TIME DELAY ESTIMATION ALGORITHMS FOR ECHO CANCELLATION

Kirill SAKHNOV[1], *Ekaterina VERTELETSKAYA*[1], *Boris SIMAK*[1]

[1] Dept. of Telecom. Engineering, Czech Technical University in Prague, Technicka 2, 166 27 Prague, Czech Republic

sakhnkir@fel.cvut.cz, verteeka@fel.cvut.cz, simak@fel.cvut.cz

Abstract. *The following case study describes how to eliminate* echo *in a VoIP network using delay estimation algorithms. It is known that echo with long transmission delays becomes more noticeable to users. Thus, time delay estimation, as a part of echo cancellation, is an important topic during transmission of voice signals over packet-switching telecommunication systems. An echo delay problem associated with IP-based transport networks is discussed in the following text. The paper introduces the comparative study of time delay estimation algorithm, used for estimation of the true time delay between two speech signals. Experimental results of MATLab simulations that describe the performance of several methods based on cross-correlation, normalized cross-correlation and generalized cross-correlation are also presented in the paper.*

Keywords

Echo delay estimation, cross-correlation.

1. Introduction

Echo phenomenon has been always existed in telecommunications networks. Generally it has been noticed on long international telephone calls. As technology advances and the data transmission methods tend more to packet-switching concepts, the traditional echo problem remained up-to-date. An important issue in echo analysis is the round-trip delay of the network. This is a time interval required for a signal from speaker's mouth, across the communication network through the transmit path to the potential source of the echo, and then back across the network again on the receive path to the speaker's ear. The main problem associated with IP-based networks is that the round-trip delay can be never reduced below its fundamental limit. There is always a delay of at least two to three packet sizes (50 to 80 ms) [1] that can make the existing network echo more audible [2]. Therefore, all Voice over IP (VoIP) network terminals should employ echo cancellers to reduce the amplitude of returning echoes. A main parameter of each echo canceller is a length of coverage. Echo canceller coverage specifies the length of time that the echo canceller stores its approximation in memory. The adaptive filter should be long enough to model an unknown system properly, especially in case of VoIP applications [3, 4]. On the other hand, it is known that the active part of the network echo path is usually much smaller compared to the whole echo path that has to be covered by the adaptive filtering algorithm inside the echo canceller. That is why the knowledge of the echo delay is important for using echo cancellers in packet-switching networks. There is a wide family of adaptive filtering algorithms that can exploit sparseness of the echo path to reduce high computational complexity associated with long echo paths [5-8].

There is an investigation of several cross-correlation-based Time Delay Estimation (TDE) algorithms in the paper. The purpose was to investigate each algorithm under a certain scenario and to determine which one is the most suitable candidate for echo control in packet-switching networks. Experimental results are presented in the context of the echo delay estimation which is described in the following section.

2. System Description

Figure 1 illustrates a block diagram of the echo delay estimator. Using a cross-correlation function, the echo delay estimator computes correlation between two voice channels for different set of delays in parallel manner [9]. The delay where the correlation is largest is selected as the echo delay estimate.

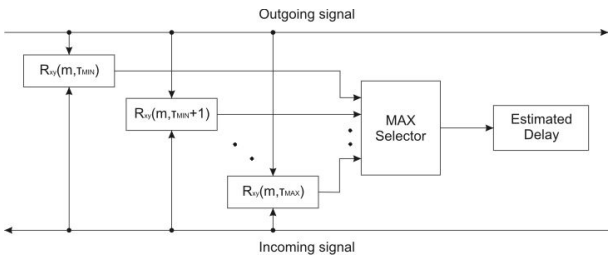


Fig. 1. A block diagram of the echo delay estimator

The idea behind TDE algorithms is to find out this delay value correctly. Signals of interest should pass through the preprocessing stage before the main procedure. First each signal is partitioned into an equal length segments (further just frames). The frames are mutually overlapped after that. The percentage of overlapping was chosen as 50% value. The length of each speech frame equals 256 samples what corresponds to 32 ms in duration. Basically, there is no strict limitation for parameters set up. Any other values can be used by the algorithm instead. It depends on a particular application, for which TDE algorithm is used.

