Enhancing Detection of Pathological Voice Disorder Based on Deep VGG-16 CNN

Agustinus Bimo Gumelar
Dept. of Electrical Engineering
Faculty of Intelligent Electrical and Informatics Technology (ELECTICS)
Institut Teknologi Sepuluh Nopember
Fakultas Ilmu Komputer
Universitas Narotama
Surabaya, Indonesia
bimogumelar@ieee.org

Eko Mulyanto Yuniarno
Dept. of Electrical Engineering,
Dept. of Computer Engineering,
Faculty of Intelligent Electrical and Informatics Technology (ELECTICS)
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ekomulyanto@ee.its.ac.id

Wiwik Anggraeni
Dept. of Information System
Faculty of Intelligent Electrical and Informatics Technology (ELECTICS)
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
wiwik@is.its.ac.id

Indar Sugiarto
Dept. of Electrical Engineering
Petra Christian University
Surabaya, Indonesia
indi@petra.ac.id

Vincentius Raki Mahindara
Dept. of Electrical Engineering
Faculty of Intelligent Electrical and Informatics Technology (ELECTICS)
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
raki@ieee.org

Mauridhi Hery Purnomo
Dept. of Electrical Engineering,
Dept. of Computer Engineering,
Faculty of Intelligent Electrical and Informatics Technology (ELECTICS)
Institut Teknologi Sepuluh Nopember
University Center of Excellence on Artificial Intelligence for Healthcare and Society (UCE AIHeS)
Surabaya, Indonesia
hery@ee.its.ac.id

***Abstract*—As a matter of fact, the system of human voice production is a sophisticated biological device that can modulate pitch and loudness. The essentials of internal and external factors often damage the vocal folds and change the vocal voice as a result. Thus, the consequences are well-portrayed in the function of the body and stand of emotion. Consequently, it is primary to identify voice changes at an early stage, deliver an opportunity to overcome any consequence, and enhance the patient's quality of life. In this case, voice disorder can be detected automatically by using Machine Learning (ML) techniques, which is, indeed, serves as a critical role. In this experiment, we specifically employ the Convolutional Neural Network (CNN), and a robust CNN model: the VGG-16. In investigating the performance of CNN in detecting disordered speech, we used the particular Pathological Voice Disorder (PVD) dataset, named the Respiratory Sound Database, which comprises hundreds of sampled PVD sound files. The experiment showed the accuracy of voice pathology detection arouses to 92.03%.**

Keywords— Pathological Voice Disorder; CNN, VGG-16, LSTM; VTLP Method

I. INTRODUCTION

The phenomenon of Pathological Voice Disorders (PVDs) occurs because of the excessive use of the voice. Various professions in various work environments have been reported due to the fact that it has been the trigger for PVD. A survey in several countries showed that 57.7% of teachers, being the profession that requires constant talking in their work hours, are exposed to voice problems and did have trouble speaking at a later age. The occurrence of voice problems for other professions is about 28.8% [1], [2]. In PVD victims, the primary malfunction caused by excessive use of vocal folds would produce different and rather defected voice. Vocal folds tend to vibrate abnormally, frequently with incomplete closure. Ultimately, the generated speech signal of PVD victims is more transient. Fig. 1 providing the speech signal (amplitudes) image visualization of healthy subjects and subjects with PVD. Therefore, an extensive and in-depth exploration of PVD in the speech signal can be done using the Automatic Voice Pathology Detection (AVPD), which is essentially developed in this study. In the initial examination of speech sound, AVPD is considered a non-invasive technique by some clinicians, therefore considered as a primary scanning tool. In a practical sense, AVPD would simplify the work of doctors specializing in otorhinolaryngology by differentiating speech with PVD for further examination and malpractice avoidance. Following this notion, researchers have found that sustained vowel in continuous speech samples with PVD and healthy classes is a rather tricky task [3]–[6]. Non-stationary and inherent behavior of speech signal in a continuous manner is a specific analysis challenge, according to Parsa and Jamieson [7] in their study in 2001. In terms of obtaining the acoustic characteristics, many other studies prefer the value of continuous speech rather than investigating deeper to sustained vowels. According to Askenfelt and Hammarberg [8], the important indicators of abnormal voice quality are based on prosodic variation (e.g. pitch and loudness) in continuous speech. Meanwhile, the agreed voice disorder examples are caused by vocal fold nodules, vocal fold paralysis, keratosis, and adductor spasmodic dysphonia[9]. Compared to the traditional methods, the Artificial Intelligence (AI) has proved itself to be superior in terms of computational problem-solving. The ability of AI to recognize patterns, fuzzy model implementation, build into Artificial Neural Network (ANN), and Support Vector Machine (SVM) has made AI the vital aspect in the fields of Biomedical Engineering [10]–[12]. The autonomous nature of SVM and PVD classification makes it appropriate to be dubbed as AVPD. And contrary to the notion of using utterance as data in speech, the Respiratory Sound Database [13], [14] provides

