ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **23**, Number S, pp. S77–88, 2020

Cough Sound Recognition in Respiratory Disease Epidemics

Gheorghe POP, Horia CUCU, Dragoş BURILEANU,
and Corneliu BURILEANU
Speech and Dialogue Laboratory, Faculty of Electronics, Telecommunications
and Information Technology, University “Politehnica” of Bucharest, Romania
gheorghe.pop@etti.pub.ro,
horia.cucu@upb.ro, dragos.burileanu@upb.ro, corneliu.burileanu@upb.ro

Abstract. The new coronavirus epidemic, which outbroke in 2019, has now grown into a full-blown pandemic, raising global concerns by its high infection speed and mortality rate. People developing the disease fill emergency hospitals, while the problem may deepen if worried general population cluster emergency rooms just for diagnosis. To control such respiratory disease epidemic, governments and medical staff usually decide to reduce virus transmission by enforcing social distance and placing in quarantine all persons suspected of carrying the virus. Everyone else is asked to stay insulated as long as possible, and refrain from calling emergency services unless relevant symptoms appear. With the medical staff shortage forced by such an epidemic, it would be very useful to have a diagnosis system capable of checking people for symptoms. As the direct contact of patients with objects used in common may raise virus transmission concerns, non-contact devices are accepted for use in evaluating a person’s health condition. Under these limitations, we present a cough sound recognition method, which, as new relevant data become available, can be extended to work more as a respiratory disease diagnostic tool.

Keywords: neural computing, respiratory disease epidemics, cough sound recognition.

1. Introduction

We are now in an ongoing coronavirus pandemic which outbroke during December 2019 in Wuhan, province of Hubei, China. The virus, known under several medical names, such as HCoV-19, SARS-nCov19, and SARS-Cov-2, inflicts a Severe Acute Respiratory Syndrome (SARS), called the *Coronavirus disease 2019*, or *COVID-19* [1]. Common symptoms include fever, cough, and dyspnea, while other symptoms may include fatigue, muscle pain, diarrhea, sore throat, loss of smell, loss of taste, and abdominal pain [2]. While most of infected individuals show mild symptoms, some develop viral pneumonia and multi-organ failure.

The coronavirus was found as most contagious during the first three days after the onset of symptoms, although both asymptomatic spread and pre-symptomatic spread are considered [3]. The virus is mainly transmitted between people, during close contact, via small droplets produced by coughing, sneezing, or talking. While these droplets are produced when breathing out, they usually fall to the ground or on surfaces rather than being infectious over large distances. People may also get infected by touching their face after having touched a contaminated surface, where the virus can survive for up to 72 hours [4]. The COVID-19 has generated worldwide concern, and the World Health Organization (WHO) has classified the epidemic as a global public health emergency. With more than 280,000 lives taken so far, it follows the path of other major respiratory virus epidemics. The SARS in 2002, the Middle East Respiratory Syndrome (MERS) in 2012, and the avian flu H7N1 in 2013, all claimed hundreds of lives, but the SARS-CoV-2 threatens with damages nearing those of the swine flu virus H1N1, which killed more than 284,500 people in 214 countries and territories between 2009 and 2010 [5].

All the above-mentioned epidemics attack the human respiratory system, and SARS-CoV-2 spreads through activities such as coughing, sneezing, and even breathing. Once the virus contamination occurred, the respiratory sounds may change, even before the onset of other observable symptoms.

The lack of a known cure, the symptoms and life-cycle of SARS-CoV-2 make the outbreak difficult to control and manage, so that authorities had to impose lockdowns on public places, cities, and even regions. The World Health Organization and healthcare agencies, as first responders to such threats, recommend governments to place symptomatic people in quarantine facilities, while asking all possible virus importers (which came lately from abroad or got in close contact to persons found positive at virus-carrying tests) to self-isolate at their home. For efficient epidemiologic inquiries to be conducted, meant to discover and isolate persons suspected of carrying the virus, movements and health status of yet asymptomatic people must be controlled. To this end, smartphones and smartphone applications play a key role, enhanced locally by some states which provided applications and networked infrastructure for citizen health state monitoring. Movement control may consist in collecting telephone call data, GPS location of personal smartphones, as well as checking of respiratory symptoms.

This creates a good base for respiratory symptoms in SARS-CoV-2 infection, such as cough sound, to be identified by analysis of audio recordings of collected respiratory sounds. Thus, cough detection, classification, and recognition became important topics in the context of respiratory epidemics, against which many researchers enrolled with the goal of fighting epidemics by signal processing.

In this paper, we present a method of cough sound recognition which can detect and classify cough sounds, by automatically and cumulatively learning adequate features for the classification task, then computing the strength of the classification decisions. The presentation is organized as follows. A review of current literature in the field is presented in section 2. Section 3 hosts the description of the problems faced by cough recognition methods in the context of respiratory disease epidemics, while section 4 describes the method we propose, the reference corpora and experiments conducted. Finally, section 5 contains the conclusions of the study.

