CONVOLUTIONAL DENSE NEURAL NETWORK BASED SPIROMETRY VARIABLE FVC PREDICTION USING SUSTAINED PHONATIONS

Shivani Yadav[1], Dipanjan Gope[2], Uma Maheswari Krishnaswamy[3], Prasanta Kumar Ghosh[4]

[1]BioSystems Science and Engineering, Indian Institute of Science (IISc), Bangalore-560012, India
[2]Electrical Communication Engineering, Indian Institute of Science (IISc), Bangalore-560012, India
[3]St. Johns National Academy of Health Sciences, Bangalore-560034, India
[4]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India

2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP) | 978-1-7281-6338-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/MLSP52302.2021.9596159

ABSTRACT

Spirometry is a lung function test used to diagnose and monitor lung diseases like asthma, pneumonia, chronic obstructive pulmonary disease, etc. Spirometry measures forced vital capacity (FVC), forced expiratory volume in 1 sec (FEV1), and their ratio to determine lung health. Spirometry is very time-consuming, strenuous, and requires proper training. Alternate methods based on voice for diagnosis and monitoring of lung health are promising because they are faster, easy to do, and require minimal training. Non-speech sounds, namely, cough and wheeze, have been used to predict spirometry variables, but the role of speech sounds that occur in natural speaking for a similar task has not been explored. In this work, the spirometry variable, FVC has been predicted from sustained phonations of vowel sounds using a convolutional dense neural network (CDNN). Mel-spectrogram has been used as a feature. An experiment conducted using 160 subjects indicates, /iː/ is the best sound and /uː/ is worst for the prediction task with an average Mean Absolute Error of $0.67l(\pm .07l)$ and $0.70l(\pm 0.13l)$ among all sustained phonations of vowels sounds considered in this work.

Index Terms— Asthma, sustained phonations, CDNN, Spirometry

1. INTRODUCTION

544.9 million people were suffering worldwide from chronic respiratory diseases [1] till 2017. Spirometry is a lung function test used to measure lung capacity to monitor and diagnose obstructive lung diseases like asthma [2], and Chronic obstructive lung diseases (COPD) [3]. Spirometry measures parameters namely, Forced vital capacity (FVC), Forced expiratory volume in 1 sec (FEV1), FEV1/FVC, vital capacity, peak expiratory flow, mid expiratory flow at 25%, 50% and 75% FVC and inspiratory vital capacity [4]. To diagnose obstructive lung diseases and grading their severity, FEV1/FVC, FVC, and FEV1 are used. FVC (l) indicates the volume of air exhaled forcefully after deep inhalation, and FEV1 (ls^{-1}) denotes the volume of air forcefully exhaled in 1 sec after a deep inhalation. Based on the height, age, weight, and gender of a person, reference values of FEV1, FVC, and FEV1/FVC ratio are predicted. During test, subject has to take deep inhalation followed by forced exhalation for at least 6 seconds into the spirometer sensor. Throughout the test, nose of the subject is closed with the nose clip to get accurate readings. The effort-dependent nature of spirometry makes it unsuitable for kids and older people. Based on the European Respiratory Society (ERS) and American Thoracic Society (ATS) [5] regulations, a subject has to meet multiple criteria while doing the test like the start of test criterion, end of test criterion, acceptability criterion, minimum of three repeatabilities of the test, etc., to obtain accurate readings of spirometry variables. Schermer et al. [6] have reported 50% of tests are rejected due to incomplete tests which leads to risks of misdiagnosis and wrong treatment. To estimate FVC value accurately, a subject should meet the end of test criteria, which is very strenuous, requires multiple test repetitions, and induce fatigue, especially in subjects with compromised lung functions [7]. Vocal sound-based method can be used as a helping hand of the spirometry.

