling all data down to the level of the patient having the least amount of data. We note that not only a balance in the number of coughs was deemed necessary, but also in acoustic environments. Since the system was placed in the home of patients, there is no control on the background level, nor on the amount and level of other sounds. In such free-living conditions, even seemingly simple concepts as background level become difficult as, at minimum, they become stochastic time-variant quantities.

	Annotations	**Coughs**	**Events**
minimum	347	81	36975
median	2808	1225	358648
maximum	23204	8302	7253234
total	124393	48633	33499103

TABLE I
STATISTICS OF THE AMOUNT OF DATA OVER THE PATIENTS.

At the start of the project, a hold-out data set was created. This consisted of 100 coughs and 100 non-coughs per patient. For two patients this could not be attained due to a lack of data: in these cases all the data of the limited category was put in the hold-out set and no data of that category was therefore present in the training.

III. SYSTEM

The audio snippets (1 s long, sampled at 8 kHz) were pre-processed before entering the deep learning. Using a Mel frequency scale, spectrograms were made. No MFCC were created as deep learning tools profit from the redundancy of information in adjacent frequency bins. The band splitting and re-sampling, using frames of 20 ms length and 50% overlap, transformed each snippet into a matrix of 98×20 samples. This is the actual input to the deep learning system.

We note that in many cases of classification of time-related events, there is an issue in terms of how to match processing frames to detected events. Especially for acoustic events, starts and stops are very hard to localize for the annotator and there is no reason why these event boundaries would match with frame boundaries introduced by the processing. In this study, we have a full congruence between the annotator task and the classifier task: both work on exactly the same data.

Several network topologies were explored during the testing phase. The k-fold cross-validation method was applied on 10 randomly selected folds of data. The metric assessed was AUC. Ultimately, it was decided that a relatively simple deep learning system would be sufficient; extra layers of complexity did not add to the performance. This is presumably due to operating on the right data substrate, i.e., the pre-processing. Various activation functions were tested as well, before going in the validation phase using the hold-out set. The final system had an average AUC value of $96.21\,\%$ with a standard deviation of $\pm 0.27\,\%$.

This ultimate system, see Fig. 1, consisted of two convolutional layers with kernel size 3×3, a ReLU activation function and stride 1. The max-pooling uses size 2×2 and stride 2. The drop-out layer has 30% drop-out, and the dense layer has a sigmoidal activation function. The two convolutional layers involved 640+18,464 parameters; the dense layer contained 2945. In total, this amounts to 22,049 trainable parameters in the system. This system is the one used for performance evaluation using the hold-out set.

The system was implemented in Python with the help of Keras, a high-level neural network API. The Matplotlib library was used to visualize figures of the training process (loss and

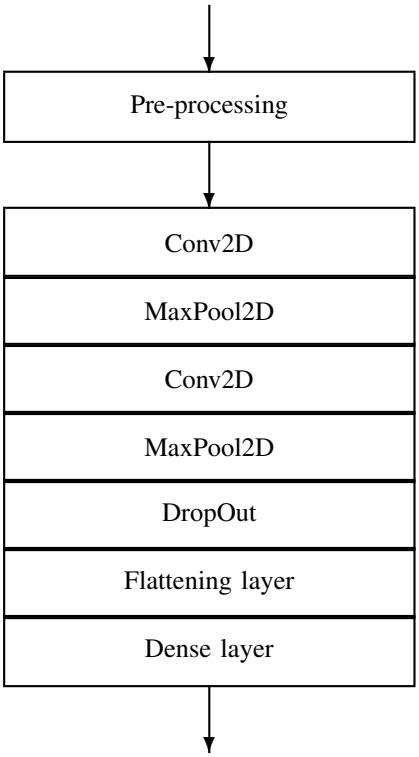

Fig. 1. Preprocessing and machine learning layers.

validation values) and performance of the model (AUC value) while validating the system.