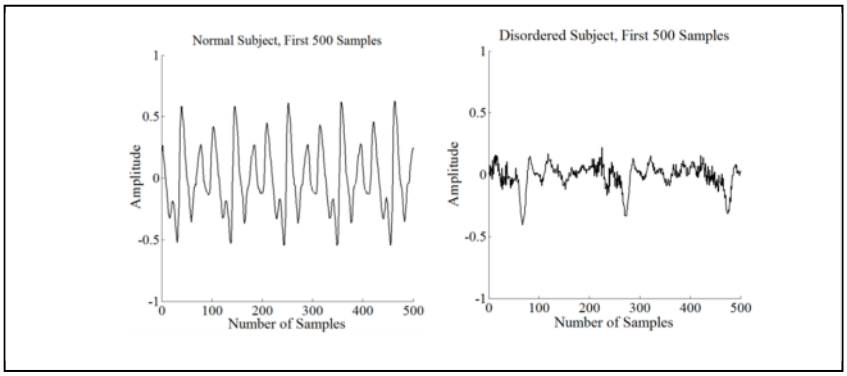

Fig 1. Amplitudes in voice samples of subjects classified as healthy (left chart), and PVD-affected (right chart) [41]

voice (hence the name Pathology Voice Disorder and "sound" in the name) as data, which contains no utterance.

This research study is composed as follows: Section I explained the mechanism effect of voice with PVD in humans, and the means using a variation of tools to identify PVD-affected speech. Secondly, Section II briefly presented previous works concerning PVD detection. The dataset material that we used in this study and some baseline features for the pre-processing technique is described in Section III. We elucidated the results from our experiment using the model we employed to train, namely, the Deep VGG-16 CNN. Finally, Section V draws out the conclusion and point of directions for future work.

II. Related Previous Works

In this study, we attempted to draw a distinction of the speech signal with PVD and the ones that were not affected by PVD. In this section, we also present the dataset being used in the experiment and compared different ML techniques in performing PVD detection.

A. Baseline Speech Features

Often reviewed speech features are the Linear Predictive Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), pitch, and frequency, as all of them have the robust representation of speech. MFCC represents humans' hearing mechanism, while LPCC represents humans' speech production. Other varying speech features that are also beneficial to be used are shimmer, jitter, HNR, glottal noise ratio, and vocal tract tube fluctuation.

Various interdisciplinary studies, specifically psychology and linguistics, are concerned with the mechanism of human speech. Many of them used the Hidden Markov Model (HMM) using prosodic features, and would later compare the performance if the spectral features are selected [15]. In a set of speech features, many speech features representing various mechanism and readings of human voice would be beneficial in the long process of understanding human voice, and discriminate binary or multi characteristics in speech, such as emotions [16], [17], PVD/healthy, humor/non-humor [18], [19], toxic/normal, etc.

In addition of this experiment, we specifically employed Cochleagram and Hilbert Spectrum (HS). A Cochleagram is the gammatone filter to display a non-linear, high-resolution image representation of the timing and frequency of the sound signal, while HS is the speech feature as a product from a non-linear and non-stationer signal data (including voice) analysis method called the Hilbert-Huang Transform (HHT) [20]–[22].

B. VTLP Method of Data Augmentation

In previous study of Jaitly and Hinton, a Vocal Tract Length Perturbation (VTLP) method was used to corrupt a sound file [23], using random warping factors chosen from range 0.9 until 1.1. There were improvements reported their experiment for TIMIT phoneme recognition [24]. Meanwhile, some study have been using VTLP in a large vocabulary continuous speech recognition task for better performance [25]–[27].

$$f' = \begin{cases} f\alpha \\ \frac{S}{2} - \frac{\frac{S}{2} - F_{hi}\min(\alpha,1)}{\frac{S}{2} - F_{hi}\frac{\min(\alpha,1)}{\alpha}}\left(\frac{S}{2} - f\right) \end{cases} \quad (1)$$

$$f \leq F_{hi}\frac{\min(\alpha,1)}{\alpha} \quad (2)$$

In this experiment, we follow the experiment of Jaitly and Hinton, only to change the purpose, not to corrupt, but to normalize. So we used the traditional warp factor which lies between 0.8 and 1.2. Equation (1) and Equation (2) shows the fundamental VTLP formula.

