Communications of the IIMA

Volume 10 | Issue 2 Article 4

2010

# Establishment of Confidence Thresholds for Interactive Voice Response Systems Using ROC Analysis

Oredola A. Soluade
*Iona College*

# Establishment of Confidence Thresholds for Interactive Voice Response Systems Using ROC Analysis

**Oredola A. Soluade**
**Iona College**
**osoluade@iona.edu**

## ABSTRACT

*An Interactive Voice Response (IVR) System is a platform for man-machine interaction. It is used for collecting and analyzing human voices so as to provide the desired response. The algorithm for collecting these utterances, analyzing them correctly, and providing the desired response to a caller, has been studied extensively (Allen, 1995). Whenever one calls most large organizations, their initial encounter is with a machine that will prompt the caller for their intent. Usually, such machines will give you options to choose from (Directed Dialog), or it may ask for your input (Open Dialog). This paper focuses on Open Dialog where the caller is free to indicate their intent. The problem is that the Voice Recognizer may misinterpret the caller intent; thereby providing the caller with the wrong information. This is because the recognizer has a threshold for recognizing any utterance, and traverses the part of the Call Flow that corresponds to what the engine recognizes. This threshold can be calibrated for optimal performance by undertaking a statistical analysis of a random sample of utterances, and based on the result, set the threshold that will be used to discriminate between caller utterances. The criteria that are used for establishing this threshold include, among others, Sensitivity, Accuracy and Specificity. The optimal threshold will be the one that optimizes the majority of these parameters.*

## INTRODUCTION

An Interactive Voice Response (IVR) System is a platform for man-machine interaction by the use of voice or keypad. Examples abound. Whenever one calls most large organizations, their initial encounter is with a machine that will prompt the caller for their intent. Usually, such machines will give you options to choose from (Directed Dialog), or it may ask for your input (Open Dialog). In the case of Open Dialog, there is the risk that the machine does not understand a caller input. This is an area where a lot of investigation takes place to deduce why this is the case. The technology for recognizing keyed input is not as challenging as speech technology because each key on the keypad corresponds to a specific sound frequency that cannot be confounded with another key. This technology is called Dual Tone Multi Frequency (DTMF[i]); and it is a mature technology due to the fact that there is little or no variability in the tone emitted by a particular key. This is not the case with speech. In the case of speech technology, there are several variables that come into play. These include whether a caller barges-into a prompt, whether there is a lot of background noise that may be of similar frequency as the spoken utterance, whether the user is using a cell phone, a speaker phone, or a computer. These, and several other factors, affect the way an IVR system recognizes the caller input. This paper is an attempt to establish guidelines for determining the best settings under which an IVR system should accept a caller input using ROC analysis.

## REVIEW OF LITERATURE

Receiver Operating Characteristics (ROC) analysis has been used in medical imaging to measure diagnostic accuracy (Metz, 2008; Pepe, 2000; Griner, Mayewski, Mushlin, & Greenland, 1981). To diagnose diseases, (McClish, 1989) used this technique to analyze the accuracy of the diagnosis. He preferred this technique because it provided the investigator with all possible combinations of sensitivity and specificity. ROC analysis has been used in the field of radiology (Metz & Obuchowski, 2003). ROC analysis was applied to biomedical informatics, (Lasko, Bhagwat, Zou, & Ohno-Machado, 2005; Brown & Davis, 2006; Hand, & Till, 2001), Signal Detection Theory (Green & Swets, 1966); it provides a precise language and graphic notation for analyzing decision-making in the presence of uncertainty. ROC curves are used extensively in epidemiology and medical research and are frequently mentioned in conjunction with evidence-based medicine (Zweig & Campbell, 1993). Bond and DePaulo (2006) used ROC analysis to study the accuracy of Deception judgments by studying over 20,000 judgments, and came to the conclusion that such analysis correlated strongly with other methods of analysis. In the field of Artificial Intelligence (Fogarty, Baker, & Hudson, 2005), ROC curves have proved useful for the evaluation of machine learning techniques (Flach, 2004; Fawcett, 2006). The approach used in this paper is to extend the use of ROC analysis to Speech Recognition. If an utterance is clearly understood (with high/medium confidence) the caller will be led further down the rest of the call flow. If, however, the IVR engine is not certain what the caller input is, it would be compelled to re-prompt the caller so as to confirm that the original intent was correctly identified. After the second attempt at recognition, for caller inputs that are still not clearly understood by the IVR engine, the caller will be transferred to a live agent. This is what the IVR engine is designed for - to minimize (and possibly eliminate) the cost of transferring to a live agent.