3. Cross-correlation algorithms

3.1 Time Domain Techniques

Time domain implementation of cross-correlation function (CCF) and normalized cross-correlation function (NCCF) is presented in the following. As it was discussed before, first we need to divide the input speech signals into overlapping frames. Denote *x(n)* and *y(n)* as an outgoing and incoming speech signal respectively. Whether parameter N be represent a signal length, L - a frame length, K – an overall number of the segments, and D – a time shift, then *x(n)* can be represented as:

$$x_1(n) = x(n),\ n = 0, ..., L-1. \quad (1)$$

$$x_2(n) = x(n+D),\ n = 0, ..., L-1. \quad (2)$$

$$x_K(n) = x(n+(K-1)\cdot D),\ n = 0, ..., L-1. \quad (3)$$

Figure 2 shows how the frames are related comparing to the full-length signal. The same principle is applied to the incoming signal *y(n)*. It is supposed that there are *K* such segments; $y_1(n)$, ..., $y_K(n)$, and that they cover the entire signal record, i.e., that *(K-1)D+L = N*.

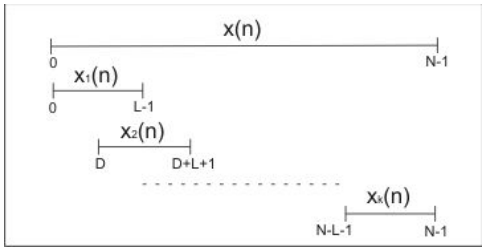


Fig. 2. Segmentation procedure

The cross-correlation function for a successive par of speech frames is then given by [10]

$$R_{xy\,\tau_{MIN}}(m) = \sum_{n=D}^{D+L-1} x(n)\cdot y(n-m),$$
$$R_{xy\,\tau_{MAX}}(m) = \sum_{n=D}^{D+L-1} x(n)\cdot y(n-m), \quad (4)$$
$$n = 0, ..., L-1,\ m = 0, ..., L-1.$$

Here, *x(n)* simply denotes a frame of the outgoing signal, *y(n)* is related to a frame of the incoming signal. According to Figure 1, the estimation of CCF is done for a range of delays. The time shift, $\tau \in [\tau_{MIN}; \tau_{MAX}]$, which causes the maximal peak value of the CC function will be an estimate of the true echo delay T_D. Similarly to the CCF, an estimate of the normalized cross-correlation function is done [11]

$$R^n_{xy\,\tau_{MIN}}(m) = \frac{\sum_{n=D}^{D+L-1} x(n)\cdot y(n-m)}{\sqrt{E_x \cdot E_y}},$$
$$R^n_{xy\,\tau_{MAX}}(m) = \frac{\sum_{n=D}^{D+L-1} x(n)\cdot y(n-m)}{\sqrt{E_x \cdot E_y}}, \quad (5)$$
$$n = 0, ..., L-1,\ m = 0, ..., L-1.$$

Here, E_x and E_y denotes a short-term energy of the outgoing and the incoming signal frames. These values are calculated using the following equations, e. g.

$$E_x = \sum_{n=D}^{D+L-1} x^2(n). \quad (6)$$

$$E_y = \sum_{n=D}^{D+L-1} y^2(n-m). \quad (7)$$

Experimental results for the CCF and the NCF are presented and discussed later in this paper. Generalized cross-correlation algorithms, which operate in the frequency domain, are further presented.

3.2 Frequency Domain Techniques

More sophisticated way how to provide TDE is to compute the cross-correlation function in the frequency domain. This process in literature is called Generalized Cross-Correlation (GCC) [12]. The idea behind this method is to provide pre-filtering of the input signals before calculating CCF. It makes possible to improve the accuracy of delay estimation. The filtering procedure is performed in the frequency domain. Describe this process in greater details.

It is well known, that the simple cross-correlation function, R_{xy}, between signals *x(n)* and *y(n)* is related to the cross-power density function (cross-power spectrum), G_{xy}, by the general inverse Fourier transform relationship

$$R_{xy}(m) = \int_{-\infty}^{\infty} G_{xy}(f)\cdot e^{j2\pi fm} df. \quad (8)$$

When $x(n)$ and $y(n)$ have been filtered with filters having transfer functions $H_1(f)$ and $H_2(f)$, the cross-power spectrum between the filter out-puts is given by

$$G_{xy}^{g}(f) = H_1(f) \cdot H_2^{*}(f) \cdot G_{xy}(f). \quad (9)$$

Consequently, the Generalized Cross-Correlation Function (GCCF) between $x(n)$ and $y(n)$ is presented by [13]

$$R_{xy}^{g}(m) = \int_{-\infty}^{\infty} \Psi_g(f) \cdot G_{xy}(f) \cdot e^{j2\pi fm} df. \quad (10)$$

$$\Psi_g(f) = H_1(f) \cdot H_2^{*}(f). \quad (11)$$

Here, Ψ_g is a generalized weighting function. Table 1 represents weighting functions that were used for the experiments with speech signals.