C. The VGG-16 CNN

Interdisciplinary study over the years have reported various experiment using ML techniques. In many of those reports, Convolutional Neural Network (CNN) stood tall as it frequently achieve predicted result, sometimes performing even better than what was predicted. The golden era of CNN begun in 2012, where eight-layered CNN was first implemented. It contains five convolutional layer and three fully-connected layers [28].

The said CNN model won the 2012 AlexNet competition (was called the ImageNet back then), with an improvement of error from 25.8% to 16.4% in the span of a year of development. According to Lettvin et al. in 1959 [29], it was reported that individual cells in visual cortex are prone to the definite regions of the visual field, by responding to the presence of edges with certain orientation. Lettvin et al. stated that the neurons working together in a columnar manner to construct a visual perception. The said behavior became the basic of CNNs, having specialized components (as neuron cells in visual cortex) looking for specific characteristics.

III. Experimental Result

Inspired by the studies that reported on the previous section, CNN has a significant ability to detect PVD. A model

of CNN were applied in this experiment, namely the VGG-16. Also, we used public dataset of RSD with healthy or PVD affected sound files. Fig. 2 shows the flow diagram of input data preparation for the experiment.

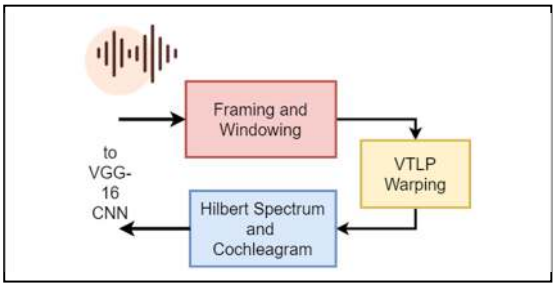

Fig 2. Input Data Preparation

A. Respiratory Sound Database

In these experiments, we used the Respiratory Sound Database (RSD) [13], [14]. It was created by annotating 920 speech recordings of varying length, more or less, 10s to 90s. These recordings were taken from 126 patients. Approximately 5.5 hours of recordings have contained 6.898 respiratory cycles: 1.864 crackles sound file, 886 wheezes sound file, and 506 of both crackles and wheezes sound file.

The dataset includes both clean respiratory sounds as well as noisy recordings that simulate real-life conditions. The patients span all age groups, namely, children, adults, and the elderly.

B. Speech Features Transformation

It is thought that emotional expression are still constructed very well by the speakers, although by defected dominant in the speech condition. In this step of experiment, we meant to classify mentioned speech from the dataset using the notion of Hamming windowing [30], [31].

Cochleagram follows the basic mechanism of human ear's, which is able to implemented Half Wave Rectifier (HWR) to replace inner hair cells to detect the output of each filter. The non-linear nature of the HWR simulates the change in motion in response to the basilar membrane in the human cochlea, into a signal that represents sound energy, while retaining temporal information [32]–[35].

On the other hand, two-stage processes of HHT requires Empirical Mode Decomposition (EMD) to breaks down the speech signal into a number of Intrinsic Mode Functions (IMF) [36], [37]. The EMD identifies and connect the dots of the maximas and minimas (both called a set of extremas) in a signal, through a process called cubic spline interpolation; eventually resulting in auxiliary lines called the envelopes. A set of envelopes is collected in an IMF, with the next derivation depends on the previous IMF [38]–[40].

All speech features were transformed into Hilbert Spectrum (HS) and Cochleagram. Fig. 3 shows the Cochleagram features of PVD-affected sound files and its healthy class, while Fig. 4 shows the HS features of sound files by the chest location.

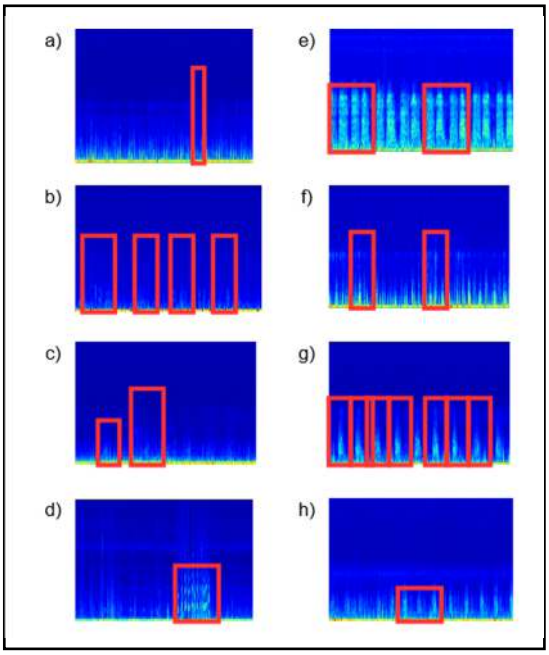

Fig 3. Cochlea features of a) URTI, b) Asthma, c) LRTI, d) COPD, e) Bronchiectasis, f) Bronchiolitis, g) Pneumonia, h) Healthy

