Training a CNN to estimate voice pathology from connected speech using EGG to automatically label the dataset for voicing

Ian S. Howard[1], Julian McGlashan[2] & Adrian J. Fourcin[3]

[1]SECAM, University of Plymouth, Plymouth, PL4 8AA UK

[2]ENT Department, Nottingham University Hospitals UK

[3]Laryngograph Ltd, Wallington UK, Emeritus Professor at UCL, UK

ian.howard@plymouth.ac.uk, julian.McGlashan@nottingham.ac.uk, fourcin@btinternet.com

ABSTRACT: We describe a new system for estimating voice pathology directly from the acoustic speech signal to assist in the diagnosis of pathological voice conditions by voice specialists. Our main novel contributions are the use of Electroglottography (EGG) in neural net training to automatically label speech acoustic signals for voicing and the generation of running estimates of pathology with high temporal resolution from the acoustic signal alone. These estimates can also be linked to the parts of speech signals where voice pathology manifests itself most strongly. By operating directly on the acoustic signal waveform without the use of any pre-processing, we avoid the use of hand-crafted features. We trained and tested a neural network using speech datasets with normal and pathological voicing and found that it can provide effective fine-grained indications of pathology. Our quantitative results show that this neural network performs well in distinguishing between speakers with normal and pathological voice conditions, achieving a recognition rate of 91%, which compares favorably with results from other studies.

1 Introduction

Voice problems, or dysphonia, affect one in thirteen adults in the UK annually, causing a significant impact on quality of life and livelihood, as well as a substantial burden on the healthcare system. The prevalence of dysphonia is higher in vocally demanding professions and among the elderly. The causes of dysphonia can range from benign to malignant, such as cancer. Different voice pathologies are typically associated with a range of effects [1]. These effects are often only manifested during specific parts of speech signals, such as voicing onsets, offsets, and specific pitch changes. For example, a vocal nodule lesion mainly affects the behavior of the vibrating edge of the vocal fold, with its effect increasing with pitch. Neurological conditions, such as those arising from a stroke, can interfere with speech timing and vocal fold palsy can affect vocal fold contact regularity and phase. Adductor and abductor spasmodic dysphonias also affect onset airflow differently [2,3]. Vocal fold paresis can cause pitch irregularity and also disrupt speech prosody, due to patients running out of breath as a result of air leakage.

2 Current approaches to dysphonia diagnosis

2.1 Clinical examination

Current approaches to the diagnosis of voice pathology [4,5,6,7] typically involve expert clinical assessment. This requires a consultant to examine a patient endoscopically, listen to their speech and carry out psychoacoustic and auditory-perceptual evaluation of the voice [8]. Since voice pathology in speech may only manifest itself in certain phonetic contexts, it is necessary to examine speech production using spoken tasks in which it is exhibited. Indeed, specifically designed passages have long been used by clinicians and researchers and are still being developed [9].

2.3 Using electroglottography

In some specialist centres, additional measurements of laryngeal function are undertaken that involve the use of electroglottography (EGG), such as using the Laryngograph [10]. This is an electrical conductance-based measurement technique that yields information on vocal fold contact and vibratory patterns using electrodes placed externally on the neck over the patient's larynx. It is currently employed worldwide to assist the diagnosis of laryngeal pathologies and provides a unique insight into vocal fold vibratory function.

To assist the clinical diagnosis of voice pathology, simple metrics have been proposed that are indicative of voice pathology, for example relating to open and closed phase of the EGG signal and shimmer [11,12]. Additionally, more sophisticated measures based on speech analysis have also been proposed [13]. Although these measures are certainly useful, there is now a strong trend towards the application of automated [14, 15] and more general machine learning approaches to the inference of voice pathology and much work has recently been carried out in this area [16,17,18,19,20].

