Ann.Bull.RILP(1985)No.19, 47-53

HIGH-SPEED DIGITAL RECORDING OF VOCAL FOLD VIBRATION USING A SOLID-STATE IMAGE SENSOR

Kiyoshi Honda, Shigeru Kiritani, Hiroshi Imagawa
Hajime Hirose and Kiyoshi Hashimoto

I. INTRODUCTION

The observation and measurement of vocal fold vibration have generally been performed by high-speed cinematographic and stroboscopic methods. High-speed cinematography can provide good resolution images of the vocal folds during vibration. However, a considerable time is required for film processing before visualizing the recorded data. Frame-by-frame analyses for the data are also time consuming. Stroboscopy, which has been mostly used for clinical purposes, has an inherent problem. In this method, the temporal information tend to be deteriorated due to the synchronous sampling principle. Particularly, when the vibration is very irregular, the observation of vibratory movement is almost impossible. These technical difficulties have restricted a large-scale data aquisition for studying the patterns of vocal fold vibration in various modes of phonation. As for the computer processing of the image data, digital processing of video pictures have become very easy through the developement of image frame memory. Unfortunately, however, ordinary video system works only at a fixed frame rate, and can not be applied to a high frame rate image recording. On the other hand, the solid-state sensors have shown remarkable progresses in recent years and have become available for practical uses. Since this sensor operates through digital scanning of picture elements, it can, in principle, adapt to image recordings of high-speed motions such as vocal fold vibration.

The present paper describes a new technique for high-speed digital recording of vocal fold vibration by means of a solid-state image sensor. The system consists of an assembly of a solid endoscope and a camera with the sensor, and of a computer image processing system. A small light source for endoscopy is used for illuminating the vocal folds. The images are digitized in real-time, and directly stored in the memory of a hardware image processor. Since the camera is hand-held and does not make mechanical noises, the view of the vocal folds is obtained easily in a quiet acoustic environment. Another advantage of this system is a capability of instantaneous image display. Digitized images are displayed on a CRT screen immediately after the end of recording.

II. EXPERIMENTAL PROCEDURES

The principle of digital image recording by a solid-state sensor is essentially equivalent to common video digitizing techniques. While a video digitizer samples considerably large number of pixels as a still picture, the present method using a solid-state image sensor handles a relatively small number of pixels per frame, and stores many frames continuously with a higher frame rate. The schematic diagram of the system is shown in Figure 1.

