Ann.Bull.RILP(1987)No.21, 25-40

HIGH-SPEED DIGITAL IMAGE ANALYSIS OF LARYNGEAL BEHAVIOR IN RUNNING SPEECH*

Hajime Hirose, Shigeru Kiritani and Hiroshi Imagawa

Abstract

For the purpose of analyzing the behavior of the vocal folds in running speech, a high-speed digital recording system was used in combination with a specially designed fiberscope having a larger number of light guide fibers than the conventional scope. Different types of Japanese nonsense CVCV words were elicited, and glottal images were recorded simultaneously with speech and EGG signals during the period including the implosion and release of consonants. During the transient period of the implosion of voiceless consonants, it was found that the degree of glottal closure at the closed phase of the vocal fold vibration became gradually less tight before the arytenoid separation was eventually achieved. In the case of voiceless fricatives, the vocal folds tended to continue vibrating even with a relatively large glottal opening. At the release of voiceless consonants, a tight glottal closure was achieved relatively quickly for the initiation of postconsonantal voicing.

Introduction

The introduction of the fiberscope to speech research opened new possibilities in the fields of laryngology and experimental phonetics especially for the analysis of laryngeal dynamics in the production of various types of speech sounds[1]. The limitation of fiberoptic observation has been that this method is hardly applicable to a precise analysis of the vibratory patterns of the vocal folds, mainly because the frame rate for recording images is too low in conventional methods of fiberoptic observation.

In recent years, a new method for the high-speed recording of laryngeal dynamics has been developed by use of a digital recording system consisting of a solid state image sensor and an image processor. By combining the new digital recording system with a solid type telescope, successful recordings of laryngeal images during sustained phonation have been made in the authors' Institute, and preliminary results have been reported elsewhere[2),3]. More recently, the possibility of applying the new recording system to fiberscopic observation has been further explored in order to analyze the behavior of the vocal folds during running speech[4),5].

* A version of this paper was presented at the fifth Vocal Fold Physiology Conference, Tokyo, January 1987.

In this paper, the general features of the recording system currently in use at the authors' Institute will be described, together with some preliminary results of an analysis of laryngeal behavior during the production of Japanese disyllabic nonsense words.

Methods

In the present study, a specially designed fiberscope was used in combination with a high-speed digital recording system employing a CCD sensor as a solid state image sensor.

The new scope, FNL-T15, was designed by the Asahi Optical Company to be equipped with a larger number of light guide fibers than the regular scope. The outer diameter is 0.49 cm. In this particular scope, the space originally used to pass the biopsy forceps (the largest circle in the diagram of the tip of the scope in Figure 1) is filled with light guide fibers. As a result, approximately a 3-4 times higher intensity of illumination is obtainable compared with a conventional fiberscope.

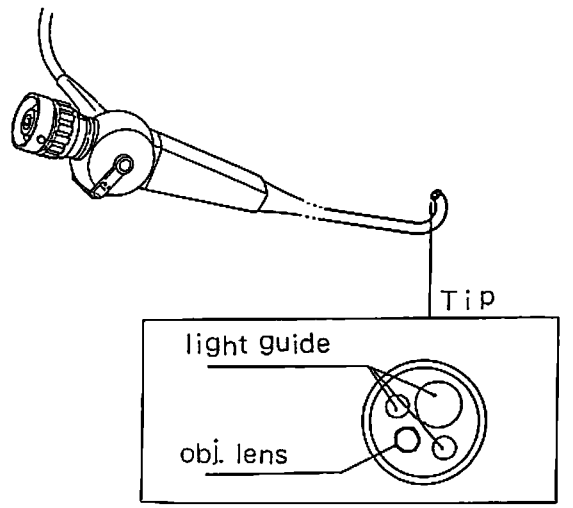

Fig. 1. A schematic view of the tip of the new fiberscope used in the present study.

Figure 2 shows a block diagram of the entire recording system. The fiberscope is attached to a single-eye reflex camera, and a 300W Xenon lamp is used as a light source. While the laryngeal view is monitored on a VTR monitor through the view finder, the subject utters CVCV test words and the laryngeal image signals are sampled simultaneously with speech and electroglottographic (EGG) signals. An array of 210 x 14 picture elements is used for recording at a rate of 2,000 frames per second. The sampling pulses for the A/D conversion of the speech and EGG signals are generated from the scanning circuit of the CCD sensor. While the original scan rate is 10 MHz, the sampling rate for the A/D conversion is counted down to 10 kHz.

Sampling of an appropriate period including the onset and release of a consonant in a test word is obtained as illustrated in Figure 3. The camera shutter is released manually to have the entire system ready for data sampling just before the production of the test utterance. Sampling is initiated with a pre-set delay after the amplitude of the acoustic envelope of the first vowel reaches a certain threshold. For recording at 2,000 frames per second, sampling continues for 117 msec so as to store 234 frames of image data at one time.

