USABILITY FIELD-TEST OF A SPOKEN DATA-ENTRY SYSTEM

Marcello Federico and Fabio Brugnara and Roberto Gretter

ITC-Irst - Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo, Trento, Italy.
{federico,brugnara,gretter}@irst.itc.it

ABSTRACT

This paper reports on the field-test of a speech based data-entry system developed as a follow-up of an EC funded project. The application domain is the data-entry of personnel absence records from a huge historical paper file (about 100,000 records). The application was required by the personnel office of a public administration. The tested system resulted both sufficiently simple to make a detailed analysis feasible, and sufficiently representative of the potentials of spoken data-entry.

1. INTRODUCTION

The main goals of the SpeeData [1] project were two. First, the development of a multi-lingual speech recognition technology suited for complex data-entry applications [1]. Second, the evaluation of the ergonomics of a spoken data-entry user interface. This work will report on the latter issue. As a user-centered project, SpeeData followed the so called *usability engineering* project life-cycle proposed in [4]. Briefly, according to [4] usability aims at characterizing the ergonomics of a user interface (UI) along five dimensions:

- *learnability*, i.e. the UI should be easy to learn;
- *efficiency*, i.e. the UI should allow high productivity;
- *memorability*, i.e. the UI should be easy to remember;
- *errors*, i.e. the UI should have a low error rate and supply easy error recovery methods;
- *satisfaction*, i.e. the UI should be pleasant to use.

Usability evaluation is introduced during the interface design and repeated at significant stages of its development. Cheap and simple *discount usability* [4] techniques can be employed, i.e. user and task observation, scenarios, thinking aloud and heuristic evaluation. For final testing, questionnaire and logging of actual use are also suggested.

This work presents results from a test performed on a personnel office data-entry task of a public administration. The historical file to be electronically stored contains about 100,000 records relative to 7,000 employees. The evaluation was not aimed at performing a competitive analysis with other input media. It is known [2] that a fair comparison would require an experimental set-up that tries to optimize the "transaction cycle" of each input medium, i.e. the number of user actions necessary to perform a task. Practically, there was not a conventional data-entry system as a reference for the evaluation. Moreover, the considered task requires the user interpreting the original documents to extract the relevant data. The speed of the data-entry process would be relevant only if it could be separated from the time required for interpreting the data, which in fact depends on the single document and the user skill.

[1]This work was supported by the European Commission, Telematics Application Programme, project reference number LE 1999.

2. TASK AND USER INTERFACE

The data-entry of absence records requires first filling in a form that identifies the employee, then a sequence of absence forms that specify the year, type, start, end, and duration of each absence. For some absences, more specific time information can be entered through a separate form. A picture of the user interface screen is shown in Figure 1. During a session, the user can fill the data fields of the current form by voice, mouse or keyboard. The user can also execute some general commands, e.g. clean or remove the current form, open a new form, etc.. Finally, the user can visit and possibly modify previously inserted forms. Fields can be filled by continuous speech specifying as many couples keyword-value as needed. Empty fields can be filled in without uttering their keywords, thus simplifying data insertion. Further, uttering an isolated field keyword selects that field. As this modality toggles a more selective language model, it is useful to perform corrections.

3. EVALUATION FRAMEWORK

A data-entry system was set-up into an office and employed for several months by 12 (female) users. All the users were familiar with computers and had used a first version of the prototype. Each speaker performed a speaker enrollment session and received some training before starting to use the system.

After the training phase, the users started to regularly use the system about 2-3 hours per week. During usage, the system produced a log file, that traced all the operations performed by the users. The users were informed about the log file. Speech signals were not recorded during this test. After 5 months of usage the users were asked to fill in a subjective satisfaction questionnaire. The users had to express their agreement with ten statements about usability of the system, by means of a 1-5 rating scale.

The log file recorded about 166,000 operations. Figure 2 shows, for each user the percentage of operations performed by speech. It results that speech was used as input medium between 90% and 97% of the times. This percentages increase considerably if only field assignments operations are taken into account (see Figure 2).

