Ann.Bull.RILP(1989)No.23, 173-185

HIGH-SPEED SPEECH ANALYSIS SYSTEM USING A PERSONAL COMPUTER WITH DSP AND ITS APPLICATIONS TO PRONUNCIATION TRAINING

Hiroshi Imagawa and Shigeru Kiritani

1.Introduction

A personal computer-based speech analysis system was constructed which employed a floating point DSP. The system performs spectrographic analysis, formant analysis and pitch analysis of the speech signal. Formant and pitch extraction can be performed almost in real time. By exploiting this real time capability, pilot systems for speech pronunciation training were developed: an intonation training system for the learners of English, a pitch accent training system for the learners of Japanese and a vowel pronunciation training system. This paper describes the basic characteristics of these systems.

2 High-speed speech analysis system

2.1 Overall characteristics of the speech analysis system

Fig.1 shows the hardware configuration of the system. The commercially available DSP-board, A/D board and D/A board are attached to the expansion bus of the personal computer (NEC PC-9801VX). The computer has a 640kbyte main memory. The DSP board contains a 32bits floating point DSP chip (NEC μPD77230) and 32kbytes of external RAM. The external RAM stores the program and the data. Nominal speed of ths DSP chip is 6.6MIPS. In this system, we use a floating point DSP instead of an integer type DSP to avoid a programming burden caused by the computational errors due to integer operations. The quantization levels of the A/D and D/A converter are 12bits. The A/D converter has a Direct

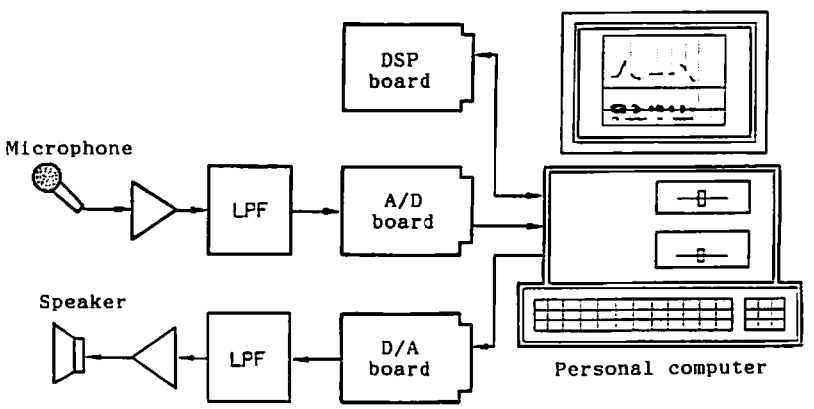


Fig. 1 Hardware configuration of the high-speed speech analysis system.

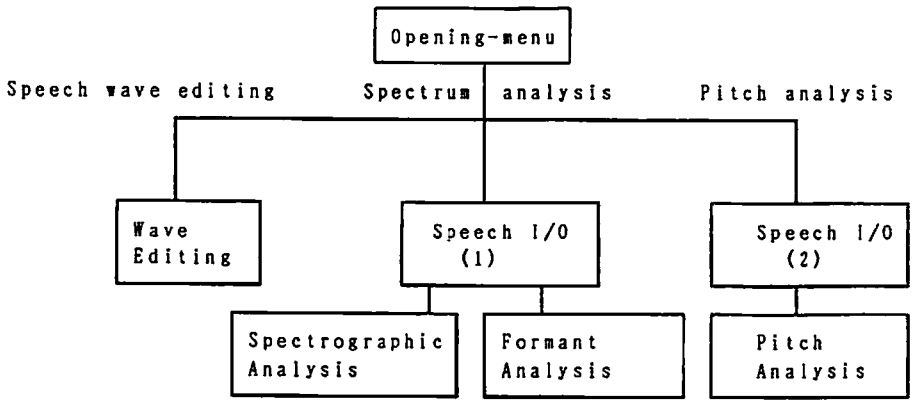


Fig. 2 Software configuration of the high-speed speech analysis system.

