The System Implementation of I-phone Hardware by Using Low Bit Rate Speech Coding

Ruei-Xi Chen, Mei-Juan Chen, Liang-Gee Chen, and Tsung-Han Tsai

Department of Electrical Engineering
National Taiwan University, Taipei, Taiwan, R.O.C.

Abstract - This paper presents a system implementation for Internet-phone communication with real-time speech coding schemes. A low-cost speech processing coprocessor is embedded. A CPLD device is used to implement the interface between the host processor and the coprocessor via conventional parallel port. At the headphone interface, there are a 16-bits PCM CODEC and an audio amplifier with acoustic echo cancellation features employed. The system consists of a mixed implementation of software and hardware. The experimental coding rate is 8.5kbps. In such rate, a 14.4kbps or higher speed modem can conform to offer full-duplex speech for the applications such as digital simultaneous voice data (DSVD).

1. Introduction

The development of speech coding techniques [1] has been an active research for more than fifty years. It has become an essential feature of daily telephone system operations. Recently, the LPC (Linear Prediction Coding) based algorithms [3] have been successfully promoted in low bit rate speech coders [2][4]. The popular applications of those coders are digital communication systems, cellular telephony, mobile radio, and secure voice systems, etc. To implement such a system, it often requires a specific architecture of Digital Signal Processor (DSP) to reduce the system cost and power dissipation. The special DSP coprocessors which code speech processing algorithms including G.723 standard [14] have aimed at this target. In the past, only software solutions are presented to realize the Internet phone (I-phone) [16]. In this paper, we propose the architecture to implement a low cost I-phone system via the conventional parallel port of desktop or portable computers. With such architecture, the host processor only acts the role of data manager. Thus, speech information can now be exchanged compatibly between different applications. For examples, it can be used for Net-games' meeting, Video-conferencing, or other multimedia applications.

2. System Configuration

As shown in Fig.1, the I-phone system consists of three major parts including the DSP coprocessor, the interface to the parallel port, and the headphone peripheral devices. The printer port is usually the most convenient parallel port of desktop or portable computers. This I-phone implementation is therefore intended to be suitable for such communication port. Unfortunately, the traditional parallel port was not so friendly designed for this purpose [5]. To design an efficient interface using printer port and avoid the collision from the printer are challenging works. An intelligent interface is proposed to solve the problem.

0-7803-3806-5/97/$10.00 ©1997 IEEE

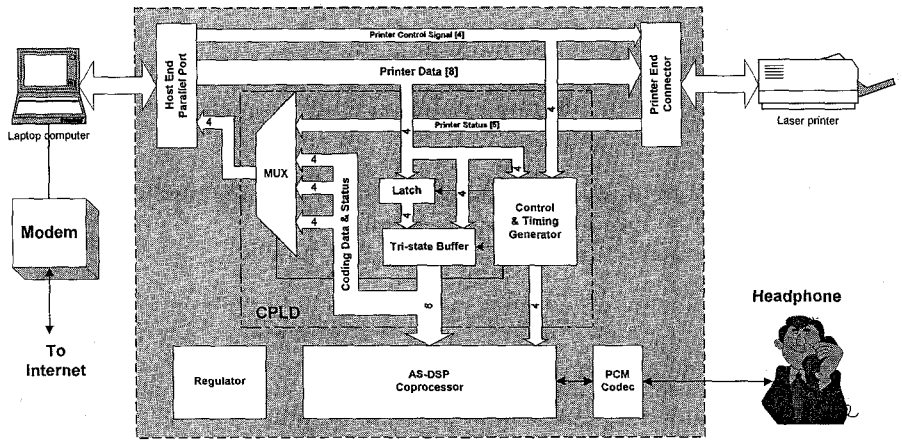


Fig.1 The block diagram of the proposed I-phone system.

The DSP coprocessor used here is a low-cost, low-power, and low complexity (about 9.5 MIPS) Application Specific Digital Signal Processor (AS-DSP). It can perform full-duplex speech compression and de-compression functions. The coding rate of 8.5kbps is selected to adapt the full-duplex speech coding to a 14.4kbps or higher speed modem.

The headphone interface works in the PCM-mode. A 16-bits μ-law/linear CODEC is introduced to provide the raw (un-compressed) speech data to the AS-DSP coprocessor, and to receive the de-compressed speech data from the same chip. Signal amplifiers are necessary to match conventional headphone set.

