ANALYSIS OF VOCAL-FOLD MOTION FROM LARYNGEAL IMAGE SEQUENCES

Jasmin Gonzalez   Sally L. Wood   Yuling Yan

Electrical Engineering and Bioengineering Departments
Santa Clara University, Santa Clara, California, 95053   USA

ABSTRACT

Analysis of vocal-fold vibration from high speed laryngeal images has been shown to have diagnostic value in the assessment of normal and abnormal voice production. This paper presents a new way to define the boundaries of the vocal-fold opening that is computationally efficient and operates successfully over a range of threshold settings. The area within these boundaries is computed for each image frame to produce the glottal area waveform (GAW*)* as a function of time. The frequency content of the GAW is analyzed to compare with previous results used for clinical evaluation. In addition, analysis of the glottal area waveform used by others is extended to include new parameters using Fourier descriptors of the boundary contours.

Index Terms— laryngeal imaging, glottal area waveform, Fourier descriptors, frequency of vocal-fold vibration

1. INTRODUCTION

High speed laryngeal image sequences showing vocal-fold motion during clinical tests can provide valuable diagnostic information for voice assessment [3-6]. From visual inspection of the vibration of the vocal-folds, irregular frequencies and asymmetries in motion of the vocal fold opening can be observed when pathological voice conditions cause abnormal voice production. Typically a high-speed camera with a frame rate of at least 2000 Hz is used to capture image sequences of the vocal folds vibrating at rates between 100Hz and 400 Hz.

Automatic processing and analysis of the image sequence is desirable to provide a quantitative assessment of voice production based on the vocal-fold motion. This would allow rapid scanning of the diagnostic image data and would also provide a basis for comparison of voice production before and after surgical corrections or therapy so that treatment progress could be monitored. Automatic processing techniques have been proposed and implemented to achieve these goals [3-6].

Previous work by Yan et al. [5-6] has demonstrated the diagnostic value of computing the glottal area, the area of the vocal fold opening, in each image frame and viewing the area as a function of time or frame number to create the GAW. The cameras are not calibrated to allow an exact computation of the area from the contour, and the diagnostic information lies in the relative changes of the area as a function of time, not the precise value of the area.

Analysis of the frequency content of the GAW during clinical evaluation procedures can identify pathological conditions or verify treatment progress. The computation methods to accomplish this must be computationally efficient if they are to be used in a real-time environment with the high frame rate camera. The methods must also be robust in the presence of noise and interference.

One approach [5] applies a global threshold to the image pixels to isolate the glottis using whole pixel classification, and then determines the bounding contour of the segmented glottis. The contour is then smoothed with morphological filtering which becomes the seed for region-growing for a more accurate delineation of the region of interest. In the region-growing method, the seed point and stopping point are very important for correct segmentation. In some cases, erroneous segmentation can be achieved due to pixel classification errors, over-erosion during morphological filtering and noise. In addition, this method requires iterative computations, which is not efficient in a real-time environment.

Another approach [6] uses a snake based tracing of vocal-fold motion. Initially, it performs global thresholding and then approximates the glottal geometry with an ellipse-shaped region. Principal component analysis (PCA) is used to find the approximate parameters of the ellipse, and then the estimated ellipse is used as the initial contour of the snake based method. The selection of the initial contour can significantly affect the segmentation results and the number of iterations needed. When the geometry of the glottis is not well approximated with an ellipse, there is a large fitting error and a high computational cost in order to achieve the desired accuracy. The computational cost, number of iterations, and convergence can vary significantly with image quality.

In this work we develop a new approach to acquiring the contours needed to obtain the GAW. The method proposed does not require iteration and is robust to threshold variation. Using an interpolated image, sub-pixel boundary point locations are determined. We explore the dependence of these results on threshold settings and show that this approach has low sensitivity to variation over a range of threshold values. The GAW measurements computed from these contours are consistent with previous work in [6]. In addition, to computing the area, the position, orientation and axes length for the approximating ellipse are also computed directly from the Fourier descriptors of the contour. Higher order approximations are also explored.

