METHODS FOR NONE INTRUSIVE DELAY MEASURMENT FOR AUDIO COMMUNICATION OVER PACKET NETWORKS

Mohammad R Zad-Issa, Norbert Rossello, Laurent Pilati
Mindspeed Technologies
4000 Mac Arthur Blvd, Newport Beach, CA, 92660, USA
1681 route des dolines, 06903 Sophia-Antipolis, Cedex, France
[Mohammad.Zadissa, Norbert.Rossello, Laurent.Pilati] @ Mindspeed.com

ABSTRACT

Measurement of the delay is an important and common problem in communication over packet networks. The end-to-end and the round trip delay are among the factors directly impacting the quality of service as well as the user satisfaction. Multimedia gateways or base stations that perform echo cancellation or suppression often rely on the round trip delay to enhance their performance or to reduce the computational complexity of echo processing logics. In this work, we present two none intrusive methods for delay estimation and tracking. Both methods find the delay using the actual audio signal that is sent through the network. The first approach uses the MDCT transformed domain coefficients of the signal while the second operates in a perceptual domain. Experiments illustrate that both schemes can track the end-to-end and the round trip delay under various network and signal conditions.

1. INTRODUCTION

Audio communication over packet network may involve long and variable delays. In the Voice over Packet (VOP) framework, increased end-to-end delay adds to the conversational effort and leads to user dissatisfaction with the service. Widespread acceptance of VOP technologies requires maintaining network delays within some specified bounds [1]. For doing so, one needs to track the delay in real time. The problem of delay estimation and time alignment is also addressed in the implementation of voice quality assessment systems [2].

In IP networks, the round trip delay may be measured by the use of RTCP services [3]. Although this method offers a close approximation of the delay, it suffers from lack of accuracy and responsiveness to network changes. In addition, RTCP packets capture the delay between two IP nodes. They do not account for the impact of adaptive jitter buffer and other factors that contribute to the user to user delay. Another important element is the presence of echoes. In measuring the round trip delay, one needs to account for the delay introduced by the echo generation process (4-wite to 2-wire hybrid, acoustic paths, etc)

The intent of this work is to estimate the delay as closely as possible to what the end user experiences. This translates to measuring the time difference between a reference and a target audio signal. For the end-to-end case, the target is a delayed, distorted (trans-coding), and possibly scaled version of the reference signal. For the round trip case, in addition to delay, distortions, and attenuations, the target contains different signals, one of which is a realization of the reference (e.g. echo).

Two methods are proposed. Both are none intrusive, i.e. they only reply on the actual audio signal that is generated or received by the end user. The first method uses on the normalized cross correlation in the DCT transformed domain. The second method maps the input signal to a set of perceptual features. Comparing these features allow for the identification of the reference realization in the target signal, and subsequently estimation of the delay.

In section two we define the framework used to implement and validate the proposed methods. Section three and four provide details on each delay estimation scheme. Simulation results are presented in section five. Section 6 offers some concluding remarks.

2. NONE INTRUSIVE DELAY ESTIMATION

Figure 1 illustrates a typical two way voice over packet communication system. The send signal on each side consists of a source audio signal that may contain some environmental noise. The blocks noted as “device” model a media gateway or a base station with various signal processing functions such as low rate coder and decoder, noise reduction, automatic level control, echo cancellation or suppression, packet loss concealment, and adaptive jitter buffer.

1-4244-0367-7/06/$20.00 ©2006 IEEE

The block "path" characterizes a network delay with possible packet loss. A different block is placed in each direction to indicate different network characteristics in the send and received paths. The dashed lines illustrate the possible presence of an echo that may be generated through a line hybrid or by an audio terminal.

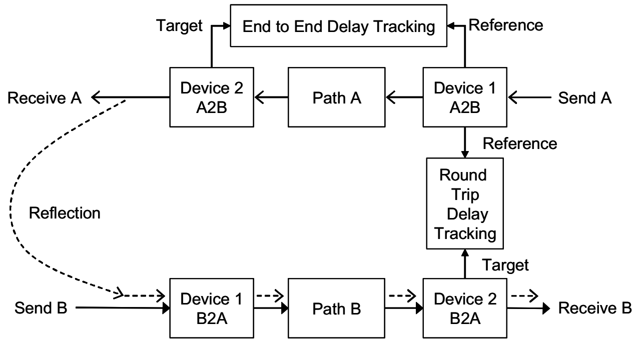

Figure 1: Block diagram of a two way audio over packet communication. Signal A is the reference.

