Computer Aided Recognition of Vocal Folds Disorders by Means of RASTA-PLP

Ali S. M. Saudi
Arab Academy for Science, Technology & Maritime Transport (AASTMT)
College of engineering Computer engineering department, Cairo, Egypt
Tel: 20-189-515-149 E-mail: eng_alisaudi@yahoo.com

Aliaa A. A. Youssif
Arab Academy for Science, Technology & Maritime Transport (AASTMT)
College of engineering Computer engineering department, Cairo, Egypt
E-mail: aliaay@helwan.edu.eg

Atef Z. Ghalwash
Department of Computer Science, University of Helwan, Helwan, Egypt
E-mail: ghalwash@cabinet.gov.eg

Received: December 12, 2011 Accepted: December 26, 2011 Published: March 1, 2012
doi:10.5539/cis.v5n2p39 URL: http://dx.doi.org/10.5539/cis.v5n2p39

Abstract

In the context of the recognition of vocal folds disorders, the systems based on acoustic analysis are being introduced as computer aided medical diagnosis tools due to its objectivity and noninvasive nature. Acoustic analysis is a complementary tool to those methods based on direct observation of the vocal folds by laryngoscopy; also, it can be used for the evaluation of surgical operation. This paper presents a novel approach in voice pathology assessment using RASTA-PLP feature extraction method in the framework of a HMM. The proposed method then compared to other feature extraction methods such as MFCC and PLP. The experimental results show that RASTA-PLP attained 92.86% correct classification rates and AUC of 0.94 compared to 0.81 and 0.79 for MFCC and PLP respectively.

Keywords: MFCC, PLP, RASTA-PLP, HMM, AUC

1. Introduction

The laryngeal pathology has received much attention nowadays due to the modern way of life which led to an increased number of professionals whose working activity greatly depends on the use of their voice such as teachers, TV presenters, and singers; also unhealthy social habits such as smoking and too much alcohol drink may cause voice disorder. People are subjected to the risk of voice problems due to errors after surgical operations such as laser cordectomy, or Para thyroidectomy, etc.

Acoustic analysis has proved to be an excellent tool for voice disorder detection and assessment. Voice assessment techniques may be categorized into two categories: subjective and objective techniques. Ear, Nose and Throat doctors use a subjective technique, which relies on the doctor's hearing to the patient's voice which may cause errors. The objective technique based on physical measurements obtained during phonation. It includes measures of vocal fold vibratory movement, such as laryngoscopy, glottography, digital stroboscopy, electromyography and videoendoscopy (Kukharchik, Martynov, Kheidorov & Kotov, 2007). These techniques are more accurate in diagnosing various laryngeal diseases due to their ability to capture the vocal folds movements. However, they are invasive, require costly resources and require experienced professionals. Also, it may cause much discomfort and sometimes generating resistance by the patients during examination, which may cause distortions in the data and thus produce false diagnoses (Adnene, Lamia & Mounir, 2003) and (Alonso, J.,

Leon, Alonso, I., & Ferrer, 2001).

In this paper, a novel approach to recognize the presence of pathology from voice records is proposed and discussed by means of short-time parameterization of the speech signal. The automatic recognition of voice alterations is addressed by means of Hidden Markov Models (HMM) and Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP) complemented with short-term energy measurements. The proposed method is compared to other well known feature extraction methods such as Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP).

2. Related Work

Over recent years, several studies have been carried out on the automatic recognition of vocal fold pathologies by means of acoustic analysis. These works can be categorized into two groups. The first group (A) concentrated on finding the most important parameters to estimate voice quality while the second group (B) concentrated on finding the best classifier to detect the vocal fold pathology.

In group (A), most of long term voice parameters that extracted from pitch data (Benesty, Sondhi & Huang, 2008) can be divided into four categories: fundamental frequency, amplitude perturbation, frequency perturbation and noise parameters.

Regarding fundamental frequency parameters: average fundamental frequency (Davis, 1979), average pitch period, highest fundamental frequency, lowest fundamental frequency, standard deviation of the fundamental frequency, phonatory fundamental frequency range (Hirano & McCormick, 1986) and (Kasuya, Endo, & Saliu, 1993).

The amplitude perturbation parameters: amplitude perturbation (shimmer) (Kasuya, Endo, & Saliu, 1993) and (Bielamowicz, Kreiman, Gerratt, Dauer & Berke, 1996), amplitude perturbation quotient (APQ) and smoothed APQ (sAPQ) (Deliyski, 1993).

Regarding frequency perturbation parameters: frequency perturbation (jitter) (Kasuya, Endo, Saliu, 1993) and (Bielamowicz, Kreiman, Gerratt, Dauer & Berke, 1996), pitch perturbation quotient (PPQ) and smoothed PPQ (sPPQ) (Deliyski, 1993).