Memory Access function. The hardware configuration described above enables simultaneous operation of three processes: a sampling of the speech signal through the A/D converter, a signal proccesing by the DSP and a display of the analyzed data by the CPU. A special program produces a hard copy of the data displayed on the monitor screen with a laser printer with in 10 seconds.

The high-speed speech analysis program contains three different modes of data processing: 1) speech wave editing, 2)spectrum analysis and 3)pitch analysis. As shown in Fig. 2, there are 7 different patterns of screen display which correspond to different stages of data processing. In each display pattern, the command menu is displayed at the top of the screen. Through the entire program, all of the operations can be performed by selecting a desired operation with the mouse cursor and clicking the mouse button. At the opening-menu screen, a user can select one of the three data processings described above.

The main program is implemented by N88BASIC(MS-DOS version). LPC analysis, FFT analysis and pitch analysis are performed by the DSP. The DSP program for these analyses was written by using the cross-assembler on MS-DOS. The program is downloaded to the external RAM when execution of a selected proccess is required. Several special routines for graphics-control, A/D and D/A control were written using a Macro-Assembler.

2.2 Speech wave editing

In this mode, the speech signals can be sampled through an A/D converter and the amplitude envelope of the speech wave displayed (Fig. 3). On the envelope curve, a part of the speech signal can be monitored through a D/A converter or can be stored in a disk file. Selection of part of the speech signal is performed by placing time marks on the speech envelope with the mouse.

Commands which can be performed in this mode are listed below.

[A/D]	Sample the speech signal at 10kHz and display the speech envelope.
[D/A]	Output a specified portion of the speech signal through a D/A.
[save]	Save a specified portion of the speech signal in a disk file.
[load]	Load speech data from a disk file and display the speech envelope.
[cut]	Cut a selected speech signal and store it in a paste buffer.
[paste]	Insert the speech signal in a paste buffer at a specified position.
[silence]	Deposit zeros at specified intervals.

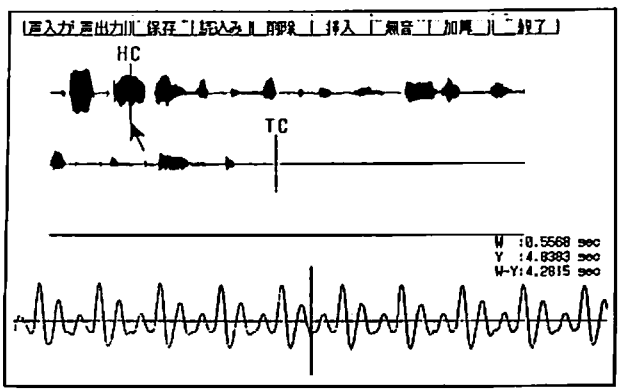


Fig. 3 Monitor screen in the speech wave editing mode. The time scale expanded speech wave around HC is displayed at bottom of the screen. HC and TC indicate Head-cursor and Tail-cursor, respectively.

2.3 Spectrographic and formant analysis program

In this mode, a spectrum analysis of the speech wave is performed and the spectrographic pattern is displayed on the monitor screen. It is also possible to perform a formant extraction and to display formant curves.

First, the program goes into the speech signal I/O mode, in which the speech signal is sampled through an A/D converter, and a part of the speech signal is selected for spectrum analysis. Several parameters for spectrographic analysis are also selected.

In the spectrographic mode, Several types of spectrogram are produced on the monitor screen: wide, narrow, section etc., as shown in Fig. 4. It is also possible to display the LPC spectrum envelope in the section mode. Spectrum analysis is performed by the FFT program on the DSP. The spectrum envelope is also calculated on the DSP. It takes about 15 seconds to produce a 3-second spectrum with wide-band mode and a frame-shift interval of 5ms. A hard copy of the spectrogram is obtained with a laser printer. Gray scale of the hard copy is produced through a dither-pattern.

In the formant analysis, the time functions of the formant frequencies are displayed. Formant frequencies are extracted by a peak picking of the LPC spectrum[1)2)]. The window and the frame-shift are the same as for the spectrographic analysis. Details of the algorithm will be described later.

