Color Display System for Connected Speech to be Used for the Hearing Impaired

AKIRA WATANABE, YUICHI UEDA, AND AKIYOSHI SHIGENAGA

Abstract—A color display system for the hearing impaired which converts connected speech signals into color pictures on a TV screen has been developed. This paper describes the principle, the function and the performance of the system, and the visual images of the patterns displayed using the system.

The system consists of a real-time formant tracker, a pitch detector, a memory system, a color coder, etc. In this system, the lowest three formant frequencies are extracted from voiced signals of connected speech by means of the formant tracker, and are converted to three primary color signals.

The three primary color signals as a time pattern are represented as a spatial color pattern on the TV screen using the memory system and the color coder. In unvoiced portions, colorless and dapple patterns can be seen.

The reproduced pattern using this system is not only beautiful, but also easy to understand intuitively. Especially, the visual experiments show that simultaneous contrast effect of colors caused by the spatial representation visually compensates for coarticulation effect on connected vowels.

I. Introduction

ALTHOUGH many electronic aids which transmit speech information to hearing impaired persons through the visual or tactile sense have been developed [1], [2], the information transmitted with such aids is much inferior to that transmitted through the auditory sense of normal hearing persons.

Those aids can be classified into speech training systems and speech reading systems. Most researchers who are engaged in the development of such aids seem to think vaguely that speech trainers (instruments) require the representation of only one or two features to be corrected, while speech reading systems need the general representation of whole speech signals.

We have doubted this idea on the basis of training experiments conducted on hard of hearing children using three kinds of speech trainers which we developed [3]. Those trainers, called the two-dimensional (2-D) pitch-intensity indicator [4], the intonation indicator, and the articulatory trainer of vowels [5], display one or two speech features on a CRT screen.

In the systematic training conducted for a year and 10 months, we concluded that each trainer was very useful for correcting speech features displayed on the CRT screen, but if the training was carried out with little attention to features other than those displayed, the undisplayed features deteriorated in the utterances. We deduced from the results that a system in which as many features as possible were displayed as a fused pattern would be better for speech training as well as for speech reading.

"Visible speech" deviced by Potter *et al.* [6] is representative of speech reading systems. It turns out to be a very good representation for speech analysis; however, contrary to expectations its pattern is not always understood by the hearing impaired. As a main reason why visible speech is hard to read, it has been suggested that speech spectrograms are modified by coarticulation effects inherent in the phoneme environment [7].

Based on the above considerations, our research has been directed toward developing new visible speech to be used for both speech training and reading.

One of our ideas was to synthesize a color pattern from speech features extracted by a speech analysis system because we expected that the visual effect on colors might compensate for the coarticulation effect on connected speech. This paper describes the system realized and specific characteristics of the visual patterns obtained by it.

II. System

A block schematic diagram of the color display system for connected speech is shown in Fig. 1. In the system, four parameters, that is, the lowest three formant frequencies (F_1, F_2, F_3) and the fundamental frequency (pitch F_0) are used to control the image displayed. In voiced portions of connected speech, color patterns determined by the three formants are displayed on a color TV screen, and in unvoiced portions a colorless and dapple pattern can be seen. The patterns are represented as spatial patterns whose time axis is given vertically on the screen. The pitch signal restricts the horizontal length of the color pattern in the voiced segment and indicates intonation.

The color display system can be partitioned into five parts consisting of a pitch detector, a formant tracker, a memory system, a color coder, and a blanking signal generator.

The four speech parameters (F_0, F_1, F_2, F_3) are extracted with the pitch detector and the formant tracker which both work in real-time and whose system parameters are automatically controlled by the input speech signal.

The extracted formants are converted into the signals corresponding to three primary color signals as follows:

$$\begin{aligned} &\text{red signal} && E_R = k(5F_1/F_3) \\ &\text{blue signal} && E_B = k(F_2/3F_1) \\ &\text{green signal} && E_G = k(3F_3/5F_2) \end{aligned} \tag{1}$$

where k is a constant.

From the above conversion equations, it can be expected that

Manuscript received April 9, 1984; revised June 11, 1984. This work was supported by the Japan International Business Machine Corporation and the Ministry of Education of Japan.

The authors are with the Faculty of Engineering, Kumamoto University, Kumamoto 860, Japan.

0096-3518/85/0200-0164$01.00

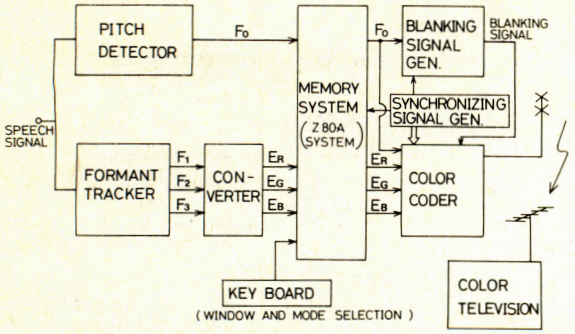

Fig. 1. Color display system for connected speech.

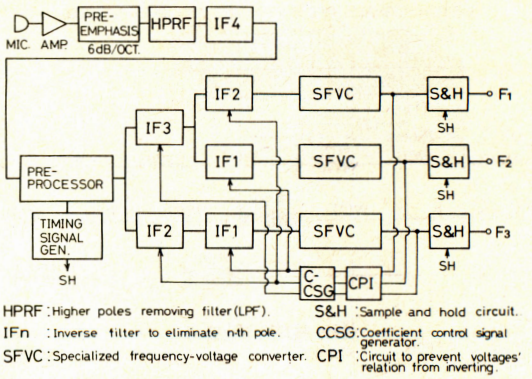

Fig. 2. Real-time formant tracker using inverse filters' control.

the circular ratios of the formant frequencies normalize influence of vocal tract lengths on the formant frequencies and that the coefficients of $\frac{1}{3}$, $\frac{3}{5}$, etc., make a neutral vowel colorless. Therefore, it will be possible in the system to display the five Japanese vowels as five different colors independently of the different vocal tract lengths due to age and sex of speakers, and to represent a more articulate vowel as a distinct color with higher saturation.

The memory system has a role in converting speech signals into a spatial pattern in which the time axis of speech is vertically represented. The four parameters, F_0, E_R, E_B, and E_G are digitized and then temporarily stored into four random access memories in this system. The stored signals can be read out in four different modes, that is, flow, storage, and cyclic flow forward and backward, selected by pushing a key on the keyboard shown in Fig. 1. Another function of the memory system is to change the display window length, i.e., the time length of the pattern capable of being displayed on the screen. Speech training and reading with the color display system become easy by the use of this memory system. It is also anticipated that the simultaneous contrast effect between adjacent colors, which are dsiplayed as a spatial patterns by the memory system, emphasizes those colors to compensate for the coarticulation effect in connected speech.

