Advanced Time-Frequency Analysis and Machine Learning for Pathological Voice Detection

Voula C. Georgopoulos
Dept. of Speech and Language Therapy
University of Patras
Patras, Greece
vgeorgop@upatras.gr

Abstract— **This paper discusses an advanced positive time-frequency analysis method with accurate zero-, first- and second-order moments combined with machine learning for detection of pathological voice signals. The time-frequency analysis is based on the Wigner-Ville distribution. The transfer learning approach is used to train a convolutional neural network. Specifically, the GoogLeNet network is trained using the normal and pathological voices of the KAY database and the results are presented.**

Keywords— time-frequency analysis, machine learning, voice disorders

I. Introduction

Voice disorders may have a significant impact on an individual's quality of life [1] and are often associated with occupational, social or health challenges. They are usually assessed by speech and language pathologists using perceptual evaluation, laryngoscopy and objective acoustic measurements (known as acoustic analysis). Perceptual evaluation, which is a primarily used by clinicians, may suffer from subjective bias. Laryngoscopy, on the other hand, is an invasive assessment procedure. Acoustic analysis of voice includes a series measures such as fundamental frequency (f0), vocal intensity (measured in dB), jitter (perturbations of the voice signal period), shimmer (perturbations of the voice signal amplitude), noise to harmonics ratio (NHR) as well as more recently the cepstral peaked prominence (CPP) measure. The appeal of the application of acoustic analysis in voice pathology detection is due to its non-invasive nature. However, limited consensus has been reached on the utility of the various acoustic measures for discriminating between normal and pathological voice samples. In other words, a patient with a diagnosed voice disorder may present with measurements within the normal range, whereas, on the contrary, a patient without a voice pathology or perceived voice problem may have measures that are out of normal range. This has led to questioning of the clinical utility of many of these measures [2].

The appeal for automated voice pathology detection on a voice signal, lies on non-invasive and non-subjective approaches. Various signal analysis approaches, such as Mel Frequency, Cepstral Coefficients (MFCC), Wavelets, Modulation Spectrum [3-9] have been used on a variety of voice databases to analyze normal and pathological voice signals. Also, research on neural network/Machine learning techniques including Multilayer Neural Networks, Support vector Machines, Probabilistic Neural Networks, Deep Learning Neural networks [10-14] have been implemented to detect voice pathologies from the analyzed signals. Most approaches rely on feature extraction of the analyzed signal with features subsequently fed into a classifier.

The approach taken here is to directly use a time-frequency distribution of the voice signal and a deep learning classification method to automatically classify voice signals as normal or pathological. The time-frequency distribution is used as an image representation of the signal. The classification method is based on transfer learning of a convolutional neural network that has been trained on a very large set of images unrelated to this problem. The convolutional neural network is a self-feature extracting method which learns to extract features while training.

A variety of time-frequency distributions have been used in signal analysis, such as the well-known spectrogram (a widely used distribution in speech analysis), wavelet transforms and the Wigner-Ville distribution (WVD). The spectrogram is a smoothed version of the WVD. The WVD is a high-resolution representation in both time and frequency of non-stationary signals, as is the voice signal. Additionally, the WVD satisfies the time and frequency marginals of the signal, i.e., the instantaneous power in time and energy spectrum in frequency and the total energy of the signal in the time and frequency plane as well as all higher-order signal moments. The disadvantage of the WVD is that it is difficult to interpret due to the cross-terms and that it also can have negative values. Smoothing of the WVD causes the conditions of the marginals to not be satisfied.

Therefore, using a time-frequency distribution that retains certain moment measures and specifically, zero-, first-, and second-order signal moments, while remaining positive and having limited cross-terms, is desirable as an input to an automated pathological voice detection based on time-frequency information and self-learning.

This paper is organized in five sections including this introduction. Section II describes the time-frequency methodology used. Section III describes the Machine Learning method. While sections IV and V describe the implementation and conclusions, respectively.