2. Review of current research

The first to use deep learning in detecting cough events were Amoh and Odame [6]. In their first work on this subject, back in 2015, they compared classical machine learning algorithms such as Hidden Markov Models (HMM) with Gaussian Mixture Models (GMMs) outputs and Support Vector Machines (SVMs) with Convolutional Neural Networks (CNNs). The classical machine learning classifiers were trained on Mel-Frequency Cepstral Coefficients (MFCCs), while the CNNs exploited a Short-Time Fourier Transform (STFT) spectrogram. Their work continued in 2016 [7], when they also explored the possibility of using Recurrent Neural Networks (RNNs) for the same task (classification of short audio files into cough vs. non-cough). All their experiments were performed on a private dataset that was recorded specifically for their study, comprising 627 cough events from 14 people. The audio recordings were performed with a wearable microphone.

Although deep learning was already being used for the task of cough detection since 2015, some new studies still resorted to traditional machine learning classifiers such as Random Forests (RF) [8], various flavors of SVMs [9, 10, 11, 12], k-Nearest Neighbors (kNN) [10], or logistic regression [13]. However, the dominant approach is by far the CNN applied either on mel-spectrograms [14, 15] or STFT spectrograms [16, 17]. The popularity of CNNs is probably based on their advantage of being able to process time-variable inputs. Vanilla Deep Neural Networks (DNN) were also used in two studies which found an innovative way to deal with the variable-length of the input. Khomsay et al. [18] extract a Fast Fourier Transform (FFT) from the full audio signal, then apply Principal Component Analysis (PCA) to lower the dimensionality of the feature vector from 1,024 to 100 and finally feed this 100-dimensional vector into a DNN. Kadambi et al [19] process fixed-size segments of 200 milliseconds. They extract MFCC-based feature sets from 4 non-overlapping windows of 50 milliseconds, concatenate all the features into an 168-dimensional feature vector and feed it to a vanilla DNN. Finally, following [7], there was a second study to use RNNs, more specifically RNNs with Long Short-Term Memory (LSTM) units, for cough detection. Miranda et al. [17] recently made a comparative study of deep architectures for acoustic cough detection and tried out DNNs, CNNs and LSTMs.

In terms of features, besides classical MFCCs, mel-scaled spectrograms and STFT spectrograms, Monge-Álvarez et al. [10] used an ensemble of multi-dimensional spectral features such as Linear Prediction Cepstral Coefficients (LPCC), GammaTone Cepstral Coefficients (GTCC), Normalized Audio Spectral Envelope (NASE), Octave Spectral Contrast (OSC) and Spectral Subband Centroid Histograms (SSCH) plus 13 additional uni-dimensional spectral features. Finally, they calculated local Hu moments as a robust candidate feature set for cough detection in noisy environments [11, 12]. Pramono et al. [13] also resorted to traditional spectral features such as the high-frequency content in B-HF ratio, the min-max ratio in B-01 (low) frequency band, and the low quantile ratio also in B-01 band. In these studies, the aforementioned spectral features were fed into traditional machine learning classifiers such as SVM, RF, or kNN.

While most of the papers on cough detection report performance metrics such as sensitivity, specificity, or area under receiver operating characteristic (AUC), we consider that these results are not comparable. The main problem is the fact that all the studies use self-created or private datasets, which they do not make available for further comparative investigation. Moreover, the datasets are very diverse. They start with as few as 7 patients [9] and go up to 43 patients [14]. Some datasets comprise just a few cough examples: 262 in [16] and around 300 in [18], while others have over 1,000 positive examples: around 1,000 in [10]; 1,500 in [15]; 5,670 in [19]; 6,737 in [14]; and 13,000 in [9].

Finally, the cough detection task is not uniformly formulated. In most studies, the dataset comprises short cough (usually 1 second long) and non-cough examples. However, the non-cough examples are speech-only in some cases and more varied (speech, laughter, sneeze, throat clearing, wheezing sound, whooping sound, machine noises, and other types of noise) in other, more rare cases. This, of course, has an impact on the difficulty of the task. More practical formulations of the task are proposed in [9] and [19]. They address cough detection as a continuous monitoring task and they continuously process 1 second, respectively 200 milliseconds long segments of audio from 24-h recordings. Finally, Bales et al. [16] proposes yet another, very different, formulation of the task. Their dataset comprises only cough sounds (a total of 262), which correspond to three different respiratory illnesses: bronchiolitis, pertussis, and bronchitis, and they aim at classifying the illness.

3. Problems of cough sound recognition in respiratory disease epidemics

One of the biggest problems with epidemics is that their earliest stages are critical to understanding what disease is being dealt with. As soon as some data are available, the first objective of the uninterrupted fight against new possible respiratory epidemic is to tag each disease case as known or unknown. If known, then the protocols and the knowledge needed to tackle it down are usually in place, while if unknown, it may be a new variant of a known disease or some totally new illness.

Whatever the cough monitoring algorithm is, the cough count output is not as informative as needed by medical doctors, because sometimes they need to diagnose faster than the time necessary for obtaining reliable measurements. Cough monitoring alone does not offer a useful diagnosis of epidemic diseases until it is known which disease has outbreak as epidemic.