The end-to-end delay tracking logic takes its input from the two edge devices on a path, while the round trip delay estimation operates in real time by taking its inputs from both devices on the reference user side. Since the same methodology applies for end-to-end and round trip delay measurements, in the subsequent sections we will refer to both cases as delay estimation.

3. TRANSFORM DOMAIN APPROACH

Identifying the delay between a signal and its realization can be carried out in the time domain using cross correlation methods [4]. However, since these methods rely on the signal full bandwidth, the none-linearity introduced by coders and other processing elements can severely impact their performance and reliability.

Several studies have demonstrated that the use of cross correlation in the transform domain leads to superior performance than operating in the time domain [5] [6]. We also adopt this general framework. Given the specifics of the problem in hand, it is important for the transform to offer high speech compaction. This offers the possibility of controlling the complexity by performing the Normalized Cross Correlation (NCC) for a subset of the transform basis functions.

In addition, it is necessary for the selected transform to be robust with respect to various perturbation sources that may exist between the reference signal and its realization. This assures that the NCC will be high between the two inputs, despite the distortions.

It has been shown that, among the signal independent transform, Discrete Cosine Transform offers the best compaction for speech and audio and is well suited for coding purposes [7] [8]. The DCT is a well known transform that is also utilized in most of the video and image compression standards. Our experiments also showed that DCT offers the desired advantages and outperforms other signal independent transforms in this delay estimation approach.

3.1. Time Varying Modified Discrete Cosine Transform

Let us define the $\mathbf{x}$ and $\mathbf{X}$ as Modified DCT pair:

$$X(k) = U(k)\sum_{n=0}^{N-1} x_w(n)\cos(\frac{(2n+1)k}{2N}\pi) \tag{1}$$

$$U(k) = \begin{cases} 1/\sqrt{2} & k=0 \\ 1 & k \neq 0 \end{cases}$$

$$n \text{ and } k \in \{0,1,2,\ \ldots,\ N-1\}$$

Where $x_w(n) = x(n) \times w(n)$, and $w(n)$ is a weighting window of length N . Using matrix notation:

$$\mathbf{m}_x = \mathbf{T}_{MDCT(N)}\ \mathbf{x}_w \tag{2}$$

Where $\mathbf{T}$ is the $N \times N$ DCT matrix.

The process of the delay estimation begins with the computation of MDCT coefficients of the input signals. Let us refer to the target signal as $\mathbf{y}$, and the reference signal sample history as $\mathbf{x}$. At the time index n, we have:

$$\mathbf{y} = [y(n), y(n-1), \cdots,\ y(n-N_y+1)]^t \qquad N_y \geq N$$

$$\mathbf{x} = [x(n), x(n-1), \cdots,\ x(n-N_x+1)]^t \qquad N_x >> N_y$$

The sub-vectors of length N are defined as follows:

$$\mathbf{y}(i) = [y(n-i)\times w(0),\ \cdots,\ y(n-i-N+1)\times w(N-1)]^t$$

$$\mathbf{x}(i) = [x(n-i)\times w(0),\ \cdots,\ x(n-i-N+1)\times w(N-1)]^t$$

The respective MDCT vectors are noted as:

$$\mathbf{m}_{\mathbf{x}(i)} = \mathbf{T}_{MDCT(N)}\ \mathbf{x}_i \tag{3}$$

$$\mathbf{m}_{\mathbf{y}(i)} = \mathbf{T}_{MDCT(N)}\ \mathbf{y}_i \tag{4}$$

Equations (3) and (4) populate the MDCT matrices for the reference and the target signals:

$$\mathbf{M_X} = [\mathbf{m}_{x(0)}, \cdots, \mathbf{m}_{x(N_{MX}-1)}] \qquad N \times N_{MX} \tag{5}$$

$$\mathbf{M_Y} = [\mathbf{m}_{y(0)}, \cdots, \mathbf{m}_{y(N_{MY}-1)}] \qquad N \times N_{MY} \tag{6}$$

The sizes of the matrices defined in (5) and (6) depend on the amount of overlap between consecutive MDCT windows. This is a design parameter offering a complexity-time resolution tradeoff. When there are no overlaps, we have:

$$N_{MY} = N_y - N + 1$$

$$N_{MX} = N_x - N + 1$$

One can observe that the row *i* of the matrices in (5) and (6) correspond to the temporal evolution of *i-th* MDCT coefficient of the corresponding signal. Hence, by analyzing a row, it is possible to determine the contribution of the associated MDCT coefficient in the time interval of interest. We select the *P* most contributing coefficients in the target signal matrix $\mathbf{M_Y}$.

$$\mathbf{P_X} = [\mathbf{M_X}(k_1)^t\ ,\cdots,\mathbf{M_X}(k_P)^t] \qquad P \times N \qquad (7)$$

$$\mathbf{P_Y} = [\mathbf{M_Y}(k_1)^t,\cdots,\mathbf{M_Y}(k_P)^t\ ] \qquad P \times N \qquad (8)$$

The matrices in (7) and (8) are composed from the columns of $\mathbf{M_X}$ and $\mathbf{M_Y}$ that correspond to those dominant coefficients.