## THE ENVIRONMENT

The Interactive Voice Response (IVR) environment consists of a platform for collecting and analyzing caller utterances using a voice recognizer. The quality of the categorization varies with the parameter settings of the recognizer. The two main parameters of the recognizer are: the energy floor and the confidence threshold. The energy floor should be set so that the recognizer can pick up faint utterances. However, if this setting is too low, the recognizer will also pick up background noise. The confidence threshold is the minimum setting below which an utterance will rejected (a NoMatch). If the confidence threshold is set very high, the recognition rate will be very low because more utterances that would ordinarily be recognized by the human ear will be rejected by the recognizer. On the other hand, if the threshold is set very low, the recognizer will tend to accept unintelligible caller inputs – thereby degrading the quality of the recognizer.

Consider a situation in which a caller accesses an IVR system. The caller could be placing the call from any communication medium such as PSTN[ii] phone, wireless phone, or VOIP[iii] phones. The IVR system receives the call and prompts the caller for their intent. Assume there are six (6) possible options available to the caller:

- I want to check my account balance
- I would like to locate a store near my home
- I would like to place an order

- I would like to return an item
- I would like to speak with an agent
- I would like to know my Promotional Code

Each of these utterances has a corresponding DTMF equivalent. The DTMF equivalent is usually invoked whenever the initial caller intent is not recognized with a high enough confidence. For example, if the caller intent is "*Place an order*" and the system wrongly interpreted that to mean "*Promotional Code*", the caller will be re-prompted for their input –but this time, the recognizer may give the caller the option of either providing speech input, or inputting a DTMF tone. The caller will most probably enter a DTMF tone if the prompt category exists for her intent; otherwise, the IVR system will reprompt at least one more time before it opts out of the call flow, and transfers the caller to a live agent. DTMF tones are usually very reliable because there can hardly be any interference between the tone generated, and background noise. The same is not true of spoken utterance. Depending on the caller's location, the ability of the recognizer to decipher the caller intent will vary accordingly. The Call Flow below shows the path of the interaction between a caller and the IVR platform.

At the beginning of the Call Flow, both the intent and error counts are initialized at zero. This is necessary to be able to keep track of how well the Speech recognizer captures the caller intent. An increase in the intent and error counters provides a clear indication that the recognizer is not identifying the caller intent correctly at the first encounter. The call flow is usually designed so that by the second or third iteration, the recognizer prompts the caller with a DTMF option; and if that fails, the recognizer then 'opts out' of the call flow and transfers the caller to a live agent. This will help to ensure that the caller gets the desired service. This predefined frustration limit is set by the Software Developer at the design stage based on previous experience with callers. It is preferable that the caller is transferred to a live agent, than have the caller go through an infinite loop. It is also possible for the caller to reach their frustration limit and request an agent well before the recognizer provides the DTMF option.

