COMMITTEE T1
CONTRIBUTION

Document Number T1Q1.5/92-113

****

STANDARDS PROJECT: Analog Interface Performance Specifications for Digital Video Teleconferencing/Video Telephony Service

****

TITLE: Spatial and Temporal Information Measures for Video Quality

****

ISSUE ADDRESSED: Video Quality

****

SOURCE: NTIA/ITS - Arthur Webster and Steve Wolf

****

DATE: January 22, 1992

****

DISTRIBUTION TO: T1Q1.5

****

KEYWORDS: Video Teleconferencing, Video Telephony, Video Quality, Subjective Quality, Objective Quality

****

DISCLAIMER:

****

Spatial and Temporal Information Measures for Video Quality

1. Introduction

As part of the work to produce an objective video quality measurement system, an effort is being made to decouple the spatial and temporal components contributing to subjective (as perceived by human beings) quality. This contribution will present the current state of this work.

Among the goals of this work is to establish quantitative measures which can be made on a video scene to determine the spatial information content (S) and the temporal information content (T) of the scene. Time histories of these two quantities can be used to determine the location of the scene in the spatial-temporal information matrix (see Table 1, T1Q1.5/91-132 R1).

From these two scalar measures, S and T, a vector may be formed as shown in Figure 1.

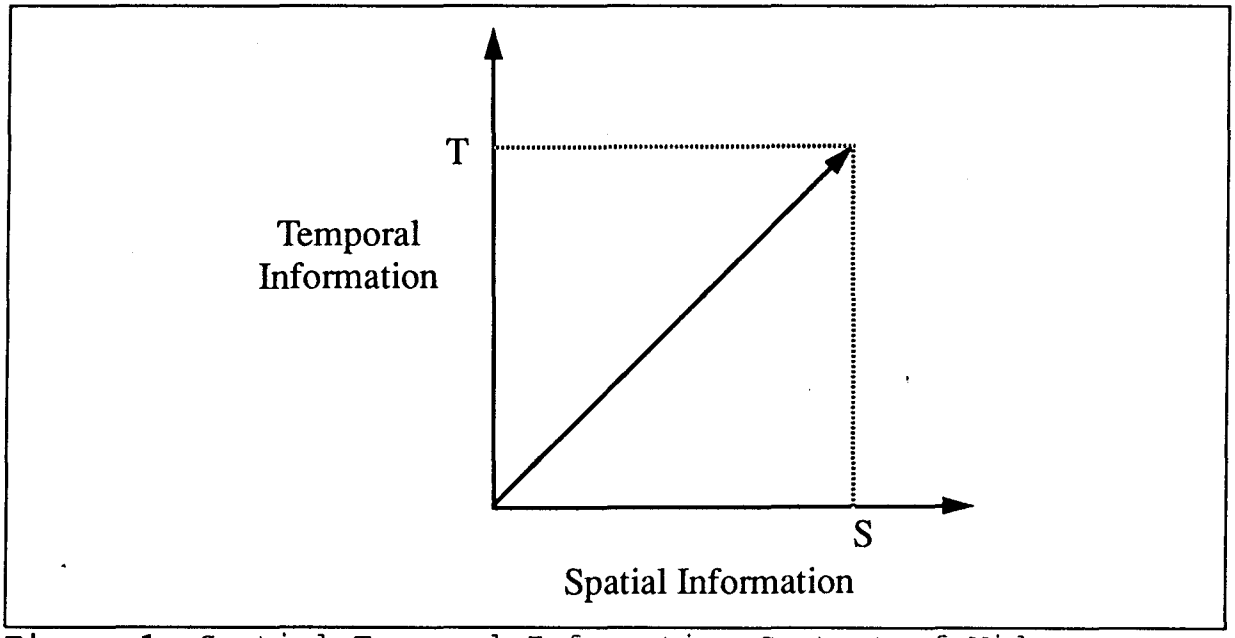

Figure 1 Spatial-Temporal Information Content of Video

If the S-T measurement is made on a source video scene and on a version of the scene which has been passed through some video system (such as a codec, tape dub, etc.) then the distance between the vectors can serve as a measurement of impairment which, hopefully, will correlate well with human perceptions of video quality. The components of the distortion vector ($\Delta S, \Delta T$), as shown in Figure 2, provide separate measures of spatial and temporal distortion.

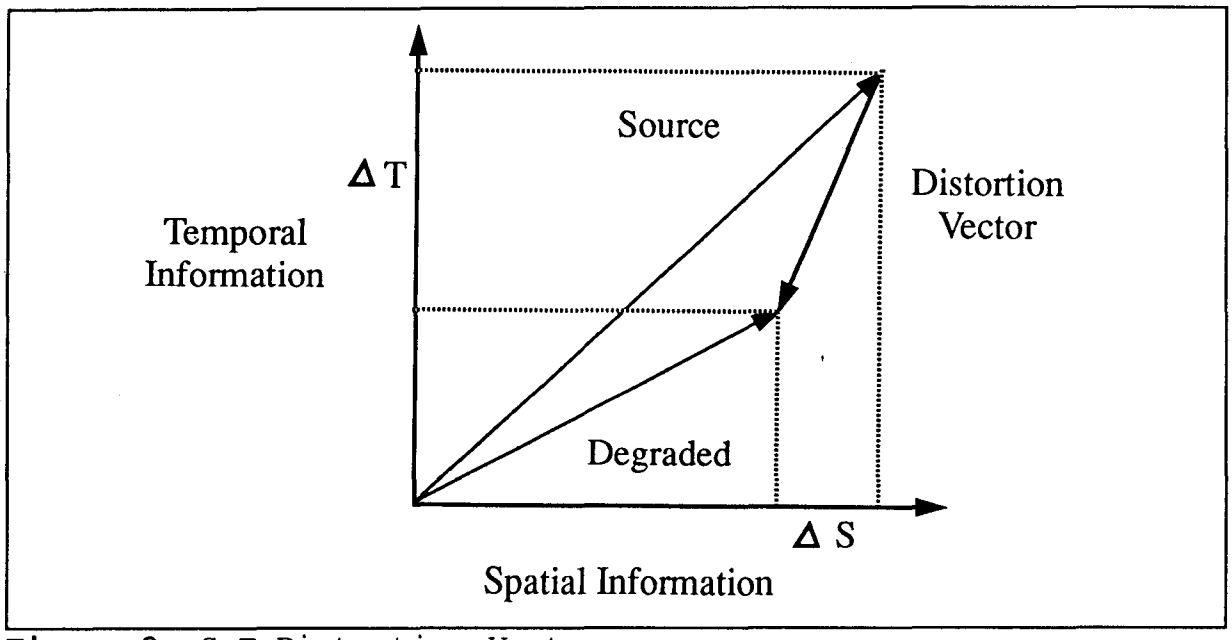

Figure 2 S-T Distortion Vector

This contribution will present spatial information content measures (section 2.1), temporal information content measures (section 3.1), spatial distortions measures between input and output video (section 2.2), and temporal distortion measures between input and output video (section 3.2). The measures presented here complement work on other video quality measures that have been presented to the T1Q1.5 VTC/VT sub-working group.

2. Spatial Measures

Several classes of spatial parameters are being considered for locating video scenes in the spatial-temporal information matrix (see Table 1, T1Q1.5/91-132 R1). This contribution will focus primarily on the class of spatial parameters based upon the radial average of the magnitude of the 2-dimensional Fast Fourier Transform (FFT) of the image. This class of parameters measures the spatial information content and distortion of the video. Another class of spatial parameters being considered is based upon the edge energy of the image (i.e., as output by the Sobel filter). Since edges contain high frequency information, this class of parameters also measures the spatial information content and distortion of the video.

For the radial average FFT parameters, a video frame, 756 pixels x 486 lines, is trimmed to a frame size of 726 pixels x 450 lines (This approximates the displayed region of an NTSC video frame). This trimmed frame is then subdivided into six slightly overlapping subregions of 256x256 pixels each (see Figure 3). The reason for

sub-dividing the image is to provide square regions (for the FFT and radial average), and to select the sub-region with the highest information content (currently determined by parameter SI2 below). The sub-region with the highest information content is the most likely to suffer the greatest distortion after passing through a video codec.