Noise parameters are important in detecting the presence of voice disorders since most pathological voices present some degree of noise. It includes signal-to-noise ratio (SNR) (Klingholz, 1997), harmonics-to-noise ratio (HNR) (Qi & Hillman, 1997) and (Kasuya, Ogawa, Mashima & Ebihara, 1986), normalized noise energy (NNE) (Michaelis, Gramss & Strube, 1997), voice turbulence index (VTI) (Qi & Hillman, 1997), soft phonation index (SPI) (Deliyski, 1993) and glottal-to-noise excitation ratio (GNE) (Michaelis, Gramss & Strube, 1997) and (Godino-Llorente, Ruiz, Lechon & Gomez-Vilda, 2008).

A study like (Kasuya, Ogawa, Mashima & Ebihara, 1986), the authors proposed the NNE parameter for acoustic discrimination of voice disorders obtained an accuracy of 78.6% for NNE and 74.1% for HNR. In (Godino-Llorente, Ruiz, Lechon & Gomez-Vilda, 2008), the authors evaluate the capabilities of the GNE ratio for the screening of voice disorders, reporting an accuracy of 95%. The authors in (Yumoto, Gould & Baer, 1982) proposed the HNR parameter for acoustic discrimination of voice disorders reporting an error rate of 16.7%. In (Godino-Llorente, Ruiz & Gomez-Vilda, 2009), the authors proposed a new parameter that correlated with the perceived hoarseness, giving an indication of the degree of normality. The proposed index has been named Pathological Likelihood Index (PLI) reported accuracy in the screening of voice disorders equal to 95%.

Other works indicate that an accurate screening can be carried out by using a combination of several of the aforementioned acoustic parameters. An approach found in (Hadjitodorov & Mitev, 2002), where the authors use several parameters and a new parameter called turbulent noise estimation to detect pathological voices , the system reached an accuracy of 96.1% using a k-means nearest neighbor (k-NN).

Regarding group (B), the pattern recognition methods used for the automatic detection of vocal folds pathologies range from a simple classifier such as (k-NN) or a Linear discriminant analysis (LDA), to more complex techniques such as Gaussian mixture model (GMM), Hidden markov models (HMM), Support vector machines (SVM) and Artificial neural networks (ANN); Other approaches use hybrid classifiers.

In (Ananthaknshna, Shama & Niranjan, 2004), the authors used a simple (k-NN) classifier for voice pathology detection, yielding a classification accuracy of 89.19%. In (Shama, Krishna & Cholayya, 2007), a modification of the standard k-NN classifier was proposed to classify a set of 53 normal and 163 pathological speakers extracted from MEEI database. The best accuracy obtained was 94.28% by using HNR. In (Hariharan, Paulraj, & Yaacob, 2009), simple k-NN and LDA based classifiers are used for testing the effectiveness of the

mel-frequency band energy coefficients (MFBECs) combined with singular value decomposition (SVD) based feature vector. The experiments were performed by using a subset of the MEEI database, with 53 normal and 657 pathological speakers; yielding classification accuracy of 99.59% for k-NN classifier and 98.48% for LDA classifier.

In (Godino-Llorente, Gomez-Vilda & Velasco, 2006) and (Godino-Llorente, Aguilera-Navarro & Gomez-Vilda, 2001), a probabilistic model GMM was used for classification between normal and pathological voices. In (Godino-Llorente, Gomez-Vilda & Velasco, 2006), the features used to train the classifier were Mel-Frequency Cepstral Coefficients (MFCC) along with their first derivative, obtained an efficiency of around 94% with 53 normal and 173 pathological speakers from MEEI database. In (Godino-Llorente, Aguilera-Navarro & Gomez-Vilda, 2001), the features used to train the classifier were MFCC and energy along with their first and second derivatives, obtained an efficiency of around 94% with 53 normal and 82 pathological speakers from MEEI database.

In (Dibazar, Narayanan & Berger, 2002), more complex probabilistic models, such as HMM have also been used for voice pathology detection reported different accuracies ranging from 97.75% to 98.3%. The features used in these cases are MFCC, the velocity and acceleration parameters, as well as different acoustic and noise measures.

Studies like (Godino-Llorente, Gomez-Vilda & Velasco, 2005) a discriminative classifier as SVM classifier was used to identify laryngeal pathologies. MFCC and noise features are used in yielding classification accuracy up to 95%. The study proposed in (Saenz-Lechon, Osma-Ruiz, Godino-Llorente, Blanco-Velasco, Cruz-Roldan, & Arias-Londono, 2008) considers a subset of the Kay database comprising 53 normal and 173 pathological sustained vowels. The authors investigate the performance of an automatic system for voice pathology detection when the voice samples have been compressed in MP3 format with different binary rates (160, 96, 64, 48, 24, and 8 kb/s). The feature set was MFCCs, HNR, NNE, GNE, energy, as well as their respective first derivative. The classification was performed using GMMs and SVMs classifiers. For these two classifiers, the best accuracy was 94.35 % for GMM and 93.01 % for SVM. The authors highlighted that there are no significant differences in the performance of the detector when the binary rates of the compressed data were above 64 kb/s.

