Probing the relationship between qualitative and quantitative performance measures for voice-enabled telecommunication services

Shrikanth Narayanan, Mani Subramaniam, Benjamin Stern, Barbara Hollister, Chih-mei Lin

AT&T Labs: Customer Labs/Research
180 Park Avenue, Florham Park, NJ 07932, USA

ABSTRACT

The relationship between objective speech recognition performance measures and perceived performance is analyzed and modeled using data obtained from a voice-dialing service trial with 798 participants. The ability of these models for predicting user perception and overall demand for such voice-enabled services is discussed.

1. INTRODUCTION

The value of automatic speech recognition (ASR) as an interface technology for telephony services clearly depends on how well the ASR works. There are methods for measuring speech recognition performance that are fairly common in the ASR community, but the question of how good is good enough is still a murky one. Of course a perfect speech recognition system had better be good enough, and many a poor one has been built that no one would use. But the state of the technology today lies in between these extremes, with ASR beginning to reach performance levels where it is acceptable in some service situations, but not in many others.

In order to direct improvements in the ASR technology, and to define success criteria of service trials, it would be therefore valuable to understand:

- the relationship between objective measures and perceived ASR performance;
- the ability of various objective measures at predicting perception;
- the relationship between satisfaction with ASR (or any interface technology) and overall demand for the service.

Recent improvements in spoken dialog systems have led to the development of several voice-enabled service prototypes. Systematic performance evaluations constitute an integral step in transforming these prototypes to real-world services [1,2]. These performance evaluations, however, typically tend to be iterative in nature [1] and require vast amounts of data to justify statistically valid results. In addition, it is our belief that significant progress in this field can only be made through collective insights gained from several such studies.

In this paper, we present some results on these questions from a consumer trial of a voice dialing service. We begin by describing the service and trial, and the trial data collection. This is followed with definitions of several different versions of objective ASR performance measures, followed by analysis relating these measures with subjective data. Finally, we provide a summary and conclusions.

2. THE SERVICE TRIAL

The service provides a single set of user-configurable names that can be used for voice dialing. In addition, the service provided voice control for system command and control with DTMF defaults. A consumer trial was conducted with 798 users to evaluate the service. The trial was conducted in three phases (1A, 1A', 1B): While all phases allowed users to enroll names in their personal dialing list by voice, only the third phase (1B) allowed the users to configure their dialing lists by text entry.

The trial evaluation involved assessment of data from multiple, typically disparate, sources such as usage session and call logs, user speech data, pre- and post-trial surveys and user demographics. In addition to evaluating the performance of the underlying speech technology, much effort went into correlating the different forms of data: for example, relating (objective) speech recognition performance with (subjective) user responses. Data collection and organization is hence a crucial pre-requisite of such performance analysis task. In this paper, we will present results from the analysis of speech data and the post-trial survey.

In order to obtain enough statistics to report per-household ASR accuracy, only those households with a daily average of ½ or more phone calls, over a period of approximately two months, are included in the analyses that follow.

2.1 System and Data

The system supported a system-initiative, small vocabulary voice-enabled application for voice dialing and menu navigation. The underlying ASR technology used context independent phone models for telephone speech and constrained grammars defining various system command features (seven in all; only the confirmation grammar used whole word models for *yes & no*). Users could configure up to 50 names in their personal dialing list by voice (all users) or by text (third phase users). In addition to the user specified entries, the "voice-label"

grammars contain a common set of system-specified commands for navigation and control.

The data set includes:

1. **User profile and demographics.**
2. **Usage Data.** These include call details, counts of feature usage, ASR results, voice label creation activity, etc.
3. **Post-trial interview data**. After several months of usage, users were surveyed on perceived quality and performance, user interface issues, likes/dislikes, purchase intent ratings, pricing alternatives, retention, etc.
4. **Speech Data.** All of the speech data from trial participant interactions with the system were recorded, along with other pertinent information such as the time of the call, a unique call identifier, the speech recognition system's result, and the grammar active at that point in the call flow.

For the purposes of this paper, all of the utterances are from points in the service call flow where the "Voice Label Grammar" was active i.e., the user-configurable grammars. This is the dominant set of data, where the main functionality of the service, voice dialing, is effected.

