Invited Paper
IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, Co, 11-15 December 1999

APPLICATIONS OF SPEECH TECHNOLOGY: DESIGNING FOR USABILITY

Lou Boves[1,2], *Els den Os*[2]

[1]**Nijmegen University**
Erasmusplein 1
6525 HT Nijmegen
The Netherlands

[2]**KPN Research**
Sint Paulusstraat 4
2264 XZ Leidschendam
The Netherlands

ABSTRACT

The number of operational applications of speech technology has been growing steadily, but slowly, over the last couple of years. This paper gives an overview of the issues that are involved in bringing application of speech technology in general and automatic speech recognition in particular, to the field. This is done by analyzing the advantages and limitations of speech relative to alternative communication modes. In addition, the performance of speech recognition under real world operating conditions is identified as an important limiting factor.

In the second part of the paper we discuss a number of operational services and field tests that illustrate the usability issues that were identified.

1. INTRODUCTION

Over the last couple of years the number of large scale applications of speech technology in general and of automatic speech recognition (ASR) in particular has grown slowly but steadily. The more cautious comments say that speech technology is getting mature; others seem to believe that ASR is essentially a 'solved problem'. Yet, there are fewer operational speech driven services than what one would expect from the persistent assertion that speech is man's most natural communication mode. Could it be that 'most natural' is not always equal to 'most effective for the purpose at hand'? Certainly, few scientists and engineers share the view that ASR is a solved problem. In actual fact, we tend to see that ASR performance in the field rarely reaches the levels that are needed for unobtrusive interaction with a speech driven service. In this paper we intend to determine 'critical success factors' for speech driven services. In doing so, we will analyze the arguments that are conventionally put forward in favor of speech interfaces, to obtain a better understanding of the conditions under which they really hold. In addition, we will analyze a number of performance issues that may have an impact on the take-up of voice driven services.

The use of speech technology in all kinds of applications is often motivated with reference to the claim that "speech is the most natural way of communication between humans". The obvious face value of this claim makes it unusual to question its real meaning. Speech undoubtedly precedes written forms of language in the evolution. However, many of the oldest written documents show that speech, despite its naturalness, may not be the most effective and efficient mode of communication for all purposes. The Sumerians did their bookkeeping in cuneiform script, most probably because it was much longer lasting than speech and human memory. The texts painted in Egyptian tombs were most probably meant to be much more permanent than spoken accounts of the deeds of the deceased. The very start of what we presently know as 'Linguistics', i.e., the works of Panini for the Sanskrit language dating back to around 600 BC, was motivated by the finding that the grammar and pronunciation of the Vedas suffered from undesirable variation, and had to be made more permanent by putting the exact syntax and pronunciation 'on paper'. Also, few people have not come across the saying that a picture is worth a thousand words. Therefore, while speech may well be a highly natural way of communication, it is necessary to ask why it is not always the most effective and useful one. All the examples of ancient written representations of information point in the same direction: in all civilized communities there is information that needs some degree of permanence in order to fulfill its purpose. Speech may be easy to produce and to understand, but by its very nature it lacks permanence. Written and graphical representations have another feature that speech has not: although writing and drawing may take a long time to complete, the results can often be processed in a glance. Moreover, the permanence and simultaneous presence of printed information items provide random access to the viewer, in contrast to the sequential access in audio material.

There is a large number of research papers that may indeed convey the suggestion that ASR performance is well beyond the critical level for real applications. Although performance is good enough for some applications, testified by their success in the field, recognition errors are still a major concern in a large number of services that could profit from ASR. In our view, the performance problem is very real, and extremely difficult to solve. It is very likely that the effectiveness of human-human communication is to a large extent due to the fact that the process can rely on highly active intelligent processing of the signals by all partners. There is a growing body of evidence that human perception is in a very essential way an active process: we do not just passively decode the information in a signal; rather, we actively recreate the message that is buried in the signal. For all but the simplest cases, the mapping of a 'message' onto a physical signal (whether acoustic or visual) is many-to-many: a single message may take many physical forms, and at the same time several different messages can take the form of essentially equivalent signals. From high school Optics we know how to compute the two-dimensional image of a two or three dimensional object projected on a screen (perhaps the retina) by a set of lenses. Only recently, however, we have begun to realize that inverting the projection is a mathematically ill-posed problem. A somewhat similar problem in speech has been known for a very long time, and frequently cited without making its full impact. Already in 1956 Peterson and Barney [1] published data for the formant frequencies of vowels, which convincingly show that we are confronted with a many-to-many mapping. The ease with which we handle the myriad of mathematically insoluble inversion problems in daily life is difficult to explain, except by assuming that we use highly

