Multi-Modal Acoustic Echo Canceller for Video Conferencing Systems

Mario Gazziro, Guilherme Almeida, Paulo Matias, Hirokazu Tanaka, *Member, IEEE,* Shigenobu Minami

***Abstract*—In video conferencing, if two people talk at the same time howling noises can appear due to failures in the echo cancellation system. In order to reduce these noise situations, a novel AEC (Acoustic Echo Cancellation) system is proposed by using audio and video information. Through image processing techniques we have developed a new DTD (Double Talk Detector) based on video information (video-DTD). A novel Acoustic Echo Cancellation system was successfully developed, integrating audio and video information. In simulated results using data from a real scenary the misalignment of the new AEC (with video-DTD) was smaller than a traditional AEC.**

***Index Terms*—Acoustic Echo Cancellation, Double Talk Detection, Multi-Modal, Video Conference, Image Processing.**

I. Introduction

WITH the advent of internet, communication has become easier in such a way that people can hold a video conference session at their own houses. But if two people talk at the same time, howling noises can appear due to failures in the echo cancelation system [3]. In order to reduce these noise situations a novel AEC (Acoustic Echo Canceller) system is proposed by using audio and video information about the user. The implementation of this system is the objective of this paper.

II. Related Work

Murata et al [1] attempted to integrate sound and image before solving echo cancellation problems, but using only binarized lip image processing [2] and performing a low integration with the AEC.

III. Acoustic Echo Canceller

In acoustic echo cancellation, a measured microphone signal contains two signals: the Near-end speech signal and the Far-end echoed speech signal. The goal is to remove the Far-end echoed speech signal from the microphone signal so that only the Near-end speech signal is transmitted. The block diagram of an AEC is showed in Figure 1.

In a full-duplex conversation, the presence of signals other than the Far-end signal convolves with the echo path, thus inhibiting the ability of the adaptive algorithm to model the system.

The presence of the Near-end talker during the Far-end speech is a source of disruption in the adaptation of the filter. Therefore, adaptation of the filter must be prevented during this double talk scenario via a double talk detector (DTD).

M. Gazziro is with the ICMC/USP, Brazil. e-mail: mariogazziro@usp.br
G. Almeida is with the Wernher von Braun Center, Brazil.
P. Matias is with the IFSC/USP, Brazil.
H. Tanaka and S. Minami are with Toshiba Semiconductors, Japan.

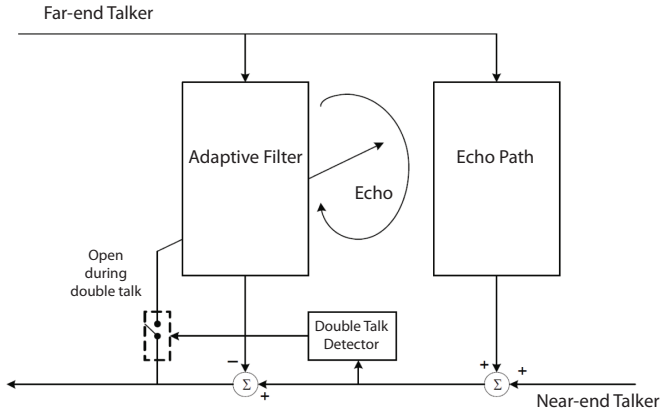

Fig. 1. Acoustic echo canceller structure.

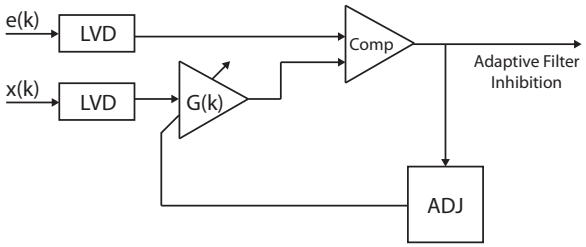

Fig. 2. Minami DTD algorithm block diagram.

