Hybridized Optimization Framework for Routing Calls in Call Centres

Mughele Ese Sophia
Department of Computer Science, Delta State School of Marine Technology Burutu, Delta State, Nigeria
Email: prettysophy99@gmail.com

Stella Chiemeke, Ph.D
Department of Computer Science
University of Benin, Benin City, Edo State, Nigeria
Email: schiemake@yahoo.com

Konyeha Susan Ph.D
Department of Computer Science
University of Benin, Benin City, Edo State, Nigeria
Email: susan.konyeha@gmail.com

Kingsley C. Ukaoha, Ph.D
Department of Computer Science
University of Benin, Benin City, Edo State, Nigeria
Email: kingsley.ukaoha@uniben.edu

Dorothy Akpon-Ebiyomare, Ph.D
Centre for Distance Learning
University of Benin, Benin City, Nigeria
Email: ebiyomare77@gmail.com

Abstract—The major challenges in a call centre with respect to customer satisfaction has to do with waiting time on the queue for long period of time before they are attended to. Beyond the problem of queue is the nature of the service itself, the effective resolution of the customer issues. The major challenge there is to determine a routing rule that can reduce waiting time and as well enhance call resolution rates. In this study, we conducted simulation analysis using Java simulation library on seven routing rules, four for waiting time and three on call resolution (CR) rate oriented routing rules, in a bid to determine the optimal rule. The data used for the simulation was collected from a call centre of a telecommunications outfit in Nigeria. The result from the simulation gave the optimal rules for waiting time and CR rate routing rule. A hybrid framework was developed from the outcome of the simulation result. The proposed routing rule will be able to achieve low wait-time and enhanced call resolution, this will improve and optimize call centre operations as well as increase customer’s satisfaction and brand loyalty.

Keywords—Optimization; hybrid rule; routing rule; queue; call centre; call resolution

I. INTRODUCTION

Businesses create value through product offerings or service delivery to their customers. For customers to have access to these products and services sometimes they spend long time waiting in the queue before the service is delivered, in some other instances, the customer is abandoned on the queue without accessing the desired services. It is therefore, the concern of every company management to render prompt delivery of service, eliminate waiting queue and give value for money so as to ensure customer satisfaction and loyalty.

Call centre can be defined as any group whose business is talking to customers or prospective customers through the telephone. A call centre is a system that offers complete management of all communication channels between a business and its customers, optimising polices, eliminating duplicated work and making better use of time [1]. But in recent times, customers waiting for so long in order to lodge a complaint or make an enquiry have become a worrisome phenomenon in the call centres especially in telecommunications. In contemporary society, satisfying customers need has become a phenomenon that is highly inevitable for businesses that wants to survive in this era of high competition amidst the global financial crisis. A customer’s experience during a call service encounter consists of two parts, namely, the time spent waiting for the service and the service itself. Call centres give priority to the two criteria with emphasis on one more than the other. Those that place more emphasis on time spent waiting for the service are more concerned with reducing the average time involved in handling a call while those that are concerned with the service itself aims at effective resolution of customer issues.

For a call centre to reduce waiting queue with emphasis on the reduction of time spent, it is best to route calls to agents who can handle customer issues the fastest, sometimes even holding a call in queue to wait for that agent than routing the call to a slower agent. This might lead to further increase in congestion, repeat calls from unresponsive issues and undue burden on some agents [2]. To reduce waiting lines, emphasis should be on the service itself, that is; call resolution. It is best to route calls to agents who resolve customer issues, sometimes holding a call in queue to wait for such agent. This might also lead to increase in congestion and undue burden on some agents [3].