The following commands are available both in the spectrographic analysis and formant analysis.

[A/D] Sample the speech signal at 10kHz and display the speech envelope.

[D/A] Output a specified portion of the speech signal through a D/A.

[save] Save a specified portion of the speech signal in a disk file.

[load] Load speech data from a disk file and display the speech envelope wave.

[narrow]/[wide] Select an analysis window width.
narrow : 51.2ms
wide : 6.4ms

[sona]/[formant] Select an analysis mode.
sona : Spectrographic analysis
formant: Formant curve analysis

[5.0ms]/[2.5ms] Select an analysis frame-shift.

[gray scale] Set the brightness and contrast of the spectrographic display.

[execute] Execute a selected analysis and display the result.

[end] Return to opening-menu.

[pitch] Superimpose the pitch curve onto the spectrogram.

[copy] Produce a hard copy of the screen.

[color] Select display mode (16 colors or 16 shades of gray).

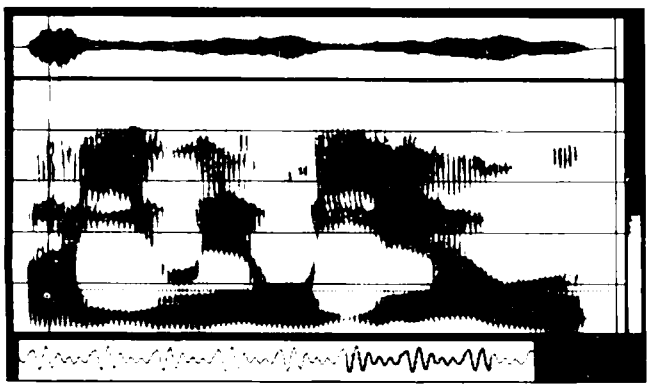
[gray scale] Set the contrast of spectrographic display.

[section] Section mode On/Off.
(In the section mode, a spectrum at a specified moment is displayed)

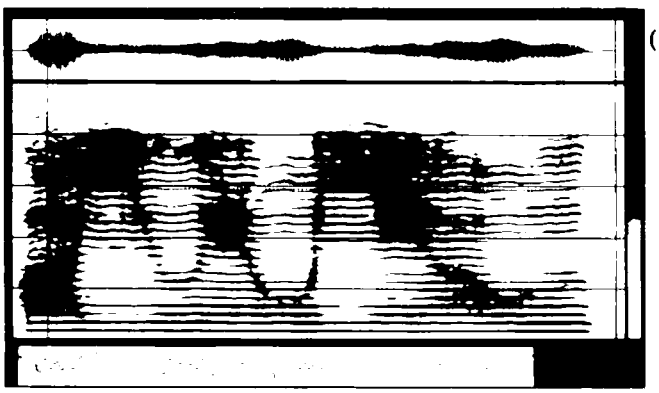
[LPC] LPC-section mode On/Off.
(In the section mode, an LPC spectrum at a specified moment is displayed)

[F1F2F3] Display the numerical values of the averaged formant frequencies of F_1, F_2 and F_3 during a specified portion.

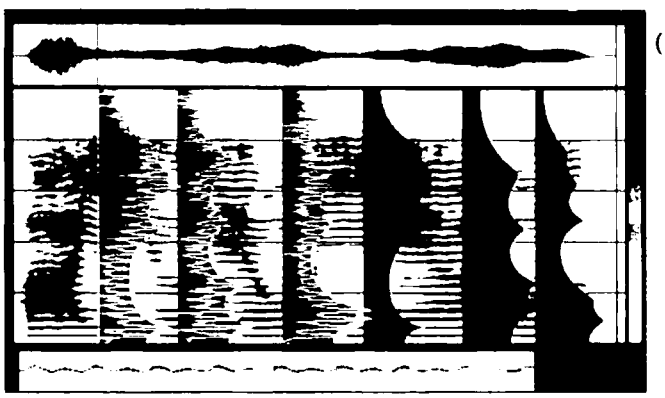
[return] Return to the speech I/O mode.