IV. Performance

The final system was used in a leave-one-patient-out validation test, i.e. the model was trained using all data except those of the considered patient (and none of the data of the hold-out). Next, the performance with the trained model was measured for the left-out patient. The amount of training data is only slightly dependent on the considered patient and the ratio of the two categories is about 40/60%.

For each patient, the convolutional neural network is trained using the Adam optimizer, with a default learning rate of 0.001, the binary cross-entropy loss function, a batch size of 64, and a validation split of $10\,\%$ using the training data from all other patients. The training converged after about 80 epochs, with training data being randomly shuffled at each epoch.

In Fig. 2 we plotted the statistics of the resulting AUC. This is done in terms of a cumulative cohort plot: given a certain AUC, we determine which fraction of the considered patients actually attains a level higher than this. Obviously, no part of the cohort (cohort fraction=0) has an AUC above 1. As can be seen from the plot, about $80\,\%$ of the cohort has an AUC above 0.9.

V. Discussion

Benchmarking of systems is typically very difficult as often different databases have been used to evaluate the performance. Apart from the technical details (concerning e.g. testing/training separation, actual usage of a hold-out data set, etc.), it is often the question how representative the sounds are for real-life, and also various performance numbers depend on the prevalences of the two classes. In training and testing, the prevalences are hardly ever in line with realistic scenarios. Considering the prevalence-independent measures (sensitivity, specificity and AUC), the general feeling is that generic cough classifiers with excellent specificity (> 0.95) and a good sensitivity (> 0.8) should be feasible [5]–[10]. Heaping up all data over all patients puts the present system in that range as well.

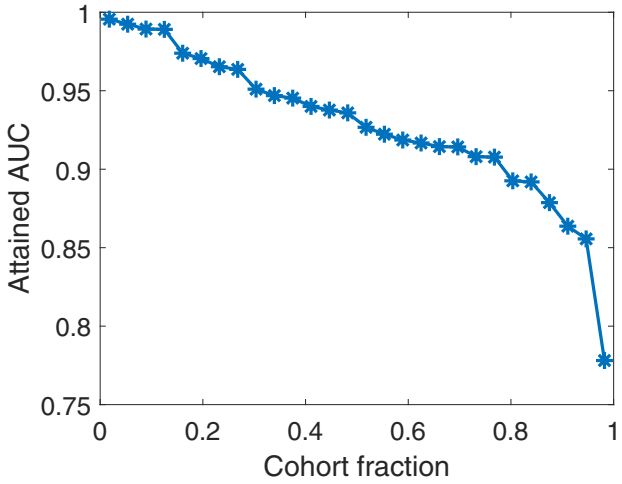

Fig. 2. Cohort fraction attaining a certain minimum AUC.

Particularly relevant is the question how such a generic system would work in a real application scenario. This question is hardly ever addressed in any detail, and we present a first quantitative analysis highlighting the harshness of the environment in which such a system would operate.

Consider a typical night, and suppose its length is 8 hours, meaning 28,800 s. Suppose furthermore we have a classifier constructed as outlined above and, therefore, use 1 s snippets as input where in the first half of the second a cough has to start. Then it would require to use an overlap of 50% leading to more than 56,000 frames. Suppose that we desire that the number of false identifications is below a certain number, than the required specificity can be determined. For simplicity, we say that if a person produces below 20 coughs over the course of a night, we would not really be concerned, even if this person is not a chronic cougher. To be below 20 would require a specificity larger than $1 - (20/56000) = 0.9996$. Such a number is with the current state of the technique an absurd high specificity only attainable with a sensitivity of about 0.