This paper tries to eliminate the challenge of wait-time oriented rule being superior to call resolution oriented routing rule and verse versa. We propose a framework that hybridizes the optimal rule for wait-time and call resolution rate routing rules in our bid to achieve low wait-time and enhanced call resolution which will improve the overall performance of the call centre and increase customer satisfaction and brand loyalty.
II. RELATED WORK

Customer service call centres have obviously become a very integral part of many organizations’ business operations today, inbound call centres employ millions of agents across the globe and serve as a primary customer-facing channel for many different industries. There has also been a great deal of research interest in call centre operations management, with the extensive and evolving literature thoroughly analysed [1]. This study determines whether average handling time and call resolution are true determinants of operational success of a call centre to reduce waiting queue. It also examine whether emphasis should be on reducing handling time or effective call resolution. In [2], it was noted that the operational challenges from call centres provide a perspective on both traditional and emerging call centre management challenges and the associated academic research. They deployed literature review method and identified a handful of broad themes for future investigation while also pointing out several very specific research opportunities.

Given the size of the call centre industry and the complexity associated with its operations, call centres have emerged as a fertile ground for academic research. A complete review of articles on First Call Resolution (FCR) is provided in [3], while also pointing out the importance of measuring and using FCR. Resolving customer queries the first time around is a commonly shared goal. A company’s business context, human resources strategy, supporting technology and budget constraints influences this key performance indicator (KPI) in many ways, and makes FCR a difficult measure to benchmark. The study established the differing views on the value and measurement of FCR, identifies the main factors affecting FCR and the relationships among these factors, and relates results in a South African context to academic and practitioner.

Operations management has paid comparatively little attention to models and methods for managing routing. However, there are many published papers that describe call routing and resource allocation rules for call centres. In [4], authors observed that customers in a call centre service experiences real time delay as a result of queue and call back delay. This metric affects customer’s perception of the product or service and this impact on customer’s loyalty. Probabilistic choice model was deployed, and the dynamics of the system are modelled as an $M/M/N$ multiclass system. The study justifies that as the number of agent increases, the system’s load approaches its maximum processing capacity.

In [5], authors noted that the challenge in call centre operation is how to determine the relevant control in the routing; that is, the decision concerning which agent should handle an arriving call when more than one agent is available. An inverted-V model setting was designed, and they formulated an optimization problem with the dual performance objective of minimizing average customer waiting time and maximizing the call resolution. They also noted that focusing on minimizing average waiting time as the sole performance objective may not deliver the best customer experience.

In [6], authors discovered that performance metric in call centres are average call handling time (AHT) and call resolution probability (RP). They observed that challenge for call centre managers is to determine how to make use of this information to determine which types of calls should be handled by which types of agents under which system conditions. Routing Rules to improve operational performance on one or both of these output dimensions was examined and simulation experiments were conducted to examine the relative performance of these routing rules.

In [7], authors posited that in a service base call centre, the two key challenges are: 1) where should a call be routed to; and 2) who should handle the call? They deployed base case FIFO approach for the simulation to analyse performance-based routing strategies in call centres. Their work shows the potential for significant improvements in call centre performance especially Average Speed to Answer (ASA). This was achieved by using rules based on historic performance data such as Average call Handling Time (AHT) and First call Resolution (FCR) rates.

In [8], authors noted that as time spent on queue at the call centres increases, it becomes unacceptable for customers, and this affect their satisfaction level. A study was conducted using Univariate Analysis of Variance (ANOVA) to determine customer’s perception of their wait experience at call centres. Their result showed that though the time spent on the queue waiting can lead to customer’s dissatisfaction, nevertheless, it is not as important as the agent’s ability. More so, the concept of routing rules to be deployed for efficient call resolution rate was not emphasised.

In [9], authors observed that the key performance indicators to measure call centre metrics performance such as average speed of answer, cost per call, agent utilization rate, first contract resolution rate, customer satisfaction and aggregate call centre performance are not effectively maximised. He used CallLogic system to improve the fundamental call routing logic of the Northeast Utilities call centres. Although the findings from the CallLogic system led to discoveries and ideas on how to improve the fundamental call routing logic of the Northeast Utilities call centres, the CallLogic project achieved high success in the average call handling time.

The quality of service accessibility and customer waiting time are dominant performance measures [10]. Hence capacity planning and call routing software system strive to minimize cost while achieving self-imposed service level constraints, though considering low average time waiting in queue, these approach do not consider the quality of service rendered to
customers [12]. Low quality of service has significant impact on the call centre operations; this operational impact of service failure is often ignored by call centre capacity planning and call routing management system. Their work was motivated by the fact that a major European telecommunications service provider discovered that customers needed to talk to more than three different agents before their problems are resolved.