intelligent processing (which very probably involves extensive use of world knowledge and common sense reasoning) to recreate the messages. At the same time we use linguistic and social intelligence to adapt our own contributions to a conversation. In our opinion the role of some kind of 'intelligence' in speech driven applications has been underestimated in the past. Therefore, we will have several occasions to show that problems in speech driven services are actually caused by the lack of intelligence of the system. Taken to its extreme, this suggests that only services for which ASR in the form of meaning-free pattern recognition suffices are suitable candidates for short term success. In our opinion, this may very well be true.

In the past the DARPA community has set highly influential standards for objective evaluation of speech technology based on experiments with pre-recorded corpora. This evaluation strategy has been decisively instrumental for the fast progress in the performance of ASR and NLP modules. However, if it comes to real applications, with paying customers, there is no simple relation between error rates and usability. The limitations of corpus-based evaluation have been clear from the very beginning, and today there is a strong trend towards the evaluation of the usability of truly interactive speech driven services. In this paper we will present our experience with the development and the evaluation of speech services that have at least reached the stage of large scale field tests. In doing so we will refer to several applications developed by KPN and field tests that we have been involved in ourselves, but also to several field tests and applications developed in other companies.

In section 2 we will introduce a number of factors that may affect the usability of a speech driven application. The presentation will be fairly general. In section 3 we will investigate a number of applications (both operational services and field tests) to illustrate the abstract issues introduced in section 2 by means of practical examples. We will focus on applications over the telephone, not because these are by necessity the most important ones, but simply because we know them best.

In section 4 we will formulate conclusions and suggestions for further research.

2. ISSUES IN USABILITY

2.1 Introduction

Designing for usability is very similar to User Centered Design (UCD). UCD attempts to involve the end user in the design process from the very inception of a product. Questions that are asked in an early stage include "Why would customers want to consider the new product?", "Do customers understand and need the functions?", "What are the essential performance factors that will make the product easy and pleasant to use?".

Although many of the concepts underlying UCD may contribute substantially to the design of successful products, this paper is not the place to introduce the approach. Rather, we will limit ourselves to a number of issues that have surfaced in UCD approaches to the design and testing of speech driven services. Most of the issues introduced below have to do with the question whether speech is indeed suitable for the application in which it has to function. Inevitably, this requires to draw up a gross classification of services and users. Other issues are more closely related to the problem of ASR performance.

In all cases questions about the 'best' ways to measure 'usability' will pop up. Designing experiments with interactive services in the laboratory is already very difficult. As soon as a service gets into field test or into production performance measurements only become more difficult and complicated. It is hardly ever possible to control decisive factors like the characteristics of the test subjects. At the same time it may be extremely difficult to decide whether a given dialog was successful from the user's point of view. These issues leave ample room for contradictory interpretation of the raw data.

Also, the question arises whether there are generic applications (like travel information services) that we know well enough to be able to skip the early stages of the design phase of new services in the same domain, or perhaps even in the same general functionality. We definitely have accumulated sizeable amounts of knowledge and experience that should allow us to have a head start [5]. However, two superficially similar applications may involve several factors that differ so much that generalization of findings in one context to the other are not warranted. In the remainder of this section we will sketch some of these factors. To complicate things further, the relevant factors are all but orthogonal.

