Wake-Cough: cough spotting and cougher identification for personalised long-term cough monitoring

Madhurananda Pahar[1], Marisa Klopper[2], Byron Reeve[2], Rob Warren[2], Grant Theron[2], Andreas Diacon[3] and Thomas Niesler[1]

[1]*Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa*
[2]*Division of Molecular Biology and Human Genetics, Stellenbosch University, South Africa*
[3]*TASK Applied Science, Cape Town, South Africa*

Email: {mpahar, marisat, byronreeve, rw1, gtheron, ahd, trn}@sun.ac.za

***Abstract*—We present 'wake-cough', an application of wake-word spotting to coughs using a Resnet50 and the identification of coughers using i-vectors, for the purpose of a long-term, personalised cough monitoring system. Coughs, recorded in a quiet (73±5 dB) and noisy (34±17 dB) environment, were used to extract i-vectors, x-vectors and d-vectors, used as features to the classifiers. The system achieves 90.02% accuracy when using an MLP to discriminate between 51 coughers using 2-sec long cough segments in the noisy environment. When discriminating between 5 and 14 coughers using longer (100 sec) segments in the quiet environment, this accuracy improves to 99.78% and 98.39% respectively. Unlike speech, i-vectors outperform x-vectors and d-vectors in identifying coughers. These coughs were added as an extra class to the Google Speech Commands dataset and features were extracted by preserving the end-to-end time-domain information in a trigger phrase. The highest accuracy of 88.58% is achieved in spotting coughs among 35 other trigger phrases using a Resnet50. Thus, wake-cough represents a personalised, non-intrusive cough monitoring system, which is power-efficient as on-device wake-word detection can keep a smartphone-based monitoring device mostly dormant. This makes wake-cough extremely attractive in multi-bed ward environments to monitor patients' long-term recovery from lung ailments such as tuberculosis (TB) and COVID-19.**

I. INTRODUCTION

Wake-words are used as trigger phrases which enable keyword spotting systems to initiate certain tasks such as speech recognition by continuously listening for specific keywords using low computational power [1]. This is the first important step between the user and the processing units on either the device or the cloud server [2] and both the near and far field wake-word detection requires to be highly sensitive in both quiet and noisy environments for better performance [3]. For example, some widely-used trigger phrases for voice assistants on smart devices are: Google's 'OK Google', Apple's 'Hey Siri', Amazon's 'Alexa' and Microsoft's 'Hey Cortana' [4]. These algorithms are highly sensitive in both quiet and noisy environments [3], making them extremely useful in hands-free situations like driving [5]. Coughing is the forceful expulsion of air to clear the airway and a common symptom of respiratory diseases, such as tuberculosis (TB) [6], asthma [7], pertussis [8] and COVID-19 [9], [10], which can be identified using machine learning classifiers. To successfully implement cough as a personalised wake-word in commercial smartphones, it is necessary to accurately identify the cougher [11] in both noisy and quiet environments and the cough among various other commonly used trigger phrases [12].

Vocal audio such as speech can be identified using i-vectors, which present a low-dimensional speaker and channel-dependant space using factor analysis proposing a speaker representation system for speaker identification [13]. The performance can be improved by using x-vectors [14] and d-vectors [15], which use the data augmentation and DNN based embeddings to map speaker embeddings.

Coughers have been identified using x-vectors on natural coughs in an open world environment for 8 male and 8 female subjects after implementing data augmentation to address the effect of background noise [16] and using d-vectors on forced coughs [17]. Here, we identify both natural and forced coughs among other trigger phrases in the Google Speech commands dataset [18] while also identifying the coughers in noisy and quiet environments using i-vectors, x-vectors and d-vectors. To accurately monitor the long-term cough rates, for example in a multi-bed ward, automatic detection of coughs among other environmental noises and classification of coughers while consuming less power and preserving privacy is extremely important. By detecting coughs among other wake-words and classifying coughers using i-vectors, wake-cough represents a personalised long-term cough monitoring system. This system is also power-efficient as specialised algorithms work on the