Few works have been done previously to predict the spirometry variables using cough and wheeze sounds. Cough is produced by the sudden release of pressure by the glottis opening. The obstruction produces wheeze sounds in the airways during obstructive diseases like COPD and asthma. Rao et al. [8] have used statistical spectrum descriptor of cough and wheeze sound to predict spirometry variables in asthmatic and healthy subjects using support vector regressor (SVR). Sharan et al. [9] have predicted FEV1, FVC, and FEV1/FVC by using cough sounds recorded at mouth using mobile phones from 322 subjects. The authors have used bispectrum scores, non-gaussianity score, formants, log energy, Shannon's entropy, zero-crossing rate, kurtosis, and

978-1-7281-6338-3/21/$31.00 ©2021 IEEE

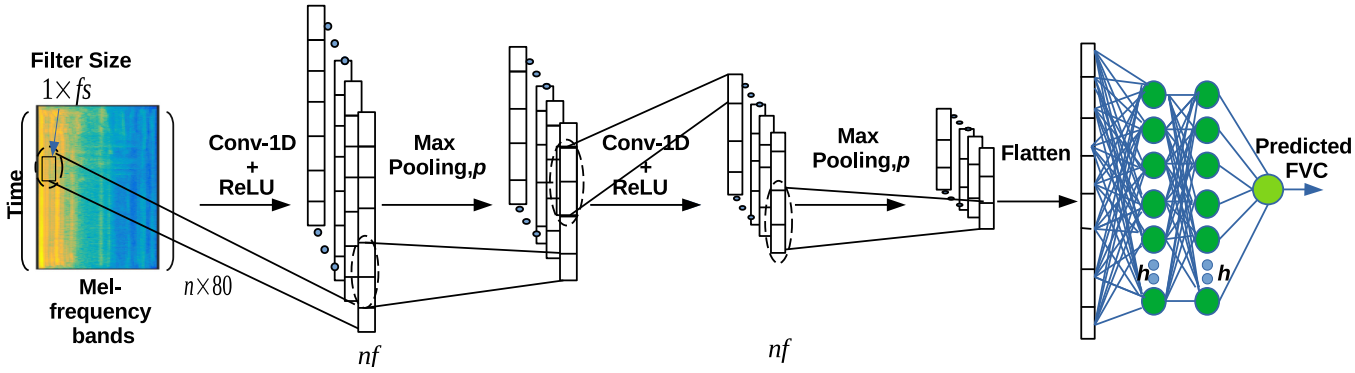

Fig. 1. Architecture of CDNN, where fs indicates filter size, n is time step, nf denotes number of filters, p indicates pooling size and h denotes hidden nodes in the deep neural network.

Mel-frequency cepstral coefficients (MFCC) for prediction. None of the works till now has explored the use of speech sounds. In this work, spirometry variable FVC has been predicted using sustained phonations namely /ɑː/ (as in 'Father'), /iː/ (as in 'Meat'), /uː/ (as in 'Suit'), /eɪ/ (as in 'May') and /oʊ/ (as 'Orange'). Iwarsson et al. [10] have shown that glottal voice source is a function of lung volume. They observed that glottal source parameters like closed quotient, peak to peak flow amplitude, and glottal leakages decrease with decreasing lung volume. According to source-filter theory by Fant [11], speech is modeled as a result of convolution of the source signal, radiation filter, and vocal tract filter. Therefore, changes in the source characteristics with varying lung volume should reflect in the speech as well. Based on this finding, we hypothesize that sustained phonations can encode information related to FVC. Constriction happens during the production of non-vowel sounds, which could modify the spectrum of the source signal in addition to vocal tract filter. Therefore, as a preliminary work, only vowels sounds are used. The following two questions are of interest in this work: a) Can FVC be predicted from sustained phonations? b) Out of five sustained phonations used in this work, which is the best for FVC prediction? Total 160 subject's data have been used to predict FVC in this work. Convolutional dense neural network (CDNN) is used for this purpose, which comprise two layers of Convolutional neural networks (CNN) with three fully connected layers at the end. Mel-spectrogram is used as a feature, and eight folds setup has been used. /iː/ is found to perform the best with mean Mean Absolute Error (MAE) of $0.67l(\pm 0.07l)$ among all sustained phonations. /uː/ have performed worst among all with mean MAE of $0.7l(\pm 0.13l)$.

2. DATASET

Dataset consists of total 160 subjects. Out of 160 subjects, 62 are healthy subjects, and 98 subjects are asthmatic patients. Description of the data set is shown in Table 1. Data

Table 1. Description of age(mean(SD)), gender distribution and spirometry variables namely, FEV1(ls^{-1})(mean(SD)), FVC(l)(mean(SD)) and FEV1/FVC(s^{-1})(mean(SD)) in patient and control group.