3 Machine learning for speech pathology estimation

Encouraging progress has been made in the direct estimation of voice pathology from the speech signal using a range of machine learning classification techniques (see Hedge [21] for a recent survey), including a study that uses connected speech [22]. Operating on speech alone removes the need for specialist equipment to directly record EGG and enhances the possibility of automatic remote screening, including the use of cloud-linked smart phones. Recent work has also compared the effectiveness of different machine learning techniques on the same datasets to ensure fair comparison can be made [23], as well as a preliminary investigation of the use of deep neural networks for fine-grained estimation of voice pathology [24].

Here we take the first steps towards building a system to assist clinicians in diagnosing voice pathology, using an analysis that operates directly on the acoustic speech signal without the use of a feature detection stage. To do this, we use a convolutional neural network (CNN), which, in addition to providing an estimate of pathological voicing, is also able to automatically learn useful features for input data processing directly from the speech waveform. We believe this approach assists in circumventing deficiencies that can arise in systems that use handcrafted features.

4 Methods

4.1 Data acquisition

We used clinical data consisting of simultaneously recorded speech and Laryngograph electroglottograph (EGG) signals from 57 speakers comprising 27 normal (male = 13, female = 14) and 30 pathological (male = 9, female = 21) voices, reading the "Arthur the Rat" passage in British English [25]. The pathological data and some normal controls (7) data were recorded in a normal clinical environment using Laryngograph Speech Studio using a standard electret condenser microphone. The pathological speakers exhibited adductor (n=25) and abductor dysphonia (n=5). To provide a balance between pathological and normal speakers, data from additional 20 normal speakers were included in the dataset. This data was recorded from students in noise-free anechoic conditions. All speech and Laryngograph data were down-sampled to 16kHz to reduce the subsequent size of pattern vectors.

4.2 Data labelling

The time derivative of the EGG signal was calculated to estimate the location of excitation points in time. These were then used to delineate regions of voicing using a 400-sample

window centred on each excitation point. The automatic determination of voicing is illustrated in Fig. 1. Three labels were used to indicate regions of: all voicing (Va); normal voicing (Vn) and pathological voicing (Vp), as illustrated in Fig. 2. Target values were set to numeric values +1 to indicate the presence of each respective label and -1 to indicate the absence. The diagnoses of the voice pathologies were made by a single clinician on the basis of full clinical examination. We note that, for these pathologies, such judgments are very reliable and rarely differ across clinicians.

