Voice Activity Statistics*

Gopi K. Vaddi, D.W.Petr

Information and Telecommunication Technology Center
Department of Electrical Engineering & Computer Science
The University of Kansas
2291 Irving Hill Dr.
Lawrence, KS 66045-2228

ITTC-TR-FY2000-15664-03

August 1999

*This work is funded by Sprint Corporation

1 Introduction

Human voice intrinsically consists of talk spurts and silence periods. The length of talk spurts and silence periods depends on various factors, the most important of which is the type of talk. For example, conversations tend to have more silence than uni-directional speech and scripted speech has lesser silence than unscripted speech. We focus here on unscripted, conversational speech typical of telephone calls. During most telephone conversations only one person speaks at a time and thus there is a lot of waiting silence time at each user end. Thus if the silence periods in between the talk can be detected and only the active speech is sent across a communication network, there would be a considerable gain in terms of network bandwidth. This gain will be prominent especially in the case of telephony over data networks in which statistical multiplexing gain is exploited. Here arises a need to find the nature of On-Off speech patterns in telephone conversations. This would also help in modeling of voice sources for simulation purposes.

2 Related Work and Problem Statement

2.1 Definitions of Terms Used

Figure 1 shows the speech pattern consisting of talkspurts and silence periods in a two way telephone conversation. The critical terms used in the rest of this report on voice statistics are defined below.

1.On-time: The length of an individual talk spurt in time units.

2.Off-time: The length of an individual silence period in time units.

3.Mean On-time: A simple mean of all On-times (talkspurts) in the conversation.

4.Mean Off-Time: A simple mean of all Off-times (silence periods) in the conversation.

5.Speech Activity Factor: The ratio of the sum of all On-times to the total length of the conversation.

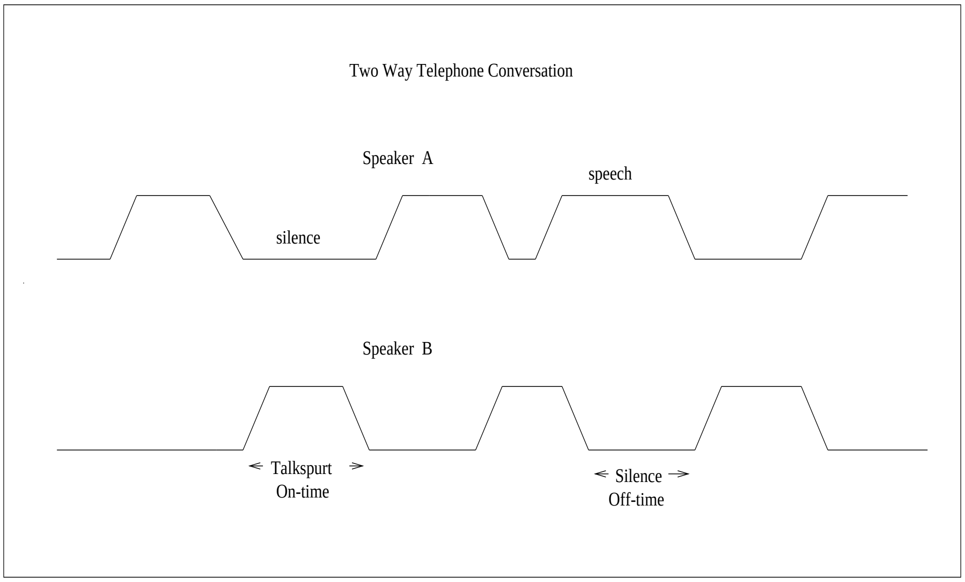

Figure 1: Two Way Telephone Conversation

It can also be obtained as the ratio of the mean On-time to the sum of mean On-time and mean Off-time in the conversation.