2.2 Dimensions of usability

Usability is a very complex concept, especially if it must be applied to a wide range of services [6]. During the Eurospeech-97 Conference Elsnet organized a comparative evaluation of spoken dialogue systems, with widely different functionalities. Although the methodology of this enterprise is easy to criticize, it was conclusively shown that 'usability' has at least three dimensions [7]. Perhaps somewhat surprisingly, the dimension that explained most of the variance in the scores was the quality of speech output. This is in line with previous experience with user tests of a train timetable information system in several European countries: the output speech is what strikes naïve users as by far the most conspicuous feature of the service. The second dimension is the functionality of the system or service. If users do not understand how they might benefit from using a service, it is likely to be rated low on usability, irrespective of its technical qualities in terms of ASR performance and output speech quality. Failure of the prospective customers to understand the functions of the service caused a field trial of a Personal Communication Assistant to fail. But if people really need the service, they even put up with extremely lengthy interactions with IVR machines. The third dimension was labeled 'interaction'. It combines issues related to ASR performance and dialog management. The focus of this paper is on this dimension, because it is the most interesting one for the ASR community.

Experiments have shown that 'usability' of a service cannot be predicted from the technical quality of its components. This is to a large extent due to the interaction between objective performance measures and functionality. A simple service, for which small vocabulary isolated word ASR is sufficient, but that is not very useful for the test subjects, tends to be rated low on 'usability', even if the ASR in the service is almost flawless. At the same time we have seen very positive scores for a service that provided free access to Directory Assistance for visually impaired customers, despite performance problems with the speech technology used in the implementation (ASR and SV) [8]. Moreover, for modal customers it is impossible and irrelevant to

distinguish between the components of a system (ASR, NLP, Dialog Management, TTS, etc.).

2.3 Types of users

One factor that complicates the interpretation of usability measures on a service is the fact that the general public is all but a population that is sufficiently homogeneous to make conventional descriptive statistics and statistical tests meaningful. In almost all relevant respects there is no such thing as an 'average' or 'median' user. The concept of modal users may make sense, but almost invariably the distribution will be multi-modal. Therefore, even findings which are true for a 'modal' user must always be qualified by adding an indication of the specific mode.
One very important difference is that between professional users and the general public. Professionals are able and motivated to take the courses which are necessary to learn how to use a service, while services aimed at the general public must be usable without any prior instruction. Therefore, one should not generalize findings for a professional service to a consumer service, even if the two services are similar at the surface, like messaging services in telecommunication or dictation in hospitals or in the home.
The impact of speaker characteristics on the performance of ASR (and even more on Speaker Verification) has appeared from many corpus based studies. More often than not, a relatively large part of the errors are due to a relatively small proportion of the speakers. In real services there is a difference between those who learn to adapt to the requirements of a system (not necessarily limited to the ASR module; design decisions in the dialogue flow may make for even more of a learning experience) and those who do not learn. In this context it is worth mentioning that it may be dangerous to generalize from success rates in a specific service: it may well be that the very high ASR performance claimed for some operational services is due to a non-negligible extent to some self selection process. Customers for whom the service works, keep using it, while those for whom the interface did not result in a rewarding experience -and who happen to have alternative ways to get the service- simply do not call again. This effect may cause a severe bias in the data on 'service success rate', reported for some operational information systems.

2.4 Types of services

This is not the place to develop a full-fledged typology of speech driven services and applications. However, a coarse classification will help in understanding critical success factors.

2.4.1 Telecommunication services

In telecommunication services it is necessary to distinguish between services that are accessed very frequently (perhaps more often than once a day, like voice dialing) and services that are only used occasionally (e.g., travel information). For the first type of service one may hope that customers learn over time, for the second type this is hardly the case. In this context the question arises when we will reach the development stage where people have had so many experiences with sufficiently standard speech driven interfaces that it is safe to rely on some kind of learning and carry over. A comparison might be drawn here with IVR services. Today most callers are so much used to this type of interface that few are really at a complete loss when confronted with a decently designed IVR menu.
Today, most telecommunication services require some kind of speaker independent ASR. However, the number of cellular users, who effectively have a personal handset, that is always identified before it can even access a speech driven service, is growing very quickly. In principle, this opens the door to speaker adaptive ASR, provided that one is able to manage all the data related to user profiles in the network.

2.4.2 Dictation

For dictation applications one must distinguish between free dictation and the generation of fairly standard and structured documents. It seems that the most successful applications of ASR in 'dictation' are in the second class. A structured report in a Radiology lab can be created by means of a small number of inputs (even if for each 'command' there is a potentially large number of options or alternatives). Radiology report generation is a good example of an application that is used with high frequency by professionals who are very motivated to learn how to comply with the standards of the hospital they work in. Similar remarks apply to report generation in other Medical disciplines like Pathology, and in the offices of legal and insurance companies. Most, if not all, dictation applications rely on speaker adaptive ASR.