The color coder shown in Fig. 1 generates a composite color video signal from the three primary color signals according to the NTSC (National Television System Committee) system [8]. The composite color video signal consists of a sine wave modulated by two chrominance signals, a luminance signal, a color burst, and vertical and horizontal driving pulse trains. As the brightness of the colors is influenced not only by the luminance signal, but also by the amplitude of the modulated wave corresponding visually to color saturation in this color coder, the luminance signal is corrected by the amplitude of the modulated wave. Hence, all voiced patterns can be seen with almost the same brightness. This luminance correction is effective in eliminating the contrast of the brightness which makes the color pattern of connected speech hard to read.

In addition to the luminance correction, two kinds of modifications are adopted in the color coder. One is the restriction of horizontal lengths of voiced patterns controlled by the pitch signal. This function is realized by controlling a composite blanking signal with the pitch signal read out from the RAM. The other is the function which makes unvoiced patterns colorless and dapple. In the unvoiced portions where the level of speech signal is beyond the threshold but the pitch signal is not, the luminance signal alone is on-off switched by the random signal and the two chrominance signals are held at a zero.

III. Real-Time Formant Tracker

The most difficult problem in developing the color display system of speech was how to realize a real-time formant tracker with high accuracy. Representative methods for formant extraction in software, for example, the linear prediction method [9] and the analysis-by-synthesis method [10], have high accuracy but do not work in real-time at present.

Therefore, our effort was directed toward developing an accurate formant tracker with hybrid hardware. Fig. 2 shows the real-time formant tracker realized.

Initially, the system was completely simulated on a digital computer. Later the system was improved to yield higher accuracy and finally realized in hardware.

Simply speaking, the system controls the zero frequencies, *IF*1, *IF*2, *IF*3, of five inverse filters on the basis of average axis-crossing frequencies extracted from those filters' outputs as shown in Fig. 2. Since the average axis-crossing frequency is approximately proportional to the first moment of the spectrum, the zero frequencies of each inverse filter converge approximately to the lowest three formant frequencies. After convergence, the average axis-crossing frequencies of the output signals approximatley indicate the lowest three formant frequencies.

The inverse filter *IF*4 has a fixed zero to eliminate the fourth formant and its zero is manually switched according to whether the speaker is a male or a female. It is shown by the computer simulation that if the fourth zero is fixed at a frequency somewhat higher than the actual fourth formant frequency, it has little influence on the accuracy of extracting the lowest three formants [13].

By the processing of the specialized frequency-voltage converter (SFVC), the average axis-crossing frequency is extracted as a voltage from the output signals of the inverse filter. If the frequency band in which this system works is constant, it takes 100 ms to detect the average axis-crossing frequency of an intermittent sinusoid whose frequency and duty cycle are 1 kHz and 0.7, respectively [11]. This response is too slow to extract formants from connected speech. Hence the need for the preprocessor shown in Fig. 2.

This unit consists of three random access memories, an A/D converter, three D/A converters, a timing circuit, etc. Before the processing with this unit, speech waves are preemphasized (6 dB/oct) and then prefiltered to remove poles higher than the third. The resultant signals fed to the preprocessor are converted by the A/D converter into 12-bit samples at a rate

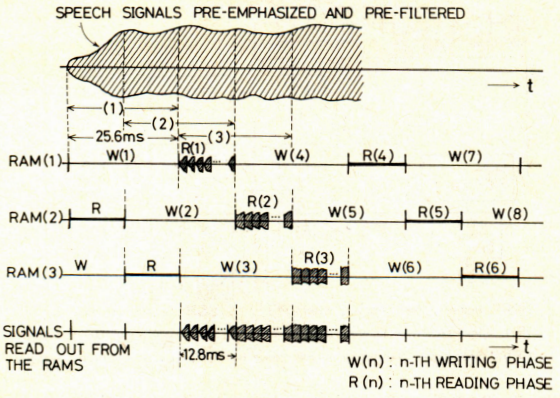

Fig. 3. Timing chart to show how to extract the repeated signals in real-time.

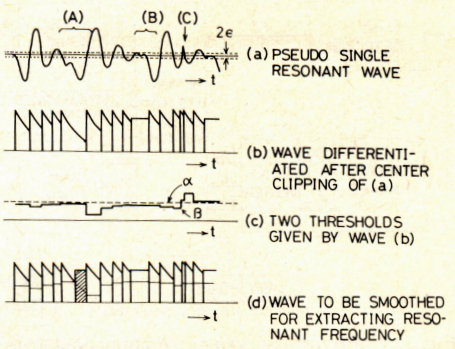

Fig. 4. Error reduction method used in the SFVC (*s*pecialized *f*requency-*v*oltage *c*onverter) of the formant tracker.

of 10 kHz. The digitized signals are temporarily stored in the three RAM's.

A relation between writing and reading phases in the RAM's is shown in the timing chart of Fig. 3. In each memory, 25.6 ms long samples are sequentially stored at intervals of 12.8 ms. The intervals between two writing periods are used to repeatedly read out the stored samples. This means that the sampling rate for reading out of the memories must be many times (16 times in this system) as high as the rate for writing.

Phase shifts for the writing and the reading between the three RAM's are 12.8 ms. That is, if the signals read out from those memories are mixed and then converted into analog signals, the resultant signals will be shaped as shown in the bottom of Fig. 3.

Thus, the frequency band of the signal fed to *IF*2 and *IF*3 connected to the preprocessor is 16 times that of the input. Thus, all of the system following the preprocessor must be designed to work in a frequency band which is 16 times as wide as that of the speech signal.

As the low-pass filters used for smoothing instantaneous axis-crossing frequency signals in the SFVC also have a high cut-off frequency, the convergence time in extracting the average axis-crossing frequency can be reduced by a factor of 16, that is, to about 6.25 ms. Thus, the time length of the frame, 12.8 ms, is enough for the axis-crossing frequency to converge. Although it takes 38.4 ms to get the formant frequencies from the instant of speech input, it may be regarded as almost real-time.

A design of the SFVC which estimates the resonant frequency from the output signal of the inverse filter is especially important to reduce the estimation error. An ordinary frequency-voltage converter consists of a center clipper with a fixed threshold, an amplifier, a differentiator, and a low-pass filter. Using those circuits, axis-crossing square waves are shaped by center clipping and amplification of the input signal, and then changed into waves with a constant exponential decay by differentiation as shown in Fig. 4(b). Voltages which correspond to axis-crossing frequencies are obtained by smoothing the differentiated waves with the low-pass filter.