	Count	Male	Female	Age mean((±)SD)	FEV1(ls^{-1})	FVC(l)	FEV1/FVC(s^{-1})
Patient	98	43	55	41(±19)	1.71(±0.77)	2.27(±0.94)	0.75(± 0.12)
Control	62	34	28	34(±10)	2.63(±0.71)	3.02(±0.76)	0.87(±0.06)

is recorded in the hospital under the guidance of the doctor. Spirometry is performed in the St John's Medical College Hospital, Bengaluru, Karnataka, India, hospital by the technicians present in the laboratory, for patients as well as healthy subjects. For each subject spirometry is done by one of the three technician present in the laboratory according. Variation in the spirometry readings, FVC and FEV1, are shown to be non-significant due to technicians [12]. A consent form signed by the subjects before the recordings.

Sustained phonation of speech sounds namely /ɑː/ (as in 'Father'), /iː/ (as in 'Meat'), /uː/ (as in 'Suit'), /eɪ/ (as in 'May') and /oʊ/ (as Orange') are recorded. Total number of recordings for /ɑː/, /iː/, /uː/ , /eɪ/ and /oʊ/ are 873, 808, 761, 815 and 810, respectively. On average, five samples of each sustained phonation are recorded from every subject. All recordings are done using ZOOM H6 handy recorder in the hospital at a sampling rate of 44.1kHz. The average duration of each sample of sustained phonation is around 8 seconds. The recording has been done in the spirometry lab of the hospital, which, in general, has a noisy background because of the conversation between patients, technicians, AC noise, and fans. Recordings are done after 15 minutes of spirometry. During recording, patients are instructed to take deep breaths and utter the sustained phonation while breathing out until they are breathless. To make sure patients breath upto their full capacity, their nose is closed with the nose clip. Sufficient breaks are given between recordings to avoid the patients' fatigue, which can affect the experiment. Boundaries of each sustained phonation are marked manually by listening and

visual inspection of spectrogram and waveform in Audacity [13].

3. METHODOLOGY

In this work Mel-spectrogram is used as the features which is shown to carry the glottal source information [14]. CDNN has been used for predicting the FVC values because features learned through CNN can be analyzed to find the cues contributing to the better FVC prediction. CDNN architecture used in this work is presented in Fig. 1.

3.1. Architecture

To extract the relevant features for the prediction task CDNN has two convolutional layers followed by three fully connected layers[1]. Input to the CDNN is a 2D-matrix of dimension $n \times 80$, where n is the time steps, and 80 is the number of Mel-filter bands. In this work, 1D-CNN is used to capture temporal relation for each Mel-band. Each $1 \times n$ dimensional row of input Mel-spectrogram has been convolved with filter of length $1 \times fs$, where fs indicates filter size. fs is varied to determine how much neighborhood information is required to learn representation for better prediction of FVC. Each convolution layer of CDNN has nf number of filters. Output of each CNN layer is passed through rectified linear unit (ReLU) activation functions [15]. Output feature maps are max-pooled with size p to decrease the dimension. Output of the 2^{nd} layer of CNN after pooling is flattened and fed into the 1^{st} layer of dense neural network. Each layer, except the last layer of the neural network, has h hidden nodes, followed by ReLU activation. The second last layer's output is passed through the last layer of CDNN, which has one node and linear activation. Linear activation at the output is used to avoid the problem of vanishing gradient. ReLU activation has been used as activations functions at every layer of CDNN except the last layer because both input and output are non-negative.

3.2. Data Augmentation

Initially, experiments are done without augmentation, but even though training and validation mean MAE are good, evaluation on the test set yields poor results. By visualizing training and validation loss, it has been observed that the network is over-fitting to the training data. One way to alleviate this problem is to have more variability of the data while training, which is achieved by data augmentation. As known in literature, lung capacity is a function of age, gender etc. Therefore, all augmentations used in this work, modify the environment or recording conditions but not the voice characteristics. For this purpose, three kinds of augmentations, namely, reverberation, additive uniform noise, and both, have been used. These augmentation methods represent the variability introduced by the environment and do not change any speaker characteristics.

[1]All network parameters used in this work are determined after several rounds of experimental investigation.