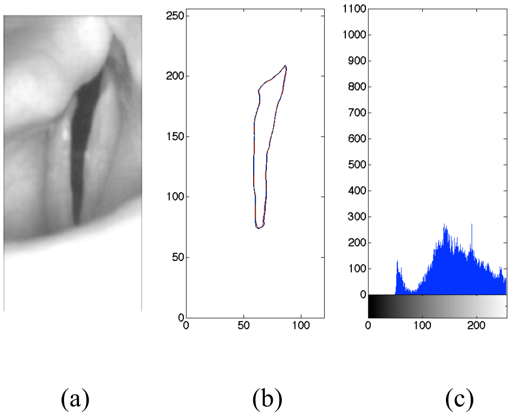
Although the boundary contours based on bilinearly interpolated image data usually have smoother edges than contours from a segmentation that uses whole pixel recruitment, we explore smoothing methods and focus on Fourier descriptors [1-2]. The Fourier descriptors are attractive because they can be computed efficiently, they can represent most biologically reasonable contours well with only a few coefficients, the descriptors change in simple ways under translation, rotation, scaling, and starting point shift [2], and they represent fundamental shape properties. In addition, the reduced coefficient approximate representation for contours of variable length can be used to specify all contours with the same number of coefficients.

2. GLOTTAL BOUNDARY DETECTION

Sequences of 256 row by 120 column gray scale images of the larynx were taken at 2000 frames per second from patients who were asked to produce a sustained vowel phonation. No calibration data was available to determine the actual size of structures in the image, and diagnostic information was based on the relative variation of imaged structures. In some image sequences a distinct interlace variation was observed. This was removed with a short low pass pre-processing filter.

Based on the image histogram, an appropriate threshold level was determined which would separate the vocal-fold opening from the rest of the image. The vocal-fold opening was very dark, but the surfaces of the surrounding soft tissues showed a wide range of lighter pixel values with well-defined edges and variable illumination. No source model seemed well matched to the distribution of gray levels observed in the images. Initial thresholds for segmentation were placed near low-valued local minima of the histogram. A threshold range was established which would separate the vocal fold opening, or glottis, from the rest of the laryngeal image.

A typical single 256 row by 120 column image of the larynx is shown in Figure 1 on the left. The dark area in the center, the vocal fold opening, changes shape and size during vocalizations. The center plot displays three contours computed for threshold values of 85, 90, and 95, but the contours are so similar that they are not easily distinguished. In the image histogram on the right, the trough between the pixel values of the vocal-fold opening and the rest of the image is evident.



(a) (b) (c)

Figure 1: (a) A single image frame (b) Contours obtained at three threshold levels : 85, 90 and 95 (c) Image histogram showing trough from which threshold values are selected

For each threshold value a contour was defined as a sequence of points on the grid-line crossings of a bilinearly interpolated image. For a fixed global threshold, let $c_m = \{x_m(n), y_m(n)\}$ be the set of N_m contour points of the vocal-fold opening of the m^{th} frame. Figure 2 shows an expanded view of a small part of the gray-scale image from Figure 1 with the three contours shown in different colors. These contours are very similar in shape, and, although the threshold values differ by 10, there is little movement on the pixel grid.

The sensitivity of the glottal area computed from the segmented image can be demonstrated by displaying the area in a single frame as a function of the threshold value. The upper plot in Figure 2 shows the area for threshold values over the very wide range of 50 to 130. Below it is the corresponding range of the histogram. From the area plot it can be seen that, in the region between 65 and 100, the area changes very slowly due to expansion of the contours as the threshold level is increased. Outside of that range there are jumps in the area due to addition of regions that are not part of the vocal fold opening. Since there is no explicit camera calibration, only the relative variation of the GAW is of interest diagnostically, and in previous work [5-6] the complete GAW was normalized to have a maximum magnitude of 1. For this reason, variations in the threshold over a range that simply expands or contracts the contour without changing its shape will result in similar GAW patterns, and the analysis of the GAW will not be highly sensitive to threshold settings.

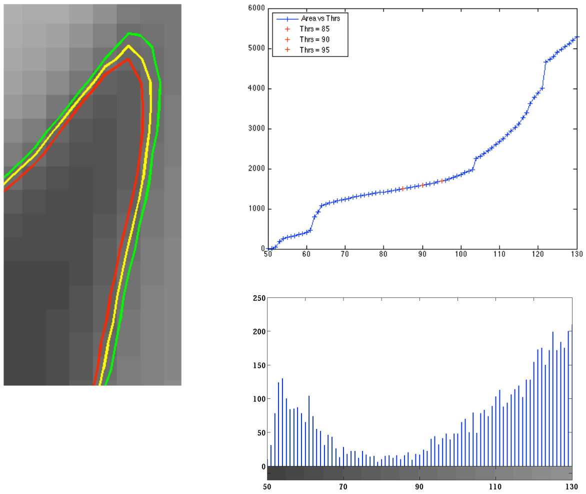


Figure 2: Detail of contours at three threshold levels on image, plot of area as a function of threshold and expanded histogram