2.2 Related Work

Research in this field was triggered by Paul T. Brady with his classical Bell System Technical Journal papers [1][2]. These papers remain an important reference for all subsequent work in voice activity detection. Though we have only a few published works in this area, there has been a lot of variation (see Table 1) in the statistics given by various works [1][3][4][5]. This is most likely because of the difference in the coding techniques employed by each and the conditions and assumptions under which the tests were performed. References [4][5] contain the critical values obtained, but they are not substantiated with research finding. Also, we do not know which of these findings would be a better approximation for current telephone network research. In this context, knowing the critical values for our own network environment assumes importance. In addition to mean values, distributions are important. An exponential model for talkspurts and silences is widely used in network analysis that involves voice transport. This model was originally proposed in [1][2] and also stated in [4], but is strongly opposed by [3] through its findings. Thus an attempt is made in this paper to find the critical values and also verify the exponential model for the On-time and Off-time distributions through mean-square error fit.

Reference	Mean On-time (s)	Mean Off-time (s)	Speech Activity
1	1.311	1.695	0.436
2	1.189	-	-
3	0.352	0.650	0.351
4	0.420	0.580	0.420

Table 1: Comparison of previous findings.

3 Technique Used for Finding Voice Statistics

3.1 Recordings

Recordings of telephone conversations were obtained from Sprint Corporation. Twelve different people were asked to speak on the telephone and one direction of each telephone conversation was recorded. Speech coded with 8 bit linear PCM at 8 KHz sampling rate was stored in Hexadecimal files. A total talk length of 248.4 minutes was analysed.

3.2 Technique

3.2.1 Frame Size and Silence Detection

Each conversation was divided into frames of size 80 samples corresponding to 10 ms conversation segments. Although a shorter frame size would have yielded more information, computation time must also be considered and a 10 ms frame size was chosen as a compromise. Each of the frames was examined for energy content and a decision was made based on an energy threshold. When 20 or more consequetive frames (corresponding to 200 ms) have energy below the threshold, it is considered as a silence period. This minimum was used to avoid clipping due to momentary interruptions in speech. The minimum silence length of 200 ms was based on [1][2][3]. The selection of frame size makes each talk spurt to be at least 10 ms long, avoiding momentary spikes.

3.2.2 Fixing the Silence Threshold

Fixing a threshold to distinguish between the talk and silence is very important as it would greatly influence the critical values that we wish to calculate. We used a fixed threshold method to eliminate silence. Though an adaptive threshold is advocated in [6], it was not required in this case because of identically recorded voice. The speech files were analysed for various thresholds. A curve of speech activity factor vs. threshold is obtained as shown in figure 2. From the figure it is clear that a threshold value near the knee of the curve would be a good choice for a threshold, because a threshold value of less than that would greatly increase the value of speech activity and a value much more than that would not substantially decrease the speech activity but might eliminate important low-energy sounds. We took values around the knee of the curve and listened to the processed (silence-eliminated) speech for various thresholds using 'Gold Wave v4.02', a digital audio editor for Windows. The lowest value in this region that is enough to obtain a good subjective quality of processed voice has been selected as the best value of threshold. The threshold value obtained by using this technique was 10^{-4} times the maximum energy level. This value was used throughout the analysis for calculating mean On-Off times.

4 Results and Discussion

4.1 Mean Values

Each of the speech files was first analysed individually based on the technique as described in section 3.2. The On-Off times and speech activity factor based on the threshold selected are tabulated for each of the files in Table 2. Table 2 shows a lot of variation in the speech charateristics of individual users. The 'average' of the mean On-times of all users was found to be 1.232 secs and average of the mean Off-times was found to be 1.373 secs. This gives a speech activity factor of 0.473. The mean values of On-Off times given above are obtained by calculating a simple mean of values obtained from individual users. The values of mean On-Off times and speech activity factors shown in figures 2, 3 and 4 are obtained for the entire duration of all the speech files. The mean values taken as the simple mean of individual users do not exactly match the values shown in figure 2, 3 and 4 due to different lengths of individual files.

All the speech files were then concatenated and analysed to give the following results.

- Speech Activity Factor vs threshold (see Fig.2) Speech activity factor decreases as threshold increases due to the increase in the number of frames that fall below the threshold.

- Mean On-time vs threshold (see Fig.3): This curve shows that the mean On-time decreases as the threshold is increased. This is because, as threshold increases, more frames fall below the threshold. In Figure 3, we see a minor deviation from this behaviour in the region pointed by the arrow. This is because, in the region marked, as the threshold is increased, some of the shorter length noise falls below the threshold and gets eliminated. The elimination of such shorter length noise results in the increase of average talkspurt length.

