16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

# DETECTION OF ECHO GENERATED IN MOBILE PHONES USING PITCH DISTANCE

*Tõnu Trump*

Ericsson AB
Tellusborgsvägen 83-87, 12 637, Hägersten, Stockholm, Sweden
phone: +372 53955958, fax: +372 6410405, email: tonu.xt.trump@ericsson.com
web: www.ericsson.com

## ABSTRACT

*This paper presents a method for detecting the presence of echo returned from a mobile phone and estimating its delay. The detector bases its decisions on the distance between pitch periods in uplink and downlink speech signals. We argue that the pitch estimation errors in AMR speech codec are well modelled by Laplacian distribution. Using a mixture of Laplacian and uniform model for errors, we then derive an algorithm to detect the presence of the echo and estimate its delay. Performance of the derived algorithm is investigated by simulations.*

## 1. INTRODUCTION

Ideally the mobile terminals should handle their own echoes in such a way that no echo is transmitted back to the telephony system. Even though many of the mobile phones currently in use are able to handle their echoes properly, there are still models that do not. ITU-T has recognized this problem and has recently consented the Recommendation G.160 that addresses these issues [1]. Following [1] we concentrate on the scenario where the mobile echo control device is located in the telephone system.

It should be noted that differently from the conventional network- or acoustic echo problem [2, 3], where one normally assumes that the echo is present, it is not given that any echo is returned from the mobile phone at all. Therefore, the first step of a mobile echo removal algorithm should be detection of the presence of the echo, as argued in [4]. A simple level based echo detector is also proposed in [4].

To design such a detector we first examine briefly the Adaptive Multi Rate (AMR) codec [5] in Section 2. In Section 3 we present derivation of the detector and some practicalities in Section 4. Section 5 summarizes our simulation study.

Following the terminology common in mobile telephony, we use the term downlink to denote the transmission direction toward mobile and the term uplink for the direction toward the telephony system.

## 2. PROBLEM FORMULATION

In order to detect the echo, which is a (modified) reflection of the original signal one needs a similarity measure between the downlink and the uplink signals. The echo path for the echo, generated by the mobile handsets is nonlinear and non-stationary due to the speech codecs and radio transmission in the echo path, which makes it difficult to use traditional linear methods like adaptive filters, applied directly to the waveform of the signals. As argued in [4], the proper echo removal mechanism in this situation is a nonlinear processor, similar to the one that is used after the linear echo cancellation in ordinary network echo cancellers. In addition, as our measurements with various commercially available mobile telephones show, a large part of popular phone models are equipped with proper means of echo cancellation and do not produce any echo at all. Invoking a nonlinear processor based echo removal in such calls can only harm the voice quality and should therefore be avoided. That's why the first step of any mobile echo reduction system that is placed in the telephone system should be detection of the presence of echo. The nonlinear processor should then be applied only if the presence of echo has first been detected.

Another important point is that speech traverses in the mobile system in coded form and that's why it is advantageous, if such a detector were able to work directly with coded speech signals. In this paper we attempt to design a detector that uses the parameters present in coded speech to detect the presence of echo and estimate its delay. Exact value of the delay associated with the mobile echo is usually unknown and therefore needs to be estimated. The total echo delay builds up of the delays of speech codecs, interleaving in radio interface and other signal processing equipment that appear in the echo path together with unknown transport delays and is typically in the order of couple of hundreds of milliseconds.

The problem addressed in this paper is that the simple level based echo detector is not always reliable enough due to the impact of signals other than echo. The signals that are disturbing for echo detection originate from the microphone of the mobile phone and are actually the ones telephone system is supposed to carry to the other party of the telephone conversation. This is usually referred to as double talk problem in the echo cancellation literature. In this paper we propose a detector that is not sensitive to double talk as shown in sequel of the paper.

Let us now examine the structure of the AMR speech codec that is the codec used in GSM and UMTS mobile networks. According to [5], the AMR codec uses the following parameters to represent speech: Line Spectrum Pair (LSP)

vectors, the fractional pitch lags that represent the fundamental frequency, the innovative codevectors that are used to code the excitation, and the pitch and innovative gains. In the detector, the LSP vectors are converted to the Linear Prediction (LP) filter coefficients and interpolated to obtain LP filters at each subframe. Then, at each 40-sample subframe the excitation is constructed by adding the adaptive and innovative codevectors scaled by their respective gains and the speech is reconstructed by filtering the excitation through the LP synthesis filter. Finally, the reconstructed speech signal is passed through an adaptive postfilter.