Another efforts are made to focus on the software enhancement, such as the device drivers for pumping the coding data through network more smoothly, and for adapting the near-end automatic echo cancellation (AEC) features.

3. Low Bit Rate Speech Coding

Typical speech signals as shown in Fig. 2 include three parts: the silence, the voiced speech, and the unvoiced speech. To code the speech signals at low bit rate, the coder must extract the features of these signals efficiently.

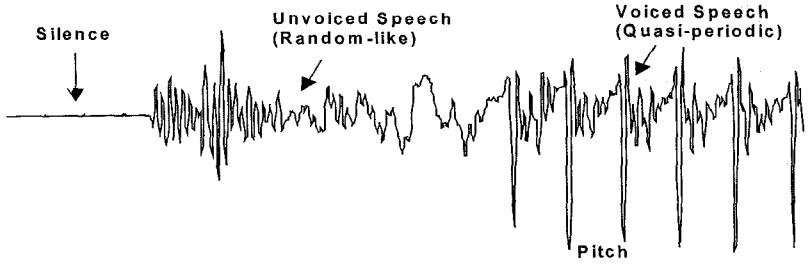


Fig. 2 Speech signal

There are three categories of speech coders, called waveform coders, vocoders, and hybrid coders. In the vocoders, often used for speech signals, the source system (speech production) model was taken account for the sake of coding efficiency. This model assumes that speech is produced by exciting a linear system, the vocal tract, with a series of pulse(voiced) or noise(unvoiced) signals. The waveform coders provide very high quality speech at medium and high bit rates (above 32kbps), but they are not suitable for coding speech at low bit rates. Combine the advantages of these former two coders, the hybrid coders was proposed by introducing the mixed excitation signals to LPC model, and have shown a high quality at low bit rates (below 8kbps).

The standards of digital representation for speech signals in waveform coders are Pulse Code Modulation (PCM) at 64kbps and Adaptive Differential PCM (ADPCM) at 32kbps (G.721). The 64kbps PCM signal is considered as "uncompressed" speech data and often used as a reference for compression. Hybrid coders are mainly divided into two sub-categories: frequency domain and time domain analysis. The latter is becomes the main stream of speech coders, which models speech production by exciting a linear time-varying filter with a periodic pulse-train for voiced speech or a random noise source for unvoiced speech. The different treatment in excitation signals reclassifies the coders into several types. The well known coders are the Adaptive Predictive Coding (APC) [9], the Residual Excited Linear Predictive Coding (RELP) [11], the Multi-pulse LPC (MPLPC) [10], the Regular-Pulse Excited LPC (RPELPC) which was chosen for the Pan-European Digital Cellular Mobile Radio System (GSM), and the Code-Excited LPC (CELP) [13].

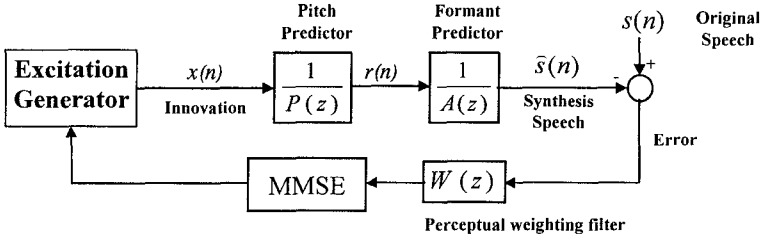


Fig. 3 General block diagram of Analysis-by-Synthesis LPC coder

The generic block diagram of hybrid coders is shown in Fig. 3 by using Analysis-by-Synthesis technologies. First, the LPC filter *1/A(z)*, also referenced as Short Term Predictor (STP), predicts the feature of vocal tract, which defines the formants structure (spectral envelope) on speech spectrum. Then, the pitch synthesis filter *1/P(z)*, also named as Long Term Predictor (LTP), predicts the feature of vocal chord, which specifies the fine harmonic structure on speech spectrum. Finally, the excitation signal *x(n)*, also called the innovation, represents the best sequences that resembles the residuals of STP and LTP. The optimal excitation signal can be obtained by using the error minimization criteria. The most popular criterion is known as Minimum Mean Square Error (MMSE) and Maximum Likelihood algorithm. In addition, the perceptual weighting filter $W(z)=A(z)/A(z/\gamma)$ is introduced to de-emphasis the frequency regions corresponding to the formants. Through this de-emphasis filter, the noise that disturbs the formants can then be reduced during analysis phase. Typically, for 8kHz sampling, γ is selected around 0.8~0.9.