The next step is to compute the NCC between the corresponding rows of the above two matrices. The indexes of columns at which the NCC are maximized, are candidates for the delay estimate. A state machine tracks these indices and associates a probability to each candidate delay. The final delay estimate is the index with the highest probability.

4. PERCEPTUAL DOMAIN APPROACH

Perceptually Matched Spectral Evolution (PMSE) is a block processing method in which the input signal is mapped to a set of features capturing its perceptual characteristics. These features and their temporal evolution provide a compact and distinctive representation of the signal in the perceptual domain. The delay is estimated by measuring the similarity, in the feature space, between the incoming target and the reference signal's history.

Figure 2 shows the block diagram of the PMSE logic. In the preprocessing bloc, we compute the short term power spectrum of the inputs using 50 % overlapping windows. We use a generic Voice Activity Detector (VAD) for the labeling of input blocks as noise or speech. During the noise segments, the level and spectral characteristics of noise is estimated. We reduced the impact of noise on speech features via spectral subtraction [9]. The signal excitation [10] is then obtained according to a model of the critical bands [11]. In addition to the excitation pattern, we identify the dominant harmonics in the power spectrum and measure an overall signal predictability index. These parameters constitute the feature set associated with the incoming block.

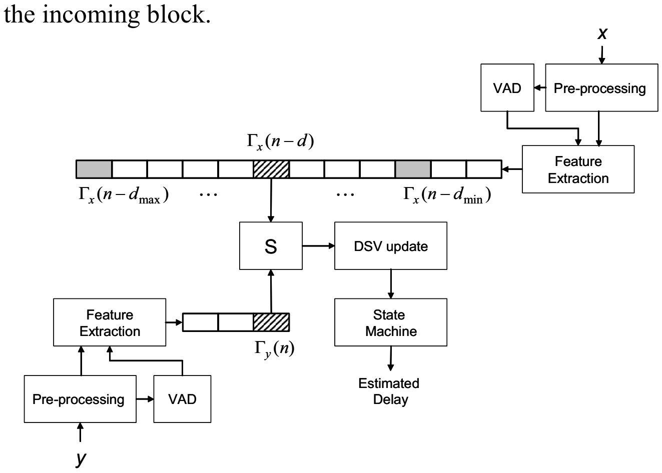

Figure 2: PMSE block diagram. The Delay Score Vector (DSV) tracks the perceptual similarity between the features of the target and the reference inputs.

The similarity between current incoming target features at time index *n* and the reference feature set at candidate delay *d* is measured by:

$$\mathbf{DSV}(d) = \mathbf{S}\ \left(\Gamma_{x(n-d)}\ ,\ \Gamma_{y(n)}\ \right) \qquad (9)$$

$\Gamma_{x(n-d)}$	Feature set of the target at time *n*
$\Gamma_{y(n)}$	Feature set of the reference at delay *d*
S	Similarity measure between two sets

Examples of similarity measures between two features are the normalized cross correlation of excitation patterns and the closeness of the dominant harmonic frequencies. The Delay Score Vector captures the similarity at each candidate delay. The length of DSV depends on the selected block size. Increasing the block size reduces the computational complexity of this algorithm. Reducing the block size, on the other hand, results in higher temporal resolution. The final delay estimate is the output of a state machine locating the maximum within the DSV.

5. EXPERIMENTS

5.1. Simulation Setup

We validated the proposed methods in a software simulation environment. The test material consisted of digitalized speech (8 KHz) of 4 male and 4 female talkers, each 10 seconds. We mixed these samples with car and babble noise, at signal to noise ratio of up to 15 dB. For the round trip delay estimation, we placed a filter followed by amplitude saturation to model the echo path. Each device in figure 2 performed low rate encoding and decoding. We selected

ITU-T G711; Adaptive Multi-Rate coder (AMR: 3GPP TS 26.071); and Enhanced Variable Rate Coder (EVRC: TIA IS-127). The tests were carried out with and without noise reduction in front of the voice coder. The end to end delay were varied from 40 to 420 ms. The reflection delay also varied up to 140 ms. The max round trip delay is therefore 980 ms. We changed the end to end delay in the middle of the test by an integer multiple of the voice coder frame size. The time resolution was set to 2.5 and 5 ms for MDCT and PMSE methods, respectively.