This project was funded by the South African Medical Research Council (SAMRC) through its Division of Research Capacity Development under the SAMRC Intramural Postdoctoral programme, the South African National Treasury, as well as an EDCTP2 programme supported by the European Union (grant TMA2017CDF-1885, grant SF1401, OPTIMAL DIAGNOSIS; grant RIA2020I-3305, CAGE-TB) and the National Institute of Allergy and Infection Diseases of the National Institutes of Health (U01AI152087). We also thank the South African Centre for High Performance Computing (CHPC) for providing computational resources on their Lengau cluster for this research, and gratefully acknowledge the support of Telkom, South Africa.

device without needing any cloud service.

II. Dataset Preparation

For the cougher identification task, two datasets which will be referred to as TASK and Wallacedene (Table I), were both manually annotated using ELAN [19]. The TASK dataset, which contains natural coughs, was collected at a TB research hospital in Cape Town, South Africa (TASK clinical trial centre). This research hospital accommodates up to 24 patients in six 4-bed wards [20], [21]. A plastic enclosure, attached to the bed-frames, holds a Samsung Galaxy J4 smartphone connected to a BOYA BY-MM1 cardioid microphone (Figure 1) and the distance between the cougher and the microphone was between 30 and 150 cm. The dataset includes 6000 cough events, sampled at 22.05 kHz and collected from 14 adult male patients over a 6 month period, totalling 3.16 hours of cough audio with an average SNR of 73±5 dB. No other information of the patients was collected due to ethical constraints. Wallacedene dataset was collected inside an outdoor booth next to a busy primary health clinic in Wallacedene, near Cape Town, South Africa representing a real-world environment where a typical TB test would likely to be deployed [22] (Figure 1). Patients were asked to count from 1 to 10, then cough, take a few deep breaths, and cough again, thus producing a bout of forced coughs. These counts were used as speech to provide a baseline to compare the performance of cougher identification in Table IV. The audio, sampled at 44.1 kHz, was recorded using a RØDE M3 condenser microphone from 38 males and 13 females, keeping a 10 to 15 cm gap between the microphone and the patients. Environmental noise was present in both cough and speech recordings, which had an average SNR of 34 dB and 33 dB respectively with a standard deviation of 17 dB (Table I).

Table I shows that the TASK dataset is less-noisy and contains much longer cough audio for each subject, whereas the Wallacedene dataset is noisier but contains both cough and speech audio from a larger number of subjects. All audio recordings were downsampled to 16 kHz, as required for the Kaldi ASR system [23].

TABLE I
Data used in cougher & speaker identification: The TASK dataset is less-noisy than the Wallacedene dataset.

Dataset	Subjects	Events	Avg SNR	Avg Length
Cougher identification				
TASK	14	6000	73±5 dB	1.87±0.2 sec
Wallacedene	51	1358	34±17 dB	0.77±0.1 sec
Speaker identification				
Wallacedene	51	510	33±17 dB	0.99±0.2 sec

For cough spotting, we randomly selected 3795 coughs from the TASK and Wallacedene datasets. Each cough was normalised to a 1-sec duration by either trimming or padding with silence. These 'cough' events were added as an extra class to the 2nd version of Google Speech Commands dataset, which contains a total of 109,624 1-sec long events, sampled at 16 kHz and belonging to 35 classes [18]. These events were mixed with the background noises (Section 5.8 of [18]) with a randomly selected SNR between 73 and 34 dB (Table I). A subset of this dataset, with only 42,341 events belonging to 10 classes, is also available for use as commands in IoT or robotics [18]. For spotting cough as a trigger phrase, we note these two datasets as SC-36 and SC-11, containing 36 and 11 classes respectively.

