



**Modeling and Designing an IVR via
Phase Type Distributions**

M.Sc. Research Proposal

Nitzan Yuviler

Advisor: Prof. Avishai Mandelbaum

**The Faculty of Industrial Engineering and Management
Technion - Israel Institute of Technology**

March 22, 2011

Contents

1 Introduction	3
2 Literature Review	4
2.1 Methodologies for evaluating IVRs.....	4
2.2 Designing and optimizing IVRs.....	6
2.3 Modeling a Call Center with an IVR.....	7
2.3.1 "Performance Analysis of a Call Center with Interactive Voice Response Units".....	7
2.3.2 "Designing a Call Center with an IVR".....	8
2.3.3 "Analysis of Customer Patience in a Bank Call Center".....	9
2.3.4 "Robust Design and Control of IVR Systems in Call Centers".....	10
2.4 Fitting Phase Type Distributions.....	11
2.4.1 "Fitting Phase-Type Distributions to Data from a Telephone Call Center".....	11
2.4.2 "A Novel Approach for Fitting Probability Distributions to Real Trace Data with the EM-Algorithm".....	12
2.4.3 "Flowgraph Models for Generalized Phase Type Distributions Having Non-Exponential Waiting Times".....	12
3 Problem definition and research objectives	13
4 Methodology	21
5 Data Source	23

1. Introduction

Call centers play an increasingly growing part in today's business world and global economy. Indeed, they serve as the primary customer-contact channel, for companies in many different industries. Employing millions of agents across the globe, call centers are highly labor-intensive operations, with the cost of staff members who handle phone calls (also known as "agents", or CSRs – Customer Service Representatives) typically comprising 60%-80% of the overall operating budget [1]. In order to reduce these operational costs, it is important to identify means for reducing agents' workload – self-service of customers is one way, and IVR (Interactive Voice Response) is presently the main self-service channel [12].

IVR systems, **if properly designed**, can increase customer satisfaction and loyalty, cut staffing costs and increase revenue by extending business hours and market reach [3]. Poorly designed IVR systems, on the other hand, will cause the opposite effect and lead to dissatisfied customers, increased call volume and even increased agent turnover, as customers take out their frustrations on the agents. A recent Purdue University study revealed that 92% of US consumers form their image of a company based on their experience using the company's call center. More strikingly, the study found that 63% of consumers stop using a company's products based on a negative call center experience [6]. Analysis of IVR transactions may shed light on the IVR service quality and efficiency. For example, are customers satisfied with the IVR service or do they opt out for agent assistance, or perhaps, leave the IVR without getting any relevant information? Are all IVR capabilities in fact being used? [12].

The goal of our proposed research is to model and analyze customers flow within the IVR. The optimization of IVR system design, management and performance can be achieved through system modeling and careful analysis of the data supporting the model. In this research we shall try to model the customer flow within the IVR via phase-type distributions, and based on real data.

Phase-type distributions are defined as distributions of absorption times T , for Markov processes with $k < \infty$ transient states (the phases) and one absorbing state Δ . Quoting from Ishay [10], "There are several motivations for using phase-type distributions as statistical models. These distributions arise from a generalization of

Erlang's method of stages in a form that is particularly well-suited for numerical computation: problems which have an explicit solution, assuming exponential distributions as their building blocks, are algorithmically tractable when one replaces the exponential distribution with a phase-type distribution. Furthermore, the class of phase-type distributions is dense and hence any distribution on $[0, \infty)$ can, at least in principle, be approximated arbitrary close by a phase-type distribution".

Customers' transactions in the IVR consist of varying sequences of tasks. Fitting phase type distributions to the IVR process will help analyze the various IVR tasks and their role and effect on process duration and success. The design of IVR service is a comprehensive process, which must account for customer needs, company preferences and human factors [12]. We believe that a stochastic model of customers flow within the IVR will contribute to better designing and modeling IVR systems, thus supplementing existing research from other fields such as Human-Factor-Engineering [5,15,18].

This research proposal is arranged as follows: in Section 2 we present a brief literature review. In Section 3 we introduce the problem and our research objectives. Our proposed methodology for solving the problem is presented in Section 4. In Section 5 we present the data source on which we shall be relying on in our research.

2. Literature Review

2.1 Methodologies for evaluating IVRs.

Suhm and Peterson [19, 20] presented a comprehensive methodology for IVR usability evaluation and re-design. It is claimed there that the standard IVR usability tests and standard IVR reports are not sufficient to assess the true performance and usability of an IVR. Most of the reports being used by call center managers are based on measures related to IVR utilization, for example, the percentage of customers that left the IVR without seeking a live agent, where this percentage is interpreted as the success rate of the IVR service. After analyzing thousands of end-to-end calls, which included the interaction with the IVR and the customer-agents dialogs, it was discovered that although 30% of the costumers completed their service in the IVR, only 1.6% of the customers actually got relevant service.

Suhm and Peterson [19] also presented a new measure for quantifying IVR usability and cost-effectiveness. This measure is defined as the agent time being saved by handling the call, or part of the call, in the IVR, compared to handling the call only by a live agent. It was also suggested that, by using User-Path Diagrams, one can identify usability problems, such as nodes in the IVR that are rarely visited, nodes with high volume of abandonment and nodes with high volume of customers seeking live agent assistance. By analyzing end-to-end calls, one can also compare the distribution of original call reasons, which is revealed through the customer-agent dialogs, to the call-type distribution that arose from the IVR logs. Differences between those two distributions indicate that the customers are not navigating correctly within the IVR. In order to identify menu navigation problems, Suhm and Peterson [20] suggested a chart that shows IVR options chosen by customers (IVR routing) as columns. Then, they break down these columns by the original call reasons, as revealed through the customer-agent dialog. In this chart, menu options that are chosen correctly are shown as columns that consist of just the matching reason, while columns with many different components indicate menu options that are frequently selected incorrectly. Identifying which menu options have usability problems, and analyzing the menu wording, will infer specific solutions for improving menu navigation. For example, the option "Leave Message", in Figure 1 below, might often be chosen incorrectly because it suggests an opportunity of speaking to an agent.