To judge how well the ASR system performed, it is essential to know what the user actually spoke and therefore the speech data had to be manually transcribed (called "labeling"). The user-configured grammars changed with time (i.e., voice labels could be added, deleted, and changed), and so the transcription process was facilitated by a dynamically changing user-configured vocabulary, duplicating the service evolution. In addition, these "labelers" also characterized the speaker, speech, and background attributes, using a common set of conventions and rules.

A great deal of effort went into pre-processing all of the various forms of data in order to reformat, join, check, correct and otherwise prepare the data. The data were finally loaded into MS Access™ databases for analysis.

With the help of the labeled data, it is possible to classify the speech into various categories, as illustrated in Figure 1. This enables us to calculate various speech recognition performance measures.

As can be seen in the Figure 1, a little over half of the utterances had in-vocabulary speech, and almost all of these were voice labels – 51.7% of all utterances were labels with no extra speech. Most of the remainder (almost one-third of the utterances) had no obvious foreground speech, while 11.5% of all utterances were "out-of-vocabulary"; that is, all of the spoken words were not part of the valid vocabulary for the voice label grammar.

3. ANALYSIS RESULTS

3.1 Observed Speech Accuracy

Conventional accuracy measures typically take a "technology-centric" view – they report how well a recognizer works given various categories of speech input (in-vocabulary, out-of-vocabulary, noisy, etc.). In addition to these, we defined two other objective measures that attempt to take a broader system view of the ASR performance. All of these measures were used to calculate average per-household accuracy, for each of the moderate to heavy calling households.

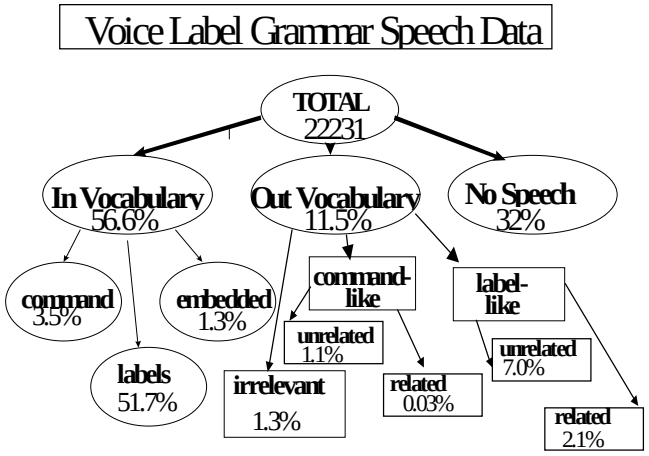

Figure. 1: Classification of speech data categories.

Following are definitions of the performance measures:

1. Voice Label Accuracy: = *correct invoc. labels/ invoc. labels*

This is a traditional measure of ASR performance, the number of in-vocabulary utterances that are correctly recognized. It focuses on labels since voice dialing was the primary "service".

2. Handled Correctly: = {*(correct invoc. Labels & commands) + (rejected out.of voc &silence)}/all utterances*

This combines the two traditional components of ASR performance measures into a single one, combining in-voc. accuracy and out-of-voc. rejection. Thus if an in-vocabulary utterance is recognized correctly, or an out-of-vocabulary or silence is rejected, then the utterance is "handled correctly".

3. User Interaction Success: = *(correct invoc.labels & commands)/(all utterances with foreground speech).*

This measure reports how often the user says the "right" thing *and* the machine recognizes it correctly, and so in effect takes into account both user errors and machine errors. ("Silence" files are not included here, as it is unclear whether this should be considered correct input.)

In addition, we defined two error measures:

1. User Error: (*misrecognized out.of.voc labels and commands)/all utterances*

The blame is assigned on the user when the system fails to reject out of vocabulary utterances *(*The rationale here is that only a small fraction of the out of vocabulary utterances are can be considered entirely irrelevant (Figure 1)).

2. System Error: (*misrecognized & rejected invoc.voc labels and commands)/all utterances*

Here the blame assignment is on the system.

3.2 Perceived Speech Accuracy

There were several questions on the post-trial survey that asked respondents to rate how well the service recognized one's speech in various situations, on a 0 to 10 scale (0 meaning "very poorly", and 10 meaning "very well"). For example, respondents were asked, "Overall, how did the service recognize your speech?" Table 1 shows the percent of survey respondents that chose one of the top three boxes in response to these perceived performance questions. In the discussion that follows, we use the "overall" result as the measure of perceived speech recognition performance.

