S11.11

ENHANCING THE PERFORMANCE OF SPEECH RECOGNITION WITH ECHO CANCELLATION

J. Picone, M.A. Johnson, W.T. Hartwell

AT&T Bell Laboratories, Naperville, IL

ABSTRACT

In most telecommunications services, there is a large probability of encountering a telephone connection with significant amounts of echo. Additive noise consisting of echoed speech severely degrades speech recognition performance. In this paper, we describe the use of echo cancellation to improve speech recognition performance over telephone channels. Echo cancellation is shown to provide an increase of 25 dB in signal to noise ratio, thereby increasing recognition performance to a level which can be attained over telephone channels with no echo. A prototype system which includes the echo canceller and an isolated word speaker independent speech recognizer has been implemented within a single AT&T WE DSP-32.

INTRODUCTION

A popular scenario for an automated telecommunications service utilizing small vocabulary, isolated word, speaker independent speech recognition technology is shown in Figure 1. In this scenario, a user is first prompted with an announcement which describes the various choices available for the service, and the appropriate responses. The user is then expected to respond within some short time interval immediately following the announcement. A major problem associated with this type of service is that a large percentage of untrained users will speak before the specified time interval begins (typically during the announcement when the desired choice is described). Similarly, experienced users would like to respond with their choice as soon as possible.

It is, therefore, desirable to allow a speech recognizer to perform recognition during the announcement, a capability known as "talk-thru". Due to the large probability of encountering a telephone connection with significant amounts of echo, maintaining good speech recognition performance during the outgoing announcement is difficult. One approach to improving the performance of current speech recognition technology in this environment is to employ echo cancellation to cancel the echoed announcement. In this paper, we describe the use of a standard echo canceller with a standard speech recognition algorithm to achieve talk-thru.

ECHO CANCELLATION

Applications of echo cancellers today are widespread, mainly because echo cancellers are available in low cost single chip packages. The simplest form of echo cancellation is the adaptive transversal filter. The properties of the adaptive transversal filter are well-understood. Several variants have been developed over the years. The particular algorithm chosen for this study is described in detail in [1,2]. Three considerations influenced the design of the echo canceller used in this study: the convergence rate of the transversal filter, the length of the correlation window used in updating filter coefficients, and the accuracy of the near-end speech detector.

The adaptation step-size controls the rate at which the adaptive filter will converge to a solution. A tradeoff usually exists between the convergence rate and the asymptotic cancellation error. A fast convergence rate can sometimes result in divergence. The convergence rate of the echo canceller averaged over several echoes collected over long distance telephone lines is shown in Figure 2. In our implementation, the echo canceller has been designed to converge within 0.5 seconds on typical echoes, and has never been observed to diverge or produce an unstable filter.

A second important consideration in the design of the adaptive transversal filter echo canceller is the correlation window length used in updating the filter coefficients. This window length is the single most important parameter in determining the computational complexity of the algorithm since it essentially determines the number of required multiply

CH2561-9/88/0000-0529 $1.00 © 1988 IEEE

adds per sample per coefficient. Figure 3 depicts convergence rate as a function of window length. While long window lengths are desirable for stationary signals such as white noise, we find echo canceller performance on echoed speech signals is best for the stochastic gradient method [2]. In the stochastic gradient method, a one sample window (125 usec window duration) is used for coefficient update. The computational advantage of this method is that filter coefficients can be updated with a single dot product.

The third component important to echo cancellation is the choice of a near-end speech detection algorithm. This is certainly the most difficult issue in echo cancellation. Speech detection, in itself, is a difficult problem in speech recognition. A standard technique to detect incoming speech in echo cancellation is to declare incoming speech when the amplitude (or energy) of the incoming speech plus echo exceeds a certain percentage of the amplitude (or energy) of the outgoing speech [2]. Unfortunately, this simple algorithm does not detect low energy portions of the speech signal, and is not optimal for speech recognition. For instance, the fricative in four is an important queue in distinguishing the word "four" from "one."

In the service scenario described above, one simple, effective solution is to merely freeze coefficient adaptation after one second. Since the convergence rate is 0.5 seconds, this typically guarantees convergence and also minimizes the computational burden of speech detection. The only weakness of this approach is that an incoming extraneous signal during the adaptation interval may result in a less than optimal cancellation filter. Speech recognition performance will degrade significantly if coefficient adaptation is allowed during incoming speech. More sophisticated speech detection algorithms have been suggested [2]. In fact, the recognition algorithm described below uses a highly sophisticated, robust speech detection algorithm. However, we find the approach of inhibiting adaptation after 1 second to be sufficient for the applications presented here.

SPEECH RECOGNITION WITH ECHO CANCELLATION

In most telecommunications applications, speaker independent speech recognition technology is most appropriate. Since performance is at a premium, small vocabulary isolated word recognition was considered. A standard recognition algorithm [3], based on the dynamic time warping paradigm and the log-likelihood distance measure, was used in this study. This algorithm is currently available in several hardware configurations, and has been shown to reasonably robust on small vocabulary, isolated word speech recognition tasks in telephone environments.

