Human-Computer Interaction -- INTERACT'03
M. Rauterberg et al. (Eds.)
Published by IOS Press, (c) IFIP, 2003, pp. 105-112

Error Resolution Strategies for Interactive Television Speech Interfaces

Aseel Berglund & Pernilla Qvarfordt

Human Centred System, Depart. of Computer and Information Science, Linköpings universitet, SE-581 83 Linköping, Sweden

{asebe, perqv}@ida.liu.se

Abstract: Using speech input to augment the remote control can be an alternative interaction technique for interactive television. However, little is known about how to design such a system that is suitable for the home environment. In this paper we explore possible error resolution strategies in the case of speech recognition errors in a TV-setting. From previous research two techniques have been identified: repetition of input by the user and choice of interpretation. From these two techniques, four alternative strategies, suitable for a TV-environment, were designed and tested with a Wizard-of-Oz method: one with repetition and three based on choosing the best alternative from an n-best list given in audio and visual mode, alone or combined. The results show that displaying an n-best list gives the most efficient interaction. Redundant audio feedback does not influence the performance.

Keywords: Speech input, error resolution, interactive television, electronic program guide (EPG), Wizard-of-Oz

1 Introduction

In the near future we will have many thousands of TV channels to choose from. Consequently, selecting TV content may become too cumbersome to handle with a remote control, since it does not support search capabilities. In the prospect of this, we investigate methods to ease the burden of the viewer to find what she is looking for. Thus, other interaction techniques for this setting are currently explored; one example of such interaction technique is speech input. Both Portolan et al. (1999) and Ibrahim, Lundberg and Johansson (2001) have shown that speech can be an attractive interaction technique for this setting. However, speech interfaces are not without problems, one of them is that the speech recognizers make errors when interpreting users' utterances and these errors can be a cause of irritation and low performance. In studies of dictation systems, Halverson et al. (1999) have shown that error correction consumes a significant portion of users' effort, even with as high recognition accuracy as 98%. In another study, they (Karat et al., 1999) showed that recovery from speech recognition errors took longer time than to enter the text. It is unlikely that speech recognizers will produce a perfect performance in the near future. Good and efficient error recovery strategies are therefore essential for viewer acceptance of speech interfaces.

Interactive television poses both opportunities and challenges for designing good speech interaction techniques. The large TV screen can be used to display visual information, but at the same time the distance between the viewer and the screen limits its resolution. The interaction techniques also have to fit the relaxed environment that surrounds the TV, as well as the tasks, which are associated with entertainment, rather than work. A large part of the current design exploration has to do with creating a natural, non-intrusive, and efficient interaction technique that matches these requirements.

In the rest of this paper we first review and synthesize the literature on error resolution strategies for speech interaction. We then propose and analyze four error resolution strategies for the interactive TV-setting. These strategies are then evaluated using Electronic Program Guide (EPG) prototypes, onscreen TV guides, in a Wizard-of-Oz study. Finally, we discuss and conclude the pros and cons of each strategy.

2 Error Resolution Strategies

Speech recognition technology contains errors since it relies on statistical interpretation of the users' utterances. This is the reason why error resolution has been explored since the invention of speech recognizers. For example, Martin and Welch (1980) developed a method based on repetition. The speech recognizer stored the preliminary interpretation results of the user input in a buffer. Then, the buffer was actively and interactively edited by the users by for example repeating their spoken input. The idea to let the user repeat the erroneous output has also been used, with some modifications such as eliminating elements from the buffer that are known to be incorrect, by Ainsworth and Pratt (1992) as well as Murray, Frankish, and Jones (1993). A second method created by further developing the former one and based on choosing an item from a list of alternative words was introduced by Murray et al. (1993). The list is alternative interpretations the speech recognizer makes from the user utterances; this list is also called an *n*-best list.

When Mankoff, Hudson, and Abowd (2000) surveyed existing commercial and research systems that utilize speech recognition, they found that these two error resolution strategies are the most commonly used. Repetition is usually done in the same modality as the original input. Choice of best alternative is usually done by selecting one of the alternatives from the *n*-best list presented for the user in form of a menu. The selection can be done in another modality than the original input, for example by selecting an item by a mouse in dictation tasks. In addition to these two strategies, they found that recognizers used a confidence level, a measure in the recognizer's interpretation confidence, to decide which interpretation to use. In this case the user was not involved in error resolution strategy.