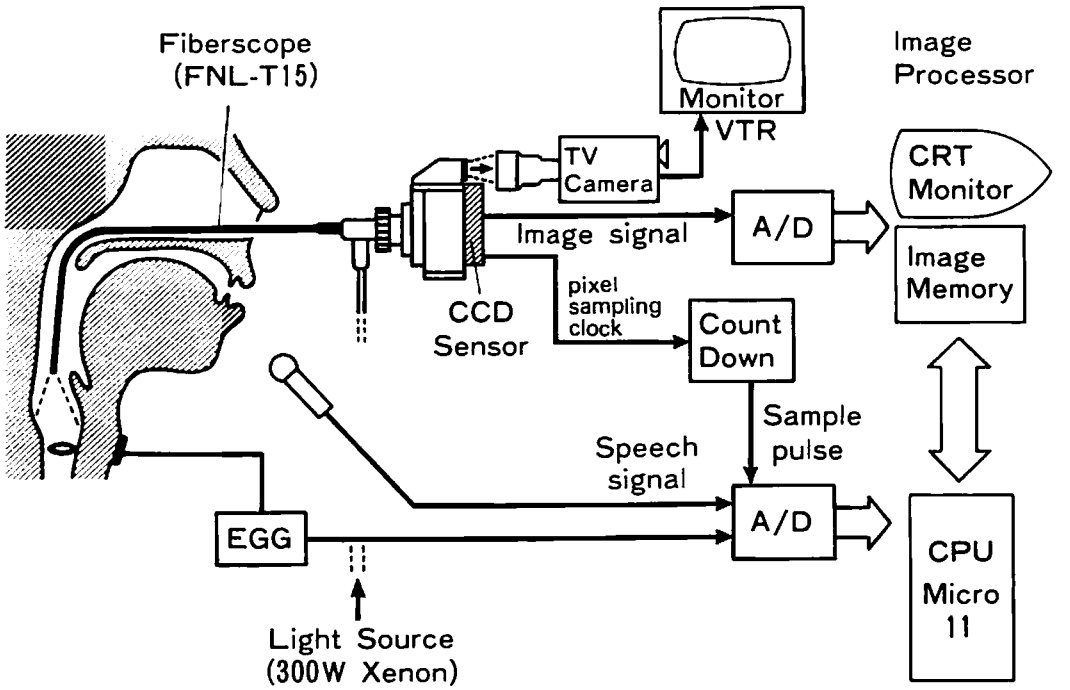

Fig. 2. A block diagram of the entire recording system.

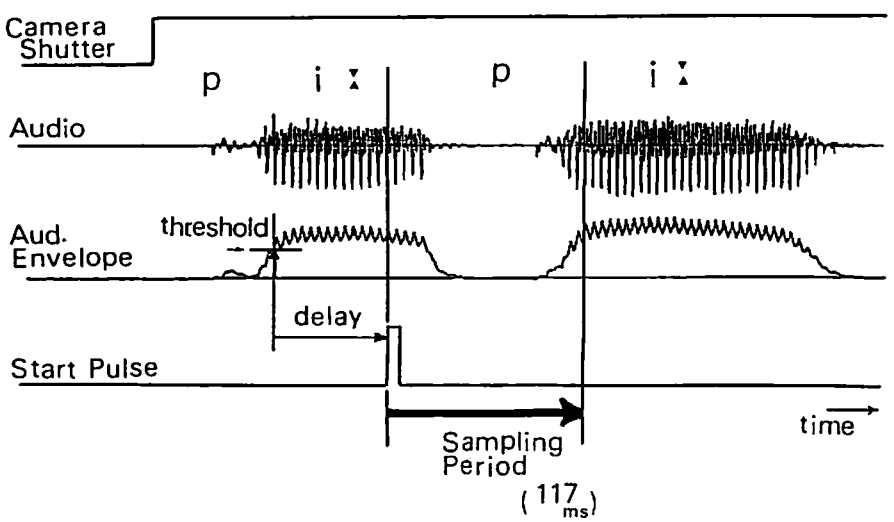

Fig. 3. A diagram illustrating the method of sampling an appropriate portion of a test utterance.

The result of EEG recordings will not be presented in this paper.

Results

Figure 4 shows an example of the laryngeal images displayed on a CRT monitor as an array of small-size images of sequential frames. The sampling proceeds from left to right, and from top to bottom. Each frame consecutively corresponds to the timing pulse displayed at the bottom, so that a temporal comparison between the laryngeal image and the speech signals can be readily made. No temporal adjustments are made in the figure for the time lag of the speech signal.

In this particular example, the first 42 frames of the stored data of the utterance /pi:pi:/ are shown, starting from where the glottis gradually begins to separate toward the medial

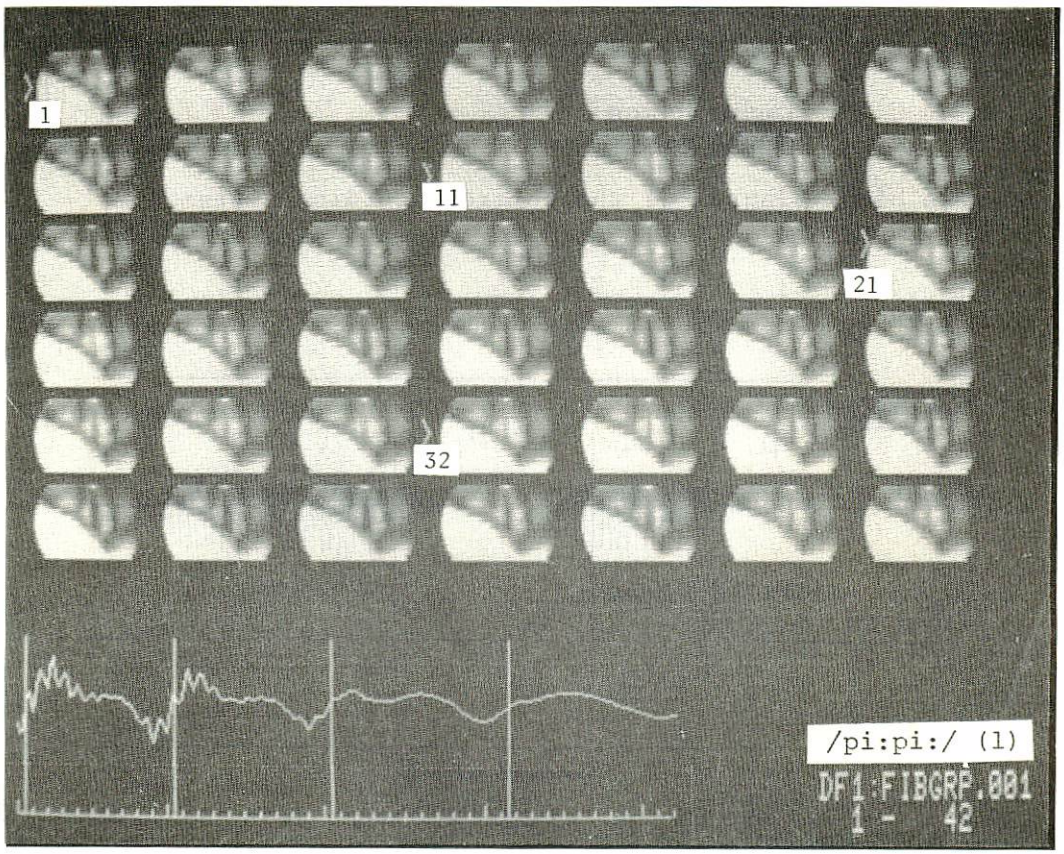

Fig. 4. The first 42 frames of recorded images for the test utterance /pi:pi:/. Speech signal is displayed below. Vertical lines on the time axis corresond to the pertinent frame numbers. Time: 0.5 msec./frame.

