Task Dependency of User Perceived Utility in Autonomic VoIP Systems

Edward Stehle, Maxim Shevertalov, Paul deGrandis, Spiros Mancoridis, Moshe Kam
Department of Computer Science
Drexel University
Philadelphia, PA 19104

Abstract

The transmission of voice-over-Internet protocol (VoIP) network traffic is used in an increasing variety of applications and settings. Many of these applications involve communications where VoIP systems are deployed under unpredictable conditions with poor network support. These conditions make it difficult for users to configure and optimize VoIP systems and this creates a need for self configuring and self optimizing systems. To build an autonomic system for VoIP communications, it is valuable to be able to measure the user perceived utility of a system. In this paper we identify factors important to the estimation of user perceived utility in task dependent VoIP communications.

1. Introduction

As the transmission of voice-over-Internet protocol (VoIP) network traffic becomes commonplace, VoIP is used in an increasing variety of applications and settings. Many current applications are outside the context of simple social conversation across dependable networks. Field applications, such as military operations, employ VoIP for task-specific communications and require VoIP to operate under poor network conditions. Emergency-response personnel may use VoIP communications to complete tasks in disaster areas where extreme weather or other adverse conditions interfere with network performance. Operations may be carried out in locations where there is little or no communications infrastructure or where the communications infrastructure has been damaged. Under these field conditions VoIP needs to be served by small, mobile, ad-hoc networks with limited resources.

VoIP systems for field communications need to be deployed quickly to minimize response time. In order to deliver the best possible support to field operations, VoIP systems must be optimized to the field conditions. This creates a difficult problem for the users of field VoIP systems. How do you quickly find an optimal configuration for a VoIP network under adverse conditions when little is known about these conditions before the system arrives in the field? How do you optimally manage a VoIP network under changing field conditions? This is an ideal application for autonomic systems. If we can produce a context aware VoIP system that can self configure when deployed and self optimize as field conditions change, we can reduce deployment time and improve overall performance in unknown and unpredictable settings.

In order to build an autonomic system for field VoIP communications, we must have a way to measure the performance of the system. Such an autonomic system must be aware of user perceived utility of the VoIP application. One approach when including "black-box" applications in an autonomic system, is to develop models for application utility estimation [1]. Autonomic systems using utility function policies [2, 3] require an estimate of an application's performance. Previous work in the area of monitoring the health of autonomic systems involved the use of a pulse to estimate the health of specific autonomic elements [4, 5, 6].

In this paper we look at methods to map network conditions to user-perceived utility as a utility function. We identify factors that need to be considered when mapping network conditions to user perceived utility. Specifically, we determine if the mapping from network conditions to perceived utility is task dependent. We also determine if the mappings for users performing different roles within the same task are affected by their roles. Finally, we wish to determine if perceived utility changes with the continued repetition of a task.

This paper is structured as follows. First we present previous work in calculating the user perceived utility of VoIP applications (Section 2). Then we will present the set up of our human subject experiments to explicitly determine user perceived utility of VoIP applications (Section 3). We will conclude by presenting our results (Section 4), concluding remarks (Section 5), and an appendix of collected data (Section 6).

2. Previous Work

Existing approaches for predicting user perception of utility in VoIP systems fall into two main categories. Some approaches base predictions on the degradation of a reference signal and other approaches map network conditions

to perception of utility based on subjective data gathered in human-subjects testing.

2.1. Reference Signal Approach

Objective systems such as the Perceptual Speech Quality Measure (PSQM) [13] and the Perceptual Assessment of Speech Quality (PESQ) [12] require a speech sample to be sent across a VoIP network. The original sample is then compared to the sample that is received on the other end of the VoIP system. A prediction of user utility is made based on the degree to which the signal has degraded.

The main criticism of the existing objective approaches is that they only consider signal distortion in one direction. They do not consider network impairments such as delay and echo [11].

2.2. Subjective Testing based Approach

The most common model for mapping network conditions to user-perceived utility for voice applications is the E-model [9]. During the mid-nineteen nineties the International Telecommunications Union (ITU) designed the E-Model to measure objectively the quality of a public-switched telephony network (PSTN). The E-Model was originally intended to be used by network planners to predict the quality of PSTNs without the need for expensive and time-consuming testing of human subjects. It has since been adapted to cellular communications and IP telephony [10, 8, 7].