In (Fraile, Saenz-Lechon, Godino-Llorente, Osma-Ruiz & Fredouille, 2009), (Godino-Llorente, Gomez-Vilda & Blanco-Velasco, 2004), (Marinus, Fechine, Gomes & Costa, 2009) and (Salhi, Talbi & Cherif, 2008) the authors have used artificial neural networks (ANN) to differentiate between different levels of pathology according to a perceptual quality voice scale. A study like (Fraile, Saenz-Lechon, Godino-Llorente, Osma-Ruiz & Fredouille, 2009) the patients were split out and differentiated by sex. The feature extraction used to train the ANN was based on MFCC yielding a classification accuracy of 88.3% with 53 normal and 173 pathological speakers from MEEI database.

In (Godino-Llorente, Gomez-Vilda & Blanco-Velasco, 2004), the authors compare between two techniques ANN and Learning Vector quantization (LVQ) in the detection of pathological voice. The feature extraction based on MFCC yielded that LVQ demonstrated to be more reliable than the MLP (Multilayer Perceptron) yielding 96% accuracy under similar working conditions with 53 normal and 82 pathological speakers from MEEI database.

In (Marinus, Fechine, Gomes & Costa, 2009), the MLP used for discrimination among normal voice, voices affected by local fold Edema and voices affected by other pathologies (nodules, cysts and paralysis). The experiments were performed by using a subset of the MEEI database with 44 pathological speakers with Edema, 23 with other pathologies such as nodules, cysts and paralysis in the vocal folds, and 53 normal. The feature extraction based on cepstral coefficients yielded a correct classification rate above 99% for normal voice, 96% for Edema and 93% for other pathologies. In (Salhi, Talbi & Cherif, 2008), the authors proposed a technique that uses wavelet analysis to extract a feature vector from speech samples, which is used as an input to a MLP classifier, yielding best accuracy of 90% with 50 normal and 50 pathological speakers from a private database.

A study like (Wang, Zhang & Yan, 2011) uses hybrid of aforementioned classifiers. The GMM-SVM is proposed and the feature set used to train the new classifier was MFCC on MEEI database yielded classification accuracy up to 96.1%.

3. Methodology

This paper proposes a system for the discrimination between normal and pathological voice based on HMM classifier. The method employed based on Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP) feature extraction technique. Then it's compared to other feature extraction methods such as MFCC and PLP. Figure 1 depicts a block diagram of the different steps carried out in the process set up for the

recognition of voice alterations. A short description of each step is presented in the following sections.

3.1 Signal Pre-processing

Before the digital speech signal can be used for feature extraction, a process called pre-emphasis is applied to emphasize the high-frequency portion of the spectrum. Pre-emphasis is accomplished by passing the signal through high-pass filter whose transfer function $H(z)$ is given by (Rabiner & Huang, 1993):

$$H(z) = 1 - az^{-1} \quad \text{where } 0.9 \leqslant a \leqslant 1 \tag{1}$$

Due to the boosting of high-frequency energy gives more information to the acoustic model, the value for the pre-emphasis parameter 'a' determined adaptively to be 0.97. Figure 2 illustrates the time representation of normal and pathological speech signal before and after pre-emphasis step.

The speech data then divided into overlapped frames of the length 20 milliseconds with frame shift interval 10 milliseconds and multiplied by Hamming windows.

3.2 Feature Extraction

Feature extraction aims at giving a useful representation of the speech signal by capturing the important information from it. A common division of the feature extraction approaches is production-based and perception-based methods. LPC is an example from the first group while MFCC, PLP, and RASTA-PLP belong to the perception-based approaches family. Since we want to simulate an experienced speech therapist who can detect the presence of a disorder just by listening to it, we'll focus on the perception-based group.

Through this approach, the recognition of voice disorders is carried out by means of short-time features. For each frame, the following features were extracted: a) 12 MFCCs, b) 12 PLPs, c) 12 RASTA-PLPs, d) the energy of the frame, e) Both, first (Δ) and second temporal derivatives (ΔΔ) extracted from each enumerated parameter. At the end, we have 9 distinct feature vectors that can be categorized into three categories according to its length. Firstly, feature vector has length 13: (12 MFCCs and Energy), (12 PLPs and Energy) and (12 RASTA-PLPs and Energy). Secondly, feature vector has length 26: (12 MFCCs, Energy, and 13 Δ), (12 PLPs, Energy, and 13 Δ) and (12 RASTA-PLPs, Energy, and 13 Δ). Thirdly, feature vector has length 39: (12 MFCCs, Energy, 13 Δ, and 13 ΔΔ), (12 PLPs, Energy, 13 Δ, and 13 ΔΔ) and (12 RASTA-PLPs, Energy, 13 Δ, and 13 ΔΔ). A brief description of these parameters is given next.