A. Double Talk Detector

There are many approaches to determining when a double talk scenario occurs, but all of them follow the same general procedure, which says that a detection statistic, η, can be formulated from the excitation, desired, and/or error signals. Then this detection statistic is compared to a threshold, in order to determine if a double talk can be declared.

The classical implementation of DTD uses Geigel [4] or NCR [5] (Normalized Cross Correlation) algorithms. But in this paper we will focus on an algorithm proposed by [6]. With this algorithm we can perform the comparison between the Far-end signal and the Adaptive Filter output. A block diagram is showed Figure 2.

The output from the Adaptive Filter $e(k)$ and the Far-end Signal $x(k)$ pass through a level detector, which converts the signal as if it was an envelope detector. The block diagram of this level detector can be seen in Figure 3. It compares the current sample of the signal with the previous one with a small delay and a small degradation. If the current sample is higher than the previous one, the output will be the current sample.

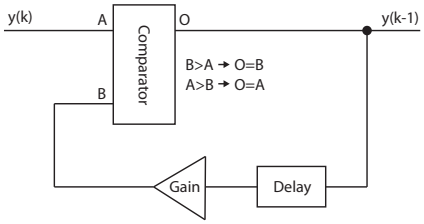

Fig. 3. Details of LVD (Level Detector) block in Minami DTD.

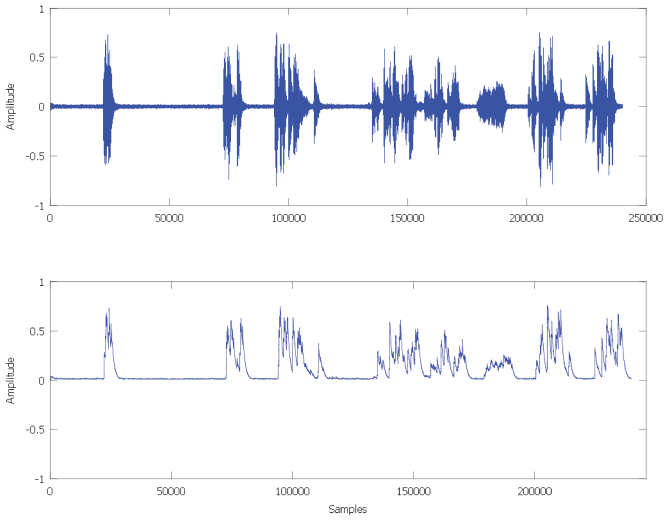

Fig. 4. Level detector results with voice sample.

Otherwise, it will be the previous one.

This LVD block was implemented with a degradation of 0.999 and it was tested with an audio sample, resulting in the plots from Figure 4.

After both signals have passed through this level detector, they follow to a comparator, which will decide if there is a double talk situation through the Equation 1.

$$L_e(k) > L_x(k) + G(k) \tag{1}$$

Where:
$L_e(k)$ = Adaptive filter output signal after level detector
$L_x(k)$ = Far-End signal after passing through level detector
$G(k)$ = Detection Gain

If this condition is satisfied, it means a double talk situation was detected and $G(k)$ is the value which determines accuracy of this algorithm. Considering a range of variation from 0 to $-40dB$, the gain $G(k)$ varies because of the block adjustment (ADJ) and according to the detection results previously obtained. In the case of a double talk situation, $G(k)$ is increased at the rate of equation 2, considering a maximum double talk duration of 3 seconds.

$$\delta G(k) = 40dB/3s = 13.333dB/s \tag{2}$$

In the case of a non-double talk situation, $G(k)$ is decreased at the rate of equation 3, considering that 10s is sufficient time for the filter to adapt.

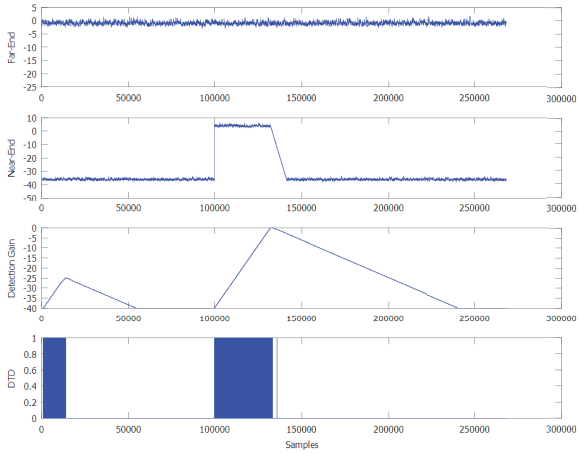

Fig. 5. Simulated results using WGN samples with a small DTD duration.