The E-Model has become a commonly used metric to predict the quality of VoIP applications for several reasons. Most models for objective quality measurement require that the received signal be compared to the sent signal. The E-Model is the only widely recognized metric that does not require a reference signal, making it computationally feasible for real time applications. In addition, the E-Model correlates well with subjective quality in situations where IP telephony functions in the same fashion as PSTN; for example in local VoIP networks where anomalous traffic conditions are minimized.

There are, however, problems with using the E-Model to predict perceived utility in general. Although the E-Model correlates well with subjective quality in situations where IP telephony functions in the same fashion as PSTN, using the E-Model outside of this context greatly decreases such correlation. The E-Model was not derived for this explicit purpose. In fact, the E-Model was not intended as a quality assessment tool, but rather as a tool for planning circuit switched networks. The E-Model was not meant to be applied to IP networks. The impairment factors that comprise it deal more with signal processing than with IP networks.

Parameter	Values
Bandwidth	25, 40, 50, 65, 80 (kbps)
Latency	0,1000, 2000, 3000, 4000 (ms)
Loss	0, 12.5, 25, 50, 60 (percent)

Table 1: 3-tuple Parameters

2.3. Problems with current approaches

Neither reference-signal based approaches nor subjective test approaches consider the impact of task on a perceived utility. Current models assume that, given network conditions, users will always perceive utility in the same manner regardless of what task they are using VoIP to perform. In tests using circuit-switched networks Kitawaki and Itoh concluded that speech quality due to propagation delay greatly depends on the kind of task [11]. Their tests showed that delay has a greater effect on tasks that are more interactive.

3. Our Tests

In our tests, subjects rate the quality of VoIP under varying network conditions. Each test involves one pair of human test subjects. The subjects carry out a series of similar tasks that require communication using a VoIP application. For all of our testing we used Gnome Meeting as the VoIP application and G.711 for our audio codec. We vary the network conditions using a FreeBSD application named Dummynet, which allows us to set the bandwidth, latency and loss of the link used by our test subjects. A single test point in our experiment is a 3-tuple (bandwidth, latency, loss). Each of these parameters can have one of five values. We test across all combinations of these values, giving us 125 points per subject. The possible values of the parameters are listed in Table 3.

We have been performing three different types of human subject tests, each with a different task. We believe that the relationship between network conditions and user satisfaction is task dependent and that using more than one test with different tasks will provide data to support this belief. All of the tests have the same basic structure. There are two roles that the subjects play during a test. One subject is a `questioner` and one subject is a `responder`. The actual duties of the `questioner` and the `responder` vary between the types of test. The subjects perform one task at each of the 125 test points. After a task is completed each subject votes on the quality of the communication. Then the network conditions are changed to the next point and the next task begins. The subjects rate the quality on a scale of one to five where one is bad, five is good, and

three is okay. The subjects alternate between the roles of questioner and responder after each task. Each test collects 250 data points and takes between 60 and 90 minutes to complete.

3.1. Simple Information Exchange Test

The first VoIP test is designed to measure perceived utility during tasks involving a simple exchange of information. The tasks in this test involve the the `questioner` asking a trivia question and the `responder` answering it. Completion of this task involves minimum back-and-forth conversation between the subjects and does not have any time constraint. We believe that this test is useful for modeling VoIP communications where the users are simply exchanging facts or instructions. For example, if VoIP is being used to convey a military target's position and instructions for engaging the target, we expect the conversation to be limited to conveying position, conveying instructions, and a confirmation that the message has been received.

In this test the `questioner` is given a trivia question and the answer to the trivia question. The `responder` is given a list of possible answers, one of which is correct. The `questioner` reads the question to the `responder`. The `responder` picks an answer from the list and reads it to the `questioner`. Then the `questioner` records whether the question was answered correctly. This requires both subjects to receive a piece of information from the other and then respond to that information.

We have conducted the simple information test with thirty human subjects and collected 3750 data points.