Tab. 1. Various GCC weighting functions

Processor Name	Weighting Function
Cross-correlation, [13]	1
ROTH, [14]	$1/G_{xy(f)}$
SCOT, [14]	$1/\sqrt{G_{xx}(f) \cdot G_{yy}(f)}$
PHAT, [15]	$1/\lvert G_{xy}(f) \rvert$
Cps-m, [13]	$1/\sqrt[m]{G_{xx}(f) \cdot G_{yy}(f)}$
HT (ML), [16]	$\dfrac{\lvert \gamma_{xy}(f) \rvert^2}{\lvert G_{xy}(f) \rvert \cdot \left[1 - \lvert \gamma_{xy}(f) \rvert^2\right]}$
Eckart, [17]	$\dfrac{\lvert G_{xy}(f) \rvert}{\left[G_{xx}(f) - \lvert G_{xy}(f) \rvert\right] \cdot \left[G_{yy}(f) - \lvert G_{xy}(f) \rvert\right]}$
HB, [18]	$\lvert G_{xy}(f) \rvert / G_{xx}(f) \cdot G_{yy}(f)$
Wiener, [15]	$\lvert \gamma_{xy}(f) \rvert^2$

4. Experimental results

We are beginning our discussion with the simulation results related to the time domain algorithms and then continue with the frequency domain. We used MATLAB software as a simulation environment, where the algorithms were implemented. For simplicity reason double-talk situation was not considered. It means that voice users do not speak simultaneously. The time difference between time when the outgoing signal leaves the voice terminal and moment when the incoming signal containing the echo of the original signal arrives back from the network is referred to as a true echo delay. This value for the first three figures that are presented below (Fig. 3 – Fig 5) equals 6ms (48 samples). For the purpose of TDE it is also necessary to specify time interval through which the value of the true delay is searched. To cover the 6ms delay we chose the interval between 0 and 10ms that corresponds to the maximum delay value of 60 samples. Afterwards we present the estimation results for a group of different delays. It helps to better understand the algorithms performance.

Figure 3 shows the results of the experiment obtained for the algorithm using the cross-correlation function. The first pair of graphs (Fig 3a) shows the curves where τ, the time shirt parameter, equals zero. The next pair of graphs (Fig 3b) has τ equals 24 samples, and the last one (Fig 3c) has τ equals the true delay value of 48 samples. It is evident when the time shift matches the true delay correctly, the maximum peak value of the CCF takes the biggest value. Consequently, this τ is a good candidate to estimate of the true delay. Unfortunately, because of the nonstationary nature of human speech, the CCF does not perform reliably at all situations. Its performance highly depends on many factors, i.e. signal strength, signal-to-noise ratio (SNR), etc [9]. The normalized cross-correlation function is supposed to eliminate some drawbacks of the CCF. Regarding to the normalization, it should not be so sensitive to sudden changes in the signal's amplitude and should outperform its counterpart when work with low level signals. Figure 4 shows the same curves, but obtained for the algorithm using the NCCF.

The advantages of the algorithms working in the frequency domain compared to the algorithms operating in the time domain are accuracy and reduced computational costs. Figure 5 contains the outcomes for the generalized cross-correlation algorithms presented in Table 1. The conventional cross-correlation function is also included in this figure (Fig 5a). It is referred to as a Standard Cross-correlation (SCC) function. The results in the figure are arranged in the following manner. The CCF curves are placed in the right column. There are corresponding echo delay time diagrams in the left column. The red line shows the true delay value, whereas the black color is used to show the estimated values.

The next two tables provide us along with the results for the algorithms from the following point of view. The joint comparison was done in terms of their estimation accuracy. The group of delays was chosen for the experiment. These delays are consistent with ones referenced in the corresponding ITU-T recommendation [19]. Once the respective cross-correlation function was calculated, its maximum peak value is detected using the searching procedure described in Figure 1. The abscissa value of the maximum peak is the estimated delay. Note that 50 trial speech records for each processor were evaluated to obtain the mean value and the Root Mean Square Deviation (RMSD) value [20]. Not only different speech signals, but various models of hybrid impulse response were used as well. The results for delays from 5

to 300 ms are presented in the tables below. Table 1 contents the mean values, whether Table 2 illustrates the estimated RMSD values.