In [13], authors also observed that when using routing rules that emphasises on reducing queues, calls are quickly routed to agents, without considering the root of the problem being fixed, to avoid a call back of such customers on the same problem. In [10], authors in a related study noted, that call centre managers and decision makers tends to only look for information that simply confirms existing beliefs and often disregard all other information, the authors believes that these will enable such call centre operators implement a convenient routing rule even if it is not the optimal rule.

In [14] and [5], authors conducted study on the concept of customer waiting time on the queue, these researchers focused on queues, staffing and performance analysis which are input into personal scheduling and rostering models. In [15], empirically study the agent’s heterogeneity in Average Handling Time (AHT). Majority of researches conducted in the domain of call centre management were focused on reducing waiting time on the queue and how it impacts on customer’s satisfaction and loyalty. In a related study by [16], they modelled a call centre as an M/M/S+M queue which is developed to determine the behavioural queue model in which customers arrive in and depart from the system based on their satisfaction with waiting time. Their model has two sectors, representing the feedbacks of repeat behaviour of customer and abandonment rate. The performance metric of abandonment is the loss of customers based on waiting time; they further explained that the metric of satisfaction with waiting experiences is used to build a link between staffing costs and call centre customer revenues. They considered a call centre model with a single class of customers made up of homogeneous and parallel agents, the analysis of Process-Related Metrics of Call Centre. The model of the abandonment behaviour was developed by the extension of the Erlang-A formula, which can be viewed as an M/M/s+M queueing system with feedback.

In [17], authors maximises CR routing rules as one of the metrics for call centre operational performance. They modelled an optimization problem that focussed on call resolution, taking into cognisance the work of [18] and [15] that minimised wait time routing rule. However, the limitation observed from their optimization is that their proposed minimised wait time and maximised call resolution was implemented using FCFS and RP. They did not consider the optimal routing rules for wait time and CR oriented rules. In practice to achieve a balanced operational performance in call centre, it requires a routing rule that will minimise waiting time on the queue and also maximise CR rate. Adapting [17], therefore this research evaluates and selects the optimal routing rules for CR and waiting time to develop a hybrid routing rule, in a bit to improve customer satisfaction and enhance operational performance of call centres operations.

## III. METHODOLOGY

A field study Simulation was conducted at a telecommunication outfit in Nigeria on September 2015, and data was collected from the organization’s call centre to carry out simulation analysis using Java Discrete Event Simulation Program. The tools used for the simulation is a collection of Java simulation libraries programs, Microsoft Excel was also used to do some basic data analysis and graphical presentation of results. The data used consist of the following:

1. Largest call types were considered.
2. Restricted the number of agents in model to include only those agents who can handle a certain amount of calls, i.e., a minimum of 100 calls.
3. The database used for simulation included records for incoming phone calls. Specifically, each tuple in the database contained the following fields:
   i. The date and time of the call.
   ii. The unique ID number for the agent who handled the call.
   iii. The Call Type for that call.
   iv. The time spent by the agent on the phone handling the call, or Handle Time (HT).
   v. The resolution status of the call.

A subset of the call types and agents are used to ensure that the run times for the simulations were fast enough to conduct numerical experiments. The same interface was used for the implementation of the simulation, as well as the input data which were also the same for all routing rule.

The input data in Table 1, shows the various service type, number of call offered, analysis of the number of calls answered, abandoned, average speed of answer, average talk duration and other report from the calls offered. The application is a standalone application.

The SSTF rule features the lowest ASA, it also results in a higher CR rate than Shortest Queue Routing (SQR), which suffers only slightly higher ASA values. Taken together, these results clearly demonstrate that optimal routing decisions can have a significant positive impact on operational performance.