As is well known in the acoustic theory of speech production [14], the resonant waveforms due to glottal excitation considerably differ from those due to impulse excitation in time intervals corresponding to open glottis. In the glottal excitation, we can often observe a tendency that the axis-crossing period of the resonant wave is prolonged during open glottis. This is accounted for by a differentiated glottal wave added to the resonant components excited by three impulses within a pitch period [14], [15].

Another factor which disturbs the axis-crossing periods is the existence of noise added to damped portions of the resonant signals [16].

These phenomena in speech waveforms will cause errors in the case where the resonant frequency is estimated as the axis-crossing frequency using the ordinary frequency-voltage converter.

The SFVC used in this system has two improvements. One is that the threshold level in the center clipping changes with the input signal level, as the threshold is determined from a constant attenuation value of the smoothed input signal. In the fixed threshold, axis-crossing frequency—exactly speaking, it should be called level crossing frequency—is dependent on the signal amplitude. The variable threshold has the effect of considerably reducing this dependence.

Another improvement is made for the low-pass filter smoothing the differentiated waves. First of all, in order to get the axis-crossing frequency in the prominent resonant periods alone, the output signal of the low-pass filter is held constant and fed to the same filter again during small resonant periods in which the input signal is not beyond the threshold ($\pm\epsilon$) determined dynamically. (See Fig. 4(a) part B.) By this operation, the internal conditions of the low-pass filter are kept as they are until the next large resonant wave comes in.

The prolonged axis-crossing period of the resonant signal caused by the glottal wave is corrected using two other thresholds. For example, assume that a long period shown in Fig. 4(a) part A has appeared. If the wave is processed by the operation mentioned above, the input signal to the low-pass filter of the SFVC will become the wave which is shown in Fig. 4(b). In the SFVC, the following two thresholds are extracted from the signal in Fig. 4(b). One of the thresholds is determined as 0.75 times the final value of the exponentially decayed signal. The threshold level is held constant from the final value to the next axis-crossing instant. (It is shown as threshold β in Fig. 4(c).) The other threshold is given as a half level of the SFVC output. (See threshold α in Fig. 4(c).)

The lower of the thresholds α and β is compared with the signal shown in Fig. 4(b) and then, if the decayed signal has a

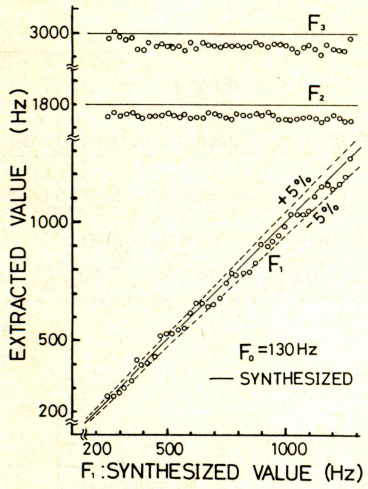

Fig. 5. Formant estimation errors.

level lower than the threshold, the input signal of the low-pass filter is held at the same value as the output level just before holding. As a result, the signal shown in Fig. 4(d) is obtained as the input signal of the low-pass filter.

Thus, we can extract the signal which essentially depends on the prominent resonant waves only. Even if a short axis-crossing interval exists as shown in Fig. 4(a) part C, the smoothed output of the signal shown in Fig. 4(d) will include little error in this system.

In order to investigate accuracy of extracting formant frequencies using this real-time formant tracker, synthetic vowels made with a computer simulation of a terminal analog system were used. The results in the case when the first formant of each stationary vowel was changed in the interval of 25 Hz are shown in Fig. 5. The first, second, and third formant bandwidths of each synthetic vowel are of 50, 80, and 115 Hz, respectively. The bandwidths of zeros of the inverse filters are fixed at 50, 100, and 150 Hz. The errors are almost all within ±5 percent and this result is the same as when the second or third formants in the synthetic vowels were changed.

Examples of formant extraction from natural speech are shown in Fig. 6(a) and (b). The continuity of the formant trajectories for the real-time system during voiced portions seem to be satisfactory in comparison with the trajectories obtained with the software formant tracker [13] and also in comparison with the sound spectrogram pattern shown in Fig. 6(c).

In Fig. 6(b), B_1, B_2, and B_3 indicate the three bandwidths of the inverse filters. As shown on the right-hand side of Fig. 6(b), reducing or augmenting the bandwidths of the inverse filters by a factor of three or so has little effect on the extracted formant frequencies.

Moreover, for the judgement by auditory perception, two short sentences and three phrases were synthesized with the computer simulation of a terminal analog synthesizer, using the formant frequencies extracted by the software method. From the first derivative of the glottal wave obtained by inverse filtering, five parameters were determined in each pitch period. The glottal wave was generated from those parameters and used as the driving source of the synthesizer. The result of preliminary listening shows that all of the synthesized speech is quite intelligible although the listeners perceive it to be synthetic.

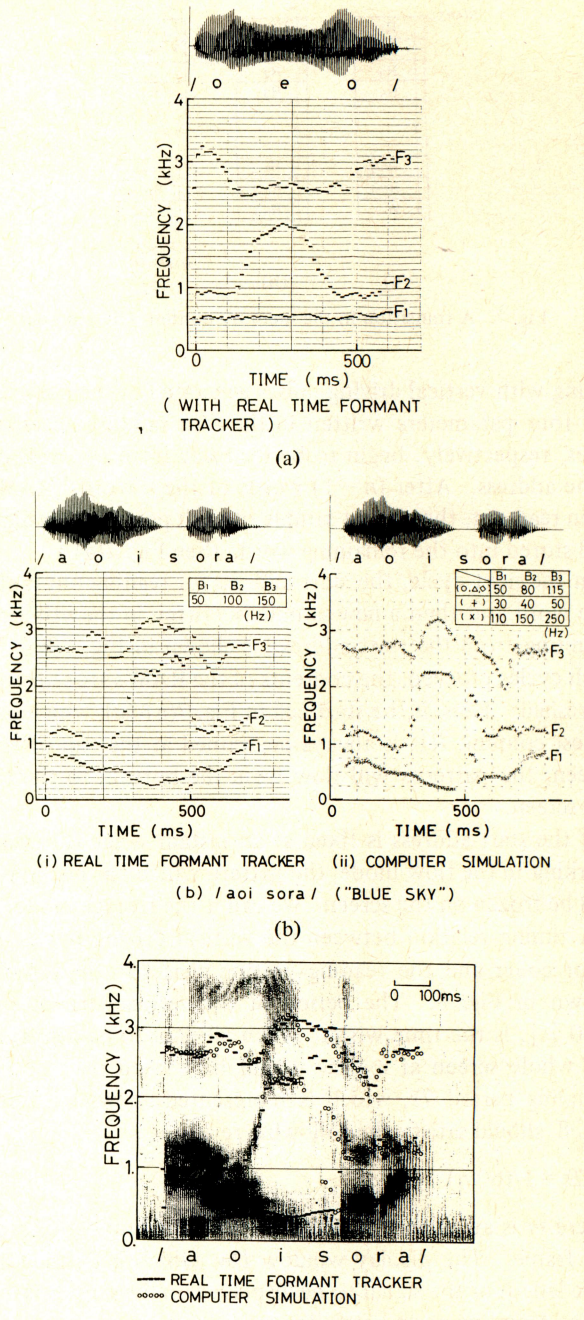

Fig. 6. Formant trajectories extracted from natural speech signals.