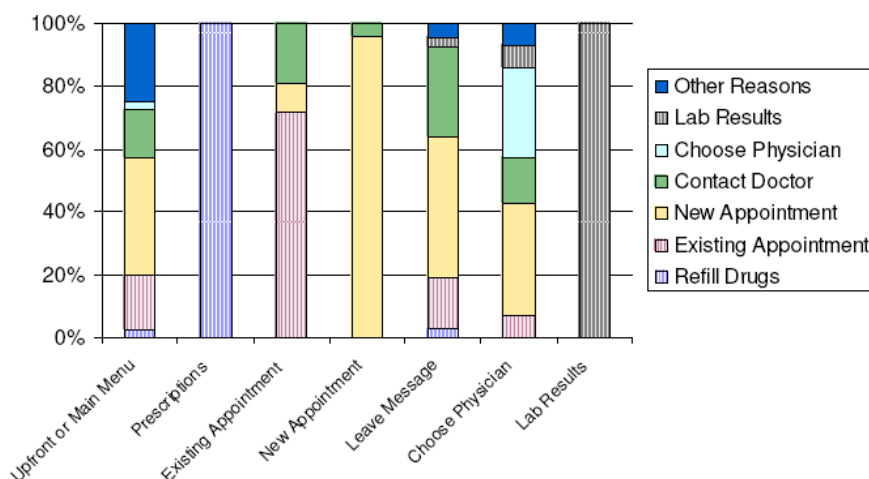


Figure 1: Analysis of menu navigation accuracy. Columns represent menu selection, broken down by call reason.

2.2 Designing and optimizing IVRs

The subject of improving and optimizing IVR design has been addressed from different aspects. From the human factors aspect, two main issues have been dealt with: a) The menu structure – mainly, comparing broad (shallow) menus and deep (narrow) menus, and b) The IVR platform – touch-tone versus natural language.

Schumacher et al. [15] recommended that menus with command-like options are to be limited to four or less items in each layer. They claimed that because all the output is auditory, broad menus, with more than four options per layer, place a heavy demand on the working memory and therefore should be avoided.

Commarford et al. [5], on the other hand, argued that broad menus do not overload the working memory. The user does not need to remember all the options in the menu but to hold the best option, compare it with the new one and then save the better option. Therefore, the user only needs to hold up to two options in the working memory. Creating a deep, divided menu in order to limit the number of options in every layer can increase the complexity and create confusion because it might not fit the user's mental model. Commarford et al. [5] have conducted an experiment which showed that users who used broad menu structure performed tasks faster and with greater satisfaction than users who used deep menu structure. Suhm, Freeman and Getty [18] showed that long menu with specific, well defined options is better than short menu with few wide options, in terms of response rate (choosing a legal option from the menu), routing rate (being routed to a specific group of agents) and re-prompt rate (need to listen to the menu again).

Suhm et al. [17] addressed the issue of touch-tone IVRs vs. natural language IVRs. It was found that although natural language menus may still have recognition problems, they give better performances in response rate, routing rate and routing time.

More researches address the issue of optimizing IVR design by presenting algorithms to reduce the service times in the IVR. Salcedo-Sanz et al. [14] introduced an evolutionary algorithm to optimally design IVRs, based on Dandelion encoding. Specifically, the IVR menu is considered to be a service tree, where each announcement (sub menu) is a node and each service is a leaf. The algorithm assumes that the number of options in each announcement is constant and that all the announcements have the same duration. The algorithm aims to reduce the average time to reach a desired service. If M

is the number of services, t_i is the time required to reach a certain service i and p_i is the probability that a customer will ask for service i , then the optimal design of the IVR will minimize the function $f(T) = \sum_{i=1}^M t_i p_i$. The basic idea behind the algorithm is to associate the most frequently requested service with the shortest path. The problem is similar to assigning code words to a set of messages to be transmitted, so that the mean length of the code word is minimized. The suggested algorithm was tested in a real call center of an Italian mobile telecommunications company and in synthetic experiments. The results showed that the algorithm is able to improve the results of Huffman approach (Huffman coding provides near optimal performance for the source coding problem) and obtain results which are very close to a lower bound (entropic bound) for the problem.

Qi, Liu and Li [13] presented a multi-dimensional regression algorithm which aims to reduce the average number of nodes a customer must go through in order to reach a desired service. The main purpose is to use actual information about customer's histories in the IVR and to place more frequently visited nodes at the top layers.

2.3 Modeling a Call Center with an IVR.

2.3.1 "Performance Analysis of a Call Center with Interactive Voice Response Units"

Srinivasan, Talim and Wang [16] used a Markovian model of a call center with an IVR to determine the number of trunk lines (N) and agents (S) required to meet a certain service level, which is defined by the probability of a wait for an agent (after the IVR service) and the probability that an arriving call will be blocked (all lines being busy). The model assumptions are as follows: The arrival rate is a Poisson process with constant rate λ . If a call arrives to the call center and all trunk lines are occupied, the call is lost. Otherwise, the call spends some time in the IVR and then can either request an agent service with probability p , or leave the system with probability $1-p$. If there is no available agent the call will join the queue. The IVR processing times are i.i.d exponential random variables with rate θ . The agent service times are i.i.d exponential random variables with rate μ . The model in [16] does not include abandonment from queue.

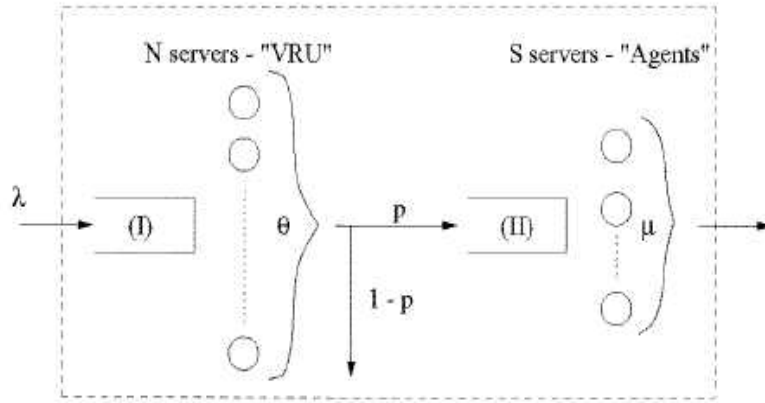


Figure 2: Call Center Model with N Trunk lines (VRU/IVR) and S Agents.

The model can be addressed as a Flow-Controlled Jackson Network (FJN) and can be converted into a three-nodes closed Jackson network, which has a product form solution for its stationary distribution. From the stationary probability P_k (for $k \leq N$), which means that there are exactly k calls in the system (in the IVR, in the queue or being served by an agent), one can derive the probability that all the lines are occupied - P_N , which, according to the PASTA property is also the loss probability. The next step is finding the distribution function of the waiting time - $W(t) := P(T \leq t)$, $t \geq 0$. With P_N and $W(\cdot)$ in hand, and specified values of λ , θ , p , and μ , one could find the number of trunk lines (N) and the number of agents (S), subject to $P_N \leq \epsilon_1$ and $P(T \leq \tau) \geq \epsilon_2$, for pre-defined ϵ_1 , ϵ_2 and τ .