The basic structure of the decoder in a simplified form but sufficient for the purposes of this paper, is shown in Figure 1 and described by the equation (1).

$$c \times g_c \times \frac{1}{1+g_p z^{-T}} \times \frac{1}{A(z)} \times \frac{A(z/\gamma_n)}{A(z/\gamma_d)} \qquad (1)$$

In the above $c$ denotes the innovative codevector, $g_c$ denotes the innovative gain (fixed codebook gain), $g_p$ is the pitch gain, $\gamma_n$ and $\gamma_d$ are the postfilter constants and $A$ denotes the LP synthesis filter coefficients. $T$ is the fractional pitch lag, commonly referred to as "pitch period" throughout this paper.

Of the parameters present in AMR coded bit-stream, the pitch period or the fundamental frequency of the speech signal is believed to have best chance to pass a nonlinear echo path unaltered or with a little modification. An intuitive reason for this is that a nonlinear system would likely generate harmonics but it would not alter the fundamental frequency of a sine wave passing it. We therefore select pitch period as the parameter of interest for this paper.

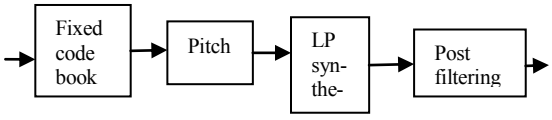


**Figure 1 Simplified structure of AMR decoder**

## 3. DERIVATION

In this section we derive a structure for the echo detector based on comparison of uplink and downlink pitch periods. The derivation follows the principles of statistical hypothesis testing theory described e.g. in [6, 7].

Denote the uplink pitch period for the frame $t$ as $T_{ul}(t)$ and the downlink pitch period for the frame $t$-$\Delta$ as $T_{dl}(t-\Delta)$. The uplink pitch period will be treated as a random variable due to the presence of pitch estimation errors and the contributions from the true signal from mobile side.

Let us also denote the difference between uplink and downlink pitch periods as

$$w(t,\Delta) = T_{ul}(t) - T_{dl}(t-\Delta). \qquad (2)$$

Then we have the following two hypotheses:

$H_0$: the echo is not present and the uplink pitch period is formed based only on the signals present at the mobile side

$H_1$: the uplink signal contains echo as indicated by the similarity of uplink and downlink pitch periods

Under hypothesis $H_1$, the process, $w$, models the errors of echo pitch estimation made by the speech codec residing in mobile phone but also the contribution from signal entering the microphone of the mobile phone. Our belief is that the distribution of the estimation errors can be well approximated by the Laplace distribution and that the contribution from the microphone signal gives a uniform floor to the distribution function. Some motivation for selecting this particular model can be found in Section 5.1.

We thus assume that under the hypothesis $H_1$ the distribution function of $w$ is given by

$$p(w|H_1) = \qquad (3)$$
$$\begin{cases} \alpha \max\left( \frac{1}{2\delta} \exp\left( -\frac{|T_{ul}(t) - T_{dl}(t-\Delta)|}{\delta} \right), \frac{\beta}{b-a} \right), & a < w < b \\ 0 & \text{otherwise.} \end{cases}$$

The constant β, in the above equation, is a design parameter that can be used to weight the Laplace and uniform components and σ is the parameter of Laplace distribution. The variables $a$ and $b$ are determined by the limits in which pitch period can be represented in the AMR codec. In the 12.2 kbit/s mode the pitch period ranges from 18 to 143 and in the other modes from 20 to 143. This gives us limits for the difference between uplink and downlink pitch periods $a$ = -125 and $b$ = 125 in 12.2 kbit/s mode and $a$ = -123, $b$ = 123 in all the other modes. α is a constant normalizing the probability density function so that it integrates to unity. Solving

$$\int_a^b p(w)dw = 1 \qquad (4)$$

for α we obtain

$$\alpha = \frac{b-a}{2\delta\beta\left( \ln\frac{2\delta\beta}{b-a} - 1 \right) + (1+\beta)(b-a)}. \qquad (5)$$

Equation (3) can be rewritten in a more convenient form for further derivation

$$p(w|H_1) =$$
$$= \begin{cases} \frac{\alpha}{2\delta} \exp\left( -\frac{\min\left( |T_{ul}(t) - T_{dl}(t-\Delta)|, -\delta \ln\frac{2\delta\beta}{b-a} \right)}{\delta} \right), & a < w < b \\ 0, & \text{otherwise.} \end{cases} \qquad (6)$$

Under the hypothesis $H_0$, the distribution of $w$ is assumed to be uniform within the interval [a, b].