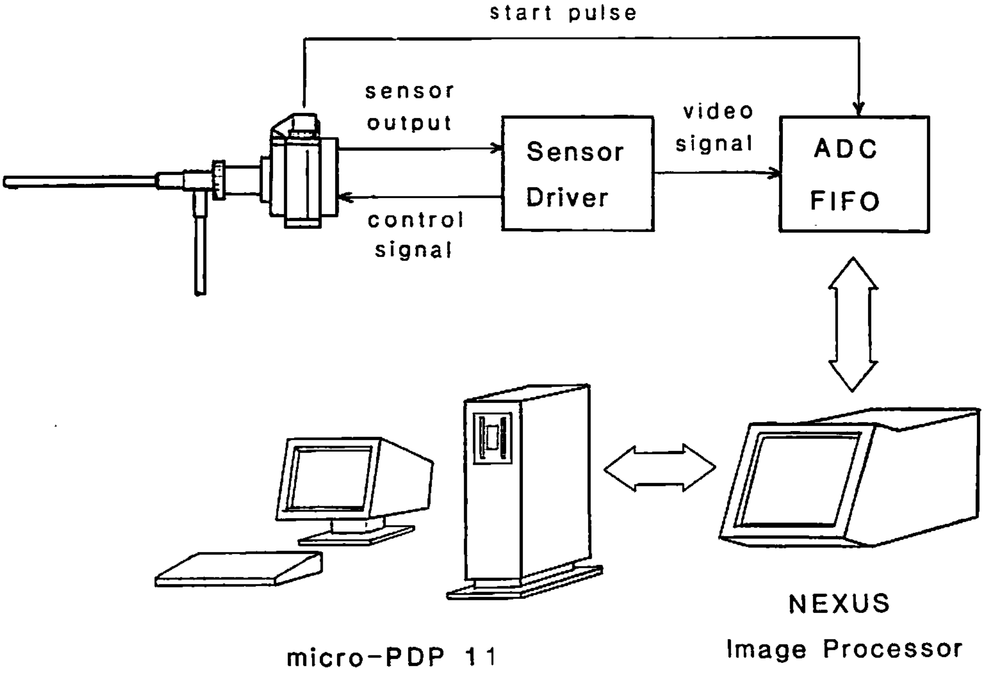

Figure [1] Schematic diagram of the high-speed image recording system.

There are many types of solid-state image sensors. Any type of image sensors, with an adequate driving circuit, equally outputs the charge of each photodiode element in the form of a common video signal. This charge is a function of the time integral of radiant power incident on the photodiode elements (Soclof, 1985). The image sensor used in this study (Reticon, RA50x50A) has a relatively simple construction. Different from a "CCD" sensor, which uses CCD shift registers between photodiode elements and the output line for storing and transfering the charge, this particular type of a "photodiode" image sensor consists of a two-dimensional array of photodiodes and two (horizontal and vertical) shift registers for multiplexing switch connections between the elements and the output line. The output signal resembles with a common video signal, but output frame rate can vary with variable clock frequencies for outputting the charge in each element. The sensor has 50 x 50 elements and the maximum clock frequency is 5 MHz. Thus, the frame rate is 2000 frames per second at the maximum.

An oblique-angled solid endoscope is used to obtain the view of the vocal folds. This endoscope is a version of the Type SFT-1 telescope (Nagashima Medical Instruments, Co.) for laryngeal observation. In this version, the lens diameter of the image-guide is enlarged so that a brighter image is obtainable. A 35mm reflex camera connects the endoscope and the image sensor, and also provides for viewing the vocal folds. The sensor is mounted on the back plate of the camera at the position of the film surface. A small light source with a 250 watts halogen lamp illuminates the vocal folds through the fiberoptic light-guide of the endoscope.

An 8 bit parallel high-speed A/D converter digitizes video signals from the sensor at a sampling rate identical to the clock frequency. Then, an FIFO memory buffer transfers the digital data to an image processor, NEXUS 6400, which has four pages of 256k byte image memory and a 512 x 480 pixels color graphic screen. A camera shutter pulse triggers the initiation of sampling. The storage of image data continues untill one page of the image memory is filled up. Since a frame consists of 50 x 50 image pixels and 10 x 50 additional pixels for horizontal blanking, 85 frames can be stored at a time. An array of images for successive frames (typically 72 frames) is displayed on the CRT screen of NEXUS 6400 immediately after the end of sampling. A small laboratory computer, micro-PDP 11 (DEC), controls basic operations of NEXUS 6400, stores the data parmanently on the disc, and serves for further data analyses.

In the present system, the sampling rate is set at 1.25 MHz, and therefore the frame rate is 500 frames per second. The maximum rate for the sensor is not attained, because the image focussed on the sensor is not sufficiently bright as to retain an adequate S/N retio at a higher rate. This frame rate is obviously too low to record vocal fold vibration. However, the frame rate can be increased by reducing the number of horizontal scan lines. While an ordinary frame is produced by 50 scan lines with 50 pixels in a line, a frame with a smaller number of scan lines can be digitized faster. A larger number of frames can also be stored in a fixed size of memory. A frame rate of 2500 frames per second is obtained, as shown later, by scanning 10 lines per frame. By doing so, the resolution in the vertical direction becomes worse, however, essential informations on the changes in the glottal shape associated with vocal fold vibration are still preserved.