Question: How would they rate the service on recognizing their speech for:	**Top 3 Box (All) (N=132)**	**Top 3 Box (Takers) (N=37)**	**Top 3 Box (Non-Takers) (N=95)**
Labels they spoke at "Call Where?"	30%	49%	23%
When calling from home	31%	47%	26%
Commands	53%	75%	45%
Overall	44%	73%	33%

Table 1: User Perception of Speech Accuracy.

Another survey question asked how likely is the respondent to purchase the service if it were offered as a product (on a 0 to 10 scale). For the purpose of this paper, the responses to this question are divided into two categories – "takers", who chose one of the top three boxes, and "non-takers".
Table 1, columns 2 and 3, show the percent of users rating performance in the top 3 boxes separately for takers and non-takers. It can be seen that the responses were significantly higher for takers, especially for the "overall" question. The box plots in Figure 2 emphasize this by showing the spread in accuracy for the two groups.

To better understand the relationship between the different observed accuracy measures and perception, models were fitted using each accuracy measure as the explanatory variable and perception as the response variable. These models are classical regression models with response variable being the 0 to 10 scale satisfaction scores for perception. A scatter plot between perception scores revealed several outliers in the data. These data points corresponded to households that experienced an accuracy of 89% or more yet rated overall ASR performance between 0 and 4 (lower end of the scale). These accounted for about 10% of the households in each of the trial phases. The models fit well in the absence of these points.

These outliers could be due to noise in the data attributable to measurement error in the data collection phase. But it is not unlikely that these consumers either did not like the service at all, and this colored their perception of ASR performance, or they are "hard to please" and would not be satisfied with less than perfect performance. Only a few (2) data points exhibited the opposite trend i.e., high perception scores corresponding to very low observed speech accuracy.

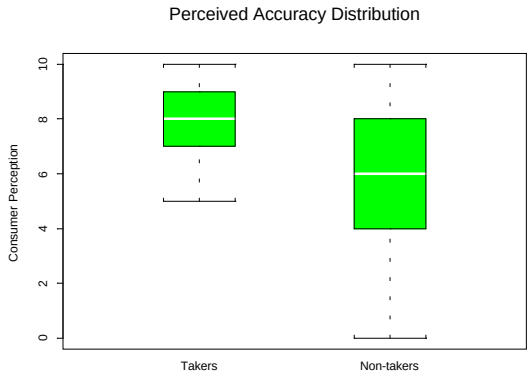

Figure 2: Distribution of Perceived Speech Accuracy for Takers & Non-Takers

The intercept term β_0 in all the 3 speech accuracy models: *mean perception* ~ (β_0+β_1*accuracy) were not significant. Thus the mean perception score can be predicted by multiplying observed speech accuracy with a parameter β_1. Table 2 gives β_1 values for the three different observed speech accuracy measurements.

The results based on error measures were, however, different: only system error correlated with user perception (β_0=7.6,β_1= -0.259). As expected, user perception scores were higher when the observed system errors were lower. However, there was no statistical relationship between user error and user perception. One possible explanation for this result is that the user interface design (error control/clarification prompts) is such that the blames were attributed to the system rather that the users making it difficult to distinguish their "mistakes", or user errors, from system errors.

Observed Accuracy	β_1
Voice Label Accuracy	0.079
Handled Correct	0.082
User-Interaction Success	0.090

Table 2: Coefficients Of the Models for the users

Figure 3 shows the 95% confidence interval for the regression fit for voice label accuracy in Phase 1B. These models help us to predict the overall consumer perception scores using these speech accuracy measurements. Table 3 shows the $\mathbf{R^2}$-coefficient (coefficient of determination, a statistical measure for degree of linear association) for each model. This measure lies between 0 and 1. $\mathbf{R^2}$ is 1 for a perfect fit and 0 when there is no relationship between variables. All three measures, *user*

interaction success, *handled correct* and *voice label accuracy*, do an equally good job of predicting the mean consumer perception ASR performance. In predicting purchase intent, again all of the speech accuracy variables were significant, but not as effectively as for the perceived performance. As one would expect, other qualitative variables like "How much do you value the convenience of being able to place a call by saying a label instead of dialing a number?" and demographic variables contribute to predicting take rate.