Now one might argue that the specificity need not be that high. The data in training and testing may actually not reflect the normal events occurring in the night and underestimate all events easily detected as silence or background noises. For

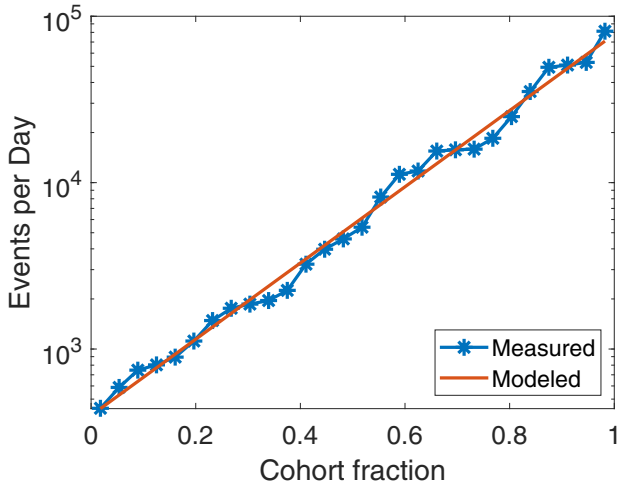

Fig. 3. Maximum number of events per day as function of cohort fraction.

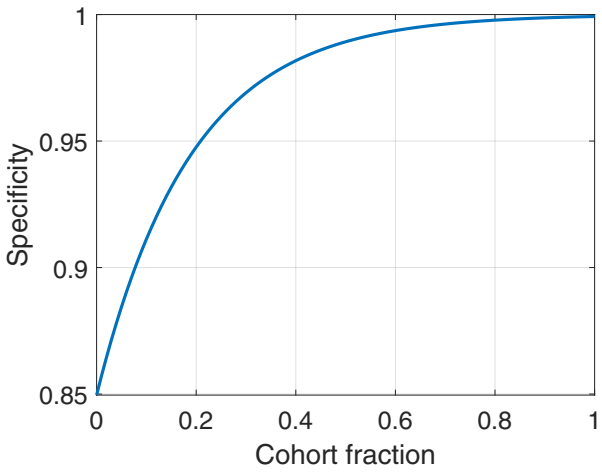

Fig. 4. Minimum required specificity and cohort fraction.

this reason, we designed a relatively simple mechanism to eliminate most of the silence periods and natural background sounds for the patient data. The level at which this happens is patient specific, and even typically time-dependent, thus a fixed level is not an option. The system we selected is one considering changes in the spectral content. This was effectuated by a low-order linear predictive coding (LPC) system similar to that discussed in [12]. Next, we determined the number of events (seconds) that could not be discarded offhand as background. As expected, this number varies largely over the population: some people live in a quiet and stable environment, some in an acoustically rich environment, i.e. with constant changes in the auditive scene. Statistics concerning the total number of acoustic events per patient over the course of the trial are provided in Table I. The number of segments needing analysis per day ranged from 450 to 27,000 and, on a log scale, is almost equally distributed over the cohort, see Fig. 3. We took this logarithmic relationship between the number of required classifications and the cohort percentage and derived the required specificity, which then ranges between 0.85 and 0.9993. The graph is shown in Fig. 4 and shows that about 50% of the cohort could be served with a classifier specificity performance of 0.99.

The performance figures obtained for personalized cough classifiers using the same data and using XGBoost, but on a different feature set [13] were found to be in the range 0.93 to 0.99 with a median of 0.97 over the patients. It demonstrates that personalization of the classifier gives a boost in performance, presumably partly due to a personalization to the cough of the targeted person, but also due to a better fit of the classifier to the particular acoustical environment the monitor is operating in.

Given the performances as indicated by the AUC and high specificity requirement due to acoustically challenging environments, it is expected that for real-world applications some form of customization will be needed. Only a small part of the population would be served with the performance of our developed generic cough classifier and, presumably, with any generic cough classifier. Nevertheless, to have a high-quality, relatively simple, generic cough classifier at one's disposal as a starting point for customization is considered key.