The inverse filter $A(z)$ is given in the FIR form:

$$A(z) = 1 - \sum_{i=1}^{P} a_i z^{-i}$$

where a_i are the P-order LPC coefficients with $a_0 = 1$. These coefficients can be obtained by auto-correlation method (Levinson-Durbin's algorithm) or covariance method. From waveform plots, the original speech and its STP residual are shown in Fig. 4(a) and Fig. 4(c). While the spectrum plot of the original speech is shown in Fig. 4(b), and its STP residual spectrum is flatten by inverse filter as illustrated in Fig. 4(d). Note the energy of the residual is reduced significantly.

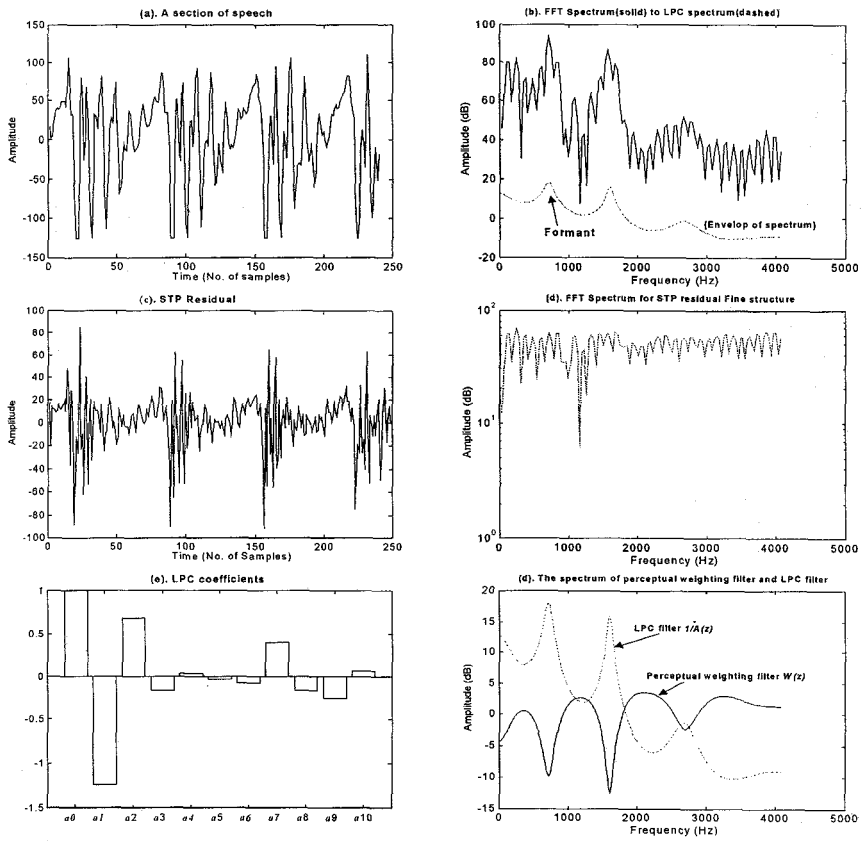


Fig. 4 Waveform plots for: (a) The original speech; (b) Its spectrum; (c) The time domain STP residual; (d) The frequency domain STP residual; (e) LPC coefficients; (f) spectrum of perceptual weighting filter (γ=0.8).

For long-term prediction, the filter $P(z)$ is defined as:

$$P(z) = 1 - \sum_{j=-M}^{M} b_i z^{-(T+i)}$$

where T is the long-term delay corresponding to pitch period, b is the pitch gain, and M is the number of filter taps. The parameters T and b can be obtained either by open loop methods or by closed loop methods[12]. Like the LPC inverse filter, $P(z)$ can extract the pitch energy from the STP residual and generate the LTP (or pitch) residual. If the filters $A(z)$ and $P(z)$ are allocated in enough orders, the LTP residual will look more like a random noise and have the cumulative probability distribution function

near the Gaussian distribution function, e.g., they have the same mean and variance [13]. Fig. 5 has shown a sample speech signal, its STP residual, and the LTP residual obtained by using one of open loop pitch predicting methods.

