Communications of the IIMA

Volume 9 | Issue 3 Article 2

2009

A Comparative Analysis of Speech Recognition Platforms

Ore A. Soluade
Iona College

Recommended Citation
Soluade, Ore A. (2009) "A Comparative Analysis of Speech Recognition Platforms," *Communications of the IIMA*: Vol. 9: Iss. 3, Article 2.
DOI: https://doi.org/10.58729/1941-6687.1108
Available at: https://scholarworks.lib.csusb.edu/ciima/vol9/iss3/2

A Comparative Analysis of Speech Recognition Platforms

Ore A. Soluade
Iona College
USA
osoluade@iona.edu

ABSTRACT

Speech recognition (also known as automatic speech recognition) converts spoken words to text. It is a broad term which means it can recognize almost any speech – such as in a call centre system designed to recognize many voices. Speech Recognition in the field of telephony commonplace; and in the field of computer gaming and simulation, is becoming widespread. People with disabilities are another part of the population that benefit from using speech recognition programs. It is becoming increasingly certain, that the interaction between humans and speech recognition engines is on the increase. In certain circumstances, the caller is directed with a series of options. This is called a Directed Dialog interaction. On the other hand, there are situations where the caller is not limited by pre-defined options; but rather given the opportunity to indicate their intent. This scenario is known as an Open Dialog interaction where the caller indicates their intent orally, and the speech platform is expected to correctly interpret the caller's intent. Such interpretations are prone to variation in recognition and classification. Even if the application software correctly classifies the caller intent, it may not adequately capture the actual utterance. This paper proposes statistical techniques for measuring the performance of three Speech Recognition engines in a directed-dialog scenario.

INTRODUCTION

Speech recognition (also known as automatic speech recognition) converts spoken words to text (Jurafsky & Martin, 2000). It is a broad term which means it can recognize almost anybody's speech – such as in a call centre system designed to recognize many voices. The performance of speech recognition systems is usually specified in terms of accuracy and speed (Allen, 1995). Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER), whereas speed is measured with the real time factor. Substantial efforts have been devoted in the last decade to the test and evaluation of speech recognition in fighter aircraft, or the training for military (or civilian) air traffic controllers (ATC). Speech Recognition in the field of telephony is now commonplace and in the field of computer gaming and simulation is becoming more widespread (Flach, 2004). People with disabilities are another part of the population that benefit from using speech recognition programs.

In telecommunications, **Interactive Voice Response (IVR)** systems allow customers to access a company's database via a telephone touchtone keypad or by speech recognition. They can respond with pre-recorded or dynamically generated audio to further direct users on how to proceed. Often they are used to control functions where the interface can be broken down into a series of simple menu choices. In telecommunications applications, such as customer support lines, IVR systems generally scale well to handle large call volumes. However, the use of such systems is significantly impacted by several extraneous factors; among which are noise, accent,

casual speech styles and medium of speaking (Handset, headset, speakerphone, and cell phone). However, there have been significant advances in recent years. Automatic speech recognition capabilities now permit us to use speech as an interface for dictation and for information access. In such applications, the interactivity of the system should be such that the user experience is as good as a human experience; otherwise users will drift away from its use. IVR technology is also being introduced into automobile systems for hands-free operation. Current deployment in automobiles revolves around satellite navigation, audio and mobile phone systems. This paper is an attempt to compare three different platforms in terms of their ability to adequately recognize the utterances made by callers into an airline agency, and classify them correctly. Essentially, we are trying to determine how well the utterances collected, are recognized, and properly classified into their stipulated categories.

Consider an airline agency in which there are six (6) possible options available to the caller as shown in Table 1 below.

Table 1: Table of caller options.

1:Reservations	To make reservations for a flight
2:Flight Status	To obtain information about a flight
3:Reconfirmation	To reconfirm a flight
4:Agent	To select a seat on a flight with existing reservation
5:Seat Selection	To speak to an agent
6: Other	Any other utterance

For each of these options, a grammar base is developed to accommodate different possible permutations for a caller to indicate their intent - so as to avoid re-prompts. An utterance like "Make a Reservation" will have several alternative forms that are deemed to be synonymous caller inputs as shown in the Table 2 below:

Table 2: Confidence thresholds for utterance classification.

Platform	High Confidence Threshold	Medium Confidence Threshold	Low Confidence Threshold
Platform A	0.4 – 1.00	0.3 – 0.4	0 – 0.3
Platform B	0.3 – 1.00	0.25 – 0.3	0 – 0.25
Platform C	0.4 – 1.00	0.25 – 0.3	0 – 0.25

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) (Metz, 1978). 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.