3.2.1 Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs have been calculated following a non-parametric modeling method, which is basically originated from knowledge on the human auditory perception system. These coefficients are computed for each speech frame by weighting the magnitude spectrum by a mel-filterbank. The term mel refers to a kind of measurement related to perceived frequency. The mapping between the real frequency scale (Hz) and the perceived frequency scales (mels) is approximately linear below 1 kHz and logarithmic at higher frequencies (Feijoo & Hernandez, 1990). The suggested formula that models this relationship is described as follows (Deller, Proakis & Hansen, 1993):

$$\mathrm{F_{mel}} = 2595 \cdot \log_{10}\left(1 + \frac{\mathrm{f}}{700}\right) \quad \text{where f is the real frequency (Hz)} \tag{2}$$

Then computing the log of each filter output and finally computing the Discrete Cosine Transform (DCT) of the log-mel-spectrum. The MFCCs are the resulting coefficients of this DCT operation.

3.2.2 Perceptual Linear Prediction (PLP)

The PLP feature extraction is similar to LPC analysis. It is based on short term spectrum of speech. In contrast to pure linear predictive analysis of speech, PLP modifies the short-term spectrum of the speech by several psychophysically based transformations in order to mimic human auditory system. In practice, PLP can give small improvements over MFCCs, especially in noisy environments and hence it is the preferred encoding for many systems.

3.2.3 Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP)

The RASTA approach (Hermansky & Morgan, 1994) is based on a band-pass time-filtering applied to a log-spectral representation of the speech as shown in Fgure 3, in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel. The PLP technique (just like most other short term spectrum based techniques) is vulnerable when the short term spectral values are modified by the frequency response of the communication channel. Hence RASTA methodology which makes PLP more robust to linear spectral distortions and yields better results for speech recognition tasks than PLP in noisy environment.

3.2.4 Temporal Derivatives

An improved representation can be obtained by extending the analysis including information about the temporal derivatives speed and acceleration of the parameters. This is especially important in the present case because it provides information about the short-term variability that is higher under pathological conditions (Childers & Sung-Bae, 1992).

To introduce temporal order into the parameter representation, we denote the *m*th coefficient at time *t* by c_m (*t*) (Rabiner & Huang, 1993):

$$\frac{\partial c_m(t)}{\partial t} = \Delta c_m(t) \approx \mu . \sum_{k=-K}^{K} k . c_m(t+k) \tag{3}$$

Where μ is an appropriate normalization constant and (2*K* +1) is the number of frames over which the computation of the derivative is performed. For each frame at time t, the result of the analysis is a vector of *L* coefficients, to which two *L*-size vectors corresponding to the first and second time derivatives have been appended as follows:

$$O(t) = \left(c_1(t), c_2(t), \dots, c_L(t), \Delta c_1(t), \Delta c_2(t), \dots, \Delta c_L(t), \Delta\Delta c_1(t), \Delta\Delta c_2(t), \dots, \Delta\Delta c_L(t)\right) \tag{4}$$

Where *O*(*t*) is a feature vector with 3·*L* elements.

3.3 Classification

The technique used for the classification stage was HMM. It is well known that the HMM are stochastic models that allow the representation of time series. The use of hidden states makes the model generic enough to handle a variety of complex real-world time series.

The proposed system uses the hidden Markov model toolkit (HTK Version 3.4). It was modified to accommodate the RASTA-PLP features as shown in Fgure 3. In addition, left to right HMMs, 3-state, 1-mixture were formed. The Expectation–maximization (EM) algorithm was used to train the HMM and a series of experiments were carried out with this HMM topology. In all of the experiments of this study, five training iterations were enough for good convergence of model likelihoods.

4. Experimental Results

4.1 Data Collection

To collect the voice data, the collection was done in a sound proof room of the Phoniatrics department of Kobri Elkobba Hospital. The acoustic samples correspond to sustained phonations (1-3 s long) of vowel /ah/ from patients (males and females) with normal voices and a wide variety of vocal folds disorders such as Cyst, Polyps, Nodules, Paralysis, Edemas and Carcinoma. Table 1 shows the database of vocal fold diseases. The files were obtained with low noise level, constant microphone distance around 15 cm from the talker's lips, and 22 kHz sampling rate then quantized at a resolution of 16 bits per sample. We have made our experiments on 35 voices. The HMM classifier has been trained with 60% of available speech records, the remaining 40% of records have been used for testing.

4.2 Performance Evaluation

In order to evaluate the performance of the detector and to enable comparisons to be made, several measurements (TP, TN, FP, and FN) and ratios (SE, SP, E, and AUC) were taken into account.

1) True positive (TP): The detector found an event (pathological voice) when one was present.

2) True negative (TN): The detector found no event (normal voice) when indeed none was present.

3) False positive (FP): The detector found an event when none was present

4) False negative (FN): The classifier missed an event.

