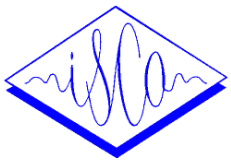

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Recognition Accuracy Methods and Measures

Frank H. Wu and Monica A. Marics

U S WEST Technologies
4001 Discovery Drive, Boulder, CO 80303, USA

ABSTRACT

This paper presents a standard data collection methodology and analysis measures which have been developed for reporting speech recognition accuracy. With this standard process, it is now possible to compare data collected at different times and on different systems. Moreover, other accuracy figures can now be interpreted based on knowing how their data were collected and analyzed.

I. INTRODUCTION

Advances in speech recognition in recent years have prompted the introduction of many automated voice-activated applications and services in areas varying from laptop PCs to telecommunication networks [1]. One of the main interests in the communications industry to date has been the development of speech recognition systems for various applications. However, as different recognition systems emerge, it is important to focus on how to assess the recognition performance of all the systems in a standard way [2].

For any speech recognition system that is either tested in a laboratory environment or deployed in the field with live users, the novel recognition accuracy methods and measures presented here provide information on (1) how users *really* use the system, and (2) how often the system responds *appropriately* to user input, which contrasts with the traditional method for assessing accuracy.

Assuming the availability of both user input (utterances spoken by the user are captured by the system via recording) and system response (system output stored in a file), the process and measures proposed here, referred to as the human "hand-coded" accuracy method, include two major steps: (1) data entry and (2) data analysis. For data entry, a spreadsheet (Fig. 1) is used to facilitate the sorting of various columns of data, with each column tracking a specific type of information. Each row in the spreadsheet corresponds to usage information obtained from the data collection and measuring process. That is, each row tracks a single user/system event. The column information includes the date of the event, subject number, event information, prompt played by system, user's action, system's response, etc.

For data analysis, information in the spreadsheet can now be sorted in various ways, depending on the characteristics of the recognition system being evaluated. Typically, the analysis includes information on a) action match to vocabulary (how well actual user commands matched the recognition vocabulary), b) timing of action (whether the subject gave speech commands after the prompt was played, which can help to determine if talk-through technology is needed for that particular application), and c) recognition accuracy (the percentage of valid commands that were correctly recognized).

The scope of this paper will be as follows: All the definitions used for measuring the system's recognition performance are listed in Section II. Two common methods (system "self-reported" accuracy and human "hand-coded" accuracy) for measuring recognition accuracy are discussed in Section 3. Section IV identifies the initial procedures required for data collection. The proposed procedures and measures (including coding user input, coding system response, coding true user attempts, data verification, and calculation of recognition accuracy) are presented in Section V. Section VI shows an example of how the proposed technique can be used, and contrasts system "self-reported" data with human "hand-coded" data. Section VII will draw conclusions.

II. DEFINITIONS

In general, the recognition performance of a speaker-dependent and/or speaker-independent recognition system can be measured using the following definitions:

in-vocabulary:The contents of a user's utterance exactly matches the content of a template pre-defined in a vocabulary directory, no more and no less. In the case of the speaker-dependent system, the speaker should be the same individual who placed the template in the directory.

mixed-vocabulary:The contents of a user's utterance has an in-vocabulary word/phrase embedded in the speech (defined as "mixed+" here), such as "I would like to talk to John Smith, please" with John Smith being the only in-vocabulary phrase in the vocabulary set. This also includes cases where the user says only part of an in-vocabulary word/phrase (defined as "mixed-" here), such as "Smith" rather than "John Smith."

out-of-vocabulary:Anything spoken by the user that does not match any template, either partially or totally, in the

vocabulary directory. In the case of the speaker-dependent system, this includes utterances spoken by an individual who did not originally place the template in the directory.

input error:Either no input is received and the system times out, or the system judges that the user's speech signal may have superimposed a defined beep (if no talk-through is allowed), too late, too long, too soft, or other errors.

valid input:The user says an in-vocabulary word and it is passed to the recognizer.

first attempt:The first time the user attempts to says a word/phrase.

second attempt:The second time the user says a word/phrase.

third attempt:The third time the user says a word/phrase.

correct recognition:The system correctly recognizes an in-vocabulary utterance spoken by the user.

correct rejection: The system correctly rejects an out-of-vocabulary utterance spoken by the user.

incorrect rejection:The system incorrectly rejects an in-vocabulary utterance spoken by the user.

misrecognition: The system incorrectly recognizes either an in-vocabulary word as another existing (in the vocabulary directory) in-vocabulary word or an out-of-vocabulary word as an existing (in the vocabulary directory) in-vocabulary word spoken by the user.