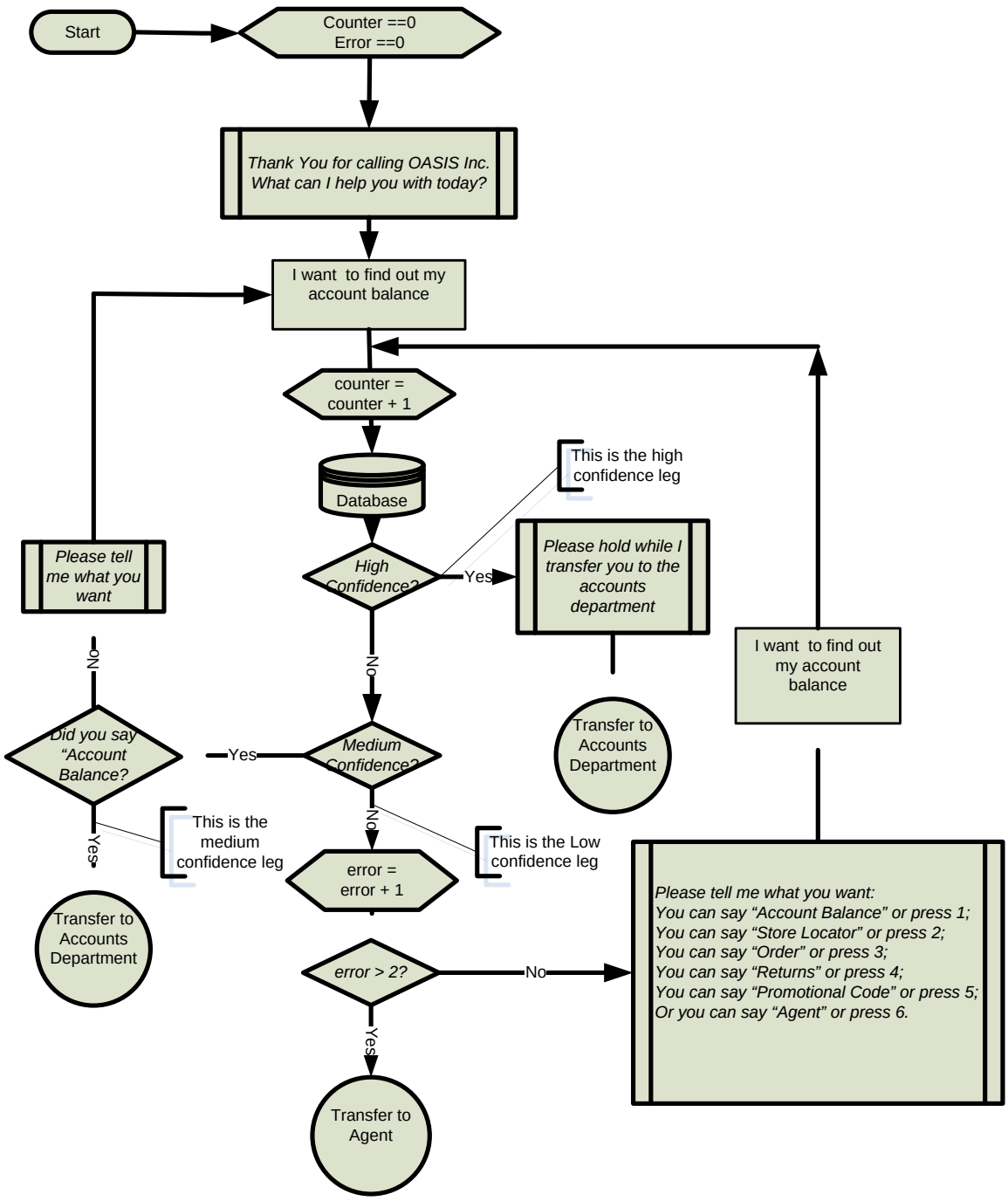



**Figure 1: Sample Call Flow.**

For each of these utterances, a grammar base is developed to accommodate different possible permutations of the caller intent - so as to avoid re-prompts. The occurrence of a re-prompt is an indication that the recognizer is not picking up the caller utterance with high enough confidence, and thus needs to re-prompt to ensure that the caller is directed to the correct destination. An utterance like "*Place an order*" will have several alternative forms that are

deemed to be synonymous caller inputs. A sample of three alternative forms of the six caller inputs analyzed in this paper is shown in the table below.

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| | **Account Balance** | **Store Location** | **Place an order** | **Returns** | **Agent** | **Promotional Code** |
| **Alternative 1** | Please give me my account balance | Store Locator Please | I would like to place an order | I would like to return an item | Please give me a live agent | I want to know my promotional code |
| **Alternative 2** | Account Balance please | I would like to find a store near me | I want to order an item | Returns please | I want to speak to a live person | What is my promotional code? |
| **Alternative 3** | What's the balance in my account | What store closest to me | Placing an order | Returning an item | Agent Please | Promotional Code Please |

**Table 1: Sample Grammar Base.**

Any of these utterances is run through a robust grammar to establish the closest approximation to the caller intent. This is where the confidence score[iv] is used for establishing the degree to which the recognizer accurately interprets the caller intent. This confidence score is based on several factors, significant among which is the energy level[v] of the volume of sound generated by the caller's utterance. If the energy level is high, the probability is high that the confidence score will also be high. On the other hand, if the energy level is low, then the recognizer will come up with a low confidence score, which may result in a *Reprompt* or a *NoMatch.* The demarcation between these three thresholds is not arbitrary, and can be established using several techniques. One such technique is known as Receiver Operating Characteristics or ROC.

## METHODOLOGY

Utterances were collected using an Automation tool - *Hammer CallMaste*r[vi] which offers an advanced user interface that allows analysts to create, schedule, and manage sophisticated voice performance tests, as well as generate Interactive Voice Response (IVR) application performance data. The utterances were collected from various sources according to the table below:

| | | Handset | Speaker phone | Cell | TOTAL |
|---|---|---|---|---|---|
| 1<br>Account Balance | I want to check my account balance | 2 | 2 | 2 | 6 |
| 2<br>Store Locator | I would like to locate a store near me | 2 | 2 | 2 | 6 |
| 3<br>Order | I would like to place an order | 2 | 2 | 2 | 6 |
| 4<br>Return | I would like to return an item | 2 | 2 | 2 | 6 |
| 5<br>Agent | I would like to speak with an agent | 2 | 2 | 2 | 6 |
| 6<br>Promotional Code | I would like to have my Promotional | 2 | 2 | 2 | 6 |
| | TOTAL | 12 | 12 | 12 | 36 |

**Table 2: Utterance-Collection Table.**

A total of 21 callers were assembled to place calls to the IVR platform. Each caller (7 male, 7 female, and 7 foreign) places a total of 36 utterances - 2 utterances for each of the telephone medium, for each utterance. So there will be 6 recorded utterances for "Account Balance"; 6 recorded utterances for "Store Locator", 6 recorded utterances for "Order", 6 recorded utterances for "Return", 6 recorded utterances for "Agent", and 6 recorded utterances for "Promotional Code" giving us a grand total of 756 recorded utterances.