It has been shown that combining recognition results from multiple input modalities, such as speech and gesture, can be effective to reduce the occurrence of recognition error (McGee, Cohen, and Oviatt, 1998, and Oviatt, 1999). Furthermore, Suhm, Myers, and Waibel (2001) studied four multimodal error resolution methods for a dictation task. The results showed that multimodal error correction was more efficient than unimodal correction; error correction was more efficient if the users changed their modalities. For example, using speech for input and gesture for correction was more efficient than using speech for both input and correction.

However, the existing studies have not considered a TV-setting where the input methods are limited to speech input and remote control device. Repetition and selecting from an *n*-best list are the techniques that can be used in a TV-domain. Repetition is called *correction* in this study, since the viewer has to correct the system's interpretation. Prompting the user to choose from the best alternatives is called *clarification*. Since a TV can present audio and visual information, clarification can utilize both of these output modalities, to be explored in the next sections. Clarification is triggered by the confidence level of the recognition system. The interpretation has a probability of correctness, and if this probability is lower than the confidence level, then the error resolution strategy is utilized. Otherwise the system just displays the right response. This is very similar to how humans resolve misunderstandings in conversations. It minimizes the user's effort in error correction.

2.1 The Correction Strategy

In the correction strategy viewers utter for example what they want to see, and the system displays the interpretation on the TV screen. An *n*-best list is generated with interpretations match viewer's utterance. The first item on the *n*-best list is always displayed. Thus, the item can either be correct or wrong. If the interpretation is correct the viewer can then confirm the interpretation by saying 'yes'. If the viewer does not say anything, it is interpreted as a confirmation. After confirmation, the TV switches to the desired channel. If the interpretation is wrong, the viewer can repeat the request, and a new interpretation is shown. The new interpretation is based on a new *n*-best list, and does not use knowledge from previous interactions with the system.

The advantage of the correction strategy is that it is more efficient than the clarification strategy when the confidence level is low, and yet the right alternative is the first item on the *n*-best list. The viewer does not need to go through an explicit clarification dialogue with the system, and still get the right answer from the system.

The drawback is that viewers have to repeat their input over and over if the speech recognizer's interpretation is wrong. Frankish, Jones, and Hapeshi (1992) have shown that repetitions tend to have lower recognition accuracy. In addition to this, users tend to adjust their way of speaking to what they believe is easier for the recognizer to interpret, which often has the opposite effect. However, Suhm et al. (1999) argue that if accuracy is significantly improved, an error recovery strategy based on repetition can outperform other strategies.

2.2 Three Clarification Strategies

In a TV-setting, it is possible to utilize both audio and graphics; therefore the *n*-best list can be presented in different modalities. This is what is done in the clarification strategy. The system could present alternatives in two different modalities, audio and visual, either alone or combined. Unlike the correction strategy, the system asks the viewer for clarification *only* when the probability of its interpretation is below the confidence level. Consequently, the viewer is aware that the system is uncertain, because it asks for advice.

Independent of which modality that is used, the drawback with clarification is that the viewer has to go through it even if the right item is the first on the *n*-best list. Another problem occurs when the right item is not on the *n*-best list.

2.2.1 Clarification, Audio

In the clarification audio strategy the clarification request is done by audio. The system speaks out the first item on the *n*-best list. The viewer then either repeats her request, or says 'yes', 'no', or 'next'. If the viewer says 'no' or 'next', the system continues to speak out the next candidate on the *n*-best list. In case the viewer repeats the original request, this will be treated as a 'no' or 'next'. In other words, the system reacts to "yes" as a confirmation and everything else as an indication to continue with the next candidate on the *n*-best list.

Speaking out the alternatives on the *n*-best list shows to the viewer that the system is not certain about its interpretation. The system does not do anything until an agreement between the viewer and the system has been established. This strategy is similar to what has been proposed by Brennan and Hulteen (1995) in order to establish a common ground between the user and the system. Their system gave evidence to the user in what stage of the processing it was uncertain about its interpretation, e.g. during the parsing or interpretation.