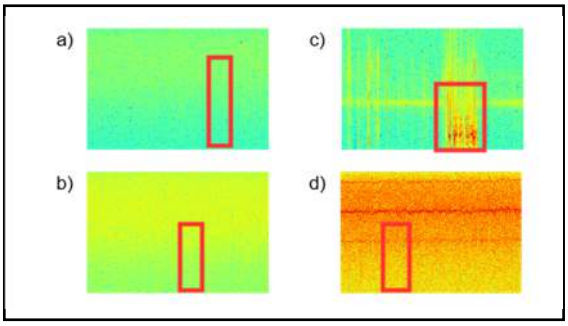

Fig 4. Hilbert Spectrum of a) Anterior Left, b) Anterior Right, c) Right Posterior, d) Trachea

C. The VGG-16 CNN

Together but in separate process, both HS and Cochleagram undergone the CNN build, resulting in different classification accuracy. We also prepared another ML techniques in comparison of the VGG-16 CNN performance, namely the Long Short-Term Memory (LSTM), Random Forest (RF), Neural Network (NN), and SVM.

This sub-section deliberately elucidate our schematic and mathematical concern behind the employed VGG-16 CNN. Consider a layer y = f(x). We wanted to discover which x components influence which y components. We also assume that this is relatable in terms of the receptive field. So, the output component yi ", j" depends only on the input components xi, j where $(i, j) \in \Omega(i", j")$. The set $\Omega(i", j")$ is a rectangle defined in the Equation (3) and Equation (4).

$$i \in \alpha(i" - 1) + \beta_h + [-\frac{\delta_h - 1}{2}, \frac{\delta_h - 1}{2}] \quad (3)$$

$$i \in \alpha(j" - 1) + \beta_v + [-\frac{\delta_v - 1}{2}, \frac{\delta_v - 1}{2}] \quad (4)$$

where (α_h, α_v) is the stride, (β_h, β_v) the offset, and (Δ_h, Δ_v) the receptive field size.

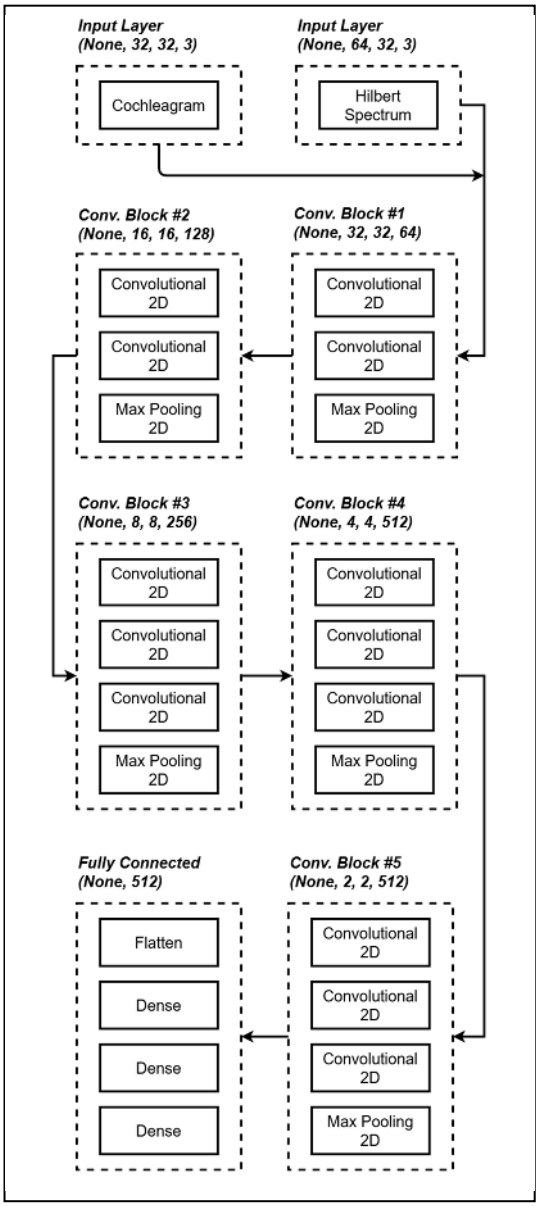

Fig 5. A schematic of the Deep VGG-16 CNN architecture

Fig. 5 shows the schematic of VGG-16 CNN used in this experiment, while Table I explained the schematic of the VGG-16 CNN in detail. Our VGG-16 CNN apply the number of Cochlea and HS frames to be tuned as a hyperparameter.