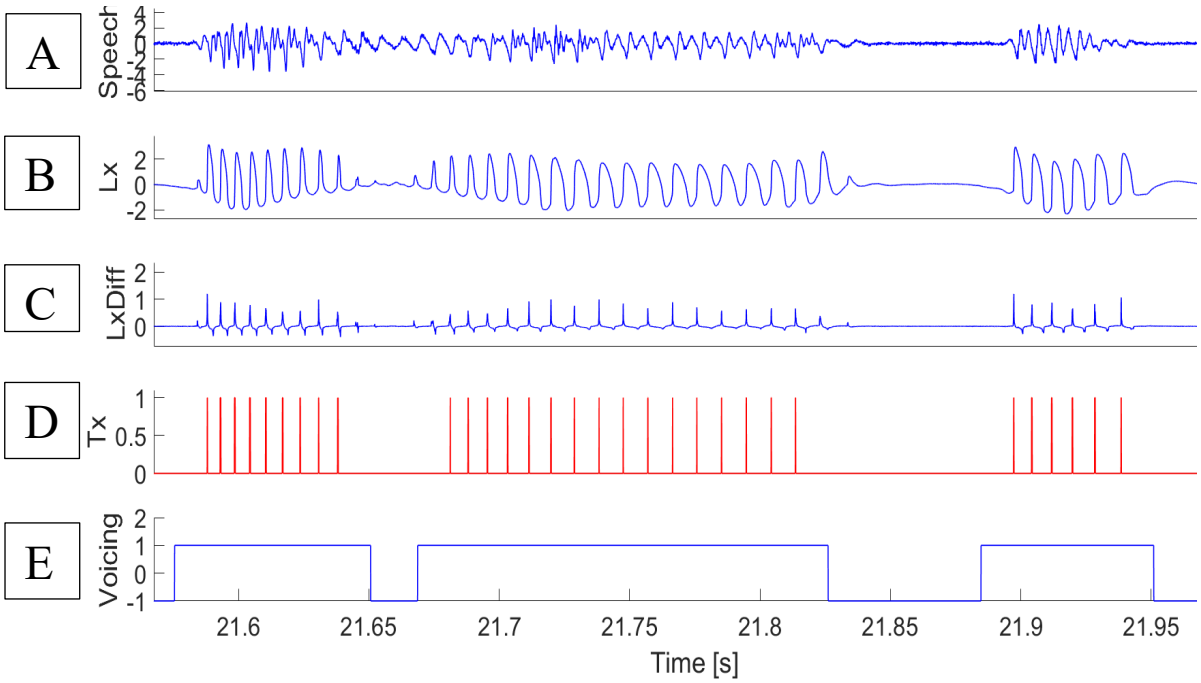

Figure 1. *Example of automatic data labelling for voicing using the Laryngograph signal training data for the except from the Arthur the Rat passage "for him if he". A Speech and B simultaneously recorded Laryngograph signal (Lx), C time derivative of the Laryngograph signal (LxDiff). D corresponding time of excitation markers derived by thresholding the differential of the Laryngograph signal (Tx). E the Tx markers are used to delineate regions of voicing (Vx) by applying a window.*

4.4 Network structure and training

We used a relatively shallow convolutional neural network (CNN) with 3 outputs to directly map input speech samples to the 3 different voice labels Va, Vn, and Vp, which were used as training targets. Network structure was motivated by previous work using CNNs in initial proof of concept analysis [24]. The CNN used an input window size of 4001 adjacent speech waveform samples (spanning 250ms), had 3 convolutional layers with an input width of 20, made use of ReLu output activations and a max-pooling factor of 2. The output layer was fully connected. The CNN was implemented in MATLAB within the Deep Neural Network Toolbox.

Although the window could be moved one sample at a time, for training it was shifted using a stride of 80 samples (5ms), to avoid generating an excessive amount of training patterns and thereby ensuring the dataset fitted within memory on the graphic card used to implement network training. During training, the order of the patterns was randomized.

