ECHO DETECTION AND DELAY ESTIMATION USING A PATTERN RECOGNTION APPROACH AND CEPSTRAL CORRELATION

Rafid A. Sukkar

Tellabs, Inc.
Naperville, IL, USA
Email: rafid.sukkar@tellabs.com

ABSTRACT

In this paper we present a method for echo detection and echo path delay estimation using a pattern recognition approach. We consider the problem of echo detection as attempting to match a speech pattern in the near-end signal to the far-end signal at a given delay. Employing features and techniques that have been successfully used in speech recognition, we define a spectral similarity function based on cepstral correlation. We show, through experimental results, that the proposed similarity function can reliably detect acoustic echoes and correctly estimate the echo path delay. Further, it is shown that the similarity function can be used in the detection of double-talk conditions. The method presented here is applicable to both electrical (hybrid) network echoes as well as to acoustic echoes.

Index Terms— Echo detection, echo control, double-talk detection

1. INTRODUCTION

The detection and suppression of acoustic echoes in telecommunication networks have become increasingly important with the proliferation of wireless networks. In non speaker-phone situations, the severity of acoustic echoes depends mainly on the design of the specific handset used during a given call. The design of the handset casing and the placement of the mouth piece relative to the earpiece play critical roles. In speaker-phone cases, the placement of the speaker and microphone as well as the room acoustics are the major factors in the level of acoustic echoes introduced. Acoustic echoes can also be present in wireline networks for the same reasons outlined above. In addition, wireline networks have to deal with electrical echoes caused by impedance mismatch at the 2-to-4 wire conversion hybrid.

In many cases, it is desirable to suppress any acoustic echoes that may be present in the voice path. In order to successfully suppress these echoes, they must, first, be detected, and the corresponding echo path delay estimated. Echo detection and delay estimation are also important in Quality of Service (QoS) monitoring applications, where telephone service providers are interested in measuring the voice path quality of their networks. In these monitoring applications, echo detection needs to apply to both acoustic and hybrid echoes.

Many methods for echo detection and suppression have been proposed [1,2]. If echoes are known to be electrical, then an adaptive linear filter can be used effectively to detect as well as cancel the echoes. In cases where acoustic echoes are to be detected and suppressed or cancelled, linear filtering may not produce adequate results and other strategies must be applied [3]. Furthermore, echoes during double-talk need to be distinguished from echoes during single-talk. In this paper we present a framework and method for echo detection and echo path delay estimation for acoustic as well as electrical echoes. In our framework, we consider the problem of echo detection as a pattern recognition problem and apply techniques and features that have been successfully used in speech recognition.

The paper is organized as follows. In the next section, the approach and framework that we employ in this work are presented. Then, in Section 3, a similarity function based on cepstral correlation is proposed as the pattern recognition measure. Experimental results are given in Section 4, followed by conclusions.

2. PATTERN RECOGNTION APPROACH

Figure 1 depicts a block diagram of an echo detection system. The far-end signal is denoted $x(k)$, and the near-end signal, $y(k)$, is composed of near-end speech, $v(k)$, near-end noise, $n(k)$, and an echo of $x(k)$. In this work we segment $x(k)$ and $y(k)$ into frames. A delay line of L frames are kept for $x(k)$, where L depends on the largest echo path delay that is to be detected. Therefore, the delay line consists of L bins, where each bin represents a delay range within the frame duration of $x(k)$. A set of spectral parameters are computed for each frame in the delay line as well as for the current $y(k)$ frame. A similarity function is defined to measure the similarity between a given $y(k)$

frame and each frame in the bins of the $x(k)$ delay line. Let $f_i(m)$ be the similarity function between the m^{th} frame of $y(k)$ and the frame in the i^{th} bin of the delay line, $1 \le i \le L$. The similarity function , $f_i(m)$, is then defined as

$$f_i(m) = f(X_i, Y_m), \tag{1}$$

where X_i is a feature vector representing parameters extracted from the frame in the i^{th} bin of the delay line of $x(k)$, and Y_m represents the feature vector for the m^{th} frame of $y(k)$. If an echo is present in a given $y(k)$ frame then the similarity function between the frame in the delay line bin corresponding to the echo delay and the $y(k)$ frame will consistently exhibit a larger value compared to the other similarity functions for the rest of the delay line bins. A short or long term average of $f_i(m)$ across the index m, when plotted as a function of the index i, $1 \le i \le L$, will exhibit a peak at the index that corresponds to the echo path delay in the near-end signal, $y(k)$. A threshold can be applied to either the instantaneous $f_i(m)$ or the averaged (smoothed) version of it to detect potential echoes. The echo path delay can be readily estimated from the delay line bin index, i^*, where

$$i^* = \arg\max_i f_i(m). \tag{2}$$

One way to view the above approach is to relate it to speech recognition. We can imagine that each bin in delay line corresponds to a word or phrase in the recognizer vocabulary set. In speech recognition, a statistical model is trained for each word in the vocabulary set. Here our model for a given word (i.e., a given delay line bin) is not statistical, but rather the exact set of frames that pass by that bin in the delay line. The unknown signal to be recognized is the near-end signal, $y(k)$. Similar to speech recognition, we use the partial or total cumulative score of the similarity function between the model and the unknown signal to determine if there is a word match (i.e., echoes), and if so, what word (i.e., echo path delay) was present in the unknown signal.