III. RESULTS

Figure 2 is a photographic picture of the CRT screen which shows 72 successive frames of digitized images of the vocal folds. These images are recorded at a rate of 500 frames per second. Repetitions of the utterance /he/ are produced by a male subject without adding a pause between a syllable to a syllable. In the figure, a portion of /e/ to /h/ transition is shown. Since the voice fundamantal frequency is about 180 Hz, the open phase of the glottis is observed in every two to three frames. The /h/ portion

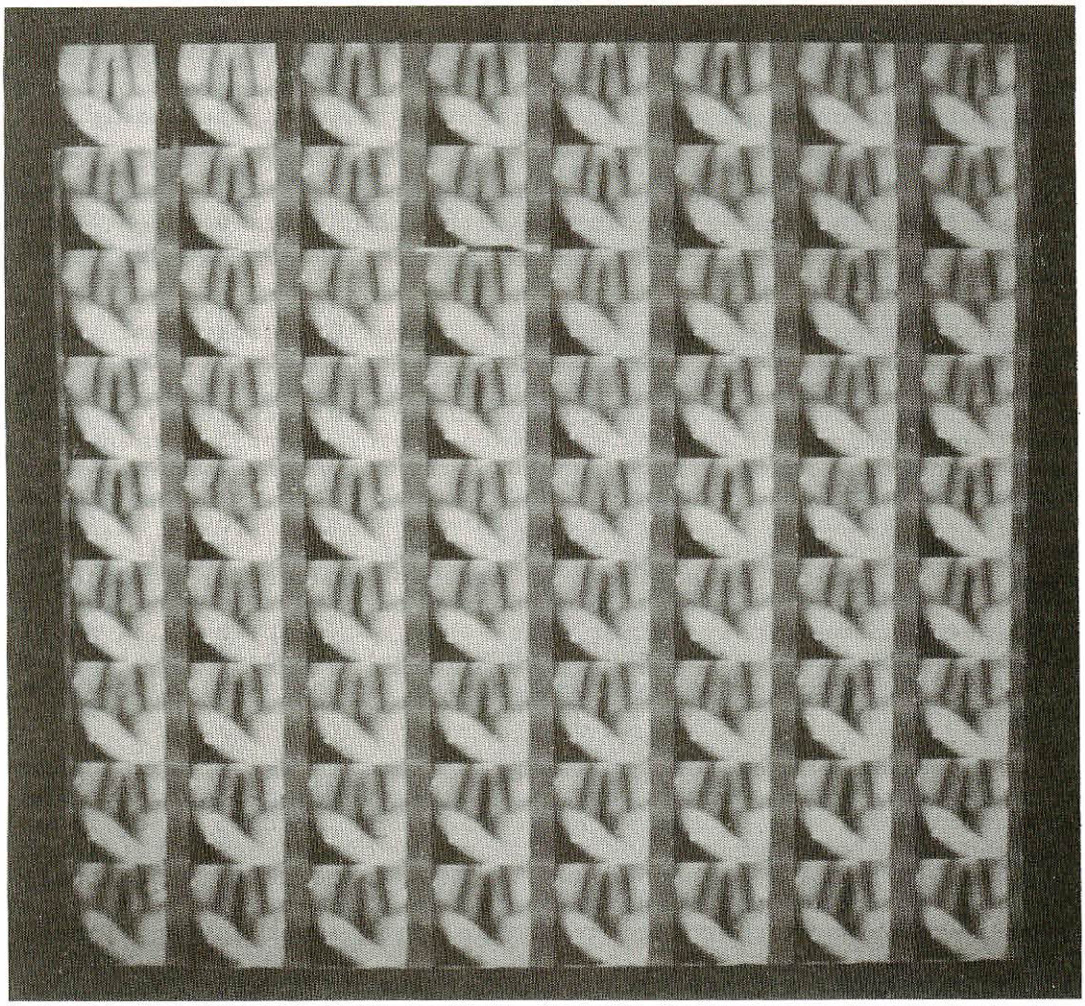

Figure [2] The CRT display of vocal fold images sampled at 500 frames per second, showing a transition fram /e/ to /h/.

is seen towards the end of the frames. These images are recorded in the same manner as taking a clinical laryngeal photography. Simply by inserting the endoscope, and by pressing the camera shutter butten after adjusting the position of the endoscope, images of the vocal folds are automatically stored in the memory. Sufficient resolution is attained for observing gross movements such as glottal articulatory gestures, although the details of the surface on the mucous membrane are hardly visible.