METHODOLOGY

In order to have a wide variety of voices, 50 persons were digitally recorded using the six designated utterances. Due to background noise some of the utterances could not be used. The recorded utterances were then played back against an application that uses the airline automation system to collect customer intents.

This is essentially an analysis of the distribution of a Multinomial Population. In this instance, the six (6) categories of utterances that a caller can make are:

1. Make Reservations,
2. Flight Information,
3. Reconfirmation,
4. Seat Selection,
5. Agent,
6. Other (Any out-of-grammar utterance different from any of the 5 above).

Three different application software are considered for comparison purposes. Each application software has unique attributes that the other application software does not have. Application software A has its confidence threshold set at 0.4; Application software B has its confidence threshold set at 0.3; while Application software C (which is a variation of Application software B) has its confidence threshold set at 0.3. For each utterance played on any application software, the degree to which the software recognizes the utterance is classified as shown below.

Table 3: Sample grammar base for Make Reservations.

Utterance	**Equivalent Grammar**
Make Reservations	Make a Reservation Reservation Reservations Give me reservation I want reservations Reservations please Ah I want reservations Ah ah give me reservations

Utterances that are recognized without a re-prompt are said to have been recognized with high confidence, and so the scores associated with such recognition will be **HIGH**. If on the other hand, the application software re-prompts the caller for the utterance, it is classified as **MEDIUM** confidence – hence the re-prompt. An utterance where the application software does not seem to recognize the caller input is classified as NoInput/NoMatch, and so will have a **LOW** Confidence.

RESULTS AND ANALYSIS

For purposes of this analysis, the classification is broken down as follows:

- High/Medium Confidence
- Low Confidence

The following categories of utterance recognition were captured:

Table 4: Categories of utterance recognition.

+low	Recognition matched expected result but with low confidence
+medium	Recognition matched expected result with medium confidence
+high	Recognition matched expected result with high confidence
-low	Recognition did not match expected result but had low confidence (so was a NoMatch condition)
-medium	Recognition did not match expected result and had medium confidence (re-prompt)
-high	-high: Recognition did not match expected result and had high confidence (false accept)
+noinput	Recognition returned noinput -- expected result
-noinput	Recognition returned noinput -- unexpected result

The results were analyzed using two statistical techniques (Anderson, Sweeney, & Williams, 2009):

1. **Chi square analysis**
   This will provide a summary statistic to determine the extent to which the *Actual Distribution* of the utterances conforms to the *Expected Distribution*.
2. **Analysis of Variance (ANOVA) using Randomized Block Design.**
   This is used to determine if there is significant variation between runs within the same environment as well as variation between environments. The experiment was designed by running the utterances in each of the three environments three times. The data collected is then analyzed using ANOVA.

Development of the Chi Square Model

Consider a multinomial population where each category is distinctly identified (Reservations, Flight Information, Seat Selection, Reconfirmation, Agent, Other). A total of 284 recordings were collected from 50 callers, each uttering the six categories of caller intent. These utterances were then transcribed so that they can be verified by listening to each utterance. This provides a basis for establishing what the Platform is expected to recognize. Based on this transcription data we obtain the distribution of the utterances as proportions (probabilities).

Let p_1 = proportion of utterances that say *Make Reservations*
p_2 = proportion of utterances that say *Flight Information*
p_3 = proportion of utterances that say *Reconfirmation*
p_4 = proportion of utterances that say *Seat Selection*
p_5 = proportion of utterances that say *Agent*
p_6 = proportion of utterances that say *Other*

The Null Hypothesis (What we know to be true) is given as:

H_0: $p_1 \leq 0.27$
$p_2 \leq 0.17$
$p_3 \leq 0.17$
$p_4 \leq 0.15$
$p_5 \leq 0.12$
$p_6 \leq 0.12$

The Alternative hypothesis is given to be:

H_1: $p_1 > 0.27$
$p_2 > 0.17$
$p_3 > 0.17$
$p_4 > 0.15$
$p_5 > 0.12$
$p_6 > 0.12$

The decision to reject or not reject the null hypothesis is a function of the significance level α, which corresponds to the point in the distribution where we conclude that the actual results are significantly different from the expected results based on the χ^2 analysis.

Thus:

If χ^2 (calculated) > χ^2□□ reject H_0
If χ^2 (calculated) ≤ χ^{22}□□ Do not reject H_0

DISCUSSION

If the sample results lead to the rejection of the null hypothesis (large value of chi-square), then we can conclude that the distribution of the utterances does not conform to the expected distribution.