Fig. 1 show the weighted average for the average speed to answer for waiting time routing rules, from the graph, SSTF has an ASA of 28 seconds, FCF has 36 seconds, HSTF has 47 seconds and FCFS/LW has 95 seconds. This clearly demonstrated that SSTF has the lowest ASA amongst the waiting time routing rules, and it is also the optimal routing rule. The highest CR rate among SQR, PR and RRPR is SQR (1795) which makes it the optimal routing rule as dipicted in Fig. 1(b), this was also observed by [19]. Fig. 1(a) and (b) show the entire results generated from the simulation in Table 2.
IV. SYSTEM FRAMEWORK

In [4], author noted the importance of Average Speed to Answer (ASA), Service Level(SL) and Abandonment Rates (ASA) for waiting time routing rules. Whereas Call resolution rates during services are key indicators that influence customers’ perceptions and quality of service. The absence of academic literature that covers both call resolutions and waiting time simultaneously was identified in [17].

A. Hybrid Heterogeneous Call Routing Rule (HHCRR)

A rule that combines the optimal rule for call resolution and waiting time routing rule is considered. As identified in simulation conducted for the waiting-time and CR routing rules, SSTF was the optimal for Waiting-Time Routing Rules and SQR for Resolution Probabilistic Routing Rules Assumptions:

A call of a particular type that arrives when agents of multiple matching groups are free will be routed to a matching agent group (j) that has the relatives Shortest Service Time and shortest queue for that call type.

Proportion of call of type i \((p_i)\)

\[
p_i = \frac{\text{Total calls of type } i}{\text{Total calls}}
\]  

Total arrival rate of call type i \((\lambda_T)\)

\[
\lambda_T = t_1\lambda_1 + t_2\lambda_2 + \cdots + t_i\lambda_i
\]

Total arrival rate of call type i in Agent group j \(\lambda_{ij} = x_{ij}\lambda_i\)

Effective arrival of unresolved calls of call type i who call back \((\beta_i)\) (Sisselman and Whitt, 2007)

\[
\beta_i = \frac{\lambda_i}{1 - \sum_j (1 - \theta_j)x_{ij}}
\]

Service rate of Agent group j for call of type i \((\mu_{ij})\)

Service rate of agent group \((\mu_{ij})\) is the reciprocal of the average handling time of all the agents [15].

\[
\mu_{ij} = \frac{1}{AHT_j}
\]

Where, \(\mu_{ij}\) is the service rate of agent group j for call type i, AHTj is the mean call handling time of all agents in group j, and AHTj is average handling time.

Total arrival rate of agent group j for call type i who call back.

\[
\beta_{ij} = \frac{\beta_i x_{ij}}{n_j}
\]

Maximal service rate of the Agent group j:

\[
\rho_j = \frac{\left(\sum_i \beta_i x_{ij}\right)}{n_j}
\]

The condition for stability is \(\rho_j < 1\), that is to say that the mean total arrival rate must be less than the mean

Total service rate of call type i \((\mu_z)\) [20].

\[
\mu_z = \sum_{j=1}^{i} n_j u_{ij} \left(\frac{\lambda_{ij}}{u_{ij}}\right) \left(\frac{u_{ij}}{\sum_i \lambda_{ij}}\right)
\]

Where, \(\mu_z\) is the total service rate of call type i, \(\lambda_{ij}\) is the arrival rate excluding call type i

Total utilization boundaries for call type i [21].

\[
\Gamma_i = \frac{\beta_i}{\mu_z}
\]

B. Model Evaluation

The model intuitively merge the maximization problem of [17] with the minimisation of [18] to form a hybrid which can be solved to determine an optimal feasible solution to
minimise waiting time and maximise call resolution. The flow diagram in Fig. 2 represents the evaluation and feasibility of the hybrid rule.

From the flow diagram, it is observed that the calls come into call queue in random fashion. Features from the call type are extracted based on some predefined logic. These features are used to compute steady state parameters using either Erlang A or B models. These parameters will be combined to form a multiple objective programming (MOP) problem and could be solved using birth-death techniques, greedy algorithm and genetic algorithm. If the max z and min c'x is achieve, then adopt and use results to route call to agent in group capable of handling such call. If the agent is busy, then route the call to the next most performing agent in that group. If the call types are prioritized, route calls to the most significant optimal solution exactly or close to the agent group with the highest service rate and minimum queue length. If the call is not resolve, the call is re-queued and solved further. Hence flow diagram represents the evaluation for the feasibility of the hybrid framework.