5.2. Results and observations

We notice there the overall performance was dependant on few specific parameters. Representing the overall results in a single table or graph would not clearly show the strengths and weaknesses of the propose methods. Hence, the results are reported based on two categories.

5.2.1 Best cases

These cases constitute about 78% of the total number of test scenarios. For the end-to-end delay estimation, both methods successfully identified the delay (the closest value according to the selected time resolution) nearly immediately after the appearance of the reference's realization in the target input. For the delay change, the new value was also correctly identified once the delay change tracking timer, in the state machine, expired. This timer is a design input parameter that determines the degree of confidence in the newly detected delay value prior to declaring a change. In the round trip case, when the level of echo remained at least 6 dB above the target input background noise, and the echo path consisted of a single reflection, both methods identified the correct delay as soon as the state machine allowed it. Similarly to the previous case, delay changes were detected in a timely manner.

5.2.2 Worse cases

These cases occurred during the round trip delay estimation. They constitute about 8 % of the total number of tests. They are characterized by the following:

- The level of the echo was smaller or equal to the level of the target background noise.
- Constant double talk: All instances of the reference's realization were masked by the target source speech.

In these cases, either the delay was not identified, or it was identified incorrectly, or identified correctly but after 5 second, i.e. half of the test duration.

5.2.3 General Observations

In all the remaining cases, i.e. 14% of total number of scenarios, both methods identified the correct delay +/- one time interval error. We also noticed that if we generated the echo using two or more reflections that were too close (e.g. < 10 ms), the algorithms detected the individual reflections only after having observed sufficient single talk periods. Also, we noticed that the time to detect the delay increased as the amount of none linear distortion from the echo generation process increased. However, both methods remained robust to random packet losses of up to 10%.

6. CONCLUDING REMARKS

We presented two none intrusive methods for measurement of the end-to-end and round trip delay for speech and audio communication over packet networks. Both methods offer user controllable parameters for a desired complexity-accuracy-tracking tradeoff. We validated these schemes, in a Voice-Over-Packet framework. Strengths and weaknesses of these methods were reported.

REFERENCES

[1] Cisco Systems: "*Understanding Delay in Packet Voice Networks*" http://www.cisco.com/warp/public/788/voip/delay-details.pdf, Document ID 5125,

[2] ITU-T Recommendation P.862: "*PESQ: Perceptual Evaluation of Speech Quality*"

[3] IETF RFC 3611: "*RTP Control Protocol Extended Reports (RTCP-XR)*", Nov 2005. http://www.ietf.org/rfc/rfc3611.txt

[4] J. Benesty, D.R. Morgan, "A New Class of Double talk Detectors Based on Cross Correlation", *IEEE Trans. Speech and Audio Processing, Vol. 8, No 2*, pp 168-172, Mar 2000.

[5] T. Gansler, J.Benesty, "A frequency-domain double-talk detector based on a normalized cross-correlation vector", *Signal Processing, Vol. 81*, pp. 1783-1787, Aug 2001.

[6] S.S Narayan, A.M. Peterson, M.J. Narasimha, "Transform Domain LMS Algorithm", *IEEE Trans. Speech and Audio Processing, Vol. Assp-31, No 3*, pp 1177-1192, Jun 1983.

[7] I.Y. Soon, S.N. Koh, C.K. Yeo, "Noisy speech enhancement using discrete cosine transform", Speech *Communication, Vol. 24*, pp. 249-257, Jun 1998.

[8] C.F. Kwong, W.M. Pang, H.C. Wu, K.P. Ho, "Simple DCT-based speech coder for Internet applications", *IEEE International Conf. on Comm.*, Vancouver, pp. 344-348, Jun 1999.

[9] S.F. Boll, "A spectral subtraction algorithm for suppression of acoustic noise in speech", *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing,* Washington DC, pp 200-203 Apr 1979.

[10] E. Zwicker, H. Fastl, "*Psycho-acoustics Facts and Models*", Springer-Verlag, 1990

[11] K. Pohlmann, "*Principals of Digital Audio*", McGraw Hill, 1995