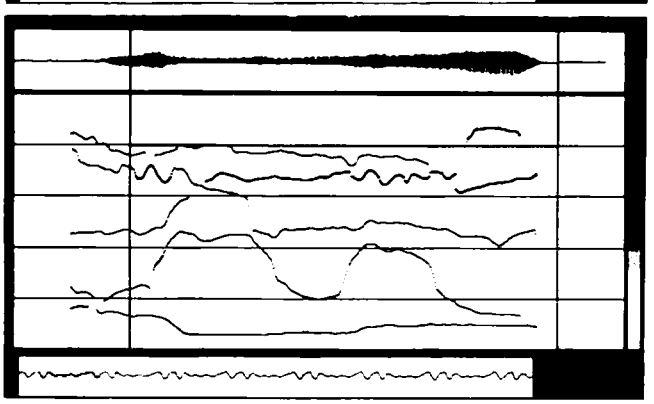
(a) Wide-band spectrogram



(b) Narrow-band spectrogram



(c) Spectrum display in the section mode



(d) Formant frequency curve display

Fig. 4
Four types of spectrogram.

2.4 Pitch analysis program

This mode performs a pitch extraction on the speech wave and displays the pitch curve. First, the program goes into the speech I/O mode, as in the spectrum analysis. The speech signal is sampled through an A/D converter, and the part of the speech signal for the pitch analysis is selected. Then, the pitch analysis is performed by executing the "pitch" command, and the time function of the pitch frequency is displayed on the monitor screen (Fig. 5).

Pitch extraction is performed through an autocorrelation method based on the 3-level clipped speech signal[3)4)]. The analysis window is 45ms, and the frame-shift is 10ms. The details of the algorithm will be described later. The numerical value of the pitch frequency at a selected point on the pitch curve can be read out by using a mouse cursor.

The following is a list of the commands in the pitch analysis mode.

[A/D]	Sample the speech signal at 10kHz and display the speech envelope.
[D/A]	Output a specified portion of the speech signal through a D/A.
[save]	Save a specified portion of the speech signal in a disk file.
[load]	Load speech data from a disk file and display the speech envelope wave.
[pitch]	Execute a pitch analysis and display the pitch curve.
[end]	Return to opening-menu.
[save]	Save the numerical values of the pitch frequencies the powers of speech signal in a disk file.
[copy]	Produce a hard copy of the screen.
[smooth]	Perform smoothing of a pitch curve with a 3-point average.
[F0 value]	Display the numerical values of the minimum, maximum and averaged pitch frequencies during a specified portion.
[return]	Return to the speech I/O mode.

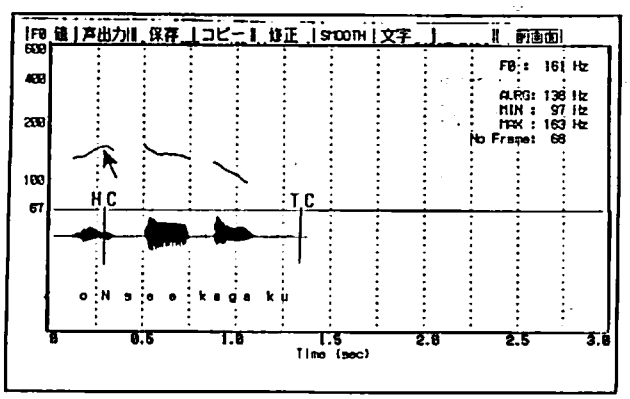


Fig. 5
Pitch curve on a monitor screen in the pitch analysis mode.

3. Applications to pronunciation training

3.1 Japanese pitch accent training system

Due to the adoption of the high-speed DSP, the present speech analysis system is capable of real time pitch extraction. This real time capability was applied to the construction of a pilot system for training of the Japanese word (pitch) accent.