II. Advanced Time-Frequency Analysis

A. The Wigner-Ville Distribution

An extensive analysis of the Wigner-Ville Distribution (WVD) and its properties was provided by Claasen and Mecklenbruker in 1980 [15]. The WVD can be interpreted as

978-1-7281-6743-5/20/$31.00 ©2020 IEEE

energy-density over time and frequency and is calculated from the analytic signal:

$$W(t,f) = \int_{-\infty}^{\infty} z\left(t+\frac{\tau}{2}\right) z^*\left(t-\frac{\tau}{2}\right) e^{-j2\pi f t} d\tau \quad (1)$$

However, it suffers from cross-terms and although the energy is positive, it can take on positive and negative values, making it difficult to interpret and process for feature extraction. The appeal to keep the salient properties of the WVD (energy and moments) and at the same time ensure that it remains positive and has limited cross-terms, is high. An approach to create a time-frequency that meets these criteria is using an iterative algorithm to transform the WVD into a positive distribution.

B. Positive Transformed Wigner Distribution

For a positive, transformed Wigner-Ville distribution (PTWD) which retains accurate zero-, first-, and second-order moments the following must hold:

- Instantaneous Power (energy per time)

$$|x(t)|^2 = \int PTWD(t,f) df \quad (1)$$

- Spectral Energy (energy per frequency)

$$|X(f)|^2 = \int PTWD(t,f) dt \quad (2)$$

- Instantaneous Frequency (Hz)

$$f_i(t) = \frac{\int f \cdot PTWD(t,f) df}{\int PTWD(t,f) df} \quad (3)$$

- Group Delay (sec)

$$\tau_g(f) = \frac{\int t \cdot PTWD(t,f) dt}{\int PTWD(t,f) dt} \quad (4)$$

- Delay Spread (sec^2) – relates to localized duration

$$\langle f^2 \rangle_t = \frac{\int f^2 \cdot PTWD(t,f) df}{\int PTWD(t,f) df} \quad (5)$$

- Frequency Spread (Hz^2) – relates to instantaneous bandwidth.

$$\langle t^2 \rangle_f = \frac{\int t^2 \cdot PTWD(t,f) dt}{\int PTWD(t,f) dt} \quad (6)$$

Also, for all values of time and frequency $PTWD(t,f) \geq 0$.

An iterative algorithm based on the well-known LMS algorithm which has been extended with error vectors for all the above criteria is used. Previous work [16] presented a time-frequency distribution using an iterative algorithm for the zero- and first-order moments. Here the second order moments of the WVD, delay spread and frequency spread, have been included to further improve the cross-term reduction.

C. Voice Signals

In this section we will view examples of PTWD for both normal and pathological voices. The voice signals used were from the KAY Elemetrics (now Pentax Medical) database, developed by the Massachusetts Eye and Ear Infirmary Voice and Speech Lab. This database contains 57 normal and 653 disordered samples. The samples used for analysis here is sustained phonation of vowel /a/.

Figure 1 shows the time-frequency analysis for a normal female voice patient (age 29 and non-smoker). On the left of the 3rd row of Fig.1 the acoustic parameters of the voice signal (f0, a RAP - measure of jitter, shimmer, and two noise measures: Noise to Harmonics Ratio and Voice Turbulence Index). The voice signal is next to the parameters and on the right is the original WVD. As a contrast the spectrogram is shown on the left of the 1st row of Fig. 1. In the center of the 2nd row the PTWD is shown. Comparing with the WVD it is much more localized due to the cross-term reduction and, as expected, it shows more detail that the spectrogram, which is a smoothed version of the original WVD. Also, shown in the figure are the zero- and first-order moments, envelope squared (1st row center), energy spectral density (2nd row left), instantaneous frequency (2nd row right) and group delay (1st row right). The second-order moments are not shown, since they are used primarily for improving localization of the PTWD in terms of cross-term reduction and are not easily interpreted within this context of time/frequency space.

Figure 2 shows the time-frequency analysis for a pathological female voice patient (age 53 and smoker). The diagnosis is hyperfunction, paresis (unilateral right), vocal tremor. Similarly, acoustic parameters of the voice signal, the voice signal, the WVD, the spectrogram, the PTWD and the zero- and first-order moments are shown.

