International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:8, No:6, 2014

Analysis of Vocal Fold Vibrations from High-Speed Digital Images Based On Dynamic Time Warping

A. I. A. Rahman, Sh-Hussain Salleh, K. Ahmad, K. Anuar

***Abstract*—**Analysis of vocal fold vibration is essential for understanding the mechanism of voice production and for improving clinical assessment of voice disorders. This paper presents a Dynamic Time Warping (DTW) based approach to analyze and objectively classify vocal fold vibration patterns. The proposed technique was designed and implemented on a Glottal Area Waveform (GAW) extracted from high-speed laryngeal images by delineating the glottal edges for each image frame. Feature extraction from the GAW was performed using Linear Predictive Coding (LPC). Several types of voice reference templates from simulations of clear, breathy, fry, pressed and hyperfunctional voice productions were used. The patterns of the reference templates were first verified using the analytical signal generated through Hilbert transformation of the GAW. Samples from normal speakers' voice recordings were then used to evaluate and test the effectiveness of this approach. The classification of the voice patterns using the technique of LPC and DTW gave the accuracy of 81%.

***Keywords*—**Dynamic Time Warping, Glottal Area Waveform, Linear Predictive Coding, High-Speed Laryngeal Images, Hilbert Transform.

I. Introduction

VOICE disorder (or dysphonia) is increasingly recognized as one of the main conditions that adversely affect an individual's quality of life [1]. Many studies suggest that voice disorders are caused by abnormal or disturbed vocal fold vibrations and functions [2]. Examination of vocal fold structures and assessment of its dynamic function are necessary for diagnosis of any voice disorders. Direct (using endoscopic laryngeal imaging) and indirect (i.e. acoustic and aerodynamic) methods of assessment provide useful information about vocal fold dynamics but question remains in terms of their interrelationship in diagnosing clinical voices or dysphonia [3].

High-speed digital imaging allows for visual observation of glottis and provides a direct means of measuring vocal fold dynamic function. The technique has the potential to describe normal and abnormal vibratory functions and few studies have acquired limited normative databases to compare abnormal functions [4], [5]. Subjective measurements such as auditory perceptual analysis and videostroboscopy, although widely used by voice therapists and physicians (collectively termed here, as voice clinicians), are less sensitive and lacks objectivity to accurately diagnose dysphonia. This problem continues to challenge voice researchers to develop more objective and practical measurement techniques or system to assist the voice clinicians in assessing voice disorders.

One of the main problems in observing vocal fold dynamics is the high speed of vocal fold movement, which is impossible for the human eye to discern. Video stroboscopy has been used to reconstruct the vibrations in slow motion but the technique is plagued with subjectivity especially with disordered voice production [6]. On the other hand high-speed recording captures thousands of frames per second and can potentially provide adequate resolution for each vibratory cycle. High-speed recording has the unrivalled advantage of capturing the actual vibrations but the challenge resides with the processing of this huge data. Recently several automated systems have been developed to overcome the drawbacks of high speed data. Objective analysis methods from Kymography [7], high speed digital recordings [8], [9] glottal area diagrams [10] and phonovibrography [11] have been shown to provide possible solutions to automated resolution of vocal fold vibrations.

An initial but vital factor to consider when automating image processing to the segmentation of the glottal area is to resolve various errors introduced during the recording like movements introduced by the recording device and/or the patient, and inclusion of onset phases of the vocal fold vibration; both of which will introduce unnecessary noise to the segment of interest. Chen et al. performed a multiple step which combined global thresholding for automatic segmentation and region-growing methods [12]. Whereas, Osma et al. used the watershed transform to segment the glottal space from unrelated laryngeal images [13]. Allin used active contour model for the segmentation of vocal folds [14]. To obtain an effective segmentation of glottal space, Mendez et al. combine the analysis of textures of Gabor Filter Bank and motion estimation techniques. [15]. Osma-Ruiz et al. filtered the images using an anisotropic diffusion technique that combined smoothing properties with image enhancement qualities [16]. Recently, methods from nonlinear systems analysis [17], Hilbert transform-based approaches [4], [5], [18] and Nyquist plots [10] were also applied, yet all the mentioned methods were limited in their ability to analyze the

This work was supported in part by the Universiti Teknologi Malaysia under Research University Grant (GUP) Tier 1 vote number Q.J130000.2545.04H21.

Ahmad Idil Abdul Rahman is with the Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia (e-mail: idil@ utem.edu.my).

Sheikh Hussain Shaikh Salleh, PhD, is with the Centre for Biomedical Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Malaysia (e-mail: hussain@fke.utm.my).