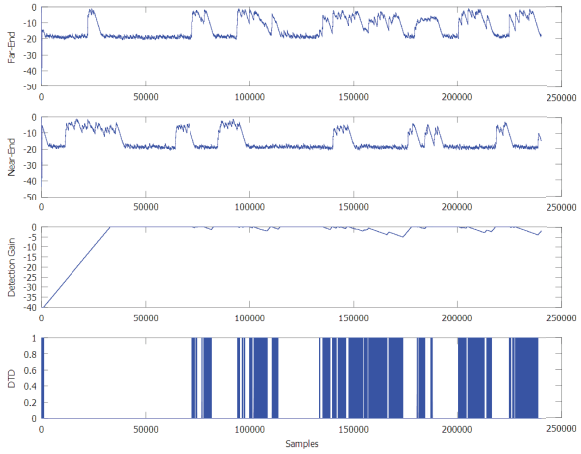

Fig. 6. Simulated results using voice samples with many DTD situations.

$$\delta G(k) = 40dB/10s = 4dB/s \tag{3}$$

After being implemented, this algorithm was simulated by using White Gaussian Noise (WGN) as input for both Far-End and Near-end talkers, simulating 3 seconds of double talk situation. The results can be seen in Figure 5 and the simulation using audio samples with several double talk situations are represented in Figure 6.

IV. Proposed Advanced AEC

In this paper a novel type of AEC has been proposed, based on the double talk detector algorithm proposed by Minami and using video information to enhance the DTD performance. The integration of this novel DTD into the AEC structure is depicted in Figure 7.

In this new proposal we improve the original Minami DTD with changes in the adjustment block, which now should modify the detection gain $G(k)$ by using a new piece of information: video double talk detection (Video-DTD). This changes are depicted in Figure 8.

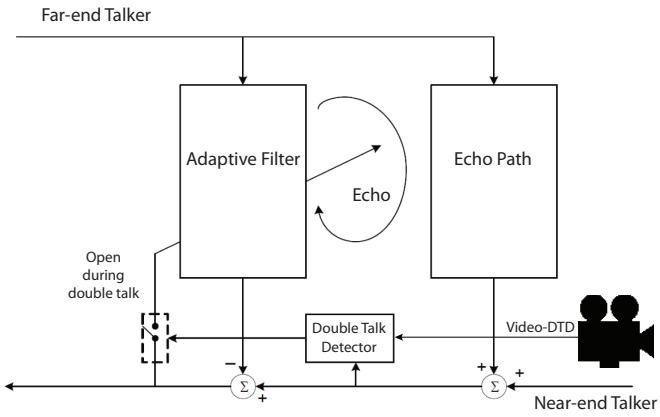

Fig. 7. Advanced echo canceller with Video-DTD.

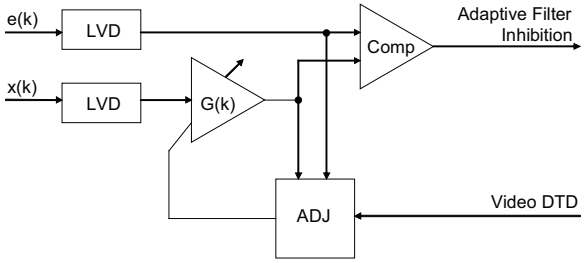

Fig. 8. Novel double talk detection with Video-DTD integration.