3.2. Time-Sensitive Collaboration Test

The second test is designed to measure perceived utility during time-sensitive tasks that involve some collaboration between subjects. The tasks in this test involve a considerable amount of back-and-forth conversation between the two subjects in order complete a time-constrained task. This test is intended to model situations where users are not trying simply to convey information but to perform some collaborative task. For example, if two military commanders need to collaborate on a plan for a time-critical task, we would expect a considerable amount of back-and-forth conversation and pressure to complete the plan quickly.

In this test the `questioner` is given a word that the `responder` must correctly guess, but the `questioner` may not explicitly state the word. The `questioner` can only describe the word and answer the questions of the `responder`. The `responder` can guess the word or ask the `questioner` for specific information about the word. Each task has a time limit of thirty seconds. The task ends when the `responder` correctly guesses the word or the time runs out.

We have conducted the time-sensitive collaboration test with 30 human subjects and collected 3750 data points.

3.3. Time-Sensitive Information Exchange

The third VoIP test is designed to measure perceived utility during time constrained tasks involving the exchange of multiple pieces of information. The tasks in this test involve the collaborative summing of a series of small integers within a limited period of time. This test is intended to model situations where users need to collaborate and the collaboration is limited to a series of simple exchanges of information. For example, in order to coordinate the response of emergency workers in separate locations of a disaster area these workers may need to combine collected data such as the number of disaster victims.

In this test the `questioner` and `responder` are each given a list of integers. The `questioner` is given a "starting number", an "ending number" and two "adding numbers". The `responder` is given three "adding numbers". The starting number is an integer from zero to ten, the adding numbers are integers from zero to five, and the ending number is the sum of the starting number and the adding numbers. The `questioner` initiates the task by reading the starting number to the `responder`. The `responder` adds his first adding number to the starting number and reads the sum to the `questioner`. The exchange continues with each subject adding one adding number to the sum until all of the adding numbers have been summed with the starting number. Once all of the adding numbers have been summed with the starting number the `questioner` checks the total against the ending number and informs the `responder` that the numbers have been summed correctly or incorrectly. Each task has a time limit of thirty seconds.

We have conducted the time sensitive collaboration test with 30 human subjects and collected 3750 data points.

3.4. User-Adjustment Tests

User-adjustment tests were designed to measure changes in perceived utility as a task is repeated. The tasks in these tests are performed over a set of network conditions, and then repeated over the same set of network conditions. The results from the first time through the set of network conditions can then be compared to the results from the second time through the same set of network conditions. These

Parameter	Values
Latency	0,1000, 2000, 3000, 4000 (ms)
Loss	0, 12.5, 25, 50, 60 (percent)

Table 2: 2-tuple Parameters

tests are designed to model situations where a user learns and adjusts to tasks.

User adjustment tests were performed using the three previously described tasks. These include the tasks described in Section 3.1 (Simple Information Exchange Test), Section 3.2 (Time Sensitive Collaboration Test), and Section 3.3 (Time Sensitive Information Exchange). In their original form, each of the previously described tests was performed over 125 network condition points. Repeating all points in a test would yield a test with 250 data points that would take two to three hours to complete. A test of this length would tire the test subject. This would corrupt the test results and create unnecessary stress for the test subjects. In order to reduce the time required to complete test trials the size of the set of network settings was reduced. The possible values of network condition parameters described in Table 3 were altered so that only the highest bandwidth value was used. A single test point in our user adjustment tests is a 2-tuple (latency, loss). Each of these parameters can have one of 5 values, giving us 25 points. These points are randomly ordered, and then repeated in the same random order, giving us 50 points per subject. The set of possible parameters for the user adjustment tests is listed in Table 2.

3.5. Our Test Bed

In order to carry out these tests we created a test bed that allows two subjects to converse using VoIP while we control the properties of the channel over which VoIP is running.

Our test bed consists of one "subject computer" for each of our two subjects, a switch partitioned into two subnets, and one "bridge computer" that is used to set the bandwidth, latency and loss of the channel over which the two subject computers communicate. Figure 1 illustrates the manner in which the test bed is connected. Each of the subject computers is connected to a different subnet and the bridge computer is connected to both of the subnets. Communications between the two subject computers are routed through the bridge computer. The bridge computer employs Dummynet to enforce the bandwidth, latency and loss on the channel connecting the two subject computers. The subject computers and the bridge computer are also connected through a back channel, which is not effected by Dummynet. This back channel is used to send messages to the bridge computer instructing it to change the Dummynet settings.