For mixed-vocabulary performance, measuring recognition, rejection, or misrecognition depends on what features and capabilities the recognition system has. For instance, if the system features key-word spotting, the rejection of mixed+ input words will be measured as incorrect rejections. On the other hand, if the system has only a discrete word/phrase recognition capability, then rejection of either mixed+ or mixed- words will be measured as correct rejections.

III. RECOGNITION ACCURACY METHODS

Two methods are commonly used for measuring the accuracy of a speech recognition system: system "self-reported" accuracy and human "hand-coded" accuracy.

III-1. System "Self-Reported" Accuracy Method

When a user accesses a speech recognition system, the system generally has the capability to record and process information on (1) how the user uses the system (e.g., what voice input the user provides to the system) and (2) how the system responds to the user's input. The system may then calculate its performance based on the "system response vs. user input" data from the collected information.

The system may report the number of events recorded at each stage of speech processing. However, the system data cannot reflect whether the system actions were appropriate based on the user input. For example, the system may record a "recognition complete" event, but this does not mean the recognition was correct!

To determine the actual recognition accuracy of the system, humans must listen to the actual recognition events. We refer to this process as human "hand-coded" accuracy.

III-2. Human "Hand-Coded" Accuracy Method

The focus of this paper is to present the human "hand-coded" accuracy method. In this method, the events occurring at each stage of speech processing are not only recorded but also hand-coded by humans via a process and measures which will be discussed in later sections. Thus, user's input is checked against the system's response to the input.

In contrast to the system "self-reported" accuracy method, the method presented in this paper can provide information on (1) how the user *really* uses the system and (2) how the system responds *appropriately* to the user's input. Moreover, the procedures and measures associated with the human "hand-coded" accuracy method can be applied to any speech recognition system that is either tested in a laboratory environment or deployed in the field with live users.

The proposed method for hand-coding the data consists of six steps: data collection, user input coding, system response coding, attempts coding, verification, and calculation. In the following sections, all the steps will be discussed in detail.

IV. INITIAL PROCEDURES

First, to conduct an analysis of the recognition accuracy of a speech recognition system, a data collection time period should be defined. This time period should be chosen so that it covers periods that have low, normal, and high usages.

Second, a group of users should be selected randomly from all users. All the callers should have used (or made an attempt to use) the system at least once during the time period that was chosen for conducting the analysis.

Once the time period and user group are determined, the information on all system uses made by the user group during that time period are collected; and the data are then analyzed using a set of pre-defined procedures, as described in Section V.

V. RECOGNITION ACCURACY MEASURES

In this section, the proposed procedures and measures (including user input coding, system response coding, true user attempts coding, data verification, and calculation of recognition accuracy) are presented.

V-1. Procedures

The recognition accuracy information is collected and analyzed using the following procedure.

1. Every time the user accesses the speech recognition system, the user's voice input to the system and the system's response may be collected and recorded in detailed session reports which are stored in a database.
2. Detailed session reports are pulled from the database for use of the system during the chosen time period.