A set of training words and sentences produced by a teacher are stored in disk files. When the user selects a training sample, its speech envelope and pitch contour are displayed on the monitor screen. An example of the display on a monitor screen is shown in Fig. 6. The pitch curves are plotted to indicate the intensity of the speech signal by the thickness of the curve. Then, the user inputs his own speech through an A/D converter. Pitch extraction is performed in real time, and the envelope and the pitch contour of the user's speech are displayed together with those of the teacher's speech. In the superimposed display, it is possible to perform a normalization of the duration of the utterance and the average pitch frequency before the data display. The user can repeat the input of his repeated attempts to reduce the difference between the teacher's speech pattern and his own.

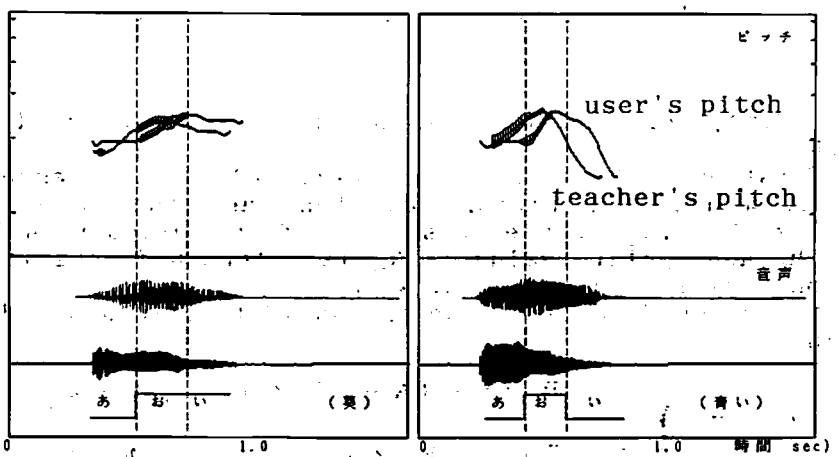


Fig. 6 Pitch pattern and speech envelope of sample words on the monitor screen of the Japanese pitch accent training system.

In most cases, training is based on a comparison of the speech patterns of minimal pairs such as a⌈me (low-high) and a⌊me (high-low). Thus, the display screen is split into two halves (left and right) to allow the simultaneous display of the speech pattern for a pair of words. In order to make possible a direct comparison of the pitch pattern of the pair of training words, the superimposed display of the pitch patterns in the left screen and the right screen can be produced.

Pitch extraction is performed based on the improved algorithm for autocorrelation pitch detection developed by Gao and Kasuya[3)4)]. The details of the algorithm are described below. First, an average absolute amplitude of N speech samples within a window of the current analysis frame is computed, and a 3-level clipping is performed on the sampled data {x(n);n=0,...,N-1}. The clipping level CL is set by Eq.(1).

$$\begin{aligned} M_1 &= \max [\,|X(0)|, \cdots\cdots\cdots\cdots, |X(N/3-1)|\,] \\ M_2 &= \max [\,|X(N/3)|, \cdots\cdots\cdots, |X(2N/3-1)|\,] \\ M_3 &= \max [\,|X(2N/3)|, \cdots\cdots\cdots\cdots, |X(N-1)|\,] \\ CL &= \min [M_1, M_2, M_3] \end{aligned} \qquad (1)$$

Next, the multiple autocorrelation function over the clipped data, defined as follows, is computed.

$$R(k) = \begin{cases} \sum_{m=0}^{N-1-k} X(m) \cdot X(m+k) & , 0 \leqq k \leqq k_b \\ \sum_{m=0}^{N-k_b-1} X(m) \cdot X(m+k) & , k_b < k \leqq 2k_e \end{cases} \qquad (2)$$

Here, {X(m)} represents a sequence of 3-level clipped data, and k_e=1/3N.

The first part of the function {R(k);k=0,...,k_b} represents the short term autocorrelation function, and the second part of {R(k);k=k_b+1,...,K_e} represents a modified autocorrelation function. The appropriate value of k_b is determined empirically. The function R(k) is continuous at k=k_b.