TABLE I. SUMMARY LOG OF BUILT VGG-16 CNN

Layer (Type)	Output Shape	Param #
input_1 (Input Layer)	None, 32, 32, 3	0
block1_conv1 (Conv2D)	None, 32, 32, 64	1,792
block1_conv2 (Conv2D)	None, 32, 32, 64	36,928
block1_pool (MaxPooling2D)	None, 16, 16, 64	0
block2_conv1 (Conv2D)	None, 16, 16, 128	73,856
block2_conv1 (Conv2D)	None, 16, 16, 128	147,584
block2_pool (MaxPooling2D)	None, 8, 8, 128	0
block3_conv1 (Conv2D)	None, 8, 8, 256	295,168
block3_conv2 (Conv2D)	None, 8, 8, 256	590,080
block3_conv3 (Conv2D)	None, 8, 8, 256	590,080
block3_pool (MaxPooling2D)	None, 4, 4, 256	0
block4_conv1 (Conv2D)	None, 4, 4, 512	1,180,160
block4_conv2 (Conv2D)	None, 4, 4, 512	2,359,808
block4_conv3 (Conv2D)	None, 4, 4, 512	2,359,808
block4_pool (MaxPooling2D)	None, 2, 2, 512	0
block5_conv1 (Conv2D)	None, 2, 2, 512	2,359,808
block5_conv2 (Conv2D)	None, 2, 2, 512	2,359,808
block5_conv3 (Conv2D)	None, 2, 2, 512	2,359,808
block5_pool (MaxPooling2D)	None, 1, 1, 512	0
flatten	None, 512	0
fc1 (Dense)	None, 4096	2,101,248
fc2 (Dense)	None, 4096	16,781,312
predictions (Dense)	None, 8	32,776
	Total params	33,630,024
	Trainable params	33,591,304
	Non-trainable params	38,720

TABLE II. F-SCORE MATRIX

Method	F-score							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
VGG-16	0.01	0.11	0.01	0.97	0.18	0.22	0.01	0.01
NN	0.01	0.09	0.01	0.88	0.13	0.17	0.00	0.00
RF	0.01	0.03	0.00	0.87	0.01	0.11	0.00	0.00
SVM	0.01	0.01	0.01	0.91	0.11	0.53	0.55	0.00
LSTM	0.00	0.00	0.00	0.82	0.00	0.01	0.00	0.00

D. Result and Discussion

At the evaluation stage, every performance result of the ML techniques is evaluated. We used the Confusion Matrix (CM) to produce the evaluation result. Fig. 7 shows the CM result of part of PVD, Chronic Obstructive Pulmonary Disease (COPD) vs non-COPD class (along with healthy class) with Cochleagram as the speech features, while Fig. 8 shows the same only with Hilbert Spectrum as the speech features. Table II elaborated each F1-score for every PVD classes (the same sequence as in Fig. 3) using every tested ML techniques. The CM shown in Fig. 7 evaluates the quality of the output of a classifier on the COPD-affected dataset. The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the confusion matrix are the better, indicating many correct predictions. In Fig. 9, the number of PVD detection accuracy has reached 92.03% using Cochleagram features, and 91.30% using HS features as in Fig. 10.