3. Speech coders listen to each recorded utterance and compare it to the detailed session report and the templates in the vocabulary directory. For each call attempt, the following information is coded onto a coding sheet (Fig. 1):
   a. the user identity and session number,
   b. the contents of the utterance including what is said, background noise, lip smacks, breathing, etc.,
   c. the type of system input error that occurs, if any,
   d. whether the utterance is in-, mixed-, or out-of-vocabulary compared to the contents of the vocabulary directory, and
   e. the *actual* system's recognition results, i.e., a correct recognition, a misrecognition, a correct rejection, or an incorrect rejection.
4. The coding sheets are then verified and checked by a second coder.
5. To find out the actual first, second, and third attempt recognition performance of the system, information collected in the spreadsheet on first, second, and third attempts are reclassified ("RE-CHECK" column in Fig. 1) based on the user's input to the system. For example, if the user speaks "over the beep" on the first try, and speaks clearly and is recognized on the second try, this would be reclassified as a correct recognition on the *first* attempt.
6. The data are entered into a computer spreadsheet and double checked.
7. The data are then analyzed for performance, user inputs, input errors, and in- and out-of-vocabulary recognition using calculation methods described in the section below.

Notice that in procedures 3-d and 3-e, the information coded onto the coding sheet will be what the user *really* said and how the system responded, thus reflecting the *actual* system recognition results rather than what the system's session report ("self-reported") data indicate. Moreover, information on in-vocabulary, out-of-vocabulary, and mixed-vocabulary of the user's input can now be obtained from the coded data, which is not possible with the system "self-reported" data. These are the key steps which separate human "hand-coded" accuracy from system "self-reported" accuracy, thus providing a true performance of the recognition system.

Also, as the example indicated in procedure 5, it is important to reclassify the attempts in the presence of user error (e.g., all the input errors defined earlier). This is to ensure fair measurement of the *system*'s performance based on appropriate usage of the system by the user.

The advantages of using the "hand-coded" method are further illustrated in the example discussed in Section VI.

V-2. Calculation Methods

Given the definitions in Section II and the coding information gathered through the process described in Section V-1, the recognition accuracy figures of a system can be calculated using the following methods.

1. The total number of usages U (in the case of a telecommunication application, U could be the total number of single calls) can be found as

$$U = \#\text{ of 1st attempts in the "ATTEMPT" column of the coding sheet} \quad (5.1)$$

2. The total number of system recognition attempts (SRA) can be defined as

$$SRA = SRA\_1 + SRA\_2 + SRA\_3, \quad (5.2)$$

where SRA_1, SRA_2, and SRA_3 are number of 1st, 2nd, and 3rd attempts in the "ATTEMPT" column of the coding sheet, respectively.

3. According to the definition of input error, the total input errors (IN_ERR) is

$$IN\_ERR = IN\_T+IN\_B+IN\_L+IN\_G+IN\_S+IN\_O, \quad (5.3)$$

where IN_T, IN_B, IN_L, IN_G, IN_S, and IN_O are input errors in the "INPUT ERROR - time out, over beep, too late, too long, too soft, and other" columns, respectively.

4. Based on the valid input (VI) definition in Section II, the total number of VIs can be sorted from the spreadsheet and computed from the coding sheet as

$$VI = VI\_R1 + VI\_R2 + VI\_R3 + VI\_M + VI\_I, \quad (5.4)$$

where VI_R1, VI_R2, and VI_R3 are number of in-vocabulary correct recognitions in the "IN - recognition" column that are 1st, 2nd, and 3rd attempts in the "RE-CHECK - 1st, 2nd, and 3rd" columns, respectively. And, VI_M and VI_I are number of in-vocabulary misrecognitions and incorrect rejections in the "IN - misrecognition and rejection" columns, respectively.

5. By applying a similar sorting technique in calculation method 4 (Eq. 5.4) to the data on the coding sheet, one can compute the percentage of correct recognition of the system on the first attempt (REG_1), by the second attempt (REG_2), and, if necessary, by the third attempt (REG_3). That is,

$$REG\_1 = VI\_R1 / VI, \quad (5.5)$$

$$REG\_2 = (VI\_R1 + VI\_R2) / VI, \quad (5.6)$$

and

$$REG\_3 = (VI\_R1 + VI\_R2 + VI\_R3) / VI, \quad (5.7)$$

where VI, VI_R1, VI_R2, and VI_R3 are defined by Eq.5.4.

Based on the data collected in the coding sheet, the performance of the system on out-of-vocabulary correct rejection and misrecognition can also be computed. Moreover, analysis can also be conducted on the system's performance for all the in-vocabulary utterances, [2].