The glottal area can be computed from the boundary sequence using the trapezoidal method for each closed clockwise contour.

$$
\begin{aligned}
A_m &= 0.5\sum_{n=0}^{N_m-1}\Big(x_m(n)-x_m\big(\langle n-1\rangle_{Mod\ N_m}\big)\Big)\Big(y_m(n)+y_m\big(\langle n-1\rangle_{Mod\ N_m}\big)\Big) \\
&= 0.5\sum_{n=0}^{N_m-1}\Big(x_m(n)y_m\big(\langle n-1\rangle_{Mod\ N_m}\big)-x_m\big(\langle n-1\rangle_{Mod\ N_m}\big)y_m(n)\Big)
\end{aligned}
\qquad (1)
$$

A typical sequence of image frames and contours is shown in Figure 3. The cycle for a vibration is 7 to 8 frames, so the four sequential frames show the change from close to maximum area to close to minimum area.

When the area function is analyzed for periodic variation, the assessment of the fundamental frequency variations and harmonic content is used to compare initial evaluations with evaluations after surgical corrections or therapy. Additional information about asymmetries is determined by analyzing the position of selected boundary points as a function of time.

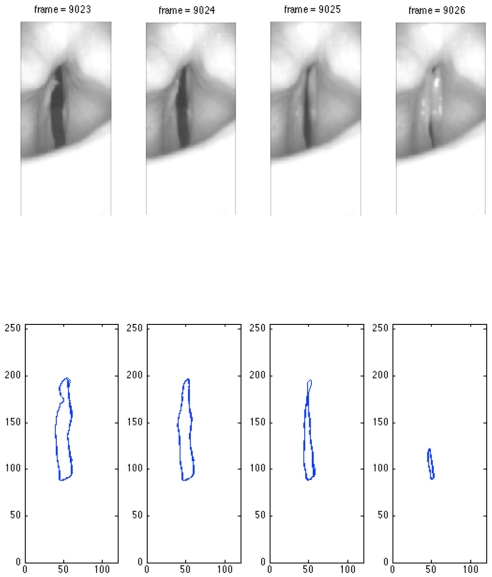
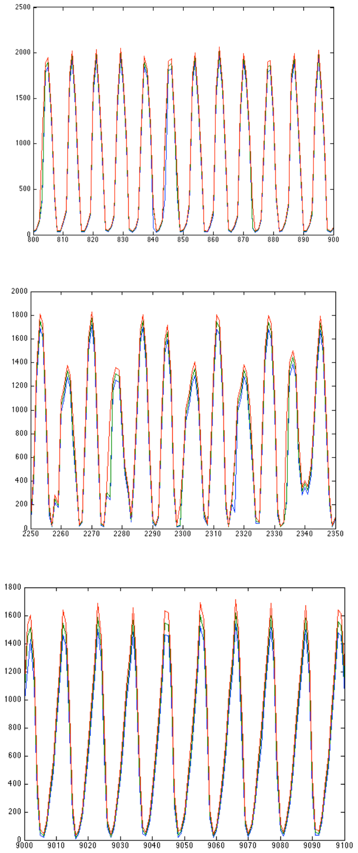


Figure 3: Sequence of four images and contours of the glottal area



(a)

(b)

(c)

Figure 4: Glottal area waveforms (in units of pixel area) as a function of frame number: (a) regular variation in voice production, (b) irregular variation in voice production and (c) regular variation after treatment

In Figure 4 three segments of GAW data are shown for a single patient with a voice production abnormality that appears soon after the examination begins. The un-normalized GAW is shown in units of pixel areas as a function of frame number. Results are shown for three threshold values to demonstrate the insensitivity to small changes in threshold settings. In (a) the GAW from the first part of the examination shows regular variation. However, about a second later the voice production becomes irregular as shown in (b). After treatment for this condition resulted in improved voice production, the GAW shown in (c) is very regular

Although the GAW waveform for a complete examination can be viewed in short segments as a function of frame number or time, it is possible to get a broader sense of context from the spectrogram. The spectrogram of the GAW from the complete examination clearly shows the intervals of phonation and rest. In Figure 5 the spectrogram is computed using a window width of 512 frames with a 50% overlap. The results are shown as a function of frame number for ease in associating the results with images and GAW plots. The frequency is shown in normalized frequency units of cycles per frame, and this value should be multiplied by the frame rate to get the actual frequency in Hz or cycles per second. The magnitude of the frequency component is coded by color with bright red representing the highest magnitude and dark blue representing the lowest magnitude.