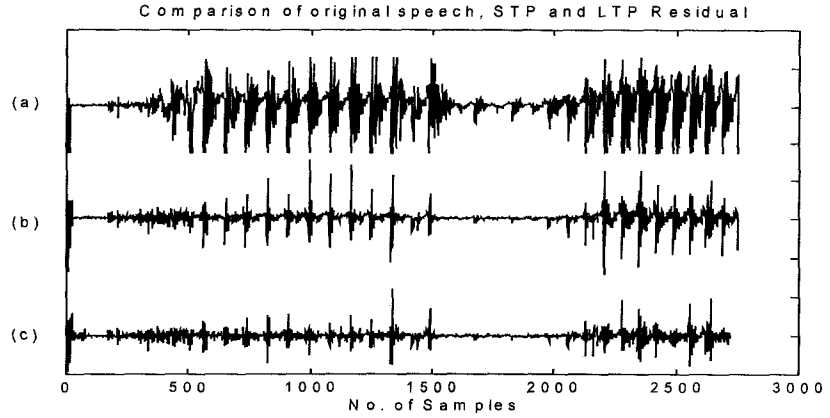


Fig. 5 Waveform plots of: (a) original speech, (b) STP residual, and (c) LTP residual

The speech processing algorithms implemented in the proposed system is the MPLPC strategies. It introduces a quantization process to approximate the LTP residual with a vector defined as:

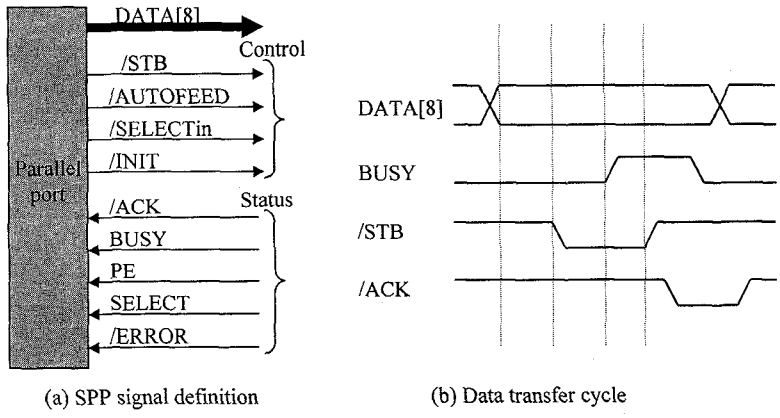
$$v(n) = G\sum_{k=0}^{M-1}\alpha_k \delta[n - p_k]$$

where G is the gain factor, $\delta[n]$ is a *Dirac* function, α_k and p_k are the sign and the pulse position of *Dirac* functions, and M is the number of pulses. The dual rate (6.3kbps/5.3kbps) version of this algorithm has recommended as the G.723 standards by ITU-T [14]. Furthermore, several other algorithms for low rate coders have already been adopted in national and international telephony standards. They are, Government FS-1015 LPC-10e at 2.4-kb/s, Government FS-1016 CELP at 4.8-kb/s, CCITT G.728 LD-CELP at 16-kb/s, and TIA/EIA IS-54 VSELP at 7.9-kb/s. Different algorithms often made the quality, efficiency, complexity, robustness, and latency trade off. The performance and quality of selected coders are listed in Table 1 [15].

Algorithm	Data Rate (kb/s)	Quality (MOS)	Delay (ms)	Prog. mem. (kB)	Data mem. (kB)	Complexity (MIPS)
FS1016	4.8/7.2	**3.1/3.5**	105/70	7.8	1.1	**19**
FR-GSM	13	**3.8**	60	5.8	0.9	**3.5**
HR-GSM	5.6	**3.8**	60	14	3.5	**30**
G.711	48/56/64	**4.3**	0	0.05	0.02	**0.5**
G.721	32	**4.1**	0	2.0	0.10	**15.9**
G.722	48/56/64	**4.5**	0	1.1	0.17	**9.5**
TrueSpeech 8.5	8.5	**3.74**	90	4~6.5	1~2	**9.5**
TrueSpeech/G.723	5.3/6.3	**4.0**	97.5	8.5	4	**25**
G.726	16/24/32/64	**4.2**	0	2.0	0.10	**15.9**
G.728	16	**4.1**	2	6.4	2.0	**37.5**
G.729	8	**4.1**	35	12.6	3.2	**34.0**

Table 1 Performance and complexity list for various coding standards.