V. CONCLUSION

The core aim in this research is to model an optimization problem that considers the optimal rules for CR rates and wait time routing rule simultaneously. The study presented a hybrid rule Hybrid Heterogeneous Call Routing Rule (HHCRR) that minimizes wait time and maximises CR. Our work also developed a proposed system framework to implement the hybrid rule. The framework was able to express a MIN|MAX mathematical programming problem. The proposed framework will not only reduce wait time on the queue and enhance CR rates, also will reduce call abandonment. This paper considered arrival rates as inputs to the proposed framework as time-independent inputs, though in practice all call centres experience different arrival rates at different times of the day, therefore, study done to determine the distribution of delay times prior to call-backs, this can have a significant impact on operational performance. Furthermore, our proposed hybrid routing rule “Hybrid Heterogeneous Call Routing Rule” (HHCRR) can be developed, tested and implemented to improve call centre operational performance.

Fig. 2. Flow diagram for MIN/MAX optimization for framework.
REFERENCES


TABLE I.  INPUT CALL TYPE DATA FOR SIMULATION

<table>
<thead>
<tr>
<th>RULE</th>
<th>CR</th>
<th>ASA (seconds)</th>
<th>Non CR</th>
<th>Resolved Calls</th>
<th>Call Backs</th>
<th>% resolved calls</th>
<th>% Call backs</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCFS/LW</td>
<td>1552</td>
<td>47</td>
<td>448</td>
<td>0.4311111111</td>
<td>0.1244444444</td>
<td>71.85185185</td>
<td>20.74074074</td>
</tr>
<tr>
<td>FCF</td>
<td>1683</td>
<td>36</td>
<td>317</td>
<td>0.4675</td>
<td>0.088055556</td>
<td>77.91666667</td>
<td>14.67592593</td>
</tr>
<tr>
<td>SSTF</td>
<td>1935</td>
<td>28</td>
<td>65</td>
<td>0.5375</td>
<td>0.018055556</td>
<td>89.58333333</td>
<td>3.00925926</td>
</tr>
<tr>
<td>HSTF</td>
<td>1268</td>
<td>95</td>
<td>732</td>
<td>0.3522222222</td>
<td>0.203333333</td>
<td>58.70370370</td>
<td>34.88888888</td>
</tr>
<tr>
<td>SQR</td>
<td>1795</td>
<td>34</td>
<td>205</td>
<td>0.4986111111</td>
<td>0.0569444444</td>
<td>83.10851851</td>
<td>9.490740741</td>
</tr>
<tr>
<td>PR</td>
<td>1775</td>
<td>39</td>
<td>225</td>
<td>0.493055556</td>
<td>0.0625</td>
<td>82.17925925</td>
<td>10.41666667</td>
</tr>
<tr>
<td>RRPR</td>
<td>1685</td>
<td>78</td>
<td>315</td>
<td>0.5236111111</td>
<td>0.0319444444</td>
<td>77.94801101</td>
<td>14.5510</td>
</tr>
</tbody>
</table>
### TABLE II. WEIGHTED AVERAGE RESULTS FOR EVALUATION OBTAINED FROM SIMULATION ANALYSIS