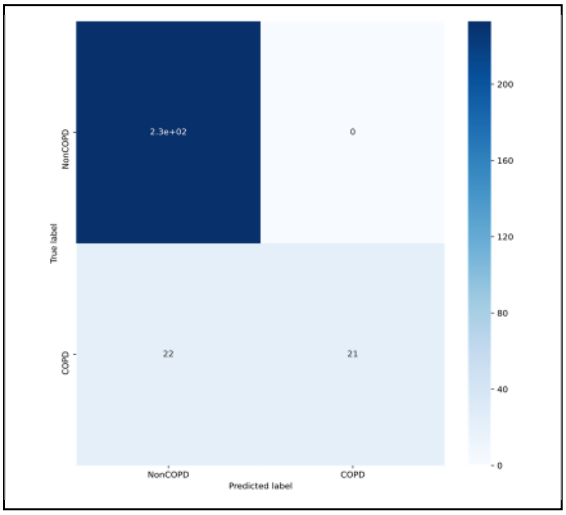

Fig 6 Confusion Matrix of COPD vs non-COPD with Cochleagram

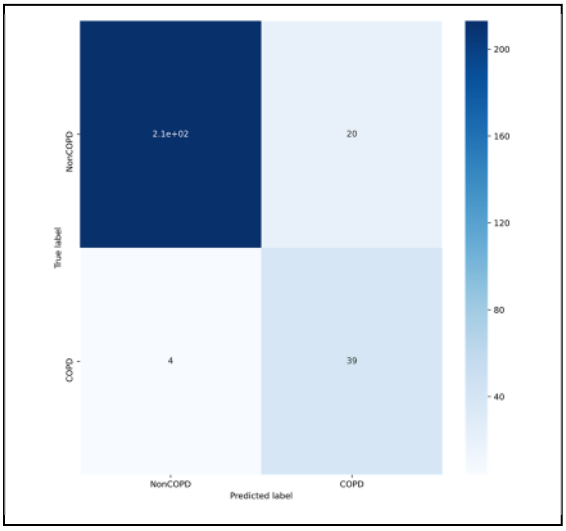

Fig 7 Confusion Matrix of COPD vs non-COPD with Hilbert Spectrum

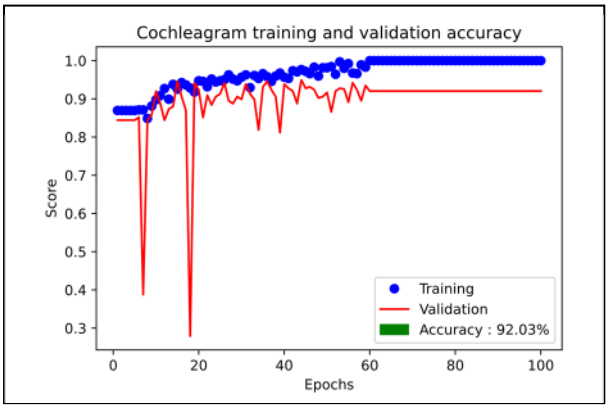

Fig 8. Training and Validation Accuracy using Cochleagram

In this work, we showed that more developed CNN for the PVD classification in the form of VGG-16 CNN are statistically significant predictors of many cases of PVD classes, including healthy class. The validation accuracy dropped by 0.73% when using different speech features. This result occurs because the visual feature might be, have an association between visual emphysema (a long-term progressive disease of the lungs) and another disease that can not be reach by voice.

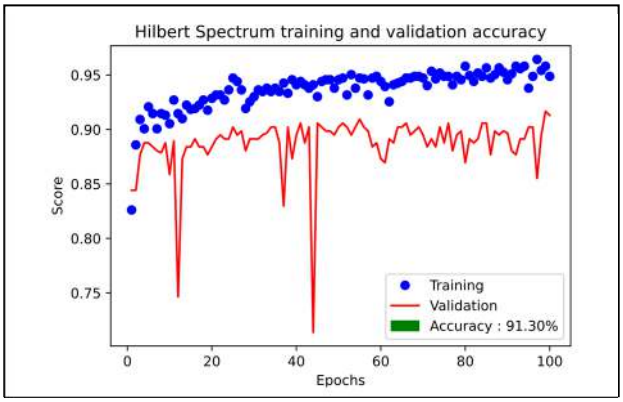

Fig 9. Training and Validation Accuracy using Hilbert Spectrum

IV. Conclusion

In this paper, we have shown the superiority and the general ability of CNN in PVD-affected speech detection. Specific for the experiment, we adopted the robust CNN model, namely the VGG-16 CNN. During the experiment, we feel the need of applying the VTLP as the warping technique, since it eliminates speaker's variability based on vocal tract length. We also prepared the data to be transformed to HS and Cochleagram, which has the sole purpose to replicate the mechanism of human hearing system.

This paper presents the result of detection accuracy achieved 92.03%, using Cochleagram and specifically designed VGG-16 CNN. In the future, we suspect the performance result can be improved in its accuracy using ML-based feature selection.

Acknowledgment

We thank the Ministry of Research and Technology / National Research and Innovation Agency (Kemenristek / BRIN) Republic of Indonesia for funding this research from BPPDN grant schemas.

References

[1] N. Roy, R. M. Merrill, S. Thibeault, S. D. Gray, and E. M. Smith, "Voice Disorders in Teachers and the General Population," *J. Speech, Lang. Hear. Res.*, vol. 47, no. 3, pp. 542–551, Jun. 2004.