3. SIMILARITY FUNCTION

Considering the speech recognition perspective, we can bring to bear some of the advances that have been made in that field to the echo detection problem. Specifically, one of the critical issues in speech recognition is what set of features to use so that the recognition results are somewhat immune to convolutional and additive noise components. The analogy in the echo detection case is that we are trying to recognize the unknown signal, $y(k)$, from the model signal, $x(k)$, where $y(k)$, has been corrupted by convolutional and non-linear noise components and additive noise components representing near-end noise and/or near-end speech. The one exception is that during near-end speech, it is desirable to detect double-talk if it occurs.

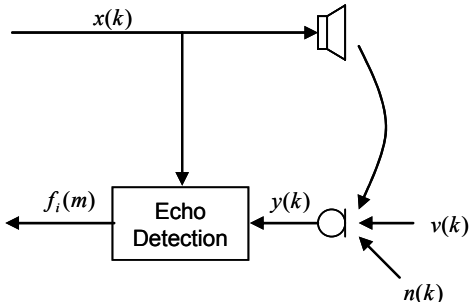

Figure 1. Echo Detection System.

In speech recognition, the use of features based on the Mel-Frequency Cepstral Coefficients (MFCC) has been almost universal [4,5]. Further, the augmentation of the MFCC's with their first and second order derivatives (i.e., delta and delta-delta cepstral coefficients) has been shown to improve accuracy [5]. These delta and delta-delta dynamic features are inherently robust against convolutional noise due to their very definition. Since echoes over short segments can be approximated as having significant linear components, these dynamic features are well suited for echo detection. Therefore, the feature vector that we will employ in this work consists of 12 MFCC's, and their first and second order derivates for a total of 36 features. Although an energy parameter is also used as a feature in speech recognition, we decided not to include it here because of the possibility of near-end speech energy as well as the echo return loss.

It has been shown that using cepstral correlations as a similarity measure is robust against additive noise and outperforms spectral distance measures based on the L2 norm [6]. It was further shown in [6] that cepstral vectors with large norms are more immune to additive noise than cepstral vectors with small norms. Therefore, we define our similarity function based on the one defined in [6]. Specifically, we define the similarity function as the correlation coefficient between X_i and Y_m weighted by the norm of X_i, as follows:

$$f_i(m) = |X_i| \, r(X_i, Y_m), \tag{3}$$

where $r(X_i, Y_m)$ is the correlation coefficient given by,

$$r(X_i, Y_m) = \frac{X_i^T Y_m}{|X_i| \, |Y_m|}. \tag{4}$$

In speech recognition, the cepstral coefficients are typically liftered before the recognition distance function is computed. It is noted that the variance of the cepstral coefficient tend to decrease with increasing quefrency index [7]. Cesptral liftering usually takes the form of normalizing the cepstral coefficients by their variance so as to equalize the contribution of each coefficient in the recognition distance function. In this work we normalize each feature

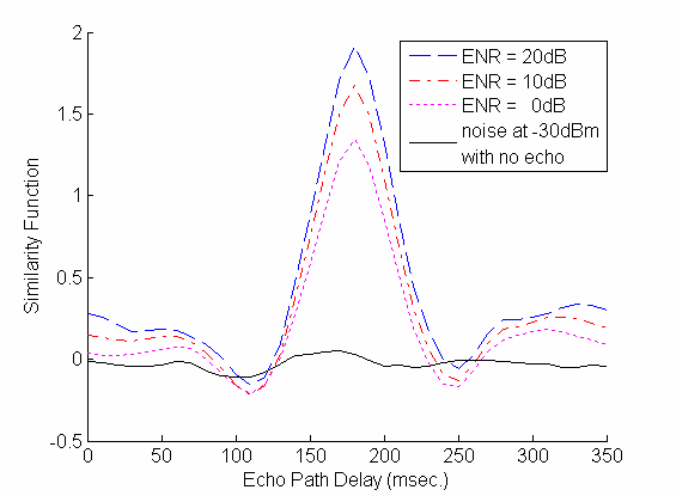

Figure 2. The similarity function mean (over speech frames) versus echo path delay for different Echo-to-Noise Ratios where the noise is car noise.