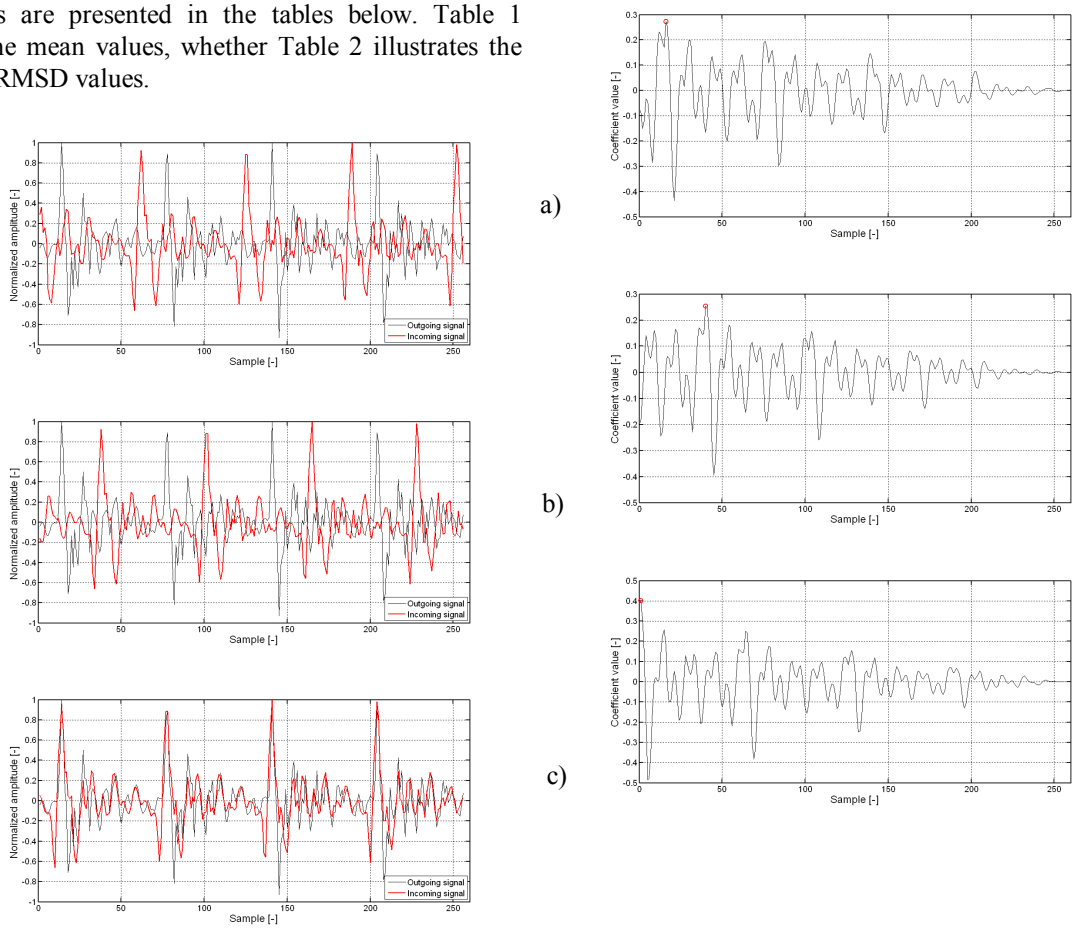


Fig. 3. Cross-correlation function for outcoming and incoming speech signals: a) $\tau = 0$, b) $\tau = 24$, c) $\tau = 48$ samples