4. EXPERIMENTAL SETUP

4.1. CDNN

Signal has been framed using 1 sec ($n = 101$) window length and 0.2 sec shift. Mel-spectrogram has been calculated at 20 msec window length and 10 msec overlap with 80 Mel-bands. Randomly ten frames have been chosen from each sustained phonation for training. For augmentation, the WavAugment package has been used [16]. Three kinds of augmentation, namely, reverberation, additive uniform noise, and their combination, have been used. From each of the three augmentation methods, ten frames have been randomly selected for each sustained phonations. Therefore, total 40 frames from each sustained phonations, ten from clean signal and ten each from augmented signal for training. During testing clean signal is used. During reverberation, augmentation, reverberance, damping factor, and room size are all set to 50. For noise addition, SNR has been randomly chosen in the range from 5 to 40. 1D-CNN has $nf = 10$ and $p = 2$. fs varies from 5 to 30 with a step size of 5. For varying fs, the network's receptive field at the last layer of CNN varies from 16 to 91 with a step size of 15. h in the first two fully-connected layers is 128. CDNN is optimized for MAE [15] using Adam optimizer [17] with learning rate 0.001. For each sound, networks are trained with the same initialization using Glorot's uniform initializer [18] to make a valid comparison. Batch size of 160 and early stopping criteria with the patience of 10 epochs are used to avoid over-fitting. Training epochs are set to be 30. Eight folds setup has been used in this work. Each fold has 20 subjects. Six folds are used for training (120 subjects), one fold for validation (20 subjects), and one for testing (20 subjects). After each epoch, MAE on the validation set has been calculated with the current model weights. Model weights corresponding to the minimum MAE have been used to get the prediction on the test set. CDNN is implemented using Keras [19] with TensorFlow [20]. As there is no existing work on the prediction of FVC using sustained phonations as a baseline scheme, the median of the training labels has been used for the FVC prediction, and baseline error is calculated using these predictions for each fold. The baseline scheme using this method is referred to as Baseline1. Wilcoxon signed-rank test has been used to test the significance [21] at the significance level of 5% in this work.

4.2. Baseline 2

Work by Rao et al. [8] used cough and wheeze sounds for the prediction task. As cough and wheeze are non-speech sounds and method used for the prediction using these sounds may not be a good choice for sustained phonations. Hence, to

Table 2. Mean(l) and SD(l) of MAE across 8 folds by using CDNN and Baseline1 method for all sustained phonation with varying filter size (fs). The minimum mean MAE on the validation set and corresponding test set is shown in bold for each sustained phonation.

fs	/ɑː/		/iː/		/oʊ/		/eɪ/		/uː/	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
5	0.651 ±.125	0.708 ±.113	0.614 ±.079	0.67 ±.1	0.645 ±.105	0.678 ±.131	**0.626 ±.077**	**0.706 ±.078**	0.638 ± .086	0.714 ± .139
10	0.66 ±.131	0.694 ±.105	0.609 ± .091	0.647 ± .11	0.642 ± .119	0.668 ± .122	0.633 ± .098	0.672 ± .109	**0.633 ±.086**	**0.7 ± .131**
15	**0.619 ±.146**	**0.683 ±.123**	**0.6 ± .095**	**0.671 ± .076**	**0.639 ±.112**	**0.677 ±.11**	0.662 ± .1	0.732 ± .145	0.685 ±.182	0.712 ±.07
20	0.658 ±.119	0.719 ±.165	0.614 ± .102	0.66 ± .098	0.669 ± .105	0.695 ± .099	0.646 ±.076	0.67 ±.083	0.703 ± .123	0.713 ± .117
25	0.668 ±.116	0.733 ±.149	0.642 ±.138	0.711 ± .151	0.682 ±.118	0.757 ± .09	0.673 ±.105	0.753 ±.086	0.687 ±.082	0.804 ±.204
30	0.666 ±.116	0.746 ±.142	0.605 ±.1	0.712 ±.072	0.69 ± .123	0.751 ± .085	0.658 ± .1	0.739 ±.074	0.658 ±.103	0.71 ±.135
Baseline1	0.792 ±.137	0.791 ±.151	0.759 ±.121	0.759 ± .125	0.784 ± .103	0.785 ± .109	0.794 ± .1	0.796 ± .114	0.779 ± .114	0.778 ± .118

check the validity of this hypothesis, FVC is predicted for our database in a manner similar to that by Rao et al. [8], and results are compared. This scheme for prediction is referred to as the Baseline 2.