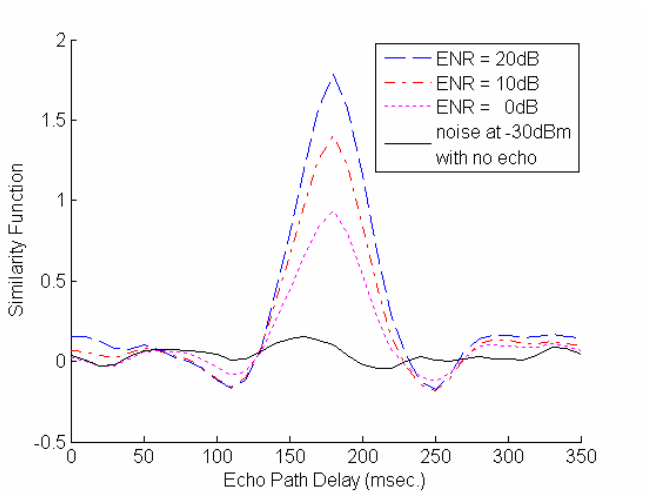

Figure 3. The similarity function mean (over speech frames) versus echo path delay for different Echo-to-Noise Ratios where the noise is mall noise.

in the feature vector by its respective variance. The feature vector variance can be determined offline using a speech database, or, in the case of processing $x(k)$ and $y(k)$ in a batch mode, by computing the feature variance over all frames that include speech in the two signals $x(k)$ and $y(k)$. The later method for estimating the feature variance is used in this work. With variance normalization, the similarity function in equation (3) can be written as

$$f_i(m) = \frac{X_i^T U^{-1} Y_m}{\left| U^{-1/2} Y_m \right|}, \qquad (5)$$

where U is a diagonal covariance matrix.

4. EXPERIMENTAL RESULTS

To test effectiveness of the proposed echo detection method, a system was set up where actual echoes over a commercial 2G GSM network can be recorded. We chose, at random, six sentences spoken by a female speaker, and concatenated them with a period of silence after each sentence. The system enabled an audio file to be played to a mobile handset over an actual call within the GSM network. Any echo suppression within the network was turned off. We then recorded any echoes that returned from the mobile handset operating in non speaker-phone mode. In this setup, no electrical echoes are possible and any echoes recorded are purely acoustic due to, among other factors, the design of the mobile phone itself. Further, due to typical 2G GSM network architecture, the recorded echoes would have gone through a double encoding/decoding using the GSM voice codec, before arriving at our recording station. Therefore, because of the acoustic nature of the echoes, and the tandem encodings, there is a significant degree of non-linearity in the recorded echoes.

To generate different echo conditions, we scale the recorded echo to a desired level and shift it to a predetermined echo path delay. We then mix it with near-end noise and/or speech to simulate a typical near-end signal, $y(k)$. We compute the similarity function, given in equation (5), over 20 msec. frames, updated every 10 msec. This results in a 10 msec granularity in estimating the echo path delay, implying a confidence interval of ± 5 msec. Figures 2 and 3 show plots of the similarity function values versus echo path delay. The similarity function value at any given delay represents the mean value over the 6-sentence utterance. However, to remove the bias caused by including silence periods in the averaging process, a Voice Activity Detector (VAD) was employed to identify non-silence periods in the far-end signal, $x(k)$. The similarity function mean was then computed only over non-silence periods as determined by the VAD. The specific VAD used in our experiment is the VAD (Option 1) that is part of the 3GPP specification for the 12.2 kpbs Enhanced Full Rate coder [8]. In figures 2 and 3, the far-end signal level is -17 dBm, and the Echo Return Loss (ERL) in the near-end signal is 25 dB. The echo path delay is 175 msec. The near-end signal was constructed by mixing the echo signal with different type noises at varying Echo-to-Noise ratios (ENR). As a baseline, we also include in Figures 2 and 3 the case where there is only noise at -30 dBm, and no echo in the near-end signal. Figure 2 shows the results when the near-end noise was recorded in a car driving on a highway, while Figure 3 shows the results when the noise was recorded in a crowded shopping mall.

It is clear from Figures 2 and 3 that even at low ENR, the proposed method results is a clear peak at the correct echo path delay despite the fact that the echoes have a degree of non-linearity due to their acoustic nature and

codec effects. Compared with the case of no echo, it is evident that a reasonable threshold can be applied to detect echoes and estimate the echo path delay correctly. It is useful to note also, that the mall noise is basically babble noise. Nevertheless, the proposed method is able to properly identify the echo, although the peak values at the correct echo path delay are somewhat smaller than the case when the noise is car noise. Also, the difference in the peak value at different ENR's are larger in the case of mall noise compared to the car noise case. This can be due to the fact that the mall noise has speech-like components.