2.4.3 Data entry and Command & Control

Hands busy - eyes busy data entry and Command & Control were the first application domains of ASR. Small vocabulary, speaker adaptive ASR is successfully used in several large scale warehouse storage and picking applications, as well as parcel sorting/routing tasks. For speech technology vendors this type of applications has not become very profitable, however, because the number of copies of the software sold are very small. Typically, even a large warehouse will not employ more than a couple of hundred workers, each of whom have a personal wearable computer and an ASR system adapted to their voice.
Command & Control (C&C) applications keep popping up. Their most recent incarnations comprise voice dialing (with ASR either in the handset or in the switch, in which case the application is often classified under telecommunication) and voice operated remote controls.

2.5 Alternative modes

Human beings are inherently multi-modal. Under normal conditions we use ears, eyes, nose and touch in parallel. It has only been after the advent of telecommunication that we have generalized audio-only communication to arbitrary contexts. Before the age of telecommunication mono-modal audio communication was limited to conversations in the dark, and -extremely importantly for survival- the processing of sounds coming from the back. It is interesting to speculate whether the telephone would have been successful if man had not learnt to carry on speech-only communication in ecologically more plausible situations.
Yet, human-human communication may very well obtain much of its efficiency and effectiveness from the availability and use of several modes in parallel. A lecture without the use of a blackboard or overhead sheets is almost unthinkable. Therefore, we must constantly be aware that it may not be a good idea to rely

too much on the exclusive use of speech in human-machine interaction. Moreover, ASR is certainly not the only input technology under development. For many simple selection tasks speech input has been superseded by the mouse. One of the things that has limited the role of ASR in parcel sorting/routing was the emergence of bar code printers and readers. Today, many information and help desk services are moving from call centers employing thousands of operators who act as a distant input device to computer programs to Internet applications where the user can interact directly with the same computer program. In other words, ASR is aiming at a moving target: a large part of the problems for which speech seems to provide the solution may very well be solved by other, new technologies, be it perhaps at the expense of a reorganization of the work flow or business process. The risk that ASR will be superseded in mobile applications by hand writing recognition is small, because hand writing suffers from the exact same problems as speech. However, the next generation of cellular telephones may very well come with touch screens and touch pens that are combined with cleverly designed procedures for menu presentation and selection, allowing applications based on form filling that will leave us speechless.

If customers have the choice, they will select the interaction modes which are most comfortable for them under the given condition. Sometimes these modes will include speech, sometimes they will not. This forces us to think about the goals that humans want to reach, and to carefully analyze the conditions under which ASR (or speech in general) can make a useful contribution. For instance, speech may be used for direct access, compared to lengthy paths through menus. Speech may prove the best mode for making selections from very long lists of items that are impractical to display and to scan, like the city names in a travel information service. However, it has appeared that both uses of speech as a shortcut may run into problems, for instance because the user may try to carry out an impossible operation or because he is trying to find a connection to a city that is not in the network.

2.6 Dialog management strategies

Despite its naturalness, speech is also error-prone, even in communication between humans. In human-human communication most of the recognition errors and production lapses are detected almost instantaneously. Perhaps speech is so effective in human communication because it is so easy -and socially accepted- to signal and repair errors. Of course, human intelligence plays a decisive role in making exceptions handling so effective. Many decodings with a high acoustic likelihood are ruled out from the very beginning, because they make no sense whatsoever. It is not unusual for comedians to fail to land their jokes, because too few persons in the audience 'hear' the alternative decoding of an otherwise bland statement. Just try the "wreck on nice beach" example with an audience that does not know about HMMs.

Machines, that do not have the knowledge of the world, nor the ability of common sense reasoning are relegated to the use of a limited lexicon and a fairly primitive language model to try and prevent misunderstandings. In human-machine communication this puts all of the burden of error detection and error handling on the human partner. This is not necessarily problematic. Computer text processing has wiped out electric typewriters long before on-line spelling checkers flagged questionable spellings immediately after the typing of a terminator. But even when spotting typos was left to the writer, computer text processing has made the correction of errors virtually effortless. This ease of error correction is probably the single most important enhancement of the text production process. Also, Command & Control applications can go a long way without 'intelligent' dialogs, as long as ASR is capable of recognizing the commands correctly (and the commands are unambiguous for the user).