/p/. Here, the closed phase of the vocal fold vibration is observed at frame numbers 1, 11, 21 and 32. It appears that the degree of glottal closure at each closed phase gradually, and very slightly, becomes less constricted toward the voiceless period. In this series, F0 is about 180 Hz.

Figure 5 shows the next 42 frames of the stored samples for the same utterance /pi:pi:/ from frame number 43 to 84. In this series, F0 is about 170 Hz, and the closed phase is observed at frame numbers 43, 55, 67 and 79. As mentioned above, the degree of glottal closure further becomes less constricted. Toward the end of this series, the glottis appears to start to open after frame number 80.

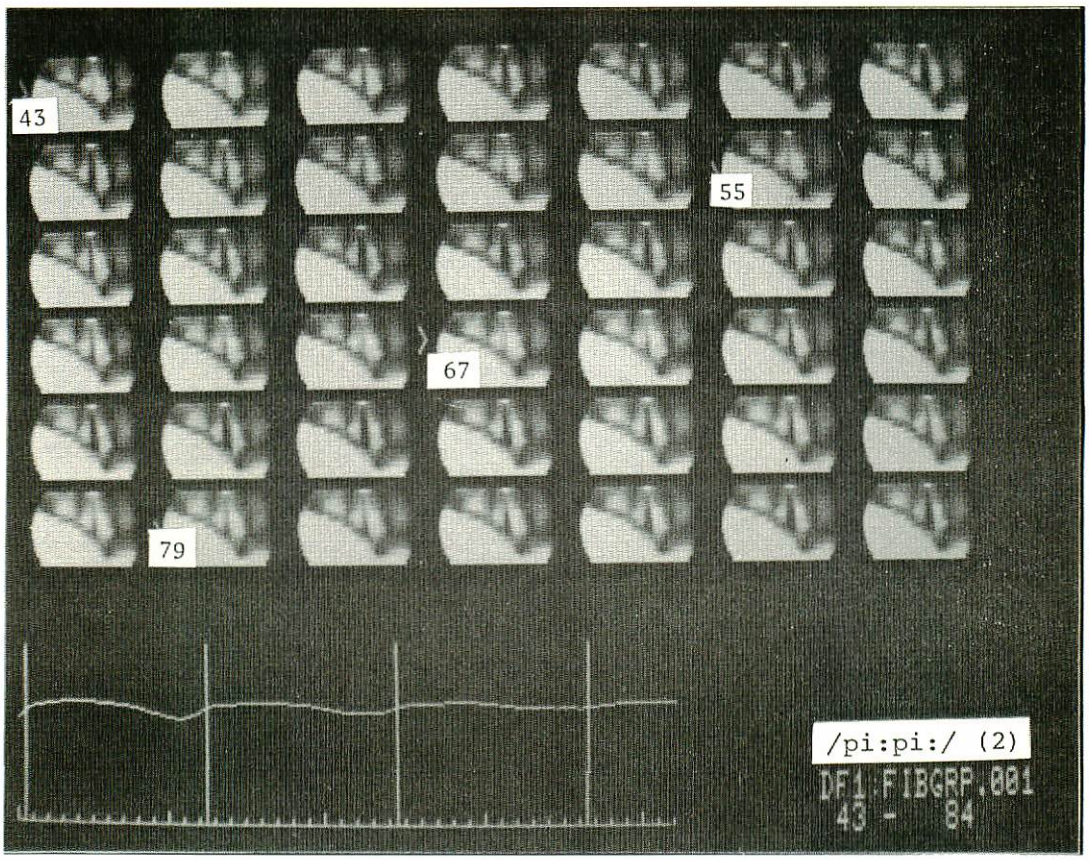

Fig. 5. The second 42 frames for /pi:pi:/.

Figure 6 shows the next 42 frames of the same utterance /pi:pi:/ from number 85 to 126. In this series, the vocal folds appear to approximate each other slightly at frame number 91 and 103. However, the audio signal stays flat, and there is no glottal closure in these samples.