Cough has a great deal of variations, such as dry and wet, with single expulsion, as opposed to several expulsions of air, while being a symptom of numerous respiratory diseases, both chronic and acute. For chronic respiratory diseases, such as the Chronic Obstructive Pulmonary Disease (COPD) and asthma, monitoring of cough has already become one of recommended measures to take to predict severe episodes. For acute conditions, as well as for episodes of chronic diseases, particular aspects of the cough are looked for, such as wheezing (monophonic or polyphonic), crackle (fine or coarse), rhonchi, stridor, pleural rub (pleural friction), and squawk sounds [20].

Illnesses such as COVID-19 are easy to mistake as a cold or flu, even by specialists, and that’s why sometimes X-ray images are necessary, besides respiratory sounds, and knowledge of results from other tests, to assist in reaching a correct diagnostic, with human intervention as a must.

Another problem is that the cough classification algorithms used in machine learning, and especially deep learning, need to see vast amounts of real data before doing reliable classification. Given that it is never useful at the time the first unknown respiratory disease case is reported, the need of data collection for deep learning algorithms is often omitted by first line caregivers, and collecting it afterwards means to let the epidemic evolve wildly in the meantime.

Specific observations on the health state of any patient are separated into several sections of hospitals’ data management systems, which makes it difficult to extract and compile time synchronous data for datasets usable in epidemic crisis management. Ideally, specific observations on the health state of any patient should fit in parts of readily prepared data point templates which would comprise respiratory sounds, thoracic X-rays, body temperature, body-worn sensors, self-assessment data on general state, and so on, all attached to the person being observed, without the real world ID data.

Lung sounds are non-stationary, with their evolutions in time leading to complicated procedures for analysis, recognition, and distinction. Sound recognition algorithms should therefore consider features based on large analysis windows, together with other recursive techniques.

Sound and image collection, whatever their type, should preferably be non-contact, to reduce the risk of transmitting contagious diseases. This raises concerns about using medical devices such as stethoscopes, including digital ones, and recommends the use of personal devices, wearable or not. Tracheal microphones and some laryngophones need specialized personnel to install them in medically correct installation points, as well as a corresponding level of biohazard security, and therefore are of limited use.

Smartphone applications are a widespread good way to securely and precisely record (or even self-record) respiratory sounds of patients. When installed on a patient’s body, it can collect various features, such as audio and accelerometer, but also body temperature, blood O_2 saturation, and so on, and inject it into the extended dataset during sleep hours.

Although data sources in respiratory disease epidemics are heterogenous and rather sparse in time, there is no overall model of the human body, as a context of respiratory symptoms. Such a model could be initialized and fine-tuned with individual data for each patient.

4. Proposed method and experiments

A patient's timeline is able to get filled in with its own data of medical importance in epidemics, such as evaluations on global risks against patient's life, as well as contagiousness to others, but the most useful conclusions are still drawn by specialized medical staff. For that, they need various views on the acquired knowledge pertaining to the patient involved, including discovery of comorbidities.

Such a task is enormous in epidemics, and it needs to be started as soon as possible, even if data from important investigations are not available initially. The low volume and lack of significance of first pieces of data collected, in case of unknown respiratory disease epidemics, do not support early intervention. We thus propose a method which can rapidly start as a recognizer of cough – a symptom in most of respiratory illnesses.

As soon as new relevant data volumes are collected, the cough sound recognition can be expanded, by re-training the feature extractors and classifiers on the evolved dataset, in order to perform better and tackle more classes of sounds of medical importance. Given that acoustic waves carrying the information of interest are exposed to contamination by unwanted sound sources, during the collection in various environments (at home, during the ambulance transportation, or during intensive therapy), there is a need to also label known classes of interferers and sounds of normal states (in our case, non-cough sounds).

This is why we started an exploratory process with sound classification experiments, in which we used feature extractors trained on sound event corpora of various number of classes, not just "cough" and "non-cough". The exploration includes three experiments with extracted features examined separately, and classifiers specially trained in each case.

The principle of the proposed method, depicted in Fig. 1, and evaluated mainly through the fourth experiment, relies on combining complementary features, extracted from the same waveform. Input signals are first split in sound events, although some corpora already are in a one-event-per-file format. The extractors used are time-aligned at the short-time analysis level, so that all feature extractors (n = 2 in our case), produce the same number of feature vectors. For a more performant classification, vectors from both extractors, for the same analysis window, are concatenated to form the input of the final classifier.

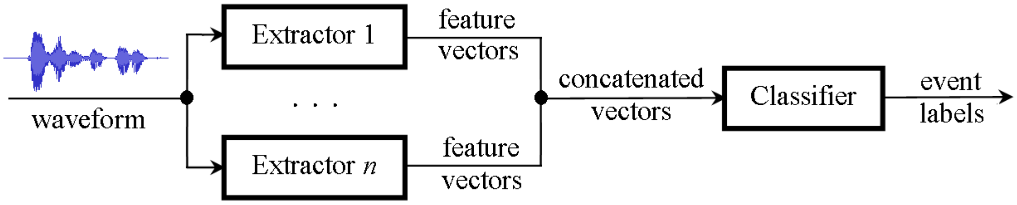

Fig. 1 – The principle of the proposed method (n = 2)

We begin the presentation with the resources used, that is, with the description of corpora used to adapt the feature extractors to input corpora, then the adaptable feature extractors will be presented, followed by the description of the sequence of experiments.