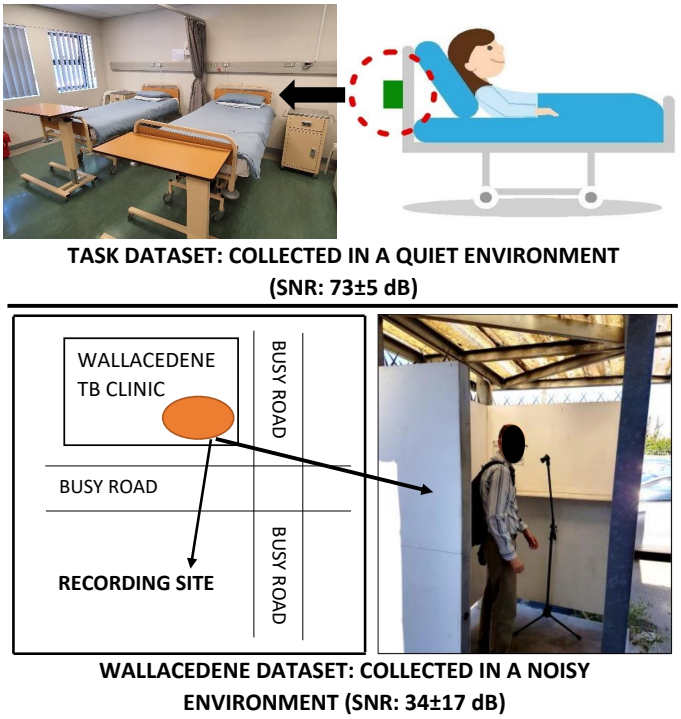

Fig. 1. **Data collection process for cougher identification:** The TASK dataset, containing only coughs, was collected in a quiet environment. The Wallacedene dataset, containing both cough and speech (counting from 1 to 10), was collected in a noisy environment.

III. Feature Extraction

For cougher identification, we have extracted x-vectors and i-vectors using extractors pre-trained on the under-resourced languages [24], which are spoken by the subjects in the TASK and Wallacedene datasets (Figure 2). Audio segments that are t-sec long from each of N coughers are concatenated by following the data preparation requirements of Kaldi ASR toolkit [23]. For each non-overlapping 0.1 sec audio, i-vectors are generated from each utterance ID, with a dimension of $(t \times 10, 100)$ for each cougher [13]. Unique x-vectors are generated for each 1.5 sec of utterance with a 0.75 sec overlap, having a dimension of $(1, 512)$ [14]. Thus for each t-sec long audio from each cougher, there are x-vectors of dimension $(\frac{t}{0.75}, 512)$. We have also extracted d-vectors using an extractor pre-trained on VCC 2018, VCTK, LibriSpeech, and CommonVoice English datasets and were generalized using the end-to-end loss function [15]. Every t sec cough is split into non-overlapping 0.5 sec segments, thus producing d-vectors of dimension $(\frac{t}{0.5}, 256)$ for every cougher and suggesting that

the i-vectors have a higher dimensionality than x-vectors and d-vectors. The number of subjects (N) and the cough-time (t) were the hyperparameters in cougher identification task (Table III). For speakers, we used all counts, having only N as a hyperparameter. For the TASK and Wallacedene datasets, N has been varied between 5 & 14 and 5 & 51 respectively in steps of 5.