4. System Specification



(a) SPP signal definition (b) Data transfer cycle

Fig. 6 The compatible mode of parallel port

The new IEEE 1284 standard for parallel port was released in 1994. This standard defines 5 modes of data transfer. They are Compatibility Mode, Nibble Mode, Byte Mode, Enhanced Parallel Port (EPP), and Extended Capability Port (ECP). Although the EPP/ECP mode announced a higher performance for bi-directional data transfer, many parallel ports can only implement a bi-directional link by using the Compatible and Nibble modes. For this reason, the design of I-phone system was first considered on this compatibility. Fig. 6 shows the specification of compatible mode for parallel port. Three groups of signals have been defined as control signals, status signals, and the data signals. Generally, the handshaking signals used for data transfer are strobe (/STB), BUSY, and acknowledge (/ACK) signals. The eight DATA signals are only forward directional, it must use the status line to transfer data in reverse direction. But this will make the data-width be limited to 4-bits. To solve these problems, the control and status signal timing should be modified. A CPLD chip employed for accommodating this feature and its architecture is described in next section.

The essential features of the speech coprocessor[6] includes: real-time full-duplex speech compression and decompression, concurrent Acoustic Echo Cancellation (AEC), automatic gain control (AGC), DTMF tone generation, 480 bytes frame buffer, 8/16-bit host interface, and provides synchronous μ-law or 8/16-bit linear PCM CODEC interface. The coding rate is 8.5kHz. If the connecting CODEC provides a 16-bit speech data sampling at 8kHz, then the compression ratio of 15:1 can be achieved.

In the headphone module, a temperature enhanced serial interface CODEC is employed. This circuit consists of μ-law and A-law monolithic PCM CODEC/filters utilizing the 16-bit A/D and D/A conversion architecture shown in Fig. 7

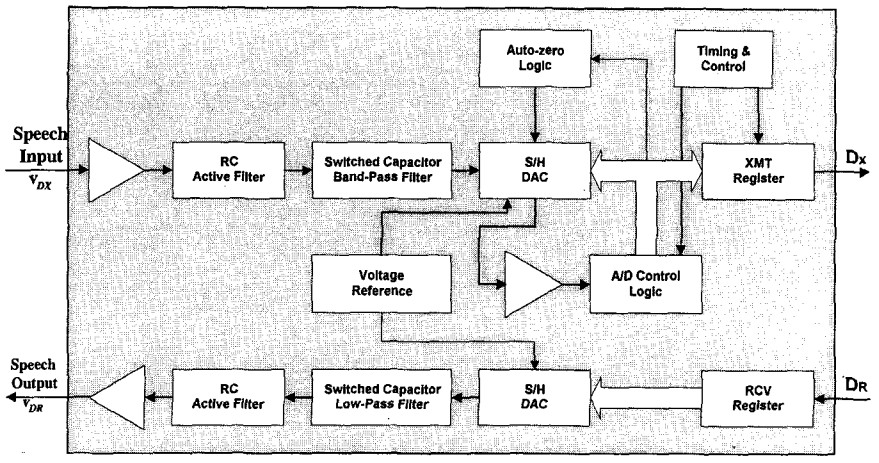


Fig. 7 Block diagram of PCM CODEC chip

The flow diagram of the whole system with software drivers is depicted in Fig. 8. In this diagram, the rectangular blocks represent the hardware devices; the circular items mean the software modules; the "U" pools store the processing data and system messages; and the arrow lines indicate the flow of control and data streams.