In the following, measures related to the usability factors are presented and discussed.

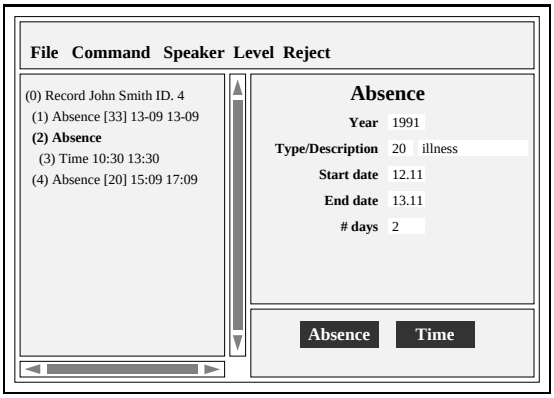

Figure 1: User interface (English translation). The screen shows the current form (right center), buttons to open new forms (right down), and a summary of the already entered forms (left).

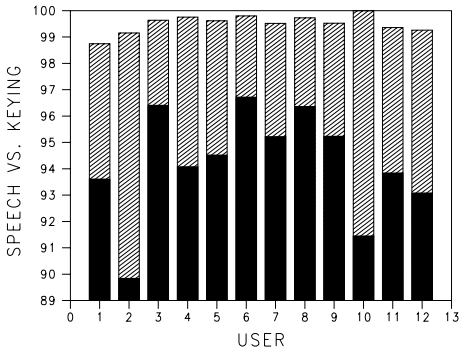

Figure 2: Percentage of operations performed by speech versus keyboard/mouse, by each user. Black bars account for all kinds of actions (i.e. button clicks, field selections, and field assignments). Gray bars only consider field assignments.

4. MEASURES OF USABILITY FACTORS

4.1. Errors

As a usability goal, a speech based data-entry system should allow low error rates and should supply easy error recovery methods. With respect to a conventional keyboard data-entry interface, interaction errors may be due to wrong operations performed by the user and to speech recognition errors. According to a rough estimate, recognition errors indeed occur about 10 times more than other errors. The analysis of errors focused on some issues that should influence usability:

- *rate*, i.e. how often do errors occur?
- *concentration*, i.e. do errors concentrate in time?
- *repetition*, i.e. how reliable is error correction?
- *patterns*, i.e. where do errors occur more often?

As a further approximation, only field assignments were considered in this analysis. The reasons are several: field assignments are the most relevant operation performed by speech, almost all speech recognition errors occur during field assignments, and, last but not least, they are more easily tracked from the log file. In fact, errors on field assignments can be assumed where already filled data-fields are re-entered. Other types of errors will however be included by the subsequent analysis of efficiency. The error rate for each user is shown in Figure 3. (Note that the user ordering is based on this parameter.) The computed error rates result between 5% and 10%. Past experience on machine dictation suggests that error rates below 10% are generally accepted, provided that error correction is easy and reliable. Moreover, we also know that it is quite difficult to subjectively feel small error rate variations (e.g. 2%-3%) over a long period. The questionnaire response fits these hypotheses quite well. Nine users of twelve scored the "error" quality over the "average" satisfaction threshold, which was set to 3.6, according to [4].

About how error concentrate in time, an interesting model is given by the error waiting-time distribution. This tells the probability of waiting less or equal than a given time λ for the next error. The time dimension is expressed in terms of progressively entered data fields. By assuming the error probability $\Pr(\epsilon) = p$ and independence among errors, the geometric distribution [3] follows:

$$F(\tau \leq \lambda) = 1 - (1-p)^{\lambda} \tag{1}$$

The α-quantile [3] of the geometric distribution is:

$$\tau_{\alpha} = \frac{\log(1-\alpha)}{log\ (1-p)}.$$

In Figure 4, the $\alpha = .3, .5, .7, .9$ quantiles of the geometric distribution are plotted versus each user. The mean waiting time of each user is also plotted, which is $\frac{1}{p}$. The geometric distribution in some way represents the ideal situation of independence among errors. For instance, by taking user 1, (error rate ≈ 0.05), the ideal error waiting time should be less or equal than 15 data fields with 50% chance. The corresponding quantiles of the empirical distribution (Figure 4), derived from the log file, show a quite different behavior. Error-to-error distances are in reality more concentrated than with the geometric distribution. For instance, the .5-quantile of the empirical distribution is located around the .3-quantile of the theoretical one. The geometric model indeed fits well the empirical distribution for the higher quantiles. From the usability point of view, this means that there are input patterns for which speech recognition is more difficult, more errors may occur and correction may need more than one attempt. Moreover, locally in time degradation of performance may also occur due to input channel problems, like microphone settings, input level settings, etc.

An interesting and related distribution to examine is that of the number of correction attempts. Assuming the geometric distribution, where this time $p = 1 - Pr(\epsilon)$, theoretical and empirical quantiles for each user were again compared. The mismatch between the theoretical and empirical models is again large. If the former gives, for all users, 90% change to repair an error with one attempt, the empirical model requires two attempts.

The last considered issue is about frequent error patterns. Strange error patterns could reveal usability problems. In Figure 5, the estimated error rates on different input data-types are plotted. Data-types have been grouped into six categories: **CD** codes, i.e. list of two digit codes, **DS** codified descriptions, e.g. illness, **DT** dates, e.g. 10.9, **DY** days, i.e. 1..365, **TM** time, e.g. 10:30, and **YR** years, i.e 1989..1995. By giving a look at Figure 5, we see that the code data-type has generally a quite high error rate. This is surprising, as numbers are generally well recognized by the system. By inspecting the log file, it seems that many of such errors are in fact due to misrecognitions of the target data-field. For instance,

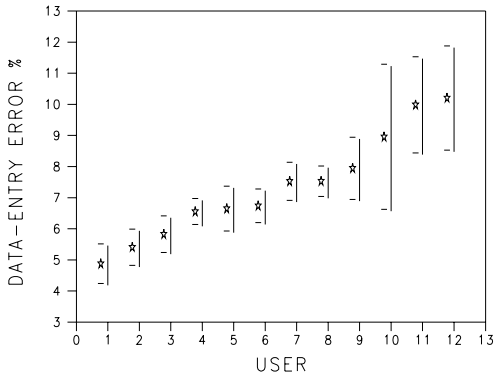

Figure 3: Percentage of wrong field assignments for each user. The 99% confidence interval is shown for each estimate.

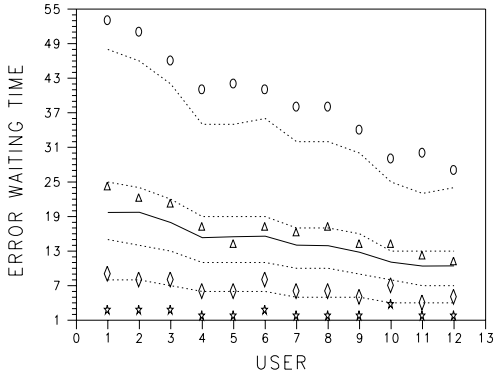

Figure 4: Distribution of the time distances between errors. The sharp line represents the empirical mean. Dotted lines show, from the button upward, the .3,.5, .7, and .9 quantiles of the geometric distribution. Symbols represent the same quantiles of the empirical distribution.

the type keyword is confused with the start-date one. This problem suggests that usability can be improved by a more careful choice of keywords.

4.2. Efficiency

Efficiency aims at evaluating if the proposed interface allows the user to reach an high productivity. Efficiency refers here to the transaction cycle, i.e. the number of simple actions required to perform a task. A task is sized by the number of data to enter plus the number of forms to fill-in. Simple operations are limited to field selections, field assignments, and form changes. It can be noticed that all possible errors made by the user impact on efficiency. Two different kinds of efficiency measures have been considered:

- *single operation efficiency,*
- *multiple operation efficiency.*

Single operation efficiency aims at evaluating the UI without considering the potential advantages of using speech as input medium. Even if more actions, i.e. field assignments, can be performed with a single utterance, the efficiency measure takes into account all the single operations. Hence, the maximum achievable efficiency is 1. Filling in a form of three field, for instance, requires at minimum four operations: three assignments plus one form change.