The drawback with clarification audio is that the viewer might have to traverse through a long clarification dialogue before coming to the right item.

2.2.2 Clarification, Visual

The second clarification strategy is visual, where the system displays an *n*-best list from which the viewer can choose the right candidate by either uttering her choice or saying the number of it.

The benefit with this strategy is that the viewer can access items further down on the *n*-best list without the need to go through a long clarification dialogue.

2.2.3 Clarification, Audio and Visual

Finally, with the third clarification strategy, audio and visual, the system both displays the *n*-best list and speaks the *first* candidate on the list. The system thereby gives partially redundant information. The viewer can act in the same way as with the visual clarification strategy.

A possible advantage with using both audio and visual feedback to the viewer is the redundancy offered by the system so the viewer can rely on either one of the two modalities to understand the top choice of the system. The drawback is that the audio feedback might be considered intrusive and disturbing, for example Ibrahim & Johansson (2002) have shown that viewers were concerned that spoken feedback would interfere with the sound from the TV.

3 Method

Which of the four strategies that will provide the best error recovery strategy for an interactive TV is difficult to know solely based on a theoretical analysis. We therefore decided to empirically evaluate the strategies before they were implemented. The different strategies were evaluated with a Wizard-of-Oz method. This approach allowed us to control and manipulate the amount of "recognition" errors and confidence levels independent of the constantly improving recognition technology. The Wizard-of-Oz system simulates an imperfect recognition process by randomly assigning low confidence level to utterances. This was partly because it is difficult for a human wizard to deliberately misunderstand spoken utterances. Furthermore, the random appearance of errors in the simulation was made to match the errors in real speech recognizers. The random error appearance was also congruent with the fact that it is usually very difficult for the viewer to understand the logic behind the misinterpretations of a speech recognizer.

Low confidence levels were only assigned to the participant's initial requests. When the participants had entered into an error recovery phase, no errors were assigned. Thus, during the test no errors occurred when the participant answered 'yes' or 'no'. To make the experiment manageable, low confidence levels were only assigned to TV program titles.

In this study we used a between-subject factorial design with four conditions: correction, clarification audio, clarification visual, as well as clarification visual and audio.

3.1 Participants

36 people, 14 women and 22 men, participated in this study. To each condition nine participants were randomly assigned.

The age span of the participants was between 16 and 50, with mean age of 24.5 years. Of the participants, 14 had used a speech recognition system before this study, 12 had used a digital TV-box before, and three had used spoken input to a digital TV-box before. The participants were familiar with computers, most of the participants (22 of them) stated that they used computers everyday, and two stated that they used computers at least once a week. The rest of the participants used computers several times each week, or once a day.

3.2 Prototype

An EPG was used to evaluate the error resolution strategies. EPSs are onscreen TV guides that are used to, among other things, view TV channel schedules, search for TV programs, and find information about TV programs. EPGs are often included in interactive TV-settings. For this study, four EPG prototypes in Macromedia's Director were developed, one for each feedback strategy. Expect for the feedback strategies, the prototypes were identical. The prototypes provided information about TV-shows and films.

The same *n*-best lists were used in all four prototypes. The *n*-best list for each program title contained three items. A commercial speech recognizer, Philips FreeSpeech 2000, was used to create the content of the lists by speaking the right answer to it. The items were picked from the recognizer's interpretations, and occasionally adjusted to resemble names of TV-shows and films. The *n*-best lists thus simulated the results from a speech recognizer with a vocabulary adjusted to the TV-domain. The three items were randomly assigned to the positions in the *n*-best list for each participant and scenario.

The audio feedback used in the prototypes was pre-recorded utterances by a human male voice. Since the intonation of the spoken feedback was considered important, a recorded human voice was used instead of synthesized voice. The intonation indicated the system responses were made as questions to participants, and that the system was not sure if it had interpreted the participants correct.