IV. Memory System

As mentioned in the Section II, speech patterns can be displayed in four modes on the screen of a color TV set. In the flow mode, for example, the patterns flow from the bottom to the top of the screen, that is, the speech patterns are vertically represented in the system. Such operations (as well as those needed for the other three modes) are realized with a microcomputer system in which four channels of input-output terminals are prepared to transmit analog signals.

The principle of the memory system is illustrated in Fig. 7 using the flow mode as an example. In the first cycle, synchro-

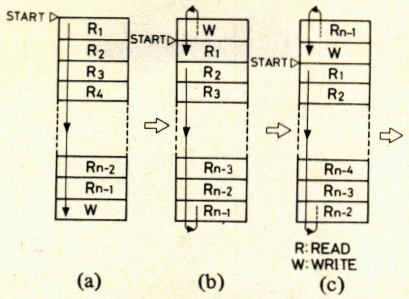

Fig. 7. A transition in the memory content (flow mode).

nizing with vertical driving pulses generated in the color coder, the four parameters written in the four random access memories, respectively, begin to be read out from the cells of the same address. After $(n-1)$ words of the n words' RAM have been read out, the newest sample of each parameter is digitized and stored into the remaining word. (See Fig. 7(a).)

In the next cycle, the start address for reading out is shifted one word and then a new sample is written in the nth location counting from the start address as shown in Fig. 7(b). When such cycles proceed, the contents of the RAM's are sequentially refreshed. Thus, if the start cell in Fig. 7(a)-(c), which has the oldest content, $R1$, is always positioned at the top of the TV screen, the patterns will flow from the bottom to the top of the screen.

If the start address is fixed at an instant when the system is working in the flow mode, the pattern which has just appeared will be frozen on the screen. We call it the storage mode.

A timing relation between the vertical driving pulses of the color coder and the reading-writing phases in the RAM's are shown in Fig. 8. The period of the vertical driving pulses, 1/60 [s], is the time within which all data to be displayed on the whole screen at a time would have been read out. The sampling period Ts, which is represented as twice a frame in Fig. 8, should satisfy the following relation:

$$Ts = 1/(60m) \quad [s] \tag{2}$$

where m is an integer which represents the number of samples per frame. The display window Dw, which is defined as the time length of the signal to be displayed on the screen, is related to the sampling period as follows:

$$Dw = Ts \times N \quad [s] \tag{3}$$

where N is an integer denoting the number of memory words to be used in a channel. From (2) and (3),

$$Dw = N/(60m) \quad [s] \tag{4}$$

Using (4), two integers m and N can be specified for the display window (Dw) appropriate to speech reading. As the sampling frequency of speech parameters is $60m$ [samples/s], the m should be larger than 1 or equal to 1 at least. The integer N is related to the information density to be displayed. Ideally, N should be as large as possible from the point of view of the continuity of the pattern visualized. In practice, $N \cong 200$ is adequate because one word of the memories occupies about two scanning lines on the screen. Therefore, if 256-byte RAM's are used, the integer N should be chosen as a maximal integer which is smaller than 257 and satisfies (4).

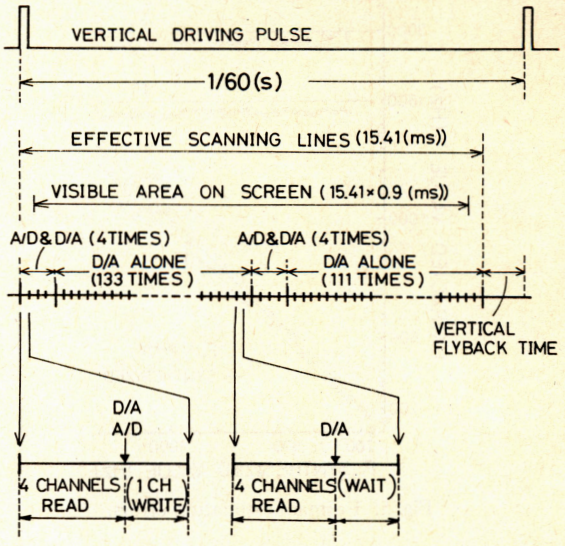

Fig. 8. An example of the timing relation between vertical driving pulses of the color coder and the reading-writing phases in the memory system.

TABLE I
RELATION BETWEEN DISPLAY WINDOW LENGTH AND EFFECTIVE MEMORY WORDS

Display window length, Dw [s]	1.00	1.25	1.75	1.92	2.00
Sampling frequency, $60m$ [Hz]	240	180	120	120	120
Number of writing samples in a frame, m [words/ch]	4	3	2	2	2
Number of samples displayed on the visible screen, N [words/ch.]	240	225	210	230	240

Table I shows five combinations of N and m which were chosen for the display windows appropriate to understand speech words or phrases.

In the memory system, it is possible to display the visible speech patterns in one of the four modes and one of the five display windows. The choice of the windows and the mode can be made instantaneously be selecting a program from the keyboard.

V. Color Coder and Blanking Signal Generator Modified by Pitch Signal

The composite color video signal in the NTSC system is basically synthesized from three primary color signals, as shown in a part of Fig. 9. That is, the three primary color signals are converted into two chrominance signals (E_I, E_Q) and a luminance signal (E_y) and then a color subcarrier of 3.58 MHz is modulated by the chrominance signals using quadrature phase modulation. The resultant signal, e_c, is added to the mixed signal of the corrected luminance signal $E_{y'}$, a color burst and a composite synchronizing signal. This yields the composite color video signal. The phase shift of the signal e_c from the color burst mainly determines the hue, and both the luminance signal $E_{y'}$ and the amplitude envelope of the signal e_c have an effect on the brightness.