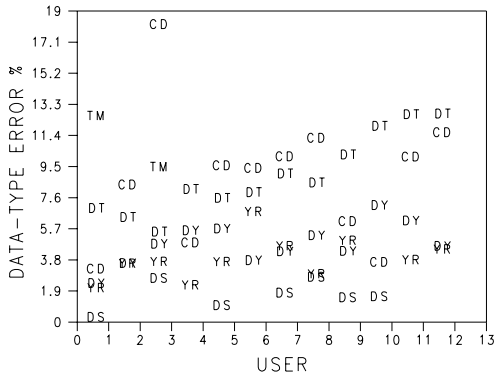

Figure 5: Error rate on different data-types by each user. Data-types are grouped into six categories: CD code, DS description, DT date, DY day, TM time, and YR year.

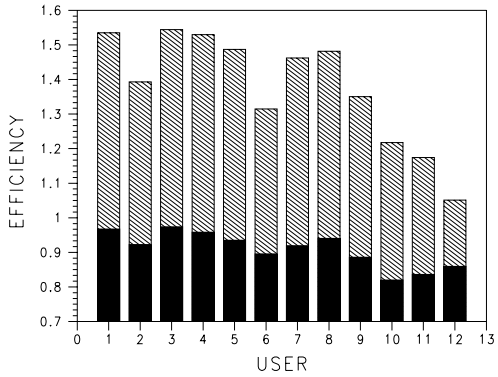

Figure 6: Efficiency measures expressed as the ratio between task size and number of performed actions. The black bar accounts for all single operations, even in case of multiple operation utterances. The gray bar counts multiple operation utterances as single operations.

Probably, even an optimized keyboard based UI would need as many operations. A bar plot of the single operation efficiency measure is shown in Figure 6. As expected, efficiency decreases along the users axis (i.e. increasing error rate order). Small oscillation are probably due to different data-entry strategies employed by the users. The first eight users scored between .90 and .96. As an absolute evaluation of the scores is not easy to carry out, something can be gained by looking at the relative variations. By reasonably assuming 95% efficiency as an optimal target, eight users of twelve scored within a 6% interval, while ten within a 10% interval. Multiple operation efficiency considers the advantage of entering more data with a single utterance, e.g. the user may avoid moving often its attention to the screen. Hence, efficiency now counts utterances with more actions as single operations. The impressive efficiency gain is shown in Figure 6. For the eight best performing users, efficiency grows by a factor between 1.5 and 1.6. Large oscillations between users can be appreciated, which probably reflect different input strategies. As a final comment, all the users exploit multiple operation utterances for field assignments. It was observed from the log file, that on the average users enter from 1.5 to 2.0 fields per utterance.

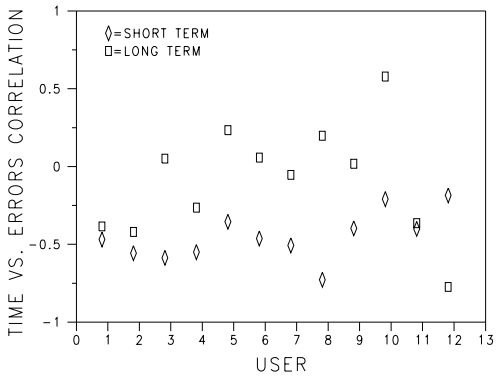

Figure 7: Correlation coefficient between errors and usage time. Squares plot the correlation over all the usage period, diamonds plot the correlation within the performed sessions.