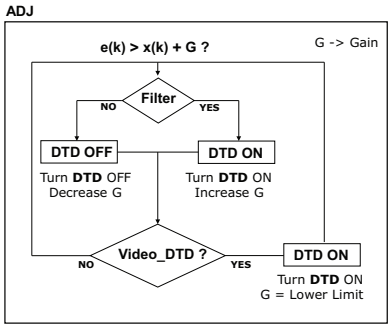

Fig. 9. Flowchart of DTD adjustment (ADJ) block.

Figure 9 presents the flowchart of the adjustment block. The DTD is the output of this block. The Gain and DTD condition are set based on Filter output and Far-end signals. If a Video-DTD signal is received, the DTD is turned ON and the Gain is set at the lower limit.

V. Experimental Evaluation of the Novel AEC

In this section we present the results of our novel AEC, using Video-DTD.

A video was recorded from a previous established scenery, and the image frames and stereo audio were extracted from the camera.

The original video was recorded in Full-HD (1920x1880) with an interlaced scan video camera. It has 30 frames per second originally, but it was resampled to 15 frames per second in order to produce a de-interlaced video (progressive). The images are then reduced about 50% (960x544 pixels) in order

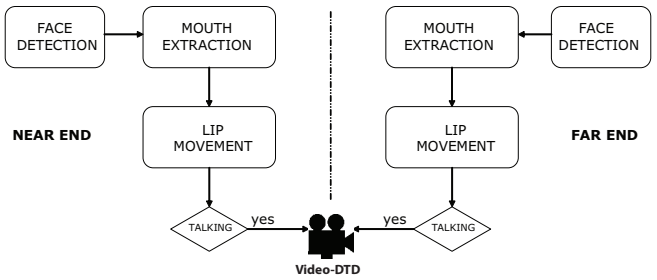

Fig. 10. Video DTD system flowchart.

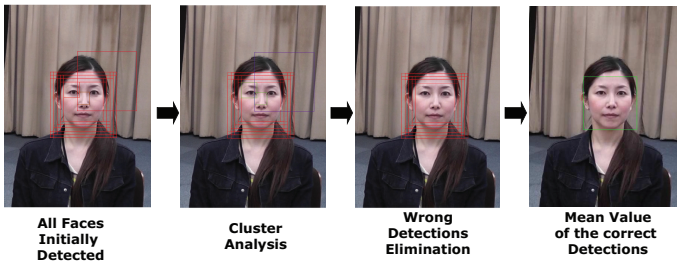

Fig. 11. Face detection stages.

to reduce the simulation time needed. In this video there are sequences of Silent, Opening, Closing and Roll mouth periods. Roll mouth are sequenced frames with a slight displacement of the whole face, but without producing any sound.

In a multi-modal approach we wanted to use video and audio information mixed in a novel technique to evaluate a double talk situation. Figure 10 shows the proposed video DTD system flowchart. For both sides (Near and Far end) we have three stages: Face Detection, Mouth Extraction and Lip Movement. On each side we determine if the speakers are talking or not. When two speakers are talking at the same time, a double-talking condition is detected by the video system and then informed to the novel DTD algorithm.

A. Face Detection

To perform face detection in a robust and reliable way, we combine the recent Viola-Jones algorithm [8] with a classical cluster analysis method called K-Means [9]. Figure 11 shows the stages in the face detection flowchart.

After applying the Viola-Jones results into a K-Means cluster algorithm, the frames detected by Viola-Jones will be sorted in different clusters. This is an unsupervised analysis and it is capable of classifying as many clusters as those initially found. Figure 12 shows the K-Means output for the frames in the first picture of Figure 11. Three clusters were classified. The cluster number one (green/cross) contains 2 frames, cluster number two (red/star) contains 31 frames and cluster number three (blue/circle) contains only 1 frame.

The second picture of Figure 11 depicts the frames classified according to the cluster analysis (red, green and blue). Afterwards, only the frames from the small clusters are removed, which is show in the third picture of Figure 11. Finally, a mean value calculation is made in order to detect the main face.