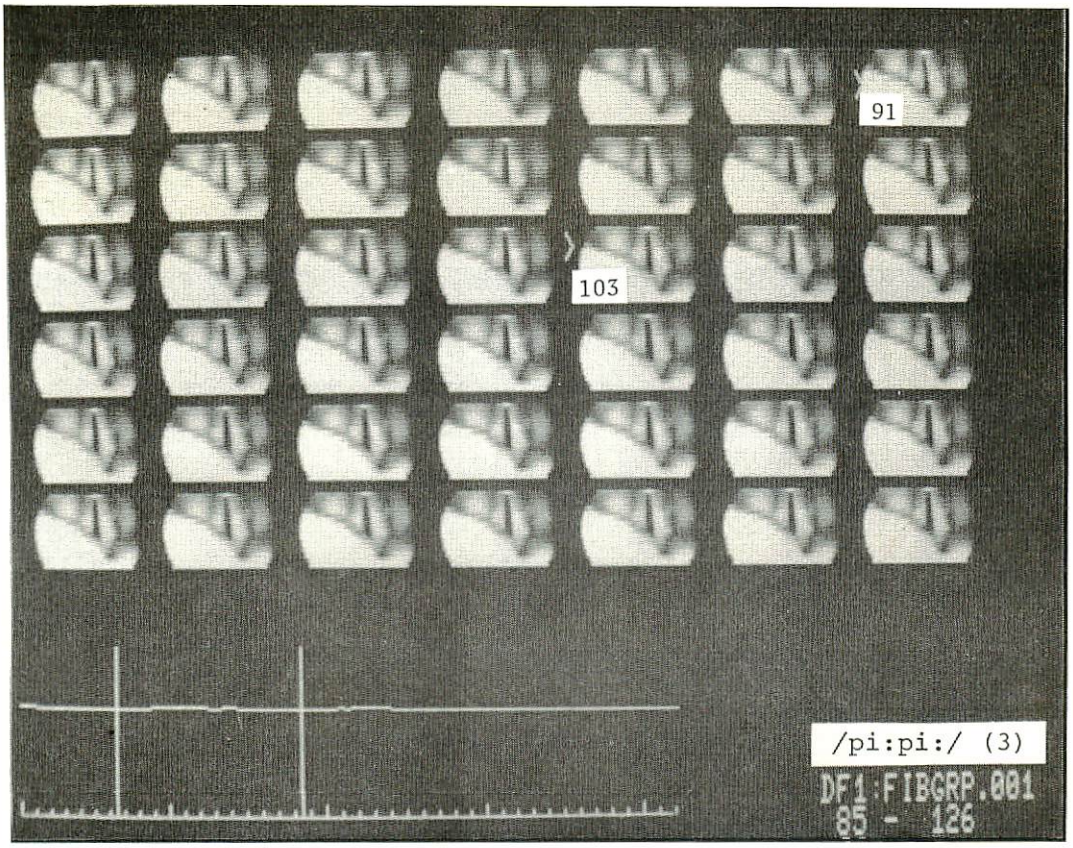

Fig. 6. The third 42 frames for /pi:pi:/.

In the next 42 frames from 127 to 168, the glottis stays open throughout, as shown in Figure 7. It can be seen that there is no appreciable variation in the glottal width in this series.

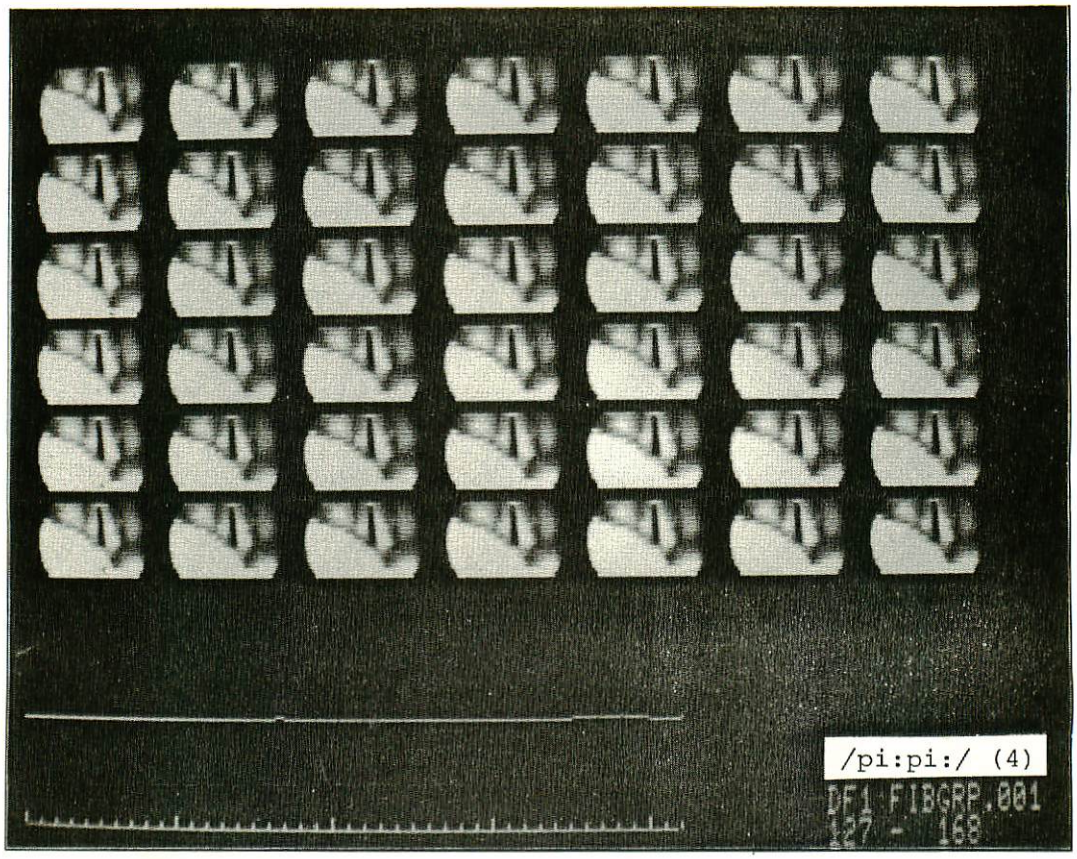

Fig. 7. The fourth 42 frames for /pi:pi:/.

In the next 42 frames from frame number 169 to 210, shown in Figure 8, it can be seen that the glottis is narrowing relatively quickly after frame number 185, and the closed phase can be observed at frame numbers 187, 197 and 207. In these 3 frames, incomplete closure is seen only at frame number 187. In other words, tight glottal closure is almost immediately achieved at the initiation of post-consonantal voicing. In this part of the utterance, F0 is about 200 Hz and is higher than that in the pre-consonantal period.