2.3.2 "Designing a Call Center with an IVR"

Khudyakov, Feigin and Mandelbaum [11] used the model presented by Srinivasan, Talim and Wang [16] and performed asymptotic analysis in the QED regime. The asymptotic analysis provided QED approximations of frequently used performance measures (such as the waiting probability, the probability of a busy signal and the mean waiting time given waiting). Those approximations can help solve the staffing and trunking problem and be used to support the operation management of a call center. The approximations were validated against data taken from a large US bank call center [7]. The comparison of the approximate performance measures derived from the model to the real data gave satisfactory results. In many intervals, the theoretical values of the model were very close to the real data.

In addition, Khudyakov, Feigin and Mandelbaum [11] expanded the model presented by Srinivasan, Talim and Wang [16] and added customer impatience to the model. Srinivasan and Wang [23] also expanded their model [16] by adding abandonment from queue. The model assumption is that each call abandons the queue after an independently exponentially distributed amount of time. Naturally, when adding abandonment to the model, the number of required lines (N) and required agents (S) will be smaller. Furthermore, customers' abandonment can decrease the probability of a busy signal and the waiting time.

2.3.3 "Analysis of Customer Patience in a Bank Call Center"

Feigin [8] presented a statistical analysis of customer patience in a call center of a US bank. The queue analyzed is the one that customers join in order to receive agent service, after completing the interaction with the IVR and the post-IVR phase. The post-IVR phase may consist of announcements, made by the system, which are designed to warn the customers of heavy load in the system. In times of heavy system load, the system announces an expected waiting time (<1 min, 1 min, 2 min, etc.) and recommends that the customer returns to the IVR. It was shown that less than 1.5% of the customers who heard the announcement actually returned to the IVR.

One of the factors that may affect the customers patience is how much time they have already invested in the call, namely, how much time they spent in the IVR and the post-IVR phases. Feigin compared the patience of customers who spent short time in the IVR (<100 sec) with the patience of customers who spent longer time in the IVR, by analyzing their survival probability. There was a clear separation between the two survival curves. Customers who have invested more time in the IVR were more patient while waiting for an agent service (see Figure 3). This fact has some ramifications on the operational management of the system, especially when aiming to minimize abandonment. It suggests that the priority of customers entering the agent queue should be inversely related to the time they had already spent in the IVR and in the post-IVR phase. This means that customers who tend to be less patient in the IVR and post-IVR phases will wait less in the agent queue, and those who tend to be more patient would be allowed to wait longer. Histograms and QQ-plots of the IVR time distribution and the

total invested time distribution (IVR + post IVR), revealed that there is quite a good fit of the IVR time to the lognormal distribution but the fit of the total invested time to the lognormal distribution is not adequate.

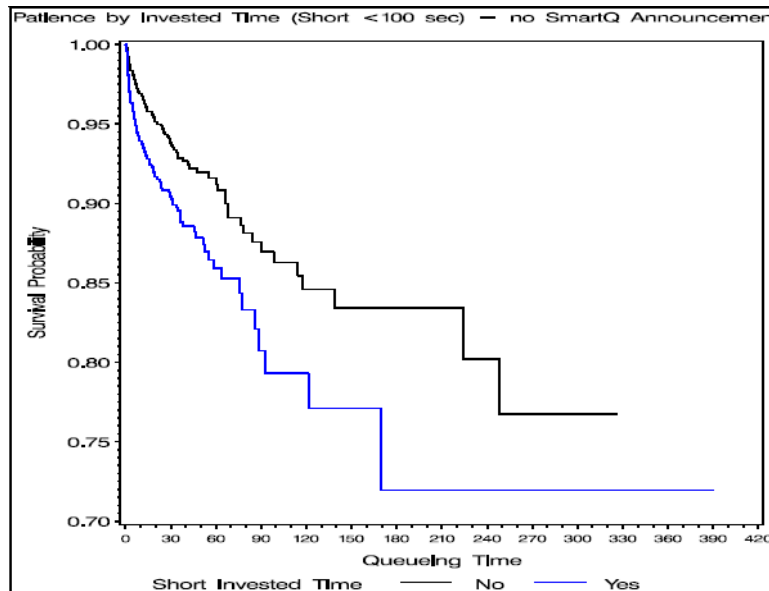


Figure 3: Survival curves for customers who invested short time (<100 sec) in the IVR and post-IVR stages compare to those who invested 100 sec or more.

2.3.4 "Robust Design and Control of IVR Systems in Call Centers"

Behzad and Tezcan [4] proposed a model of a call center, that has two IVR systems with different performance measures. The performance measures are: call resolution probability – proportion of calls completing the service in the IVR; Opt-out probability – proportion of calls that were transferred to a live agent; Abandonment probability – proportion of calls abandoning from the IVR. The IVR performance measures are effecting the call center staffing. For example, if the call resolution is high, then fewer agents are needed and vice versa. Therefore, IVR design must be synchronized with call center staffing. In the model, every call can be routed into one of the two IVR systems. A two-stage stochastic program was formulated in order to determine the optimal staffing level and the proportion of calls that should be routed to each IVR system, in a way that will minimize total costs. The total costs consist of agents cost and abandonment cost. The most interesting scenario is when one IVR system is more efficient than the other, meaning that the call resolution probability is higher and

the abandonment probability is higher. The intuition behind this scenario originates from the fact that if it is more difficult to reach the option of opting out to an agent, it is more likely that customers will find the answer to their problem while navigating through the IVR but may also lead to higher abandonment probability. In this scenario, there are two critical values of arrival rate λ_1, λ_2 which determine the routing to each one of the IVR systems. When the arrival rate is lower than λ_1 , calls will be routed only to the less efficient IVR system. When the arrival rate exceeds λ_2 , calls will be routed only to the more efficient IVR system. When the arrival rate is between λ_1 and λ_2 , the routing should bring to a full utilization of the agents. The numerical experiments showed that using the presented model can bring to an average reduction of 8% in the total costs.

2.4 Fitting Phase Type Distributions.

2.4.1 "Fitting Phase-Type Distributions to Data from a Telephone Call Center"

Ishay [10] analyzed service times and customer patience by modeling and fitting phase-type distribution to data from an Israeli bank call center. By analyzing system data and modeling the system, one could then optimize the system performances. Ishay fitted several phase type distributions to the data, according to different customer priorities and service types. The empirical survival, density and hazard functions were plotted against the fitted functions to visually examine the fit. Kolmogorov-Smirnov and Anderson-Darling goodness-of-fit tests were implemented to assess whether a particular phase-type distribution provides an adequate fit to the data. Two different customer priorities and four different service types were examined. It was shown that the best fitted phase type model is different for each priority and each service type. Fitting the phase type distribution to the data was done via the EM-algorithm, using the EMPHT-program as describes by Asmussen, Nerman and Olsson [2]. The EM (expectation-maximization) algorithm is an iterative maximum likelihood method for estimating the parameters of the phase-type distribution – (q, R) . The basic idea of this approach is that the class of phase-type distributions, for a fixed number of transient states, can be viewed as a multi-parameter exponential family, assuming that the whole underlying absorbing Markov process is observed. The data, in practice, consists of i.i.d replications of the absorption

times Y_1, \dots, Y_n and therefore is incomplete in the sense that it only tells us when the process reached the absorption state but it does not provide any information about where it started and which of the states it visited and for how long. In this setting of incomplete observations it is natural to implement the EM algorithm. The EMPHT-program is a program for fitting phase type distributions either to a sample or to another given continuous distribution, using the EM-algorithm.