4.3. Speaker verification

FVC value of the subject depends on height, weight, age, sex, and racial or ethnic background [22]. In our database a total of 127 FVC values are unique. The following experiment has been performed to find out if CNNs are learning representation about lung volume indeed, and not speaker characteristics like vocal tract shape, pitch, etc. i-vectors, the state-of-the-art features used for speaker recognition [23] are used as input to 3 layer neural network having the same configuration as fully connected layer of CDNN with FVC as target variable. If the performance obtained using i-vectors is identical or better than the CDNN learned features, it will indicate that features learned by CDNN are not lung volume specific, rather, they are doing speaker classification. On the other hand, if the performance of CDNN is better than that using i-vectors, it shows that learned features are indeed carrying information about the lung volume. For this experiment, same eight fold structure with augmentation is used to train neural networks. Open source Kaldi toolkit is used to calculate i-vectors [24] at 1 sec chunk with 0.2 sec shift. In most of the existing works, i-vector is computed using speech length greater than 5 seconds but in this work, sustained phonations are used, so we decided to experiment with 1 second chunks.

5. RESULTS & DISCUSSION

5.1. Comparison of sustained phonations for spirometry predictions

Results comparing the performance of the proposed CDNN and the Baseline1 are given in Table 2. It is observed that MAE has reduced using the CDNN method as compared to Baseline1 in all sounds. Maximum improvement (0.173l) in mean MAE is observed over the Baseline1 in the validation set for sound /ɑː/. For test sets, the maximum improvement (0.108l) in mean MAE is observed for sound /oʊ/ and /ɑː/. Minimum change of 0.145l is observed for sound /oʊ/ in the validation set over the baseline and 0.078l for /uː/ in the test set. The best performing sound is /iː/ using both baseline and CDNN in validation and test case. With varying fs as shown in Table 2, minimum MAE in the validation set is observed for fs less than or equal to 15 in all sustained phonations. It is interesting because it means that the local features learned over receptive field less than 460 msec signal is enough to predict FVC, even though the average duration of sustained phonation is 8 secs. Three sounds namely, /ɑː/, /iː/ and /oʊ/ have minimum validation error by using $fs = 15$. Similar trends have been noticed for the test sets for all sounds except /eɪ/ where SD is lower for $fs = 20$ as compared to $fs = 10$ even though mean MAE values are similar. For /ɑː/ and /uː/, the minimum MAE occur at the same fs for both validation and test sets. Maximum improvement of 0.07l and 0.104l in the mean MAE over baseline has been observed for sound /uː/ in validation and test set, respectively among all sustained phonations with varying fs. Among all the sustained phonations, /iː/ performed the best with a mean MAE of 0.671l(±.07l), which is an improvement of 0.088 over baseline in the test set by using $fs = 15$. On the other hand, the worst performing sound is /uː/ with a mean MAE of 0.7l(±.013l) by using $fs = 10$. All sounds have shown significant improvement over Baseline1 in both validation set and test set in best performing fs case (shown in bold in the table).

5.2. CDNN filters analysis

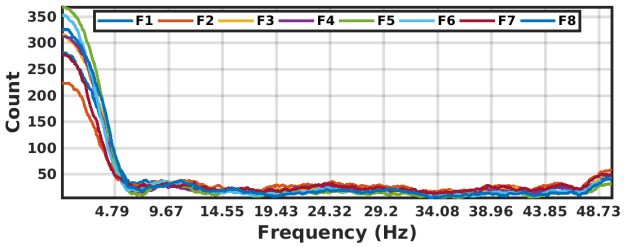

Fig. 2. Frequency versus count above 95 th percentile of the CNN filter spectrum with fs=15 for the best performing speech sound /iː/.

Spectral analysis of first layer CDNN filters is done for the best performing sound /iː/ for $fs = 15$. First layer of CDNN has $nf = 10$ each of size, $fs \times 80$. For every filter, 1024 points spectrum is calculated for each column. Hence 1024×80 dimensional spectrum matrix is estimated. For every filter spectrum, all frequencies with a magnitude greater

than 95^{th} percentile are equated to 1 and others to 0. Hence, we will get 10 (nf =10) such matrices of 1024×80 dimension with 0 and 1 entries. All 10 filters are summed up along nf, which leads to one spectrum matrix of size 1024×80. To see which frequencies are having majority of the weights, 1024×80 dimensional spectrum matrix is summed up along Mel-bands to get 1024×1 dimensional vector. This analysis will give us which frequencies in each fold has the most of the high gain for the filters. Similar steps have been repeated for each fold. Thus, we obtain a total of eight, 1024×1 dimensional vectors. Each of these vectors is plotted in Fig. 2. Only the first 512 points are plotted, corresponding to 50Hz, due to the symmetric property of the spectrum. i^{th} fold is referred as Fi, where $1 \leq i \leq 8$. Fig. 2 shows that all learned filters are low-pass in nature with a cut-off around 5Hz.