For the pitch detection, two threshold functions are defined as follows.

$$T_1(k) = \begin{cases} a_1 \cdot R(0) \cdot (N-k)/N & , k < k_e \\ a_1 \cdot R(0) \cdot k_e/N & , k \geqq k_e \end{cases} \qquad (3)$$

$$T_2(k) = \begin{cases} a_2 \cdot R(0) \cdot (N-k)/N & , k < k_e \\ a_2 \cdot R(0) \cdot k_e/N & , k \geqq k_e \end{cases} \qquad (4)$$

Here, a_1 and a_2 are constant values, and a_2<a_1. The search range of the pitch period is determined based on pitch values in past frames. (This criterion will be explained later.) Using the delay value k_p, which gives the largest peak value in the search range, The pitch period P in the current frame is determined as follows.

$$P = \begin{cases} k_p & , R(k_p) \geqq T_1(k_p) \\ 0 & , R(k_p) \leqq T_2(k_p) \end{cases} \quad (5)$$

Here, P=0 indicates that the current frame is an unvoiced segment. In case of $T_1(k_p)>R(k_p)>T_2(k_p)$, we determine if there exists any significant peak around $2k_p$. Let k_{pp} to the delay value which gives the peak value in the range $2k_p+k_p/b_1$, then the pitch period is determined as follows.

$$P = \begin{cases} k_p & , \quad |2 \cdot k_p - k_{pp}| \leqq k_{pp}/b_2 + 2 \\ & \quad \text{and} \quad R(k_{pp}) \geqq T_2 k(k_{pp}) \\ 0 & , \text{otherwise} \end{cases} \quad (6)$$

Here, b_1 and b_2 are constant parameters.

If the average absolute amplitude of the present analysis window is less than the background noise level, P is determined as P=-1. P=-1 indicates that the current frame is a silent segment.

The search range of the pitch period $k_1<k<k_2$ is determined by Eqs.(7) and (8).

$$k_m = \begin{cases} 2 \cdot IP_1 - IP_2 & , |IP_1 - IP_2| < c_0 \\ 0 & , |IP_1 - IP_2| \geqq c_0 \end{cases} \quad (7)$$

Here, $c_0=IP_1/8+2$.

$$\begin{cases} k_1 = c_1 k_m \\ k_2 = c_2 k_m \end{cases} \quad (8)$$

Here, IP_1 and IP_2 are the pitch periods for the past 2 frames, and c_1 and c_2 are constant parameters. If $k_m=0$ or $IP_1<=0$ or $IP_2<=0$, then the search range is set to the default value, i.e. from k_s to k_e.

The parameter values of a_1, a_2, b_1, b_2, c_1 and c_2 are also empirically determined. The actual values of the parameters in the present system are N=450, k_b=106, k_e=150, k_s=20, a_1=0.6, a_2=0.125, b_1=4, b_2=8, c_1=0.75 and c_2=1.5.

The process of real-time pitch analysis in the present program is shown in Fig. 7. When the program is started, the executable DSP program for pitch extraction is read from the disk and loaded into the external RAM(program region) of the DSP board. To input the speech signal, the user presses a function key and begins his utterance. The A/D conversion is started and

the sampled data are stored in a cyclic buffer. A CPU monitors the amplitudes of the sample data, and when the amplitude of the sampled data reach a prespecified threshold, the transfer of the DMA data from the A/D converter to memory is triggered. Continuous data of 15000 samples are stored. When the speech data for the one frame analysis is sampled, the data are transferred to the external RAM(data region) of the DSP board, and the DSP pitch extraction program is initiated by the CPU through an I/O port. Then, the CPU plots the pitch data (logarithmic scale) and the speech envelope data of the previous frame. Then, the CPU waits for the done-flag which indicates the completion of the DSP program. When the flag is set, the CPU resets the DSP and saves the computed values. The monitoring of the done-flag and the resetting of the DSP are performed through an I/O port. When the speech samples for the next frame are all sampled, the CPU again transfers the data from main memory to the external RAM of the DSP board and starts the DSP program for the next frame's pitch extraction. The frame-shift of analysis is 10ms. Pitch frequencies of 1.5-second periods can be analyzed in about 1.9seconds.