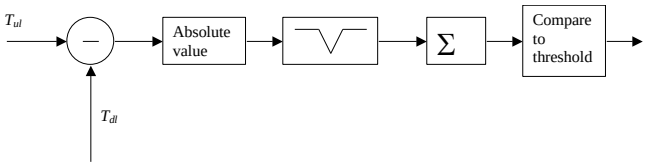


**Figure 2 Structure of the detector**

$$p(w|H_0)=\begin{cases}\dfrac{1}{b-a}, a<w<b\\ 0, \quad \text{otherwise.}\end{cases} \tag{7}$$

We assume that the values taken by the random processes $w(t)$ at various time instances are statistically independent. Then the joint probability density is product of the individual densities

$$\begin{cases} p(\mathbf{w}|H_1)=\prod_{t=1}^{N} p(w(t)|H_1) \\ p(\mathbf{w}|H_0)=\prod_{t=1}^{N} p(w(t)|H_0). \end{cases} \tag{8}$$

Let us now design a likelihood ratio test [5] for the hypotheses mentioned above. We assume that the cost for a correct decision is zero and the cost for any fault is one. We also assume, that both hypotheses have equal a priori probabilities. Then the test is given by

$$\Lambda(T_{ul})= \frac{\prod_{t=1}^{N}\frac{\alpha}{2\delta}\exp\left(-\frac{\min\left(|T_{ul}(t)-T_{dl}(t-\Delta)|,-\delta\ln\frac{2\delta\beta}{b-a}\right)}{\delta}\right)}{\prod_{t=1}^{N}\frac{1}{b-a}} \underset{H_0}{\overset{H_1}{\gtrless}} 1, \tag{9}$$

where $\overset{H_1}{\underset{H_0}{\gtrless}}$ denotes that the hypothesis $H_0$ is accepted if the left side of the inequality is smaller than the right side and the other way around, the hypothesis $H_1$ is accepted if the left side of the inequality is greater than the right side. Taking the logarithm and simplifying the above we obtain the following test

$$\sum_{t=1}^{N} -\min\left(|T_{ul}(t)-T_{dl}(t-\Delta)|,-\delta\ln\frac{2\delta\beta}{b-a}\right) \underset{H_0}{\overset{H_1}{\gtrless}} \delta N\left(\ln\frac{2\delta}{\alpha}-\ln(b-a)\right). \tag{10}$$

The decision device thus needs to compute the absolute distance between the uplink- and downlink pitch periods for all delays, Δ, of interest, saturate the absolute differences at $-\delta\ln\frac{2\delta\beta}{b-a}$, sum up the results and compare the sum with a threshold. The structure of the decision device is shown in Figure 2.

## 4. PRACTICAL CONSIDERATIONS

The detector, as given by (10) is not very convenient for implementation, as it needs computation of a sum over all subframes with each new incoming subframe. To give formula (10) a more convenient, recursive, form, let us first denote the constants appearing in (10) as $c=\delta\left(\ln(b-a)-\ln\frac{2\delta}{\alpha}\right)$ and $d=-\delta\ln\frac{2\delta\beta}{b-a}$. Let us further define a set of distance metrics $D$, one for each echo delay Δ of interest

$$D(t,\Delta)=tc-\sum_{i=1}^{t}\min(T_{ul}(i)-T_{dl}(i-\Delta),d). \tag{11}$$

The distance metrics are functions of time $t$ or more precisely the subframe number. Computation of the distance metric can now easily be reformulated as a running sum i.e. at any time $t$ we compute the following distance metric for each of the delays of interest and compare it with zero

$$D(t,\Delta)=D(t-1,\Delta)+c-\min(|T_{ul}(t)-T_{dl}(t-\Delta)|,d) \underset{H_0}{\overset{H_1}{\gtrless}} 0. \tag{12}$$

Note that a large distance metric means that there is a similarity between the uplink and downlink signals and the other way around, a small distance metric indicates that no similarity has been found. Also note, that one can easily introduce a forgetting factor to the recursive detector structure in order to gradually forget old data as it is customary in e.g. adaptive algorithms [8]. We are, however, not going to do this in this paper. The echo is detected if any of the distance metrics exceeds a certain level. The echo delay corresponds to Δ with largest associated distance metric, $D(t,\Delta)$.