In Figure 5a the initial vocalization shows a steady response in the red/orange band at about 240Hz (0.12 on the plot). There is a strong second harmonic with a thin orange band and a visible third harmonic. However, around frame 2000 a lower frequency output at about 120Hz (0.06 on the plot) appears. This is an indication of a specific voice production abnormality for which treatment is possible. There is a stronger second harmonic at 240 Hz and a third harmonic at about 360 Hz.

In this examination, shortly after the appearance of the lower frequency sound, the vocalization is halted and the camera drifts so that the glottal area falls outside the field of view of the camera for a while. Then the camera is realigned and the vocalization resumes. This appears as an unstructured interval in the center of the spectrogram. The rest of the test shows that the abnormality causing the 120 Hz signal and harmonics persists.

In Figure 5b the spectrogram of the GAW after treatment shows steady patterns of sustained vocalization at about 200 Hz with brief intervals of rest. The harmonic structure is well defined. Toward the end of this spectrogram there is a brief appearance of a lower frequency at about 100 Hz.

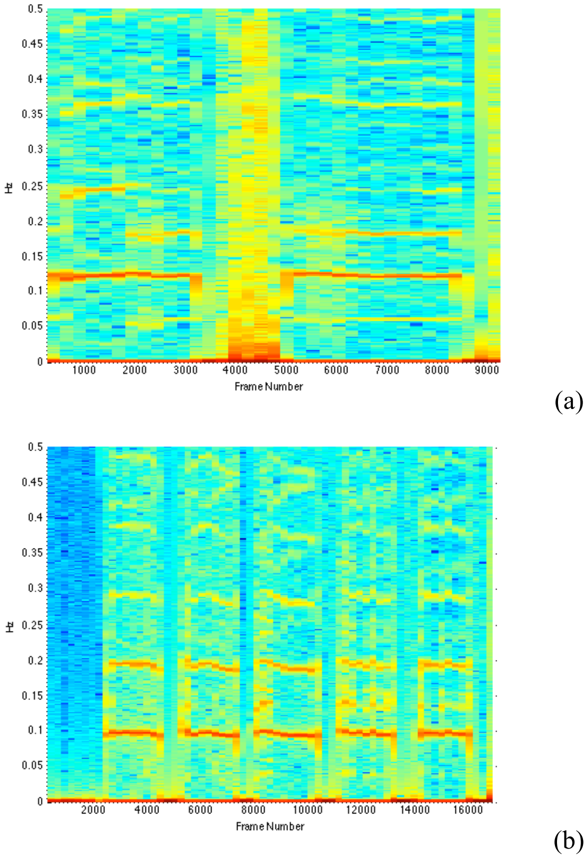


Figure 5: Spectrogram (a) before and (b) after treatment. Note: The frequency is shown in normalized frequency units. This value should be multiplied by the frame rate to get the actual frequency in Hz.

3. FOURIER DESCRIPTORS

Fourier descriptors [1-2] were applied to sequences of contour points to obtain additional information for diagnostic use. The descriptors are attractive because they can represent shapes with relatively few coefficient and they exhibit simple and easily detected reactions to transformations such as translation, rotation, shifting the starting point of a contour, or scaling. They can be used to identify shape parameters and also to smooth a contour by computing a limited frequency reconstruction of a sequence of boundary points. In addition, the area inside the original contour or any of the smoothed approximations can be computed directly from the Fourier descriptors.