## THE ROC SPACE

The contingency table for this analysis is as shown in the table below, and can be used to derive several evaluation "metrics".

| | **High Confidence** | | | **Medium/Low Confidence** | | **Total** |
|---|---|---|---|---|---|---|
| **Call** | **In-Grammar** | | | **Out-Of-Grammar** | | |
| **Positive** | True Positive | a | | False Positive | c | a + c |
| **Negative** | False Negative | b | | True Negative | d | b + d |
| **Total** | | a + b | | | c + d | |

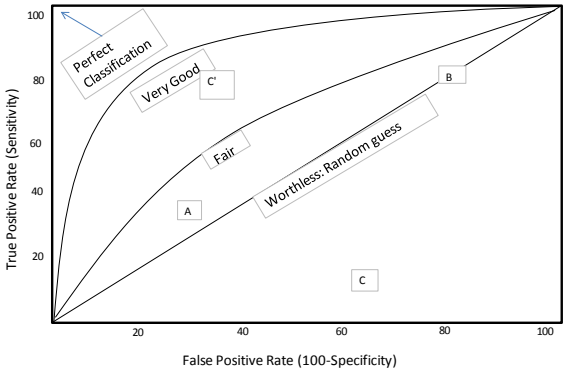
**Table 3: Schematic Outcomes of an utterance.**

To draw an ROC curve, only the true positive rate (TPR) and false positive rate (FPR) are needed. TPR determines a classifier or a diagnostic test performance on classifying positive instances correctly among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test.

An ROC space is defined by FPR and TPR as *x* and *y* axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent with sensitivity and FPR is equal to (1 - specificity), the ROC graph is sometimes called the sensitivity vs. (1 - specificity) plot. Each prediction result or one instance of a confusion matrix represents one point in the ROC space. The best possible prediction method would yield a point in the upper left corner or coordinate (0,1) of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). The (0,1) point is also called a *perfect classification*. A completely random guess would give a point along a diagonal line (the so-

called *line of no-discrimination*) from the left bottom to the top right corners. An intuitive example of random guessing is a decision by flipping coins (head or tail). The diagonal line divides the ROC space in areas of good or bad classification/diagnostic. Points above the

**Figure 2: Interpretation of the ROC Space.**




diagonal line indicate good classification results, while points below the line indicate wrong results.

Let us look into four prediction results from 100 positive and 100 negative instances:

| A | | |
|---|---|---|
| TP=63 | FP=28 | 91 |
| FN=37 | TN=72 | 109 |
| 100 | 100 | 200 |
| TPR = 0.63 | | |
| FPR = 0.28 | | |
| ACC = 0.68 | | |

| B | | |
|---|---|---|
| TP=77 | FP=77 | 154 |
| FN=23 | TN=23 | 46 |
| 100 | 100 | 200 |
| TPR = 0.77 | | |
| FPR = 0.77 | | |
| ACC = 0.50 | | |

| C | | |
|---|---|---|
| TP=24 | FP=88 | 112 |
| FN=76 | TN=12 | 88 |
| 100 | 100 | 200 |
| TPR = 0.24 | | |
| FPR = 0.88 | | |
| ACC = 0.18 | | |

| C' | | |
|---|---|---|
| TP=76 | FP=12 | 88 |
| FN=24 | TN=88 | 112 |
| 100 | 100 | 200 |
| TPR = 0.76 | | |
| FPR = 0.12 | | |
| ACC = 0.82 | | |

Plots of these four results are indicated in the ROC space in the figure. The result **A** clearly shows the best among **A**, **B**, and **C**. The result **B** lies on the random guess line (the diagonal line), and it can be seen in the table that the accuracy of **B** is 50%. However, when **C** is mirrored onto the diagonal line, as seen in **C'**, the result is even better than **A**. The relationship between C and **C'** is derived from C by simply reversing the predictions of whatever method or test produced the **C** contingency table. When the **C** method predicts **p** or **n**, the **C'** method would predict **n** or **p**, respectively. In this manner, the C' test would perform the best. While the closer a result from a contingency table is to the upper left corner the better it predicts, the distance from the random guess line in either direction is the best indicator of how much predictive power a method has, albeit, if it is below the line, all of its predictions including its more often wrong predictions must be reversed in order to utilize the method's power.[vii]