5) Sensitivity (SE): Likelihood that an event will be detected given that it is present

$$\mathrm{SE} = 100 . \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

6) Specificity (SP): Likelihood that the absence of an event will be detected given that it is absent

$$SP = 100.\frac{TN}{TN + FP} \tag{6}$$

7) Efficiency (E): Likelihood that the classification is correct

$$E = 100.\frac{TN + TP}{TN + TP + FN + FP} \tag{7}$$

8) Area under curve (AUC): is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1.0.

4.3 Results

Table 2 represents the results corresponding to three independent feature extraction techniques MFCC, PLP and RASTA-PLP obtained from our private database. With respect to accuracy, it can be shown that RASTA-PLP parameters complemented with their first derivative are considered the best solution for our purpose where the accuracy reached to 92.86% and AUC equals 0.94 while the AUC of MFCC and PLP equals 0.81 and 0.79 respectively.

Looking at the results observed in Table 2, it is possible to infer that the behavior of the recognition system gets better when it is trained with RASTA-PLP features compared to MFCC and PLP features, where the recognition accuracy is reduced when the dimension of features was increased.

5. Discussion and Conclusions

The proposed scheme may be used for laryngeal pathology recognition. RASTA-PLP, PLP and MFCC feature extraction methods were used. The features are then tested with a Hidden Markov Model (HMM) classifier. Short-term RASTA-PLP complemented with the first derivative is revealed as a good parameterization approach for the recognition of voice diseases. We can conclude that the combination of the second derivatives do not show relevant influence on the results.

Anyway, a wider database of pathological voices is needed which it is not an easy work.

6. Future Work

Due to the fact that it seems to be easy to recognize voice disorders, the future work will be to identify the type of pathologies. For this purpose, the system should pass through two main steps: the first one deals with the recognition of voice disorder; once the presence is confirmed, the second step it will be voice disorder type identification.

Acknowledgment

The authors would like to thank Dr. Ahmed Eldemerdash, Phoniatrics department, Kobri Elkobba Hospital for his constant encouragement and for the grants given to this work.

References

Adnene, C., Lamia, B., & Mounir, M. (2003). Analysis of pathological voices by speech processing. *Proceedings of the Seventh International Symposium on Signal Processing and Its Applications*, 1, 365-362. http://dx.doi.org/10.1109/ISSPA.2003.1224716

Alonso, J. B., Leon, J., Alonso, I. & Ferrer, M. A. (2001). Automatic detection of pathologies in the voice by HOS based parameters. *EURASIP Journal on Applied Signal Processing*, 4, 275-281. http://dx.doi.org/10.1155/S1110865701000336

Ananthaknshna, T., Shama, K. & Niranjan, U. C. (2004). k-Means Nearest Neighbor Classifier for Voice Pathology. IEEE INDIA Annual Conference, 352-354.

Benesty, J., Sondhi, M. M. & Huang, Y. (2008). Springer handbook of speech processing, Springer-Verlag Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-540-49127-9

Bielamowicz, S., Kreiman, J., Gerratt, B. R. Dauer, M. S. & Berke, G. S. (1996). Comparison of Voice Analysis Systems for Perturbation Measurement. *Journal of Speech and Hearing Research*, 39, 126-134.

Childers, D. & Sung-Bae, K. (1992). Detection of laryngeal function using speech and electroglottographic data. *IEEE Transactions Biomedical Engineering*, 39, 19-25. http://dx.doi.org/10.1109/10.108123

Davis, S. B. (1979). Acoustic Characteristics of Normal and Pathological Voices. Speech and Language Research and Theory. Academic Press. N. J.

Deliyski, D. D. (1993). Acoustic model and evaluation of pathological voice production. in Proceedings of Eurospeech'93, Vol. 3, Berlin, Germany, 1969-1972.

Deller, J. R., Proakis, J. & Hansen, J. (1993). *Discrete-Time Processing of Speech Signals*. New York: MacMillan.

Dibazar, A. A., Narayanan, S. & Berger, T. W. (2002). Feature analysis for automatic detection of pathological speech. In Proceedings of the Second Joint EMBS/BMES Conference, vol. 1.

Feijoo, S. & Hernandez, C. (1990). Short-term stability measures for the evaluation of vocal quality. *J. Speech Hearing Res.*, 33, 324-334.

Fraile, R., Saenz-Lechon, N., Godino-Llorente, J. I., Osma-Ruiz, V. & Fredouille, C. (2009). Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficients parameters and differentiation of patients by sex. *Folia Phoniatrica et Logopaedica*, 3, 146-152. http://dx.doi.org/10.1159/000219950

Godino-Llorente, J. I., Aguilera-Navarro, S. & Gomez-Vilda, P. (2001). Automatic detection of voice impairments due to vocal misuse by means of gaussian mixture models. *IEEE 2001 Proceedings of the 23rd Annual EMBS International Conference*, 2, 1723-1726.