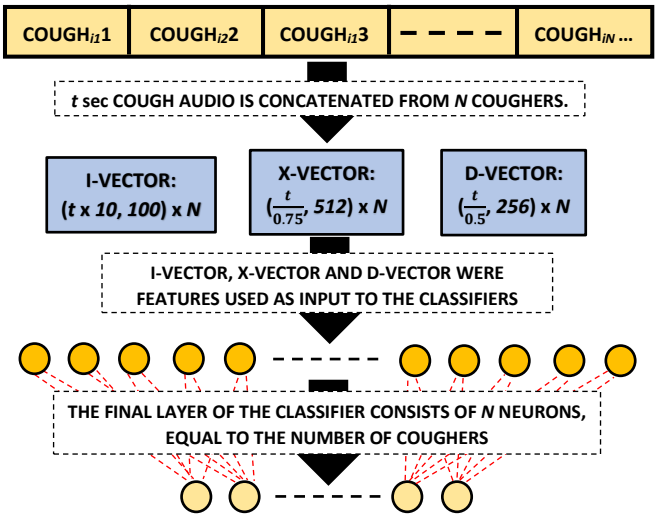

Fig. 2. **Feature extraction for cougher identification:** t-sec long cough segments (COUGH_{i1}, COUGH_{i2}, COUGH_{i3}, ..., where, $1 \leq i \leq N$) from each cougher are concatenated as they appear in the audio recording for N coughers. i-vectors, x-vectors and d-vectors are extracted from this $t \times N$-sec long audio and presented to the classifiers, which have N neurons in the final layer to distinguish the cougher using a cross-validation scheme.

For spotting cough as a trigger phrase, we have extracted STFT, ZCR and kurtosis from overlapping frames ($\mathcal{F}$) of the audio, where the frame overlap is computed to ensure that the audio signal is always divided into exactly $\mathcal{S}$ frames, so that the entire audio event is always captured within a fixed number of frames, allowing a fixed input dimension to be maintained while preserving the general overall temporal structure of the event. Such fixed two-dimensional features are particularly useful for the training of DNN classifiers [9]. Table II shows that in our experiments each audio signal is divided into between 70 and 150 frames, each between 512 and 4096 samples i.e. 32 msec and 256 msec long, thus varying the spectral information extracted from each event in the SC-11 and SC-36 datasets.

LR, LDA, SVM and MLP classifiers were used to identify coughers and CNN, LSTM and Resnet50 were used to spot coughs as a trigger phrase. Table III lists the hyperparameters considered and the ranges considered during the 5-fold cross-validation. The standard deviation among the outer folds is noted as σ_{ACC} in Table IV. For Resnet50, the 50-layer architecture described in [25] has been used.

IV. Results and Discussion

Table IV shows the results using the best two features for both TASK (less-noisy) and Wallacedene (noisier) datasets.

TABLE II
Feature extraction hyperparameters. Table IV and V show classification results for these hyperparameters.

Hyperparameter	Description	Range
Cougher identification		
Subject (N)	no. of coughers or speakers	5 to 51 with step of 5
Cough-time (t)	cough from each subject	2, 5 to 100 with step of 5
Cough spotting		
Frame length ($\mathcal{F}$)	used to extract features	2^k, $k = 9, \ldots 12$
No. of frames ($\mathcal{S}$)	extracted from audio	$10 \times k$, $k = 7, 10, 12, 15$

TABLE III
Classifier hyperparameters used in both identifying 'coughers' and spotting 'cough'.

	Hyperparameters	Classifier	Range
coughers	Regularisation	LR & SVM	10^i where $i = -7, \ldots 7$
	$l1$ penalty	LR	0 to 1 in steps of 0.05
	$l2$ penalty	LR, MLP	0 to 1 in steps of 0.05
	Kernel coeff.	SVM	10^i where $i = -7, \ldots 7$
	No. of neurons	MLP	70 to 150 in steps of 20
cough	Batch size	CNN & LSTM	2^k where $k = 6, 7, 8$
	No. of epochs	CNN & LSTM	10 to 200 in steps of 20
	No. of conv filters	CNN	3×2^k where $k = 3, 4, 5$
	kernel size	CNN	2 and 3
	Dropout rate	CNN & LSTM	0.1 to 0.5 in steps of 0.2
	Dense layer size	CNN & LSTM	2^k where $k = 4, 5$
	LSTM units	LSTM	2^k where $k = 6, 7, 8$
	Learning rate	LSTM	10^k where $k = -2, -3, -4$

The highest accuracy (99.78%) has been achieved by an MLP when using i-vectors to identify coughers from 100-sec ($t = 100$) long cough collected from each of 5 coughers. By increasing the number of coughers to 10 and 14, the performance of the MLP classifier decreased to 98.87% and 98.39% respectively for i-vectors (Table IV and Figure 4).