The recognition system divides the recognition problem into two steps, an energy-based endpointing which locates the endpoints of an incoming speech utterance, and a dynamic time warping pattern matching which compares the incoming utterance to a database of reference templates. A simple acceptance/rejection decision criterion is used in which the ratio of the best to the next best score for competing word candidates is tested against a threshold. Robust, high performance, speech recognition can be achieved through careful construction of the reference templates. Reference templates are built using an algorithm based on statistical clustering techniques [4]. In this study, 8 templates per vocabulary word were used.

The database used for evaluation of the impact of echo cancellation consisted of a database collected during a field trial of an automated reservations service. In this field trial, users were prompted with a 25 second announcement, and asked to choose between one of four items by speaking the words one, two, three, and four. Users from all parts of the country participated in this field trial, and thus, the database contains a reasonable sampling of long distance telephone connections. The total database consists of over 15,000 utterances. Approximately 20% of the transactions in the database contained some form of echo of the announcement. A smaller database of 3000 test tokens was constructed which had an equal distribution of vocabulary items, and an equal number of males and females.

Evaluation databases were constructed by by mixing several representatives of relatively clean echoes with the test database at varying signal to noise ratios. A set of 10 examples of clean echoes of the prompting announcement were selected. The actual distribution of types of echoes was not examined, since there were many forms of signal degradation present simultaneously in many of the transactions. Each test token was added to a randomly selected example of echo at a random location in the example echo file. A fixed signal to noise ratio (SNR) was maintained on a per file basis. Thus, performance of a recognition system can be benchmarked versus SNR, and the improvement of the echo canceller can be described in terms of the equivalent boost in SNR.

be described in terms of the equivalent boost in SNR.

A performance summary is shown in Figure 4. First, four curves are shown indicating performance versus SNR for the recognition system with no echo cancellation. There is a dip in performance at the 5 dB point. This is primarily due to the inability of the endpointer to to accurately detect endpoints for low SNRs. The energy-based endpointer used in this study relies explicitly on the change in energy relative to an estimate of the background energy level.

The next two curves indicate performance with echo cancellation for the 10 dB SNR case and the 5 dB SNR case. For SNR values of 10 dB and above, recognition performance is equivalent to that for telephone connections with no echo. At a 5 dB SNR, there is a minor degradation in performance. Below 5 dB, performance gets progressively worse. The echo canceller typically delivers an improvement of approximately 25 dB in SNR, as demonstrated in Figure 4.

AN INTEGRATED HARDWARE IMPLEMENTATION

The echo cancellation algorithm described above requires significant computational resources. The bulk of the processing in the echo canceller is spent executing two dot products. The first dot product produces an estimate of the echo, while the second dot product generates the updates to the filter coefficients. A single echo canceller today typically handles a maximum of 16 ms of delay. At a sample rate of 8 kHz, the filter is 128 taps long, and one dot product of length 128 must be executed to update the filter coefficients. One divide operation is also required along with a general assortment of moves and adds. Each of these operations are required on a per sample basis until adaptation occurs, at which time the filter adaptation dot product is omitted.

The ATT WE DSP32 was chosen to implement the echo canceller primarily because its fast floating point mulitplier. Internally the DSP32 consists of two: an integer processor capable of 4 MIPS, and a floating point processor capable of 8 MFLOPS (using an instruction cycle time of 250 nsec). Assuming no inefficiencies or I/O waits, the DSP32 is able to do the required dot product operations in 26% of real time.

There are two aspects of the implementation of the echo canceller which are somewhat unique. The first involves the DSP32's serial port. Both an outgoing and incoming speech signal must be input every 125 us, even though the DSP32 has only one serial port. The most obvious solution is to memory map a shift register (SIPO) and read it every time the serial port is read.

This solution will not work in other situations in which processing of blocks of data is necessary. The reason is that the dma feature of the DSP32 will ignore the SIPO and no interrupts exist with which to read it. An alternative is to multiplex the serial input. By supplying the codecs with sync pulses which are 180 degrees out of phase and reading the serial port at double the sampling rate, one serial port can read both codecs via dma.

A second feature of the echo canceller implementation involves the minimization of wait states. To avoid generating wait states, the DSP32 must access its high and low banks of memory alternately. The best method of avoiding wait states is to locate often used variables so that bank interleaving is maximized. In the echo canceller implementation, filter coefficients and speech samples are the most often used variables. To achieve zero wait states in both of these computations it was necessary to place one of the dot products in high memory. Since high ram already exists and the rom is located in the low bank, it was only a matter of relocating several bytes of the program at initialization time.

The echo cancellation algorithm as implemented on the DSP32 uses very little memory: 512 bytes of ROM, 254 bytes of LRAM, and 278 bytes of HRAM. Ignoring I/O times, the speed of operation per sample (125 usec) is: 83.25 usec for coefficient adaptation, and 50.90 usec after adapting. Since the DSP32 has an internal 4096 bytes of RAM and 2048 bytes of ROM, it is possible to use the less expensive 40 pin version. In this case memory mode 1 would have to be used to split the RAM between the two banks.

