ECHO CANCELLATION FOR EVALUATING SPEAKER IDENTIFICATION TECHNOLOGY

Aravind Ganapathiraju, Joseph Picone
Institute for Signal and Information Processing
Mississippi State University, Mississippi State, Mississippi 39762, USA
Ph (601) 325-3149 - Fax (601) 325-3149
{ganapath, picone}@isip.msstate.edu

***Abstract* - Echo cancellers using adaptive filtering techniques have traditionally found application in solving a wide variety of communications systems problems. We present here a novel application of an FIR echo canceller to be used in evaluating speaker identification technology on conversational speech data collected over the public telephone network. Various modifications to the standard LMS echo canceller are required to deal with double-talk, a time-varying background channel, and stability of the adaptive filter. Our implementation of the echo canceller has been optimized for two-way telephone conversations and has been tested extensively on the SWITCHBOARD corpus.**

INTRODUCTION

Echo cancellers using adaptive filtering techniques have enjoyed widespread success in a variety of communications systems problems. Cancelling echoes in long distance telephone conversations is by far the most direct and widely used application of echo cancellers. Cancelling echoes from speech recorded in noisy environments is another common application of echo cancellers.

Speaker identification technology (SID) has its main application in the area of security. Since this technology is still in its infancy, there is a need for benchmarking the technology on common data. Two such examples of common evaluations being conducted annually are the annual Human Language Technology workshops sponsored by DARPA [1], and the large vocabulary continuous speech recognition workshops (LVCSR) sponsored by DoD. These evaluations have been conducted on the SWITCHBOARD corpus [2], a spontaneous two-way telephone speech data corpus.

In order that we make our terminology consistent with the adaptive filtering literature, we view this two-channel data as a reference channel consisting of the far-end talker, and a test channel consisting of the originating "near-end" caller. Due to the irregularities of the analog telephone network, an echo (a filtered and delayed version of an incoming signal) of the far-end speech gets added to the near-end speech. This echo could be used by the speech recognizer to gather important cues regarding the identity of the speaker or channel, thereby making the SID task artificially easy. To eliminate this problem, we need an efficient echo cancellation technology. A block diagram of the overall system discussed in this paper is presented in Figure 1.

ALGORITHM SPECIFICS

The most commonly used algorithm for echo cancellation is the Least Mean Square Error (LMS) algorithm [3] based on a finite impulse response (FIR) filter.

The LMS Algorithm

The algorithm uses an FIR filter to model the channel which has caused the echo. If the estimate of the channel is accurate, we can pass the reference signal through this filter to produce a replica of the echo. A simple time-domain subtraction of the signals will then eliminate the echo from the speech signal. Since we do not have direct access to the channel, we must adaptively estimate the filter coefficients using a gradient-descent approach. The adaptation algorithm is based on minimizing the mean square error (MSE) between the reference and the echo-cancelled signal [3].

Let $y(i)$ represent the far-end speech, $x(i)$ the near end

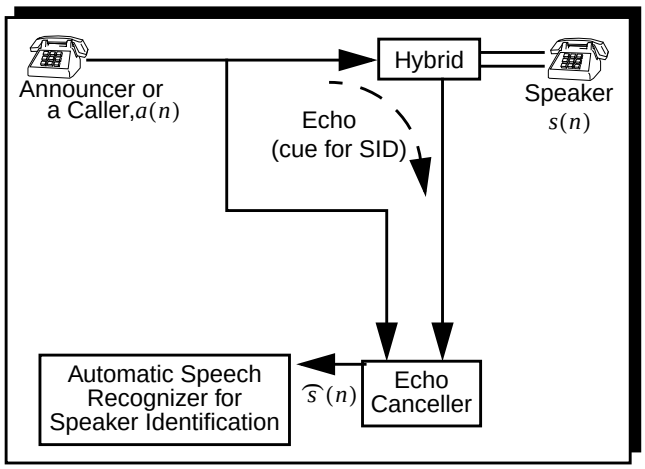

Figure 1. A demonstration of the application of echo cancellation to speaker identification technology evaluation.

speech and $r(i)$ the reflected echo. Let h_k be the discrete representation of the channel impulse response:

$$r(i) = \sum_{k=0}^{N-1} h_k \times y(i-k) \quad (1)$$

We assume the echo is generated by a linear system, and is of finite duration. An echo canceller with N filter coefficients adapts to produce a replica of $r(i)$ defined as:

$$\hat{r}(i) = \sum_{k=0}^{N-1} a_k \times y(i-k) \quad (2)$$

N, the length of the filter, is typically set to be greater than twice the time in samples corresponding to the maximum distance over which a telephone call can travel. Some knowledge of the communications media is required for this calculation (typically an electrical signal's propagation speed on coaxial cable).