[2] N. Roy, R. M. Merrill, S. Thibeault, R. A. Parsa, S. D. Gray, and E. M. Smith, "Prevalence of Voice Disorders in Teachers and the General Population," *J. Speech, Lang. Hear. Res.*, vol. 47, no. 2, pp. 281–293, Apr. 2004.

[3] G. Muhammad *et al.*, "Automatic Voice Pathology Detection and Classification using Vocal Tract Area Irregularity," *Biocybern. Biomed. Eng.*, vol. 36, no. 2, pp. 309–317, 2016.

[4] J.-W. Lee, H.-G. Kang, J.-Y. Choi, and Y.-I. Son, "An Investigation of Vocal Tract Characteristics for Acoustic Discrimination of Pathological Voices," *Biomed Res. Int.*, vol. 2013, pp. 1–11, 2013.

[5] J.-W. Lee, S. Kim, and H.-G. Kang, "Detecting Pathological Speech using Contour Modeling of Harmonic-to-Noise Ratio," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5969–5973.

[6] G. Muhammad and M. Melhem, "Pathological Voice Detection and Binary Classification using MPEG-7 Audio Features," *Biomed. Signal Process. Control*, vol. 11, pp. 1–9, May 2014.

[7] V. Parsa and D. G. Jamieson, "Identification of Pathological Voices Using Glottal Noise Measures," *J. Speech, Lang. Hear. Res.*, vol. 43, no. 2, pp. 469–485, Apr. 2000.

[8] B. Hammarberg, B. Fritzell, J. Gaufin, J. Sundberg, and L. Wedin, "Perceptual and Acoustic Correlates of Abnormal Voice Qualities," *Acta Otolaryngol.*, vol. 90, no. 1–6, pp. 441–451, Jan. 1980.

[9] G. Muhammad, T. A. Mesallam, K. H. Malki, M. Farahat, A. Mahmood, and M. Alsulaiman, "Multidirectional Regression (MDR)-Based Features for Automatic Voice Disorder Detection," *J. Voice*, vol. 26, no. 6, pp. 817.e19-817.e27, Nov. 2012.

[10] B. S. Aghazadeh and H. K. Heris, "Fuzzy Logic based Classification and Assessment of Pathological Voice Signals," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2009, pp. 328–331.

[11] M. P. Paulraj, S. Bin Yaacob, A. N. Abdullah, and S. K. Natraj, "Fuzzy Voice Segment Classifier for Voice Pathology Classification," in *2010 6th International Colloquium on Signal Processing & its Applications*, 2010, pp. 190–195.

[12] M. Al Mojaly, G. Muhammad, and M. Alsulaiman, "Detection and Classification of Voice Pathology using Feature Selection," in *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, 2014, pp. 571–577.

[13] B. M. Rocha *et al.*, "A Respiratory Sound Database for the Development of Automated Classification," 2018, pp. 33–37.

[14] Marsh, "Respiratory Sound Database," *Kaggle Database*. [Online]. Available: https://www.kaggle.com/vbookshelf/respiratory-sound-database. [Accessed: 10-Feb-2020].

[15] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic Segmentation and Labeling of Speech based on Hidden Markov Models," *Speech Commun.*, vol. 12, no. 4, pp. 357–370, Aug. 1993.

[16] E. Franti, I. Ispas, V. Dragomir, M. Dasc, E. Alu, Zoltan, and I. C. Stoica, "Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots," *Rom. J. Inf. Sci. Technol.*, vol. 20, no. 3, pp. 222–240, 2017.

[17] N. Sundarprasad, "Speech Emotion Detection Using Machine Learning Techniques," *Master's Proj.*, 2018.

[18] S. Attardo, L. Pickering, and A. Baker, "Prosodic and Multimodal Markers of Humor in Conversation," *Pragmat. Cogn.*, vol. 19, no. 2, pp. 224–247, Aug. 2011.

[19] J. Juckel, S. Bellman, and D. Varan, "A Humor Typology to Identify Humor Styles Used in Sitcoms," *HUMOR*, vol. 29, no. 4, p. 583, Jan. 2016.

[20] B. Gao, W. L. Woo, and L. C. Khor, "Cochleagram-based Audio Pattern Separation using Two-Dimensional Non-negative Matrix Factorization with Automatic Sparsity Adaptation," *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1171–1185, Mar. 2014.

[21] M. Russo, M. Stella, M. Sikora, and V. Pekić, "Robust Cochlear-Model-Based Speech Recognition," *Computers*, vol. 8, no. 1, p. 5, Jan. 2019.