All classifiers performed well in identifying both coughers and speakers on the noisier the Wallacedene dataset. The speaker identification is used as the baseline and Table IV shows that using x-vectors produced better classification scores

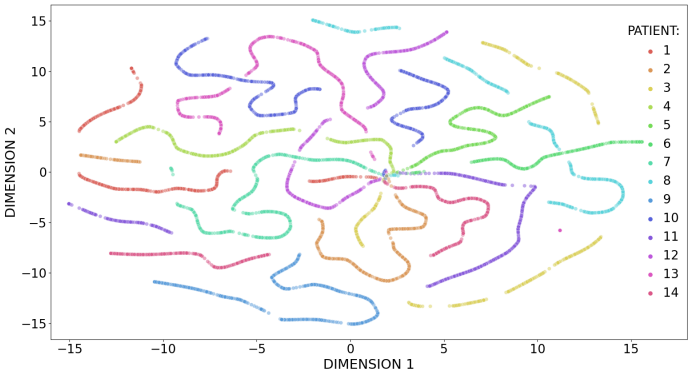

Fig. 3. **The t-SNE cluster of i-vectors extracted from 2-sec long cough audio from 14 coughers in TASK dataset**. The MLP produces 95.11% accuracy using these i-vectors in discriminating 14 coughers (Table IV).

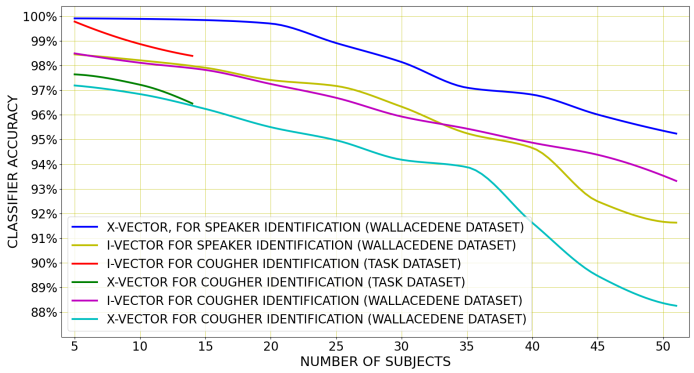

Fig. 4. **Classifier performance.** The accuracies from the MLP classifier decrease while discriminating more subjects (Table IV).

than using i-vectors for speaker identification, as also found by others [14]. The highest accuracy (99.91%) has been achieved using the MLP and x-vectors while discriminating among only 5 speakers. This accuracy drops to 98.14% using MLP while differentiating between 30 speakers and to 95.24% when discriminating among all 51 speakers in the Wallacedene dataset. For a smaller number of coughers, such as 5, the MLP classifier has achieved the highest accuracy of 98.49% using i-vectors. As the number of coughers is increased to 15, 25, 40 and 51, the accuracy of the MLP has dropped to 97.82%, 96.69%, 94.87% and 93.32% respectively and the σ_{ACC} has increased sharply. These scores show that although cougher identification is not as accurate as speaker identification, the performance is close, especially for a small number of subjects.

The results also show that, unsurprisingly, cougher identification on the less-noisy TASK dataset is more accurate than the noisier Wallacedene dataset. Although longer coughs from each subject improve the classifier accuracy in general, similar performance is achieved (accuracies of 95.11% & 90.02% on the less-noisy & the noisy data) for coughs as short as only 2 sec (Figure 3). Although the performance is close, i-vectors performed better than x-vectors in cougher identification. The MLP is the classifier of choice as it shows a lower σ_{ACC} across the cross-validation folds for the less-noisy data than noisier data. d-vectors are outperformed by i-vectors and x-vectors for both speech and cough, as also found by [26], and thus excluded from Table IV.