Figure 4 shows the behavior of $f_i(m)$ during periods of single-talk, double-talk, and no speech. In the top part of the figure, $f_i(m)$ is plotted as a function of the time index, m. The middle plot is the near-end signal, while the bottom plot is the far-end signal. The near-end signal is constructed by mixing the following three signals:

1. Echo of the far-end at 25 dB ERL and 175 msec. delay.
2. Near-end car noise at Echo-to-Noise ratio of 5 dB.
3. Near-end speech at -17 dBm.

The near end speech starts around 17 seconds into the signal and consists of four sentences spoken by a male speaker. The first two sentences do not overlap with far end speech, while the last two sentences do overlap, producing double-talk condition. The top plot of Figure 4 represents a smoothed version of the similarity function, $f_i(m)$, at index, i, corresponding to an echo path delay of 175 msec. The smoothing is performed using

$$f_i'(m) = \alpha f_i'(m-1) + (1-\alpha) f_i(m), \qquad (6)$$

where $f_i'(m)$ is the smoothed similarity function, and α is constant set to 0.95.

Comparing regions where there is echo, to regions where there is only near-end noise or near-end noise plus near-end speech, we can see that the smoothed similarity function is able to discriminate well echo and no echo regions. Further, when comparing double-talk regions to single-talk regions, we can see that the value of the similarity function values are lower than the values in regions where only the far end is talking and higher in regions where there is no echo. These results show that the similarity function can reliably detect echoes and can help identify double-talk regions.

5. CONCLUSIONS

We presented a method for echo detection and echo path delay estimation based on a pattern recognition approach. The problem of echo detection is cast as a problem of speech pattern recognition where the specific delay interval in the far-end signal represented a pattern to be matched with the near-end signal. Using features and techniques that have been widely used in speech recognition, we defined a similarity function based on cepstral correlations. We showed, through experimental results, that the similarity function is able to detect echoes and correctly estimate echo path delay. Further, it was shown that the proposed method is useful in detecting double-talk conditions.

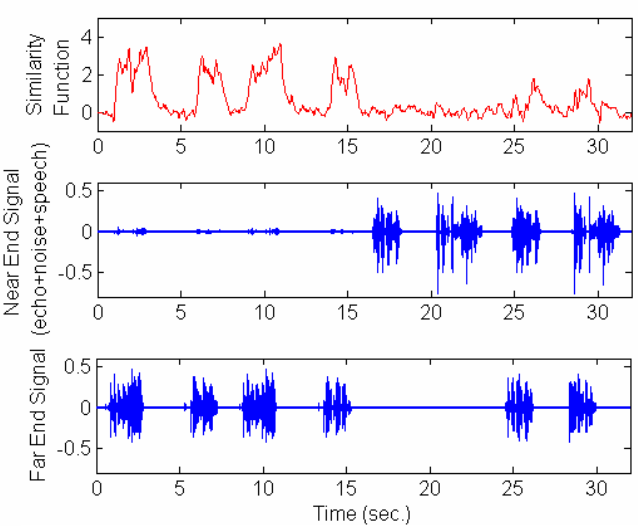

Figure 4. The similarity function performance when the near-end signal consists of echo, car noise, and near end speech in both single-talk and double-talk conditions.

6. REFERENCES

[1] J. Benesty, T. Gansler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer-Verlag, Berlin, 2001.

[2] E. Hansler and G. Schmidt, *Acoustic Echo and Noise Control. A practical Approach,* Wiley, New Jersey, 2004.

[3] F. Kuech, A. Mitnacht, W. Kellermann, "Nonlinear Acoustic Echo Cancellation Using Adaptive Orthogonalized Power Filters," in Proc. *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP),* pp. 18-23, Vol. 3, March 2005.

[4] ETSI, "ETSI ES 202 050 V.1.1.4, Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression algorithms," October 2005.

[5] B. Milner, "Inclusion of Temporal Information Features for Speech Recognition," in Proc. *Int. Conf. on Spoken Language Procession (ICSLP),* pp. 21-24, Vol. 1, October 1996.

[6] D. Mansour, and B. H. Juang, "A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing,* pp. 1659-1671, Vol. 37, Nov. 1989.

[7] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing,* pp. 947-954, Vol. 32, Jul. 1987.

[8] 3GPP, "3GPP TS 26.094 V6.0.0, Voice Activity Detector (VAD)," Dec. 2004.