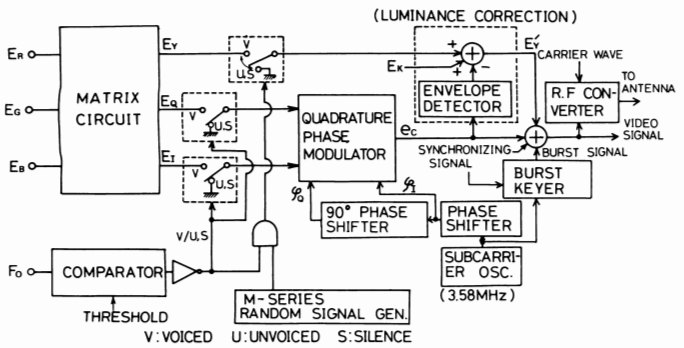

Fig. 9. Color coder modified by voiced/unvoiced detection and luminance correction.

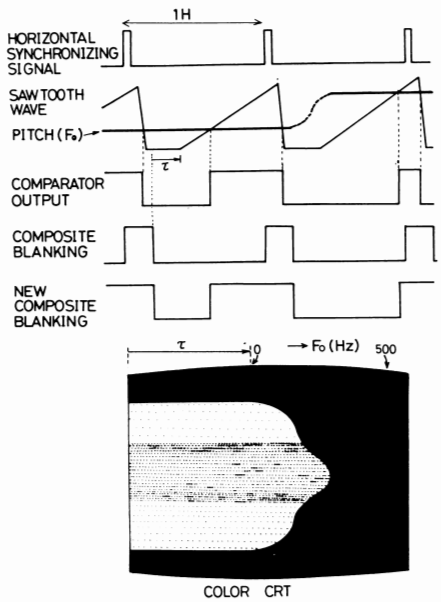

Fig. 10. Principle of the pitch display.

When it is desired to receive the speech patterns on two or more television sets, a high-frequency signal modulated with the composite color video signal is broadcast from an antenna.

In order to get speech patterns modified by pitch, the composite blanking signal is controlled by the pitch signal based on the principle shown in Fig. 10. First of all, a sawtooth wave train synchronized to the horizontal driving pulses is compared with the pitch signal. If the sawtooth wave level is higher than the pitch level, the output of the comparator holds a high level and if not, a low level. This output signal and the proper composite blanking signal are simultaneously fed to an OR gate, and its output signal is used as a new composite blanking signal.

Thus, the horizontal length of the speech pattern is controlled by the pitch signal so that we can get patterns similar to those shown at the bottom of Fig. 10.

VI. Display Examples

To give an idea of the patterns realized by the color display system mentioned above, some examples are shown in the color pictures in Fig. 11(a)-(d).

All of the utterances were pronounced by an adult male. Fig. 11(a) shows the five Japanese vowels /a/, /i/, /u/, /e/, and /o/ from top to bottom. The color image of each vowel is almost the same, independent of the voice quality of normal hearing speakers.

We notice from the examples of the short phrases in Fig. 11(b)-(d) that each vowel looks categorical and the same vowels in the different phrases have colors which belong to an identical category as a color, although the representation of consonants are yet incomplete. The categorical aspect of the connected vowels is caused by simultaneous color contrast. If the articulatory movement is quite slow (e.g., if the transition period between two adjacent vowels corresponds to the whole time interval displayed on the screen), the color change also will be gradual.

As the voiced consonants are displayed with colors only, it makes the short phrases hard to read. In particular, a very short voiced consonant, whose chromaticity is close to that of surrounding vowels may not be represented as a different color because of the assimilation effect of colors. This defect should be solved by overlapping other parameters for the voiced consonants, such as the gross features of the spectrum and its time changes as a fine change of luminance.

As a general rule, however, we believe that the patterns obtained with this system are very intuitive, beautiful, and interesting for children.

VII. Discussions—Visual Image of Isolated and Connected Vowels

A. Chromatic Distribution of the Five Japanese Vowels

Since the three primary colors change with formant frequencies, as shown in (1), individual differences of the utterances appear on the screen as differences of the chromaticity. We can estimate by simple calculation what chromaticity is reproduced by the NTSC system for given formant frequencies.

Fig. 12 shows a distribution of five Japanese vowels on a chromaticity diagram obtained by such calculations. The utterances were collected from 130 men, women, and children whose ages were between 7 and 23. From the speech signals of their utterances, the lowest three formant frequencies were extracted with computer analysis using the analysis-by-synthesis method. Furthermore, three primary colors were calculated with (1) and then the following equations were used to get CIE chromaticity (x, y) [17]:

$$\begin{aligned} X &= [0.61(E_R/E_{Rm})^{2.2} + 0.17(E_G/E_{Gm})^{2.2} \\ &\quad + 0.20(E_B/E_{Bm})^{2.2}]\, Yc \\ Y &= [0.30(E_R/E_{Rm})^{2.2} + 0.59(E_G/E_{Gm})^{2.2} \\ &\quad + 0.11(E_B/E_{Bm})^{2.2}]\, Yc \\ Z &= [0(E_R/E_{Rm})^{2.2} + 0.07(E_G/E_{Gm})^{2.2} \\ &\quad + 1.11(E_B/E_{Bm})^{2.2}]\, Yc \\ x &= X/(X+Y+Z), \qquad y = Y/(X+Y+Z). \end{aligned} \tag{5}$$

Here $0 < E_R \leqq E_{Rm}$, $0 < E_G \leqq E_{Gm}$, $0 < E_B \leqq E_{Bm}$, and Yc is the luminance of a standard light source. Actually, the luminance correction has an influence on chromaticity as shown in the Appendix. However, as the chromaticity change due to the corrected lumimance is small, it is neglected here.