- Mean Off-time vs threshold (see Fig.4). The mean Off-time increases with the threshold.

- Histograms of On-time and Off-time distributions (see Fig.5 and Fig.8)

These curves give us information about the distribution of On-time and Off-time. The bin size taken in the histogram is 10 frames corresponding to a time length of 100 ms.

User	Mean On-time	Mean Off-time	Speech Activity
1	1.25	1.82	0.41
2	0.82	1.21	0.40
3	1.2	1.4	0.46
4	0.58	1.57	0.27
5	1.15	1.38	0.45
6	1.36	1.28	0.51
7	1.4	0.98	0.59
8	0.8	2.6	0.24
9	1.42	0.9	0.61
10	1.3	0.82	0.61
11	1.25	1.45	0.46
12	2.25	1.07	0.68

Table 2: Statistics of all users

4.2 Models for On-time and Off-time distributions

Two models were considered for the On-time distribution: an exponential model of the form xe^{-xt} for $t \geq 0$, and an alternative model that deviates from exponential of the form $ye^{-xt} + z$ for $0 \leq t \leq 20s$. A mean-square error fit was done and the alternative form was found to be

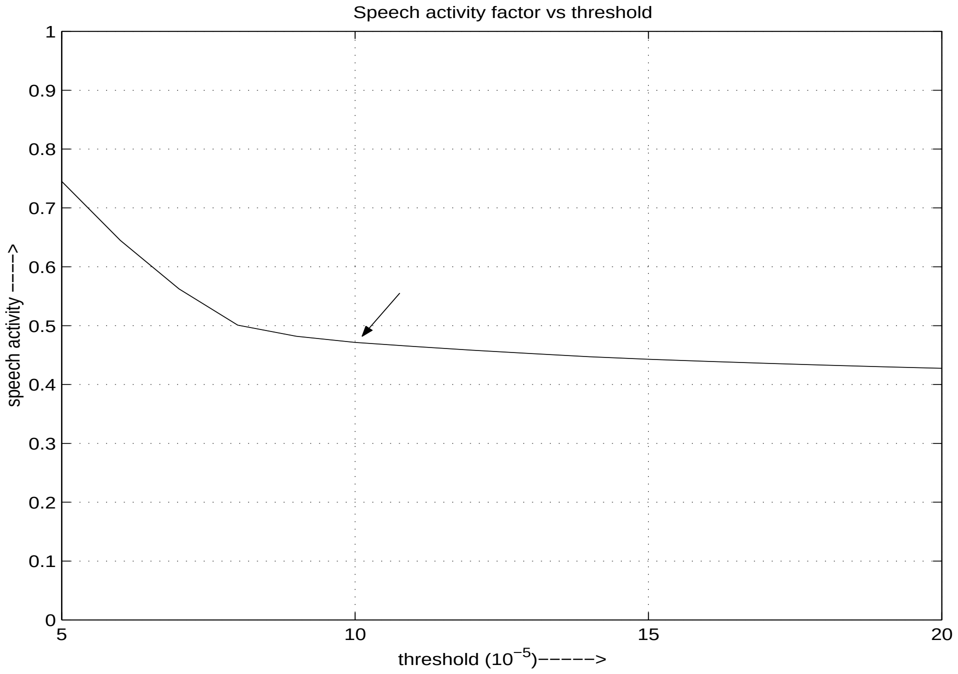