4.1 Sound event corpora

4.1.1 ESC-50 v2.0 Dataset [21]

The ESC-50 dataset consists of 2,000 labeled environmental sounds, equally balanced between 50 classes (40 clips per class). For convenience, they are grouped in 5 loosely defined major categories (10 classes per category): animal sounds, natural soundscapes, and water sounds, human (non-speech) sounds – including coughing, interior/domestic sounds, exterior/urban noises. The goal of the database build was to keep sound events exposed in the foreground with limited background noise when possible.

However, naturally occurring events are far less clean. A variety of sound sources are present, from some very common (laughter, cat meowing, dog barking) to some quite distinct (glass breaking, brushing teeth), while some class differences are more nuanced (helicopter and airplane noise). The events contained in the dataset come from an initial collection of files, some of which contain more than one sample of the event, while yet others contain several samples from different sound classes This is why sound event databases often come with metadata files, where data such as the label, validation fold, event ID and position in the contents of initial files are stated for each sample.

4.1.2 ICBHI'17 Respiratory Sound (IRS17) Database [22]

The database consists of a total of 5.5 hours of audio containing 6,898 respiratory cycles, of which 1,864 contain crackles, 886 contain wheezes, and 506 contain both crackles and wheezes, stored in 920 annotated audio files from 126 subjects.

The cycles were annotated by respiratory experts with four labels: "crackles", "wheezes", "both", and "none". The recordings were collected using heterogeneous equipment and their duration ranged from 10 s to 90 s. The sensors and chest locations from which the recordings were acquired are also provided in the metadata. Noise levels are often high, like in real life conditions.

4.1.3 Sweet Home Cough (SHC) corpus [23]

A number of 882 sound events were identified in 240 event recordings, consisting in separate short and long cough sound events, as well as voice-fix sounds, voluntarily produced by 9 women and 6 men of ages between 20 and 41. The recordings were made using a SWEET_HOME facility, under the supervision of Michel Vacher and François Portet, from the GETALP group at Laboratoire d'Informatique de Grenoble, France [24]. In view of the experiments conducted, we placed the sound events in categories labelled as "short_cough", "long_cough", and "voicefix".

4.2 Features and feature extractors used

As we train feature extractors to discover the most discriminative features in a given corpus, different features are discovered in different training corpora. The trainable feature extractors we selected are taken as complementary.

4.2.1 SoundNet type features [25]

The first feature type we used was produced by a SoundNet neural network, shown in Fig. 2, in which a deep 1-D Convolutional Neural Network (CNN) was trained by authors of [25] in a student-teacher architecture, using a corpus of more than 2,000,000 unlabeled videos, in 1,000 sound event categories (including cough sounds), and 400 different environment categories (*Places CNN*).

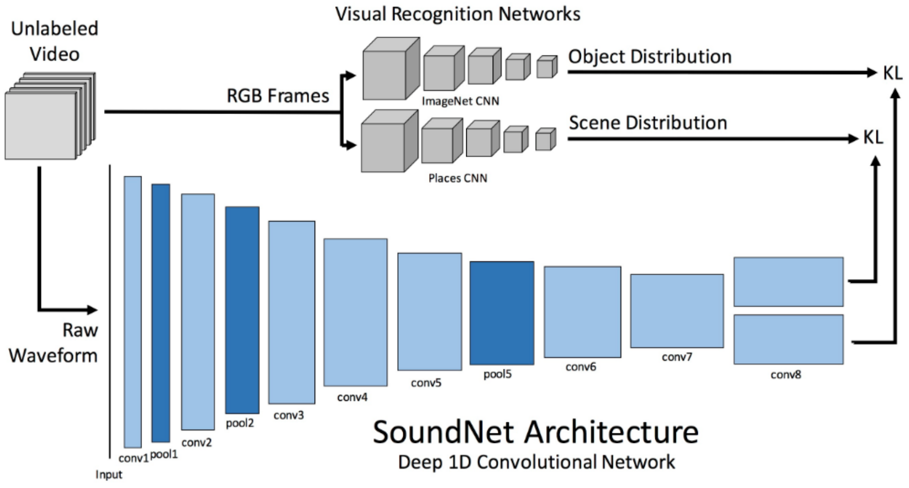

Fig. 2 – SoundNet architecture, from [25]

Natural synchronization between image and sound is leveraged by learning acoustic representations from videos, using a student-teacher training procedure. From well-established CNN visual models (i.e. *ImageNet* and *Places*), the discriminative visual knowledge is transferred into the sound modality using unlabeled video as a bridge [25]. Extracted features are the layer coefficients from *conv5* to *conv8*, respectively, as shown in Fig. 2, which correspond to layers 14 to 24 in the computation graph. We only used the *conv8* features, which are the coefficient set from layer 24.