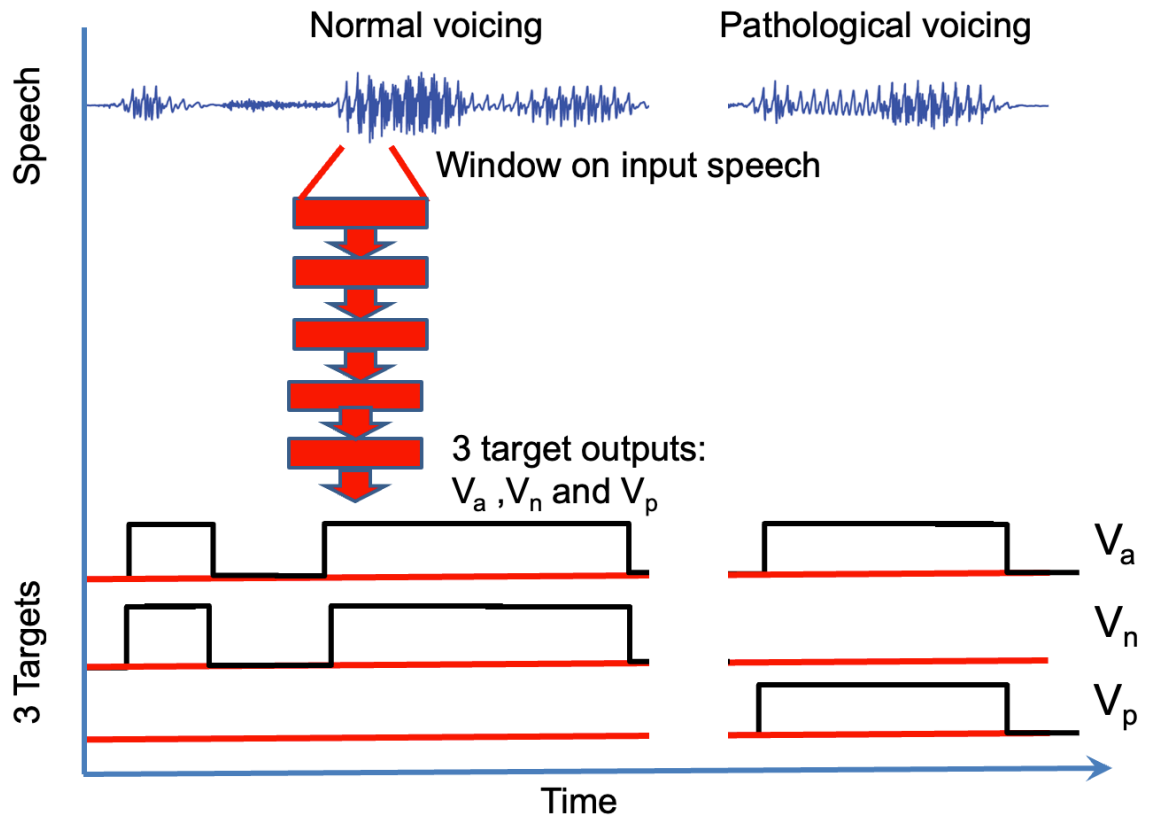

Figure 2. *Schematic illustrates training a neural network to estimate voice pathology by mapping a sliding window of speech signal samples to multiple output labels representing 3 target outputs: all voicing, normal voicing and pathological voicing. Two examples of speech are shown, one with normal and the other with pathological voicing, along with their respective output targets.*

Figure 3. *Deep neural network structure implemented using MATLAB deep learning toolbox.*

Training was performed on a Windows 10 PC fitted with an NVIDIA GEFORCE RTX 2080TI graphics card. Cross-fold validation was used to estimate task performance, to which end the data files were randomly split into 5 sets. The network was trained on 4 sets of data and performance then estimated using the 5th set. This procedure was carried out 5 times, each time using a different validation set and the overall mean performance across the 5 sessions was calculated. Each run took about 25 minutes to complete. After training, EGG was no longer required to estimate pathological voicing, since it was only used here to label the training dataset.

5 Results

5.1 Operation on normal speech

During testing, data was processed one participant at a time, and pattern order was maintained (not randomized) to ensure CNN output could be related to its corresponding speech and EGG

signals. Processing a single data file at a time enabled the window stride to be reduced to 40 samples (2.5 ms) to increase output resolution. The trained CNN provided a fine-grained indication of the presence and nature of normal and potentially pathological phonation. Outputs from the network on previously unseen speech data from a normal participant are shown in Fig. 4. It can be seen there is a strong agreement between the corresponding EGG signal (trace B) and the estimation of all voicing by the network (trace C) and generally with normal voicing (trace D). In this specifically chosen speech section to illustrate the phenomenon, the trained CNN pathological output (trace E) also responded to features in normal speech in which there is, amongst other things, vibrational irregularity. This illustrates the fact that even during normal speech production by healthy individuals, there are occasionally excitations that show similarity to pathological conditions.