VI. COMPARISON OF ACCURACY MEASUREMENT METHODS

The example below illustrates how this method can be used, and contrasts system "self-reported" data with human

"hand-coded" data.

Consider a multi-language World Cup ticket ordering system that uses a speech recognition component as a front end. After prompting the system user to speak a country's name, the system recognizes the name and sends it to an operator who speaks the language of that country to process the order. In this case, the recognition system is speaker independent. It has no talk-through and no key-word spotting. It provides the user with up to three tries (attempts) before routing to a human operator. And, it captures each utterance for later playback.

The following tables, with 100 calls processed, contrast the two methods. (Note: there could be up to three attempts made by a user during a single call.)

Table 1: System "Self-Reported" Accuracy

total # of calls	100
# of calls w/ valid (in-vocabulary) input	90
# of calls successfully completed (country names recognized)	88
# of calls rejected as invalid input (country names are not on the list)	10
Performance (in-vocabulary recognition)	97.8%

With the system "self-reported" accuracy, the recognition performance was 97.8%.

First, with the "hand-coded" accuracy method, it was possible to find out that there were 117 input attempts made by users during the 100 calls, and of all the attempts, only 85 were in-vocabulary inputs that had the chance to be correctly recognized. Secondly, the system performance on all the in-vocabulary first attempts was only 78.8%. And, finally, by-the-third attempt recognition was 91.8% which reflects the true *successful calls* rather than the 97.8% reported by the system, which also included all the misrecognitions.

VII. CONCLUSIONS

A standard data collection methodology and analysis measures for reporting recognition accuracy have been developed. This includes procedures for those who are hand coding the data.

Table 2: Human "Hand-Coded" Accuracy (Fig. 1)

total # of calls	100
# of attempts made by users	117
# of attempts w/ valid (in-vocabulary) input	85
# of attempts w/ out-of- & mixed vocabulary	15 & 3
# of attempts w/ input errors	4
correct recognition (on 1st, 2nd, & 3rd attempts)	78 (67, 8, & 3)
misrecognition (in-, out-of-, & mixed-vocabulary)	17 (3, 9, & 5)
correct rejection (out-of- & mixed-vocabulary)	14 (11 & 3)
incorrect rejection (in-vocabulary)	4
Performance (in-vocabulary recognition) on 1st, by 2nd, & by 3rd	78.8%, 88.2%, & 91.8%

The methods and measures presented here provide information on (1) how speech recognition system users *really* use the system, and (2) how often the system responds *appropriately* to the users' input.

With this standard process, it is now possible to compare the actual recognition performances measured, based on data collected at different times and on different systems. Moreover, other accuracy figures can now be more realistically interpreted by using the information on how their data were collected and analyzed.

References

[1] TechLink (Monitor: P. Gonzalez), "1993: The Year in Review," TechMonitoring, SRI International, December 1993 - January 1994.

[2] M. Marics and F. Wu, "Recognition Accuracy Methods and Measures," The 5th Telco Speech Research Workshop (St. Louis, MO), February 1994.

GENERAL		ATTEMPT			RE-CHECK			CONTENT	INPUT ERROR						IN			MIXED+			MIXED-			OUT		misrecognition*	COMMENTS
user ID	session #	1st	2nd	3rd	1st	2nd	3rd	spoken by user	time out	over beep	too late	too long	too soft	other	recognition	misrecognition*	rejection	recognition	misrecognition*	rejection	recognition	misrecognition*	rejection	misrecognition*	rejection	misrecognized as "name"	
caller #1	w0001	1			1			Germany							1												
caller #2	w0002	1			1			America																1		Cameroon	"USA" on the list
•	•							•																			
•	•							•																			
•	•							•																			
caller #99	w0099	1			-			(input error)		1																	
	w0099		1		1			Brazil							1												
caller #100	w0100	1			1			Saudi															1				should be "Saudi Arabia"
TOTALS		100	10	7	106	8	3		2	1	0	0	1	0	78	3	4	0	5	1	0	0	2	9	11		

Figure 1 Recognition Analysis Coding Sheet