According to the experimental needs, the CNN audio feature extractor was used both as given and after fine-tuning, with audio data alone. For consistency reasons, the SoundNet fine-tuning will be still called "training". The fine-tuned extractor was denoted as SoundNet-FT, because of changes incurred in the type of features.

4.2.2 auDeep type features [26]

AuDeep project aims to provide unsupervised feature learning with DNNs, for sequence-to-sequence classification This is paramount in case of unknown respiratory illnesses, because of the classification of unlabeled events. In Fig. 3., the internals of the recurrent autoencoder (RAE) are presented, where t_0, t_1, t_2, …, t_{n-1}, and t_n are the center timestamps of the successive signal analysis windows, "0" is an initialization all-zero spectrum, and the extracted features are the final states of the encoder.

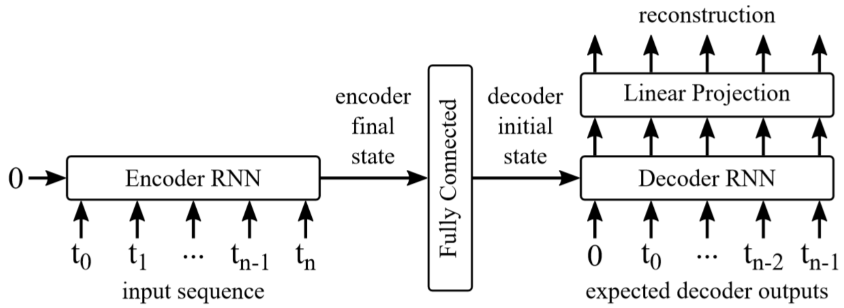

Fig. 3 – auDeep diagram of the recurrent autoencoder feature extraction [26]

In the auDeep Python toolkit [27], the input audio sample is first converted into a configurable log-mel-spectrogram. We chose to use 256 mel-frequency bands, over 0.16 s long analysis windows, with 50% overlap, and normalized to [–1; 1]. The auDeep allows the transformation of spectrograms so that values under a specified threshold are clipped. The features extracted by auDeep with different spectrogram thresholds can be combined using an internal mechanism of feature fusion, presented in Fig. 4.

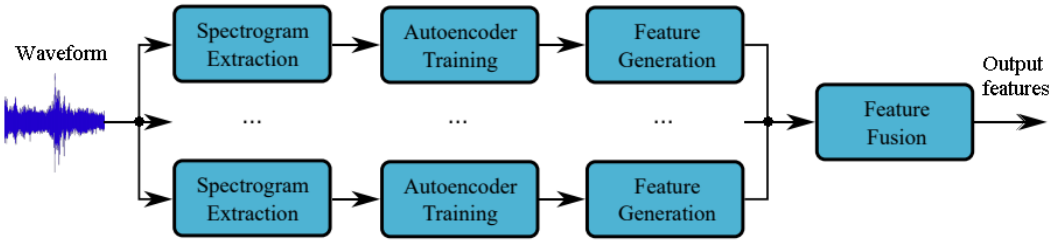

Fig. 4 – auDeep feature fusion mechanism [26]

The length of output feature vectors is adjustable for both extractors. For auDeep, the internal feature fusion mechanism can be used, while for the SoundNet, the coefficients from a different range of network layers can be selected as trained features. Nevertheless, we used a single spectrogram clipping threshold in auDeep, at –60 dB (no internal fusion), while from the SoundNet we retained the layer 24 coefficients as output features.

4.3 Sound classification and cough sound recognition experiments

As we needed to examine sound events specific to at least home, transportation, and hospital environments, we considered finding a suitable feature set, by training feature extractors on more or less adequate datasets. Given that the selection of the corpora may have a tremendous effect on the performance of the classification, three exploratory experiments were conducted on selected corpora, before the cough sound recognition performed in the fourth experiment. The exploration was aimed at checking cough sound recognition performance using Multi-Layer Perceptron (MLP)-type classifiers, especially because of their good discrimination power.

All classifiers used the same hidden structure, consisting in two layers of 150 ReLU-activation units each, while the Softmax output size was different, according to the number of classes in the training corpus considered. The classifiers were trained using an Adam optimizer for 50 epochs, with a 0.001 learning rate and 40% dropout regularization.

The MLP for each classification task was trained on half the corpus involved, selected randomly, while the evaluation was performed on the remaining half. Considering the degree of similarity between the structure, training, and testing of all MLP classifiers intended, in the remaining part of the present paper we call it a *standard MLP*.

4.3.1 Sound classification with feature extractors trained on ESC-50

For this experiment we trained the auDeep and SoundNet feature extractors, along with SoundNet-FT, on the whole ESC-50 corpus. They were next used to produce features for all files in three corpora: ESC-50, IRS17, and SHC. Nine standard MLP classifiers had to be trained, one for each corpus-extractor pair. The accuracy of all classifiers is presented in Table 1, per feature type.