Godino-Llorente, J. I., Gomez-Vilda, P. & Blanco-Velasco, M. (2004). Automatic Detection of Voice Impairments by Means of Short-Term Cepstral Parameters and Neural Network Based Detectors. *IEEE Transactions on Biomedical Engineering*, 51(2), 380-384. http://dx.doi.org/10.1109/TBME.2003.820386

Godino-Llorente, J. I., Gomez-Vilda, P. & Velasco, M. B. (2005). Support Vector Machines Applied to the Detection of Voice Disorders. *Lecture Notes in Computer Science*, 3817, 219-230. http://dx.doi.org/10.1007/11613107_19

Godino-Llorente, J. I., Gomez-Vilda, P. & Velasco, M. B. (2006). Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters. *IEEE Transactions on Biomedical Engineering*, 53(10), 1943-1953. http://dx.doi.org/10.1109/TBME.2006.871883

Godino-Llorente, J. I., Ruiz, O. V., Vilda, P. G. & Gomez-Vilda, P. (2009). Pathological Likelihood Index as a Measurement of the Degree of Voice Normality and Perceived Hoarseness. *Journal of Voice*, 24(6), 667-677. http://dx.doi.org/10.1016/j.jvoice.2009.04.003

Godino-Llorente, J. I., Ruiz, O. V., Lechon, N. S. & Gomez-Vilda, P. (2008). The Effectiveness of the Glottal to Noise Excitation Ratio for the Screening of Voice Disorders. *Journal of Voice*, 24(1), 47-56. http://dx.doi.org/10.1016/j.jvoice.2008.04.006

Hadjitodorov, S. & Mitev, P. (2002). A computer system for acoustic analysis of pathological voices and laryngeal disease screening. *Medical Engineering &Physics*, 24(6), 419-429. http://dx.doi.org/10.1016/S1350-4533(02)00031-0

Hariharan, M., Paulraj, M. P. & Yaacob, S. (2009). Identification of Vocal Fold Pathology based on Mel Frequency Band Energy Coefficients and Singular Value Decomposition. *IEEE International Conference on Signal and Image Processing Applications*, 514-517. http://dx.doi.org/10.1109/ICSIPA.2009.5478710

Hermansky, H. & Morgan, N. (1994). Rasta Processing of Speech. *IEEE Transactions on Speech and Audio Proc.*, 2(4), 578-589. http://dx.doi.org/10.1109/89.326616

Hirano, M. & McCormick, K. R. (1986). Clinical Examination of Voice. *The Journal of the Acoustical Society of America*, 80(4), 1273. http://dx.doi.org/10.1121/1.393788

Kasuya, H., Endo, Y. & Saliu, S. (1993). Novel acoustic measurements of jitter and shimmer characteristics from pathological voice. *Proceedings of EUROSPEECH'93*, 1973-1976.

Kasuya, H., Ogawa, S. Mashima, K. & Ebihara, S. (1986). Normalized noise energy as an acoustic measure to evaluate pathologic voice. *The Journal of the Acoustical Society of America*, 80(5), 1329-1334. http://dx.doi.org/10.1121/1.394384

Klingholz, F. (1997). The measurement of the signal-to-noise ratio (SNR) in continuous speech. *Journal of Speech Communication*, 6, 15-26. http://dx.doi.org/10.1016/0167-6393(87)90066-5

Kukharchik, P., Martynov, D., Kheidorov, I., & Kotov, O. (2007). Vocal fold pathology detection using modified wavelet-like features and support vector machines, Proceedings of 15th European Signal Processing Conference.

Marinus, V. M. L., Fechine, J. M., Gomes, H. M. & Costa, S. C. (2009). On the Use of Cepstral Coefficients and Multilayer Perceptron Networks for Vocal Fold Edema Diagnosis. Proceedings of the 9th International

Conference on Information Technology and Applications in Biomedicine, ITAB 2009, Larnaca, Cyprus.

Michaelis, D., Gramss, T. & Strube, H. W. (1997). Glottal-to-noise excitation ratio—a new measure for describing pathological voices. *Acustica/acta acustica*, 83, 700-706.

Qi, Y. & Hillman, R. E. (1997). Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals. *J Acoust Soc Am.*, 102, 537-543. http://dx.doi.org/10.1121/1.419726

Rabiner, L. & Huang, B. H. (1993). *Fundamentals of speech recognition.* Englewood Cliffs, NJ: Prentice-Hall.

Saenz-Lechon, N., Osma-Ruiz, V., Godino-Llorente, J. I., Blanco-Velasco, M., Cruz-Roldan, F. & Arias-Londono, J. D. (2008). Effects of audio compression in automatic detection of voice pathologies. *IEEE Transactions on Biomedical Engineering*, 55(12), 2831-2835. http://dx.doi.org/10.1109/TBME.2008.923769

Salhi, L., Talbi, M. & Cherif, A. (2008). Voice Disorders Identification Using Hybrid Approach: Wavelet Analysis and Multilayer Neural Networks, World Academy of Science, Engineering and Technology.