## ESTABLISHMENT OF THRESHOLDS

Two settings of the TellMe[viii] Speech Recognition engine were tested using the same 756 recorded utterances on each engine. Using the call flow in Figure 2, an application was developed where each utterance was recorded as a .wav file and played back to each of the two TellMe settings. The application is supposed to collect the caller input (speech utterance or DTMF) and direct the caller to the desired department if the caller input is correctly identified, or else reprompted or redirected if the recognition level is low. The discriminate factor as to whether an utterance is recognized or not is based on the energy level of the utterance (or in the case of DTMF, the tone). The thresholds established for categorizing an utterance into High, Medium, and Low levels is set at varying levels, and the percentage of caller inputs that are correctly recognized is captured. Initially, the breakdown is set at High: 40% and above; Medium/Low: Below 40%. Based on the Central Limit Theorem, it can be assumed that the distribution of the utterance recognition will follow a normal distribution. The results are then used to calculate four (4) parameters:

***True Positive (TP)*: The recognizer correctly identifies the caller input with high confidence**

***True Negative (TN)*: The recognizer correctly rejects an out-of-grammar utterance**

***False Positive (FP)*: The recognizer incorrectly identifies the caller input**

***False Negative (FN)*: The recognizer incorrectly rejects a correct (in-grammar) caller input**

**Assume there are N utterances. Then we have: TP + FP + TN + FN = N**

**Accuracy is defined as: TP + TN. The Total Error is defined as: FP + FN**

| | | | | | | Threshold 0.4 | | Threshold 0.3 | |
|---|---|---|---|---|---|---|---|---|---|
| (N = a+b+c+d)--> Utterance | (a) True Positive | (b) False Negative | (c) False Positive | (d) True Negative | TOTAL | % Accuracy $\frac{a+d}{N}$ | % Error $\frac{b+c}{N}$ | % Accuracy $\frac{a+d}{N}$ | % Error $\frac{b+c}{N}$ |
| 1 Account Balance | 462 | 70 | 56 | 168 | 756 | 83.33 | 16.67 | 81.48 | 18.52 |
| 2 Store Locator | 518 | 42 | 28 | 168 | 756 | 90.74 | 9.26 | 85.19 | 14.81 |
| 3 Order | 546 | 84 | 56 | 70 | 756 | 81.48 | 18.52 | 79.63 | 20.37 |
| 4 Return | 560 | 56 | 28 | 112 | 756 | 88.89 | 11.11 | 85.19 | 14.81 |
| 5 Agent | 546 | 42 | 70 | 98 | 756 | 85.19 | 14.81 | 85.19 | 14.81 |
| 6 Promotional Code | 644 | 28 | 28 | 56 | 756 | 92.59 | 7.41 | 88.89 | 11.11 |

**Table 4: Threshold Comparison of Accuracy and Percent Error.**

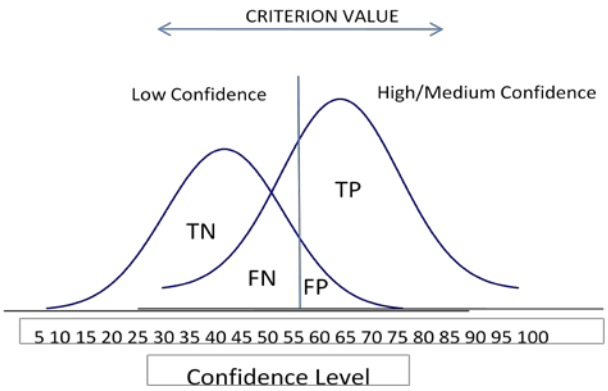



**Figure 3: ROC Curve for Analysis of Utterances.**

The quality of the recognizer is enhanced by a large grammar base. The larger the grammar base, the more efficient the recognizer, resulting in a higher probability that the recognition will occur at the first attempt - thereby reducing the number of re-prompts. This is then used as a basis for establishing a confidence score for each utterance. The confidence score is calculated based on the volume and energy level of the caller input. In most IVR Systems, the categorization of the confidence score into High, Medium, or Low level is based on an analysis of the Operating Characteristics of the Recognition software – known as Receiver Operating Characteristics (ROC). Our focus here is on how well the platform performs for the given confidence thresholds. On most platforms, the threshold for high confidence is set at between 0.3 and 0.4. For every possible threshold set for categorizing the recognition as High, Medium/Low, any of the four scenarios shown below, is bound to occur.

## ROC ANALYSIS

We want to be able to tell, without actually knowing the truth, if the recognition result is likely to be correct or not. If it is incorrect or likely to be incorrect, we want to reject it. Rejection relies on the confidence score assigned to each utterance which is then used as a criterion for accepting or rejecting a caller input. The four categories of acceptance or rejection can be classified into several measures of the recognizer performance.

| | | | |
|---|---|---|---|
| Sensitivity | $\frac{a}{a+b}$ | Specificity | $\frac{d}{c+d}$ |
| Accuracy | $\frac{a+d}{a+b+c+d}$ | False Positive Rate | $\frac{c}{b+d}$ |
| Positive Predictive Value | $\frac{a}{a+c}$ | Negative Predictive Value | $\frac{d}{b+d}$ |
| Positive Likelihood Ratio | $\frac{Sensitivity}{1-Specificity}$ | Negative Likelihood Ratio | $\frac{1-Sensitivity}{Specificity}$ |

**Table 5: Summary of Recognizer Performance Metrics.**

***SENSITIVITY (TRUE POSITIVE RATE):*** Probability that an utterance will be positively recognized with high confidence when the utterance in in-grammar. This is expressed as a percentage of all the in-grammar utterances.