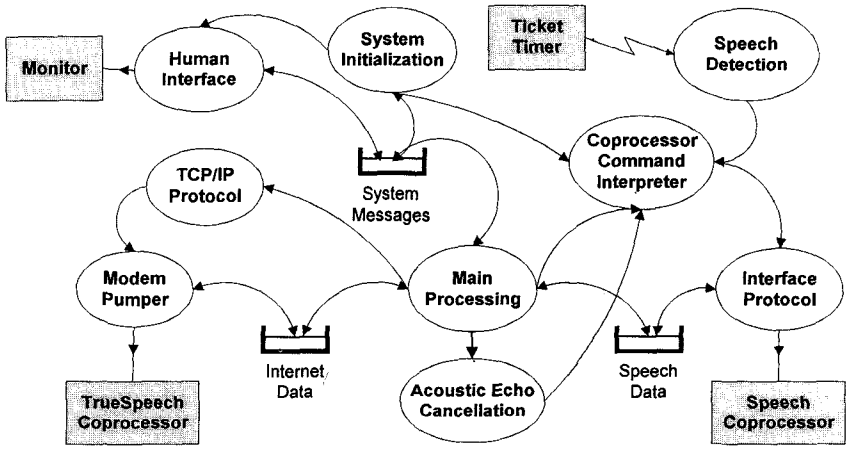


Fig. 8 System data-flow diagram

5. CPLD Design

The CPLD introduced in this system acts as a role of interface converter. It connects the host parallel port to speech coprocessor, and preserves the printer functions. In this feature, the speech coprocessor may co-exist with the printer. The CPLD always monitors the port messages and dynamically switches the bus connection to the correct device. As shown in Fig. 9, the CPLD accommodates a check logic circuit, a command register, a data register, a timing generator, and a multiplexer in one chip.

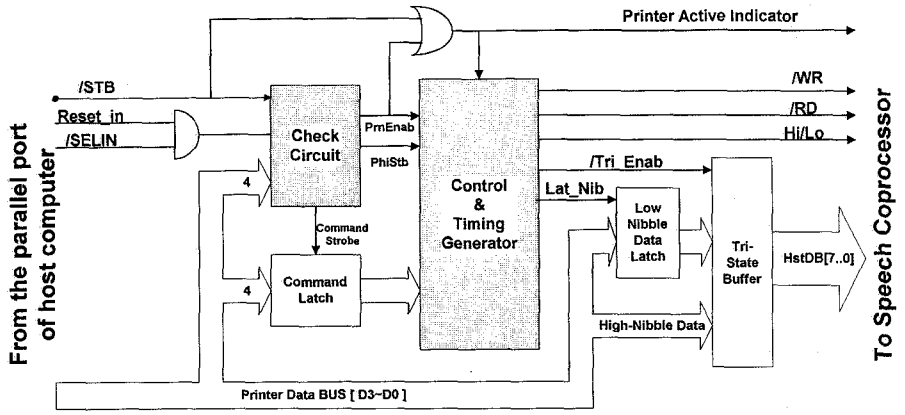


Fig. 9 CPLD internal structure: command and control part

The check circuit is a simple finite state machine and has implemented by the Hardware Description Language shown below:

```
SUBDESIGN prn_chks
(
    /Preset, /STB, d[3..0]      : INPUT;
    Prn_Enab                    : OUTPUT;
)
VARIABLE
    checks : MACHINE            % Create state machine with bits: %
        OF BITS (q1,q0)         % q1 & q0 as outputs of register   %
        WITH STATES (
            S0 = B"00",                 % initial state %
            S1 = B"01",                 % first step %
            S2 = B"10",                 % second step %
            S3 = B"11");                % final state %
BEGIN
    checks.clk   = /STB;        % input pin "clk" connects to state machine clk    %
    checks.reset = !/Preset;    % input pin "reset" connects to state machine Preset  %

    CASE checks IS
        WHEN S0 =>
            Prn_Enab = VCC;
            IF ( d[] == B"0011" ) THEN checks = S1;
            ELSE checks = S0;
            END IF;
        WHEN S1=>
            Prn_Enab = VCC;
            IF ( d[] == B"1101" ) THEN checks = S2;
            ELSE checks = S0;
            END IF;
        WHEN S2=>
            Prn_Enab = VCC;
            IF ( d[] == B"1010" ) THEN checks = S3;
            ELSE checks = S0;
            END IF;
        WHEN S3=>
            Prn_Enab = GND;
            checks = S3;
    END CASE;
END;
```

The condition of state transition is determined by a code sequence presented on the data bus of the parallel port. In this example, only the sequence of "3,D,A," is the legal sequence. This legal sequence will enable access interface of the coprocessor's and disable the printer. The result is shown in Fig. 10. [Note that the time scale is μs]