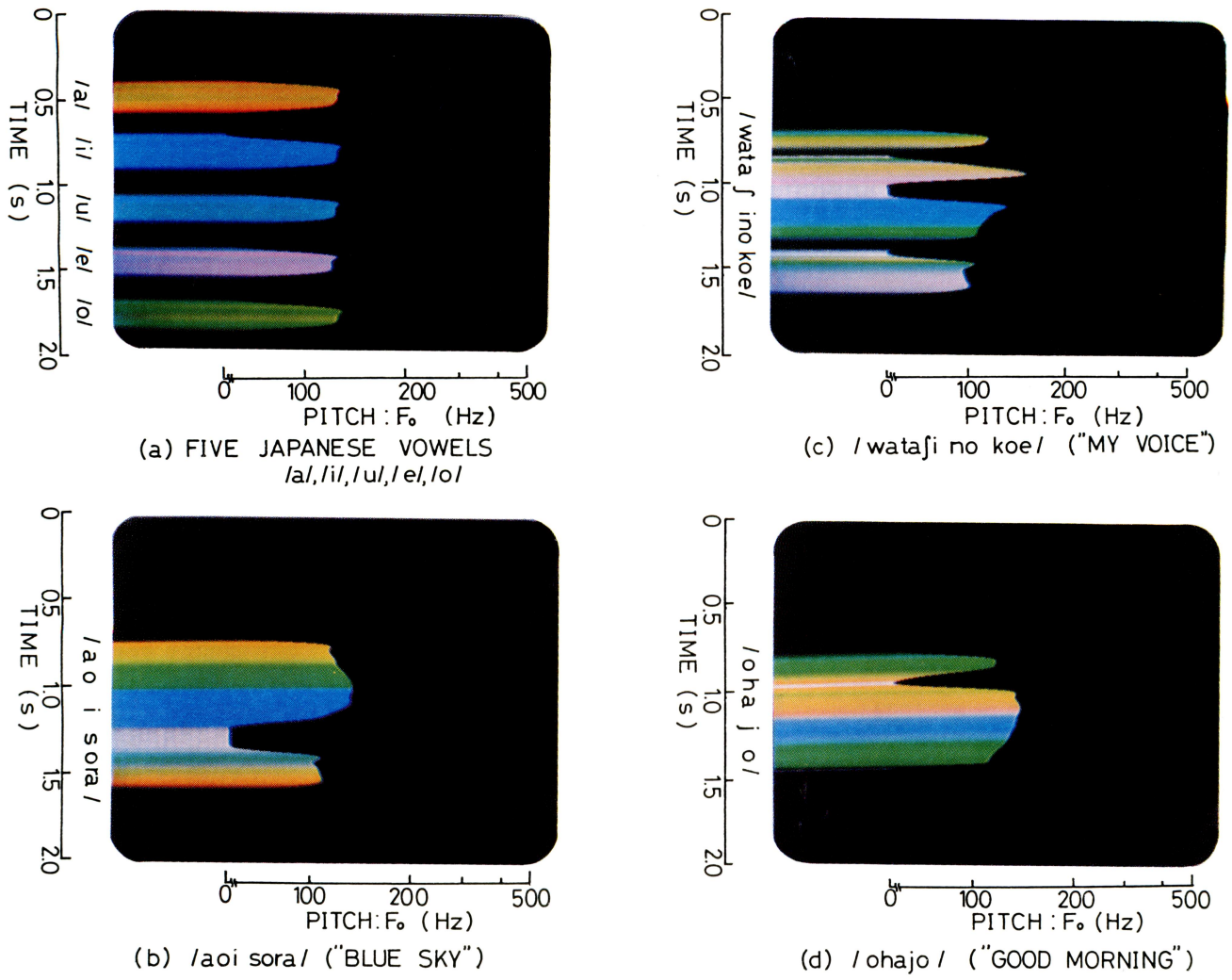

Fig. 11. Examples of speech patterns reproduced by the color display system.

The CIE chromaticities (x, y) which were calculated using (1) and (5) are represented on a two-dimensional orthogonal space in Fig. 12(a).[1]

Representative colors of the five vowels can be expected from the figure to be as follows:

/a/; orange-yellow-greenish yellow with high saturation
/e/; magenta with low saturation
/i/; blue with very high saturation
/o/; green-cyanic green with high saturation
/u/; cyan with middle saturation.

The elliptic regions surrounded by the dotted lines on the CIE chromaticity diagram show the areas of 2 times the standard deviation 2σ. The CIE chromaticity diagram is a physical space and not a uniform color space. Equally noticeable chromaticity differences on the diagram were investigated using color matching by MacAdam [18].

[1] Equation (5) represents the computation in a standard color TV monitor. The monitor used in this system, Tektronix 670A-1, somewhat differs from the standard in the coefficients and the exponents of (5). So empirical equations are obtained by measurement and used here instead of (5).

The equally noticeable chromaticity differences are very small in the vicinity of blue and become larger as the color approaches green, yellow, or red. Roughly speaking, the ratio of green to blue in the difference reaches about 3 in the range of colors reproducible by a color television. Therefore, even though the clusters of /i/, /u/, and /e/ are close to each other on the CIE diagram, the color differences among the three are not as close as those that can be differentiated.

In comparison with the F_1-F_2 diagram in Fig. 12(b), the five Japanese vowels on the CIE chromaticity diagram are distributed more separately as a result of the transformation by (1). Thus, (1) the normalizing equation based on the theory of wave propagation acoustic tubes which is easy to be realized with real-time hardware, is very efficient.

B. Compensation for Coarticulation Effect by Contrast of Colors

An expectation that coarticulation effect in connected speech might be compensated for by the contrast effect of colors has been investigated through visual experiments. In the experiments, the speech materials used are four sets of three con-

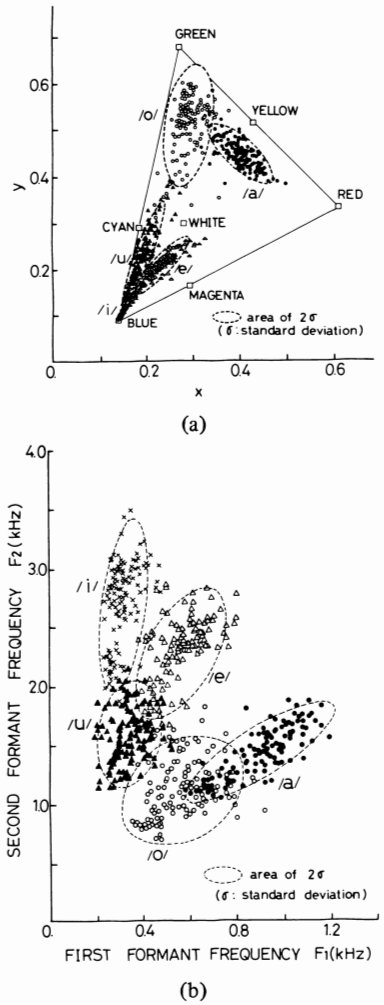

Fig. 12. Distribution of five Japanese vowels uttered by 130 speakers (age 7–23, males and females).

nected vowels /aua/, /eue/, /iui/, and /ouo/ uttered by an adult male in which the first and the third vowels are the same.

Subjects can select two kinds of patterns: one of the four speech sets displayed in the storage mode and a uniform color on the whole screen.

The uniform color can be changed by manually controlling the three primary color signals and the luminance signal. The task of the subjects is to manually adjust the uniform color so that it corresponds explicitly with the color of the central vowel of the three connected vowels.

Two subjects took part in the experiment. One is the experimenter who knows clearly the aim, and the other is an adult female who is quite ignorant of the visualization of speech.

The chromaticities were calculated from the three primary color signals and the luminance signal in the case where the uniform color has been adjusted. (See the Appendix.) Those chromaticities are compared with the physical chromaticity changes of the three connected vowels on the chromaticity diagram shown in Fig. 13.