It should be noted that the acoustic measurements f0, jitter RAP%, Shimmer and NHR are in the normal range for both patients. Therefore, these typically used acoustic measures do not indicate pathology in the voice sample of the second patient. This illustrates the need for alternative signal analysis methods.

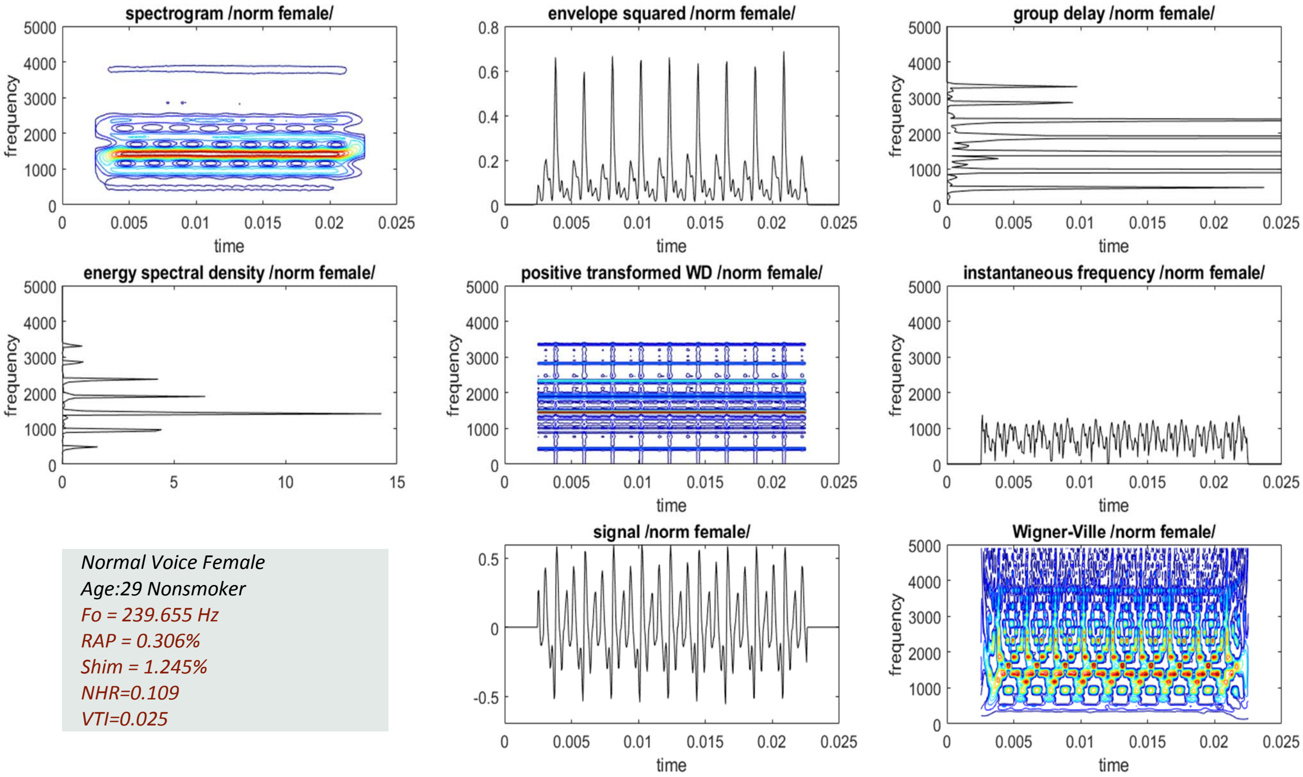
Since the PTWD is a 2-D distribution, it can be considered an image which can be classified using machine learning techniques.

III. Machine Learning

Machine learning algorithms can figure out how to perform important tasks by generalizing from examples and a task that they are extensively used for, is classification [17]. Deep learning (DL) is a subset of machine learning dealing with algorithms of many stages of artificial neural networks [18]. A type of deep learning network, particularly suitable for image classification, is a Convolutional neural network (CNN). CNNs can learn extremely complex mapping functions if provided with sufficient (large) amount of data. In essence, a convolutional network performs self-feature extraction while learning to classify data [19].