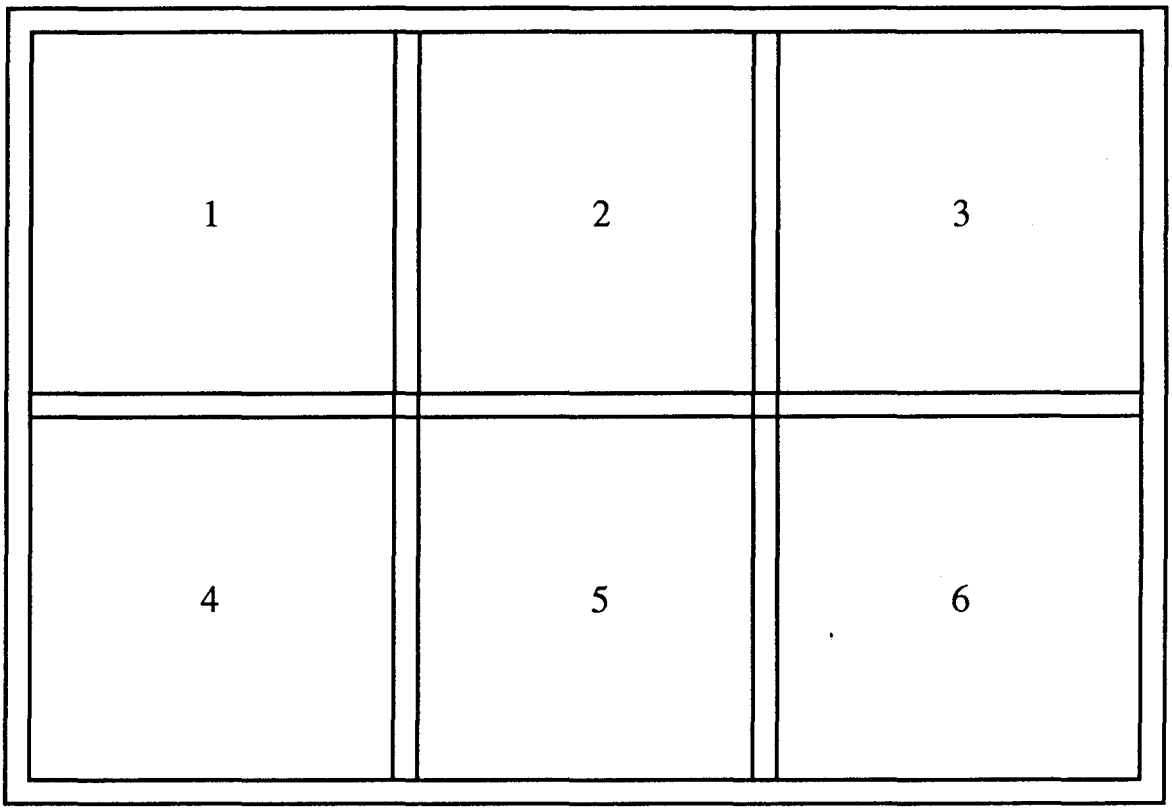

Figure 3 Image sub-regions

Each subregion is Fast-Fourier-Transformed and the quadrants shifted so that the direct current (DC) term is at the center of the resulting two-dimensional complex array. The magnitude of the 2 dimensional Fourier spectrum is then averaged radially to result in a vector of 127 points (DC term is discarded). The radial averaging produces a vector, each element of which is the average of the pixel values which lie within a circular band of inner radius r_i and outer radius r_{i+1} ($1<=i<=127$) as measured from the center of the array. See Figure 4 for a diagram of the radial averaging process.

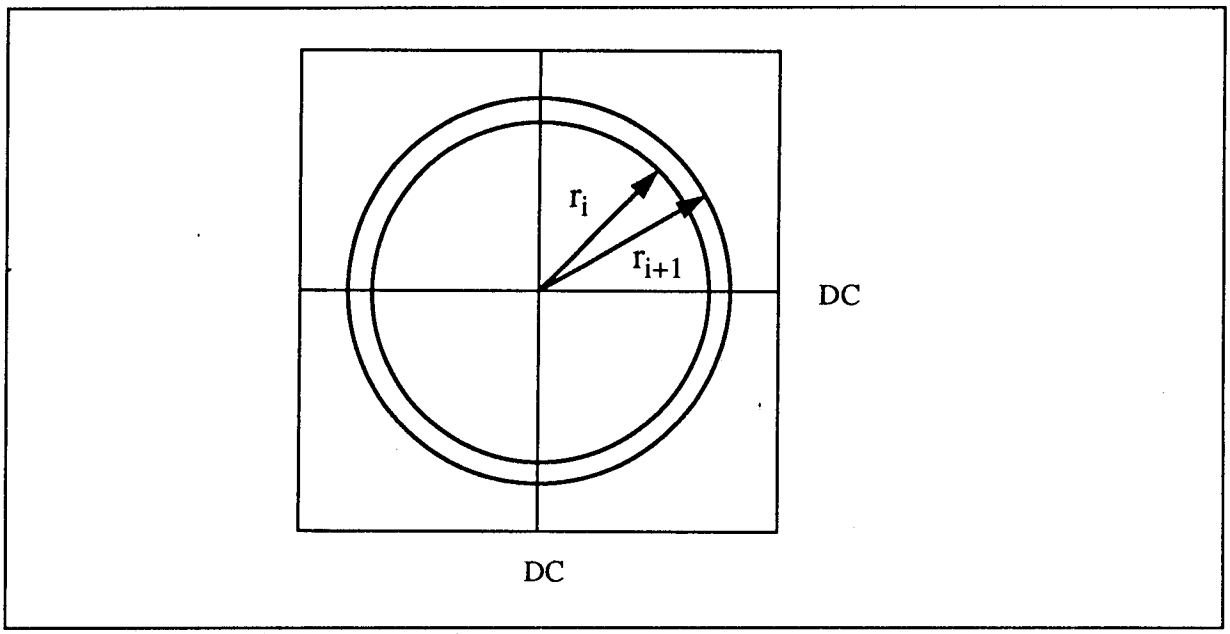

Figure 4 Radial Averaging of 2-D FFT Spectrum

$\mathbf{R}_k(f)$ can be defined as

$$\mathbf{R}_k(f) = \textit{radial average of } |\mathcal{F}\{ I_k(x,y) \}| ,$$

where $\mathcal{F}$ denotes the 2-dimensional spatial Fourier transform and I_k is the kth subregion of the video image, I. 'f' is the bin number of the FFT. An 'f' of 127 corresponds to approximately 25 cycles per degree in our subjective viewing configuration.

$R_k(f)$ is then normalized for unit energy (not counting DC), such that

$$R_k(f) = \frac{\mathbf{R}_k(f)}{\left\{ \sum_{f=1}^{127} \mathbf{R}_k^2(f) \right\}^{\frac{1}{2}}} \quad (\textit{DC term is ignored})$$

Note: As mentioned above, only the $R_k(f)$ from the subregion which is determined to have the maximum spatial content is used, hence the subscript, k, will be dropped.

Figure A1 shows plots of R(f) for 3 different scenes. One can see from Figure A1 how different scenes distribute spatial energy. As noted in Figure A1, scene Boblec has the highest spatial

information content (due to the detailed chalkboard diagram). The increase in energy at the high frequency end of the curve for scene Steve1 represents interlace energy and indicates motion (a zoom in this case).

Figure A2 shows R(f) for an original scene (Boblec) and 3 degradations of the same scene. Figure A2 illustrates the spatial frequency distortions produced by different video systems. Note that Codec B (384 Kbps) has lost more high frequency spatial information than Codec I (45 Mbps). Also note that RF transmission noise has added high frequency components to the original. Thus, blurring shows up as a decrease in spatial frequency content and noise as an increase. Both are to be viewed as impairments.

2.1 Spatial Information Content Measures

To calculate measures of spatial information content, R(f) is weighted by f^3 ($1 <= f <= 127$) and integrated (summed) over four spatial frequency bands. The pre-weighting before summation tends to flatten the spectrum by weighting higher frequency components more heavily. Figures A3 and A4 show the effects of f^3 spectral weighting on Figures A1 and A2, respectively.