In Fig. 13, hatched areas represent isolated vowels pronounced by the same speaker and the areas surrounded by

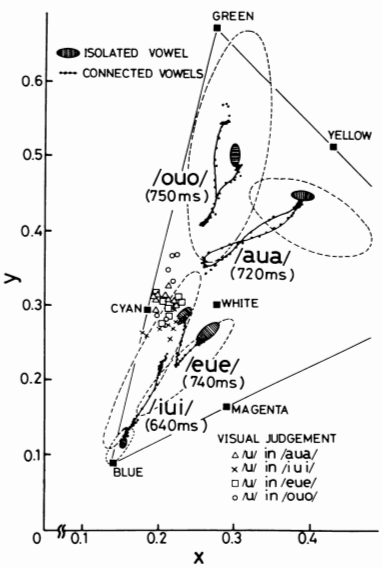

Fig. 13. Compensation for the coarticulation effect by contrast effect of colors.

dotted lines are those pronounced by the 130 speakers. Solid curves in the figure indicate physically estimated trajectories of the four sets of the three connected vowels, and the symbols (△, ×, □, ○) show visually evaluated chromaticities of the central vowels. Since there is no significant difference in the results between the two subjects, the same symbols are used in the figure for simplicity.

We notice from the figure that the physical chromaticities of the central vowels do not reach that of the isolated vowel /u/ except in the case of the /eue/. This is hypothesized to be due to the coarticulation effects of connected speech.

However, we can also see in the figure that in spite of such chromaticity changes, the visually estimated chromaticities are very close to that of the isolated vowel /u/. The visually estimated values are approximately located on the extension of the physical chromaticity trajectories.

From those results it appears that visual judgement is influenced by simultaneous contrast between two colors which represent two adjacent vowels.

Thus, it is expected that smoothing of speech parameters due to coarticulation may be compensated for by the contrast effect of colors. This effect would play an important role in understanding intuitively the visualized speech.

C. Recognition of Three Connected Vowels–A Comparison with a Formant Display

In a practical case, recognition of speech patterns must be carried out in a flow mode and in real-time. So as to investigate the practical usefulness of the color display system in the flow mode, we tried to compare the system with a formant display system which gives patterns similar to "visible speech."

An example of three connected vowels displayed by the formant display system is shown in Fig. 14. The three bright lines in the figure represent the lowest three formant frequencies and they evolve in real-time from the right to the left of

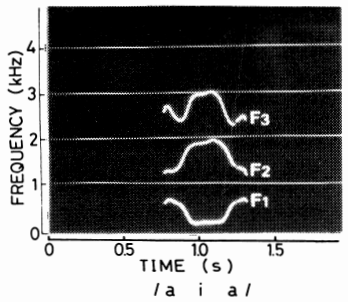

Fig. 14. An example of the pattern with the formant display system. (Three bright lines indicate $F1$, $F2$, and $F3$, respectively, from the bottom to the top.)

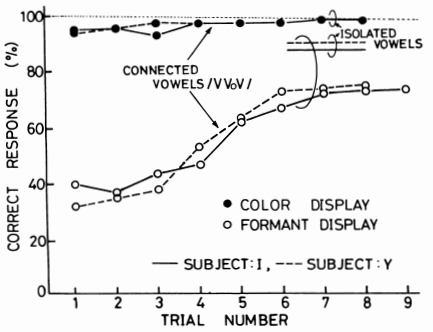

Fig. 15. Learning curves of the visual experiment for reading connected vowels using the color display system and the formant display system. (The learning was conducted 5 times for an isolated vowel in each trial and the connected vowels were not learned at all.)

the screen. Speech materials used for the tests are the five isolated Japanese vowels and all sets of three connected vowels (20 sets per speaker) which have the form $/V_0V_1V_0/$ uttered by two adult males and an adult female.

The learning was carried out using the isolated vowels alone and then the randomized sets of the three connected vowels were tested. Fig. 15 shows the learning curves obtained with the two systems. As we can point out immediately from the figure, the rate of correct responses in the color display system is much higher than it is in the formant display system. Furthermore, in the color display system, the correct responses saturate at a very high score (98 percent), almost without learning. In contrast, the formant display requires much learning time to get a saturated score and the score is not beyond 72 percent.

This excellent performance in the color display system is considered to be based on the normalization of the formant frequency dispersion due to differences of the vocal tract lengths and the compensation for coarticulation effect by the simultaneous contrast effect of colors.

VIII. Conclusion

A color display system for the hearing impaired which converts connected speech signals into color pictures on a TV screen has been developed. The most difficult problem in the development of the system was how to design a real-time formant tracker with high accuracy.

The new formant tracker was realized using the inverse filters' control system. The lowest three formant frequencies extracted by the formant tracker are converted into three primary color signals so that five Japanese vowels have different colors and so that individual differences of the vocal tract lengths of the speakers are normalized according to the theory of wave propagation acoustic tubes. The three primary color signals as a time pattern are represented as a spatial color pattern on the TV screen, using a memory system and a color coder.

The spatial representation of connected vowels causes a simultaneous contrast effect of colors so that the coarticulation effect which appears in the formant shifts of the connected vowels is visually compensated for by the contrast effect.

Based on the recognition experiments compared with a formant display system, it appears that the color display system has much better performance at least in the display of connected vowels.

One of our aims for the future development of this system is the representation of voiced consonants. However, even if its realization is difficult for the present, this system would be of considerable use for speech training of the hearing impaired.

Appendix

A. Influence of Luminance Correction on Chromaticity

In the transmitter of the NTSC system, as shown partially in Fig. 9, the three primary color signals are converted into two chrominance components (E_I, E_Q) and a luminance signal E_y using the following equation, and then transmitted [19].

$$\begin{bmatrix} E_y \\ E_I \\ E_Q \end{bmatrix} = \begin{bmatrix} 0.30, & 0.59, & 0.11 \\ 0.60, & -0.28, & -0.32 \\ 0.21, & -0.52, & 0.31 \end{bmatrix} \begin{bmatrix} E_R \\ E_G \\ E_B \end{bmatrix}. \tag{A1}$$

In the receiver, the three transmitted signals (E_y, E_I, E_Q) are inversely transformed to the three primary colors by the equation

$$\begin{bmatrix} E_{Rr} \\ E_{Gr} \\ E_{Br} \end{bmatrix} = \begin{bmatrix} 1, & 0.96, & 0.63 \\ 1, & -0.28, & -0.64 \\ 1, & -1.11, & 1.72 \end{bmatrix} \begin{bmatrix} E_y \\ E_I \\ E_Q \end{bmatrix} \tag{A2}$$

and then a color is synthesized as a mixture of the three primary colors.