VI. Conclusions

Due to the low prevalence of cough events, classifier performance requirements for unobtrusive real-life applications are high. We have shown that a general cough classifier based on deep learning provides a system which achieves high performance for a major part of the participants as evidenced by the high AUC. We also argued that for real-world applications some form of customization is likely to be needed in view of the low prevalence of coughs relative to other acoustic events.

Acknowledgments

The authors wish to thank R. van Dinther and R. Rietman for their comments on a first version of this paper.

References

[1] T. A. Seemungal, G. C. Donaldson, A. Bhowmik, D. J. Jeffries, and J. A. Wedzicha, "Time course and recovery of exacerbations in patients with chronic obstructive pulmonary disease," *Am. J. Respir. Crit. Care Med.*, vol. 161, no. 5, pp. 1609–1613, 2000. [Online]. Available: https://doi.org/10.1164/ajrccm.161.5.9908022

[2] C. M. Parker, N. Voduc, S. D. Aaron, K. A. Webb, and D. E. O'Donnell, "Physiological changes during symptom recovery from moderate exacerbations of COPD," *Europ. Respir. J.*, vol. 26, no. 3, pp. 420–428, 2005.

[3] A. Oliveira and A. Marques, "Understanding symptoms variability in outpatients with AECOPD," *Pulmonology*, vol. 24, no. 6, pp. 357 – 360, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2531043718301557

[4] M. G. Crooks, A. den Brinker, Y. Hayman, J. D. Williamson, A. Innes, C. E. Wright, P. Hill, and A. H. Morice, "Continuous cough monitoring using ambient sound recording during convalescence from a COPD exacerbation," *Lung*, vol. 195, no. 3, pp. 289–294, Jun 2017. [Online]. Available: https://doi.org/10.1007/s00408-017-9996-2

[5] S. Larson, G. Comina, R. H. Gilman, B. H. Tracey, M. Bravard, and J. W. López, "Validation of an Automated Cough Detection Algorithm for Tracking Recovery of Pulmonary Tuberculosis Patients," *PLoS ONE*, vol. 7, no. 10, p. e46229, Oct 2012.

[6] H.-H. Wang, J.-M. Liu, M. You, and G.-Z. Li, "Audio signals encoding for cough classification using convolutional neural networks: A comparative study," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov 2015, pp. 442–445.

[7] Y. A. Amrulloh, U. R. Abeyratne, V. Swarnkar, R. Triasih, and A. Setyati, "Automatic cough segmentation from non-contact sound recordings in pediatric wards," *Biomedical Signal Processing and Control*, vol. 21, pp. 126 – 136, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809415000804

[8] J. Amoh and K. Odame, "Deep neural networks for identifying cough sounds," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 5, pp. 1003–1011, Oct 2016.

[9] M. You, Z. Liu, C. Chen, J. Liu, X.-H. Xu, and Z.-M. Qiu, "Cough detection by ensembling multiple frequency subband features," *Biomedical Signal Processing and Control*, vol. 33, pp. 132 – 140, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1746809416301835

[10] J. Monge-Álvarez, C. Hoyos-Barceló, P. Lesso, and P. Casaseca-de-la-Higuera, "Robust detection of audio-cough events using local hu moments," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 184–196, Jan 2019.

[11] M. G. Crooks, Y. Hayman, A. Innes, J. Williamson, C. E. Wright, and A. H. Morice, "Objective measurement of cough frequency during COPD exacerbation convalescence," *Lung*, vol. 194, no. 1, pp. 117–120, Feb 2016. [Online]. Available: https://doi.org/10.1007/s00408-015-9782-y

[12] L. Oudre, "Automatic detection and removal of impulsive noise in audio signals," *Image Processing On Line*, vol. 5, p. 267–281, 2015.

[13] L. Di Perna, G. Spina, S. Thackray-Nocera, M. G. Crooks, A. H. Morice, P. Soda, and A. C. den Brinker, "An automated and unobtrusive system for cough detection," in *2017 IEEE Life Sciences Conference (LSC)*, Dec 2017, pp. 190–193.