Fig.2 Speech Activity vs threshold for all users

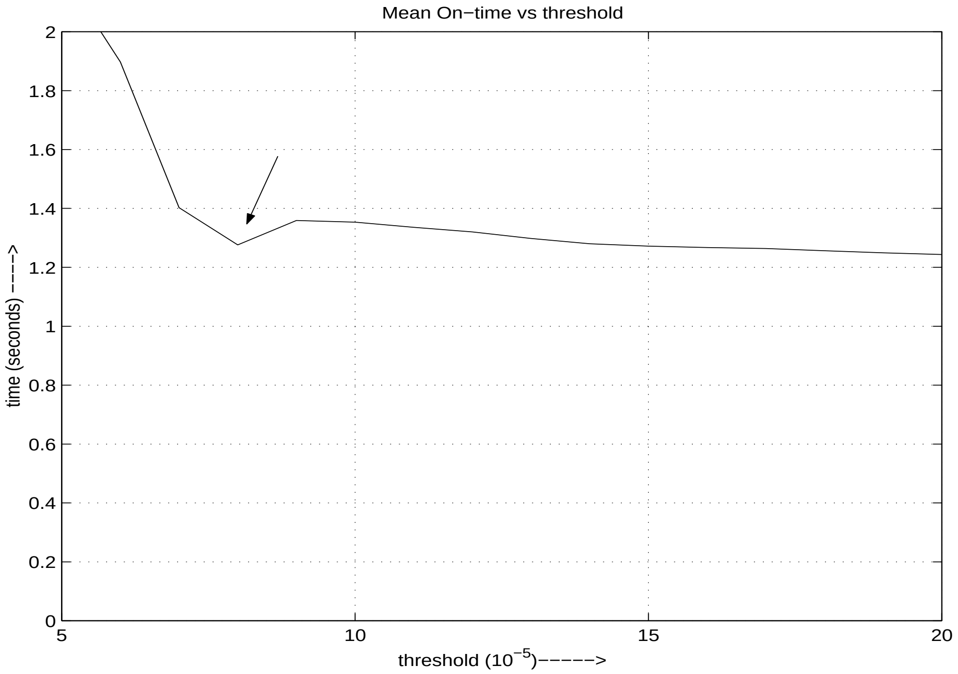

Fig.3 Mean On-time vs threshold for all users

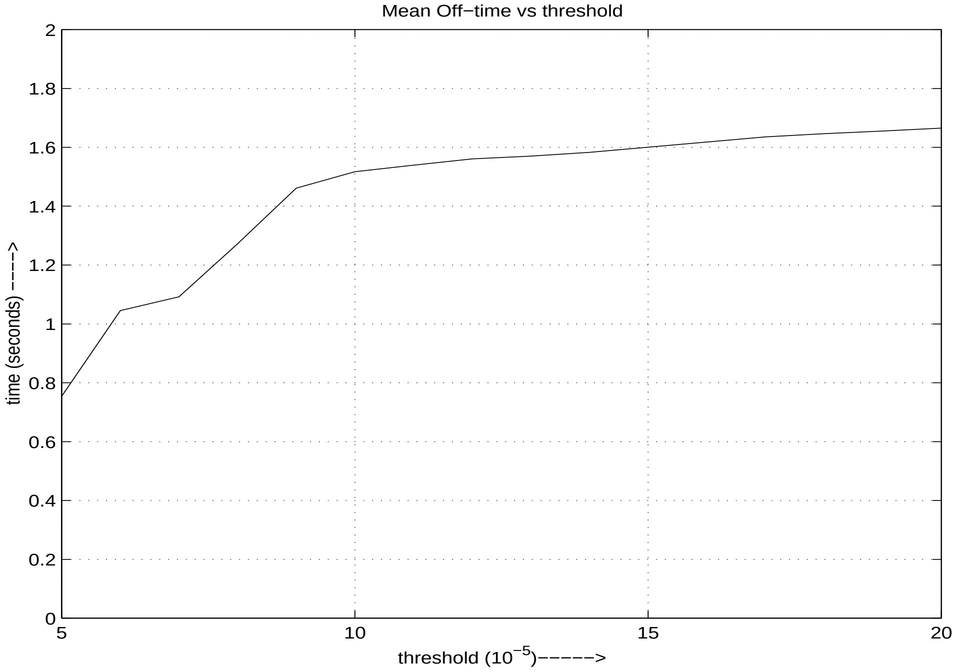

Fig.4 Mean Off-time vs threshold for all users

having a slightly better fit than the exponential model. The mean-square error value obtained using the exponential model was 0.0055 and using the alternate model was 0.0047. The value $x = 1.1480$ gave the closest exponential fit and the values $y = 1.2290$, $x = 1.5480$, $z = 0.0103$ gave the closest fit of the alternative form. A brute force method was employed to find the values corresponding to the best fit for both the models. Both the fits match the real curve fairly well along the body of the curve, but the alternate model matches better along the tail of the curve. These curve fits derived out of the On-time distribution over the entire length of all conversations is shown over the On-time distribution curves for users 3 and 7 in figures 6 and 7.