SUMMARY

We have presented a successful application of echo cancellation to improving speech recognition performance in telephone environments. The echo canceller was shown to improve the SNR by approximately 25 dB for relatively simple echoes. Recent advances in DSP chip technology allow both the echo canceller and a small vocabulary isolated word speech recognizer to be implemented using a single DSP chip and appropriate external memory.

REFERENCES

[1] D.L. Duttweiler and Y.S. Chen, "A Single-Chip VLSI Echo Canceller," Bell System Technical Journal, Vol. 59, No. 2, pp. 149, February 1980.

[2] D. Messerschmitt, D. Hedberg, C. Cole, A. Haoui, and P. Winship, "Digital Voice Echo Canceller With A TMS32020," Digital Signal Processing Applications With The TMS320 Family, pp. 415-437, Texas Instruments, Inc., 1986.

[3] L.R. Rabiner, S. E. Levinson, A.E. Rosenburg, and J.G. Wilpon, "speaker Independent Recognition Of Isolated Words Using Clustering Techniques," IEEE Transactions On Acoustics, Speech, and Signal Processing, Vol. ASSP-27, No. 4, August 1979.

[4] J.G. Wilpon and L.R. Rabiner, "A Modified K-Means Clustering Algorithm For Use In Speaker Independent Isolated Word Recognition," IEEE Transactions On Acoustics, Speech, and Signal Processing, Vol. ASSP-33, No. 3, June 1985.

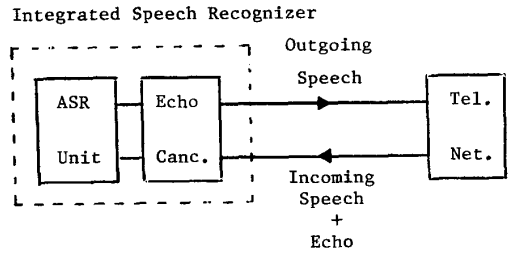

Figure 1. Speech Recognition Enhanced With Echo Cancellation

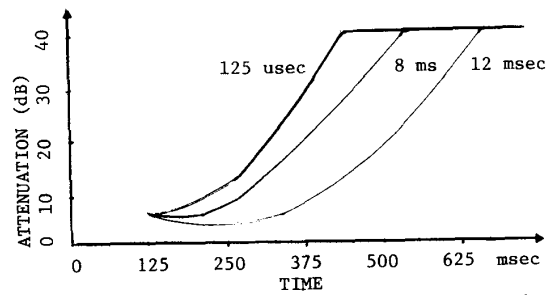

Figure 2. Convergence Rate Versus Window Duration (Beta = 2**(-10))

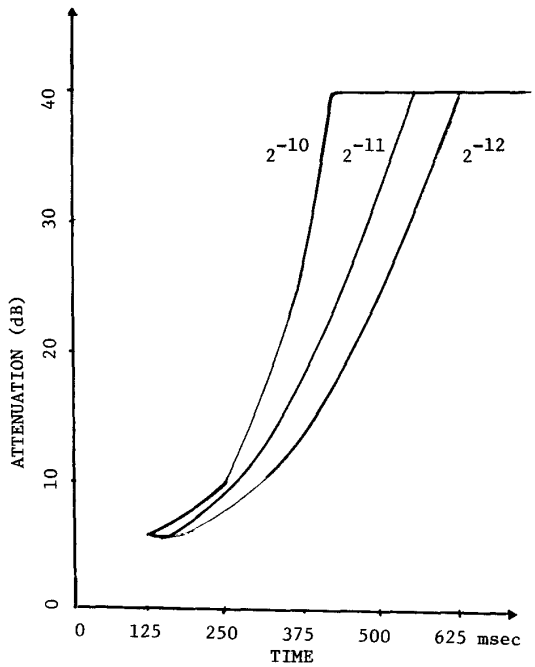

Figure 3. Convergence Rate Vs. Adaptation Speed (Window Duration = 2 msec)

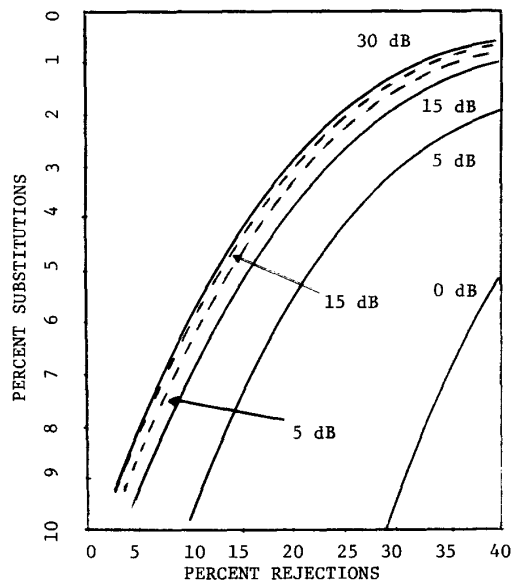

Figure 4. Performance Without Echo Canc. (solid lines) and With Echo Canc. (dashed lines).