2.4.2 "A Novel Approach for Fitting Probability Distributions to Real Trace Data with the EM-Algorithm"

Thummler, Buchholz and Telek [22] presented an EM-algorithm for fitting hyper-Erlang distributions. It was shown that mixtures of Erlang distributions of unlimited order are, theoretically, as powerful as acyclic or general phase type distributions, in the sense that any probability density function of a non-negative random variable can be approximated by a hyper-Erlang distribution. Nevertheless, this class of distributions allows the realization of a very efficient fitting algorithm, due to the fact that the fitting time is independent of the number of states and depends only on the number of Erlang branches, which may be significantly lower than the number of states. Experiments with the proposed EM-algorithm showed that hyper-Erlang distribution can be fitted more efficiently and, in most cases, more accurately than acyclic phase type distributions. These results are achieved probably due to the fact that the hyper-Erlang class has more restricted structure which does not reduce its fitting flexibility. The presented EM-algorithm is implemented in a software package called G-FIT, which is available for download at: <http://ls4-www.cs.tu-dortmund.de/home/thummler/>.

2.4.3 "Flowgraph Models for Generalized Phase Type Distributions Having Non-Exponential Waiting Times"

Huzurbazar [9] presented a generalization of the exponential based phase-type models to semi-Markov processes. The generalization is based on flowgraph models which provide greater generality in model specification by allowing semi-Markov processes with non exponential waiting times. These models also accommodate feedbacks and loops as well as multiple inputs and outputs. In flowgraph models, nodes

represent actual system states and branches represent transitions between nodes. Each branch is labeled with a transmittance. A transmittance consists of the probability of taking that branch times the moment generation function (MGF) of the waiting time before reaching the next state. The use of transmittances enables one to solve the overall MGF of the process and also provides the MGF of the phase-type model for any initial state to any end state. The analysis of simple Markov models can be performed by exact inversion of the MGF, while for complicated Markov and semi-Markov processes an approximate inversion using saddle point method is suggested. Even though these approximations are restricted to distributions having tractable MGFs, this is a rich class and includes commonly used distributions such as exponential, gamma, Weibull and inverse Gaussian.

3. Problem definition and research objectives

We consider a call center with an IVR system. A customer usually starts the service in the IVR and then, if necessary, continues to an agent service. The call may either be connected immediately or queued. If the customer's waiting time in the queue exceeds the customer's patience he will abandon. After receiving a service from an agent, the customer might be transferred to a different agent, back to the IVR or simply exit the system. A customer service in the IVR usually constitutes several "events" such as identification and different queries.

Modeling and analyzing the customer flows within the IVR is important in order to optimize IVR design – discover usability problems, shorten service durations, raise the proportion of customers completing the service successfully, decrease the proportion of customers seeking an agent service and improve routing to the agents so that customer actions in the IVR will route them directly to the right group of agents. One way to analyze customer flows within the IVR is via user-path diagrams as described by Suhm and Peterson [19]. Similarly to the user-path diagrams, Khudyakov et al. [12] constructed a flow chart describing the transition in the IVR, based on one day data from an Israeli commercial call center (see Figure 4). The chosen day was a typical Friday in August 2008. Only the first visit of each customer to the IVR was taken into account (as opposed to a returning visit). There were 50,085 such calls on that day. Each customer transaction

in the IVR is referred as an “event”, and there are 14 different “event” types. When customers complete one of the events, they can continue to another one. These possible transitions are denoted by arrows in the chart. The number near each arrow reflects a number of such transitions during the Friday being analyzed. A distribution of the number of visits to different events and the mean and standard deviation of the event durations are presented in Table 1.

The flow chart in Figure 4 and Table 1 reveal interesting issues that could be addressed. First, we observe that some events occur much more frequently (2000 times and up) than others (less than 50). There are several explanations for events with small number of visits: low demand for such operations; customers do not know about these services; customers prefer to implement these services in another way (on the website, in a branch, with a live agent, etc); the way to achieve these services is unsuccessful [12]. Each one of these reasons may be a trigger to a change in the IVR design. Second, almost 10% of the customers end their call directly after the "Identification" phase. Those customers leave the system without getting any useful information, because they are, probably, unable to identify themselves. This is also a problem that should be considered in the IVR design. Moreover, when calculating the fraction of calls completing their service in the IVR, these calls can not be defined as completed in the IVR only, but are actually uncompleted. Third, we note that a large number of customers reach event 5, which records transition to an agent. Some customers reach this event without visiting any other event (298 instances) or directly after the identification (4346 instances). As agents' salaries constitute about 2/3 of the call center costs, one of the goals in the IVR design is to reduce the transitions to agents as much as possible, but without hurting customer satisfaction.

Type of event	# of events	% of events	Mean	St. Dev.
7	72866	48.72%	53	67
3	52468	35.08%	39	93
5	14062	9.40%	62	84
10	5857	3.92%	55	74
11	2173	1.45%	102	140
4	1488	0.99%	26	97
8	187	0.13%	124	91
15	155	0.10%	102	111
1	133	0.09%	129	86
9	65	0.04%	188	97
2	39	0.03%	137	72
6	29	0.02%	117	68
12	18	0.01%	220	88
16	13	0.01%	154	91

Table 1: A distribution of the number of visits to different events (1.08.08)

Table 1 shows that some of the events could be exponentially distributed, as their coefficient of variation (CV) is quite close to one (for example events 15 and 7), while others are certainly not.

IVR service durations and customer paths in the IVR depend on various factors, such as the customer type or priority, the service the customer requires, the customer familiarity with the system and so on. These factors also effect the durations of IVR events.

The following graphs demonstrate the dependency of the IVR service durations as well as the duration of specific IVR events (Identification, Query, etc.), upon various factors [12]. The data was taken from an Israeli commercial call center. The graphs were created by Khudyakov and the SEELab team, using SEEStat program (for more information regarding SEELab and SEEStat, refer to Section 5). Figures 5-6 shows the distribution of service times for calls served only by the IVR (as opposed to calls that received agent service after the IVR). Figure 6 reveals that customers with different priorities have different service distributions. Figures 7-9 present the distribution of service time provided by the IVR for customers of different priorities requesting different types of service. A customer can have one of the following priorities, defined as the customer's type (in descending order): Private; Star; Rainbow; Unknown. There are also

several types of services that the agents can provide. The customer's problem, or request, defines the service type required. From Figures 7-9 it is clear that the IVR service time distribution is influenced not only by the customer types, but also by the service they are requesting.