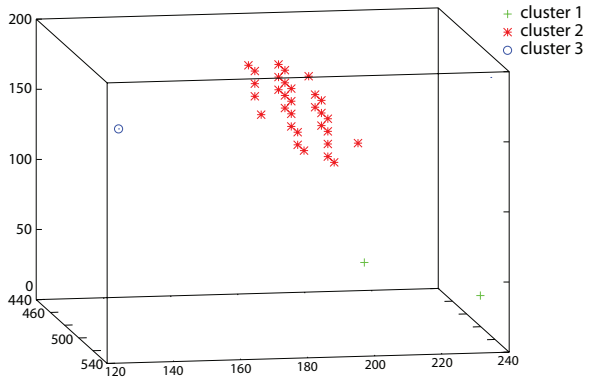

Fig. 12. Cluster being sort out using the K-Means algorithm. X and Y axes: bottom-left coordinates of square frames; Z axis: size of the square frame.

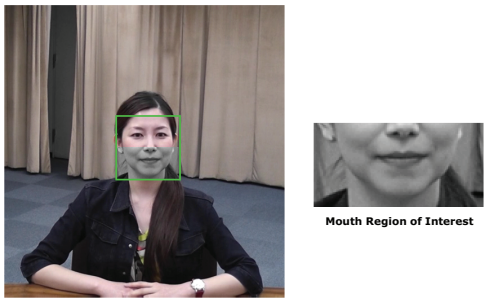

Fig. 13. Mouth extraction.

B. Mouth Extraction

The mouth area search is the simplest technique in all proposed Video-DTD. We only compute the bottom half of the face frame detected, as shown in Figure 13.

C. Lip Movement

In order to detect the mouth movement, we will adopt the approach developed by Tamura et al [2] using optical flow analysis. The optical flow vectors (horizontal and vertical) are added, which generates an increase of their intensity when the mouth is opening, as well as a decrease when the mouth is closing.

The optical flow was computed only in the Region of Interest, around the mouth, acquired according to the descriptions in the section V-B. We use three algorithms to compute the optical flow: Lucas-Kanade (LK) [10], Horn & Schunck (HS) [11] and Brox (BROX) [7]. Each one of the four situations (Silent, Opening, Closing and Roll) to be classified was tested with these four algorithms, and the results are showed in Figures 14, 15 and 16.

Figure 17 shows the optical flow variance for each mouth condition using all algorithms. In this figure we see that it is not possible to use variance only as a classification factor. However, the variance of the Silent condition was close to zero (as presented by Tamura); the variance of the Roll condition is higher, even bigger than the Opening or Closing situations (for most algorithms).

An extension to Tamura's proposal was performed here, in order to also classify periods of face rolling. When the

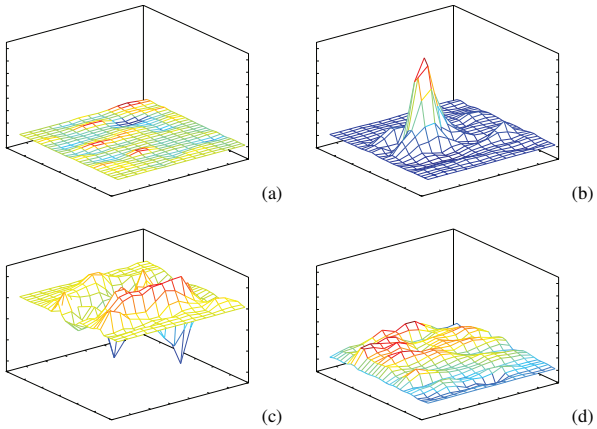

Fig. 14. Mouth situations: (a) Silent, (b) Opening, (c) Closing and (d) Roll using Lucas-Kanade algorithm.