<table>
<thead>
<tr>
<th>Service Type</th>
<th>Calls Offered</th>
<th>Calls Answered</th>
<th>Calls Abandoned</th>
<th>Calls Abandoned in Queue</th>
<th>Calls Abandoned in Ringing</th>
<th>Avg. Speed of Answer</th>
<th>Avg. Abandoned Duration</th>
<th>Avg. Ringing Duration</th>
<th>Avg. Talk Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>3G HIS</td>
<td>672</td>
<td>557</td>
<td>115</td>
<td>83</td>
<td>32</td>
<td>0:00:52</td>
<td>0:01:22</td>
<td>0:00:06</td>
<td>0:03:24</td>
</tr>
<tr>
<td>Blank Calls</td>
<td>6601</td>
<td>2234</td>
<td>4367</td>
<td>4339</td>
<td>28</td>
<td>0:00:53</td>
<td>0:00:17</td>
<td>0:00:04</td>
<td>0:03:25</td>
</tr>
<tr>
<td>Conoil</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0:00:54</td>
<td>0:00:09</td>
<td>0:00:09</td>
<td>0:03:26</td>
</tr>
<tr>
<td>Glo 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0:00:55</td>
<td>0:00:00</td>
<td>0:00:07</td>
<td>0:03:27</td>
</tr>
<tr>
<td>HNI</td>
<td>443</td>
<td>392</td>
<td>51</td>
<td>40</td>
<td>10</td>
<td>0:00:56</td>
<td>0:01:37</td>
<td>0:00:08</td>
<td>0:03:28</td>
</tr>
<tr>
<td>JustDialNew</td>
<td>39</td>
<td>31</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>0:00:57</td>
<td>0:00:27</td>
<td>0:00:02</td>
<td>0:03:29</td>
</tr>
<tr>
<td>NBC</td>
<td>32</td>
<td>24</td>
<td>8</td>
<td>8</td>
<td>0</td>
<td>0:00:58</td>
<td>0:00:34</td>
<td>0:00:04</td>
<td>0:02:40</td>
</tr>
<tr>
<td>Others</td>
<td>38</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0:00:59</td>
<td>0:00:00</td>
<td>0:00:01</td>
<td>0:02:15</td>
</tr>
<tr>
<td>PREMIUM</td>
<td>1552</td>
<td>1330</td>
<td>222</td>
<td>196</td>
<td>26</td>
<td>0:00:60</td>
<td>0:02:43</td>
<td>0:00:03</td>
<td>0:02:16</td>
</tr>
<tr>
<td>Prepaid BroadAccess</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0:00:61</td>
<td>0:01:30</td>
<td>0:00:00</td>
<td>0:02:17</td>
</tr>
<tr>
<td>Postpaid Blackberry</td>
<td>75</td>
<td>55</td>
<td>20</td>
<td>19</td>
<td>1</td>
<td>0:00:62</td>
<td>0:01:18</td>
<td>0:00:04</td>
<td>0:02:18</td>
</tr>
<tr>
<td>Postpaid BroadAccess</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0:00:63</td>
<td>0:00:12</td>
<td>0:00:05</td>
<td>0:02:19</td>
</tr>
<tr>
<td>Postpaid new</td>
<td>857</td>
<td>743</td>
<td>114</td>
<td>101</td>
<td>13</td>
<td>0:00:64</td>
<td>0:00:18</td>
<td>0:00:03</td>
<td>0:02:20</td>
</tr>
<tr>
<td>Prepaid</td>
<td>1348,30</td>
<td>30794</td>
<td>104036</td>
<td>103856</td>
<td>177</td>
<td>0:00:65</td>
<td>0:06:02</td>
<td>0:00:03</td>
<td>0:02:21</td>
</tr>
<tr>
<td>Prepaid Blackberry</td>
<td>3455</td>
<td>2634</td>
<td>821</td>
<td>780</td>
<td>41</td>
<td>0:00:66</td>
<td>0:01:59</td>
<td>0:00:03</td>
<td>0:04:43</td>
</tr>
<tr>
<td>SIMREG</td>
<td>19663</td>
<td>9563</td>
<td>10100</td>
<td>10019</td>
<td>80</td>
<td>0:00:67</td>
<td>0:01:42</td>
<td>0:00:02</td>
<td>0:01:66</td>
</tr>
<tr>
<td>Shell</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0:00:68</td>
<td>0:00:32</td>
<td>0:00:05</td>
<td>0:01:56</td>
</tr>
<tr>
<td>Topup</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0:00:69</td>
<td>0:00:00</td>
<td>0:00:01</td>
<td>0:01:57</td>
</tr>
<tr>
<td>Total</td>
<td>168282</td>
<td>48412</td>
<td>119870</td>
<td>119456</td>
<td>409</td>
<td>0:00:70</td>
<td>0:05:25</td>
<td>0:00:03</td>
<td>0:01:58</td>
</tr>
</tbody>
</table>