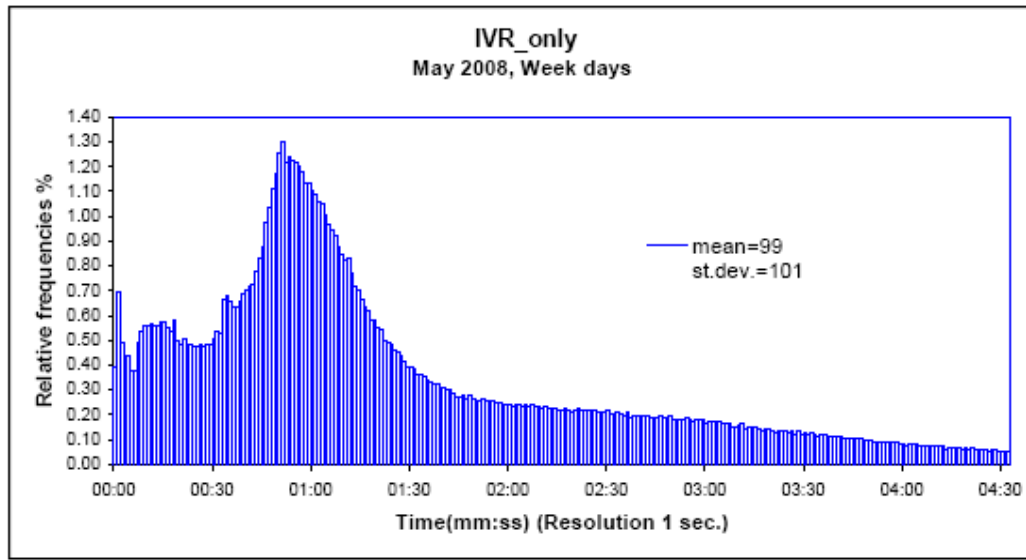


Figure 5: Distribution of service time for calls served only by the IVR.

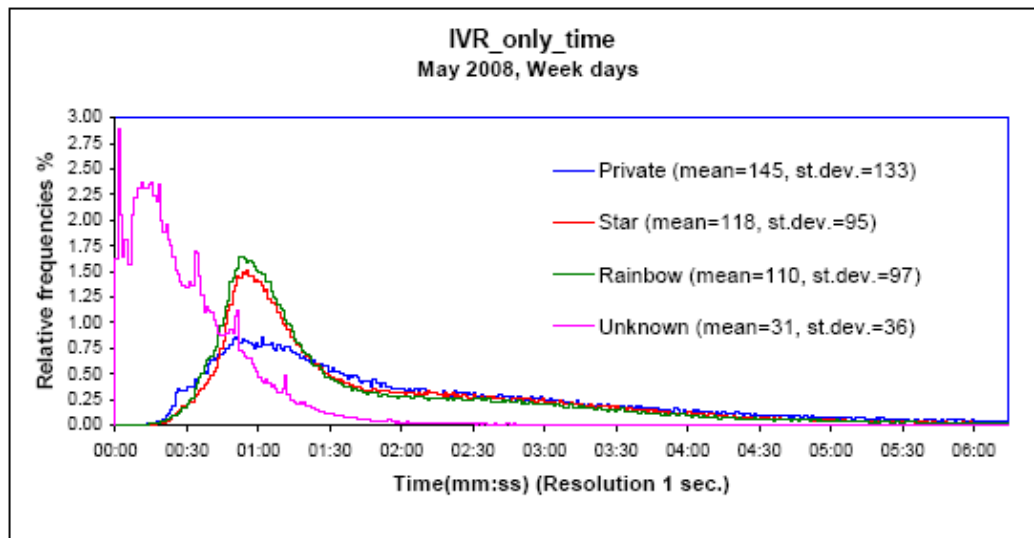


Figure 6: Distribution of service time for calls served only by the IVR, for each type of customers

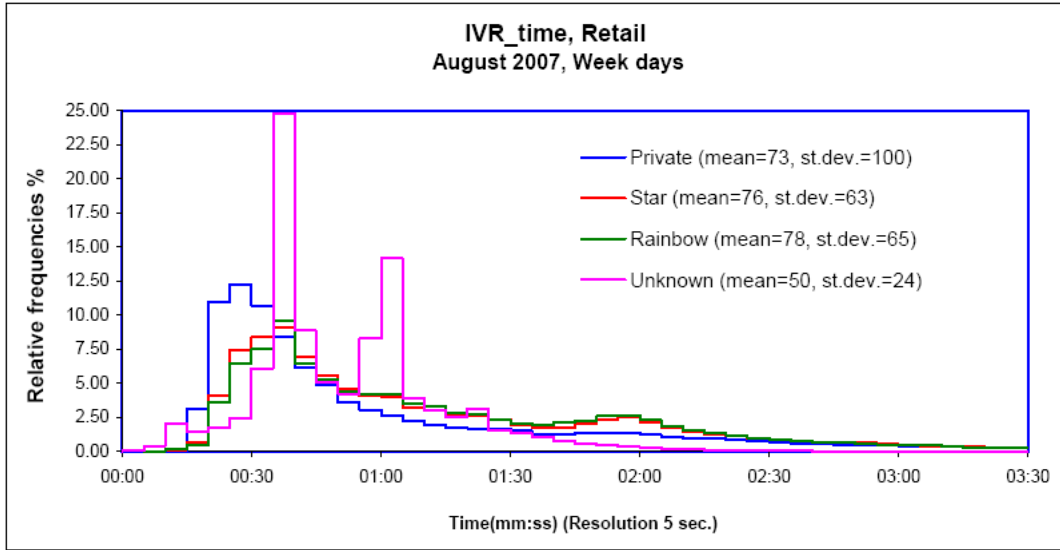


Figure 7: Distribution of service time provided by the IVR, for customers of different types requesting "Retail" service.