In the luminance correction we used in this system, ($E_y + \Delta E_y$) should be used in (A2) instead of E_y because the luminance signal is modified by the amplitude envelope of the signal e_c shown in Fig. 9. Therefore, the reproduced primary colors (E_{Rr}, E_{Gr}, E_{Br}) all include an increment ΔE_y. Thus, the chromaticity has to be calculated according to the equations in which $E_R + \Delta E_y$, $E_G + \Delta E_y$, and $E_B + \Delta E_y$ are subsitituted instead of E_R, E_G, and E_B, respectively, in (5). In the equations, ΔE_y is given by

$$\Delta E_y = E_k - a(E_I^2 + E_Q^2)^{1/2} \tag{A3}$$

where E_k and a are constants which are experimentally determined.

As mentioned above, the chromaticity of the displayed colors is influenced by the luminance correction. Therefore, if the luminance signal is changed independently of the three primary color signals as conducted in the experiment, which has been descibed in Section VII-B, the chromaticity needs to be calculated taking the luminance signal into account.

ACKNOWLEDGMENT

The authors wish to deeply thank their research associate, the late S. Kisu, for his comments and assistance in the early phases of this work. They also thank M. Isayama, O. Matsuno, T. Ikeda, and Y. Morikawa, graduate students who collaborated in the experiments and in the development of the system.

REFERENCES

[1] H. Levitt, J. M. Pickett, and R. A. Houde, Eds., *Sensory Aids for The Hearing Impaired.* New York: Wiley, 1980.
[2] A. Risberg, "Speech coding in aids for the deaf: An overview of research from 1924 to 1982," *STL-QPSR*, vol. 4, pp. 65-98, Jan. 1983.
[3] S. Kisu and A. Watanabe, "Systematic use of three types of visual trainers to hard of hearing children" (in Japanese), *J. Acoust. Soc. Japan*, vol. 34, pp. 576-584, Oct. 1978.
[4] A. Watanabe and H. Okamura, "Effect of speech training by two-dimensional pitch-intensity indicator on deaf children" (in Japanese), *J. Acoust. Soc. Japan*, vol. 31, pp. 179-188, Mar. 1975.
[5] A. Watanabe and S. Kisu, "Three types of speech trainers by a visual display system" (in Japanese), *J. Acoust. Soc. Japan*, vol. 34, pp. 569-575, Oct. 1978.
[6] R. Potter, G. Kopp, and H. Green, *Visible Speech.* New York: Van Nostrand, 1947.
[7] A. M. Liberman, F. S. Cooper, and M. Studdert-Kennedy, "Why are speech spectrograms hard to read," *Amer. Ann. Deaf*, vol. 113, pp. 127-133, 1968.
[8] NHK, Japan Broadcasting Corporation Ed., *Color Television* (in Japanese). Japan: Japan Broadcasting Publishing Corporation, 1971, pp. 11-12.
[9] J. D. Markel and A. H. Gray, *Linear Prediction of Speech.* Amsterdam, The Netherlands: Springer Verlag, 1976.
[10] C. G. Bell, H. Fujisaki, J. M. Heinz, K. N. Stevens, and A. S. House, "Reduction of speech spectra by analysis-by-synthesis techniques," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1725-1736, 1961.
[11] A. Watanabe, "A real-time formant tracker using inverse filters," *STL-QPSR*, vol. 3-4, pp. 1-30, Jan. 1980.
[12] J. L. Flanangan, *Speech Analysis Synthesis and Perception.* Amsterdam, The Netherlands: Springer-Verlag, 1965, pp. 142.
[13] Y. Morikawa and A. Watanabe, "Formant extraction method by inverse filters' control using selected waveform information" (in Japanese), Tech. Group Rep., EA 82-57, IECE Japan, Jan. 1983.
[14] G. Fant, "Glottal source and excitation analysis," *STL-QPSR*, vol. 1, pp. 85-107, 1979.
[15] ——, "Vocal source analysis–A progress report," *STL-QPSR*, vol. 3-4, pp. 31-53, Jan. 1980.
[16] J. N. Holmes, "Formant excitation before and after glottal closure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Philadelphia, PA, 1976, pp. 39-42.
[17] NHK, Japan Broadcasting Corporation, Ed., *Color Television* (in Japanese). Japan: Japan Broadcasting Publishing Corporation, 1971, pp. 179. .
[18] D. L. MacAdam, "Visual sensitivities to color differences in daylight," *J. Opt. Soc. Amer.*, vol. 32, pp. 247-274, 1942.
[19] NHK, Japan Broadcasting Corporation, Ed., *Color Television* (in Japanese). Japan: Japan Broadcasting Publishing Corporation, 1971, pp. 220.

Akira Watanabe was born in Sendai, Japan, in 1937. He received the B.E. degree in electrical engineering in 1962, and the M.E. and D.E. degrees in electrical and communication engineering from Tohoku University, Sendai, Japan, in 1964 and 1968, respectively.

He served as a Research Associate from 1967 to 1968, and as an Associate Professor from 1968 to 1981 with the Faculty of Engineering, Kumamoto University, Kumamtoto, Japan. He investigated a real-time formant tracking system as a Guest Researcher from 1978 to 1979 at the Royal Institute of Technology (KTH), Stockholm, Sweden. He is now a Professor with the Department of Information Engineering, Kumamoto University, Kumamoto, Japan, and is researching the application of speech information processing to aids for the deaf.

Dr. Watanabe is a member of the Institute of Electronics and Communication Engineers of Japan and the Acoustical Society of Japan.

Yuichi Ueda was born in Kumamoto, Japan, in 1955. He received the B. E. degree in electronics and the M.E. degree in electrical engineering from Kumamoto University, Kumamoto, Japan, in 1978 and 1980, respectively.

In 1980, he joined the Kumamoto Radio Technical College, Kumamoto, Japan, as a Research Associate. Since 1982 he has been with the Faculty of Engineering, Kumamoto University, as a Research Associate. His research interests are in the area of applications of speech processing to electronic aids for the hearing impaired.

Mr. Ueda is a member of the Institute of Electronics and Communication Engineers of Japan and the Acoustical Society of Japan.

Akiyoshi Shigenaga was born in Kumamoto, Japan, in 1958. He received the B.E. degree in electronics and the M.E. degree in electrical engineering from Kumamoto University, Kumamoto, Japan, in 1981 and 1983, respectively.

He joined NHK, Kagoshima Station, Kagoshima, Japan, in 1983. His work has been in the field of automated meteological data acquisition systems developed by NHK.

Mr. Shigenaga is a member of the Institute of Electronics and Communication Engineers of Japan.