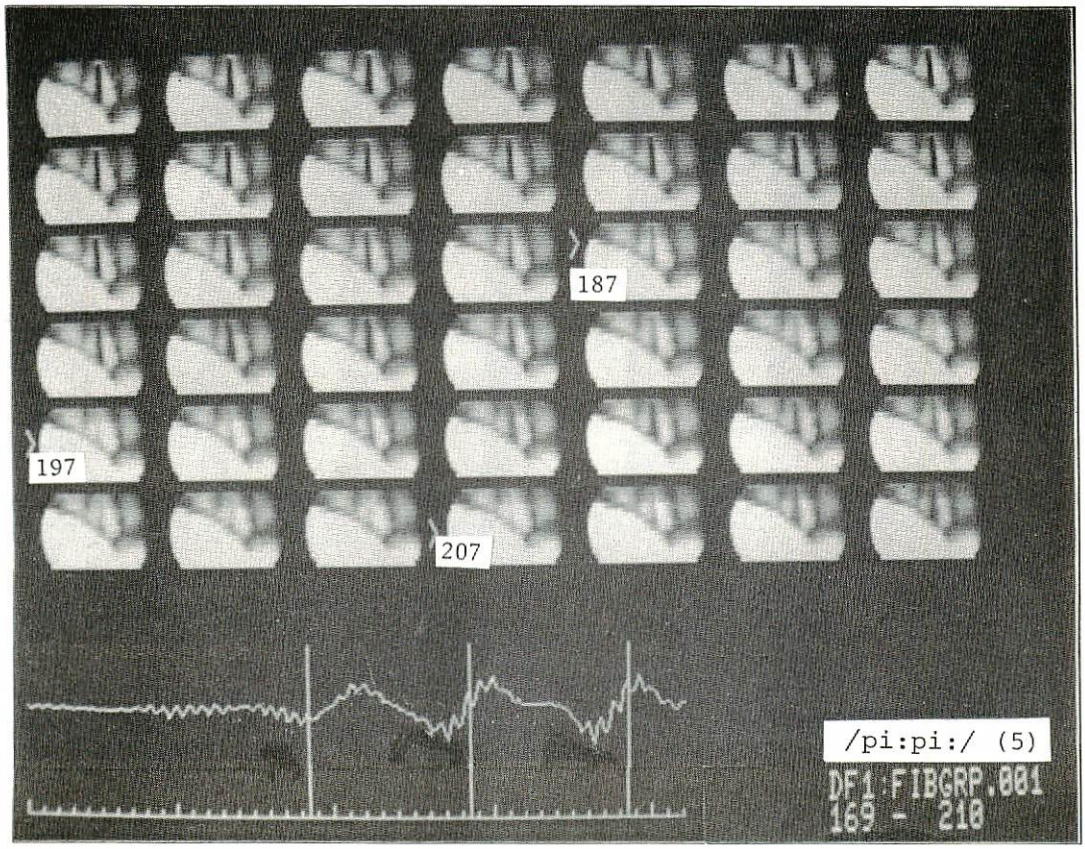

Fig. 8. The fifth 42 frames for /pi:pi:/.

In the following illustrations, the glottal behavior during the production of utterances containing an intervocalic fricative will be shown. Figure 9 shows the first 42 frames of the glottal images recorded immediately before the onset of /ʃ/ during the production of the utterance /pi:ʃi:/. In this series, the vocal fold approximations occur at frame numbers 9, 19, 29 and 39. In the latter two, an incomplete closure of the glottis with arytenoid separation can be seen. The degree of glottal opening during the open phase in this series appears to be wider than that observed in the previous series embedding the stop consonant /p/.

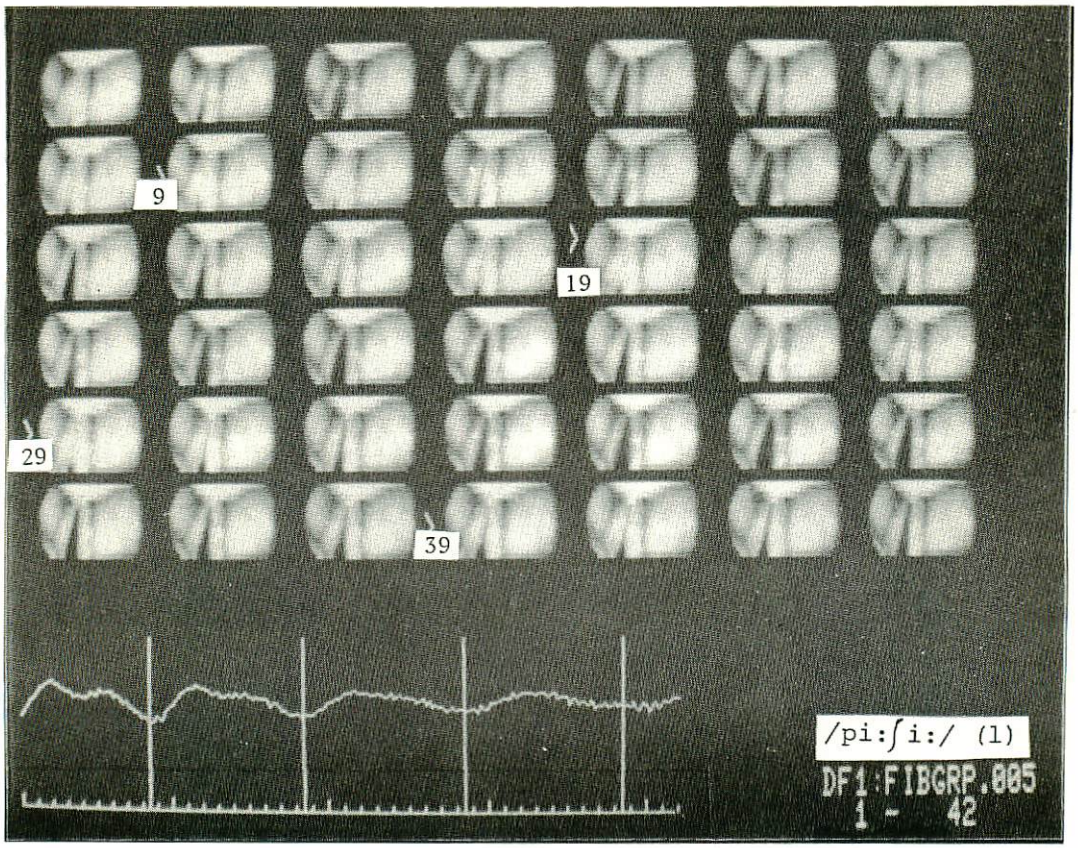

Fig. 9. The first 42 frames for the test word /pi:ʃi:/.

In the next 42 frames from number 43 to 84, shown in Figure 10, the arytenoids remain separated. However, a periodic glottal narrowing can be seen at frame numbers 50, 61, 73 and very slightly at 84. Corresponding to these narrowings, there is a tendency toward periodic amplitude modulation on the audio trace.