Kartini Ahmad, MB.BCh,. PhD, is with the Department of Audiology & Speech Sciences, Faculty of Health Sciences, Universiti Kebangsaan Malaysia, 50300, Kuala Lumpur, Malaysia (e-mail: kartini.ahmad@gmail.com).

Khairy Anuar Mohd Khairuddin is with the School of Health Sciences, Universiti Sains Malaysia, 16150, Kubang Kerian, Kelantan, Malaysia (e-mail: khairy@usm.my).

2-second vocal fold recording in one go without the need to manually select the image-stable segments [19].

Hilbert transform has been successfully applied to the analysis of nonlinear systems and non-stationary signals. Yuling Yan et al. [10] used the Hilbert transform as a tool to analyze and visualize the pattern of clinical voice conditions such as breathy, pressed, fry, stage whisper and hyperfunctional. The study concluded that subjects having voice problems related to their inability to sustain the oscillation of the vocal fold vibration exhibit an increased cycle to cycle variation in the amplitude or /and in the phonatory frequency, appearing in the analytic phase plot as radial scattering.

This study aims to utilize the available information thus far and perform a computer-aided decision support system to categorize groups of voice production patterns by analyzing the GAW. LPC coefficients were derived from the GAW and used as input to the DTW classifier. The classification was performed objectively using DTW with the five reference vibratory templates of voice productions such as clear, breathy, fry, pressed and hyperfunctional, simulated by a professional speaker. The result was then compared with the visual pattern of analytic signals generated through the Hilbert transform of GAW of those productions. The accuracy of DTW was subsequently tested on a group of normal speakers.

II. Methods and Theory

A. Glottal Area Waveform

The High-speed laryngeal imaging acquires images at a rate of 2000 frames/second with a spatial resolution of 160 x 140 pixels. In this study, the glottal area waveform (GAW) was used to analyze the characteristics of vocal fold vibrations for different voice productions. The GAW has been effectively used to analyze the characteristics of vocal fold vibrations. [20], [21]. The GAW is derived from images of the vocal folds and is a plot of the area of the vocal fold opening which is the glottis area, over time. To obtain the GAW an algorithm was developed using image edge detection to delineate the glottal space on a frame-by frame basis from the HSLI data sets. Segmentation of image was done by setting all pixels whose intensity values are above a threshold to a foreground value and all the remaining pixels to a background value as shown in Fig. 1.

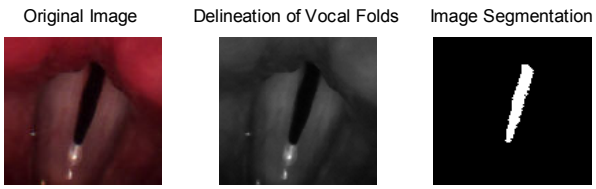

Fig. 1 Image segmentation using thresholding

B. Linear Predictive Coding

Linear predictive coding (LPC) has also been successfully used as a tool for signal pitch period estimation, voice/unvoiced determination, analysis/encoding of speech signal and vocal tract filter. The applications of LPC cover the speech recognition, text to speech synthesis, telephone systems and multimedia.

LPC can be defined as a signal processing method for encoding an analog signal in which a particular value is predicted by a linear function of the past values of the signal. Solutions of LPC equations were derived using autocorrelation method and Levinson-Durbin algorithm [22]. It was an efficient algorithm for finding the prediction coefficients [23]. Log area ratios (LAR) are used to represent the LPC or reflection coefficients. Let k_i be the *i*th reflection coefficient of a filter, the *i*th LAR is defined as

$$g_i = \log\left(\frac{1+k_i}{1-k_i}\right) \tag{1}$$

This LAR parameter, g_i is used as the reference templates as well as the test templates.

C. Dynamic Time Warping

Vocal fold vibration is a time-dependent process. GAWs of the same speaker are likely to have different in magnitudes and frequencies. Dynamic time warping (DTW) is a well-known frame matching technique to find an optimal alignment between two given (time-dependent) sequences under certain restrictions [24]. GAW patterns were represented as a sequence of vectors (LAR parameter) and warped in a nonlinear way to match each other. A time alignment performed to obtain a global distance measure between these two GAW patterns.. The aims of the DTW alignment is to discover the optimal warping path, which is a curve relating the *j* time axis of the reference pattern, represented by $R = [R(1),R(2),\ldots,R(N_R)]$ to the *i* time axis of the test pattern represented by $T = [T(1),T(2),\ldots,T(N_T)]$, where N_T is the number of frames or vectors in the test signal and N_R is the number of frames or vectors in the reference signal. This warping path takes the form $W = \{w(1),w(2),\ldots,w(k)\}$ where each *w* is a pair of pointers to the samples being matched (i.e., $w(k) = [i(k),j(k)]$). The warping function is required to minimize the overall cost function [25].