The same models were also considered for mean Off-time distribution, but the range of the alternate model was changed to $0.2 \leq t \leq 20s$. This was done because the mimimum length of silence period is 200ms. The value of $x = 0.8090$ gave the closest exponential fit with a mean square error of 0.0241, and the values $y = 2.1970$, $x = 1.83$ and $z = 0.0085$ gave the closest fit of the alternate form with a mean square error of $4.9488*10^{-4}$. The alternate fit in this case proved to be a very close match to the original curve. It is to be noted that the probability of occurance of Off-times less than 200ms long is zero in the real curve. But the exponential model has non zero probabilities of occurance of these values. This explains why the exponential model is a better fit for On-time distribution than for the Off-time distribution. These curve fits derived out of the Off-time distribution over the entire length of all conversations is shown over the Off-time distribution curves for users 3 and 7 in figures 9 and 10.

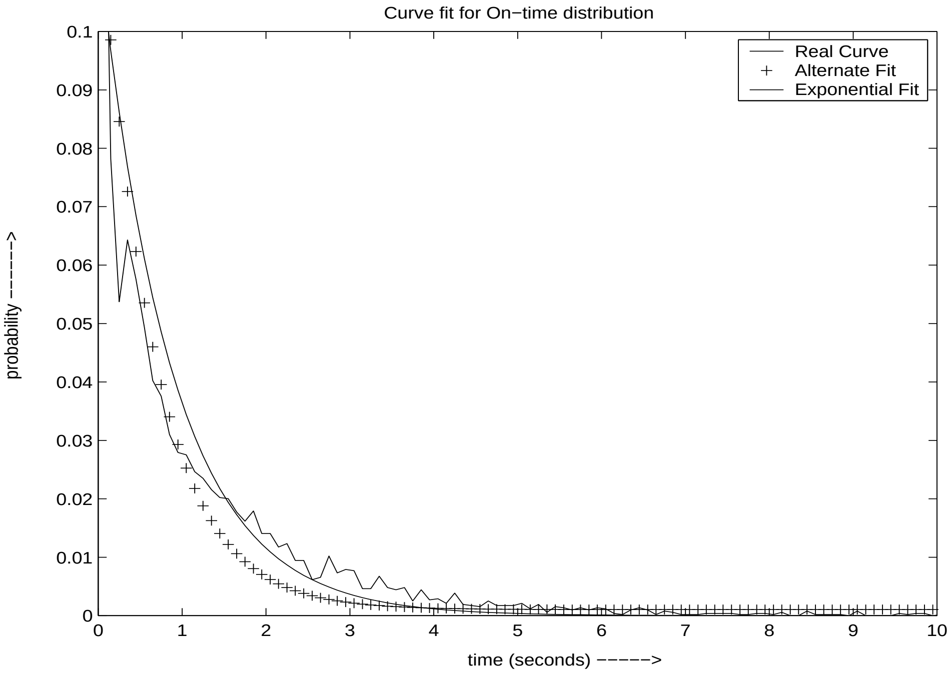

Fig.5 Curve fit for Histogram showing On-time distribution

5 Conclusions

The desired speech characteristics of telephone conversations are obtained. It has been found that there is generally good agreement between [1][2] and our own work in terms of speech activity factor and mean On-Off times, even though the values are not exactly the same in all works. However,