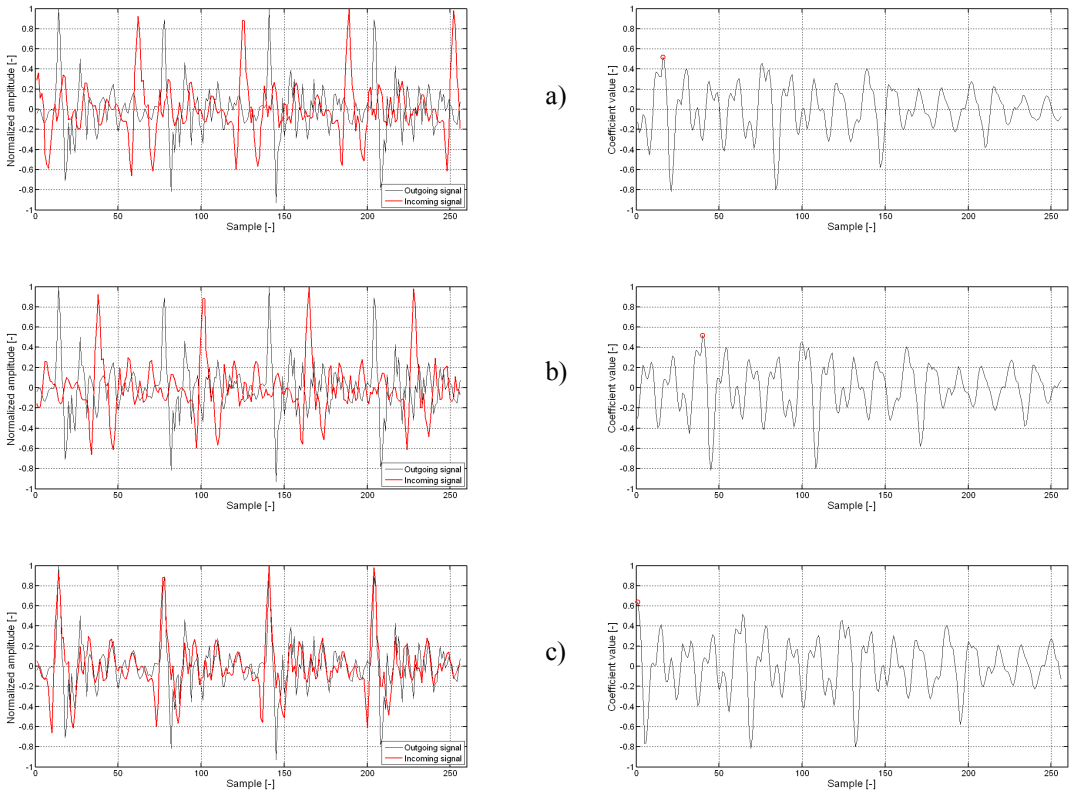
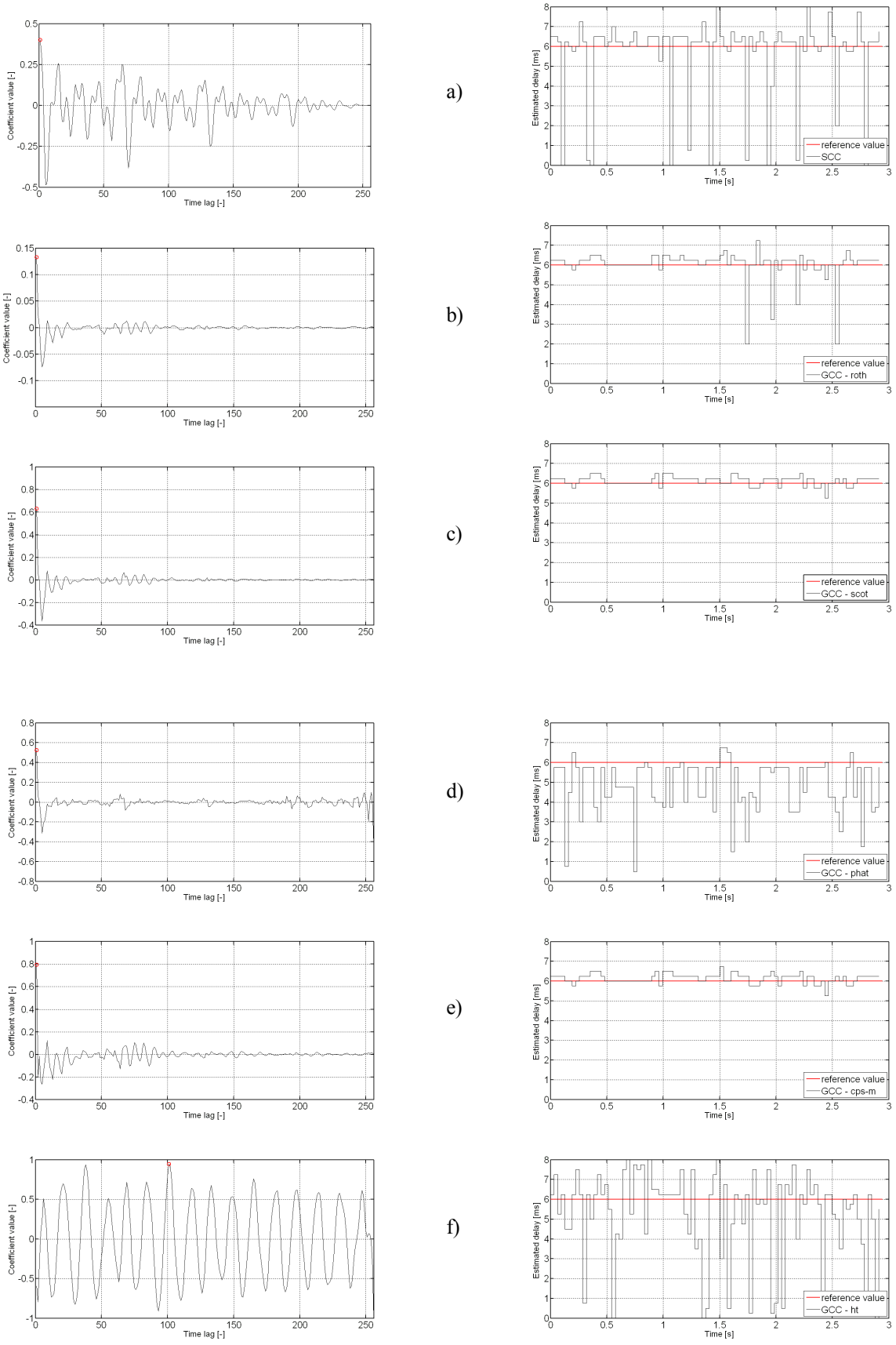