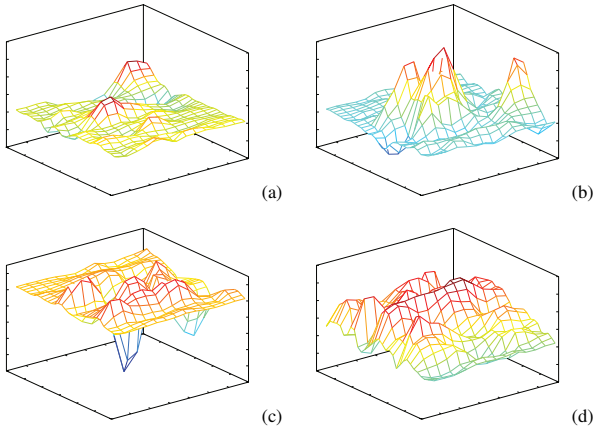

Fig. 15. Mouth situations: (a) Silent, (b) Opening, (c) Closing and (d) Roll using Horn & Schunck algorithm.

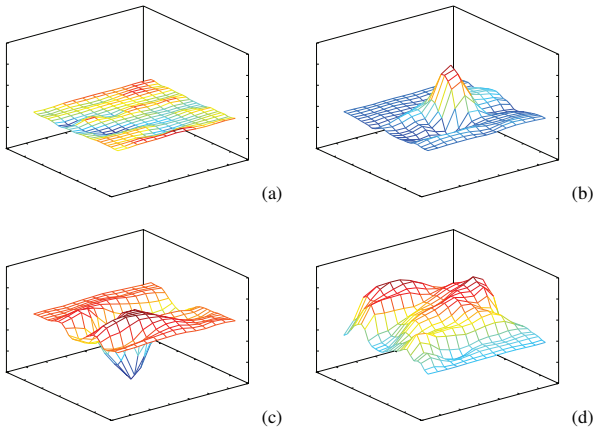

Fig. 16. Mouth situations: (a) Silent, (b) Opening, (c) Closing and (d) Roll using Brox algorithm.

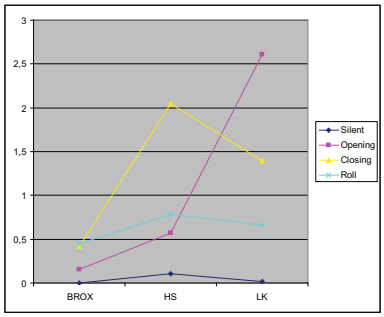

Fig. 17. Variance analysis of the optical flow algorithms in the four situations.

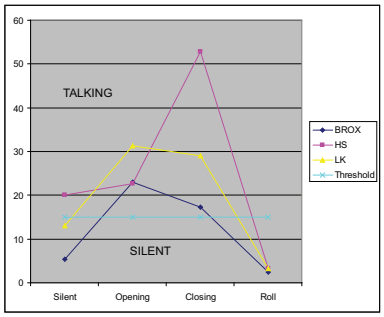

Fig. 18. Peak above mean using optical flow algorithms.

face is inclined a little (clockwise or counter-clockwise), the optical flow suffers small increases or decreases as a whole. An analysis of maximum and minimum peaks above the mean was performed.

Figure 18 depicts the final comparison between results for three algorithms. If a threshold was established in order to separate which algorithm correctly classifies each situation, we can see that only Horn & Schunck (HS) is wrong.

Silent and Roll are non-talking conditions, indicated in the region below the threshold (SILENT). Opening and Closing are talking conditions, indicated in the region above the threshold. With this in mind one can see that Horn & Schunck algorithm is not suitable for detecting talking conditions. There is no significant difference between the Silent and Opening mouth conditions. Brox and Lucas-Kanade are suitable for detecting talking conditions. Brox is faster than LK in simulations, but LK is more suitable to speed up when implemented in hardware. Brox provides a better classification for Silent conditions than LK.

VI. Results

As described in detail in section III-A, a White Gaussian Noise analysis is the first step in order to test the new Echo Cancellation System. Figure 19 shows the behavior of Gain and DTD output when DTD-Video signal is set.

The small delay between the start of DTD situation and the application of DTD-Video signal is related with the inherent delay of the image processing. Results using real voice data are depicted in the Figures 21 and 22.