In the prototype with correction strategy, the confirmation was implemented as a screen that displayed the interpretation with text and graphics, see Figure 1. The graphics showed a screen shot from the TV-show/film, and the text provided a description of the content. Each item on the *n*-best lists had a confirmation screen attached to it.

Figure 1. The confirmation screen, with a screen shot and a description of the TV-show are displayed.

In the clarification audio strategy, the prototype was implemented to speak items on the *n*-best list one at a time if the confidence level was low. An example on how a clarification dialogue looked like with this strategy is shown in Figure 2.

V1:	show me Friends?
S1:	Finale?
V2:	No, Friends
S2:	Finds?
V3:	No
S3:	Friends?
V4:	yes
S4:	(Displays the confirmation screen with the TV program Friends)

Figure 2. An example of clarification audio dialogue (V=viewer, S=system).

Figure 3 shows an example of a screen with an *n*-best list that was used in the clarification visual, and clarification audio and visual. The participant could in these conditions either repeat the required item, or say the number of it. In the clarification visual and audio prototype, the first item on the *n*-best list was uttered at the same time as the *n*-best list became visible for the participant.

3.3 The Wizard and His Environment

The wizard was located in an adjacent room to the room where the study was conducted. The wizard had two screens in front of him. One showed the participants' view of the system and the other showed the participants sitting in a sofa.

The wizard interpreted the participants' input, and displayed either one of the feedback strategies,

or the required item. He controlled the prototypes responses with hot keys. The wizard followed a pre-defined randomized schedule to simulate the low confidence levels.

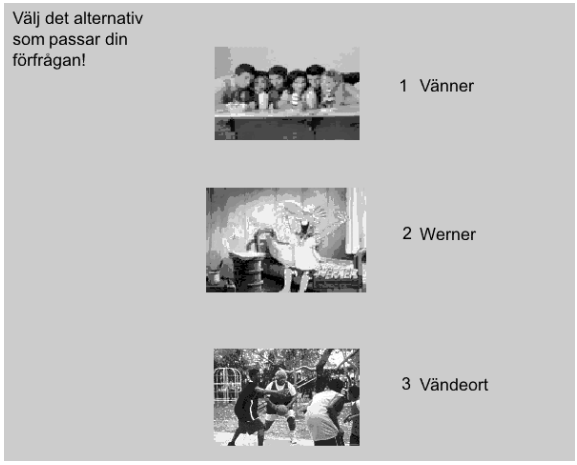

Figure 3. Example of a visual *n*-best list.

3.4 Scenarios

Each participant completed ten scenarios, four correctly interpreted and six with low confidence level. The ten scenarios corresponded to various situations from real life, and gave a motivation to see a certain TV-show or film, see Figure 4. The scenarios to be correctly interpreted were randomly assigned for each participant. The order of the scenarios as well as the order of the *n*-best list were also randomly assigned to each participant.

> You just had dinner and are now sitting in your sofa. You consider watching a film recommended by a colleague earlier today. You don't remember the exact name of the film, but think it was 'Fanny and Alexander' by the famous director Ingemar Bergman. Use the program guide to find the film.

Figure 4. An example of a scenario.

3.5 Procedure

The experiment was conducted in a home-like environment. The participant was seated in a sofa in front of a TV set, which showed the EPG. The participants had a remote control with a microphone dummy, into which they were requested to speak.

The experiment consisted of three parts: an introduction, the main study, and a post-test interview.

During the introduction, the purpose and the procedure of the study were explained. The participants were asked to fill in a pre-test questionnaire to collect background information. Then, each participant was given the opportunity to try the prototype in two scenarios, one illustrating correct interpretation (high confidence level), and the other low confidence level, where the participants were presented to the error resolution strategy used in their condition.

In the main study, the participants were given the scenarios one at a time. When the participants had finished a scenario, they were asked to fill out a questionnaire measuring their attitude towards the concept. When they had completed all ten scenarios, they were asked to fill in a post-test questionnaire, which asked for their general opinions of the system. All questionnaires in this study used a seven-grade Likert-scale.

The session ended with an interview. The purpose of the interview was to collect the participants' assessments of the concept, their satisfaction and experience from talking to the EPG.

