DETECTION OF PATHOLOGICAL DISEASES USING A PARAMETRIC MODEL OF VOCAL FOLDS AND NEURAL NETWORKS

P. Chytìl1,2, C. Jo3, K. Drake4, D. Graville4, M. Wax4, M. Pavel1
1Biomedical Engineering Department, Oregon Health & Science University, Portland, Oregon, USA
2Facility of Electrical Engineering and Communications, Brno University of Technology, Czech Republic
3School of Mechatronics, Changwon National University, Changwon, Gyeongnam, Korea
4Department of Otolaryngology, Oregon Health & Science University, Portland, Oregon, USA

Abstract: There are a number of clinical conditions that affect directly or indirectly the function of the vocal folds and thereby the pressure waveforms of elicited sounds. If the relationships between the clinical conditions and the voice quality are sufficiently reliable, it should be possible to detect these diseases or disorders. The focus of this paper is to determine the set of features and their values that would characterize the speaker’s state of vocal folds. To the extent that these features can capture the anatomical, physiological, and neurological aspects of the speaker they can be potentially used to mediate an unobtrusive approach to diagnosis. We will show a new approach to this problem, supported with results obtained from two disordered voice corpora.

Keywords: Model, glottal pulse, pathological voice

I. INTRODUCTION

Production of voice is influenced by the cognitive, neurological and physical state of speaker. In fact, voice production depends on the precise interaction of many components including anatomical, physiological and neural aspects of the body. It is, therefore, not surprising that voice characteristics would be affected by a wide range of disorders and diseases. Hormonal imbalance, neurological disturbances, lung disease, and mental functioning can influence and often interfere with the ability to produce a clear and intelligible voice. Conversely, it should be possible to use acoustical analysis of signals generated by patients to assess the health and the mental state of the patient.

Existing attempts for voice-based diagnosis have been based on features which are only remotely connected to the physical characteristics of the vocal folds. We describe a new method to estimate vocal fold dynamics using a parametric model of glottis movements in order to assess the health of the vocal folds and detect pathological conditions of the larynx. This approach would ultimately enable clinicians to assess and diagnose individuals using only their vocalizations. Although the sensitivity and specificity of the diagnosis are likely to be limited, this is a very feasible approach for triaging individuals for further testing and treatment. We envision that in the future this diagnosis can be performed over the telephone. Therefore, the analysis would be conducted as an unobtrusive exam and would contribute to the comfort of the patient.

II. METHODS

Our general approach to the diagnosis of larynx pathologies consisted of two phases: (1) estimation of the vocal tract transfer function \( H(\omega) \) and the pitch \( F_0 \) and (2) estimation of the parameters of the best fitting glottal pulse generating model. This approach is similar to the previous work on the characterization of the quality of voice using parameters of the Fant model [1]. The vocal tract transfer function was estimated assuming that the frequency distribution of the glottal pulse within the relevant region was approximately constant. Given this estimate, we found the parameters of a mathematical model that would maximize the correspondence between the observed and synthetically generated utterances filtered by \( H(\omega) \). The resulting computed speech signal model was fitted to the speech signal. In particular, the best-fitting set of models’ parameters was then estimated for each subject’s data by maximizing the correlation between the computed and the actual signals prior to the lip transformation. The optimization was performed using the Nelder-Mead simplex search method, because of the complex error surface due to nonlinearities, discontinuities and the complex interactions among the model parameters. The block diagram of the process of estimation of the parameters is shown in Fig. 1.

In this study we report the results based on the

![Image](http://www.isca-speech.org/archive)

Fig. 1. Block diagram of the parameters estimation process.
Fujisaki-Ljungqvist (FL) model [2]. The glottal flow and its derivative of this model are represented by polynomial segments. The choice of a polynomial model provides a convenient way to vary the number of parameters, which is useful for evaluating their relative importance. In its most elaborate form, the model has three timing parameters controlling open phase duration, pulse skew and the time interval from glottal closure to maximum negative flow. In addition, there are three amplitude-related parameters controlling the slope at glottal opening \( A \), the slope prior to closure \( B \) and the slope following closure \( C \). Although the offset parameter \( A \) (see Fig. 2) has not been in prior applications, we have included it since a secondary excitation can often be observed at glottal opening. The rounded closure, that is often evident in the glottal flow waveforms, is sometimes attributed to the lowering of the vocal cords following the glottal closure. The mathematical representation of the glottal flow in the FL model is given by:

\[
E(t) = \begin{cases} 
\frac{A - 4ARt + AR^2}{t^2} & 0 < t < R \\
\frac{\alpha(t-R)^2 - 2\beta(t-R)^3 - 2Bt - F_0 t^3}{F_0^2} & R < t < W \\
\frac{C(t-W)^2 - \beta (t-W)^3}{D(t-W)^2} & W < t < W+D \\
W+D < t < T 
\end{cases}
\]

where \( W = R + F \) and \( \alpha, \beta \) are defined by:

\[
\alpha = \frac{4AR + 6FB}{2R^2 - F_0^2}, \quad (2)
\]

\[
\beta = \frac{CD}{D - 3(T - W)}. \quad (3)
\]

![Fig. 2. One period of FL model showing glottal flow \( U(t) \) and glottal flow derivative \( E(t) \) and its parameters.](image)

We have evaluated three different methods of obtaining the estimation of the glottal pulse from the speech signal:

1) Estimation process of the vocal tract using the LPC coefficients [3], based on a questionable assumption that the glottal pulse is a pulse train with a uniform spectrum. The order of the LPC coefficients was selected to 32 (for sampling frequency of 16 kHz) to be sufficiently high to characterize the vocal tract, and yet sufficiently low not alter the shape of the glottal pulses.