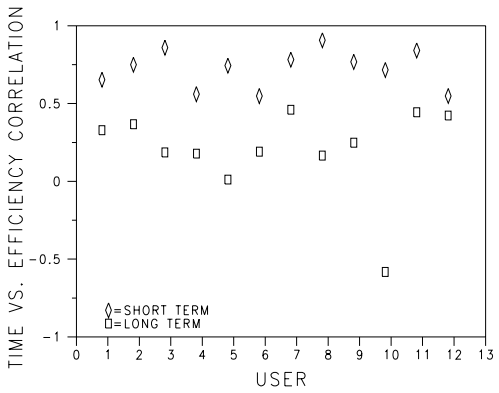

Figure 8: Correlation coefficient between efficiency and time. Squares show the correlation computed overall the usage period, diamonds show the within-session correlation.

The subjective evaluation of the efficiency quality told that only four speakers of twelve believe that efficiency is good and higher than by keying. However, nine users of twelve declared themselves satisfied with the achieved performance.

4.3. Learnability/Memorability

As described in Section 1, learnability and memorability describe respectively how hard it is for a new user to become acquainted with the system, and how much of the proficiency acquired through a period of usage is kept after the user has been away from the system for some time.
By the questionnaire, all users gave very high ratings in this respect. This means that the system is easy to learn and remember, and offers the users a consistent interface, well aligned with the perception that they have of the task itself.
An analysis of the log file was also carried out, to see how error-rate and efficiency are correlated with time. The choice of these measures was motivated by the observation that the performance of a speech-based system is sensitive to several aspects of the user behavior that are implicitly learned by the user during usage (e.g. voice loudness, absence of extraneous utterances, positioning of the microphone, etc.).

In Figure 7, the correlation between the error-rate and time is plotted, for the different users. The short-time correlation considers within sessions variations, while the long-time one runs over the whole usage period. A negative value indicates that error tends to decrease during usage, and this is the expected behavior. As can be seen in the figure, short-term correlation is negative for each user, showing that the users actually tend to improve their dictation during a session. As for long-time correlation, this is somewhat worse, with three users having a significative positive correlation. This could be due to the fact that the proper "dictation modality" is not well remembered across session. Moreover, the long-term index could be biased by the presence of some session that is exceptionally bad due to improper microphone setting.
Concerning efficiency (Figure 8), with the exception of the long-time index of a single user, both short-time and long-time indexes are positively correlated with time. This suggests that the users actually learn to use the system features that help in reducing their work. Also in this case, the short-time indexes are better that the long-term ones, and the reasons are probably the same given for the error-rate correlation.

4.4. Satisfaction

This quality could be only estimated by means of the questionnaire. All the users scored over the neutral satisfaction threshold, i.e. 3.6, and the average score was 4. This confirms that the speech input medium is well accepted by the users as an alternative to keying.

5. CONCLUSIONS

A usability evaluation of a speech based data-entry system was presented. Evaluation was based on a questionnaire and a log file containing statistics relative to 5 months of real usage by 12 users. Approaching objective measures related with usability qualities requires a significant amount of work: setting up a system into a real world environment, gathering a sufficiently large user base, collecting usage data by logging, and defining a set of measurable quantities that are representative of the qualities to be evaluated. This paper goes through all the above phases and discusses a set of measures with respect to a system which has been deployed by real users.

6. ACKNOWLEDGEMENTS

The authors thank B. Angelini and D. Giuliani of ITC-Irst, G. Farace, M. Stambul, and D. Zocchi of Informatica Trentina Spa, U. Ackermann of FORWISS, M. Miorandi of P.A.T. for their contribution to the SpeeData project.

7. REFERENCES

[1] U. Ackermann, B. Angelini, F. Brugnara, M. Federico, D. Giuliani, R. Gretter, and H. Niemann. SpeeData: a prototype for multilingual spoken data-entry. In *Proc. of EUROSPEECH*, pp. 1807-1810, 1997.

[2] R. I. Damper and S. D. Wood. Speech versus keying in command and control applications. *International Journal of Human-Computer Studies*, 42:289–305, 1995.

[3] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, Singapore, 1974.

[4] J. Nielsen. *Usability Engineering*. Academic Press, San Diego, CA, 1995.