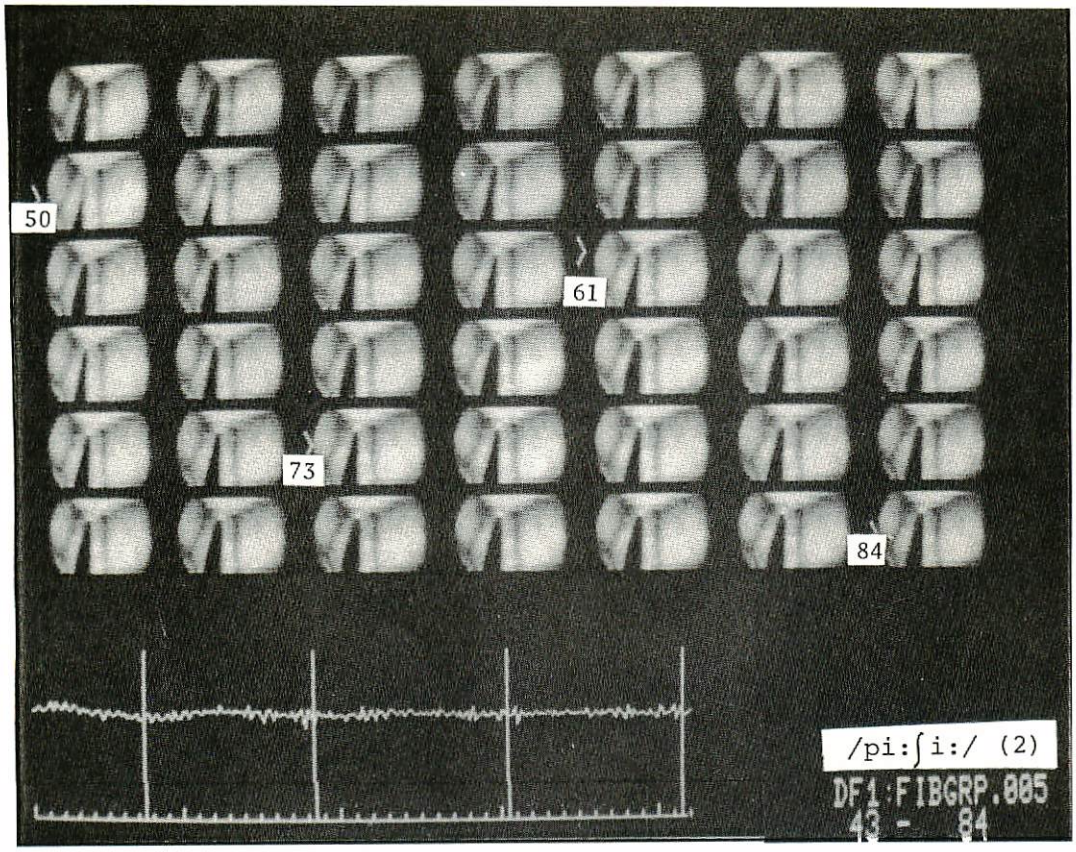

Fig. 10. The second 42 frames for /pi:ʃi:/.

Figure 11 shows the next series from frame number 85 to 126. There is no appreciable change in the glottal width during this period. The arytenoid separation appears to be wider than that in the series for the voiceless stop /p/, and the glottis shows a triangular shape throughout.

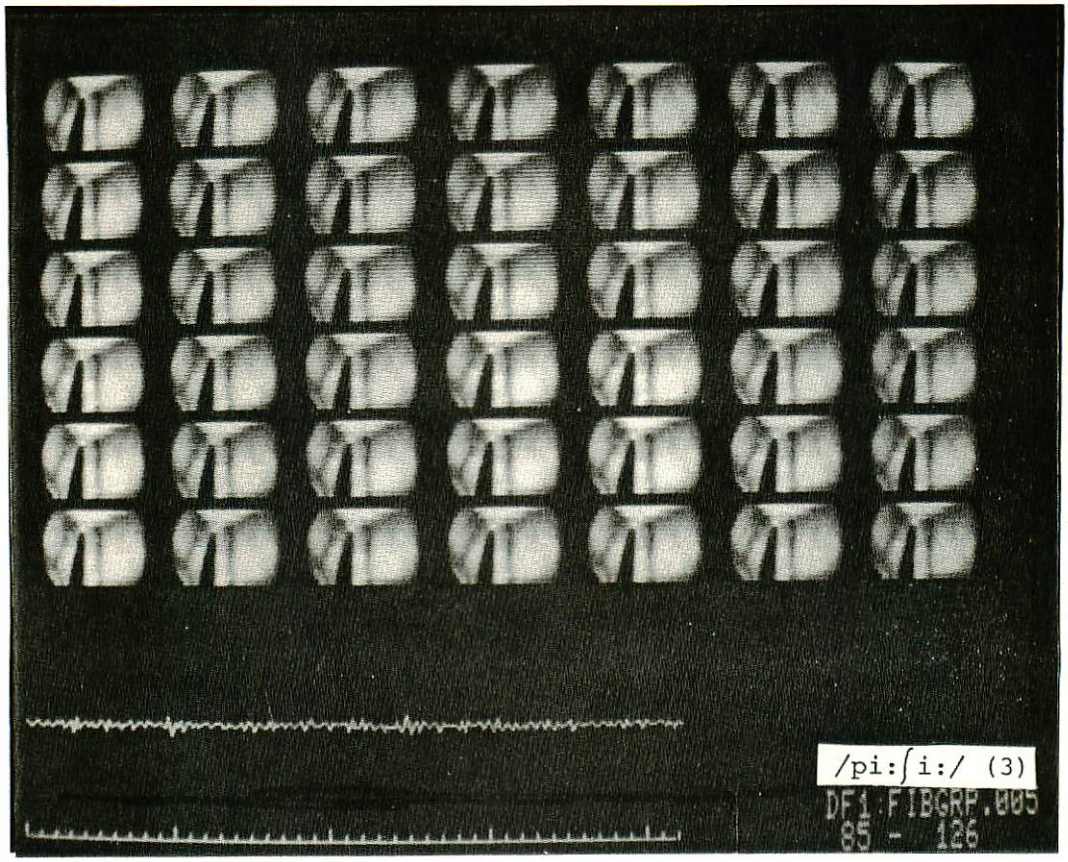

Fig. 11. The third 42 frames for /pi:ʃi:/.

In the next 42 frames from frame number 127 to 168, shown in Figure 12, glottal adduction appears to start around frame number 145 and progresses slowly.