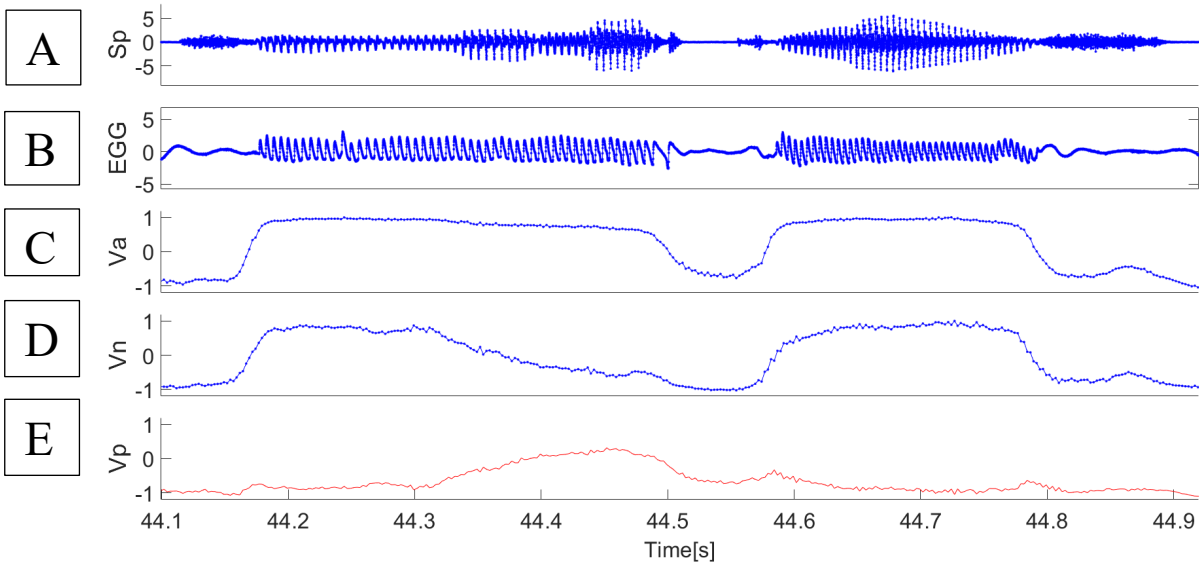

***Figure 4**: Voice quality estimation for a normal voice male speaker for the except from the Arthur the Rat passage "till then at last". **A** Speech waveform, **B** EGG signal, **C** CNN estimation of all voicing **D** CNN estimation of normal voicing, **E** CNN estimation of pathological voicing, which sometimes indicates irregular voicing consistent with pathology e.g., from about 44.3 to 44.5s.*

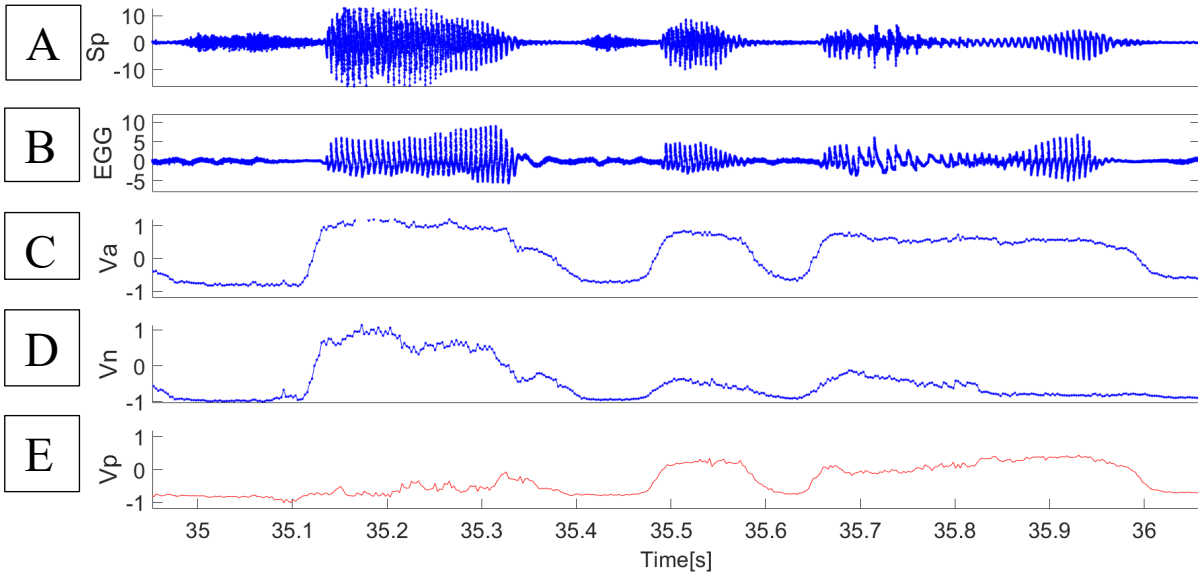

***Figure 5**: Voice quality estimation for a pathological voice female speaker for the except from the Arthur the Rat passage "say to him". (**A-E** as in Fig. 4). Some regions of voicing also contained regions where phonation appears normal exhibiting regular excitation, as indicated in the normal voice CNN output trace e.g., from 35.1 to 35.4s.*