***SPECIFICITY (OR TRUE NEGATIVE RATE):*** Probability that an utterance will be recognized as out-of-grammar when it is indeed out-of-grammar and is therefore not accepted by the recognizer. This is expressed as a percentage of the of all the out-of-grammar utterances.

***ACCURACY:*** This is a percentage of all the utterances that were correctly classified.

***FALSE POSITIVE RATE:*** This is equivalent to a false alarm rate.

***POSITIVE PREDICTIVE VALUE:*** The probability that the utterance is in-grammar when the recognizer accepts the caller input.

***NEGATIVE PREDICTIVE VALUE:*** The probability that the utterance is out-of-grammar when the recognizer rejects the caller input – expressed as a percentage.

***POSITIVE LIKELIHOOD RATIO:*** The ratio between the probability of positively recognizing an utterance with high confidence when an in-grammar utterance is spoken, and the probability of positively recognizing an utterance with high confidence when an out-of-grammar utterance is spoken. This is basically the *True Positive Rate/False Positive Rate.*

$$\text{NOTE: } \frac{True\ Positive\ Rate}{False\ Positive\ Rate} = \frac{Sensitivity}{(1-Specificity)}$$

***NEGATIVE LIKELIHOOD RATIO:*** The ratio between the probability of rejecting an in-grammar utterance and the probability of rejecting an out-of-grammar utterance. So we have:

$$\text{NOTE: } \frac{\textit{False Negative Rate}}{\textit{True Negative Rate}} = \frac{(1 - \textit{Sensitivity})}{\textit{Specificity}}$$

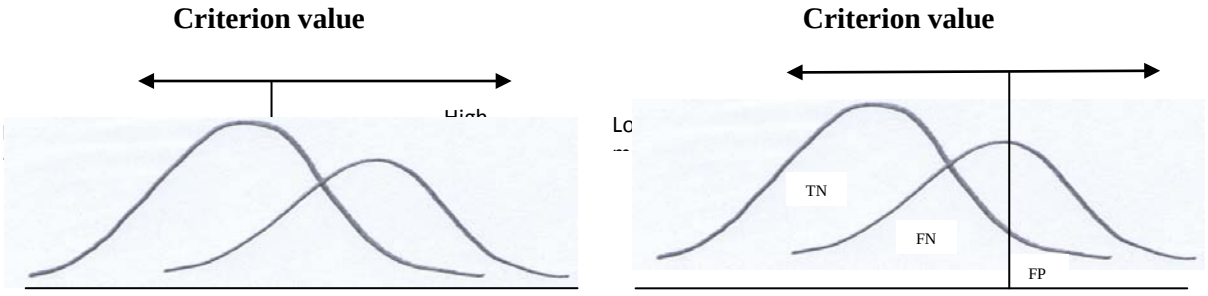



**Figure 4: High Vs Low Confidence Thresholds.**

## TEST RESULTS

There is always a trade-off between sensitivity and specificity. If the criterion value is increased, (shifted to the right), the *False Positive* fraction will decrease with increased specificity. On the other hand, the *True Positive* fraction will increase with increased specificity. As the criterion value is decreased, (shifted to the left), the True Positive fraction will increase with increased sensitivity. On the other hand, the *False Positive* fraction will also increase thereby decreasing the *True Negative Fraction* and Specificity.

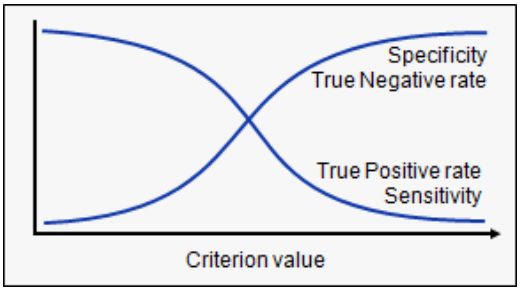



**Figure 5: Sensitivity Vs Specificity.**

The threshold set depends on the objective of the experiment. If the objective is to MAXIMIZE the percentage of utterances that can be categorized as True Positive, then, the threshold should be as low as possible. On the other hand, if the objective is to MINIMIZE the percentage of utterances that turn out to be False Positives, then the threshold should be set as high as possible.