Table 5: Calculating chi-square statistic on First Platform.

	Expected Distribution		Actual Distribution	
Category	**Frequency**	**Probability**	**Frequency**	**Probability**
1:Reservations	58	0.27	52	0.28
2:Flight Status	36	0.17	32	0.17
3:Reconfirmation	37	0.17	35	0.19
4:Agent	34	0.15	29	0.15
5:Seat Selection	26	0.12	20	0.11
6: Other	26	0.12	21	0.11
Total	217	1.00	189	1.00

The distribution of the ACTUAL utterances on the first platform is shown in Table 3 above. Chi-square Value on First Platform: = 4.25

The null hypothesis states that the population distribution is defined by the probability values associated with each utterances as shown in the above table.

At the $\alpha = 0.05$ level of significance, we will reject the null hypothesis if the difference between the observed and expected frequencies is *large*, i.e. if the calculated χ^2 is greater than the critical χ^2 obtained from the table. Checking the chi-square distribution table, we find that with *k-1* = 5 degrees of freedom, the critical $\chi^2 = 11.07$.

Since the calculated χ^2, 4.25, is less than the critical χ^2 obtained from the table, 11.07, we DO NOT reject the null hypothesis about the distribution of the utterance classification for the First application software. In other words, the First application software utterance classification is in agreement with the null hypothesis.

Table 6: Calculating chi-square statistic on second platform.

	Expected Distribution		Actual Distribution	
Category	**Frequency**	**Probability**	**Frequency**	**Probability**
1:Reservations	60	0.28	45	0.27
2:Flight Status	36	0.16	29	0.18
3:Reconfirmation	37	0.17	31	0.19
4:Agent	33	0.15	28	0.17
5:Seat Selection	25	0.12	16	0.10
6: Other	26	0.12	15	0.09
Total	217	1.00	164	1.00

The distribution of the ACTUAL utterances on the second platform is shown table 4 above.
Chi-square Value on Second Platform: = 14.74

The null hypothesis states that the population distribution is defined by the probability values associated with each utterances as shown in the above table.

At the $\alpha = 0.05$ level of significance, we will reject the null hypothesis if the difference between the observed and expected frequencies is *large*, i.e. if the calculated χ^2 is greater than the critical χ^2 obtained from the table. Checking the chi-square distribution table in the appendix, we find that with *k-1* = 5 degrees of freedom, $\chi^2 = 11.07$.

Then at the $\alpha = 0.05$ level of significance, we will reject the null hypothesis if the difference between the observed and expected frequencies is *large.* Checking the chi-square distribution table in the appendix, we find that with *k-1* = 5 degrees of freedom, the critical $\chi^2 = 11.07$. Since the calculated χ^2, 14.74, is greater than the critical χ^2 obtained from the table, 11.07, we REJECT the null hypothesis about the distribution of the utterance classification on the Second application software. In other words, the Second application software utterance classification does not conform to the null hypothesis.

Table 7: Calculating chi-square statistic on Third Platform.

	Expected Distribution		**Actual Distribution**	
Category	**Count**	**Probability**	**Frequency**	**Probability**
1:Reservations	58	0.27	37	0.24
2:Flight Status	36	0.17	25	0.16
3:Reconfirmation	37	0.17	31	0.20
4:Agent	34	0.15	26	0.17
5:Seat Selection	26	0.12	17	0.11
6: Other	26	0.12	19	0.12
Total	217	1.00	155	1.00

The distribution of the ACTUAL utterances on the third platform is shown in table 5 above.
Chi-square Value on Third Application software: = 18.82

The null hypothesis states that the population distribution is defined by the probability values associated with each utterances as shown in the above table.

At the $\alpha = 0.05$ level of significance, we will reject the null hypothesis if the difference between the observed and expected frequencies is *large*, i.e. if the calculated χ^2 is greater than the critical χ^2 obtained from the table. Checking the chi-square distribution table in the appendix, we find that with *k-1* = 5 degrees of freedom, the critical $\chi^2 = 11.07$.

Since the calculated χ^2, 18.82, is greater than the critical χ^2 obtained from the table, 11.07, we REJECT the null hypothesis about the distribution of the utterance classification on the Third application software. In other words, the Third application software utterance classification does not conform to the null hypothesis.