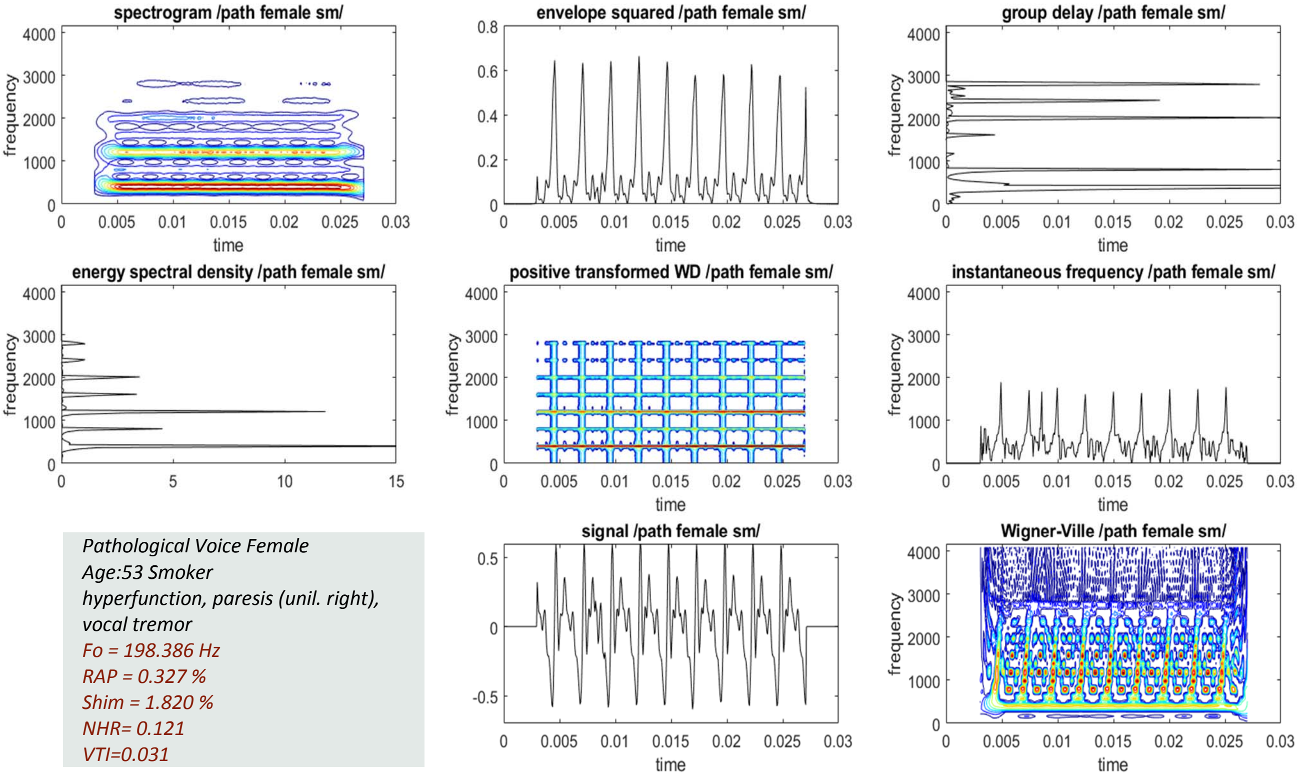
The basic building blocks of convolutional networks are weights that consist of filters. The filters that are of matrix size $n \times n$ perform convolutions through the image. The training of the CNN involves finding the values of the $n \times n$ filters so that at the output the correct classification is achieved. Deep learning CNN approaches require a large amount of data and high-performance hardware.