There are several practicalities that need to be added to the basic detector structure derived in the previous section:

- Speech signals are non-stationary and there is no point in running the detector if the downlink speech is missing or has too low power to generate any echo. As a practical limit, the distance metric is updated only if the down-link signal power is above –30 dBm0.
- By a similar reason there is a threshold on the down-link pitch gain. The threshold is set to 10000.
- The detection is only performed on "good" uplink frames i.e. SID frames and corrupted frames are excluded.
- It has been found in practice that $c$ = 7 and $d$ = 9 is a reasonable choice.
- To allow fast detection of a spurious echo burst, the distance metrics are saturated at –200 i.e. we always have $D(t,\Delta)\geq -200$.

Additionally one can notice that the most common error in pitch estimation results in double of the actual pitch period. This can be exploited to enhance the detector. In the particular implementation this has been taken into account by adding a parallel channel to the detector where the downlink pitch period is compared to half of the uplink pitch period

$$D(t,\Delta) = D(t-1,\Delta) + c_1 - \min\left(\left|T_{ul}(t) - \frac{T_{dl}(t-\Delta)}{2}\right|, d_1\right) \underset{H_0}{\overset{H_1}{\gtrless}} 0, \quad (13)$$

where the constants $c_1$ and $d_1$ are selected to be smaller than the corresponding constants $c$ and $d$ in (12) to give a lower weight to the error channel as compared to the main channel. Only one of the updates given by (12) and (13) is used each time $t$. The selected update is the one that results in a larger increase of the distance metric.

## 5. SIMULATION RESULTS

### 5.1 Distribution of pitch estimation errors

In this section we investigate the distribution function of the pitch estimation errors via simulations. The main question to answer is if the distribution function adopted in the previous section is in accordance with what can be observed in the simulations.

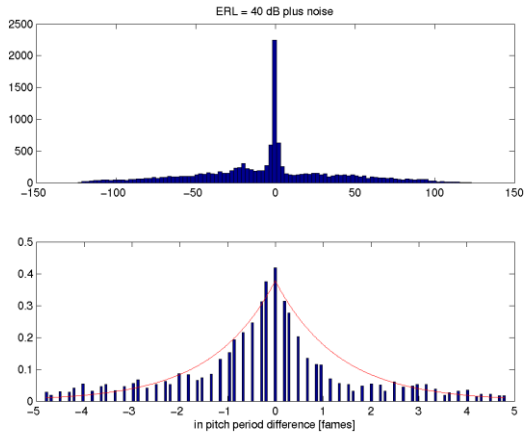


**Figure 3 Histogram of pitch estimation errors. Echo path: single reflection and IRS filter, ERL = -40dB. Near end noise at -60 dBm0**

To answer this question a two minute long speech file that includes both male and female voices at various levels was first coded with the AMR12.2 kbit/s mode codec and then decoded. Then a simple echo path model consisting either of a single reflection or the IRS filter [9] was applied to the signal and the signal was coded again. Echo return loss was varied between 30dB and 40 dB. The estimated pitch was registered from both codecs and compared. The pitch estimates were used only if the downlink power was above –40 dBm0 for the particular fame. A typical example is shown in Figure 3. The upper plot shows the histogram of pitch estimation errors. A narrow peak can be observed around zero and the histogram has long tails ranging from –125 to 125 (which are the limiting values for pitch period). The lower plot shows the Laplace probability density function fitted to the middle part of the histograms. One can see that there is a reasonable fit.

### 5.2 Detection performance

Recordings made with various mobile phones were used to examine the detection performance. All the distance metrics were initialized to -50 and the echo was declared if at least one of the distance metrics became larger than zero. Validity of the detection was verified by listening to the recorded file and comparing the listening and detection results. The two were found to be in a good agreement with each other..

The delay was estimated as the one corresponding to the largest distance metric. As the experiments were done with signals recorded in real mobile systems, the author lacks knowledge of the true echo delays in the test cases. However, the estimates were proven in practice to provide good enough estimates for usage in a practical echo removal device.
Let us finally note that the resolution of the delay estimate is 5 ms due to the 5 ms subframe structure of the AMR speech codec.