5.3. Comparison between Baseline 2 and CDNN method

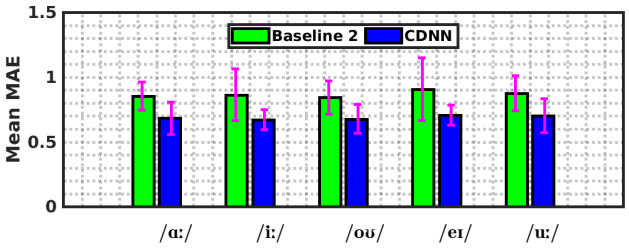

Fig. 3. Error bar graph plot to compare the Baseline 2 and CDNN method.

Rao et al. [8] (referred to as Baseline 2) have done FVC prediction using non-speech sounds, namely, cough and wheeze. The same method is used to predict FVC using sustained phonations. This experiment shows how well a method developed for cough and wheeze sound can perform for sustained phonations used in this work. This experiment's result is shown in the error bar graph of Fig. 3. Significant difference has been observed between proposed CDNN and Baseline 2 method in all sounds. An error bar graph is plotted for the mean(SD) of MAE of the test set which is corresponding to best performing fs on validation set (it is shown in bold in Table 2). Fig. 3 shows that the performance of CDNN is better than Baseline 2 across all sounds. Best performing sound using Baseline 2 is /oʊ/ with an MAE of $0.842l(\pm 0.129l)$, whereas for the same sound, MAE using CDNN is $0.677l(\pm 0.110l)$. Similarly, poor performing sound using Baseline 2 /eɪ/ has MAE of $0.905l(\pm 0.241l)$, whereas using CDNN, MAE becomes $0.706l(\pm 0.078l)$ for the same sound. Minimum and maximum improvement of 0.170 and 0.199 is observed in mean MAE over Baseline 2 for sound /ɑː/ and /eɪ/, respectively. This experiment suggests that an approach reported in the literature for cough and wheeze may not be suitable for the prediction of FVC using sustained phonations.

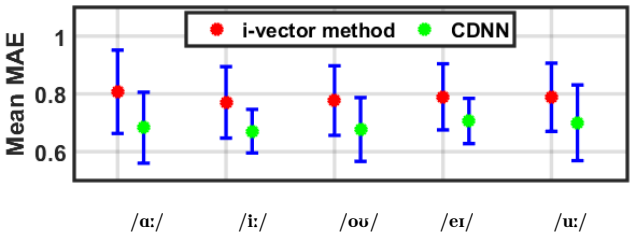

Fig. 4. Comparison of mean MAE and SD using i-vector features and CDNN.

5.4. Comparison between CDNN and i-vector based prediction

With varying fs, model corresponding to minimum mean MAE on validation set is selected to obtained the predictions on test set. Fig.4 is plotted between these test set values (values are shown in bold in Table 2) and corresponding mean MAE using i-vector features. Significant difference has been observed between proposed CDNN and i-vectors performance in all sounds. From fig. 4, it can be observed that features learned using CDNN performed better than the i-vector features across all sounds. By using i-vector features, minimum and maximum MAE are $0.77l(\pm .123l)$ and $.80l(\pm .144l)$ for /iː/ and /ɑː/, respectively. For the same sounds, MAE using CDNN are, $0.671l(\pm 0.075l)$ and $0.682l(\pm 0.122l)$. Interestingly, the best performing sound using i-vectors and CDNN is /iː/, whereas worst performing are /ɑː/ and /uː/, respectively. This experiment suggests that features learned using CDNN encodes information related to lung volume and not speaker characteristics captured using i-vectors.

6. CONCLUSIONS

From this work we concluded, FVC can be predicted using sustained phonations /ɑː/, /iː/, /oʊ/, /eɪ/, and /uː/ with mean MAE ranges from $0.671l(\pm .076l)$ to $0.7l(\pm 0.13l)$. Among all the sounds /iː/ performed the best with the minimum MAE and /uː/ performed the worst among all. In future, we plan to work on techniques to improve prediction of FVC and predicting FEV1 and FEV1/FVC by using sustained phonations and continuous speech. Standardization to predict spirometry variables using sounds also needs to be done in work future.