Fig. 4. Normalized cross-correlation function for outcoming and incoming speech signals: a) $\tau = 0$, b) $\tau = 24$, c) $\tau = 48$ samples



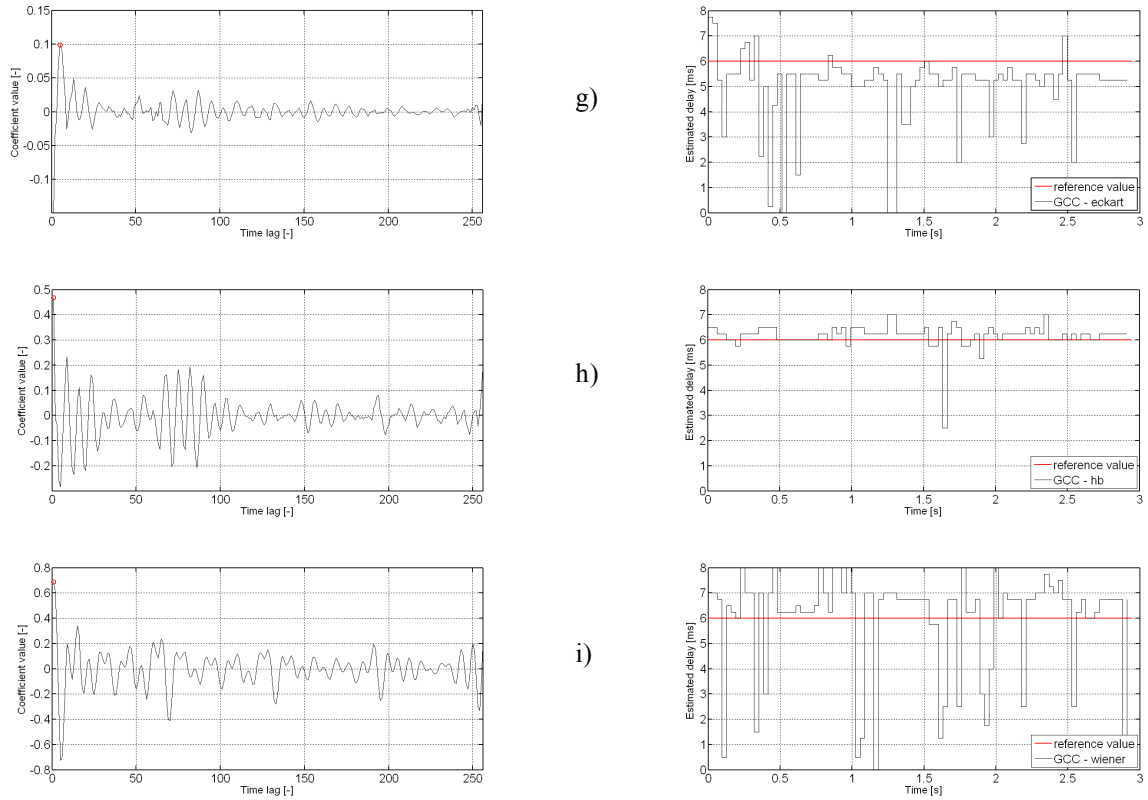


Fig. 5. Generalized cross-correlation function with the different weighting functions: a)standard cross-correlation , b) ROTH, c) SCOT, d) PHAT, e) Cps-m, f) HT (ML), g) Eckart, h) HB, i) Wiener