$$D = \sum_{k=1}^{K} d[(w(k))] \tag{2}$$

where

$$d[w(k)] = d(T(i(k)), R(j(k)))$$

is the distance between frame $i(k)$ of the test pattern, and frame $j(k)$ of the reference pattern. Fig. 2 illustrates the concept of the dynamic time warping path. Detailed explanation of the DTW algorithm can be found in [25].

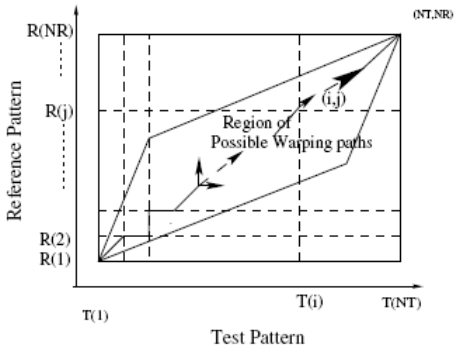

Fig. 2 Dynamic time warping path [25]

III. Data Collection

Images of the vibrating vocal folds were acquired using Kay Elemetrics High-Speed Digital Imaging system (Kay Elemetrics Corp, Lincoln, Park, NJ) at the Audiology and Speech Sciences Clinic, Universiti Kebangsaan Malaysia (UKM). A 90-degree/70 degree rigid endoscope (Kay 9106) which coupled to a camera (Model 9700) was used in the recording. The system recorded the images with a spatial resolution of 160 x 140 pixels at a rate of 2000 frames per second. The sampling rate of 2000 frames per second is considered adequate to resolve the actual vocal fold vibrations from speakers in this study. Fig. 3 shows a representation of performing endoscopic high-speed recordings.

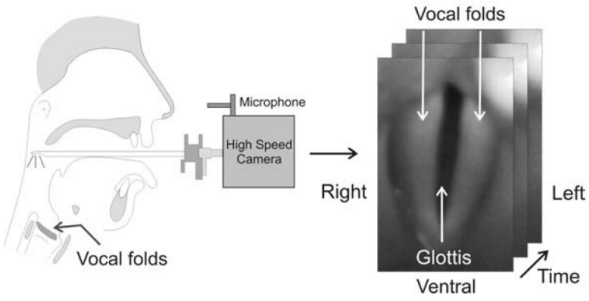

Fig. 3 High-speed recording of vocal folds

The recording procedure involves inserting the endoscope into the subject's open mouth into the posterior pharynx at a ~70-degree angle. To avoid eliciting a gag, care was taken to avoid touching either the posterior pharyngeal wall or the velum. The best view of the larynx is obtained with the subject seated forward with the chin tilted slightly upward. The subject is instructed to produce the vowel /i/ while the tongue is held by the clinician. At this stage, the endoscope is adjusted so that it is parallel to the superior surface of the vocal folds to minimize image distortion. The vowel /i/ is chosen because when it is phonated, the laryngeal inlet is opened by retraction of the epiglottis and base of the tongue providing the best view of the larynx.

The study was divided into two stages; the training and testing stages. In the training stage, a professionally trained speaker was instructed to phonate /i/ perceived as clear, breathy, fry, pressed and hyperfunctional, at her comfortable pitch and loudness. The samples of GAW from this speaker were used as the reference template for the classification. In the testing stage, 42 samples of GAW recorded from a group of 14 normal speakers production of /i/ were used to test the DTW classifier. The phonations were also produced at normal pitch and normal loudness. Two tokens of /i/ productions from each speaker were analyzed. In reference to the previous study [4], 600 frames of images were used for each phonation.

IV. Results and Discussion

Figs. 4 (a)–(e) show sets of Nyquist plot patterns derived from the professionally trained female speaker when simulating phonated /i/ with clear, breathy, fry, pressed, and hyperfunctional voice at comfortable pitch and loudness. These five samples of GAW have been used as reference template for the classification using DTW.