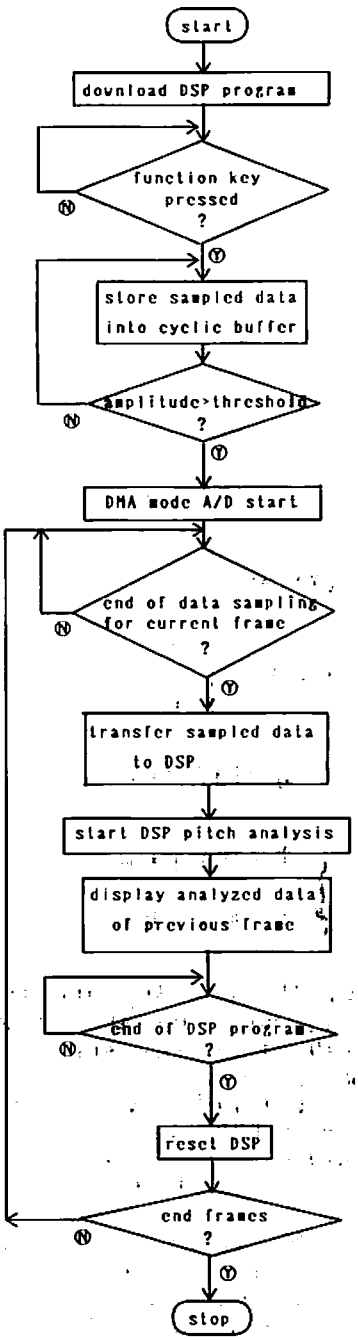


Fig. 7 Flow-chart of the real-time pitch analysis program.

The following operations can be performed with the function keys on the key-board.

f.1 / f.6	Output the teacher's utterance through a D/A.
f.2 / f.7	Output the user's utterance through a D/A.
f.3 / f.8	Display the list of sample words.
f.4 / f.9	Re-display the normalized pattern.
f.5 / f.10	Execute a pitch extraction of the user's utterance and display the result.
→ / ←	Superimpose the teacher's patterns on the left screen or on the right screen.

3.2 English intonation training system

A similar system was developed to train English intonation. The primary objectives of the training are sentence intonation patterns and sentence rhythmic patterns (durational patterns). In order to display a speech pattern of longer duration, the screen is split vertically (top and bottom) in this system (Fig. 8). This also makes it possible to compare the rhythmic (durational) patterns of a pair of sentences.

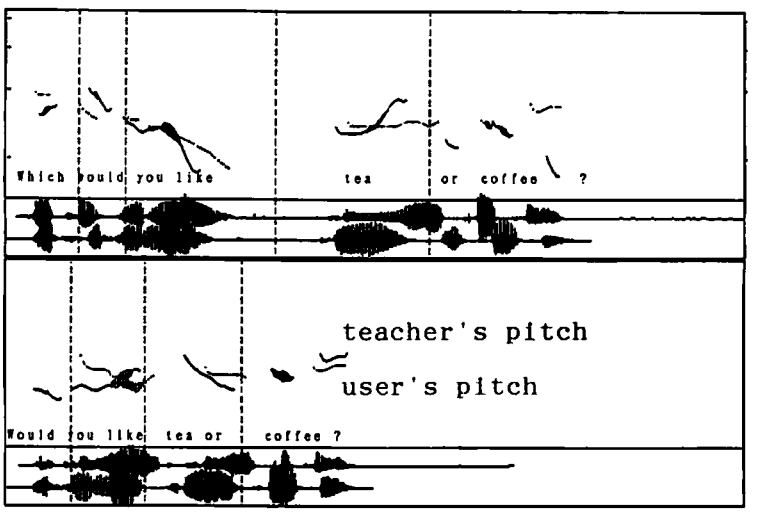


Fig. 8 Intonation pattern in the English intonation training system.

3.3 Vowel articulation training system

Ultrasonic monitoring of the tongue image is very useful for articulation training in speech disorders[5]. It is expected that the monitoring of the formant characteristics of output speech simultaneously with the tongue image will be highly valuable for speech training. For this purpose, formant frequencies are extracted by the speech analysis system, and the F_1-F_2

pattern is displayed in real time on the monitor screen of a personal computer. This display is superimposed on an ultrasound image, so that the subject can get simultaneous feed-back of the tongue image and the formant pattern.