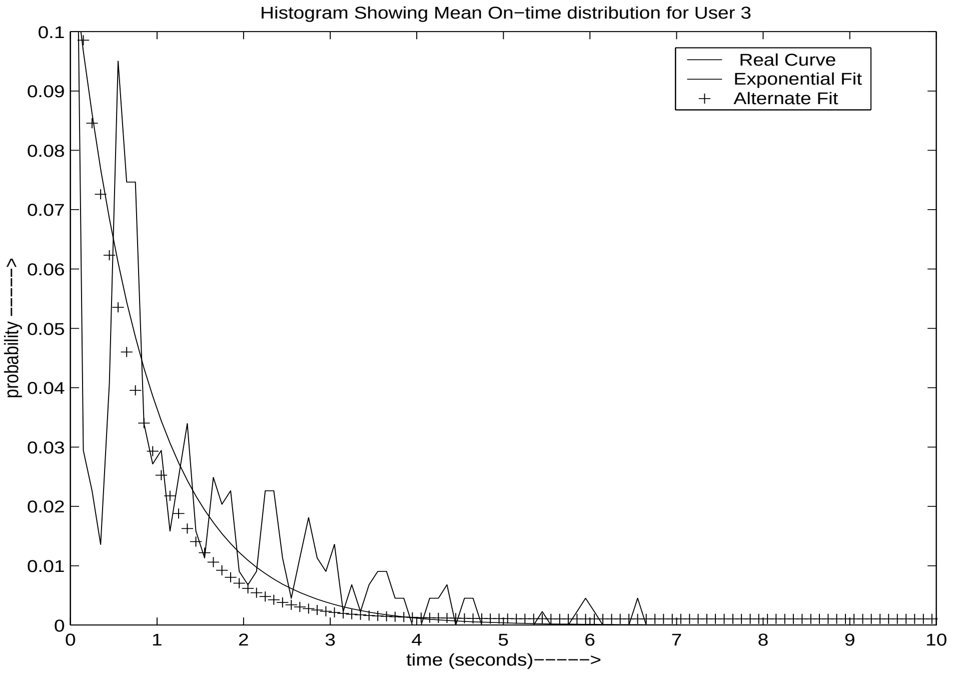

Fig.6 Histogram showing Mean On-time distribution for User 3

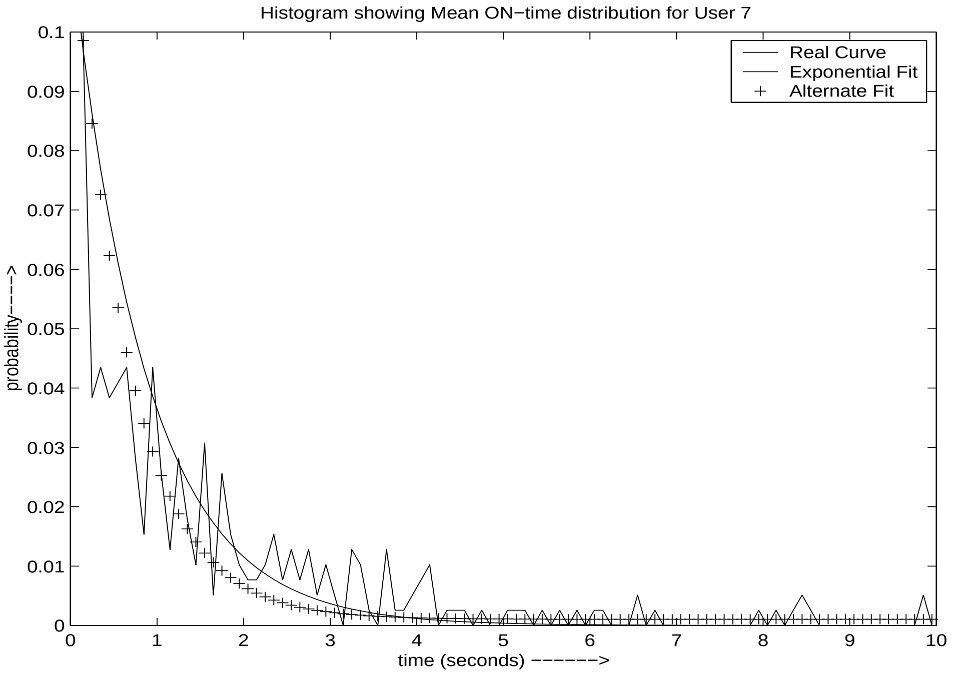

Fig.7 Histogram showing Mean On-time distribution for User 7

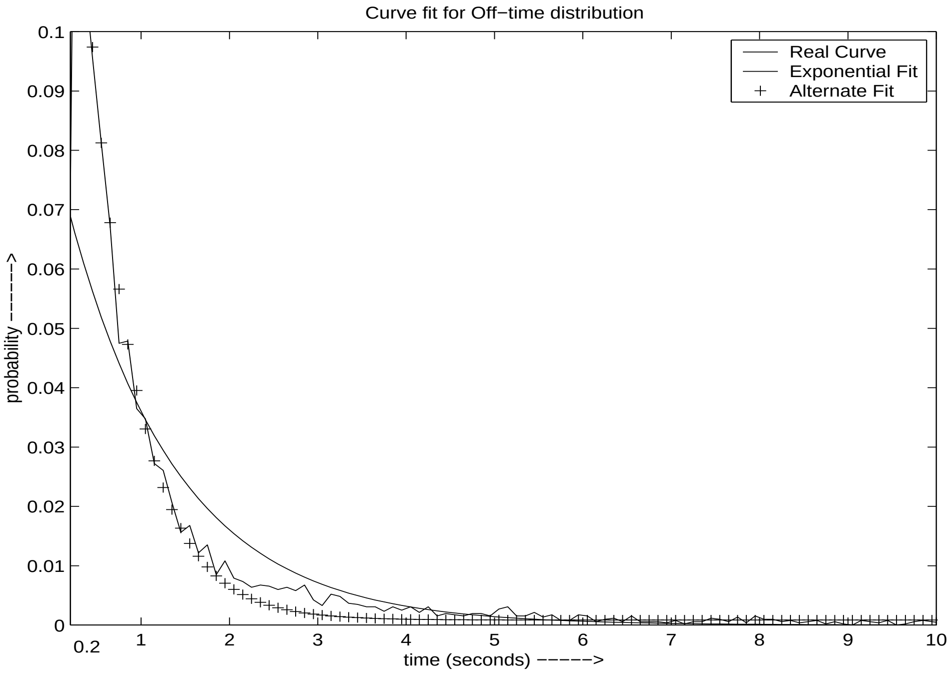

Fig.8 Curve fit for Histogram showing Off-time distribution

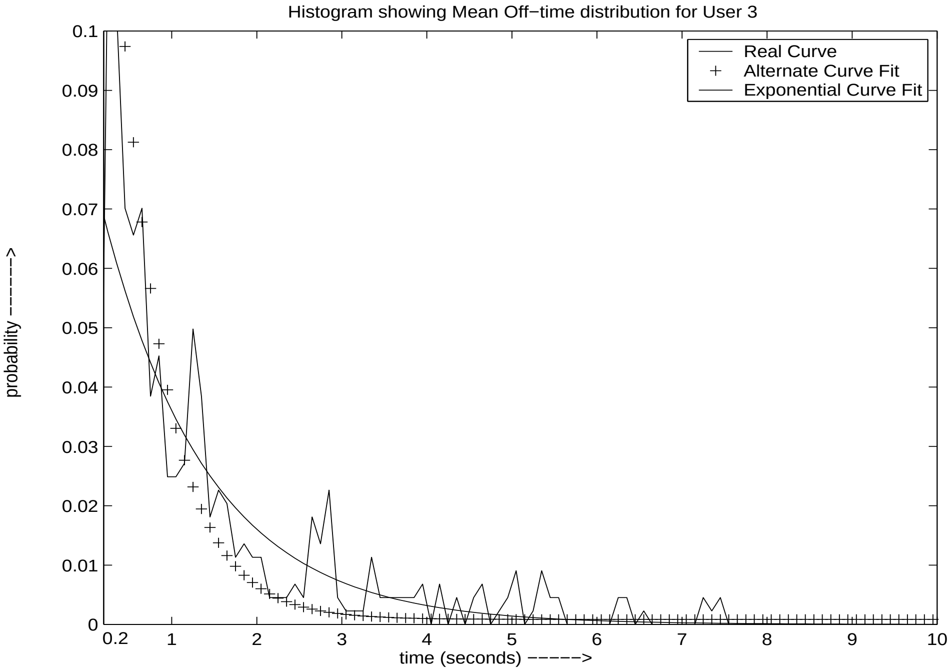

Fig.9 Histogram showing Mean Off-time distribution for User 3

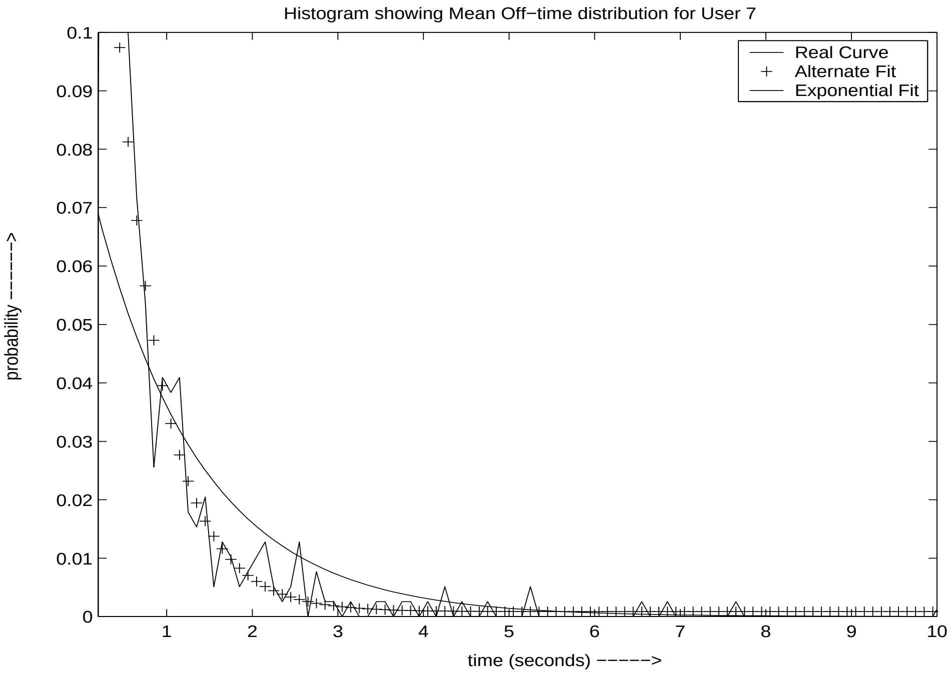

Fig.10 Histogram showing Mean Off-time distribution for User 7

the mean On-Off times obtained in [3][4] are quite different from ours. The agreement with [3][4] is better in terms of the speech activity factor, but this is because most of the time only one of the two speakers in a telephone