| Utterance | (a) True Positive | (b) False Negative | (c) False Positive | (d) True Negative | TOTAL | Sensitivity $\frac{a}{a+b}$ | Specificity $\frac{c}{c+d}$ | Positive Likelihood Ratio $\frac{Sensitivit\,y}{1-Specificit\,y}$ | Negative Likelihood Ratio $\frac{1-Sensitivit\,y}{Specificit\,y}$ | Positive Predictive Value $\frac{a}{a+c}$ | Negative Predictive Value $\frac{d}{b+d}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Account Balance | 66 | 10 | 8 | 24 | 108 | 0.868 | 0.250 | 1.158 | 0.526 | 0.892 | 0.706 |
| 2 Store Locator | 74 | 6 | 4 | 24 | 108 | 0.925 | 0.143 | 1.079 | 0.525 | 0.949 | 0.800 |
| 3 Order | 78 | 12 | 8 | 10 | 108 | 0.867 | 0.444 | 1.560 | 0.300 | 0.907 | 0.455 |
| 4 Return | 80 | 8 | 4 | 16 | 108 | 0.909 | 0.200 | 1.136 | 0.455 | 0.952 | 0.667 |
| 5 Agent | 78 | 6 | 10 | 14 | 108 | 0.929 | 0.417 | 1.592 | 0.171 | 0.886 | 0.700 |
| 6 Promotional Code | 92 | 4 | 4 | 8 | 108 | 0.958 | 0.333 | 1.438 | 0.125 | 0.958 | 0.667 |
| TOTAL | 468 | 46 | 38 | 96 | 648 | 0.911 | 0.284 | 1.271 | 0.316 | 0.925 | 0.676 |

**Table 6: Test Results for Confidence Threshold of 0.4.**

| Utterance | (a) True Positive | (b) False Negative | (c) False Positive | (d) True Negative | TOTAL | Sensitivity $\frac{a}{a+b}$ | Specificity $\frac{c}{c+d}$ | Positive Likelihood Ratio $\frac{Sensitivit\,y}{1-Specificit\,y}$ | Negative Likelihood Ratio $\frac{1-Sensitivit\,y}{Specificit\,y}$ | Positive Predictive Value $\frac{a}{a+c}$ | Negative Predictive Value $\frac{d}{b+d}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 Account Balance | 54 | 14 | 6 | 34 | 108 | 0.794 | 0.150 | 0.934 | 1.373 | 0.900 | 0.708 |
| 2 Store Locator | 60 | 12 | 4 | 32 | 108 | 0.833 | 0.111 | 0.938 | 1.500 | 0.938 | 0.727 |
| 3 Order | 62 | 16 | 6 | 24 | 108 | 0.795 | 0.200 | 0.994 | 1.026 | 0.912 | 0.600 |
| 4 Return | 66 | 12 | 4 | 26 | 108 | 0.846 | 0.133 | 0.976 | 1.154 | 0.943 | 0.684 |
| 5 Agent | 56 | 10 | 6 | 36 | 108 | 0.848 | 0.143 | 0.990 | 1.061 | 0.903 | 0.783 |
| 6 Promotional Code | 74 | 10 | 2 | 22 | 108 | 0.881 | 0.083 | 0.961 | 1.429 | 0.974 | 0.688 |
| TOTAL | 372 | 74 | 28 | 174 | 648 | 0.834 | 0.139 | 0.968 | 1.197 | 0.930 | 0.702 |

**Table 7: Test Results for Confidence Threshold of 0.3.**

## CONCLUSIONS AND RECOMMENDATIONS

A comparison of these results indicates as follows:

i. ***SENSITIVITY*** *(True Positive Rate):* The overall sensitivity of the recognizer was better with the confidence level set at 0.4 (0.911) than at 0.3 (0.834). One might be inclined to conclude that a confidence setting of 0.4 will always be superior to a 0.3 confidence setting. This is not the case. Other factors have to be put into consideration, and a conclusion made, based on the aggregate of the settings of the parameters of the recognizer.

ii. ***SPECIFICITY*** *(True Negative Rate):* The Recognizer performed better at the 0.4 confidence level (0.284) than at 0.3 (0.139). This means that out-of-grammar utterances are rejected at a higher rate for the 0.4 confidence level, than for 0.3. This is another factor that needs to be combined with the *Sensitivity* in order to determine the optimal

settings of the recognizer. In situations where there is a high risk of extraneous grammars, it would be preferable to have the confidence level set at 0.4 than at 0.3.