Recordings of vocal fold vibration, the primary concern of this study, are performed at a rate of 2500 frames per second. This frame rate is obtained by modifying the driving circuit so that every fifth lines are sampled out of 50 horizontal scan lines. Since originally recorded images have, therefore, vertically compressed frames consisting of 50 x 10 pixels, the frames are then expanded five times by displaying the same scan lines repeatedly so that the image of 50 scan lines is constructed. The dynamic range of recorded image are also affected by the reduction of exposure per frame associated with the increased frame rate. The brightness and contrast of the images are enhanced in the display through the function of NEXUS 6400. Figure 3 shows the images of the vocal folds which are processed in the way described above. These images are taken

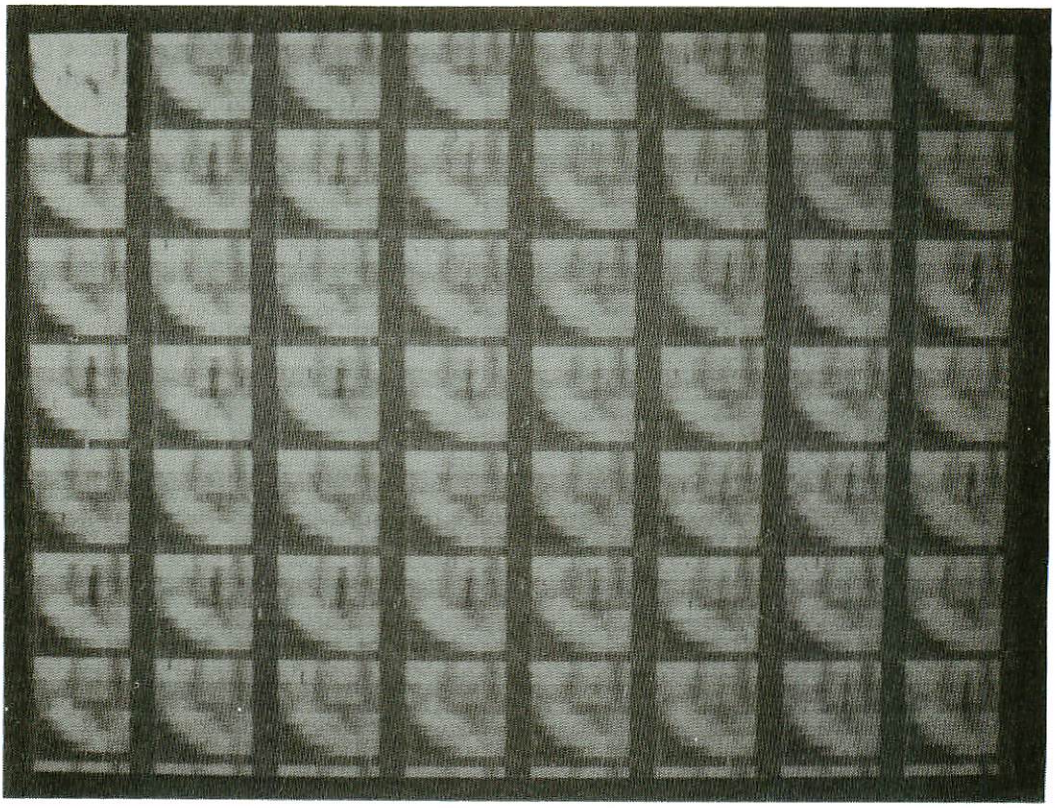

Figure [3] High-speed recording of vocal fold vibration, sampled at 2500 frames per second.

during a sustained phonation. Vibratory cycles are observed as a temporal change in the glottal shape across frames. Low resolution of the image in the vertical direction does not significantly matter, since movements of the glottis in the horizontal direction (the direction which is nearly perpendicular to the long axis of the vocal folds) are the most important factor for producing glottal area variations.