Formant frequencies are estimated by the following method. The sampled speech data is first pre-processed for pre-emphasis and Hamming windowing. The window length is 51.2 ms. The linear predictor coefficients are derived through an LPC anlysis (Durbin-levinson-Itakura method)[1], and the LPC spectrum envelope is calculated by the FFT of the linear predictor coefficients. The formant frequencies are determined by the peaks in the spectrum envelope. The major disadvantage of the peak-picking metod is that closely spaced formants may not be extracted from the spectrum. To avoid this problem, an off-axis spectral enhancement procedure[2] is adopted. The new linear predictor coefficients $\tilde{\alpha}_i$, modified by the off-axis spectral enhancement, are defined as follows.

$$\tilde{\alpha}_i = \alpha_i \cdot \{ \exp(-\pi \cdot B_0 \cdot T) \}^i \qquad (9)$$

for (i=1,2,...,p)

Here, P is an order, and B_0 is the bandwidth reduction. In this system, p=12 and B_0=100Hz.

The process of the real-time formant frequency analysis in the present program is basically the same as in Fig. 7, although the executable DSP program which is downloaded into the external RAM of the DSP board is a program for formant analysis. Formant analysis can be performed at a rate of about 6 frames per second. An example of display screen is shown in Fig. 9.

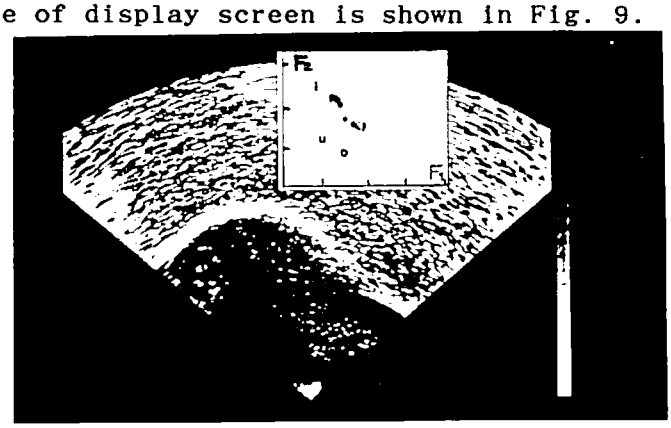


Fig. 9 CRT display superimposed the ultrasonic tongue image and formant frequencies (F_1-F_2 patterns).

Acknowledgements

The authors would like to thank Prof. Hideki Kasuya and Mr. Atsuyoshi Kawaura of Utsunomiya Univercity for their advice on an improved algorithm for autocorrelation pitch detection.

References

1) Itakura, F., and S. Saito: Speech Analysis-Synthesis System based on the Partial Autocorrelation Coefficient, Reports of the 1969 Autumn Meeting of the Acoust.Soc. Japan,199-200 (in Japanese).
2) Markel, J.D., and A.H. Gray:Linear Prediction of Speech, Springer-Verlag Berlin Heidelberg New York, 161-163, 1976.
3) Gao, X., Y. Kikuchi and H. Kasuya: An Improved Algorithm of Autocorrelation Pitch Detection, The Trans. of IECE of Japan, Vol.E67, No.5, 291-292, 1984.
4) Gao, X., Y. Kikuchi and H. Kasuya: An Improved Autocorrelation Pitch Detector and Its Evaluation with Chinese, Report of the 1985 Spring Meeting of the Acoust. Soc. Japan, 157-158 (in Japanese).
5) Masaike, H., A. Nakagawa, S. Ohno, H. Yamagami, K. Fukuyama, T. Hoshi, S. Saitoh, Y. Shimizu, S. Matsuki and J. Imai: Trial Production of Pronunciation and Speech Trainer by making use of Tongue Monitor, Report of the 1989 Spring Meeting of the Acoust. Soc. Japan, 305-306 (in Japanese).