Clearly, if the estimation of the echo path is accurate, the echo is cancelled perfectly. In general, since the impulse response of the echo path is unknown and may vary slowly with time, coefficient adaptation is necessary to minimize the MSE between the echo and its estimate.The equation used to estimate the coefficients is:

$$a_k(i+1) = a_k(i) + \frac{2\beta}{M} \sum_{m=0}^{M-1} e(i-m)y(i-m-k) \quad (3)$$

Here β controls the rate of adaptation of the filter coefficients, M is used for updating the coefficients in a block update mode (rather than at every sample). β is estimated from the error signal, and hence is inversely proportional to the power of y(i). It has been observed that the algorithm converges very slowly for low-power signals. To counter this problem, the loop gain β is usually normalized by the estimate of the power of y. Figure 2 presents the algorithm as applied to our conversational speech problem. For the echo canceller to perform satisfactorily in the context of a speech research database, we incorporate a number of heuristics into the algorithm.

Modifications to the LMS Algorithm

Since the error signal, $u(i)$, shown in Figure 2, contains a component of the near-end speech $x(i)$, in addition to the residual echo cancellation error, it is necessary to disable adaptation when near-end speech is present. This is achieved by using a simple speech detector, which compares the signal level of the near-end speech to the reference far-end speech. Also, it is desirable to continue declaring near-end speech for some time after the initial detection, since the speech detector fails when there are short-term noise bursts, or sudden drops in the speech energy level. We introduce a parameter, which we refer to as the hang-time, that controls the amount of time after the initial detection of speech for which we assume speech activity. This parameter is currently set to 75 ms.

We need to be very careful in declaring near-end speech in an event of double talk. If the adaptation is not disabled when this occurs, the filter diverges, and the result is the introduction of audible clicks and pops in the output speech. Hence, a conservative speech detection threshold is used. This threshold has been set to be 20% of the amplitude of the maximum sample value in the previous 32 ms of reference signal (256 samples).

Another important modification made to the basic algorithm is the choice of a residual error suppression algorithm. Due to nonlinearities in the echo path, the amount of achievable suppression by the echo canceller is limited. Typically, we drop the level of the outgoing echo cancelled signal when the suppression is not perfect. One way to implement this is to detect when the outgoing signal power falls below the reference signal by a specified amount, and to zero the output signal when this condition has been satisfied. This tends to cause choppiness in the output signal, particularly during intervals where only background channel noise is present.

A convenient solution to this problem is to vary the level of suppression depending on the ratio of power levels of the outgoing signal and the reference signal. Suppression is not required when this power ratio is greater than -24 dB (a double-talk condition) or less than -60dB (near-end silence). The suppression threshold α, was implemented as a linear function in terms of the floor and ceiling values of this ratio (all values expressed in dB).

$$\alpha = \frac{ratio - ceil}{floor - ceil} \quad (4)$$

A fourth modification to the standard algorithm is based on the important observation that the rate of adaptation is implicitly dependent on the length of the filter. This involves some judgement on the part of user in choosing the length of the filter and other related parameters of the echo canceller. Our rule of thumb is to decrease the adaptation constant by a half when we increase the number of taps by half.

The processed data consists of 10 to 30 minute conversations collected over a fixed telecommunications link, the channel characteristics tend to be quite stationary. Hence, after allowing the echo canceller to converge by rapidly adapting