conversation speak at any given time, which gives a speech activity of somewhat less that 50 percent in all the cases. References [1][2] propose an exponential model for distribution of On-Off times but the model is rejected by [3]. Our analysis shows that an exponential model is a good approximation for the On-Off time distribution for telephone conversations though an alternate model that was proposed proves to be a better fit than the exponential model. The exponential model was found to be a better approximation for On-time distribution than for the Off-time distribution.

References

[1] ”A Statistical Analysis of On-Off Patterns in 16 Conversations,” Paul T.Brady, Bell System Technical Journal, pp 73-91, Sept 1967.

[2] ”A model for generating ON-OFF speech patterns in two-way conversations,” Paul T.Brady, Bell System Technical Journal, Vol 48, pp 2445-2472, Sept 1969.

[3] ”Traffic Characteristics of Packet Voice,” Shuang Deng, IEEE International Conference on Communications, Vol 3, pp 1369-1374, Sept 1995.

[4] ”Traffic Smoothing Effects of Bit Dropping in a Packet Voice Multiplexer,” K.Sriram and D.M.Lucantoni, INFOCOM, Vol 48, pp 759-766, 1988.

[5] ”Packing Density of Voice Trunking using AAL2,” Chunlei Liu, Sohail Munir, Raj Jain, Sudhir Dixit, ATM Forum Document Number: ATM Forum-98-0830, 1998.

[6] ”An Improved Endpoint Detector for Isolated Word Recognition,” L.F.Lamel, IEEE Transactions on Acoustics., Speech, Signal Processing, Vol.29 pp.777-785, Aug 1981.