Since these images are stored in computer memory, a variety of image processing operations are easily performed. Figure 4 demonstrates one simple example. One line is sellected out of 10 scan lines in the frames which are recorded at a rate of 2500 frames per second, and the gray level variation in the line is drawn as a pattern using digital values of the pixels in the line. This gray level pattern is displayed successively across frames so that frame-by-frame changes in the glottal width are visualized. In the figure, the fourth scan line from the top of a frame in Figure 3 is sellected to show temporal changes in glottal movements at the point where the edge of the vocal folds produces the maximum excursion. Two glottal openings are demonstrated as two separate eminences in the figure.

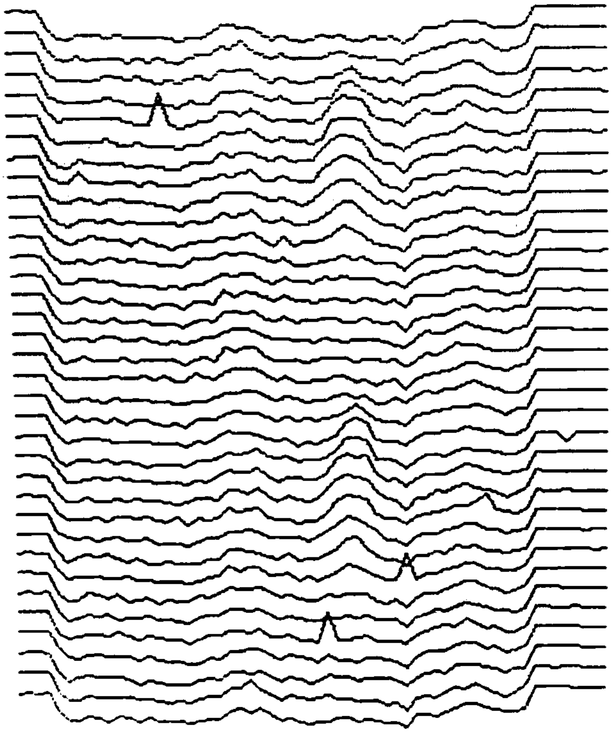

Figure [4] Frame-by-frame display of one scan line, demonstrating two glottal openings.

IV. DISCUSSION

A preliminary system has been developed for real-time digital recording of the images of vocal fold vibration. A solid-state image sensor with a laryngeal endoscope are used for obtaining vocal fold images in this system, and the images are directly stored in image memory. Therefore, the procedures for recordings and analyses with this system are very simple as compared to those with conventional high-speed filming methods. While the present system is usefull for some practical purposes, a few technical points can be noted regarding to possible improvements on the system performance. The maximum frame rate of the present system is basically restricted by two factors. One is the brightness of the image on the sensor produced by the optic system, and another is the maximum sampling rate of the sensor. It is expected that the brightness of the image can be incresed several times by minor modifications of the endoscope and the light source. Then, the maximum frame rate of 10 kHz can be realized. Horizontal resolution of the image is fixed in the system, but it can be changed by using a different type of image sensors. A frame of 5 scan lines with 100 pixels in a line is recorded at the same frame rate using a commercially available sensor. However, in order to increase the number of scan lines, the sensor which has a higher sampling rate is required. With respect to the size of storage memory, the present system is capable of storing the data for 50 msec in a page of 256k byte memory at 10 kHz frame rate and 500 pixels per frame. Several vibratory cycles are recorded in this condition. In order to increase the sampling duration, the size of memory has to be increased.

From a technical point of view, the image sensor used in the present system does not seem to have the best sensitivity characteristics among the various types of the image sensor. If it becomes prectical to use a sensor with higher sensitivity, observation of vocal fold vibration during running speech will become possible, by applying this method to fiberoptic endoscopy using a flexible bundle. Contrary to these modifications for a larger scale system, a simplification of the total system is obtainable by further reducing the number of scan lines. For example, the array display of the line images shown in Figure 4 can be made by using a line sensor consisting of one dimentional photodiode array. Since the use of a line sensor can make the total system compact, it seems to be more usefull for clinical examination.

REFERENCE

Soclof, S. (1985). Applications of Analog Intergrated Circuits (pp. 447-532). Englewood Cliffs, N.J.: Prentice Hall.