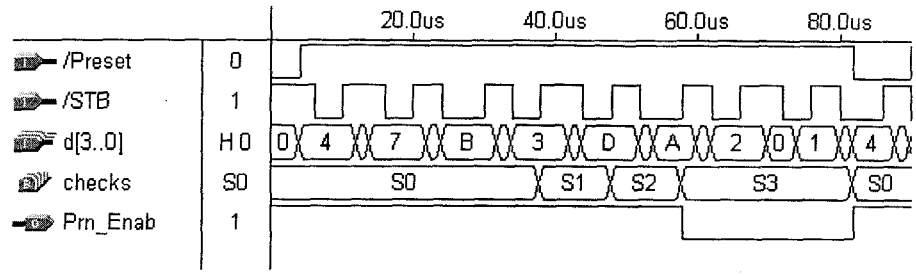


Fig. 10 The enable sequence of speech coprocessor

The timing generator is composed of the combinational logic circuits and the delay line. These combinational logic circuits are triggered by the control command presented in the data line. More than two commands of writing data to and reading data from the coprocessor are designed. For example, command code "1" defines pre-latch low nibble; command code "2" defines write whole byte, etc. The timing of this feature is shown in Fig. 11.

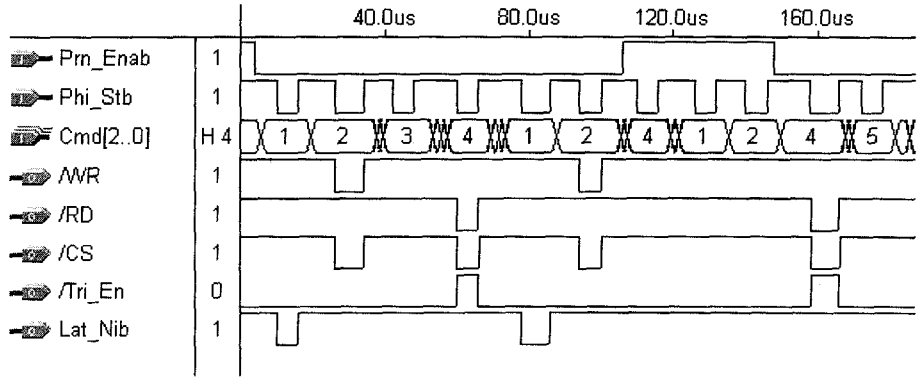


Fig. 11 Command and timing for coprocessor interface generated by control circuits

6. Performance

The overall system and its human interface are illustrated in Fig. 12. It shows the outlook of the hardware structure and the friendly software control panel. The performance of the proposed system is demonstrated by the waveform comparison of the original signal and the reconstructed signals. Fig. 13(a) shows the original signal, and (b) is the reconstructed signal produced by one of CELP coding algorithm, while (c) is the reconstructed signal produced by this I-phone system with acoustic echo cancellation (AEC) feature enabled. From this comparison, it is evident that our I-phone system has similar profiles. It has achieved good subjective quality proved by a group of users.