The integration bands are as follows:

Low Band: $1 <= f <= 5$,
Mid Band: $6 <= f <= 80$,
High Band: $81 <= f <= 127$,
Total Band: $1 <= f <= 127$.

The four spatial information (SI) content measures, SI1, SI2, SI3, and SI4 are computed as,

$$SI1 = \sum_{f=1}^{5} R^2(f) \cdot f^3$$

$$SI2 = \sum_{f=6}^{80} R^2(f) \cdot f^3$$

$$SI3 = \sum_{f=81}^{127} R^2(f) \cdot f^3$$

$$SI4 = \sum_{f=1}^{127} R^2(f) \cdot f^3$$

The mid band parameter (SI2) emphasizes spatial frequencies perceived by the viewer since neither interlace effects or very low frequencies are very noticeable to the human viewer. The high band contains interlace energy and hence, high values of SI3 correlate

well with the presence of motion. The low band contains information which might be significant in combination with other parameters. SI4 integrates over the entire band and might be used as measure of spatial frequency content as seen by hardware.

For initial tests, the spatial information content measures have been calculated every tenth frame. Therefore SI1, SI2, SI3, and SI4 are functions of time where the time interval is one third of a second. The time histories will be further processed to obtain a set of scalar values. These will be combined with other parameters and tested for correlation with subjective quality scores.

2.2 Spatial Distortion Measures

Spatial distortion (SD) measures are conceptually based on the change in spatial information Measures and are represented in Figure 2 as ΔS. This class of spatial distortion measures is calculated by integrating (over the same bands as spatial information measures above) the difference between the logarithms of the R(f) for the source image and the R(f) for the degraded image. The positive values of this difference is an indicator of blurring. The negative values of the difference is an indicator of additive energy such as blocking or noise. Therefore the positive and negative portions are accumulated separately. They are defined as,

$$SD1 = \sum_{f=1}^{5} \begin{cases} [\, LOG(R_O^2(f)) - LOG(R_D^2(f)) \,] & , \; if \; [\,\cdot\,] > 0 \\ 0 & , \; otherwise. \end{cases}$$

$$SD2 = \sum_{f=1}^{5} \begin{cases} -[\, LOG(R_O^2(f)) - LOG(R_D^2(f)) \,] & , \; if \; [\,\cdot\,] < 0 \\ 0 & , \; otherwise. \end{cases}$$

where $R_O(f) = R(f)$ for the original scene and $R_D(f) = R(f)$ for a degraded scene.

SD3 and SD4 are the same as SD1 and SD2, except integrated over the mid band. Similarly, SD5 and SD6 are integrated over the high band, and SD7 and SD8 are integrated over the entire frequency band.

As with the spatial information content measures, the spatial distortion measures are time histories with a time interval of one third of a second. Figure A5 shows the time history of SD3 and 10*SD4 for scene Steve1, Codec A (384 Kbps). Note at about 8 seconds the positive difference parameter, SD3 (which is an indicator of blurring), decreases and the negative difference parameter, SD4 (which is an indicator of additive energy such as blocking) increases. This point in the video sequence is where a fairly fast zoom takes place.

The time histories will be further processed to obtain a set of scalar values which will then be tested, in combination with other parameters, for correlation with subjective scores.

3. Temporal Measures

The term 'temporal' herein refers to the video's representation of the natural flow of time. This class of temporal measures is based upon the motion difference output feature, $M_{ij}(n)$. $M_{ij}(n)$ is the difference between the pixel values at the same location in space but at successive times or frames. $M_{ij}(n)$ as a function of time (n) is defined as,

$$M_{ij}(n) = I_{ij}(n) - I_{ij}(n-1)$$

where $I_{ij}(n)$ is the pixel of image I at the ith row, jth column, and nth frame in time.

3.1 Temporal Information Content

Temporal information content, as defined here, increases with increased 'motion' in the video sequence. Motion in this sense refers to any change in pixel values such as a pan, zoom, subject motion, etc. or any combination of these.

One measure of temporal information content, TI(n), is computed as the standard deviation of $M_{ij}(n)$ over all i and j for a given n.

$$TI(n) = STD[M_{ij}(n)]$$

$$TI(n) = \left\{ \frac{1}{NR \cdot NC} \sum_{i=1}^{NR} \sum_{j=1}^{NC} [M_{ij}(n)]^2 - \left[\frac{1}{NR \cdot NC} \sum_{i=1}^{NR} \sum_{j=1}^{NC} M_{ij}(n) \right]^2 \right\}^{1/2}$$

where NR is the number of rows and NC is the number of columns.

Figure A6 shows TI(n) plotted for two source video sequences, 5row2 and Mntbik. Scene 5row2 contains two pans, which are evident at higher values of TI(n), and, in contrast, scene Mntbik has very little motion.

When TI(n) is plotted for various degraded video scenes, temporal impairments become evident. Figure A7 shows the scene 5row2 passed through two codecs: **a**) codec A at 1536 Kbs and a bit error rate (ber) of 10^{-5} and **b**) codec B at 56 Kbs. Note the obvious effects of frame repetition on TI(n) for the two codecs.

3.2 Temporal Distortion Measures

Temporal distortion (TD) measures are conceptually related to the change in temporal information content (ΔT) as shown in Figure 2. One class of TD measures are calculated from the time histories, TI(n). From TI(n), the maximum frame repetition rate (max rep rate) can be calculated. A jerkiness measure and an overall temporal distortion measure can be also be estimated for the sequence as a whole.

Figures A8 and A9 show the same codec configurations as above, along with a plot of TD1(n), which is the positive portion of the difference between TI(n) for the source scene and the TI(n) for the degraded scene. When TD1(n) is greater than zero, the degraded has less motion than the source. The equation for TD1(n) is

$$TD1(n) = \begin{cases} [TI_O(n) - TI_D(n)], & if\ [\cdot] > P_N \\ 0, & otherwise \end{cases}$$

where TI_O and TI_D denote TI for the original source scene and the degraded scene, respectively, and P_N is the system noise level.

One temporal distortion measure can be calculated by integrating each pulse (given in Figures A8b and A9b), selecting the pulse with the most area, and multiplying that area by its width (i.e., repetition rate). It is clear that the temporal distortion measure will be higher for codec B at 56 Kbs than for codec A at 1536 Kbs.

The negative portion of the difference between TI(n) for a source sequence and a degraded sequence is also useful for estimating jerkiness. This quantity multiplied by -1 is defined as TD2(n) and is given by,

$$TD2(n) = \begin{cases} -[TI_O(n) - TI_D(n)], & if\ [\cdot] < 0 \\ 0, & otherwise \end{cases}$$

When TD2(n) is greater than zero, the degraded has more motion than the source.

Again, further processing of the time histories of TD1(n) and TD2(n) yields single numbers for each scene-degradation pair. These numbers are then combined with other parameters to determine correlation with subjective scores.

4. Conclusions

We have discussed some of the information content measures (SI and TI) that may be used for determining the location of a video scene

in the spatial-temporal information matrix (as given in Table 1, T1Q1.5/91-132 R1). The location of the video scene within the spatial-temporal matrix is important because the quality of the transmitted video scene is highly dependent on this location. We have also discussed a set of distortion measures (SD and TD) which can be used for determining the spatial-temporal error vector associated with degraded video scenes. The measures presented here complement work on other video quality measures that have been presented to the T1Q1.5 VTC/VT sub-working group.

ITS is currently in the process of selecting a small optimal set of video quality parameters. Recommendations to the T1Q1.5 VTC/VT sub-working group for inclusion of this set of video quality parameters into the draft VTC/VT standard will be forthcoming.

The work presented in this contribution is the result of the efforts of the Video Quality Project team (NTIA/ITS.N3) consisting of Steve Wolf, Arthur Webster, Steve Voran, Coleen Jones, Margaret Pinson, and Paul King.