5.2 Operation on pathological speech

Output of the network on previously unseen speech data from a participant who exhibits pathological voicing is shown in Fig. 5. It can be seen that for this unseen speaker there is a significant indication of voicing consistent with pathology (trace E). Interestingly, again in this specifically chosen speech section to illustrate the phenomenon, there is also an indication of normal voicing for regions of the speech signal as well (trace D). This illustrates the fact that speech data from patients who suffer from pathological voice conditions may also include output in which phonation appears normal.

5.3 Estimating neural network detector performance

We used a simple method to quantify voice pathology for each individual speaker in the evaluation data. To do so, for each evaluation participant recording in turn, the CNN all voicing detector output Va was first compared with a threshold value of zero (the mid-range of the training target detector outputs) to select CNN output samples for which there was evidence of voicing. For voiced regions, we then calculated the area under each output detector curve from its lower -1 target boundary. This involved adding 1 to each sample value and then calculating the mean across all values for a given participant. For each individual participant this yielded single-value estimates of any voicing present (area under Va curve), normal voicing present (area under Vn curve) and pathological voicing present (area under Vp curve).

To make an automatic judgement of voicing pathology for a given participant, we only made use of the area under Vp curve. To quantify detector performance, the receiver operating characteristic (ROC) of the CNN pathological speech detector estimates across the evaluation dataset was generated (see Fig. 6). The area under the ROC curve (AUC) was calculated by numeric integration yielding an AUC value of 0.95, indicating good detector performance. The classification of pathological versus normal voicing in the validation sets was made when the pathological detection estimate for a given speaker exceeded a fixed threshold value. It achieved an overall best detector recognition rate of 91% for the distinction between normal and pathological voicing for a threshold value of 0.342. This recognition rate compares well with studies made in other investigations [26].

6 Discussion

6.1 Summary

We trained a neural network (CNN) to estimate voicing pathology from speech, innovatively making use of EGG to label the training data for the presence of voicing. We found that we were able to generate a running estimate of both normal voicing and voice pathology with high temporal resolution. We suggest that a running display of such values may be useful clinically to help make a diagnosis. We found that there are features in normal speech which arise in specific prosodic and discourse structures, such as vibrational irregularity or breathiness, that appear to be markers for pathology. The trained CNN indicated the occasional presence of such features even in speech from speakers with normal voicing. Conversely, we also observed that speech from patients with voice pathologies also often generate spoken output where phonation appears normal.

By integrating the estimates of pathological voicing over the recordings of each individual participant, we were able to generate an overall estimate of pathology for a given speaker. This could form the basis of a system that can generate an automatic estimate of voice pathology just on the basis of a spoken text. Taken together, these results demonstrate the potential of the analysis for detailed clinical assessment and quantification of pathological voicing in connected speech. Overall, we believe our analyses provide a means to facilitate discrimination between normal and pathological voice conditions, with the final diagnosis left to the clinician. We

achieved an overall recognition rate of 91% for the distinction between speakers with normal and pathological voicing.

6.2 Future work

We note that in the current work, voice pathology estimations are only made on the basis of voiced regions automatically identified in the acoustic speech signal and without regard to the phonetic content of the spoken material. It would also be helpful to examine the results of analysis for specific phonetic contexts that are known by speech professionals to best show the manifestations of voice pathologies. The ability to perform contextually sensitive analysis has recently been shown to assist the discrimination of voice pathology [27] and future work will examine our CNN pathology detector within specific phonetic contexts.