The main result of the system can be seen in Figure 22. The misalignment of the novel AEC, using the Video-DTD is represented in the red (thin) line. The same audio samples and

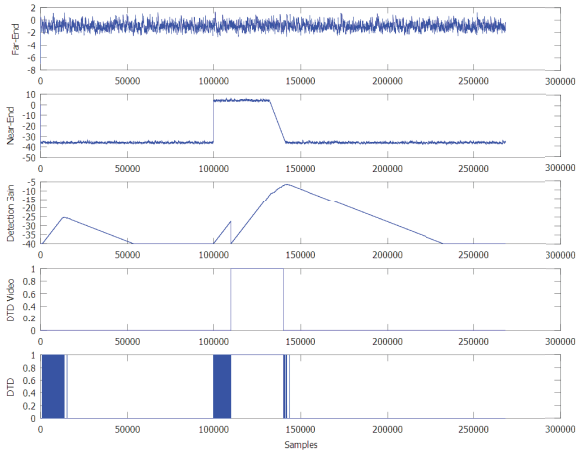

Fig. 19. DTD result using Video-DTD with WGN.

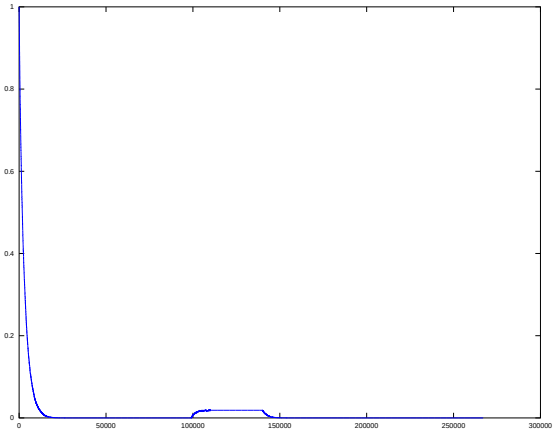

Fig. 20. Misaligment of White Gaussian Noise simulation.

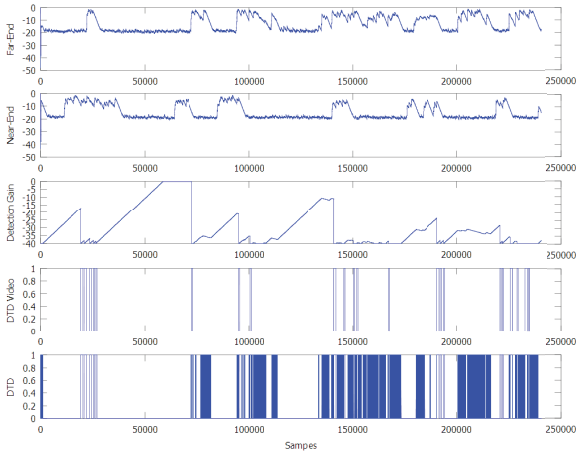

Fig. 21. DTD result using Video-DTD with voice data.

room impulse response were used in a traditional AEC using

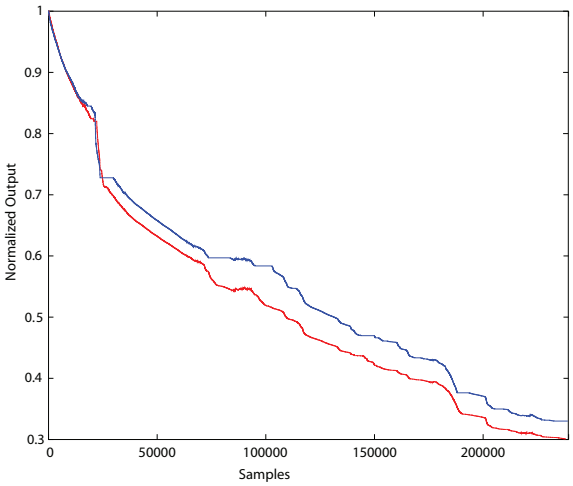

Fig. 22. Misaligment produced by Video-DTD (red/thin) slightly better than NCR-DTD (blue/thicker).

NCR-DTD algorithm and is represented by the blue (thicker) line. The novel DTD is slightly better than audio-only DTD.