4 Results

The feedback strategies were evaluated by measuring the total task completion time, and by collecting the participants' subjective evaluation of the strategies. In this study all the participants successfully completed all the scenarios.

In the following section, we present the results from measuring the total task completion time, general questionnaire, and post-test interviews.

4.1 Total Task Completion Time

The total task completion time showed that the error recovery strategy influenced the participants' performance, see Figure 5. Note that the task completion time was based on the six scenarios that involved error recovery, which was triggered by low confidence. The two conditions using a visual *n*-best list, *clarification visual* and *clarification audio and visual*, showed the shortest task completion times, while *clarification audio* had the longest task completion time. The general difference between the conditions was significant ($F(3, 32)=8.870$, $p=.0002$, one-way ANOVA).

The largest differences were between the two conditions that used a visual *n*-best list, and *clarification audio*, where the *n*-best list was spoken, as shown in Figure 5. These differences were significant (Scheffé post-hoc test, *clarification audio and visual* vs. *clarification audio*, mean diff.=78.2 s., $p=.001$, *clarification visual* vs. *clarification audio*, mean diff.=72.3 s., $p=.002$). On average, a

visual *n*-best list makes each question slightly more than seven seconds faster than a spoken *n*-best list. Although none of the other comparisons showed any significant differences, the differences are noticeable between *correction* and *clarification audio and visual* (mean diff.=36.7 s., p=.221) and between *correction* and *clarification audio* (mean diff.=41.6 s., p=.135). These differences were on average, 3.7 s. and 4.2 s. respectively for each question to the system.

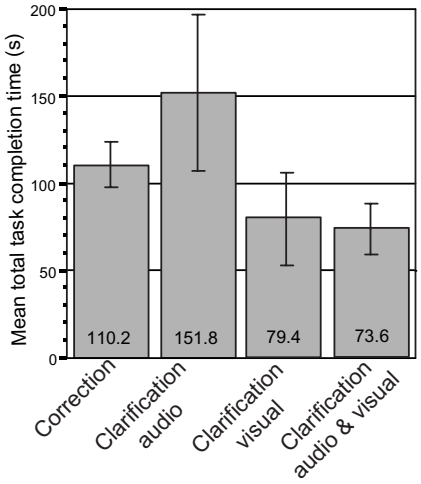

Figure 5. The mean total task completion time for each feedback strategy, with 95% error rates.

4.2 General Questionnaire

The results from the general questionnaire showed no statistically significant differences between the four conditions. In general the participants liked the EPG prototypes. On the question if they would like to use the guide in the future the average rating were between 5.78 and 6.67 on a seven-grade scale, where the lowest rating was given to the *correction strategy* and the highest to *clarification visual and audio*, see Table 1.

The participants, independent of condition, also believed the EPG was easy to use, see Table 1. Also at this question the *clarification audio and visual* received the highest ratings, however, the differences were not significant.

4.3 Post-Test Interviews

The participants in the *clarification audio* condition expressed that it was tedious to repeatedly request the system to give alternatives. They felt that it was their fault that the system made errors. Furthermore, the participants preferred to get some visual confirmation before switching to a program that could be the wrong one. They suggested that the alternatives should be presented textually, for example in a list. Another comment from the participants was that they believed that the system should say a complete sentence instead of just a TV program title, for example "Would you like to see Friends?" instead of "Friends?"

Condition	Likeability		Ease of use	
	Mean	SD	Mean	SD
Correction	5.8	1.093	6.3	.866
Clarification, audio	6.2	1.302	5.6	1.944
Clarification, visual	6.4	1.130	5.9	.601
Clarification, audio and visual	6.7	.500	6.4	.527

Table 1: Mean and standard deviation of the two questions on likeability and ease of use.

Although *correction* was not the slowest strategy, the participants found this strategy very tedious also. They tended to speak louder and articulate more thoroughly than the participants in all other conditions. All of them also stated that it would be better if the alternatives were presented in a list. The participants were very negative about the possibility that the TV would display a TV show they had not requested.

In comparison with the participants in the *correction* and the *clarification audio* conditions, all the participants in the *clarification visual* condition were very satisfied with the error recovery strategy. They especially emphasized the flexibility of either repeating the request, or saying the number of the required item in the presented list.