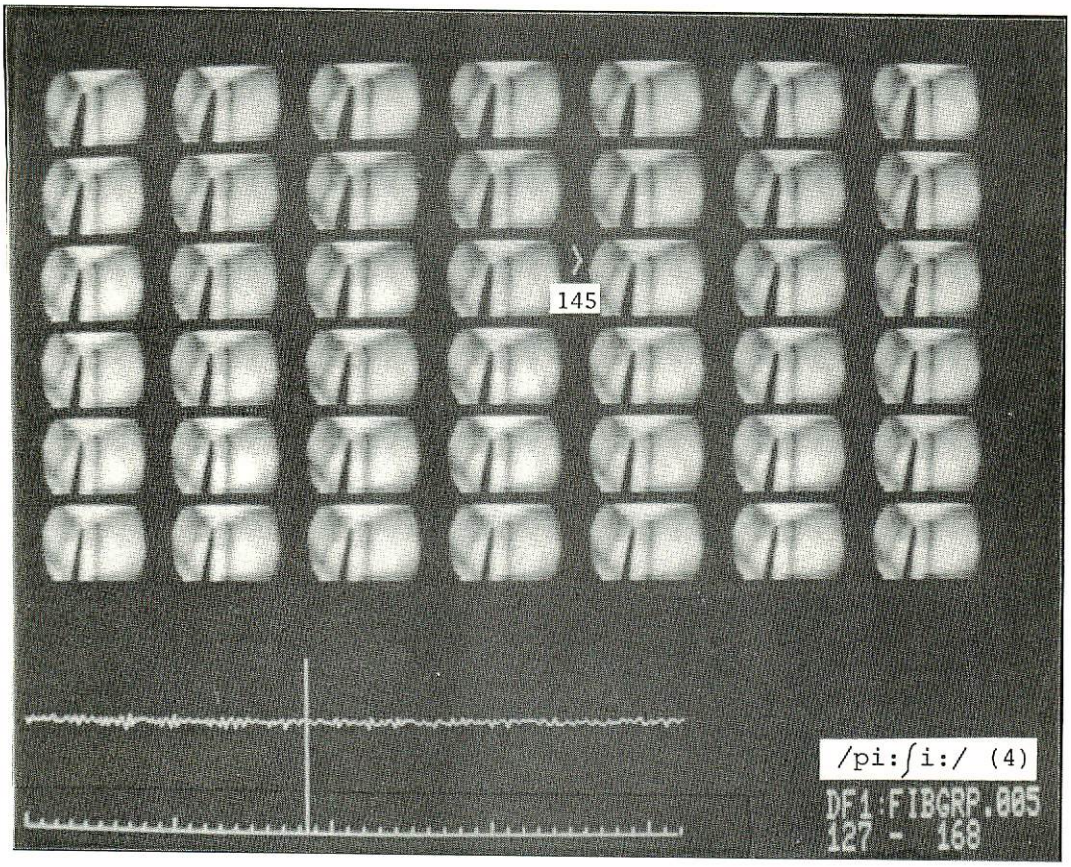

Fig. 12. The fourth 42 frames for /pi:ʃi:/.

Figure 13 shows the next 42 frames for /pi:ʃi:/ from frame number 169 to 210. Here, vocal fold vibration starts fairly suddenly. Although the glottal closure in the first cycle is still incomplete, as can be seen at frame number 174, it is complete from the next cycle, and a closed phase can be observed at frame numbers 183, 192, 201 and 210. In the last two cycles, there is no longer any arytenoid separation.

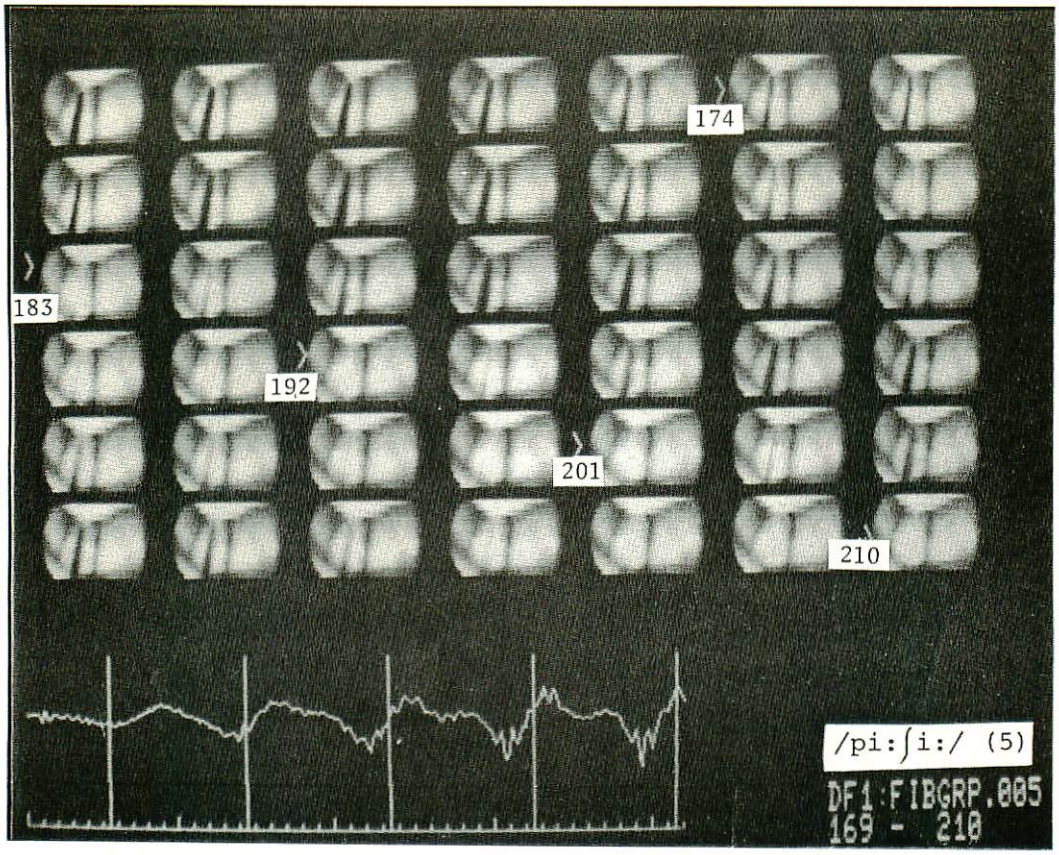

Fig. 13. The fifth 42 frames for /pi:ʃi:/.