Shama, K., Krishna, A. & Cholayya, N. U. (2007). Study of harmonics-to-noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology. *EURASIP Journal on Advances in Signal Processing*, 2007(1), 1-9.

Wang, X., Zhang, J. & Yan, Y. (2011). Discrimination Between Pathological and Normal Voices Using GMM-SVM Approach. *Journal of Voice*, 25(1), 38-43. http://dx.doi.org/10.1016/j.jvoice.2009.08.002

Yumoto, E. Gould, W. J. & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America*, 71 (6), 1544-1550. http://dx.doi.org/10.1121/1.387808

Table 1. Pathologies of Experiments

Disease	**Number**
Cyst	2
Carcinoma	9
Edema	2
Nodules	2
Normal voices	15
Paralysis	2
Polyp	3
Total	**35**

Table 2. Performance of different feature extraction methods with varying feature vector length

Vector Length	Description	Sensitivity	Specificity	AUC	Efficiency
13	12 (MFCC) + Energy	75%	83.3%	0.79	78.57%
	12 (PLP) + Energy	75%	100%	0.88	85.71%
	12 (RASTA-PLP) + Energy	87.5%	83.3%	0.85	85.71%
26	12 (MFCC+Δ)+(Energy+Δ)	62.5%	100%	0.81	78.57%
	12 (PLP+Δ)+ (Energy + Δ)	75%	83.3%	0.79	78.57%
	12 (RASTA-PLP+Δ)+(Energy+Δ)	87.5%	100%	0.94	92.86%
39	12 (MFCC+Δ+ΔΔ)+(Energy+Δ+ΔΔ)	62.5%	100%	0.81	78.57%
	12 (PLP+Δ+ΔΔ)+(Energy+Δ+ΔΔ)	75%	83.3%	0.79	78.57%
	12 (RASTA-PLP+Δ+ΔΔ)+(Energy+Δ+ΔΔ)	75%	100%	0.88	85.71%
This table contains the experimental results of vocal folds pathology detection using three different feature extraction techniques MFCC, PLP and RASTA-PLP with varying feature vector length.					

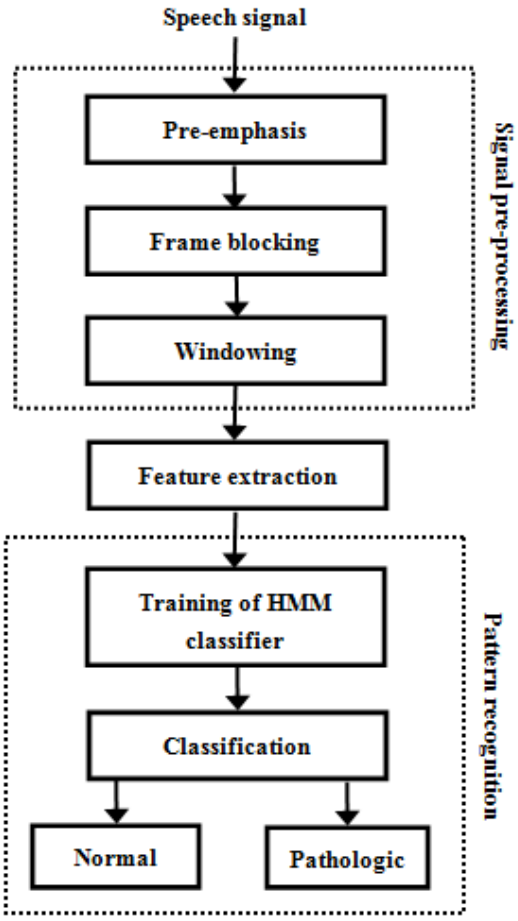
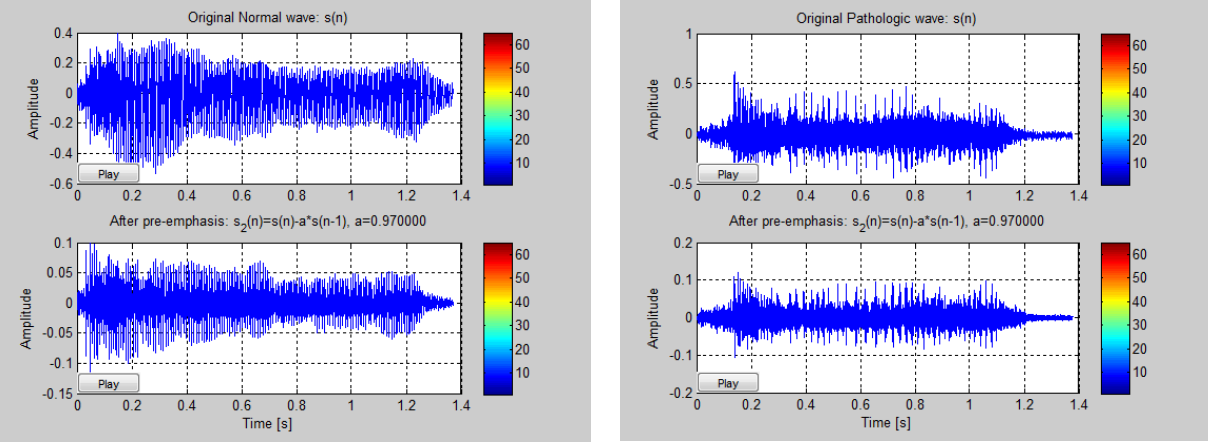