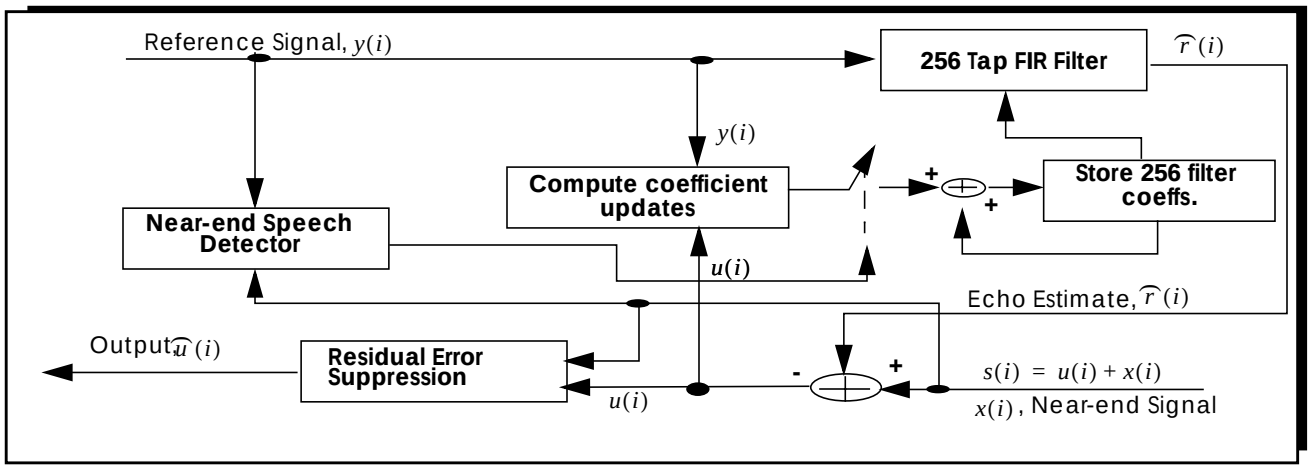

Figure 2. The basic least mean square error (LMS) echo-canceller based on a finite impulse response (FIR) filter.

over the first few seconds of data, it is not necessary to adapt any further. Hence, we exponentially reduce the rate of adaptation to a certain minimum value as we process the data. This helps prevent divergence and needless fluctuations of the filter during the adaptation process.

EXPERIMENTATION AND RESULTS

The SWITCHBOARD [2] corpus is one of more recently collected conversational speech telephone corpora. We used 40 samples of these conversations for our experiments. These samples had varying levels of echo, starting at very faint, and ranging to conditions where the echo is as loud as the near-end speech.

Most of the modifications we incorporated into our algorithm were done after listening to data processed with the basic algorithm. The performance of our echo canceller is summarized in Table 1, which shows performance as a function of the echo. We report the reduction of echo for the channel on which the echo-canceller performs the worst. Note that, when the echo is very low, 10 dB or less, echo cancellation is perfect. As the level of echo increases, we see that the performance degrades, indicating that the echo might be a complex echo, unlike the simple scaled version of the far-end speech we assume.

Echo-level	Echo-reduction
< 10dB	>40dB
10 - 20dB	17dB
20 - 30dB	22dB

Table 1. Performance of the echo-cancellation algorithm on speech data collected over long-distance telephone lines with varying levels of echo content.

CONCLUSIONS AND FUTURE WORK

We have developed an echo cancellation algorithm which can be used to preprocess two-way telephone data. Its intended application is the development of speaker identification technology. We have incorporated a number of heuristics into the algorithm to make the quality of the processed speech acceptable for speech technology development. Our results also indicate that the assumption of linearity in the production of echoes is not always true, especially in cases where the echo-level is high. Our code is presently being used extensively in the speech research community, and can be downloaded from the ISIP web site at the URL: *http://www.isip.msstate.edu/software*. It is written in C++ and is intended to be easy to use and understand.

ACKNOWLEDGEMENTS

We are grateful to Dr. George Doddington of the NSA for his careful review of many versions of the algorithm, and Mr. David Graff of the LDC for supplying us with an excellent development database.

REFERENCES

1 Pallett, D., et. al., "1995 Benchmark Tests for the ARPA Spoken Language Program", *Proc. of the APRA Spoken Language Systems Technology Workshop*, Harriman, NY, February 1996.

2. Godfrey, J., et.al., "SWITCHBOARD: Telephone Speech Corpus for Research and Development", *Proc. of ICASSP'92*, vol. 1, pp. 517-520, San Francisco, CA, March 1992.

3. Messerschmitt, D., et. al. "Digital Voice Echo Canceller With A TMS32020," *Digital Signal Processing Applications with the TMS320 Family*, pp. 415-437, Texas Instruments, Inc., 1986.