*iii.* ***POSITIVE LIKELIHOOD RATIO***: This is superior at the 0.4 level (1.271) than at 0.3 (0.968). Since the *Positive Likelihood Ratio* is a comparison of the *True Positive Rate* to the *False Positive Rate*, it means that more in-grammar utterances will be accepted than out-of-grammar utterances. Any recognizer that has a ratio less than 1 should not be adopted because it means that more out-of-grammar utterances are being accepted by the recognizer than in-grammar utterances.

iv. ***NEGATIVE LIKELIHOOD RATIO***: This metric is lower at the 0.4 confidence level (0.316), than at the 0.3 level (1.197). Since this metric measures the ratio of *False Negatives* to *True Negatives*, one would expect that the recognizer performance will increase as the *Negative Likelihood* Ratio decreases.

v. ***POSITIVE PREDICTIVE VALUE***: The recognizer performance at the 0.3 level (0.930) was better than at the 0.4 level (0.925). This means that the probability that the recognizer will accept a caller input is higher when the confidence threshold is set at 0.3 than at 0.4. In this situation, a resolution has to be made regarding the trade-off between the result and the other conflicting results. However, since the difference in the predictive values is so small (0.005), one can ignore this result and conclude that the confidence threshold should be set at 0.4.

vi. ***NEGATIVE PREDICTIVE VALUE***: The recognizer performance at the 0.3 level (0.702) was better than at the 0.4 level (0.676). The same argument can be made for this metric as the *Positive Predictive Value.* The difference between the recognizer performance at the two confidence levels is so small (0.026), one can conclude that the confidence threshold should be set at 0.4.

| Utterance | Sensitivity | Specificity | Positive Likelihood Ratio | Negative Likelihood Ratio | Positive Predictive Value | Negative Predictive Value |
|---|---|---|---|---|---|---|
| Account Balance | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 |
| Store Locator | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 |
| Order | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 |
| Return | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 |
| Agent | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.4 |
| Promotional Code | 0.4 | 0.4 | 0.4 | 0.4 | 0.3 | 0.3 |

**Table 8: Summary of Threshold Analysis of Test Results.**

Overall, based on the analysis of these confidence thresholds, it can be concluded that setting the threshold at 0.4 is superior to setting it at 0.3. There are trade-offs for arriving at this conclusion. It can be seen from the data that for a confidence score of 0.3, the *Positive Predictive Value* and the *Negative Predictive Value* is better than at the 0.4 level. However, most of the parameters indicate that setting the confidence threshold at 0.4 is superior to setting it at the 0.3 level.

## REFERENCES

Allen, J. (1995). *Natural language understanding* (2nd ed.). San Francisco, CA; Benjamin-Cummings.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review,* 10(3), 213-234.

Brown, C. D., & Davis, H. T. (2006). Receiver Operating Characteristic curves and related Decision Measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 80, 24-38.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861–874.

Flach, P. A. (2004). *The many faces of ROC analysis in machine learning.* Retrieved from *http://www.cs.bris.ac.uk/~flach/ICML04tutorial/*

Fogarty, J., Baker, R., & Hudson, S. (2005). Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. *ACM International Conference Proceeding Series, Proceedings of Graphics Interface 2005*. Waterloo, Ontario, Canada; Canadian Human-Computer Communications Society.

Green, D. M., & Swets, J. M. (1966). *Signal detection theory and psychophysics*. New York, NY; John Wiley and Sons Inc.

Griner, P. F., Mayewski, R. J., Mushlin, A. I., & Greenland, P. (1981). Selection and interpretation of diagnostic tests and procedures. *Annals of Internal Medicine,* 94, 555-600.

Hand, D. J., & Till, R. J. (2001). A simple generalization of the area under the ROC curve to multiple class classification problems. *Machine Learning*, 45, 171-186.

Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics,* 38(5), 404-415.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making*, 9(3); 190–195. *doi:10.1177/0272989X8900900307*

Metz, C. E. (2008). ROC analysis in medical imaging: A tutorial review of the literature. *Radiological Physics and Technology* 1, 2-12.

Metz, C. E., & Obuchowski, N. A. (2003). Receiver operating characteristic curves and their use in radiology. *Radiology*, 229(1). 3-8.

Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics,* 56(2), 352-359.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots; A fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39, 561-577.

## ENDNOTES

[i] DTMF: Dual Tone Multi Frequency Tones - are two different tones at two ends of a spectrum that are used to send information in telephonic communication media.

[ii] PSTN: Public Switched Telephone Network

[iii] VOIP: Voice Over Internet Protocol

[iv] Confidence Score is a measure of the probability that the recognizer finds the caller input in its database

[v] Energy Level is a measure of the sound associated with an utterance measured in decibels (dB). It is usually set at a level different from background noise, so as to filter out the effect of extraneous sounds.

[vi] Hammer CallMaster is an Automation software that is used to analyze calls, and generate reports on IVR Performance. Empirix; 20 Crosby Drive Bedford, MA 01730, United States.

[vii] Excerpted from the web. http://en.wikipedia.org/wiki/Receiver_operating_characteristic

[viii] TellMe Studio is a VXML platform that is commercially available for analysis of IVR systems.