Since the number of images in the KAY Elemetrics database is limited, it would be very challenging to design a CNN to perform the required classification task. This issue however can be surpassed by using transfer learning. In transfer learning, a pretrained deep learning network on a large data set can be used by finetuning it to learn a new task [20]. In addition to not requiring a large dataset, transfer learning uses pretrained weights and the only learning weights that need to be learned are those in the final layers that are adjusted according to the problem. Transfer learning techniques have



Normal Voice Female
Age:29 Nonsmoker
Fo = 239.655 Hz
RAP = 0.306%
Shim = 1.245%
NHR=0.109
VTI=0.025

Fig. 1. Advanced Signal Analysis Normal Voice Sample



Pathological Voice Female
Age:53 Smoker
hyperfunction, paresis (unil. right),
vocal tremor
Fo = 198.386 Hz
RAP = 0.327 %
Shim = 1.820 %
NHR= 0.121
VTI=0.031

Fig. 2. Advanced Signal Analysis Pathological Voice Sample

been used in a variety of diagnosis medical fields such as voice pathology [21], mammography [22] and electrocardiograms [23].

GoogLeNet [21] is a well-trained CNN on large-scale natural images available in ImageNet (a large visual database designed in visual object recognition research consisting of 1.2 million images of 1000 classes). It contains 22 layers and 5 million parameters. The output layer of GoogLeNet consists of 1000 neurons that correspond to the 1000 classes it has been pretrained for.

To use GoogLeNet for our application the following steps are required:

1. Replace of the ouput layer to correspond to the number of desired classes.
2. Use the initial values of weights and bias for the pretrained part of the network
3. Set the training parameters for the GoogLeNet.
4. Train the "new" network with the training, validation and testing datasets and evaluate the performance of the network.

IV. Implementation

For all the voice samples from the KAY database a PTWD was performed and stored into a database of images. Since the database consists of 57 normal and 653 disordered patients, which is a disproportionate difference in size of normal to disordered patients, 50 normal subjects and 100 pathological patients were chosen randomly from the five most frequent diagnoses in the database. The image database of PTWD was used to train the GoogLeNet using transfer learning. To meet the requirements GoogLeNet the images were resized images from 125×501 to 224×224.

The output (last) layer of GoogLeNet was changed from 1000 nodes to 2 to correspond to the two decisions normal voice and pathological voice.

During training, 70% of normal and pathological voices were randomly chosen within each category, while similarly, 15% were used for validation and 15% for testing. Figure 3 shows the block diagram of the procedure.

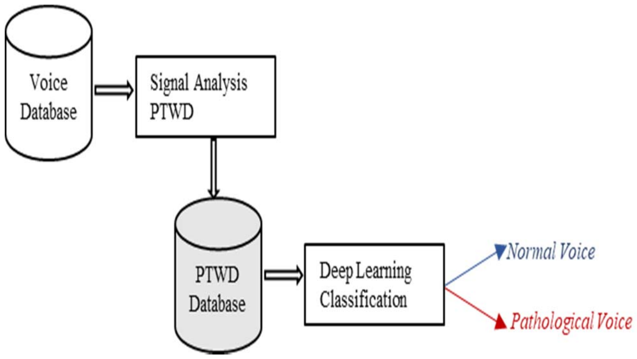


Fig. 3 Block Diagram of the Pathological Voice Detection System

The results of the training set were found 74% whereas for the validation training set 69%.

V. Conclusion

An algorithm for pathological voice detection was presented based on advanced time frequency signal analysis and transfer deep learning. The time frequency analysis was based on a positive transformed WD maintaining temporal and frequency localization while reducing cross-terms. The results show that a deep neural network, well-trained on a high volume of natural images, can be used for pathological voice detection. The results are in close agreement with deep learning approaches using other databases for voice pathology classification [14].

Although the performance is adequate as a first approach, improvements will be made by changing additional the layers of the CNN as well as fine tuning parameters. Additionally, other databases with larger sets will be tried in order to extend the classification to types of voice disorders.

References

[1] S.M. Cohen, "Self-reported impact of dysphonia in a primary care population: an epidemiological study," Laryngoscope, vol 120, pp. 2022-32, October 2010.

[2] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: A meta-analysis," The Journal of the Acoustical Society of America, vol. 126, pp.2619-2634, November 2009.