Coughs were successfully spotted among other trigger phrases in both the SC-11 and the SC-36 dataset. Table V shows that although LSTM and CNN have performed well, the best performance of 92.73% accuracy ($\mathcal{ACC}$) & mean Cohen's Kappa ($\mathcal{K}$) of 0.9218 on SC-11 and 88.58% accuracy & $\mathcal{K}$ of 0.8757 on SC-36 have been achieved using a Resnet50. The confusion matrix of the best SC-11 system exhibits an excellent performance for spotting coughs among the other trigger phrases in Figure 5. Table V also shows that the best CNN and Resnet50 results were obtained mostly when using 1024 sample (64 msec) long frames and 100 segments.

TABLE IV
CLASSIFIER ACCURACIES IN IDENTIFYING COUGHERS FOR BOTH TASK AND WALLACEDENE (WD) DATASETS.

Dataset	N	t	Feature	LR	LDA	SVM	MLP	σ_{ACC}
TASK	5	100	i-vector	98.91%	98.87%	99.44%	*99.78%*	0.0007
		100	x-vector	96.71%	96.73%	97.54%	97.64%	0.0009
	10	80	i-vector	97.54%	97.88%	98.19%	*98.87%*	0.0006
		80	x-vector	96.31%	96.24%	96.55%	97.22%	0.0005
	14	2	i-vector	94.41%	94.51%	94.55%	**95.11%**	0.0005
		100	i-vector	96.46%	96.71%	97.48%	*98.39%*	0.0006
		100	x-vector	97.26%	97.54%	98.78%	96.46%	0.0008
WD (Cougher)	5	20	i-vector	97.23%	97.19%	97.77%	*98.49%*	0.0054
		20	x-vector	95.54%	95.97%	96.72%	97.19%	0.0078
	15	20	i-vector	97.16%	97.14%	97.31%	*97.82%*	0.0061
		20	x-vector	95.41%	95.30%	95.72%	96.24%	0.0068
	25	20	i-vector	95.04%	95.18%	95.94%	*96.69%*	0.0072
		20	x-vector	93.31%	93.55%	94.07%	94.97%	0.0082
	40	20	i-vector	93.38%	93.62%	94.09%	*94.87%*	0.0091
		20	x-vector	90.23%	90.07%	90.97%	91.62%	0.0102
	51	2	i-vector	89.26%	89.38%	89.22%	**90.02%**	0.0178
		20	i-vector	90.27%	90.49%	91.89%	*93.32%*	0.0301
		20	x-vector	84.61%	84.74%	85.83%	88.26%	0.0247
WD (Speaker)	5	—	x-vector	98.57%	98.64%	99.48%	*99.91%*	0.0018
		—	i-vector	97.21%	97.17%	97.70%	98.45%	0.0027
	30	—	x-vector	96.81%	96.85%	97.42%	*98.14%*	0.0081
		—	i-vector	94.81%	94.87%	95.18%	96.33%	0.0078
	51	—	x-vector	99.44%	99.44%	99.44%	*95.24%*	0.0229
		—	i-vector	90.01%	90.05%	90.34%	91.63%	0.0274

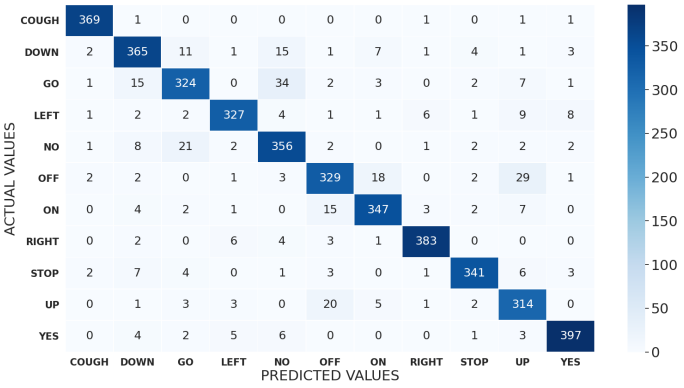

Fig. 5. **The confusion matrix** of detecting coughs among 10 other trigger phrases in SC-11 dataset using the best Resnet50 classifier in Table V.