The participants in the *clarification audio and visual* condition were also satisfied with the numbered *n*-best list, but they found the spoken feedback unnecessary and disturbing while they were reading the content of the list. According to them, the audio feedback did not provide any benefits. One suggestion from some of them was that the system could speak out all the items in the list instead of just the first one.

Some participants expressed that it was unnatural to speak to a TV, while others found that it was fun and an effective way to interact with the TV. The majority of the participants in all the conditions were interested in using a remote control in a combination with speech input. Overall, speech was considered an effective input technique that gave them flexibility to express their request.

5 Discussion

The results from the evaluation show that a visible *n*-best list gives the shortest total task completion time, the *clarification visual* as well as *clarification audio and visual* were the fastest strategies. This is because a visible *n*-best list gives the user a shortcut to items further down the list. However, note that the advantage of displaying an *n*-best list depends on the correct answer's probability distribution among the candidates in the list. The user benefits when the right answer is *not* the first candidate on the list. Therefore, the confidence levels' threshold of displaying the *n*-best list has to be set not too high to miss any potential error or too low to have too many correct answers on the first position of the *n*-best list. The threshold and the correct answer's probability distribution among the candidates on the *n*-best list have to be analyzed and tested before implementing a clarification strategy with a visual *n*-best list.

The possible advantage of partially redundant audio feedback in the *clarification audio and visual* condition did not effect the task completion time. The presence of the audio output was experienced as unnecessary and disturbing. Negative attitude towards spoken feedback have been expressed elsewhere, for example Ibrahim & Johansson (2002). Thus, *clarification visual* seems to be a better design alternative for error recovery in an interactive TV-setting, than with a combined audio and visual output. Also, the participants in the conditions without a visual *n*-best list indicated that they preferred a visible list over audio output. Despite of the negative comments on the audio feedback, participants in the clarification audio and visual condition rated this strategy equally satisfying as *clarification visual*. One of the main objections to the audio output in the clarification audio and visual condition was that it did not add any additional benefits. This indicates that it was more the content of the audio feedback than the format the participants disliked. A change in content could therefore also change the attitude towards the audio feedback. Furthermore, spoken output can have other effects, for example promoting the viewers to use speech as input. This is an interesting issue to investigate, especially if the EPG becomes more similar to a dialogue system.

The *correction* strategy performed between the clarification strategies that displayed a visual *n*-best list and the clarification audio strategy. The relative fast performance in this condition can be explained by that the participants in this condition experienced fewer error recoveries than the participants in the other conditions. When the confidence level was low but the correct answer was still the first one the *n*-best list, the participants did not go through an error recovery process.

Although the participants in the *correction* strategy experienced less error recovery processes than those in the *clarification* strategies, they preferred another error resolution strategy when error did occur. They also changed their way of speaking to the system when they had to repeat their utterances, and this will make the new utterances more difficult to recognize. These findings confirm earlier research done on correction strategies, such as Frankish et al. (1992).

Slowest was the *clarification audio* strategy. The slow performance of this strategy confirmed that traversing a list by audio takes more time than necessary. The slow performance in this condition also affected the participants' attitude towards it. The participants were least satisfied with this strategy. They felt it was tedious, and they also felt that it was their fault when the system made errors. Feelings of guilt do not fit well with the purpose of entertainment associated with watching TV. It is therefore essential that the interaction afford positive rather than negative feelings. The speech recognition engine will make errors, and the recovery of these errors have to be as pleasant as possible. This was not the primary focus in this study, but is something that designer of speech interaction with interactive TV has to take in consideration.

The type of the performed task is an important aspect when it comes to which error recovery strategy that is most efficient, as demonstrated by Suhm et al. (1999). The results in this paper are likely to be relevant for other entertainment information search systems where information can be presented visually and the possibility of using multimodal input is limited. Error recovery strategies in these kinds of systems are more efficient if the strategies use visual feedback. The purpose of this study was to comparatively evaluate and contrast four different error recovery strategies. In practice, it is possible to use a combination. For example, it could be beneficial to use the *clarification* strategies when the probability of the correct answer is relatively evenly distributed in the *n*-best list. When the probability is concentrated on the first item, a correction strategy can be used.