Tab. 2. Mean values of the estimated delays

[ms]	SCC	ROTH	SCOT	PHAT	CPS-2	HT	ECKART	HB	WIENER
5	4,9	5,1	5,2	3,7	5,2	4,4	4,2	5,2	5,2
10	9,7	10,3	10,3	7,5	10,3	8,8	8,3	10,4	10,3
20	19,5	20,6	20,7	15,0	20,7	17,7	16,7	20,8	20,7
30	29,2	30,9	31,0	22,4	31,0	26,5	25,0	31,2	31,0
50	48,7	51,4	51,6	37,4	51,6	44,2	41,7	52,1	51,7
100	97,3	102,9	103,3	74,8	103,3	88,5	83,3	104,2	103,4
200	194,6	205,7	206,6	149,6	206,6	177,0	166,6	208,3	206,8
300	292,0	308,6	309,9	224,4	309,9	265,4	249,9	312,5	310,3

Tab. 3 Root mean square deviation of the estimated delays

[ms]	SCC	ROTH	SCOT	PHAT	CPS-2	HT	ECKART	HB	WIENER
5	1,4167	0,5889	0,5092	1,7025	0,5092	1,7635	1,3637	0,6119	1,3889
10	1,4257	0,6160	0,5500	2,0571	0,5499	1,8706	1,6079	0,6670	1,4057
20	1,4611	0,7141	0,6892	3,0925	0,6892	2,2488	2,3425	0,8526	1,4710
30	1,5183	0,8530	0,8733	4,2953	0,8733	2,7664	3,2131	1,0940	1,5739
50	1,6883	1,1932	1,2973	6,8474	1,2973	3,9951	5,0795	1,6435	1,8654
100	2,3252	2,1649	2,4490	13,4232	2,4489	7,4097	9,9173	3,1227	2,8599
200	3,9536	4,2119	4,8223	26,7090	4,8223	14,5147	19,7119	6,1606	5,1938
300	5,7159	6,2845	7,2123	40,0252	7,2123	21,6863	29,5337	9,2172	7,6356

5. Conclusion

The current paper provides reader along with the up-to-date correlation-based $_{TD}$E algorithms. The problem associated with long delays taking place in the packet-switching networks was considered as a primary purpose of this research. It is important to continuously monitor a telephone conversation so as to guarantee a required quality of service to VoIP users. Because of the nature of the human ear, the increasing transmission delay associated with packet data transmission can make a negligible echo perceptible. Therefore, we suggest using the echo assessment algorithm based on cross-correlation. If the estimated echo is considerably delayed, it can be annoying to the user. The decision should be put an additional attenuation to the particular channel or to activate an echo canceller to remove the echo. The experiments show that the algorithms precision decreases with increasing transmission delay. The generalized cross-correlation algorithms operating in the frequency domain provide more reliable result comparing to the standard cross-correlation and normalized cross-correlation algorithms. As an alternative to correlation-based methods, techniques using adaptive filtering algorithms may be applied. It will be a topic for our future investigation.

Acknowledgements

Research described in the paper was supervised by Prof. Ing. B. Simak, CSc., FEL CTU in Prague and supported by Czech Technical University grant SGS10/275/OHK3/3T/13 and the Ministry of Education, Youth and Sports of Czech Republic by the research program MSM 6840770014.

References

[1] PERIAKARRUPPAN, G.; LOW, A.L.Y.; AZHAR, H.; RASHID, A. Packet Based Echo Cancellation for Voice Over Internet Protocol Simulated with Variable Amount of Network Delay Time. In Processing of TENCON 2006. 2006 *IEEE Region 10 Conference* . Article ID 10.1109/TENCON.2006.344052.

[2] GORDY, J.D., GOUBRAN, R.A., On the Perceptual Performance Limitations of Echo Cancellers in Wideband Telephony. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, vol. 14, no. 1, pp.33-42.

[3] YOUHONG, L., FOWLER, R., TIAN, W., THOMPSON, L. Enhancing Echo Cancellation via Estimation of Delay. *IEEE Transactions on Signal Processing*, 2005, vol. 53, no. 11, pp.4159-4168.

[4] NISAR, K., HASBULLAH, H., SAID, A.M., Internet Call Delay on Peer to Peer and Phone to Phone VoIP Network. In *Proceedings of ICCET '09, International Conference on Computer Engineering and Technology*, 2009, vol. 2, pp.517-520.