[22] V. K. Rai and A. R. Mohanty, "Bearing Fault Diagnosis using FFT of Intrinsic Mode Functions in Hilbert–Huang Transform," *Mech. Syst. Signal Process.*, vol. 21, no. 6, pp. 2607–2615, Aug. 2007.

[23] N. Jaitly and G. E. Hinton, "Vocal Tract Length Perturbation (VTLP) Improves Speech Recognition," in *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[24] J. Garofolo, L. Lamel, W. ~M. Fisher, J. ~G. Fiscus, and D. ~S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1." p. 27403, Feb-1993.

[25] I. Rebai, Y. BenAyed, W. Mahdi, and J.-P. Lorré, "Improving Speech Recognition using Data Augmentation and Acoustic Model Fusion," *Procedia Comput. Sci.*, vol. 112, pp. 316–322, 2017.

[26] Xiaodong Cui, V. Goel, and B. Kingsbury, "Data Augmentation for Deep Neural Network Acoustic Modeling," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1469–1477, Sep. 2015.

[27] A. Ragni, K. Knill, S. Rath, and M. J. F. Gales, "Data Augmentation for Low Resource Languages," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 810–814, 2014.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[29] J. Lettvin, H. Maturana, W. McCulloch, and W. Pitts, "What the Frog's Eye Tells the Frog's Brain," *Proc. IRE*, vol. 47, no. 11, pp. 1940–1951, Nov. 1959.

[30] M. Pereira, S. Chapaneri, and D. Jayaswal, "Analysis of Windowing Techniques for Speech Emotion Recognition," in *2016 International Conference on Information Communication and Embedded Systems (ICICES)*, 2016, pp. 1–6.

[31] P. Podder, T. Zaman Khan, M. Haque Khan, and M. Muktadir Rahman, "Comparative Performance Analysis of Hamming, Hanning and Blackman Window," *Int. J. Comput. Appl.*, vol. 96, no. 18, pp. 1–7, Jun. 2014.

[32] M. Slaney and R. F. Lyon, "On the Importance of Time - A Temporal Representation of Sound," in *Visual Representations of Speech Signals*, 1993, pp. 95–116.

[33] Y. K. Muthusamy, R. A. Cole, and M. Slaney, "Speaker-independent Vowel Recognition: Spectrograms Versus Cochleagrams," in *International Conference on Acoustics, Speech, and Signal Processing*, 1990, pp. 533–536.

[34] R. V. Sharan and T. J. Moir, "Cochleagram Image Feature for Improved Robustness in Sound Recognition," *Int. Conf. Digit. Signal Process. DSP*, vol. 2015-Septe, pp. 441–444, 2015.

[35] M. K. I. Molla and K. Hirose, "Single-Mixture Audio Source Separation by Subspace Decomposition of Hilbert Spectrum," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 3, pp. 893–900, Mar. 2007.

[36] H. Huang and X.-X. Chen, "Speech Formant Frequency Estimation based on Hilbert-Huang Transform," *Zhejiang Daxue Xuebao (Gongxue Ban)/Journal Zhejiang Univ. (Engineering Sci.*, vol. 40, pp. 1926–1930, 2006.

[37] A. B. Gumelar, Eko Mulyanto Yuniarno, Wiwik Anggraeni, Indar Sugiarto, A. A. Kristanto, and M. H. Purnomo, "Kombinasi Fitur Multispektrum Hilbert dan Cochleagram untuk Identifikasi Emosi Wicara [Spectrum Features Combination of Hilbert and Cochleagram for Speech Emotion Identification]," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 9, no. 2, pp. 180–189, May 2020.

[38] N. Sharma and T. Gedeon, "Objective Measures, Sensors and Computational Techniques for Stress Recognition and Classification: A Survey," *Comput. Methods Programs Biomed.*, vol. 108, no. 3, pp. 1287–1301, Dec. 2012.

[39] R. Sharma, R. K. Bhukya, and S. R. M. Prasanna, "Analysis of the Hilbert Spectrum for Text-Dependent Speaker Verification," *Speech Commun.*, vol. 96, no. December, pp. 207–224, 2018.

[40] N. E. Huang *et al.*, "The Empirical Mode Decomposition and the Hubert Spectrum for Non-linear and Non-stationary Time Series Analysis," *Proc. R. Soc. A Math. Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998.

[41] Z. Ali *et al.*, "Voice Pathology Detection based on the Modified Voice Contour and SVM," *Biol. Inspired Cogn. Archit.*, vol. 15, pp. 10–18, Jan. 2016.