Let $(x(s), y(s))$ be a point on a closed contour. A set of Fourier descriptors can be defined for the contour as the Fourier series coefficients of the complex valued $z(s) = x(s) + jy(s)$. For a discrete set of N contour points, the DFT can be used to compute a set of N Fourier descriptors as shown in Equation 2.

$$Z(k) = \frac{1}{N}\sum_{n=0}^{N-1} z(n) e^{\frac{-j2\pi nk}{N}} = \frac{1}{N}\sum_{n=0}^{N-1}\left(x(n) + jy(n)\right) e^{\frac{-j2\pi nk}{N}} \quad (2)$$

The smoothed contour using L pairs of Fourier descriptor can be defined by

$$\hat{z}_L(n) = \sum_{k=-L}^{L} Z(k) e^{\frac{+j2\pi nk}{N}} = \hat{x}_L(n) + j\hat{y}_L(n) \quad (3)$$

and the area inside the smoothed contour can be computed by

$$\hat{A}_L(n) = \frac{N}{2}\sum_{k=1}^{L}\left(\left|Z(-k)\right|^2 - \left|Z(k)\right|^2\right)\left(\sin(2\pi k/N)\right) \quad (4)$$

For L=1, $\hat{z}_1(n)$ is the ellipse defined by

$$\hat{z}_1(n) = Z(-1)e^{\frac{-j2\pi n}{N}} + Z(0) + Z(1)e^{\frac{+j2\pi n}{N}} \quad (5)$$

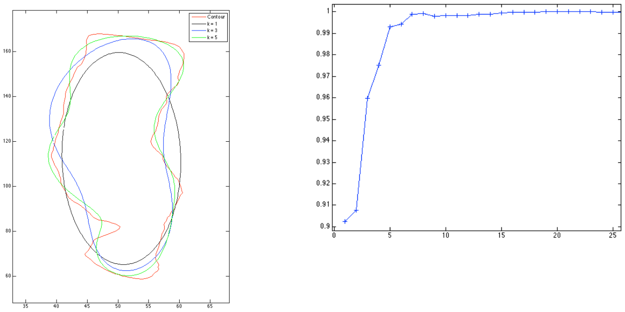
and the phase of Z(-1) and Z(1) can be used to compute the angle of rotation of the major axis of the ellipse from the horizontal axis. In [6] an ellipse was computed using PCA so that the ellipse could be the starting contour for the snake-based iterative approximation method.

In Figure 6 on the left, a typical contour is shown along with the approximation curves for L = 1, 3, and 5. The L=1 shape is an ellipse and the L=5 shape is close to the original contour in shape. On the right a plot shows the fraction of the area of the original contour that is contained within the approximation curves as a function of L. The fraction is above 99% for L=4 and is very close to 100% when L=6.

In addition to providing a simple method for smoothing contours and computing the area within the contour, a number of shape parameters can be computed from the Fourier descriptors. The center or average value, $Z(0)$, provides location information but does not include any information about the size or shape of the boundary contour. The remaining N-1 descriptors are insensitive to translation.

The magnitude of the Fourier descriptors is also insensitive to rotation of the contour or changes in the starting point index. Consider the Fourier descriptors for z'(n) which is the same shape as z(n) but rotated by angle ϕ and with a starting point shifted by n_0.

$$z'(n) = \left(x(n-n_0) + jy(n-n_0)\right)e^{j\phi}$$
$$Z'(k) = \frac{1}{N}\sum_{n=0}^{N-1} z'(n) e^{\frac{-j2\pi nk}{N}} = Z(k)e^{j\phi}e^{\frac{-j2\pi n_0 k}{N}} \quad (6)$$



(a)

(b)

Figure 6: (a) Approximation of the vocal fold opening shape using a selected number of Fourier Descriptors (b) Approximation of the vocal fold opening area using a selected number of Fourier Descriptors

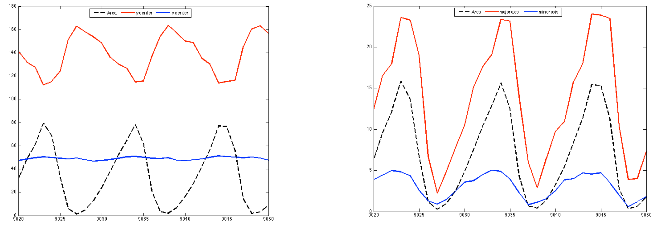


Figure 7: Center location and major and minor ellipse axes

The shape parameters can also be observed as a function of frame number. In addition to the area, the center point, the orientation, and the ratio of the lengths of the major and minor axes can be analyzed as a function of time and frequency. Higher order frequency components can be used to determine shape asymmetries.

In Figure 7 the left plot shows the value of the y-center and the x-center as a function of frame number. A scaled down plot of the area is shown with a dashed line for reference. The x-center shows almost no movement indicating the left and right sides of the vocal-fold opening are moving together in opposite directions. The y-center shows a periodic variation synchronized with the area. Since y is the row number that is increasing in the downward direction, the peak of the y-center coincides with the area peak.