6 Conclusion

We have explored the error recovery problem in an interactive TV setting by evaluating four EPG

prototypes. Our evaluation of four error recovery strategies showed that people preferred clarification over correction for error correction. Between visual and audio clarification strategies, visual *n*-best list provided better performance because it offered the flexibility to say the number of the item or to repeat the input while audio clarification was found tedious by the participants. When audio and visual clarification was combined, no further performance advantage was found and participants felt the additional audio information distracting.

Acknowledgements

This work is a result from a project supported by Center for Industrial Information Technology (Ceniit), Nokia Home Communications, and Santa Anna IT Research Institute AB. We would like to thank our wizard Hampus Jönsson and the participants.

References

Ainsworth, W. A. and Pratt, S. R. (1992). Feedback strategies for error correction in speech recognition systems. *International Journal Man-Machine Studies*, 36, 6 (June), 833–842.

Brennan, S. E., and Hulteen, E. A. (1995). Interaction and feedback in a spoken language system: a theoretical framework. *Knowledge-Based Systems*, 8(2–3): 143–151.

Frankish, C., Jones, D., and Hapeshi, K. (1992). Decline in accuracy of automatic speech recognition as a function of time on task: fatigue or voice drift? *International Journal of Man-Machine Studies*, 36, pp. 797–816.

Halverson, C., Horn, D. B., Karat, C., and Karat, J. (1999). The Beauty of Errors: Patterns of Error Correction in Desktop Speech Systems. In *Proceedings of INTERACT'99*. Edinburgh, Scotland, August 30 – September 3, pp. 133–140.

Ibrahim, A., and Johansson, P. (2002). Multimodal dialogue systems for interactive TV applications. In: *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, Pittsburgh, USA, pp. 117–122.

Ibrahim, A., Lundberg, J., and Johansson, J. (2001). Speech enhanced remote control for media terminal. In *Proceeding of Eurospeech 2001*, pp. 2685–2688.

Karat, C.-M., Halverson, C., Horn, D., and Karat, J. (1999). Patterns of entry and correction in large vocabulary contentious speech recognition systems. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI'99*, pp. 568–575.

Mankoff, J., Hudson, S. E., and Abowd, G. D. (2000) OOPS: A Toolkit Supporting Mediation Techniques for Resolving Ambiguity in Recognition-Based Interfaces. *Computers and Graphics* 24(6), 819–834.

Martin, T. B., and Welch, J. R. (1980). Practical speech recognizers and some performance effectiveness parameters. In W. A. Lea (Ed.) *Trends in Speech Recognition*. Upper Saddle River, NJ: Prentice Hall Press.

McGee, D. R., Cohen, P. R., Oviatt, S. (1998). Confirmation in Multimodal Systems. In *Proceedings of the International Joint Conference of the Association for Computational Linguistics and the International Committee on Computational Linguistics (COLING-ACL '98)*. : Montreal, Quebec: Association for Computational Linguistics Press, pp. 823–829.

Murray, A. C., Frankish, C. R., and Jones, D. M. (1993). Data-entry by voice: Facilitating correction of misrecognitions. In C. Baber and J. M. Noyes, (Eds.) *Interactive Speech Technology: Human Factors Issues in the Application of Speech Input/Output to Computers*, Bristol, PA: Taylor and Francis, pp. 137–144.

Oviatt, S. (1999). Mutal disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI'99*, pp. 576–583.

Portolan, N., Nael, M., Renoullin, J.L., and Naudin, S. (1999). Will we speak to our TV remote control in the future? In: *Proceedings of the 17th International Symposium on Human Factors in Telecommunication*, HFT'99, Copenhagen, Denmark.

Suhm, B., Myers, B., and Waibel, A. (1999). Model-based and Empirical Evaluation of Multimodal Interactive Error Correction. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI'99*, pp. 584–591.

Suhm, B., Myers, B., and Waibel, A. (2001). Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction* 8, 1, pp. 60–98.