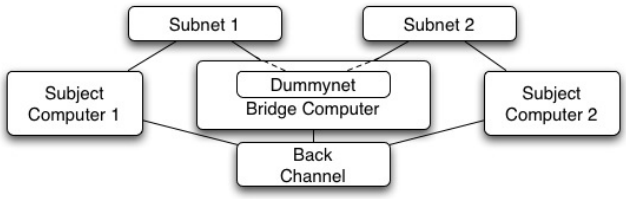

Figure 1: Architecture of the Test Bed

4. Results

The results of our experiments can be seen in Figures 2-16 found in the Appendix (Section 6). There are two types of figures: 1 tests in which the test points are 3-tuples (bandwidth, latency, loss) that are represented by three-dimensional plots, tests in which the test points are 2-tuples (latency, loss) that are represented by two-dimensional plots.

The three-dimensional plots show the space defined by bandwidth, loss and latency measurements. Within this space color is used to represent a user-satisfaction rating. The darkest red represents the areas that were rated best, and the darkest blue represents the areas that were rated worst. In each of these figures our test space is represented by three plots, each sliced along a different axis. One is cut along bandwidth, one along latency, and one along loss.

The two-dimensional plots show the space defined by our loss and latency measurements. The same color convention is used to represent user satisfaction rating.

In the three-dimensional plot section of our results there are three categories of plots. There are plots where both `questioner` and `responder` perceived utility is averaged for each test point, plots where only the `questioner` perceived utility is averaged for all points, and plots where only the `responder` perceived utility is averaged for all points. There are combined `questioner`-and-`responder`, `questioner`-only, and `responder`-only plots for each of our tests.

The two-dimensional plots section show the results of the user adjustment tests. For each of the three tests there are two plots. One plot showing the perceived user utility averaged over the first time through the test points, and one plot showing the perceived utility averaged over the second time through the test points

The average variance, minimum variance, maximum

Average Variance	0.732
Maximum Variance	2.193
Minimum Variance	0.216
Variance of Variance	0.080

Table 3: Variance of User Perceived Utility for Simple Information Exchange

Average Variance	0.610
Maximum Variance	1.140
Minimum Variance	0.127
Variance of Variance	0.040

Table 4: Variance of User Perceived Utility for Time Sensitive Collaboration

Average Variance	0.740
Maximum Variance	2.187
Minimum Variance	0.187
Variance of Variance	0.101

Table 5: Variance of User Perceived Utility for Time Sensitive Information Exchange

variance and the variance of the variance for all test points is shown in tables 3 through 5.

User utility ratings for the test points in the simple information-exchange test have an average variance of 0.732, a minimum variance of 0.216, a maximum variance of 2.193, and the variance of the variance is 0.080. The user utility ratings for test points in the time sensitive collaboration test have an average variance of 0.610, a minimum variance of 0.127, a maximum variance of 1.140, and the variance of the variance is 0.040. The user-utility ratings for test points in the time sensitive collaboration test have an average variance of 0.740, a minimum variance of 0.101, a maximum variance of 0.187, and the variance of the variance is 2.187.

As expected, the results vary somewhat for different tasks. One obvious difference between the results for different tasks is the effect of latency on utility. In the time-sensitive collaboration test and in the time-sensitive information exchange test, latency had a greater effect on perceived utility than in the simple information-exchange test. These results make intuitive sense. Tests in which the tasks are subject to time constraints show a greater user reaction to latency. We believe that this is caused not only by the addition of the time constraints, but also by the collaborative nature of the communication. During this type of collaboration, subjects spend more time speaking back-and-forth than they do during the simple information exchange test. Greater latency can cause this back-and-forth communication to fall out of sync, creating additional difficulties in communication.

Another obvious difference is the effect of bandwidth and loss. Bandwidth has the greatest effect on the simple information-exchange test. We believe that the collaborative nature of the time-sensitive tests helps users adjust to poor voice quality. Because these tests involve more back-and-forth communication, the users have more opportunity to recognize poor quality. Once poor voice quality is recognized, users may begin to employ strategies such as repeating messages without being asked. The back-and-forth communication also gives users more opportunity to recognize conversational context. Recognizing conversational context can be helpful for filling in portions of messages which cannot be understood.