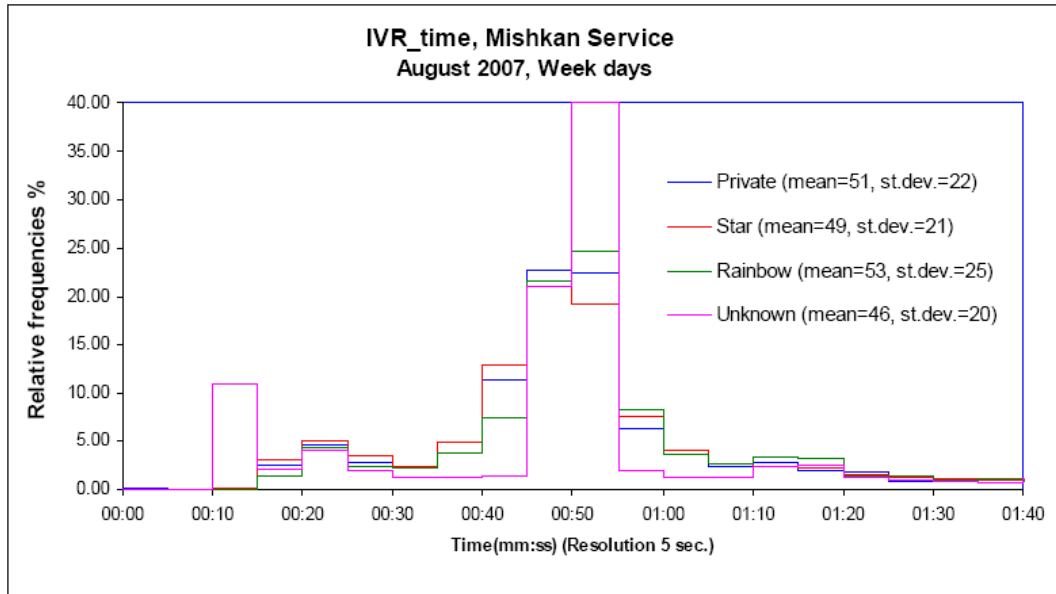


Figure 8: Distribution of service time provided by the IVR, for customers of different types requesting "Mishkan" service.

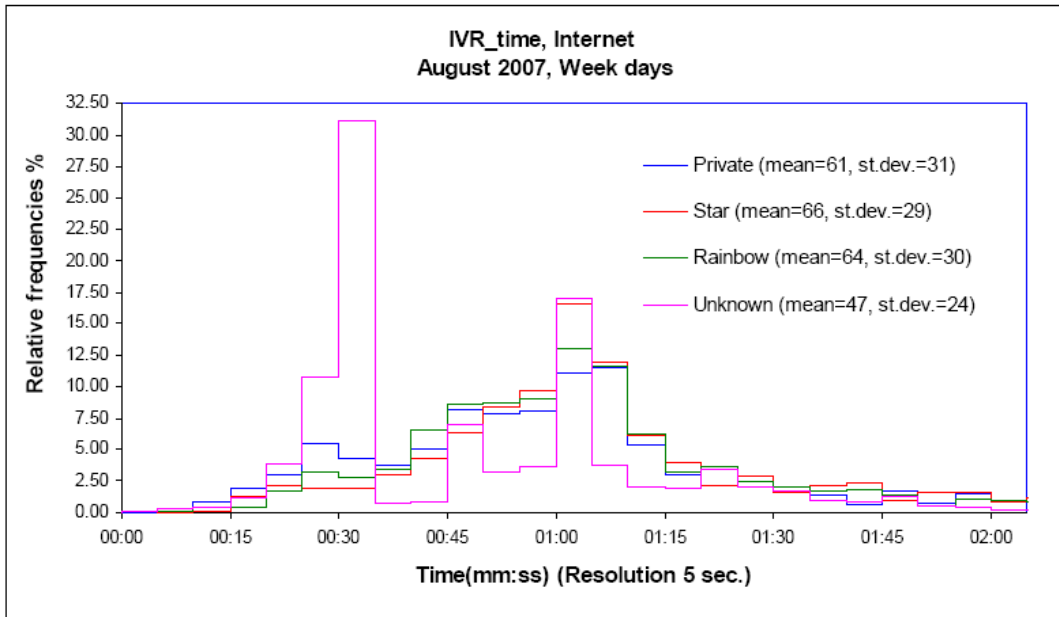


Figure 9: Distribution of service time provided by the IVR, for customers of different types requesting "Internet" service.

Figure 10 presents the distribution of event durations for different types of customers. It depicts that different events have different distributions and that, within each event, different customer types have different distributions. It is also clear that the event durations, for most of the events, are not exponential.