Table 1 – Classification accuracy for extractors trained on ESC-50

Corpus	**Samples (classes)**	**Accuracy in [%] by feature type**		
		auDeep	**SoundNet**	**SoundNet-FT**
ESC-50	2,000 (50)	70.4	74.2	**74.9**
IRS17	6,898 (4)	67.3	64.6	74.5
SHC	882 (3)	68.5	62.7	69.7

When tested on corpus ESC-50 with features trained on the same corpus (see Table 1), the SoundNet-FT performed best, as expected, with a **74.9%** accuracy. Given that audio event samples in ESC-50 are clipped at 5 seconds, and some cough sounds may occur more than once, the identification performance does not allow a precise discrimination between files containing one, two, or more cough sounds. Respiratory sounds in IRS17 and cough sounds in SHC were classified with a similar accuracy because they show some similarity with classes already present in ESC-50. The superior performance on ESC-50 for all the feature extractors used is a direct result of matching the corpora for training of feature extractors and the training and testing of the classifiers.

4.3.2 Sound classification with feature extractors trained on SHC

The auDeep and SoundNet feature extractors were trained for this experiment on the entire SHC corpus. Both feature extractors, along with SoundNet-FT, were then used to produce features for all files in the three corpora: ESC-50, IRS17 and SHC. Nine more standard MLP classifiers had to be trained, with 3 output units, this time (“short_cough”, “long_cough”, and “voicefix”). Classification accuracy, as evaluated for each corpus-feature type pair, is shown in Table 2.

Table 2 – Classification accuracy for extractors trained on SHC corpus

Corpus	Samples (classes)	Accuracy in [%] by feature type		
		auDeep	SoundNet	SoundNet-FT
ESC-50	2,000 (50)	27.4	74.2	46.9
IRS17	6,898 (4)	21.3	64.6	24.5
SHC	882 (3)	88.4	62.7	**95.5**

Classifiers trained with features adapted to SHC performed worse than in experiment 4.3.1, except for those trained on features from pre-trained SoundNet, which performed similarly. This could be explained in part by the presence in SHC corpus of sounds dissimilar to most of ESC-50 and IRS17 classes. Thus, all feature extractors performed better on "coughing" and "voicefix" classes, but worse on most of the other classes. One notable exception was produced by SoundNet-FT, which reached **95.5%** accuracy. This result, although at the state-of-the-art level, may be a circumstantial one, most probably explained by the repeatability of characteristics of the cough sounds from each corpus contributor. The need of a supplemental balance between sound source persons is important when gathering datasets for cough recognition on a large scale. On the other hand, it shows a real potential for particularization of cough recognition to each person.

4.3.3 Sound classification with feature extractors trained on a special 10-class corpus

A new 10-class training corpus was compiled, by putting together categories from ESC-50 named "breathing", "coughing", "sneezing", "drinking, sipping", and "laughing", with other categories, namely "short_cough", "long_cough", and "voicefix" from SHC, as well as "wheezing" and "crackles", from IRS17, formed in a similar way to dataset ESC-50 (40 files per class).

For this experiment we trained the auDeep and SoundNet-FT feature extractors on the complete special 10-class corpus. The feature extractors were next used to produce features from all files in IRS17 and SHC corpora, leaving the ESC-50 corpus aside. For each corpus-feature type pair, a standard MLP was trained, with 10 output units. The accuracy of the classifiers is presented in Table 3.

Table 3 – Classification accuracy for extractors trained on a special 10-class corpus

Corpus	Samples (classes)	Accuracy in [%] by feature type	
		auDeep	SoundNet-FT
IRS17	6,898 (4)	54.3	64.5
SHC	882 (3)	78.4	**81.5**

The compilation of a special corpus was necessary to explore the discrimination power of features extracted from classes of sounds which are rather similar. As the results in Table 3 show clearly, the balance in structure between train and test corpora leads to results that are close to the native performance allowed by extracted feature vectors, as evaluated in [25] and [26].

4.3.4 Cough recognition with the proposed method

Based on the exploration results from the experiments 4.3.1 to 4.3.3, we used the same 10-class dataset as in sub-section 4.3.3, in the training of a feature extractor for the method we proposed, and then conducted the cough recognition experiment which is described next.

It should first be noted that the three experiments described above used feature vectors extracted with discriminative training on specific datasets, for which classifiers were tested on corpora very unbalanced, in terms of cough/non-cough classification. Any sound corpora with cough-related categories, such as the ESC-50 corpus, which has one cough category and 49 non-cough categories, can be used in training classifiers for cough/non-cough classification. However, cough recognizers based on features learned from ESC-50 will only be able to use about 2% of the information learned, because much of discrimination data is associated to sound pairs not connected to cough, which will never be used.

The special 10-class corpus is better from this point of view, as the rate of learned information used, although computed differently for a different corpus structure, is expected to raise at roughly 30%.