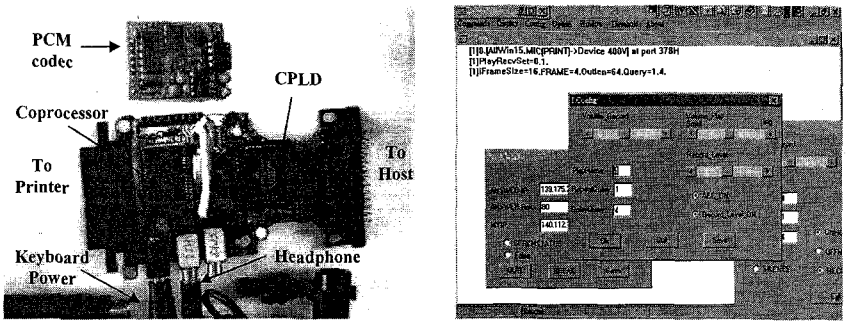


Fig. 12 System prototype and human interface

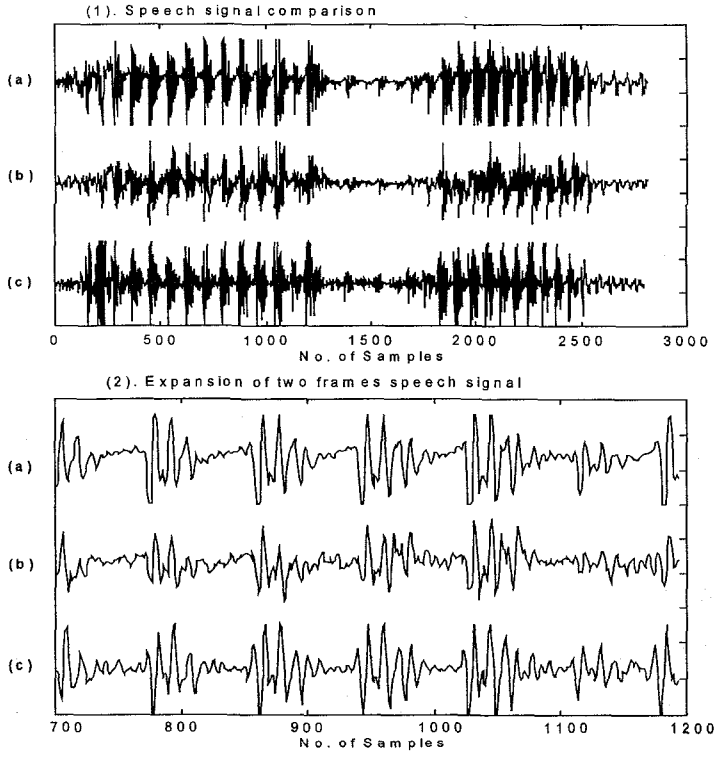


Fig. 13 Comparison of speech signals. (a) The original speech signal. (b) The reconstructed speech signal by using one of CELP algorithms. (c). The reconstructed speech signal by using this I-phone system with AEC feature enabled.

7. Conclusion

We proposed an I-phone system implementation by using speech coding algorithm, specific DSP coprocessor, high level hardware design language and CPLD device. The CPLD device is adopted as an interface between coprocessor and host computer via conventional parallel port. The design consideration for this CPLD has been described. The performance of the whole system is demonstrated. It is shown that the

proposed design can produced good quality for real-time I-phone applications. In the future, this portable system can be used to enhance many applications especially for Net-meeting, video conferencing, etc.

Reference

[1] Andreas S. Spanias, "Speech Coding: A Tutorial Review," *Proceeding of the IEEE*, vol. 82, no. 10, Oct. 1994.

[2] A. M. Kondoz, "Digital Speech – Coding for Low Bit Rate Communications Systems," *John Wiley & Sons Ltd.*, 1994

[3] John. D. Markel, and Augustine. H. Gray, Jr., "Linear Prediction of Speech," *New York*, 1980.

[4] Bishnu. S. Atal, Vladimir Cuperman, and Allen Gersho, "Speech and Audio Coding for Wireless and Network Applications," *Kluwer Academic Publisher*, 1993.

[5] Jay Lowe and Don Schuman, "Parallel Port Information System," *Parallel Technologies, Inc.* version 1.45, Oct. 1994.

[6] ____, "TrueSpeech® G.723/DSVD Co-Processor," *DSP Group, Inc*.

[7] B. S. Atal, "Predictive coding of speech at low bit rates," *IEEE Trans. Commun.*, vol. COM-30, no. 4, p. 600, Apr. 1982.

[8] J. Campbell and T. E. Tremain, "Voiced/unvoiced classification of speech with applications of the U.S. Government LPC-10e algorithm," in *Proc. ICASSP-86* (Tokyo, Japan, 1986), pp. 473-476.

[9] B. S. Atal, "Predictive coding of speech at low bir rates," IEEE Trans. On Comm., April 1982, 600-614.

[10] B. S. Atal and J. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," Proc. of ICASSP, pp. 614-617, 1982.

[11] C. K. Un and D. T. Magill, "The residual-excited linear prediction vocoder with transmission rate below 9.6kbits/s," , *IEEE Trans. Commun.*, vol. COM-23, no. 12, p. 1466, Dec. 1975.

[12] R. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-37, no. 4, pp. 467-478, Apr. 1989.

[13] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High quality at very low bit rates," in Proc. ICASSP-85 (Tampa, FL, 1985), p. 937.

[14] Draft Recommendation G.723, "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 & 6.3 kbit/s," ITU-T, Oct. 1995.

[15] URL, "Speech Coding Algorithms," http://www.sasl.demon.co.uk/speech.htm, Sep. 1996.

[16] "The Top VON (voice on network) products of 1996," http://www.von.com/1996vonvote.htm.