VII. Conclusion

A novel Acoustic Echo Cancellation system was successfully developed, integrating audio and video information. For video-DTD analysis we verify that the relation peak above the mean is more indicated due to the rolling face problem. The performance of this new system is slightly better than a traditional audio-only one. Using a DTD based on NCR has indicated that the video information applied in a double talk detection could improve the performance of the Acoustic Echo Cancellation system.

References

[1] Murata, N.; Kajikawa, Y.; *A Double-Talk-Detector Integrating Sound and Image Information*, Proceedings of IEICE SIP 108, 213-218, 2009

[2] Tamura, S., Iwano, K.; Furui, S., *Multi-Modal Speech Recognition Using Optical-Flow Analysis for Lip Images*, Journal of VLSI Signal Processing 36, pp. 117-124, Feb 2004

[3] Ahgren, P.; Jakobsson, A.; *A study of double-talk detection performance in the presence of acoustic echo path changes*, Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on , vol.3, no., pp. iii/141- iii/144 Vol. 3, 18-23 March 2005

[4] Duttweiler, D.; *A Twelve-Channel Digital Echo Canceler*, Communications, IEEE Transactions on , vol.26, no.5, pp. 647- 653, May 1978

[5] Benesty, J.; Morgan, D.R.; Cho, J.H.; *A new class of doubletalk detectors based on cross-correlation*, Speech and Audio Processing, IEEE Transactions on , vol.8, no.2, pp.168-172, Mar 2000

[6] Minami, S.; Kawasaki, T.; *A Double Talk Detection Method for an Echo Canceller*, IEEE International Communications Conference (ICC"85), pp. 1492-1497, 1985

[7] Brox, T; Bruhn, A; Papenberg, N.; Weickert, J. *High Accuracy Optical Flow Estimation Based on a Theory for Warping*, Proceedings 8th European Conference on Computer Vision, Springer LNCS 3024, T. Pajdla and J. Matas (Eds.) (4), 25-36, Prague, Czech Republic, 2004

[8] Viola, P.; Jones, M.; *Robust real-time object detection* International Journal of Computer Vision 57(2), 137154, 2004

[9] MacQueen, J.; *Some methods for classification and analysis of multivariate observations*, in Proc. 5th Berkeley Symp. Math. Stat. Probab., L. M. L. Cam and J. Neyman, Eds. Berkeley, CA: Univ. California Press, 1967

[10] Lucas, B.D.; Kanade, T.; *An Interative Image Registration Technique with an Application to Stereo Vision*, DARPA Image Understanding Workshop, pp 121-130, 1981

[11] Horn, B.K.P.; Schunck, B.G.; *Determining Optical Flow*, Artificial Intelligent, vol 9, pp 185-203, 1981

Mario Gazziro received the BSc degree (2003) and MSc degree (2005) in computer science at Instituto de Ciencias Matematicas e de Computacao-USP, Brazil. He received the PhD degree (2009) in applied physics at Instituto de Fisica de Sao Carlos-USP, Brazil (with internship at Instituto Superior Tecnico de Lisboa, Portugal). He is also a specialist in microelectronis by Cadence (2008), USA, and by Toshiba Semiconductors (2010). Currently he is a professor at ICMC/USP, Brazil.

Guilherme Almeida is an Audio Engineer from Federal University of Minas Gerais, Brazil. He is a microelectronics engineer with training from Cadence Design Systems, USA and from Toshiba Semiconductors, Japan. Currently he is an engineer at Wernher von Braun Center, Brazil.

Paulo Matias received the BSc degree (2009) in Computational Physics at Instituto de Fsica de So Carlos/USP, Brazil. Working towards the PhD degree in applied physics at the same institution. Research includes scientific instrumentation and neurobiophysics.

Hirokazu Tanaka has a PhD in Communication Engineering from the Osaka University. Currently he is a specialist at Toshiba Corporation Semiconductor Company, Japan.

Shigenobu Minami has a PhD in Information Technology from Hokkaido University. Currently he is an assistant to CTE at Toshiba Corporation Semiconductor Company, Japan.