Figure A1: Radial Averaged Spectra of 3 Scenes

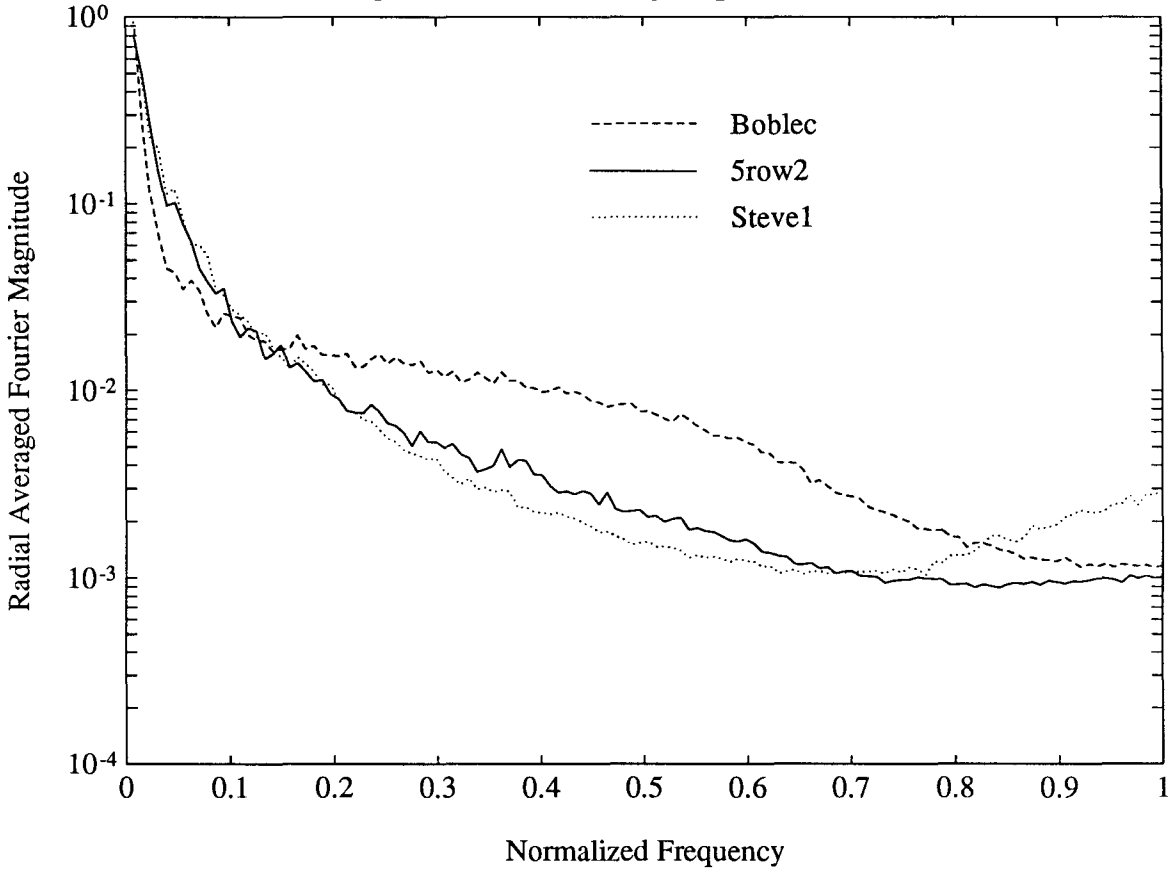

Figure A2: Radial Avg. Spectra of Boblec Source & 3 Degs.