TABLE V
COUGH SPOTTING: THE BEST-THREE RESULTS FOR EACH CLASSIFIER SHOWS RESNET50 HAS PERFORMED THE BEST BY ACHIEVING 92.73% & 88.58% ACCURACY ON THE SC-11 & SC-36 DATASET.

Classifier	SC-11 Dataset				SC-36 Dataset			
	$\mathcal{F}$	$\mathcal{S}$	$\mathcal{ACC}$	$\mathcal{K}$	$\mathcal{F}$	$\mathcal{S}$	$\mathcal{ACC}$	$\mathcal{K}$
LSTM	*512*	*150*	*88.09%*	*0.8767*	*512*	*120*	*80.74%*	*0.7937*
	2048	120	87.66%	0.8614	1024	120	80.40%	0.7931
	512	70	87.09%	0.8598	512	100	80.11%	0.7902
CNN	*1024*	*100*	*91.25%*	*0.9007*	*1024*	*120*	*86.74%*	*0.8592*
	2048	100	90.72%	0.8981	1024	70	85.98%	0.8463
	1024	70	90.11%	0.8945	2048	100	85.22%	0.8411
Resnet50	***1024***	***100***	***92.73%***	***0.9218***	***2048***	***100***	***88.58%***	***0.8777***
	2048	120	92.69%	0.8733	2048	70	87.94%	0.8729
	2048	100	92.55%	0.8715	1024	120	87.68%	0.8702

V. CONCLUSION

We propose a system using cough as a wake-word to spot coughs among other trigger phrases and identify the cougher.

A less-noisy and noisier dataset, containing 14 and 51 subjects respectively, were used to extract i-vectors, x-vectors and d-vectors, to classify the cougher. The best performance was achieved using an MLP, showing coughers as many as 51 can be distinguished from one another with 90.02% accuracy using i-vectors from as short as 2-sec long audio from each cougher in the noisy environment. We also found that, unlike speakers, coughers were better identifiable using i-vectors. Coughs were also spotted as wake-words using a Resnet50 on features keeping end-to-end time-domain information among 35 other keywords in the Google Speech Commands dataset with 88.58% accuracy. Wake-cough represents a means of personalised, long-term cough monitoring system that is able to discriminate between coughers, non-intrusive and, due to the use of wake-word detection methods, power-efficient since a smartphone-based monitoring device can remain mostly dormant. Thus, it is an attractive and viable means for monitoring a patient's long-term recovery from lung ailments such as TB and COVID-19 in multi-bed ward environments.

In our future work, we aim to include more recent architectures and extend the dataset to investigate wake-cough's performance across age, gender etc. of the subjects and compare it with metric learning-based cougher identification [27].

REFERENCES

[1] Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope, "Your Word is my Command": Google search by Voice: A Case Study," in *Advances in Speech Recognition*, pp. 61–90. Springer, 2010.

[2] Minhua Wu, Sankaran Panchapagesan, Ming Sun, Jiacheng Gu, Ryan Thomas, Shiv Naga Prasad Vitaladevuni, Bjorn Hoffmeister, and Arindam Mandal, "Monophone-Based Background Modeling for Two-Stage On-Device Wake Word Detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5494–5498.

[3] Yixin Gao, Yuriy Mishchenko, Anish Shah, Spyros Matsoukas, and Shiv Vitaladevuni, "Towards Data-Efficient Modeling for Wake Word Spotting," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7479–7483.

[4] Tara Sainath and Carolina Parada, "Convolutional Neural Networks for Small-footprint Keyword Spotting," in *INTERSPEECH*, 2015, pp. 1478–1482.

[5] Guoguo Chen, Carolina Parada, and Georg Heigold, "Small-footprint keyword spotting using deep neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4087–4091.