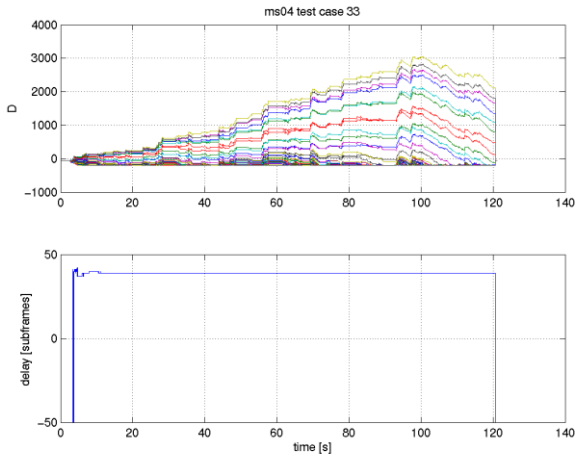


**Figure 4 Distance metrics (upper plot) and estimated delay (lower plot)**

A typical case with a mobile that produces echo is shown in Figure 4. One can see that in this example echo is detected and the delay estimate stabilizes after a couple of seconds to 165 ms. which is a reasonable echo delay for a GSM system.

### 5.3 Resistance to disturbances

Finally we check that the detector is not unnecessarily sensitive to disturbances. In echo context the most common disturbance is so called double talk i.e. the situation when both parties of the telephone conversation are talking at the same time. In this situation speech from the near end side forms a strong disturbance to any algorithm that needs to cope with echoes. The proposed detector algorithm was verified in a large number of simulations involving speech signals from both sides of connection and its double talk performance was found to be good. This is expected as robustness against disturbances is built in the algorithm in form of limited distance measure update in equation (12).

As another and perhaps somewhat spectacular demonstration of double talk performance, we used two speech files with male and female voices speaking exactly the same sentences the same time. The result is shown in Figure 5 with female voice from network side and male voice from mobile side. In this case there are some fault echo detections initially, partly caused by initialization of the distance metrics to –50. Duration of the fault detection is, however, limited to the first two sentences (14 seconds) of the double talk. There was no echo detected in the opposite scenario i.e. male voice from the mobile side and female voice from the network side. Taking into account that it is very unlikely that in an actual telephone call both sides would talk the same sentences simultaneously we conclude that the detector is reasonably resistant to double talk.

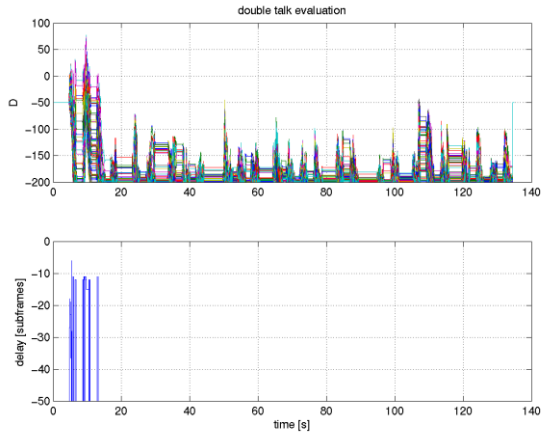


**Figure 5 Evaluation of the detector in severe double talk conditions, mobile side male, network side female**

## 6. CONCLUSIONS

This paper proposes a detector that can be used to detect presence of echo generated by mobile phones and estimate its delay. The detector uses saturated absolute distance between uplink and downlink pitch periods as a similarity measure. A good performance of the detector was demonstrated via simulations.

## REFERENCES

[1] ITU-T Recommendation G.160 *Voice Enhancement devices*, 2008

[2] M. Sondhi and D. Berkley "Silencing Echoes in Telephone Network," *Proc. IEEE* vol. 68, No. 8, pp. 948-963, Aug. 1980

[3] *Signal Processing*, Volume 86, Issue 6, June 2006, *Special issue on Applied Speech and Audio Processing*

[4] A. Perry *Fundamentals of Voice-Quality Engineering in Wireless Networks*, Cambridge University Press, 2007

[5] 3GPP TS 26.090 V6.0.0 (2004-12) 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Mandatory Speech Codec speech processing functions; *Adaptive Multi-Rate (AMR) speech codec; Transcoding functions (Release 6)*

[6] L. Van Trees, *Detection, Estimation, and Modulation Theory*, Wiley & Sons, 1971

[7] S. Kay, *Fundamentals of Statistical Signal Processing, Volume II, Detection Theory*, Prentice Hall, 1998

[8] S. Haykin, *Adaptive Filter Theory, fourth edition*, Prentice Hall, 2002

[9] ITU-T Recommendation G.191, *Software Tools for Speech and Audio Coding Standardization*, 2005