For cough recognition, two classes are formed from the files in the special 10-class corpus. The files from classes "coughing" (ESC-50), "short_cough" and "long_cough" categories (both from SHC), were set in the class "cough", while all the other classes in the special corpus ("voicefix" from SHC, "wheezing" and "crackles" from IRS17, and "breathing", "sneezing", "drinking, sipping", and "laughing" from ESC-50) were included in the class "non-cough".

The feature vectors we used in this experiment consist of feature vectors from the SoundNet and auDeep, combined as shown in Fig. 1, with both feature extractors trained on the 2-class corpus described in this subsection. During training, the class upsampling technique was used to achieve balance. A standard MLP classifier with 2 output units was trained for each corpus-feature type pair. The obtained accuracy is shown in Table 4.

Table 4 – Cough recognition accuracy of the proposed method

Corpus	Examples (cough/non-cough)	Accuracy [%]
IRS17	6,818 (0/6,818)	86.6
SHC	802 (457/345)	**91.6**

According to the accuracy values in Table 4, corpora built on purpose for training feature extractors are better in adapting classifiers to real-life respiratory sounds. The 86.6% accuracy on IRS17 is based more on correctly classifying the respiratory sounds, and less on recognition of cough sounds. A classifier with several types of cough and non-cough classes can better discriminate target sound classes from non-target ones.

The proposed method for cough sound recognition classifies sounds in "cough" and "non-cough" classes, which benefits from both the reduction in dataset perplexity (by allowing less non-target classes) and from the use of more discrimination information learned. The method combines the strengths of the two feature extractors used, as auDeep is more adaptable, while the SoundNet is more discriminative, and demonstrates a **91.6%** accuracy, which qualifies as state-of-the-art performance.

5. Conclusions and future work

In this work, a classification method was proposed for cough sound recognition. Two DNN-based feature extractors were used: the auDeep [26], which learns from normalized log-mel-spectrograms of input audio, and the SoundNet [25], fine-tunable, which works on raw waveforms. Based on their concatenated features, extracted from a corpus assembled for maximized discriminative power, a MLP-type classifier was trained on a half and tested on the other half of the corpus. The accuracy figures in Table 4, obtained on the most non-cough corpus (IRS17) as well as on the most cough-like corpus (SHC), show that the proposed method is well suited for cough classification. On the other hand, classification accuracies in Tables 1 to Table 3 show that the method may be expanded progressively to other unseen patient sounds, in case of respiratory disease epidemics.

The cough recognition accuracy of the proposed method overcomes the classification accuracy attained on corpus ESC-50 by the proponents of auDeep and SoundNet. No comparison was made to other systems, as most published research is based on scarce data, and offer results which are not comparable.

The feature extractors and the final classifier can be retrained and fine-tuned on real world data, with an extensible number of classes. Classification performance was observed to depend on separability of sound sources in audio signals, which suggests the use of several audio recording channels, especially in cough sound monitoring methods.

The real-world data corpora are made of all available data, except for some ground truth cases, left aside for method validation purposes. Problems with method validation are stated in most published literature because of the lack of data, which made the Leave One Out Cross-Validation (LOOCV) one of most often used validation methods. Performance measures obtained from cross-validation, especially those used for parameter tuning and model selection, can introduce high variance thus making the models unreliable. In epidemics, classifier validation must be based more on medical diagnostic data, and so the resolution of the problem must be postponed after the deployment stage.

As a future work, we intend to simultaneously train the feature extractors to extract the most discriminative features for our classification task instead of separate stages, and standard MLP classifiers. Knowledge transfer from stills was also considered, not yet implemented, from images such as thermal forefront images, chest X-ray films, as well as 3D head and chest models from Computerized Tomography (CT) or Magnetic Resonance Imaging (MRI) scans.

Acknowledgement

This work was partly supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI – UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818 / 73PCCDI, within PNCDI III.

References

[1] World Health Organization, *Naming the coronavirus disease (COVID-19) and the virus that causes it*, February 28, 2020.

[2] Centers for Disease Control and Prevention, *Symptoms of Coronavirus*, U.S.A., February 10, 2020.

[3] World Health Organization, *Coronavirus disease 2019 (COVID-19) Situation Report – 73*, April 2, 2020.

[4] National Institutes of Health, *New coronavirus stable for hours on surfaces*, March 17, 2020.

[5] — *Coronavirus in the top 10 worst epidemics in the last 50 years*, Atlas Magazine – Insurance Around the World, https://www.atlas-mag.net/en/article/coronavirus-in-the-top-10-worst-epidemics-in-the-last-50-years, April 2020.

[6] AMOH J. and ODAME K., *DeepCough: a deep convolutional neural network in a wearable cough detection system*, IEEE Biomedical Circuits and Systems Conference (BioCAS), pp. 1-4, 2015.

[7] AMOH J. and ODAME K., *Deep neural networks for identifying cough sounds*, IEEE Transactions on Biomedical Circuits and Systems, **vol. 10** (2016), no. 5, pp. 1003-1011.

[8] NEMATI E., RAHMAN M., NATHAN V., and KUANG J., *Private audio-based cough sensing for in-home pulmonary assessment using mobile devices*, EAI International Conference on Body Area Networks, Springer, Cham, pp. 221-232, 2018.