When the results of our test are split into questioner-only and responder-only plots it is clear that the role played within a task has an effect on perceived utility. Again, this is an expected result. Different roles within a single test can be thought of as different sub-tasks, and we have already illustrated that perceived utility is task dependent.

The results of our user adjustment tests show perceived utility changes as users repeat a task. In each of the tests the variance of the perceived utility decreased during the second time through the test points. At the same time the average perceived utility stayed approximately the same. It appears that as users repeat a task over different network conditions they "get used to it". They perceive fewer extremes in utility and tend to perceive a larger portion of the test space as "okay".

5. Summary and Conclusions

Knowledge of network conditions, such as bandwidth, latency and loss, is not sufficient to predict the performance of a VoIP system adequately. The predictor must also have knowledge of the task being performed over the VoIP system. Our tests show that user perceived utility may be very different for users performing different tasks even if network conditions are the same.

Many tasks performed over VoIP systems involve multiple users playing different roles within the tasks. Our tests show that perceived utility may be very different for users performing different roles within a task. When determining what network resources are required to complete a task, it may be necessary to base predictions on the most constrained role within a task.

While carrying out a task, a user may adjust to a task and network condition combination. Our tests show that user perception of utility changes as a user repeats tasks over the same network conditions. Users my benefit by starting to talk over a VoIP connection before beginning a task. Users may also benefit by training over simulated bad network

conditions.

References

[1] M. Karlsson and M. Covell "Dynamic Black-Box Performance Model Estimation for Self-Tuning Regulators" *Proc. Sixth International Conference on Autonomic Computing*, 2005.

[2] J. O. Kephart and W. E. Walsh "An Artificial Intelligence Perspective on Autonomic Computing Policies" *Proc. Fifth IEEE International Workshop on Policies for Distributed Systems and Networks*, 2004.

[3] W.E. Walsh, G. Tesauro, J.O.Kephart, and R. Das "Utility Functions in Autonomic Systems" *Proc. First International Conference on Autonomic Computing*, 2004.

[4] R. Sterritt "Pulse Monitoring: Extending the Health-check for the Autonomic GRID" *Proc. IEEE INDIN*, 2003.

[5] R. Sterritt and D. Bustard "A health-check model for autonomic systems based on a pulse monitor" *Knowl. Eng. Rev.*, Cambridge University Press , vol.21, no.3pp.195-204, 2006

[6] H.L. Truong, T. Fahringer, F. Nerieri, and S. Dustdar "Performance Metrics and Ontology for Describing Performance Data of Workflows" *Proc. IEEE International Symposium on Cluster Computing and the Grid*, 2005.

[7] A. Markopoulou, F. Tobagi and M. Karam "Assessment of VoIP Quality over Internet Backbones" *Proc. IEEE INFOCOM*, 2002.

[8] T. A. Hall "Objective Speech Quality Measures for Internet Telephony" *Proceedings of SPIE Voice over IP VoIP Technology*, vol. 4522, pp. 128-136, July 2001

[9] N. O. Johannesson "The ETSI Computation Model: A Tool for Transmission Planning of Telephone Networks" *Communications Magazine*, IEEE , vol.35, no.1pp.70-79, Jan 1997

[10] R. G. Cole, and J. H. Rosenbluth "Voice Over IP Performance Monitoring" *SIGCOMM Computer Communication* Rev. 31, 2, Apr. 2001

[11] N. Kitawaki, and K. Itoh "Pure Delay Effects on Speech Quality in Telecommunications" *IEEE Journal on Selected Areas in Communication* Vol. 9 NO. 4, May 1991

[12] A.W.Rix, J.G. Beerends, M.P. Hollier, and Hekstra, A.P. "Perceptual evaluation of speech quality (PESQ)-a new method forspeech quality assessment of telephone networks and codecs" *Proc. IEEE Acoustics, Speech, and Signal Processing*, 2001.