Although here we only investigated the presence or absence of two conditions of voice pathology, in the future it would be possible, using our new approach, to examine the presence of and discriminate between different pathological conditions, as well as differentiating between different levels of severity of pathology.

As a final point it is clear that machine learning approaches to voice pathology estimation rely on the availability of large amounts of representative data. Currently the data available for continuous speech is limited, and we suggest the generation of a large database of normal and pathological speech will be needed to accelerate future progress in this area. Very large databases would also enable training of more sophisticated neural network models, including much deeper ones than the CNN used here, as well as transformer architectures, which have been shown to generate state-of-the-art performance in many application areas [28].

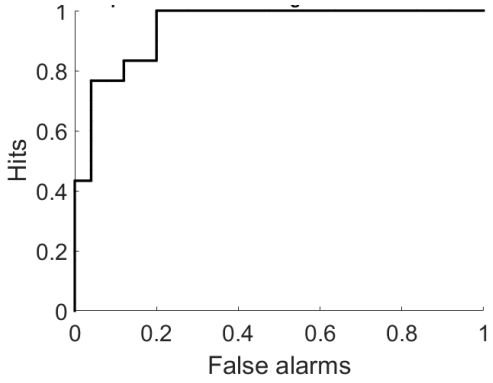

Figure 6. *ROC for all patricipant validation datasets calculated over the 5 validation runs. The AUC values for the pathology detector was 0.95.*

References

[1] J.R. OROZCO-ARROYAVE ET AL., (2015). *Characterization methods for the detection of multiple voice disorders: neurological, functional, and laryngeal diseases. IEEE journal of biomedical and health informatics, 19(6), 1820-1828.*

[2] M.R. HOFFMAN, J.J. JIANG, A.L.RIEVES, K.A. MCELVEEN, & C.N. FORD, (2009). *Differentiating between adductor and abductor spasmodic dysphonia using airflow interruption. The Laryngoscope, 119(9), 1851-1855.*

[3] B.C. WATSON, S.D. SCHAEFER, F.J. FREEMAN, J. DEMBOWSKI, G. KONDRASKE & R, ROARK (1991). *Laryngeal electromyographic activity in adductor and abductor spasmodic dysphonia. Journal of Speech, Language, and Hearing Research, 34(3), 473-482.*

[4] N. ROY, (2010). *Differential diagnosis of muscle tension dysphonia and spasmodic dysphonia. Current Opinion in Otolaryngology & Head and Neck Surgery, 18(3), 165-170.*

[5] N. ROY, S.C. MAUSZYCKI, R.M. MERRILL, M. GOUSE & M.E. SMITH (2007). *Toward improved differential diagnosis of adductor spasmodic dysphonia and muscle tension dysphonia. Folia Phoniatrica et Logopaedica, 59(2), 83-90.*

[6] C.J REES, P.D. BLALOCK, S.E. KEMP, S.L. HALUM, & I.A. KOUFMAN, (2007). *Differentiation of adductor-type spasmodic dysphonia from muscle tension dysphonia by spectral analysis. Otolaryngology—Head and Neck Surgery, 137(4), 576-581.*

[7] J.M. BARKMEIER & J.L. CASE, (2000). *Differential diagnosis of adductor-type spasmodic dysphonia, vocal tremor, and muscle tension dysphonia. Current Opinion in Otolaryngology & Head and Neck Surgery, 8(3), 174-179.*

[8] J. OATES, (2009). *Auditory-perceptual evaluation of disordered voice quality. Folia Phoniatrica et Logopaedica, 61(1), 49-56.*

[9] PATEL, R., CONNAGHAN, K., FRANCO, D., EDSALL, E., FORGIT, D., OLSEN, L., RAMAGE, L., TYLER, E., & RUSSELL, S. (2013). *"The Caterpillar": A novel reading passage for assessment of motor speech disorders. American Journal of Speech-Language Pathology, 22(1), 1–9. https://doi.org/10.1044/1058-0360(2012/ 11-0134).*

