Queuing Theory Assignment - Writeup on "Optimizing Automated Call Routing by Integrating Spoken Dialog Models with Queuing Models" by Tim Paek & Eric Horvitz

Diya Gangopadhyay
Roll No: 04020507

Area of Application of Queuing Model:
The problem discussed in the paper is that of automatic transfer of a customer's call from an mechanised interaction mode to a manual operator. A predicting algorithm has been used to decide if transferring to a manual mode of interaction has higher likelihood of success in terms of accuracy and waiting time. The mechanised interaction which is a cost efective solution for handling customer calls has its limitations with respect to speech recognition accuracy and its failures often result in prolonged waiting time and poor user experience. It is therefore important to predict such failures and accordingly decide when reverting to the more expensive but reliable manual voice operator can be beneficial.

Method of Use:
The logs generated by *VoiceDialer* - the voice spoken dialogue system used in the application were analysed in order to build a predictive model. the paper lists down definitions and distributions of final **outcomes** of sessions as follows:
• SpeakFound (45%): System finds the correct name in the directory, as confirmed by a transfer.
• OperatorRequest (23%): Caller presses '0' for an operator.
• HangUp (13%): Caller hangs at some point in the session.
• MaxErrors (12%): System reaches threshold of allowed misrecognitions and routes the call to an operator.
• SpeakNotFound (6%): System concludes that the name is not in the directory and routes the call to an operator.
• Undefined (1%): Caller presses other numeric keys.
• HelpRequest (<1%): Caller requests help by pressing '*' or '#'.
• NotReady (<1%): System is temporarily out of service.

The purpose is to find the likelihood of success and failure of the voice based interaction as well the distribution of the total operation time.
Features from the spoken dialogue system in the following four categories:
1) Action sequences by the user or the interactive system such as button press, repeating a name etc
2) Whole dialouge features: Outcome, total time (as defined above)
3) ASR (Automatic Speech recognition) features : Number of hypotheses in the n-best list, range of confidence scores, mode etc
4) Pairwise ASR features: Number of recurring first/last/full names that match in consecutive n best lists.

For building the predictive model *Bayesian networks* was used to perform inference over the joint distributions needed for the decision-theoretic procedure,and also to determine what spoken dialog features would comprise the local structure of three primary variables of interest: Outcome
Failure, a binary recoding of Outcome with Speak-Found as '1' and '0' for everything else
Duration, the expected completion time of the session.

The tables record the marginal outcomes and failures of each distribution and the lifts above the marginal outcome for each case and the corresponding time distributions.

Optimization:
The principle of maximum expected utility (synonimous to minimum cost) is used for optimization which states that the action A = a that maximizes its expected utility, EU(a|ξ) should be selected.
If ξ denotes all background information and H represents all possible states of the world, then the selection of actions is guided by the following optimization:

$$\arg\max_a EU(a|\xi) = \arg\max_a \sum_h P(H=h|\xi)u(a,h) \quad (1)$$

where u(a,h) expresses the utility of taking action a when the state of the world is h.
Define: d= the action of dispatching a call
S= set of possible outcomes, for simplicity equal to the binary variable failure
O= the state of the operators (busy/not busy)

Using (1) and the above definitions the following dispatch procedure is obtained:
Dispatch a call to an operator only when the expected utility of d, given the state of the operator O and call routing system S, exceeds that of any dialog action a.

That is for all a not equal to d; $[EU(d|S,O) > EU(a|S,O)]$ (2)

Using the definition of conditional probability:
$EU(d) = \sum_s \sum_o P(S,O|d)u(d,S,O)$

$$EU(d) = \sum_s \sum_o P(O|S,d)P(S)u(d,S,O) \quad (3)$$

The only dialog action that affects O, whether or not the operator is busy, is d since a transfer increases the number of callers waiting to be serviced by the operator. The effect of all other dialog actions taken by a call routing system remain within the system.

$$P(O|S,\neg d)P(S) = P(O|\neg d)P(S) \quad (4)$$

Call Centre Queue:
Data collected from the call centre were modelled to find the best fit for a Poisson process. It was desirable to model the data as an exponential distribution due to the advantage of its *memoryless* property and ease of analysis. Two months of call centre data with >1700 calls was collected. The average rates, which represent both the maximum likelihood estimate and method of moments estimate for the exponential distribution, were 4.41 seconds (λ) between calls and 28.22 seconds (μ) to dispense a call. To make sure that the distributions were indeed exponential, a log transformation of the empirical distributions was performed and regression lines fit to estimate the correlation coefficients. The fit with a Poisson process was found to be reasonable.

The call centre queue was modelled as an M/M/z queue where z = the number of operators (servers).
The likelihood that all the operators are busy in a call center if a call is dispatched for an M/M/z queue is:

$$P(O|S,d) = P(n >= z) = 1 - \sum_{n=0}^{z} P_n \quad (5)$$

where n is the number of callers
According to (5), an operator is busy when the number of callers in the queue, including the dispatch call from the automated system, exceeds the number of operators at the call center. According to the distribution of the likely number of callers, the expected number at any given time is around 7.
The M/M/z model was checked and found to be accurate for the given scenario.

Cost Assessments:
The utility of a dialog action a given the operator state O and system state S can be approximately decomposed as follows:
$u(a,S,O) = u(a,S) + a(a,O)$ (6)
For a=d i.e. dispatch to an operator, (6) can be written as:

$c(d,S,O) = customerCost . (t+W) + operator\ cost . z$ (7)

where t is the time a caller has already spent in the automated system, and W is the predicted "dwell time," which includes both the time waiting in line and the time being served (Gross & Harris, 1998). W is derived from the M/M/z queuing model.
W can bedropped out when the state of O is "not busy".

In order to calculate ROI for the call center, the inter-arrival rate of calls entering the spoken dialog system were modelled as a Poisson process. The average arrival rate was 1.51 seconds, and a linear regression line fit to the log transformation of the empirical distribution revealed a significant correlation indicating an exponential distribution.
The optimum number of operators that would have been required to field the calls to the automated call routing system was found to be 25 as shown in the figures given in the paper. Using the inter arrival rates of the call centre customers, the optimum number of operators was found to be 11, 1 more than the number currently employed.

Evaluation:
To assess the effectiveness of the procedure, the same test data was used for evaluating the spoken dialog models to examine how many calls the procedure would have dispatched after each ASR output given the incrementally growing number of spoken dialog features. As a comparison, an alternative baseline procedure of using the marginal distribution was considered to decide whether to dispatch:
viz, if the most likely state of S, or the binary variable Failure, is "System Not Failing," then the call is to be kept in the system; otherwise, dispatched.
In using the marginal distribution, the alternative procedure represents a more intelligent way of transferring a call than simply dispatching when the system reaches a failure, the most prevalent procedure in automated call routing.
The results of the analysis on each test set comparing the decision theoretic (DT) procedure against the marginal procedure have been displayed in the table.
It can be inferred that Optimal performance benefits from the incorporation of both a model of the operator queue and the stakes involved.

Cost Savings:

Average savings = $(\sum_n |ec(d) - ec(\neg d)|)/n$

where ec = expected cost
The individual cost savings in each recognition has been analysed using the decision-theoretic process. The procedure is seen to cut cost when the probability of success or failure of the recognition is remarakably high or low.