Figure 1. Block diagram of the computer aided recognition of vocal folds disorders



(a) Sustained vowel /ah/ said by a normal speaker (b) Sustained vowel /ah/ said by a pathologic speaker

Figure 2. Examples of time representation of speech signal before and after pre-emphasis

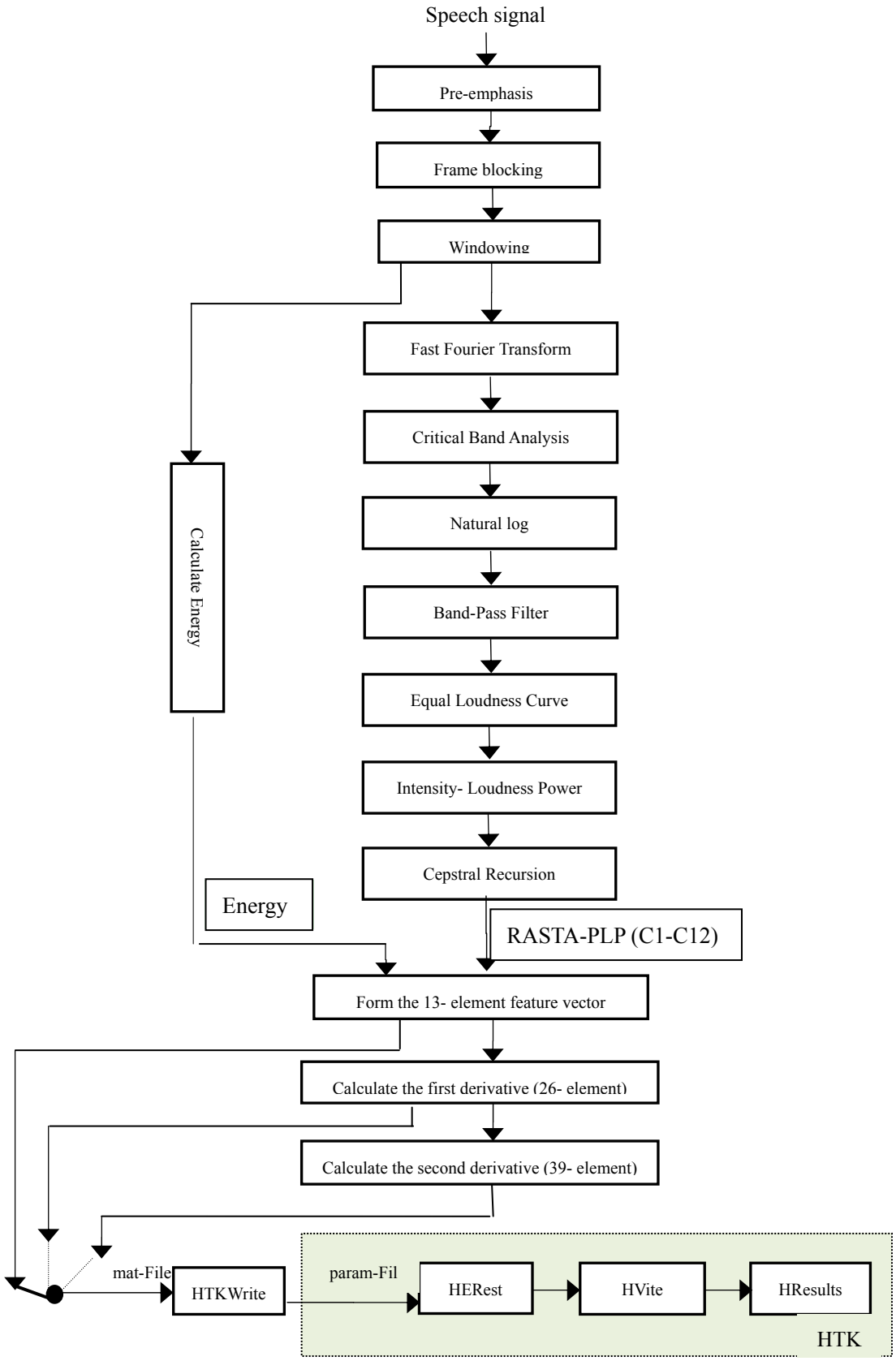


Figure 3. Block diagram of the proposed work