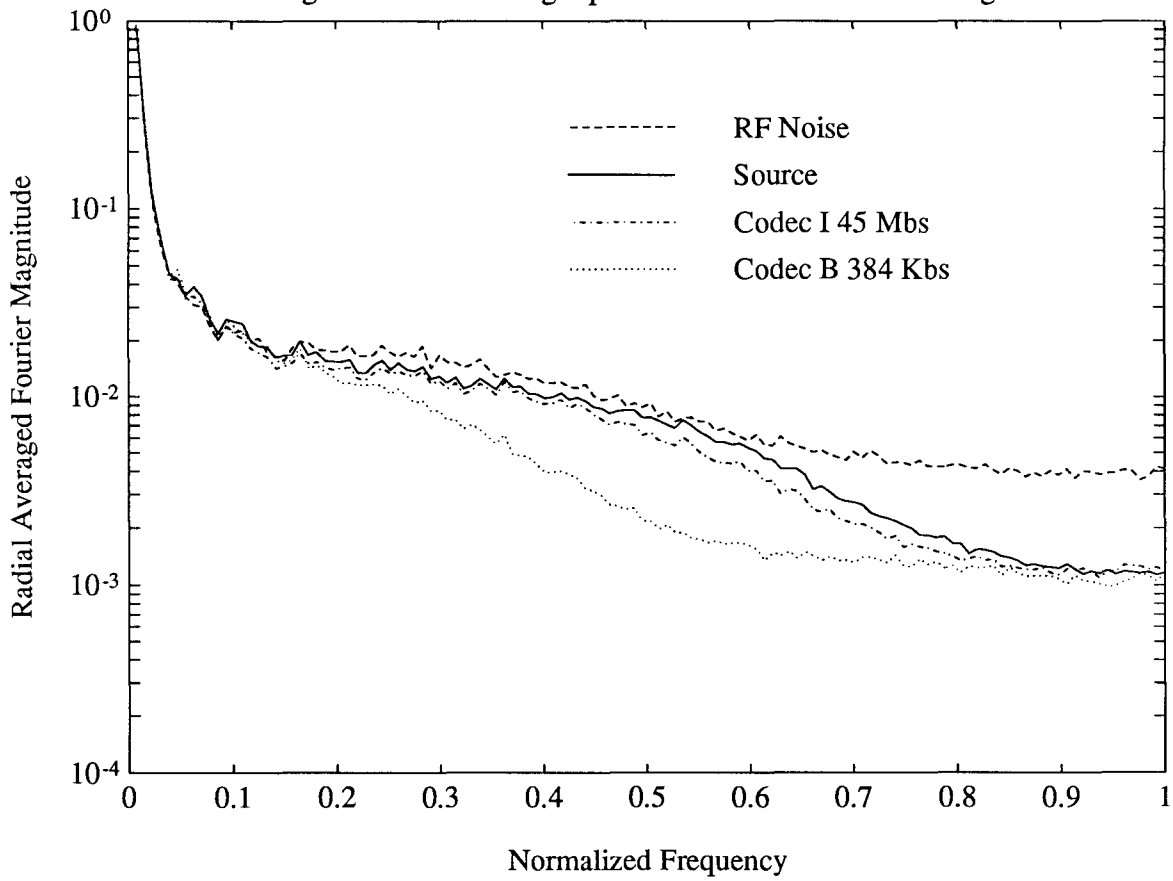

Figure A3: Radial Avg. Spectra Weighted by f^3

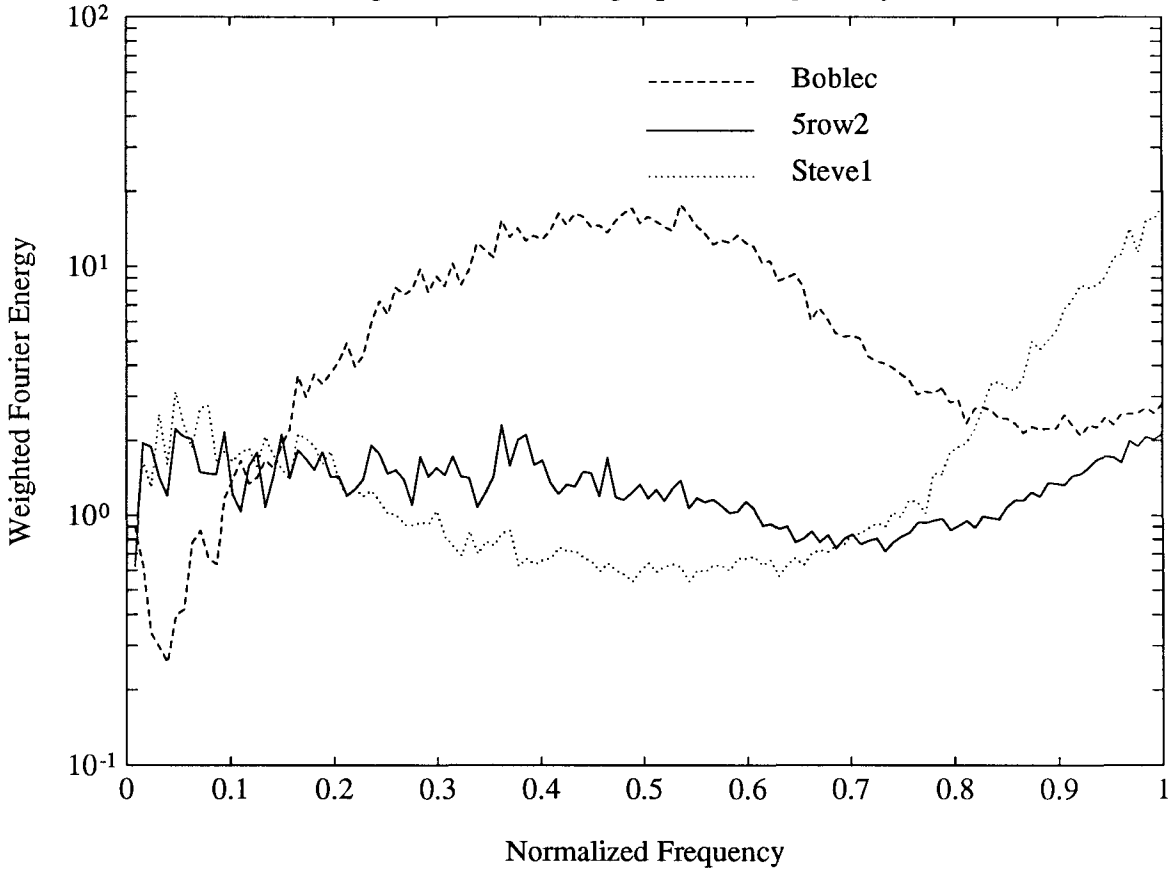

Figure A4: Radial Avg. Spectra Weighted by f^3

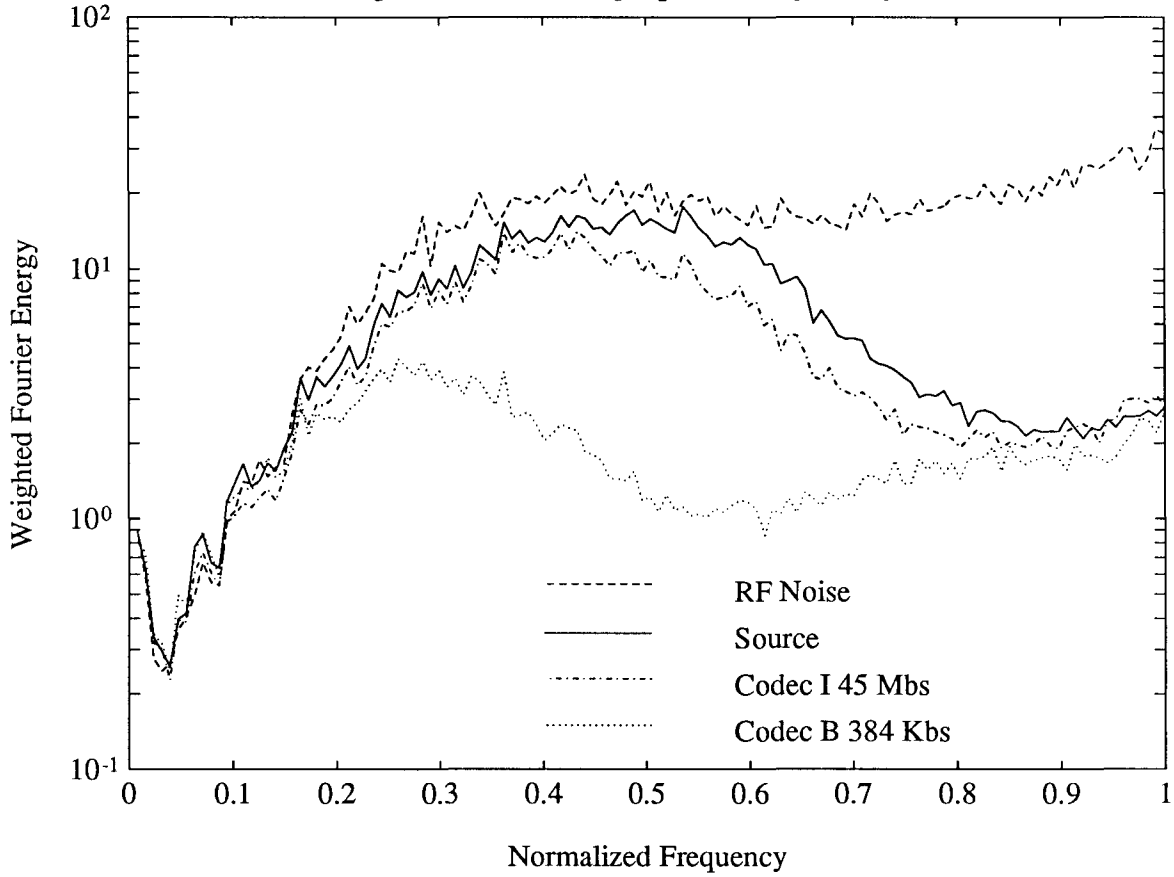

Figure A5: Two Spatial Measures for Steve1: Codec A 384 Kbs