[6] Renier Botha, Grant Theron, Robbin Warren, Marisa Klopper, Keertan Dheda, Paul Van Helden, and Thomas Niesler, "Detection of tuberculosis by automatic cough sound analysis," *Physiological Measurement*, vol. 39, no. 4, pp. 045005, 2018.

[7] Mahmood Al-khassaweneh and Ra'ed Bani Abdelrahman, "A signal processing approach for the diagnosis of asthma from cough sounds," *Journal of Medical Engineering & Technology*, vol. 37, no. 3, pp. 165–171, 2013.

[8] Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas, "A cough-based algorithm for automatic diagnosis of pertussis," *PLOS ONE*, vol. 11, no. 9, pp. e0162128, 2016.

[9] Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas Niesler, "COVID-19 cough classification using machine learning and global smartphone recordings," *Computers in Biology and Medicine*, vol. 135, pp. 104572, 2021.

[10] Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas Niesler, "COVID-19 detection in cough, breath and speech using deep transfer learning and bottleneck features," *Computers in Biology and Medicine*, vol. 141, pp. 105153, 2022.

[11] Fengpei Ge and Yonghong Yan, "Deep neural network based wake-up-word speech recognition with two-stage detection," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2761–2765.

[12] Veton Z Këpuska and TB Klein, "A novel Wake-Up-Word speech recognition system, Wake-Up-Word recognition task, technology and evaluation," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 71, no. 12, pp. e2772–e2789, 2009.

[13] Andrew Senior and Ignacio Lopez-Moreno, "Improving DNN speaker independence with i-vector inputs," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 225–229.

[14] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[15] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized End-to-End Loss for Speaker Verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[16] Matt Whitehill, Jake Garrison, and Shwetak Patel, "Whosecough: In-the-Wild Cougher Verification Using Multitask Learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 896–900.

[17] Miao Zhang, Yixiang Chen, Lantian Li, and Dong Wang, "Speaker recognition with cough, laugh and "Wei"," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 497–501.

[18] Pete Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[19] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes, "ELAN: a professional framework for multimodality research," in *5th International Conference on Language Resources and Evaluation (LREC 2006)*, 2006.

[20] Madhurananda Pahar, Igor Miranda, Andreas Diacon, and Thomas Niesler, "Deep Neural Network based Cough Detection using Bed-mounted Accelerometer Measurements," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8002–8006.

[21] Madhurananda Pahar, Igor Miranda, Andreas Diacon, and Thomas Niesler, "Automatic Non-Invasive Cough Detection based on Accelerometer and Audio Signals," *Journal of Signal Processing Systems*, pp. 1–15, 2022.

[22] Madhurananda Pahar, Marisa Klopper, Byron Reeve, Rob Warren, Grant Theron, and Thomas Niesler, "Automatic cough classification for tuberculosis screening in a real-world environment," *Physiological Measurement*, vol. 42, no. 10, pp. 105014, oct 2021.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011, IEEE Catalog No.: CFP11SRW-USB.

[24] Trideba Padhi, Astik Biswas, Febe de Wet, Ewald van der Westhuizen, and Thomas Niesler, "Multilingual bottleneck features for improving ASR performance of code-switched speech in under-resourced languages," in *Proceedings of the First Workshop on Speech Technologies for Code-switching in Multilingual Communities (WSTCSMC)*, Shanghai, China, 2020.

[25] 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.

[26] Lantian Li, Yiye Lin, Zhiyong Zhang, and Dong Wang, "Improved deep speaker feature learning for text-dependent speaker recognition," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015, pp. 426–429.

[27] Stefan Jokic, David Cleres, Frank Rassouli, Claudia Steurer-Stey, Milo A. Puhan, Martin Brutsche, Elgar Fleisch, and Filipe Barata, "TripletCough: Cougher Identification and Verification from Contact-Free Smartphone-Based Audio Recordings Using Metric Learning," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2022.