For further analysis, we created 0-1 indicator variable for the responses to the perceived performance question; 1 if consumers gave an 8, 9 or 10; 0 otherwise. Therefore we have two groups of consumers; one with high and the other with moderate to low perception scores. A logit model was fitted to predict the fraction of consumers with high perception scores as a function of the mean value of per-household accuracy. Figure 4 shows the estimated models for the trial.

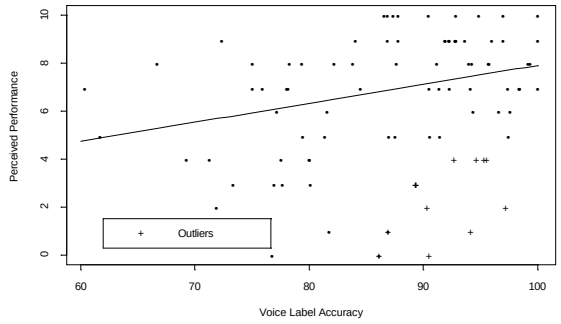

Figure 3: Household Accuracy Vs. Ratings, with Regression Line

Objective Measure	R^2	
	Perceived Performance	Take Rate
Voice Label Accuracy	0.900	0.698
Handled Correct	0.893	0.701
User-Interaction Success	0.897	0.691

Table 3: R^2 For The Models

The model is of the form $\frac{e^{\beta_0 + \beta_1 * accuracy}}{1 + e^{\beta_0 + \beta_1 * accuracy}}$ which seems to be consistent for all trials. There is a 12% increase in Phase 1A and Phase 1A' and about 16% increase in Phase 1B when accuracy goes from 80% to 90%. It is interesting to note that the accuracy was considerably higher in Phase 1B than 1A or 1A'. This was expected, as Phase 1B predominantly used "text-based" labels, rather than "voice-based" labels for configuring the user's personal list.

4. SUMMARY AND CONCLUSIONS

Three ASR accuracy measures were defined, one being simply the fraction of in-vocabulary voice labels that are recognized correctly, while the other two attempt to include most or all of utterances in a single measure of how well the service performed. All three measures do a fair job of predicting mean performance rating by a user, and there is no significant difference between them. A major limitation, however, is that in order to model this relationship, we ignore users with low/negligible usage. Of course, may factors can contribute to perceived performance, including user interface design, interest in the service, the feature set, etc.

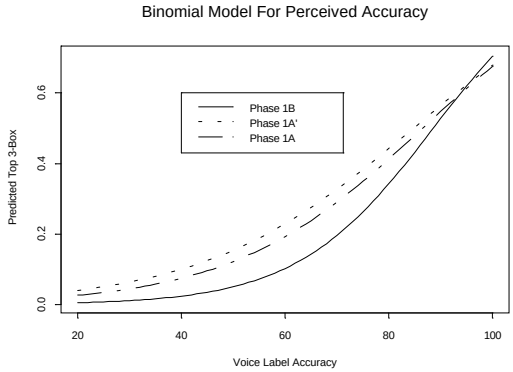

Figure 4: Binomial Model Relating Perceived Accuracy and Voice Label Accuracy

Further, all three accuracy measures have a more limited but significant predictive power of take rate for the service; as expected, other factors also contribute strongly to take rates.

Two error measures were defined, one that assigns blame on the user while the other, on the system. System error correlated with user perception while user error did not, perhaps due to the fact that the user interface did not distinguish between user errors and system errors. Perceived performance is also directly correlated with take rate. A model was developed to predict the fraction of users that would give a high rating of ASR performance as a function of the mean per-household accuracy. This model was found to be fairly generalizable across the various phases of the trial.

Acknowledgments: The authors are grateful to Dr. Sunil Dhar (Statistics Dept., NJIT) for his assistance in refining the models.

5. REFERENCES

[1] Kamm C., Narayanan S., Dutton D., and Ritenour R. *Evaluating spoken dialog systems for telecommunication services*. Proc. Eurospeech 1997 (Rhodes, Greece), pp. 2203-2206, Sept. 1997.

[2] Leonardi F., Micca G., Militello S., and Nigra M. *Preliminary results of a multilingual interactive voice activated telephone service for People on the move*. Proc. Eurospeech 1997 (Rhodes, Greece), pp. 1771-1773, Sept. 1997.