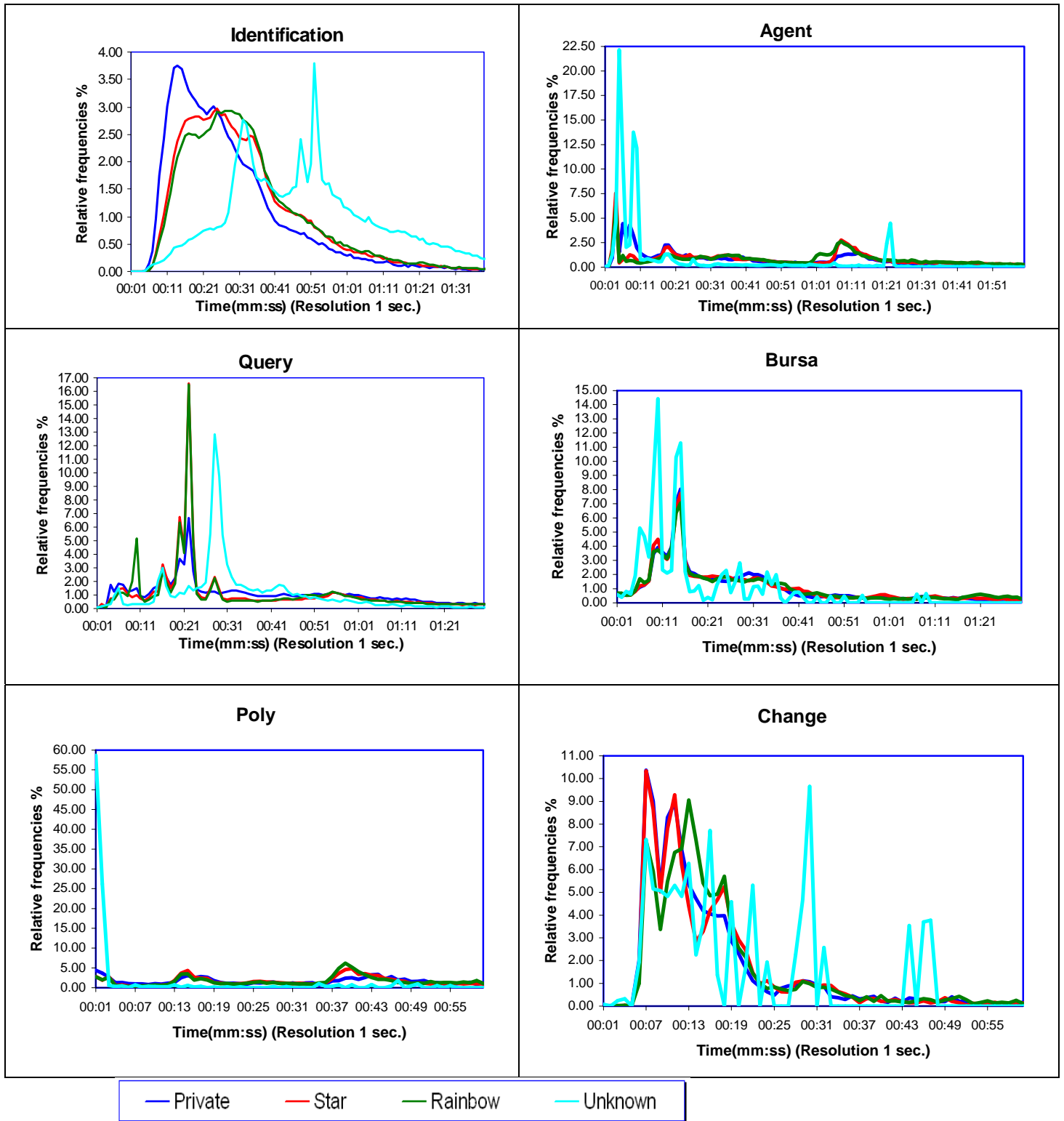


Figure 10: Distribution of event duration for different types of customers

In our research we shall first create a stochastic model for customer flow within the IVR. In the second step, the model will be applied to IVR design. So far, most research on operational design has been deterministic. A comparison between the deterministic and stochastic-based design is thus of interest, and perhaps new design methodologies can be considered. Another existing research involves Human-Factor-Engineering (HFE). We shall target our design research so that it also incorporates and addresses HFE issues.

4. Methodology

Our objective in this research is to first create a stochastic model for customer flows within the IVR. We shall then use this model, combined with HFE principles and, possibly, known deterministic-based design, to seek optimal IVR designs.

The properties of phase-type distributions make it a natural candidate for a stochastic model of customer flows within the IVR. There are some challenges that must be addressed in the research:

1. Service times in the IVR depend on various factors such as customer types, customer needs, customer experience and others.
2. In some states, there is a dependency between the choices in state i and the history (the states that the customer visited prior to reaching state i). Therefore, direct Markov chain assumption does not hold [12].
3. Event durations are not necessarily exponential and are also depend on customer type and experience.

Our analysis will be based on real data from a call center with an IVR (as described in Section 5). In order to address the first challenge, we shall start by sorting the data according to the following categories: customer types; customer call count for a specific period of time (e.g. first customer calls, second customer calls, etc., for a specific day); the call result – whether the call was completed in the IVR or transferred to an agent; and the service type the customer received from an agent (Retail, Mishkan, etc.). We can then identify, for each group, the most common customer paths in the IVR. Using these paths will help setting the initial parameters for fitting each of the data groups to some phase-type distribution via the EM algorithm, using the EMPHT-program. Another possible

fitting may be attempted with the G-FIT approach, using the EM-algorithm for fitting hyper-Erlang distributions [22]. Additional procedures for distribution fitting, developed in the SEELab, could be applied.

The other challenges mentioned above suggest that reaching a useful fit with a reasonable number of phases, which may also have a physical interpretation, might be difficult. One possible way to overcome the fact that event durations are not necessarily exponential is by using flowgraph models, as suggested by Huzurbazar [9]. In order to do so, we shall first analyze the event durations. This analysis should also be done after sorting the data into different categories, similar to the categories we will use to analyze the complete IVR service times. Using SEESat (see Section 5), we shall try to fit each event to a known distribution or a mixture of distributions. Another approach that we shall explore is fitting the data to a mixture of phase-type distribution with deterministic phase durations.

Fitting phase-type models or equivalent models to customer flows within the IVR will enable us to analyze the different states (events) in the IVR and their influence on service duration and success. We can examine the influence of changing the initial distribution, the transition rates or the menu structure. For example, the flow chart in Figure 4 reveals that, on the day analyzed, more than 98% of the customers started the IVR service in the "Identification" event. Therefore, it could be interesting to analyze how a change in the initial distribution would effect the total service time and the process success. On one hand, there might be IVR services that do not require identification, and providing direct access to them will lead to shorter service duration, and perhaps even decrease in abandonment rate from the IVR. On the other hand, lowering the probability to start the service in the "Identification" event may cause other problems in the customer flows within the IVR and lead to opposite results. Another possibility that should be examined is shortening certain state durations, especially if those states are visited with high probability.

HFE research addressed the issue of the IVR menu structure, mainly comparing between broad (shallow) menus and narrow (deep) menus [5, 15, 18]. Later research concluded that broad menus are better than deeper ones in terms of response rate, routing rate, task performance speed and user satisfaction. An interesting research

direction would be to compare those results with our model insights regarding the optimal states position and the preference of one menu structure over the other. In addition, a combination of HFE recommendation regarding menu structures, combined with the conclusion of the model analysis, could result in a better, more comprehensive, methodology to optimally design IVRs.

Combining the model insights with deterministic-based models can also lead to useful results. We have surveyed some algorithms that intend to reduce the IVR service times, based on placing the most popular services at the top layers [13, 14]. These algorithms assume deterministic and equal state durations and consider only the probability that a customer asks for a certain service, but do not consider the relation between the states as described in the transition rates. Expanding these algorithms by adding the distribution of states duration and the transition rates, as revealed from the stochastic model, or combining the algorithm results in the stochastic based IVR design, can lead to an improved design.

The stochastic model can also help identify usability problems and improvement opportunities. For example, identify calls that leave the IVR without getting any information and identify states that have high probability to lead the customers to an agent service or to abandonment.

5. Data Source

The Center for Service Enterprise Engineering (SEE) was established in February 2007, within the Faculty of Industrial Engineering and Management, at the Technion. The goal of SEE is the development of engineering and scientific principles that support modeling, design and management of Service Enterprises, for example: financial services (banking, insurance); health services (hospitals, clinics); government and teleservices (telephone, internet). SEE's main activity, through the SEE Laboratory (SEELab), consists of designing, maintaining and analyzing an accessible repository of resources and data from telephone call-centers, which has been called project DataMOCCA (Data MOdels for Call Center Analysis). The scope has now been expanded to cover also hospitals, and first steps have been taken to expand into internet services as well. The aims of the DataMOCCA research project are to provide the

infrastructure for, as well as to conduct, analyses based on individual call histories [21]. The graphical user interface, SEEStat, provides real-time statistical analysis, spanning seconds-to-months resolutions, with typical few seconds response time. The SEEStat tool also enables statistical analysis of imported raw data. The SEEStat tool is available at: <http://seeserver.iem.technion.ac.il/see-terminal/>.