[13] N. Kitawaki "Perceptual QoS assessment methodologies for coded speech in networks" *Proc. IEEE Workshop on Speech Coding*, 2002.

6. Appendix

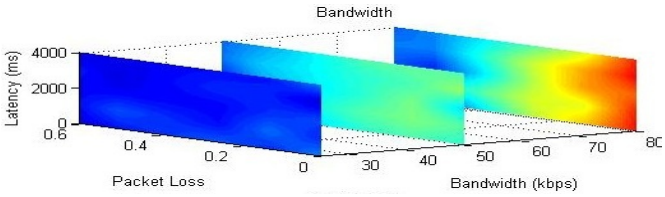

Figure 2: Simple Information Exchange

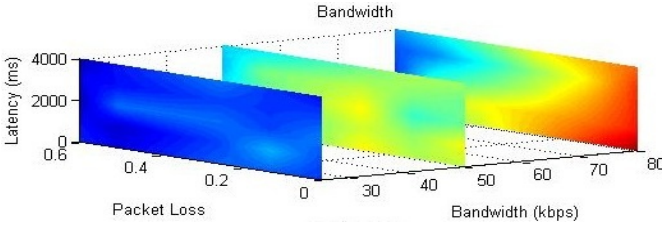

Figure 3: Time Sensitive Collaboration

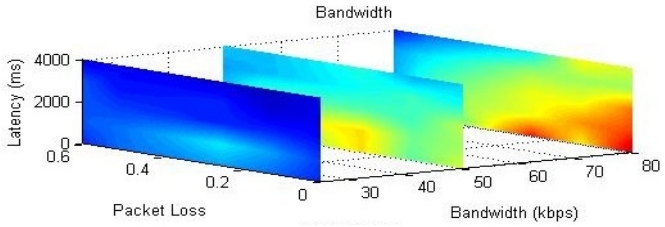

Figure 4: Time Sensitive Information Exchange

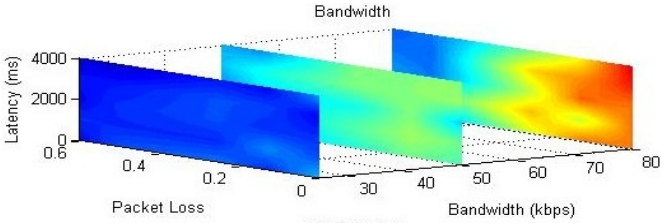

Figure 5: Simple Information Exchange: Questioner Only

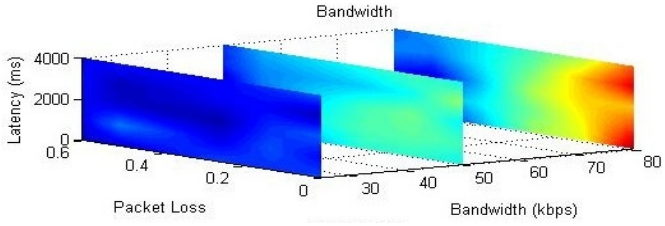

Figure 6: Simple Information Exchange: Responder Only

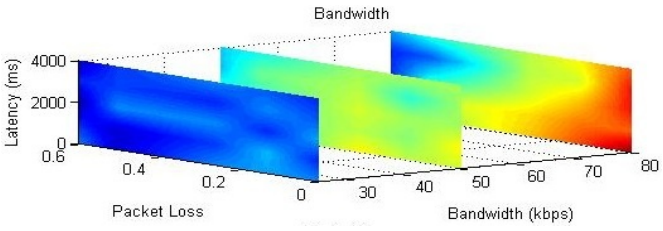

Figure 7: Time Sensitive Collaboration: Questioner Only

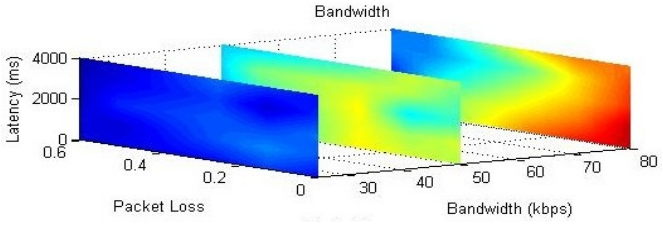

Figure 8: Time Sensitive Collaboration: Responder Only