7. REFERENCES

[1] Joan B Soriano, Parkes J Kendrick, Katherine R Paulson, Vinay Gupta, Elissa M Abrams, Rufus Adesoji Adedoyin, Tara Ballav Adhikari, Shailesh M Advani, Anurag Agrawal, Elham Ahmadian, et al., "Prevalence and attributable health burden of chronic respiratory diseases, 1990–2017: a systematic analysis for the global

burden of disease study 2017," *The Lancet Respiratory Medicine*, vol. 8, no. 6, pp. 585–596, 2020.

[2] Sunil K Chhabra, "Clinical application of spirometry in asthma: Why, when and how often?," *Lung India: Official Organ of Indian Chest Society*, vol. 32, no. 6, pp. 635, 2015.

[3] David P Johns, Julia AE Walters, and E Haydn Walters, "Diagnosis and early detection of copd using spirometry," *Journal of thoracic disease*, vol. 6, no. 11, pp. 1557, 2014.

[4] VC Moore, "Spirometry: step by step," *Breathe*, vol. 8, no. 3, pp. 232–240, 2012.

[5] Martin R Miller, JATS Hankinson, V Brusasco, F Burgos, R Casaburi, A Coates, R Crapo, Pvd Enright, CPM Van Der Grinten, P Gustafsson, et al., "Standardisation of spirometry," *European respiratory journal*, vol. 26, no. 2, pp. 319–338, 2005.

[6] TR Schermer, JE Jacobs, NH Chavannes, J Hartman, HT Folgering, BJ Bottema, and C Van Weel, "Validity of spirometric testing in a general practice population of patients with chronic obstructive pulmonary disease," *Thorax*, vol. 58, no. 10, pp. 861–866, 2003.

[7] WT Ulmer, "Lung function–clinical importance, problems, and new results.," *Journal of physiology and pharmacology: an official journal of the Polish Physiological Society*, vol. 54, pp. 11–13, 2003.

[8] MV Achuth Rao, NK Kausthubha, Shivani Yadav, Dipanjan Gope, Uma Maheswari Krishnaswamy, and Prasanta Kumar Ghosh, "Automatic prediction of spirometry readings from cough and wheeze for monitoring of asthma severity," in *25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 41–45.

[9] Roneel V Sharan, Udantha R Abeyratne, Vinayak R Swarnkar, Scott Claxton, Craig Hukins, and Paul Porter, "Predicting spirometry readings using cough sound features and regression," *Physiological measurement*, vol. 39, no. 9, pp. 095001, 2018.

[10] Jenny Iwarsson, Monica Thomasson, and Johan Sundberg, "Effects of lung volume on the glottal voice source," *Journal of voice*, vol. 12, no. 4, pp. 424–433, 1998.

[11] Gunnar Fant, "The source filter concept in voice production," *STL-QPSR*, vol. 1, no. 1981, pp. 21–37, 1981.

[12] N Kunzli, U Ackermann-Liebrich, R Keller, AP Perruchoud, and CH Schindler, "Variability of fvc and fev1 due to technician, team, device and subject in an eight centre study: three quality control studies in sapaldia. swiss study on air pollution and lung disease in adults," *European Respiratory Journal*, vol. 8, no. 3, pp. 371–376, 1995.

[13] D Mazzoni and R Dannenberg, "Audacity [software]. pittsburg," 2000.

[14] MV Achuth Rao and Prasanta Kumar Ghosh, "Pitch prediction from mel-generalized cepstrum—a computationally efficient pitch modeling approach for speech synthesis," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 1629–1633.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux, "Data augmenting contrastive learning of speech representations in the time domain," *arXiv preprint arXiv:2007.00991*, 2020.

[17] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[18] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[19] François Chollet et al., "Keras," https://keras.io, 2015.

[20] Martín Abadi, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[21] RF Woolson, "Wilcoxon signed-rank test," *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007.

[22] Timothy Barreiro and Irene Perillo, "An approach to interpreting spirometry," *American family physician*, vol. 69, no. 5, pp. 1107–1114, 2004.

[23] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[24] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number no. CONF.