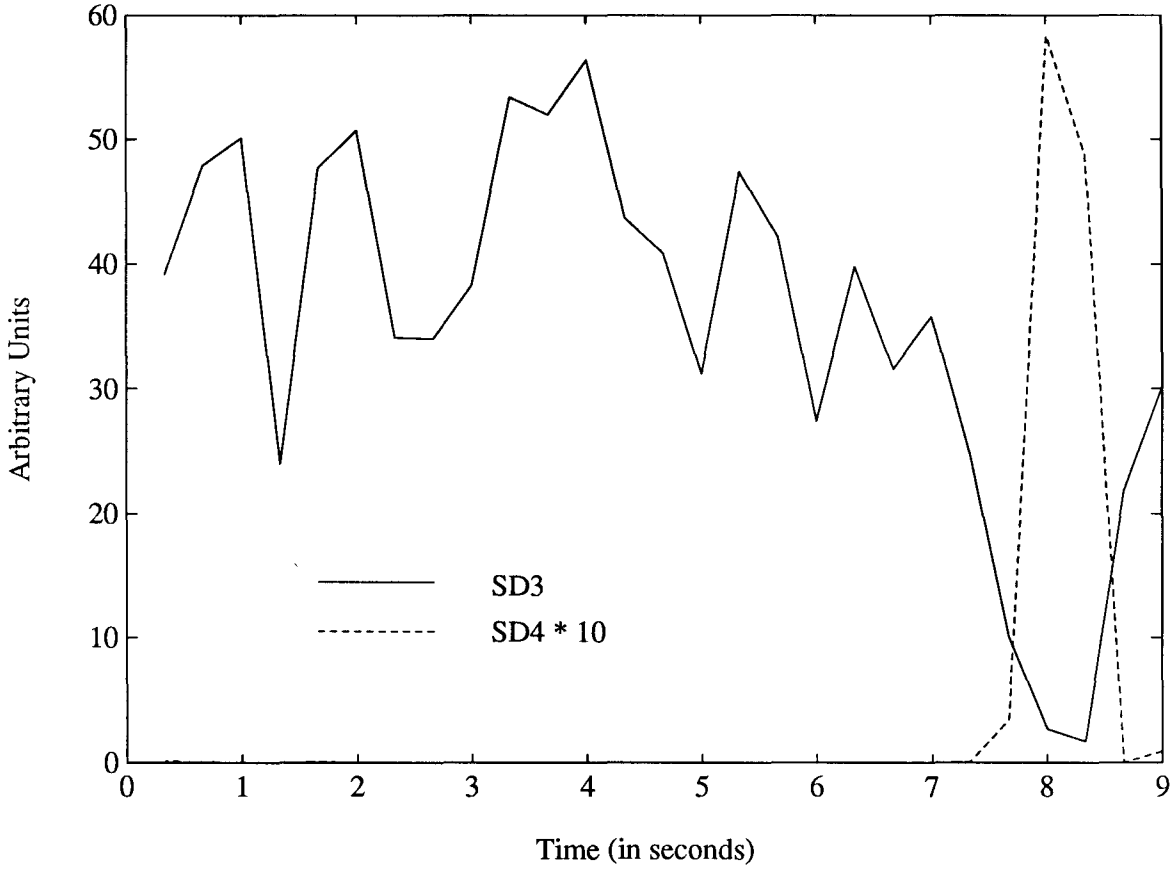

Figure A6a: TI(n) for 5row2-Source

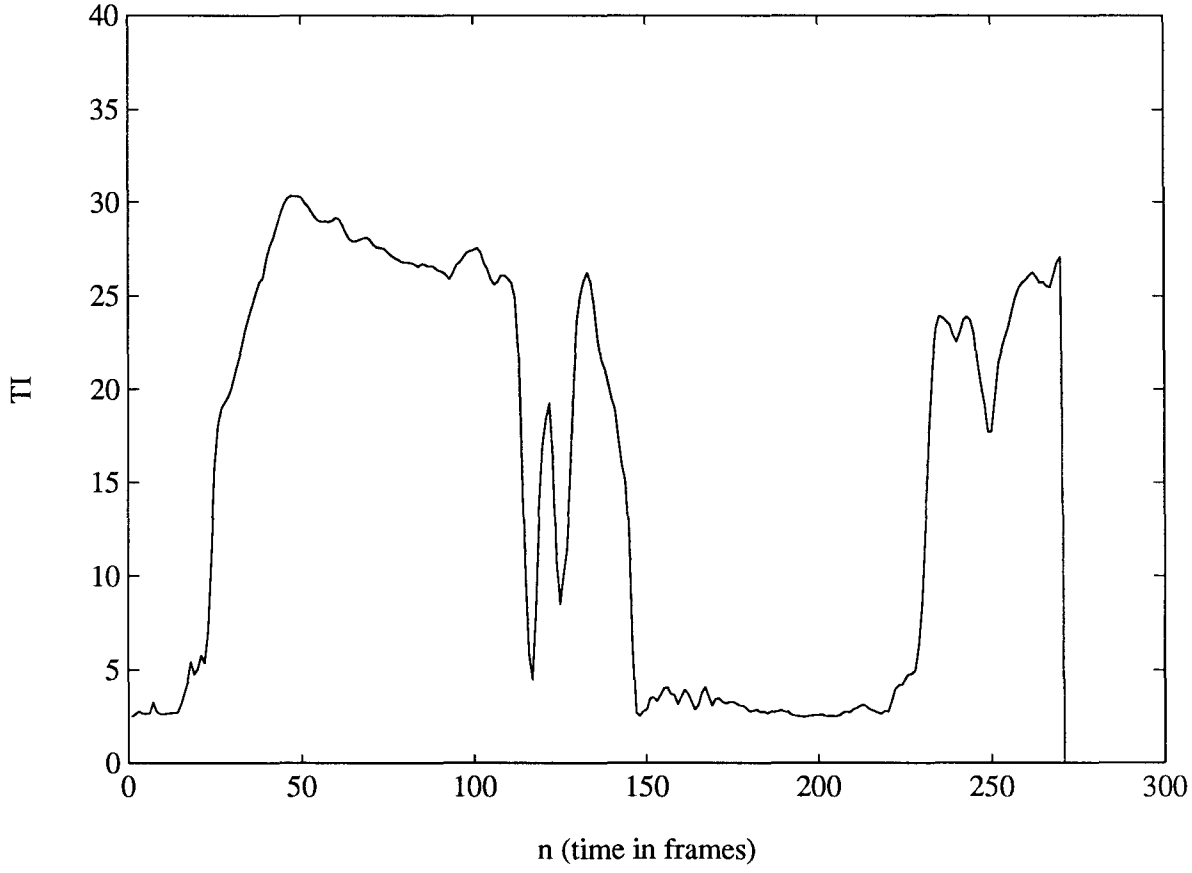

Figure A6b: TI(n) for Mntbik-Source

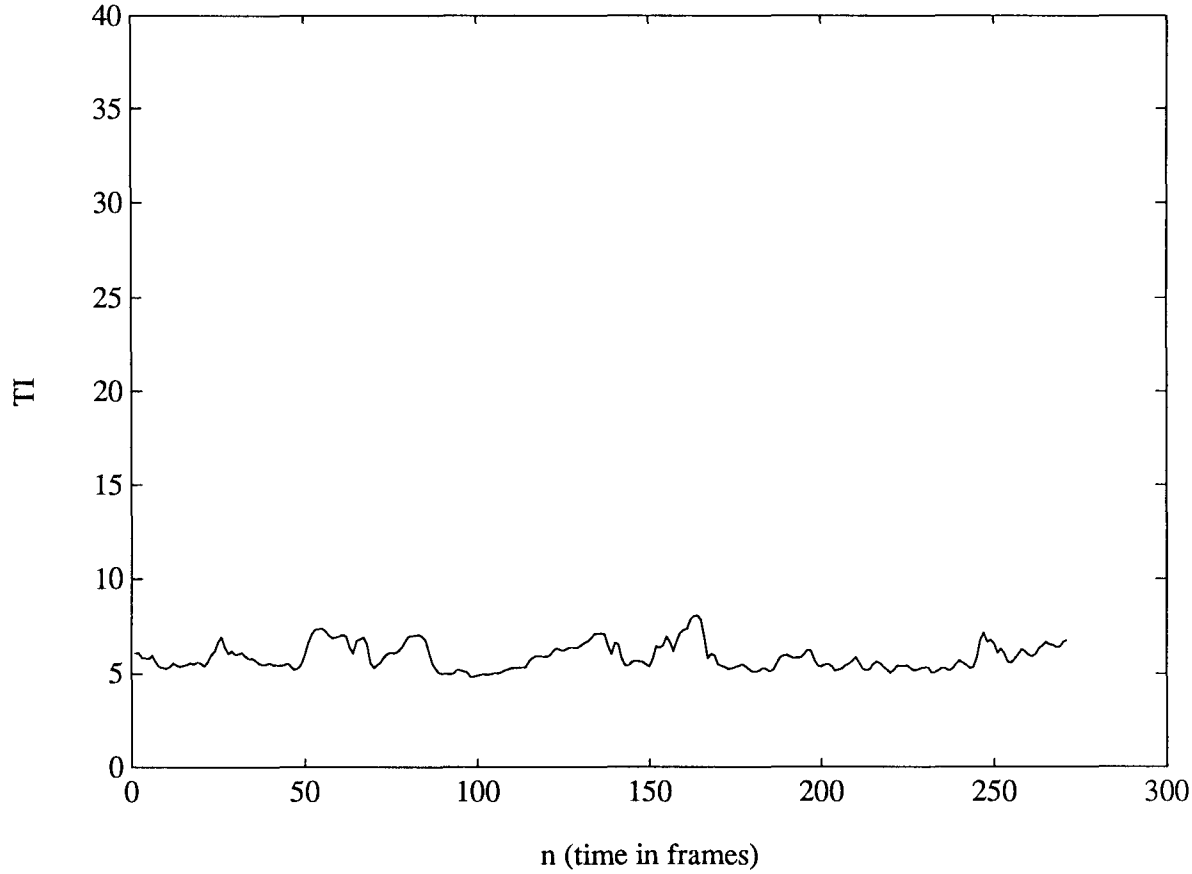