[3] J.L. Godino-Llorente, P. Gomez-Vilda and M.B. Velasco, "Support Vector Machines Applied to the Detection of Voice Disorders," Lecture Notes in Computer Science, vol. 3817, pp. 219-230, April 2005.

[4] E.S. Fonseca and J.C. Pereira, "Normal versus pathological voice signals," IEEE Engineering in Medicine and Biology Magazine, vol. 28, pp.44-48, 2009.

[5] M. Markaki and Y. Stylianou, "Voice Pathology Detection and Discrimination Based on Modulation Spectral Features," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, pp. 1938-1948, September 2011.

[6] L. Salhi, M. Talbi, S. Abid, and A. Cherif, "Performance of wavelet analysis and neural networks for pathological voices identification," International journal of electronics, vol.98, pp.1129-1140 A, September 2011.

[7] X. Wang, J. Zhang, and Y. Yan, "Discrimination between pathological and normal voices using GMM-SVM approach," Journal of Voice, vol.25, pp.38-43, January 2011.

[8] M.K. Arjmandi and M. Pooyan, "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine," Biomedical Signal Processing and Control, vol. 7, pp. 3– 19, January 2012.

[9] N. Cavalcanti, S. Silva, A. Bresolin, H. Bezerra, and A.M.G. Guerreiro, "Comparative analysis between wavelets for the identification of pathological voices," in Iberoamerican Congress on Pattern Recognition, pp. 236-243. Springer, Berlin, Heidelberg, November 2010.

[10] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A Survey on Machine Learning Approaches for Automatic Detection of Voice Disorders," Journal of Voice, vol. 33, pp.947.e11-947.e33, November 2019.

[11] R. Islam, M. Tarique and E. Abdel-Raheem, "A Survey on Signal Processing Based Pathological Voice Detection Techniques," IEEE Access, vol. 8, pp. 66749-66776, 2020, doi: 10.1109/ACCESS.2020.2985280.

[12] S. H. Fang, Y. Tsao, M. J. Hsiao, J. Y. Chen, Y. H. Lai, F. C. Lin, and C. T. Wang, "Detection of pathological voice using cepstrum vectors: A deep learning approach," Journal of Voice, vol.33, pp. 634-641, September 2019.

[13] V.C. Georgopoulos and F. Bokari, "Advanced Time-Frequency Analysis & Support Vector Machines for Pathological Voice Detection," Poster presented at the 31st World Congress of the International Association of Logopedics and Phoniatrics (IALP) Taipei, Taiwan, 18-21 August 2019.

[14] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget and Z. Smekal, "Voice Pathology Detection Using Deep Learning: a Preliminary Study," 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), Funchal, July 2017, pp. 1-4, doi: 10.1109/IWOBI.2017.7985525.

[15] T. A. C. M. Claasen and W. F. G. Mecklenbrauker, (1980). The Wigner distribution—A tool for time-frequency signal analysis. Philips J. Res, 35, 217-250, 276-300, 372-389.

[16] D. Preis and V.C. Georgopoulos, "Wigner distribution representation and analysis of audio signals: An illustrated tutorial review," Journal of the Audio Engineering Society, vol.47, pp.1043-1053, December 1999.

[17] P. Domingos, "A few useful things to know about machine learning," Communications of the ACM, vol. 55, pp.78-87, October 2012.

[18] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural networks, vol. 61, pp. 85-117, January 2015.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[20] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[21] H. Wu, J. Soraghan, A. Lowit, and G. Di-Caterina, "A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks," Proc. Interspeech 2018, pp. 446-450, DOI: 10.21437/Interspeech.2018-1351. .

[22] B. Q. Huynh, H. Li, and M. L. Giger. "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," Journal of Medical Imaging, vol. 3, pp. 034501:1 -6, July 2016.

[23] J. H. Kim, S. Y. Seo, C. G. Song, and K. S. Kim, "Assessment of Electrocardiogram Rhythms by GoogLeNet Deep Neural Network Architecture," Journal of healthcare engineering, 2019. https://doi.org/10.1155/2019/2826901