In order to model an IVR, we shall analyze data from a call center with an IVR, using DataMOCCA and its user interface, SEEStat. We shall analyze data from a call center of an Israeli Bank. This database fits our needs since it covers more than two years data of daily IVR logs, including the customer type; the service required; how many calls the customer had already made that day; the different events the customer experienced in the IVR and their durations. The database includes also complete calls data and agents data.

Table 2 presents a numerical summary of the database for a one-year period, from May 2007 to April 2008. Table 3 presents, for the same time period, a distribution of the average number of calls for each service type [12].

	Total	% out of total	Average per Weekday
Total # of arriving calls	25,990,281	100%	90,696
# Requesting agent service	9,041,425	35%	31,987
# Served only by IVR	16,948,856	65%	58,709

Table 2: Overall summary of calls.

Service Type	Average # of calls	% calls
Retail	25,169	78.68%
Securities	2,548	7.97%
Russian	1,602	5.01%
Mishkan Service	1,194	3.73%
Internat	773	2.42%
Mishkan Confirmation	378	1.18%
Exchange	141	0.44%
New customer	111	0.35%
Pilot	61	0.19%
Private Interaction	10	0.03%
Total	31,987	100%

Table 3: A distribution of the average number of calls for each service type.

Table 4 presents a distribution of calls made during Friday, August 1, 2008, according to different customer types.

Customer Type	Number of calls	% calls
Private	7,253	14.46%
Star	19,476	38.83%
Rainbow	15,363	30.63%
Unknown	8,064	16.08%
Total	50,156	100%

Table 4: Distribution of calls during August 1, 2008, according to different customers types.

References

- [1] Aksin Z., Armony M. and Mehrotra V. **The modern call-center: A Multi-Disciplinary Perspective on Operations Management Research**. Production and Operations Management, 16, 665-688, 2007.
- [2] Asmussen S., Nerman O. and Olsson M. **Fitting phase-type distributions via the EM algorithm**. Scandinavian Journal of Statistics, 23, 419– 441, 1996.
- [3] Aspect Communications Corporation. **Why Your Customers Hate Your IVR Systems**. White paper, 2003. Available at: http://www.apaccustomerservices.com/uploadedFiles/knowledge/White_Papers/2338-A_Aspect_IVR_wp.pdf
- [4] Behzad B. Tezcan T. **Robust Design and Control of IVR Systems in Call Centers**. Submitted July 2010.
- [5] Commarford P.M., Lewis J.R., Smither J.A. and Gentzler M.D. **A Comparison of Broad Versus Deep Auditory Menu Structures**. Human Factors, 50(1), 77-89, 2008.
- [6] Delorey E. **Correlating IVR Performance and Customer Satisfaction**. 2003. Available at: http://www.easyivr.com/tech-ivr-applications_108.htm
- [7] Donin O., Feigin P.D., Ishay E., Khudyakov P., Mandelbaum A., Nadjarov E., Trofimov V. and Zeltyn S. **DATA-MOCCA: Data Model for Call Center Analysis. The Call Center of “US Bank”**, vol. 4.1, Technical report, Technion, Israel Institute of Technology, 2006. Available at: http://ie.technion.ac.il/Labs/Serveng/SEE_Documents_List.php

- [8] Feigin P.D. **Analysis of Customer Patience in a Bank Call Center**. Working Paper, Technion, Israel Institute of Technology, 2006.
- [9] Huzurbazar A.V. **Flowgraph Models for Generalized Phase Type Distributions Having Non-Exponential Waiting Times**. Scandinavian Journal of Statistic, 26, 145-157, 1999.
- [10] Ishay E. **Fitting Phase-Type Distributions to Data from a Telephone Call Center**. M.Sc. Thesis, Technion IE&M, 2003.
- [11] Khudyakov P., Feigin P.D. and Mandelbaum A. **Designing a Call Center with an IVR**. Queuing Syst, 66, 215-237, 2010.
- [12] Khudyakov P., Mandelbaum A., Trofimov V., Maman S., Nadjharov E., Gavako I., Liberman P. and Kutsi K. **Empirical Analysis of a Call Center in an Israeli Commercial Company**. Technical report, Technion, Israel Institute of Technology, 2008. Available at http://ie.technion.ac.il/Labs/Serveng/SEE_Documents_List.php
- [13] Qi C.H., Liu J. and Li H. **Application Research of a Statistical Regression Algorithm in the IVR System**. International Conference on Educational and Network Technology (ICENT), 358-360, 2010.
- [14] Salcedo-Sanz S., Naldi M., Perez-Bellido A.N., Portilla-Figueras J.A. and Ortis-Garcia E.g. **Evolutionary Optimization of Service Times in Interactive Voice Response Systems**. IEEE Transactions on Evolutionary Computation, 14(4), 602-617, 2010.
- [15] Schumacher R.M., Hardzinski M.L. and Schwartz A.L. **Increasing the Usability of Interactive Voice Response Systems: Research and Guidelines for Phone-Based Interfaces**. Human Factors, 37(2), 251-264, 1995.
- [16] Srinivasan R., Talim J. and Wang J. **Performance Analysis of a Call Center with Interactive Voice Response Units**. Top, 12(1), 91-110, 2004.
- [17] Suhm B., Bers J., McCarthy D., Freeman B., Getty D., Godfrey K. and Peterson P. **A Comparative Study of Speech in the Call Center: Natural Language Call Routing vs. Touch-Tone Menus**. Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 2002.

- [18] Suhm B., Freeman B. and Getty D. **Curing the Menu Blues in Touch-Tone Voice Interfaces.** CHI '01 extended abstracts on Human Factors in Computing Systems, 2001.
- [19] Suhm B. and Peterson P. **A Data-Driven Methodology for Evaluating and Optimizing Call Center IVRs.** International Journal of Speech Technology, 5, 23-37, 2002.
- [20] Suhm B. and Peterson P. **Call Browser: A System to Improve the Caller Experience by Analyzing Live Calls End-to-End.** Proceedings of the 27th International Conference on Human Factors in Computing Systems, 2009.
- [21] Trofimov V., Feigin P.D., Mandelbaum A., Ishay E. and Nadjarov, E. **DATA-MOCCA: Data Model for Call Center Analysis. Model Description and Introduction to User Interface**, vol. 1, Technical report, Technion, Israel Institute of Technology, 2006. Available at:
http://ie.technion.ac.il/Labs/Serveng/SEE_Documents_List.php
- [22] Thummler A., Buchholz P. and Telek M. **A Novel Approach for Fitting Probability Distributions to Real Trace Data with the EM-Algorithm.** Proceedings of the 2005 International Conference on Dependable Systems and Networks.
- [23] Wang J., Srinivasan R. **Staffing a Call Center with Interactive Voice Response Units and Impatient Calls.** Proceedings of 2008 IEEE International Conference on Service Operations and Logistic and Informatics, 2, 514-519, 2008.