Figure A7a: TI(n) for 5row2: Source & Codec A T1 ber 10^-5

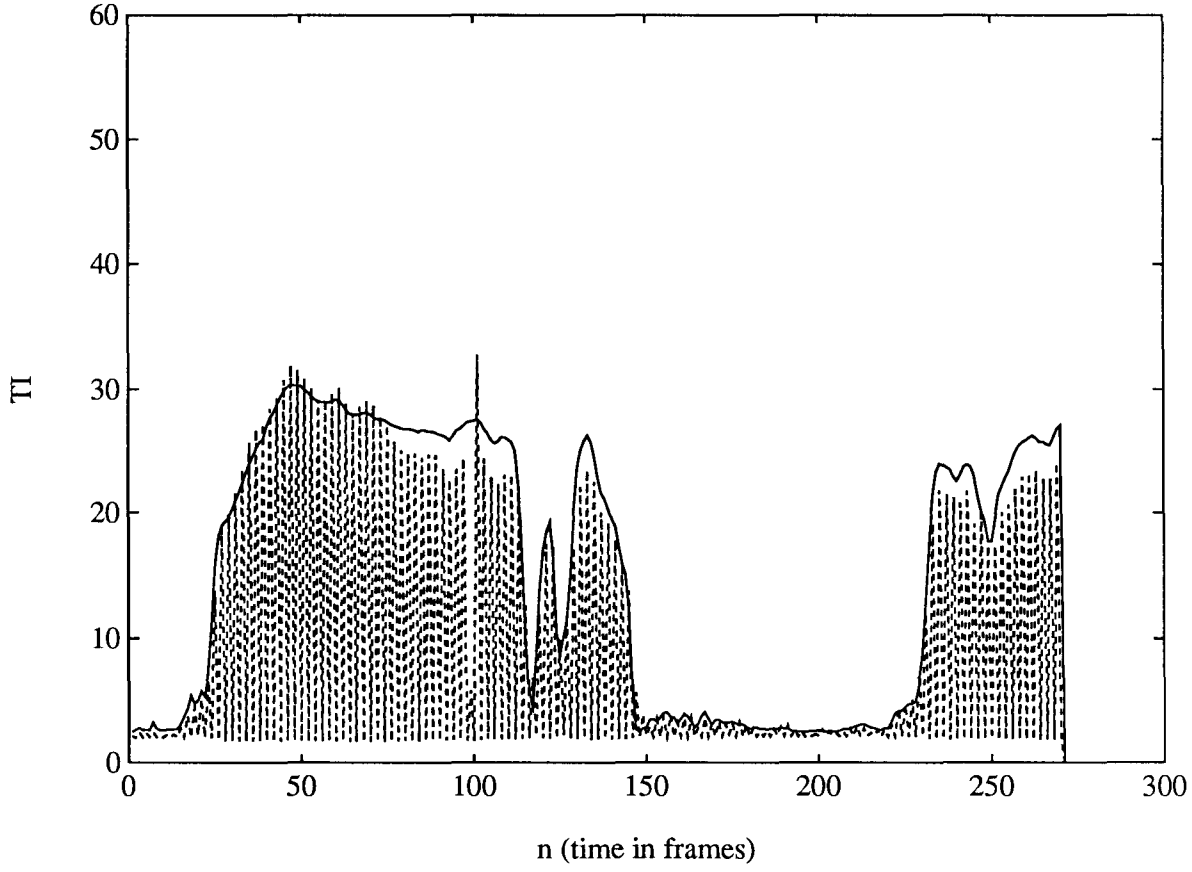

Figure A7b: TI(n) for 5row2: Source & Codec B 56 Kbs

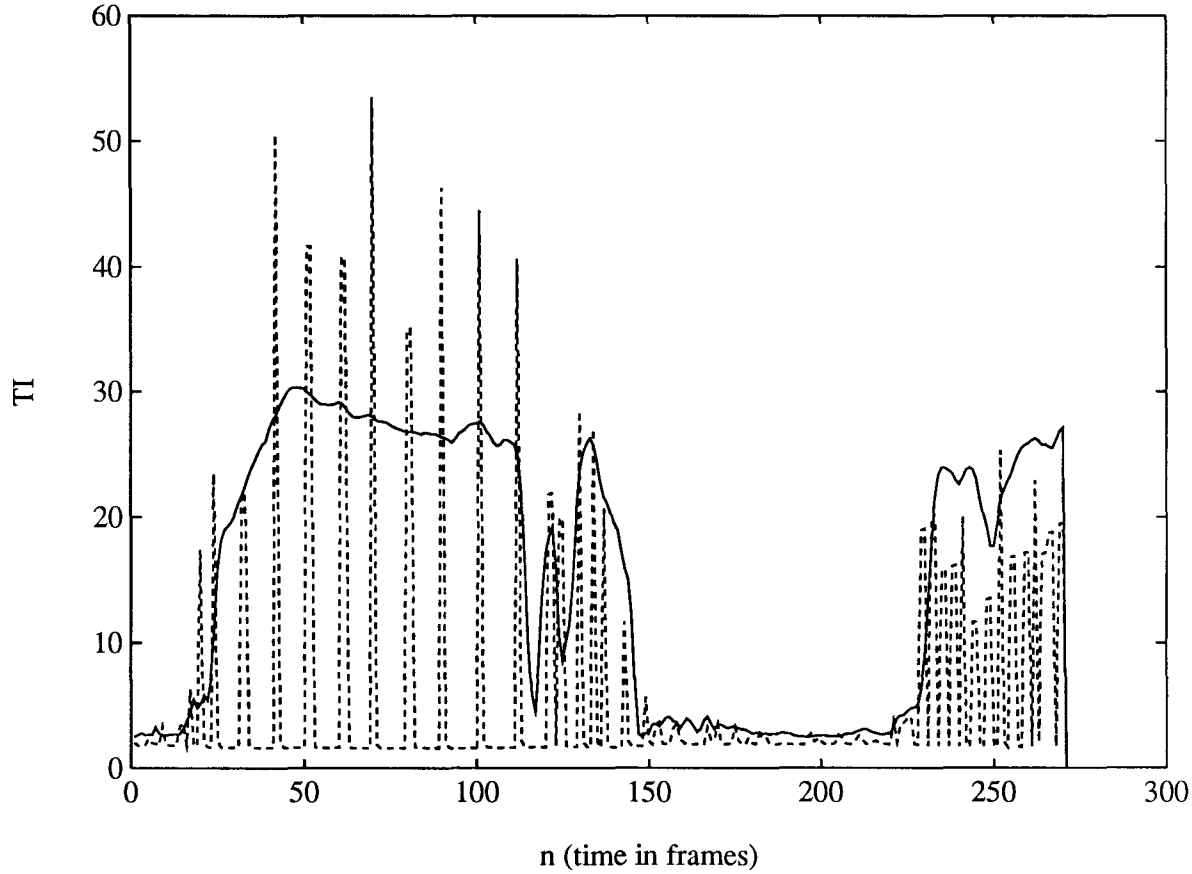

Figure A8a: TI(n) for 5row2: Source & Codec A T1 ber 10^-5