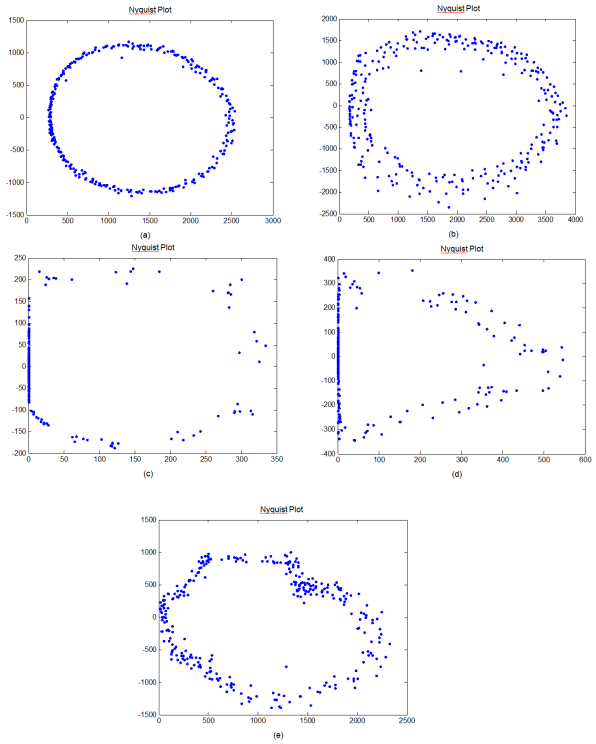

Fig. 4 (a)-(e) Nyquist plot obtained from the reference speaker's production of a clear, breathy, fry, pressed and hyperfunctional voice

A LPC vocoder processes input signals in separate signal blocks and computes a set of filter coefficients for each block. Normally, the coefficients of a block are stable in short periods that are about 20ms [23]. Because the sampling frequency of the GAW was 2000Hz, it takes 40 sample point of GAW to calculate for each block or window. The number of LPC coefficient is adjusted to get the optimum result of the classification. Table I shows the accuracy of classification using the LPC and DTW.

TABLE I
ACCURACY OF CLASSIFICATION

Number of LPC coefficient	Accuracy of classification (%)
10	69.0
11	69.0
12	71.4
13	69.0
14	81.0
15	81.0
16	81.0
17	81.0
18	78.6
24	78.6
34	76.2

There were 42 samples of GAW that have been used for the testing session. The classifier gave an accuracy of 81% when the number of LPC coefficients used was in the range of 14-17.

V. CONCLUSION

Glottal area waveform obtained from direct observation of vocal fold vibration can be used as a reliable indicator of normal and abnormal function of vocal folds. In this paper the dynamic time warping classifier is presented as a potential assessment tool to aid voice clinicians in the analysis of voice disorders. However, like the others before, these tools generally do not yield results with 100% accuracy. The accuracy of the tools depend on several factors, such as the size of the reference template, the quality of GAW that derived from the HSLI and also the parameters chosen to represent the input. However from the analysis of the results listed in Table I, it was shown that the classifiers proposed are effective to the tune of about 69 to 81% accuracy.

ACKNOWLEDGMENT

This research was supported by Universiti Teknologi Malaysia (UTM) through Research University Grant (GUP) Tier 1 vote number Q.J130000.2545.04H21, Ministry of Higher Education (MOHE) Malaysia, Universiti Teknikal Malaysia Melaka (UTeM) and Universiti Kebangsaan Malaysia (UKM). The authors are grateful to all parties for supporting the present work.

REFERENCES

[1] P. Gomez, R. Fernandez, A. Nieto, F. Diaz, F.J. Fernandez, V. Rodellar, A. Alvarez, and R. Martinez, Evaluation of Voice Pathology Based on the Estimation of Vocal Fold Biomechanical Parameters, *Journal of Voice*. 2006, 21(4), 450-476.

[2] M. Dollinger, J. Lohscheller, J. Svec, A. McWhorter, and M. Kunduk, "Support Vector Machine Classification of Vocal Fold Vibrations based on Phonovibrogram Features", Advances in Vibration Analysis Research, 2010, pp. 435-456.

[3] U. Hoppe*, Mechanisms of Hoarseness–Visualization and Interpretation by Means of Nonlinear Dynamics*. Aachen, Germany: Shaker, 2001.

[4] K. Ahmad, Y. Yan, and D. M. Bless, Vocal Fold Vibratory Characteristics in Normal Female Speakers From High-Speed Digital Imaging. *Journal of Voice*. 2011,1-15.

[5] K. Ahmad, Y. Yan, and D. M. Bless, Vocal Fold Vibratory Characteristics of Healthy Geriatric Females - Analysis of High-Speed Digital Images. *Journal of Voice.*, 2012.

[6] I. R. Titze, *Workshop on Acoustic Voice Analysis: Summary Statement.* Denver, CO: National Center for Voice and Speech, Wilbur James Gould Research Center; February 17, 1994.