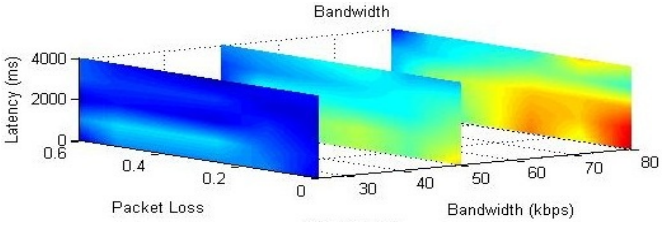

Figure 9: Time Sensitive Information Exchange: Questioner Only

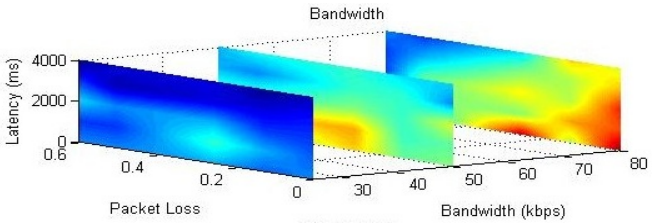

Figure 10: Time Sensitive Information Exchange: Responder Only

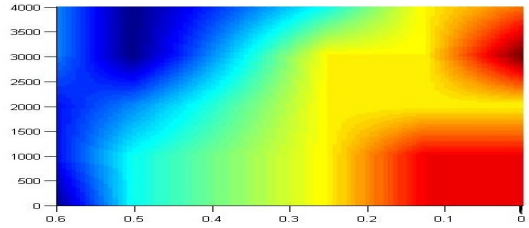

Figure 11: Simple Information Exchange: First 25

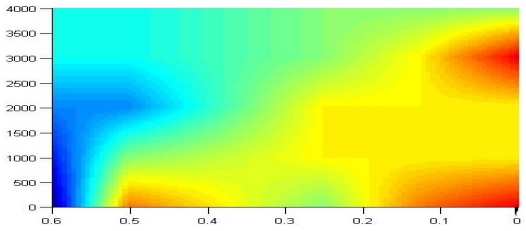

Figure 12: Simple Information Exchange: Second 25

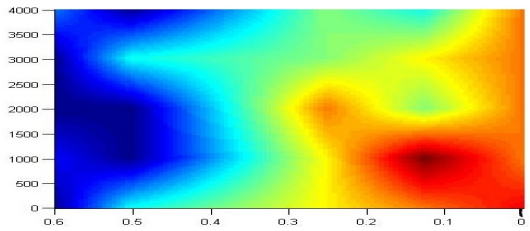

Figure 13: Time Sensitive Collaboration: First 25

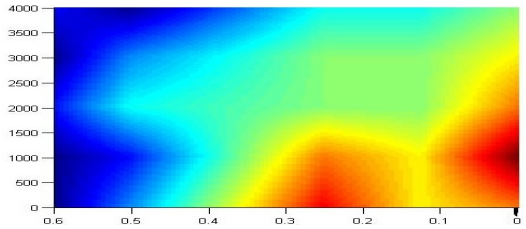

Figure 14: Time Sensitive Collaboration: Second 25

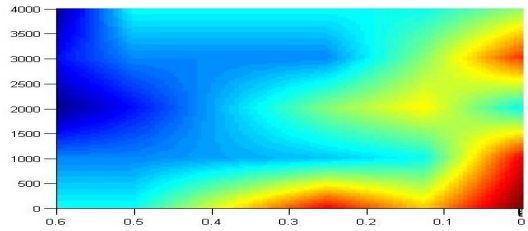

Figure 15: Time Sensitive Information Exchange: First 25

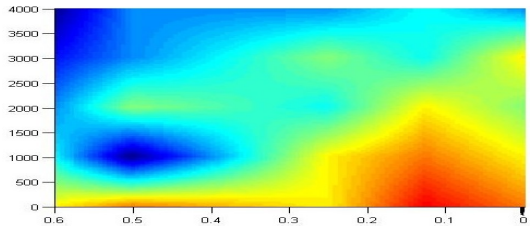

Figure 16: Time Sensitive Information Exchange Second 25