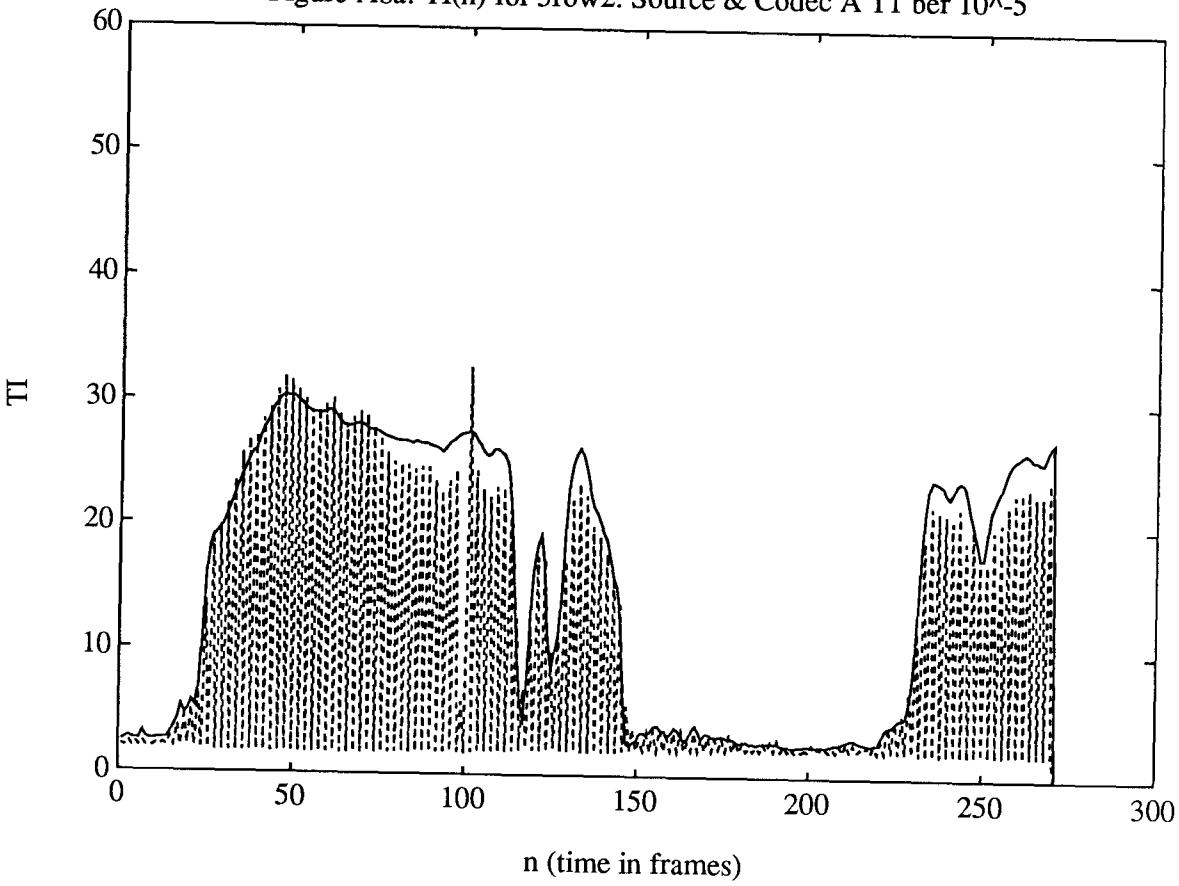

Figure A8b: TD1(n) for 5row2 - Codec A T1 ber 10^-5

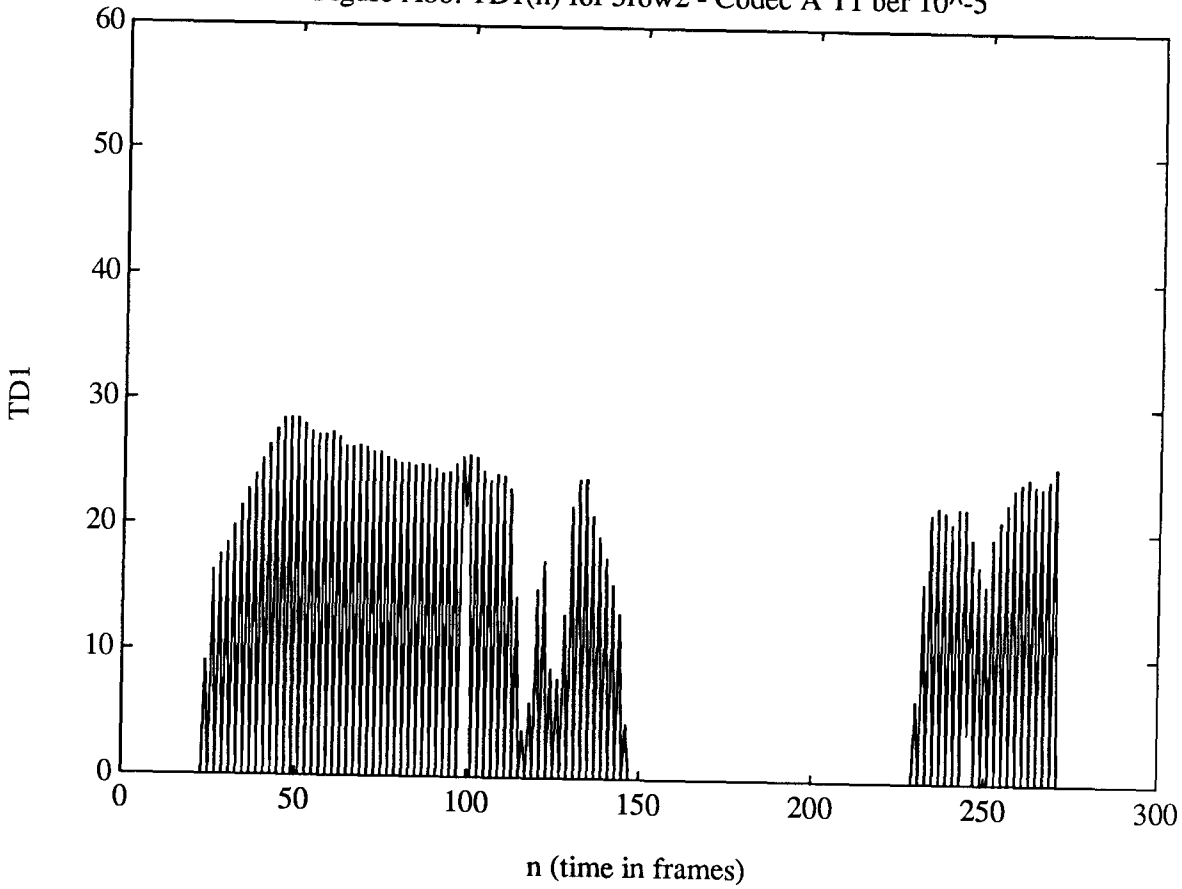

Figure A9a: TI(n) 5row2: Source & Codec B 56 Kbs

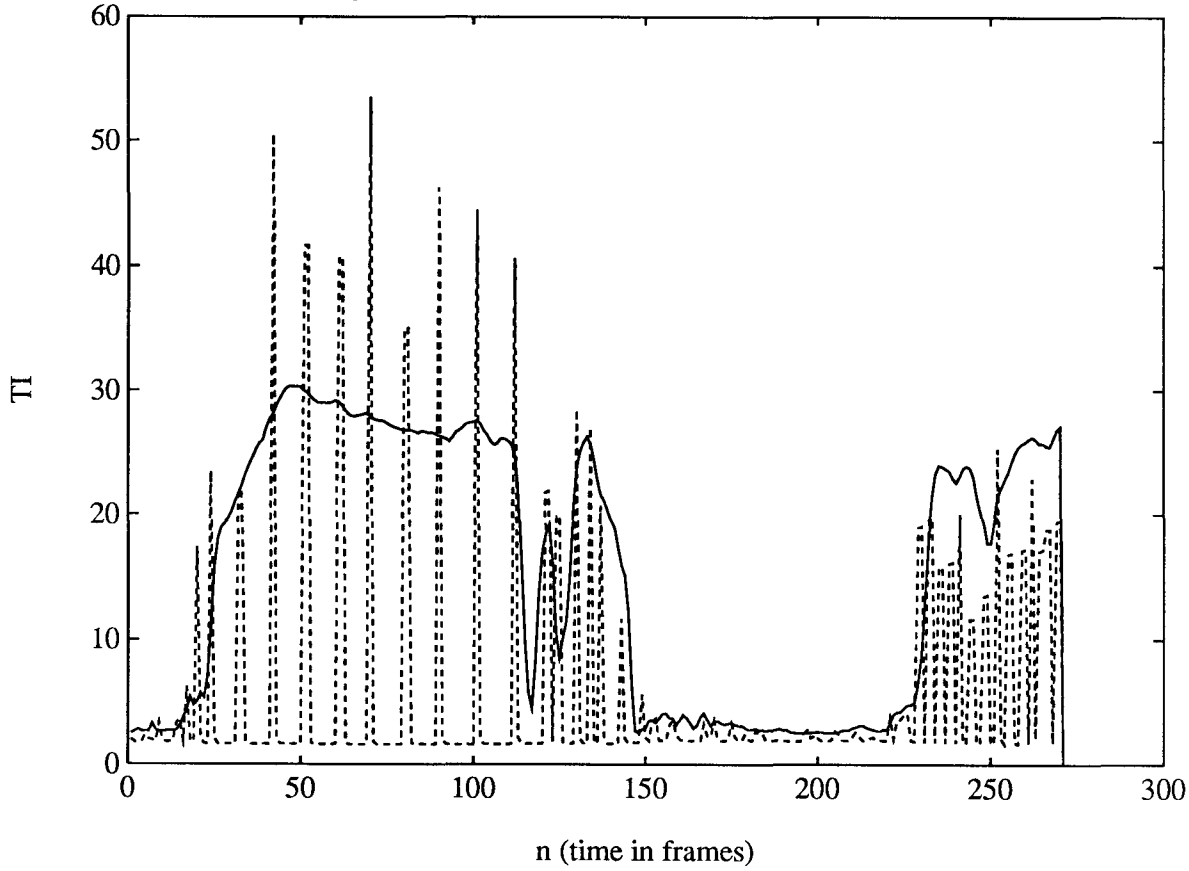

Figure A9b: TD1 for 5row2 - Codec B 56 Kbs