[5] TINGCHAN, W.; BENJANGKAPRASERT, C.; SANGAROON, O.; Performance comparison of adaptive algorithms for multiple echo cancellation in telephone network . In processing of International Conference on Control, Automation and Systems, 2007. ICCAS '07. 2007 , pp 789 - 792.

[6] DYBA, R.A. Parallel Structures for Fast Estimation of Echo Path Pure Delay and Their Applications to Sparse Echo Cancellers. In *Proceedings of CISS 2008, 42nd Annual Conference on Information Sciences and Systems*, 2008, article ID 10.1109/CISS.2008.4558529, pp.241-245.

[7] HONGYANG, D., DYBA, R.A. Efficient Partial Update Algorithm Based on Coefficient Block for Sparse Impulse Response Identification. In *Proceedings of CISS 2008, 42nd Annual Conference on Information Sciences and Systems*, 2008, article ID 10.1109/CISS.2008.4558527, pp.233-236.

[8] HONGYANG, D., DYBA, R.A. Partial Update PNLMS Algorithm for Network Echo Cancellation. In *Proceedings of ICASSP 2009, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, article ID 10.1109/ICASSP.2009.4959837, pp.1329-1332.

[9] CARTER, G.C. Time Delay Estimation. *Ph.D. dissertation,* University of Connecticut, Storrs, CT, 1976, pp.67-70.

[10] MUELLER, M. Signal Delay. *IEEE Transactions on Communications*, 1975, vol. 23, no. 11, pp.1375-1378.

[11] BUCHNER, H., BENESTY, J., GANSLER, T., KELLERMANN, W. Robust Extended Multidelay Filter and Double-talk Detector for Acoustic Echo Cancellation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, vol. 14, no. 5, pp.1633-1644.

[12] HERTZ, D. Time Delay Estimation by Combining Efficient Algorithms and Generalized Cross-correlation Methods. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1986, vol. 34, no. 1, pp.1-7.

[13] KNAPP, C., CARTER, G.C. The Generalized Correlation Method for Estimation of Time Delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1976, vol. 24, no. 4, pp.320-327.

[14] YOUN, D.H., AHMED, N., CARTER, G.C. On the Roth and SCOTH Algorithms: Time-Domain Implementations. In *Proceedings of the IEEE*, 1983, vol. 71, no. 4, pp.536-538.

[15] ZETTERBERG, V., PETTERSSON, M.I., CLAESSON, I. Comparison Between Whitened Generalized Cross-correlation and Adaptive Filter for Time Delay Estimation. In *Proceedings of MTS/IEEE, OCEANS*, 2005, vol. 3, article ID 10.1109/OCEANS.2005.1640117.

[16] WILSON, K.W., DARRELL, T. Learning a Precedence Effect-Like Weighting Function for the Generalized Cross-Correlation Framework. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, vol. 14, no. 6, pp.2156-2164.

[17] TIANSHUANG, Q., HONGYU, W. An Eckart-weighted adaptive time delay estimation method. *IEEE Transactions on Signal Processing*, 1996, vol. 44, no. 9, pp.2332-2335.

[18] CHEN, J., BENESTY, J. and HUANG, Y.A. The SCOT Weighted Adaptive Time Delay Estimation Algorithm Based on Minimum Dispersion Criterion. *Journal of EURASIP on Applied Signal Processing*, 2006, vol. 2006, article ID 26503.

[19] Talker Echo and its Control. *ITU-T Recommendation G.131*, 2003.

[20] ANDERSON, M.P., WOESSNER, W.W. *Applied Groundwater Modeling: Simulation of Flow and Advective Transport.* Academic Press (2nd Edition ed.), USA, 1992.

About Authors ...

Kirill SAKHNOV was born in Uzbekistan. He was awarded an MSc degree from the Czech Technical University in Prague in 2008. He is currently a PhD student at the Department of Telecommunication Engineering of CTU in Prague. His current activities are in the area of adaptive digital signal processing, focused

on problems of acoustical and network echo cancellation in telecommunication devices.

Ekaterina VERTELETSKAYA was born in Uzbekistan. Currently she is a Ph.D. student at the Department of Telecommunication Engineering, CTU in Prague. Her research activities are in the area of speech signal processing, focused on noise reduction.

Boris ŠIMÁK is actively involved in research of digital signal processing in the area of speech and image processing. Since 2007, he is the dean of Faculty of Electrical Engineering, CTU in Prague.