As another example of a voiceless fricative, an utterance having an intervocalic voiceless /h/ (/pi:hi:/) was examined. Figure 14 shows a series of glottal images corresponding to the voiceless period of the intervocalic /h/. Even during this period, periodic glottal narrowings can be seen at frame numbers 141, 153 and 165. Corresponding to these narrowings, amplitude modulations in the noise signal on the audio trace can be clearly noted. It can also be seen that the glottal narrowing is limited at its anterior portion, and that the arytenoids remain separated from each other. A periodic glottal narrowing of this type continues throughout the entire voiceless period until arytenoid approximation starts, after which a complete glottal closure for full vocal fold vibration is achieved relatively quickly.

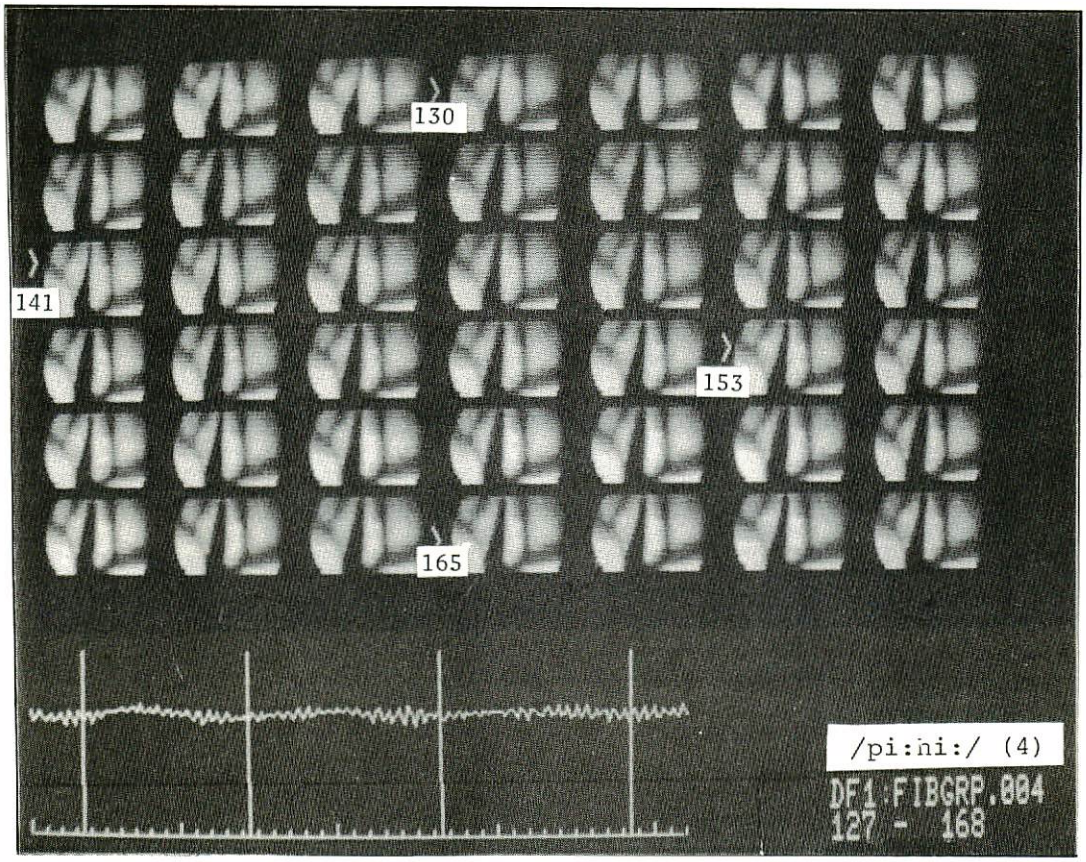

Fig. 14. The fourth 42 frames for the test word /pi:hi:/.

Comments

In this paper, preliminary results of observations on laryngeal behavior during running speech recorded by means of a high-speed digital image recording system were presented. In general, it was found that the glottis showed periodic narrowings during the period of the onset of a voiceless consonant, even after the separation of the arytenoids associated with the cessation of voicing. During the production of intervocalic voiceless /h/, in particular, periodic vocal fold approximation at the anterior portion continued throughout. At the release of the voiceless consonants, vocal fold vibration with complete glottal closure was achieved relatively quickly. These results are compatible with our previous findings obtained by means of photoglottography combined with measurement of the transglottal pressure difference, in which it was found that vocal fold vibration tends to be maintained at the onset of obstruents with less favorable physiological conditions for oscillation, while the vibration does not start after the voiceless period until more favorable conditions are obtained[6].

The data analysis in the present study is still at a stage of qualitative evaluation based on a visual observation of the photographed records of the monitor display. Although more precise and quantitative studies of the stored data must be made in the future, the new method of high-speed digital imaging is promising for analyzing the nature of laryngeal behavior during running speech. Simultaneous recordings of photoglottography and intraoral pressure are now being considered.

Acknowledgement

The present study was supported in part by a Grant-in-Aid for Scientific Research (No. 61480355) from the Japanese Ministry of Education, Science and Culture.

References

1) Sawashima, M. and Hirose, H.: Laryngeal gestures in speech production. In P.F. MacNeilage (Ed), Production of Speech (pp. 11-38), New York, Springer-Verlag, 1983.
2) Honda, K., Kiritani, S., Imagawa, H., Hirose, H. and Hashimoto, K.: High-speed digital recording of vocal fold vibration using a solid-state image sensor. Ann. Bull. RILP, 19, 47-53, 1985.
3) Kiritani, S., Imagawa, H. and Hirose, H.: Simultaneous high-speed digital recording of vocal fold vibration, speech and EGG. Ann. Bull. RILP, 20, 11-15, 1986.
4) Imagawa, H., Kiritani, S., Honda, K. and Hirose, H.: Improvements in the high-speed digital image recording system for observing vocal fold vibration. Ann. Bull. RILP, 20, 17-22, 1986.
5) Kiritani, S., Honda, K., Imagawa, H. and Hirose, H.:

Simultaneous high-speed digital recording of vocal fold vibration and speech signal. Proceedings ICASSP 86, Tokyo, Vol. 3, 1633-1636, 1986.
6) Hirose, H. and Niimi, S.: The relationship between glottal opening and the transglottal pressure difference during consonant production. In T. Baer, K. Sasaki and K. Harris (Eds.), Laryngeal function in phonation and respiration (pp. 381-390), Boston, College-Hill Press, 1987.