2) Cepstral [4], in which the vocal tract estimation is based on the notion that the frequency ranges of the vocal cord filtering action and the glottal forcing functions do not overlap. This method uses homomorphic filtering, whereby the multiplication of the transfer functions is transformed into an addition as a consequence of the logarithmic transformation. In particular, this method is based on cepstral filtering – liftering.

3) Interactive Adaptive Inverse Filtering [5], where the vocal tract transfer function is estimated by minimizing the contribution of the average glottal pulse. This method iterates two phases. The first phase generates an estimate of the glottal excitation, which is subsequently used as input of the second phase that generates a more accurate estimate. Typically, the inverse filtered signal is no longer than a couple of hundreds of milliseconds to ensure minimal changes in the vocal tract transfer function.

The evaluation process involved estimation of the model parameters, constructing feature vectors and using those for the classification of the voice samples. The vector of features included the maximum value of the correlation function, the amplitude normalized parameters of a model \((A, B, C)\) and the temporal parameters \((R, F, D)\) multiplied by \(F_0\). These features were estimated using three samples of 32ms window from a modeled subject at the 100ms, 200ms and 300ms from the beginning of the utterance. The classification was implemented with a feed-forward back-propagation network using gradient descent error for learning. The topology of the neural network comprised one input layer, one layer of hidden units and one output layer. A separate network was used for each estimation technique: number of inputs depends on the model, \(m\) hidden units and 1 output unit. The number of hidden units is in the range of 5-54. This neural network approach was chosen because of its computational efficiency, performance and simplicity.

III. RESULTS

In order to evaluate this approach, we used the Kay Elemetrics Disordered Voice Database [6], that comprises over 1,400 voice samples of approximately 700 subjects and includes sustained phonation and running speech samples from patients with a wide variety of organic, neurological, traumatic, and psychogenic voice disorders, as well as from 53 normal speakers. We used only utterances with steady pronounced vowel /a/. In addition,
we used the Korean Disordered Speech Database [7] that consists of 28 benign and 31 malignant pathological speakers and 41 normal speakers. This database was collected using the Kay Elemetrics database as a template. The utterances in this database are vowels /a/, /e/, /i/, /o/, /u/. Again we used only the vowel /a/. The sampling frequency and the bit resolution is the same as in Kay Elemetrics. However we have down-sampled all the data to 16kHz for both databases. In this paper we describe the classification results obtained with databases combined.

![Image](image1.png)

Fig. 3. Maximum value of correlation for FL models with respect to method of inverse filtration.

Since the classification is based on the correspondence between the models and the data, we first present the frequency distribution of the correlations between the model and the data shown in Fig. 3. The graph shows the distribution of the maximum values of correlation of the model and particular inverse filtering method. Although this model generally fits a large proportion of the speakers, there was small number of cases with only marginal fit to the model. This was mostly due to the effects of the pathology of the glottal signal generation process. An example of the ability of the model to fit pathological speakers is shown in Fig. 4-6.

![Image](image2.png)

Fig. 4. Signal fit for a pathological speaker, solid line is real speech and dashed line is re-synthesized speech from the model (max of correlation value is 0.972).

![Image](image3.png)

Fig. 5. Signal fit for a pathological speaker, solid line is real speech and dashed line is re-synthesized speech from the model (max of correlation value is 0.932).

![Image](image4.png)

Fig. 6. Signal fit for a pathological speaker, solid line is real speech and dashed line is re-synthesized speech from the model (max of correlation value is 0.804).

The classification was performed using feed-forward neural networks trained individually for each type of diagnosis. In order to prevent over-fitting, we used a cross-validation approach to train the classifier [8]. The results of test sets are shown in Fig. 7.

IV. DISCUSSION

The results of the binary classification process are shown in Fig. 7 in terms of the proportion of correct discrimination between pathological and healthy speakers (sensitivity and specificity). We have used a confusion matrix to determine accuracy of the methods. In each case the neural network was determined using binary classification of specific pathology vs. normal. The resulting performance of the glottal pulse model in conjunction with the simple neural network classification process is commensurate with many clinical tests.

In case of “A-P squeezing” and “A-P squeezing (mild)” we found that the results of a mild case of this pathology yields worse accuracy compared to the fully
develop pathology. This is analogical to general findings, since the mild case of diseases is closer to a healthy state, therefore, it is harder to recognize it as a disease.

Also results for the case of “pathological voice – diagnosis N/A” would confirm that this approach is suitable for general detection of the pathologies since the group consists of variety of pathological voices without known diagnosis.

V. CONCLUSION

These results suggest that this method has a potential to triage pathologies in human voice and moreover, relate the values of the parameters to the state of the speech generation mechanisms. The average accuracy of detection across the pathological voice and normal voice was for LPC method 88.7%, for cepstral method 90.83% and for IAIF method 92.42%. We achieved the best average accuracy of detection across the pathological voice and normal voice using FL model with IAIF method.

REFERENCES