The general steps for conducting a goodness of fit test for any hypothesized multinomial population is outlined as follows:

1. Formulate a null hypothesis indicating a hypothesized multinomial distribution for the population.
2. Use a simple random sample of n items and record the observed frequencies for each of the k classes or categories.
3. Based upon the assumption that the null hypothesis is true, determine the probability or proportion associated with each of the classes.
4. Determine the expected class frequencies.
5. Use the observed and expected frequencies to compute the χ^2 value for the test.
6. Complete the test by using the following Decision Strategy:

   If χ^2 (calculated) > $\chi^2_{\alpha,n-1}$ reject H_0

   If χ^2 (calculated) $\leq$ $\chi^2_{\alpha,n-1}$ Do not reject H_0

   Where α is the level of significance for the test.

In our case, we have:

Sample size:	284
Degrees of Freedom:	5
Significance Level (α):	0.05
Critical $\chi^2_{0.05,5}$	11.07 (from tables)

Table 8: Summary of Chi Square Analysis.

Platform	Chi-square value	Decision	Conclusion
First Platform	4.25	Conforms to the hypothesis	Do Not Reject H_0
Second Platform	14.74	Does not conform to the hypothesis	Reject H_0
Third Platform	18.82	Does not conform to the hypothesis	Reject H_0

CONCLUSION

In general, the χ^2 value can be used as a criterion for determining whether or not the Speech Application performance is within acceptable limits. Once the critical χ^2 value is established, the calculated χ^2 score for any test will be compared with the critical χ^2 value; and a "GO"/"NO-GO" decision is made based on the Strategy. Based on the Chi square analysis, it can be seen that the first application Software is the only one that results in the Null Hypothesis not being rejected. The second and third software had chi square values significantly higher than the critical value at the 5% significance level. It can therefore be concluded that the First Application

software is superior to the second and third software.

Development of the ANOVA Model

Repeated recordings were made for each utterance on each of the three platforms and the average number of correct recognitions is computed. Instead of comparing the proportion of utterances that fall within a particular category, the average confidence score is calculated for each type of utterance, and subjected to ANOVA (See Table 9 below).

Table 9: Average Confidence Score per platform per category.

	Average of Errors		
Observation	**First Platform**	**Second Platform**	**Third Platform**
1:Reservations	2.67	14.67	16.11
2:Flight Status	3.33	21.33	19.56
3:Reconfirmation	3.00	12.00	12.67
4:Agent	2.67	21.00	22.89
5:Seat Selection	4.00	25.00	27.33
6: Other	6.00	18.67	16.56

The Null Hypothesis is given by:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

(There is no difference in the mean of the errors between the three Platforms)

H_1: *The platform means are not equal*

The setup of the **AN**alysis **O**f **VA**riance table based on **Completely Randomized Design** is shown in the table below:

Table 10: ANOVA Table.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Between	945.6	2	472.8	**27.27**
Error	260.1	15	17.34	
Total	1206	17		

Summary of ANOVA Results

From the above ANOVA table, the calculated F value is 27.27 with degrees of freedoms 2 and 15.

From the F-table, the critical F statistic is 3.68.

Since the calculated F value is greater than the critical F statistic, we conclude that we have enough reason to reject the Null Hypothesis that the Platform means are equal.

Interpretation of ANOVA Results

Based on the ANOVA calculations there does seem to be a significant difference (at the 5% level), between the recognition in the First, Second and Third platforms for the in-grammar utterances. This means that the variance of the differences between the actual utterances and the Application recognition is not acceptable. This confirms the conclusion obtained based on the chi square analysis which indicates that the overall recognition in both Second and Third environments exceeded the critical chi-square value.

Areas for Further Research

This exercise was performed under certain constraints:

1. Only in-grammar utterances were considered. This does not have to be the case. Both in-grammar and out-of-grammar utterances could have been used for this research. It will be interesting to see the impact of out-of-grammar utterances on the operating characteristics of the application.
2. The tests were conducted with BARGE_IN turned ON. This can also be expanded to include BOTH barge-in ON and barge-in OFF scenarios.
3. The medium used for this exercise is limited to one medium – speakerphone. This can be extended to other mediums – cell phones, headsets, etc.
4. A more exhaustive study using Receiver Operating Characteristics (ROC) analysis can be performed to determine the optimal setting of the confidence thresholds.

REFERENCES

Allen, J. (1995). *Natural Language Understanding;* (2nd edition): Benjamin-Cummings Publishing Company, Inc.

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2009). *Statistics for Business and Economics* (10th edition); West Publishing Company.

Flach, P. A. (2004). Tutorial on The Many Faces of ROC Analysis in Machine Learning.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing* (2nd edition), Prentice-Hall.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, 283-298.