[9] Di PERNA L., SPINA G., THACKRAY-NOCERA S., CROOKS M.G., MORICE A.H., SODA P., and Den BRINKER A.C., *An automated and unobtrusive system for cough detection*, IEEE Life Sciences Conference (LSC), Sydney, Australia, pp. 190-193, 2017.

[10] MONGE-ÁLVAREZ J., HOYOS-BARCELÓ C., LESSO P., and CASASECA-de-la-HIGUERA P., *Robust detection of audio-cough events using local Hu moments*, IEEE Journal of Biomedical and Health Informatics, **vol. 23** (2018), no. 1, pp. 184-196.

[11] MONGE-ÁLVAREZ J., HOYOS-BARCELÓ C., SAN-JOSÉ-REVUELTA L.M., and CASASECA-de-la-HIGUERA P., *A machine hearing system for robust cough detection based on a high-level representation of band-specific audio features*, IEEE Transactions on Biomedical Engineering, **vol. 66** (2018), no. 8, pp. 2319-2330.

[12] MONGE-ÁLVAREZ J., HOYOS-BARCELÓ C., DAHAL K., and CASASECA-de-la-HIGUERA P., *Audio-cough event detection based on moment theory*, Applied Acoustics, **vol. 135**, pp. 124-135, 2018.

[13] PRAMONO R.X.A., IMTIAZ S.A., and RODRIGUEZ-VILLEGAS E., *Automatic cough detection in acoustic signal using spectral features*, The 41st International Conference of the IEEE Engineering in Medicine and Biology (EMB), Berlin, Germany, pp. 7153-7156, 2019.

[14] BARATA F., KIPFER K., WEBER M., TINSCHERT P., FLEISCH E., and KOWATSCH T., *Towards device-agnostic mobile cough detection with convolutional neural networks*, IEEE International Conference on Healthcare Informatics (ICHI), Oldenburg, Germany, pp. 1-11, 2019.

[15] KVAPILOVA L., BOZA V., DUBEC P., MAJERNIK M., BOGAR J., JAMISON J. et. al, *Continuous sound collection using smartphones and machine learning to measure cough*, Digital Biomarkers, **vol. 3** (2019), no. 3, pp. 166-175.

[16] BALES C., JOHN C., FAROOQ H., MASOOD U., NABEEL M., and IMRAN A., *Can machine learning be used to recognize and diagnose coughs?*, arXiv Preprint, 2020.

[17] MIRANDA I.D.S., DIACON A.H., and NIESLER T.R., *A comparative study of features for acoustic cough detection using deep architectures*, The 41st International Conference of the IEEE Engineering in Medicine and Biology (EMB), Berlin, Germany, pp. 2601-2605, 2019.

[18] KHOMSAY S., VANIJJIRATTIKHAN R., and SUWATTHIKUL J., *Cough detection using PCA and deep learning*, International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, pp. 101-106, 2019.

[19] KADAMBI P., MOHANTY A., REN H., SMITH J., McGUINNESS K., HOLT K. et al. *Towards a wearable cough detector based on neural networks*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Alberta, Canada, pp. 2161-2165, 2018.

[20] PRAMONO R.X.A., BOWYER S., and RODRIGUEZ-VILLEGAS E., *Automatic adventitious respiratory sound analysis – a systematic review*, in PLOS One, T. Penzel (Editor), May 26, 2017.

[21] PICZAK K., *ESC: dataset for environmental sound classification*, The 23-rd ACM International Conference on Multimedia, Brisbane, Australia, pp. 1015-1018, 2015.

[22] ROCHA B.M., FILOS D., MENDES L., VOGIATZIS I., PERANTONI E., KAIMAKAMIS E. et al., *A respiratory sound database for the development of automated classification*, Precision Medicine, Powered by pHealth and Connected Health, Springer, Singapore, pp. 51-55, 2018.

[23] DOGARIU M., CUCU H., BUZO A., BURILEANU D., and FRATU O., *Speech database acquisition for assisted living environment applications*, The 8-th International Conference on Speech Technologies and Human-Computer Dialogue (SpeD), Bucharest, Romania, pp. 191-196, 2015.

[24] VACHER M., ISTRATE D., PORTET F., JOUBERT T., CHEVALIER T., SMIDTAS S. et al., *The SWEET-HOME project: audio technology in smart homes to improve well-being and reliance*, French National Research Agency (Agence Nationale de la Recherche/ANR09–VERS-011).

[25] AYTAR Y., VONDRICK C., and TORRALBA A., *SoundNet: learning sound representations from unlabeled video*, The 29th Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, pp. 1-9, 2016.

[26] AMIRIPARIAN S., FREITAG M., CUMMINS N., and SCHULLER B., *Sequence to sequence autoencoders for unsupervised representation learning from audio*, Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, Munich, Germany, pp. 17-21, 2017.

[27] HU H.-N., *TensorFlow implementation of «SoundNet» that learns rich natural sound representations*, https://github.com/eborboihuc/SoundNet-tensorflow.