[7] T. Wittenberg, M. Tigges, P. Mergell, U. Eysholdt, "Functional Imaging of Vocal Fold Vibration: Digital Multislice High-Speed Kymography" *Journal of Voice. 2000* ;14(3):422–442.

[8] R. Schwarz, U. Hoppe, M. Schuster, T. Wurzbacher, U. Eysholdt, and J. Lohscheller, Classification of unilateral vocal fold paralysis by endoscopic digital highspeed recordings and inversion of a biomechanical model. *IEEE Transactions on Biomedical Engineering*, 2006, 53(6), 1099-1108.

[9] Y. Zhang, E. Bieging, H. Tsui, and J. J. Jiang, Efficient and effective extraction of vocal fold vibratory patterns from high-speed digital imaging. *Journal of Voice.* In press, 2009.

[10] Y. Yan, K. Ahmad, M. Kunduk, D. Bless, Analysis Of Vocal-Fold Vibrations From High-Speed Laryngeal Images Using A Hilbert Transform-Based Methodology. *Journal of Voice*,2005, 19(2), 161–175.

[11] J. Lohscheller, U. Eysholdt, H. Toy, and M. Dollinger, Phonovibrography: mapping high-speed movies of vocal fold vibrations into 2D-diagrams for visualizing and analyzing the underlying laryngeal dynamics. *IEEE Transactions on Medical Imaging*,2008, 27(3), 300-309.

[12] X. Chen, D. Bless, and Y. Yan, A Segmentation Scheme Based on Rayleigh Distribution Model for Extracting Glottal Waveform from High-speed Laryngeal Images. *Proc. of the IEEE Eng. in Medicine and Biology*. 2005, 6269-6272.

[13] V. J. Osma-Ruiz, J. I. Godino-Llorente, N. Sáenz-Lechón, N., and R. Fraile, Segmentation of the glottal space from laryngeal images using the watershed transform. *Computerized Medical Imaging and Graphics*.2008, 32(3), 193-201.

[14] S. Allin, J. Galeotti, G. Stetten, S. H. Dailey, Enhanced snake based segmentation of vocal folds. *In proc. ISBI.*1, 2004, 812- 815.

[15] A. Mendez, E. M. I. Alaoui, B. Garcia, E. Ibn-Elhaj, and I. Ruiz, Glottal Space Segmentation From Motion Estimation and Gabor Filtering. *International Conf. of the IEEE EMBS*. 2009, 5756-5759.

[16] V. J. Osma-Ruiz, J. M. Gutierrez-Arriola, J. I. Godino-Llorente, N. Saenz-Lechon, R. Fraile, and J. D. Arias-Londono, Advanced Preprocessing of Larynx Images To Improve the Segmentation of Glottal Area. ICSA. 2009, 129-132.

[17] Y. Zhang, C. Tao, J. J. Jiang, Parameter estimation of an asymmetric vocal fold system from glottal area time series using chaos synchronization. Chaos.16. 2006.

[18] Y. Yan, E. Damrose, D. Bless, Functional analysis of voice using simultaneous high-speed imaging and acoustic recordings. *Journal of Voice.* 21, 2007.

[19] D. Voigt, M. Dollinger, T. Braunschweig, A, Yang, U. Eysholdt, and J. Lohscheller, Classification of Functional Voice Disorders Based on Phonovibrograms. Artificial Intelligence in Medicine. 2010, 51-59.

[20] J. P. Noordzij, and P. Woo, Glottal Area Waveform Analysis of Benign Before and After Surgery. Ann. Otol. Rhinol. Laryngol. 2000, 105, 441-446.

[21] J. R Booth, and D. G. Childers, Automated Analysis of Ultra High-Speed Laryngeal Films. *IEEE Trans. Biomedical Eng.* 1979, 26(4), 185-192.

[22] W. C. Chu, Speech Coding Algorithms: Foundation and Evolution of Standardized Coder, Wiley-Interscience,2003.

[23] P.R. Cook, Real Sound Synthesis for Interactive Applications, A K Peters, 2002.

[24] Rubita, Sh-Hussain Salleh, Shahrudin, NN with DTW-FF conficients and pitch feature for speaker recognition, *Regional postgraduate conference on Engineering and Science (RPECES 2006)*, 2006.

[25] A.M. Youssef, T.K. Abdel-Galil, E.F. El-Saadany and M.M.A. Salama, Disturbance Classification Utilizing Dynamic Time Warping Classifier, *IEEE Transactions on Power Delivery,* 2004, 19(1), 272-278.