[10] A.J. FOURCIN & E. ABBERTON, *First applications of a new laryngograph, The Volta Review, 74(3), 161–176.*

[11] A.J. FOURCIN & M. PTOK, (2003). *Closing and opening phase variability in dysphonia. Fraunhofer IRB Verlag.*

[12] A.J. FOURCIN, J. MCGLASHAN, & R. BLOWES, (2002). *Measuring voice in the clinic-Laryngograph® Speech Studio analyses. In Proceedings 6th Voice Symposium of Australia.*

[13] D. MICHAELIS, T. GRAMSS & H.W. STRUBE, (1997). *Glottal-to-noise excitation ratio–a new measure for describing pathological voices. Acta Acustica united with Acustica, 83(4), 700-706.*

[14] G. SCHLOTTHAUER, M.E. TORRES, & M.C. JACKSON-MENALDI. (2006). *Automatic diagnosis of pathological voices. WSEAS Trans. on Signal Proc. 2, 1260-1267.*

[15] G. SCHLOTTHAUER, M.E. TORRES & M.C. JACKSON-MENALDI, (2010). *A pattern recognition approach to spasmodic dysphonia and muscle tension dysphonia automatic classification. Journal of voice, 24(3), 346-353.*

[16] P. HARAR, J.B. ALONSO-HERNANDEZY, J. MEKYSKA, Z GALAZ, R. BURGET & Z. SMEKAL, (2017). *Voice pathology detection using deep learning: a preliminary study. In 2017 international conference and workshop on bioinspired intelligence (IWOBI) (pp. 1-4). IEEE.*

[17] L. VERDE, G. DE PIETRO & G. SANNINO, (2018). *Voice disorder identification by using machine learning techniques. IEEE access, 6, 16246-16255.*

[18] S. HEGDE, S. SHETTY, S. RAI & T. DODDERI, (2019). *A survey on machine learning approaches for automatic detection of voice disorders. Journal of Voice, 33(6), 947-e11.*

[19] H. WU, J. SORAGHAN, A. LOWIT, & G. DI CATERINA, (2018). *A deep learning method for pathological voice detection using convolutional deep belief networks. Interspeech 2018.*

[20] M.A. MOHAMMED ET AL., (2020). *Voice pathology detection and classification using convolutional neural network model. Applied Sciences, 10(11), 3723.*

[21] S. HEDGE, S. SHETTY, S. RAI, AND T. DODDERI, (2019) *A survey on machine learning approaches for automatic detection of voice disorders," Journal of Voice, vol.33, no. 6, pp. 947.e11-947.e33. Vol. 6, pp. 4850-4854*

[22] S.S. WANG, CT WANG, C.C. LAI, Y. TSAO, S..H FANG., (2022) *"Continuous Speech for Improved Learning Pathological Voice Disorders". IEEE Open J Eng Med Biol.*

[23] M.A. HUCKVALE, & C. BUCIULEAC. (2021). *Automated detection of voice disorder in the Saarbrücken voice database: Effects of pathology subset and audio materials. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (Vol. 6, pp. 4850-4854). International Speech Communication Association (ISCA).*

[24] I.S. HOWARD. J. MCGLASHAN, E. ABBERTON & A.J. FOURCIN, (2021) *Automatic identification of voice pathology using deep neural networks, AQL 2021.*

[25] D. ABERCROMBIE, (1964). *English phonetic texts. Faber & Faber.*

[26] P. BARCHE, K. GURUGUBELLI & A.K. VUPPALA. (2020). *Towards Automatic Assessment of Voice Disorders: A Clinical Approach. In INTERSPEECH (pp. 2537-2541).*

[27] Z. LIU, M.A. HUCKVALE, .J MCGLASHAN, (2022) Automated *Voice Pathology Discrimination from Continuous Speech Benefits from Analysis by Phonetic Context, Interspeech Korea*.

[28] A. VASWANI ET AL. (2017). Attention is all you need. Advances in neural information processing systems, 30.1