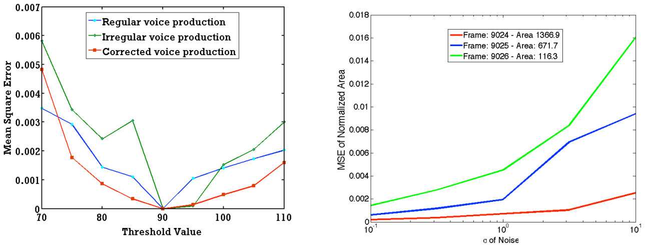
The right side of Figure 7 shows the variation of the major and minor ellipse axes as a function of frame. The major axis is abs(Z(1)) + abs(Z(N-1)) and the minor axis is abs(abs(Z(1)) - abs(Z(N-1))).

5. ERROR ANALYSIS

Since the cameras are not calibrated to allow precise computation of area, comparison of results across methods or under different noise conditions requires comparison of a normalized area as a function of time. The mean squared error (MSE) of a normalized GAW compared to a normalized reference GAW is used to demonstrate that the proposed method has low sensitivity to threshold variation over a significant range and to reasonable levels of added noise.

In Figure 8(a) the MSE of a 100-frame sequence from three different diagnostic image sequences is plotted as a function of threshold value over a wide range of threshold levels. The full range of the pixel values is 0 to 255. A threshold of 90 was used for the reference GAW.

In Figure 8(b) the MSE for three single frames is plotted as a function of the standard deviation of added white Gaussian noise. The area of the contour from the noise free image is shown in units of pixel areas in the legend. As might be expected, the frame with the smallest area was most sensitive to variations caused by added noise. These results were computed without any preprocessing noise suppression filtering to show the effect of image variations on the GAW.



(a) (b)

Figure 8: (a) MSE of GAW as a function of threshold level for three image sequences; (b) MSE of a GAW as a function of added noise level.

6. CONCLUSIONS

A new approach to detecting and analyzing the GAW was presented and demonstrated. Contours obtained from bilinearly interpolated images were smoother and less sensitive to threshold values that contours based on whole pixel classification. The results of the area analysis and frequency analysis of the GAW compared well with previously published results using other approaches requiring multiple levels of processing and iterative procedures. The method presented here relied on the insensitivity of the relative area over a significant range of threshold values and did not require iterative computation.

In addition, the use of Fourier descriptors provided multiple benefits. The Fourier descriptors are easily computed from a sequence of boundary points using efficient algorithms. An approximate curve created from a small number of descriptors is a smoothed version of the original curve, and the descriptors can be used directly to compute the area of the approximate curve. In addition, other characteristics of the GAW can be easily computed and displayed as a function of frame number. This includes motion of the center point of the vocal-fold opening, the ratio of the major and minor ellipse axes, and the orientation of the ellipse approximation. The use these other shape parameters from the descriptors should be explored for potential diagnostic value.

11. REFERENCES

[1] Chellappa, R. and R. Bagdazian, “Fourier Coding of Image Boundaries,” *IEEE Trans. PAMI*-6(1):102-105, 1984

[2] Jain, Anil K, *Fundamentals of Digital Image Processing*, Prentice Hall, New Jersey, 1989, pp370-392.

[3] Larson, H., S Hertegard, P. A.I. Indestad, B. Hammarberg, “Vocal-fold Vibrations: High Speed Imaging Kymography and Acoustic Analysis,” *Laryngoscope* vol. 100, pp 2117-2122, 2000.

[4] Manfredi, C., L. Bocchi, S. Bucchi, N. Migali, and G. Cantarella, “ Objective Vocal-fold Vibration Assessment from Videokymographic Images,” *Biomedical Signal Processing and Control*, Vol. 1, pp 129-136, 2006.

[5] Yan, Yuling, Xin Chen, and Diane Bless, “Automatic Tracing of Vocal-fold Motion from High-Speed Digital Images,”, *IEEE Trans. BME*-53(7):1394-1400,

[6] Yan, Yuling, Gan Du, Chi Zhu, and Gerard Marriott, “Snake-based Automatic Tracing of Vocal-fold Motion from High-speed Digital Images,” *IEEE ICASSP-2012*, Kyoto, Japan, pp593-596, March 25-30, 2012.