It has been customary to design spoken dialog systems such that they seem to emulate an intelligent human. Despite the face value appeal of the belief that human-machine dialogs are best modeled after human-human dialogs, it is time for a reconsideration. In human-machine dialogues the detection of exceptions (errors committed by either speaker or machine, unexpected situations arising from unexpected answers, etc.) is one of the most difficult things to do well. If we are given the impression that the machine is conducting a dialogue, it is only natural to expect that it will display 'natural' dialogue behavior. In other words, naive users will expect that the machine takes its role in detecting potential errors and in correcting them if necessary.

There have been several attempts to develop a comprehensive dialog theory, mainly from linguistics and philosophy. Most of these proposals are very idealistic, in the sense that they try to draw up models of the 'ideal' dialogue. The treatment of turn-taking is a case in point: almost all descriptions of the structure of a dialog suggest that half duplex communication is the norm, rather than the exception. This is reflected by the 'speak after the beep' strategy that is characteristic of most speech driven services. Although this strategy may work quite well for making menu selections, it fails miserably as soon as a caller thinks that he can conduct a true dialog. The most influential theory of communication in modeling human-machine dialogue, viz. Grice's Maxims, is a set of overly idealistic (and extremely vague) statements [2]. If at all, these maxims are only applicable if both partners in the dialog are goal directed and co-operative. How else could one adhere to both the requirement that one must give all relevant information and that one must not be too verbose at the same time? David Sadek [3] has proposed a different way to model dialogs, where the emphasis is on 'intelligence', rather than on co-operativity. It remains to be seen how effective this approach will be.

Even if human communication is extremely robust despite the inevitable errors, there is a threshold beyond which people run into problems. As the proportion of potential or true errors becomes too high, communication starts to become cumbersome and eventually it will break down. We do not know what this threshold is. Most likely, it is not even a single number, but rather a complicated function of the situation, the urgency to get some task completed and the a number of personality traits of the human participant that decides where one will give up.

2.7 Cost-effectiveness

Despite attempts of technology providers to offer re-usable building blocks, many speech driven services are still one-of-a-kind designs, that are intrinsically extremely expensive. This forces service providers who consider to enhance IVR services by means of ASR to be very cautious. The development of completely new services that rely on ASR is even more risky, because it combines two problems: customers must learn the new serv-

ice's functionality and they must cope with the inherently error-prone input device.
Customers are not willing to pay for services they do not need or understand. Service providers do not want to waste money and resources developing services that will hardly be used – if at all. Above, it has already been pointed out that subjects tend to give low ratings to services if they do not understand how these could help them achieving their private goals. One way to help customers to understand the functionality of new telephone services, is to build speech driven versions of applications that are already known from desktop PCs. Below we will show that this is easier said than done: speech and text/graphics behave very different from a human factors point of view.

2.8 Practical issues

In experiments with laboratory systems it is normal practice to collect detailed logs of all actions of the users and the system. These logs include the recording of all user utterances, the output of the ASR and NLP modules, the activity and decisions of the Dialog Manager, interactions with the application data base, and system prompts.
In operational services it is not possible to advise callers that their speech is being recorded by the system for the purpose of system development or evaluation. In the European Union privacy protection laws allow the unilateral recording of telephone calls, provided that the recordings will only be used by the party who was involved in the call, and under the restriction that the recordings will not be made public in any way. This allows service providers (and probably also the speech technology provider who builds an application for them) to record calls and use them for improvement and evaluation of the application. However, it is illegal to make the recordings available for research and development purposes outside the company. This severely limits the possibility of using invaluable data about the behavior of real users for the advancement of the field. Apart from the fact that evaluations of system components by means of pre-recorded speech is not very informative in the first place, this severely limits the possibility of using expensive data for the advancement of the field.
Logging of the system's actions may require a substantial part of the processing power that is available in the computer platform that implements a service. For this reason it may not be possible to record all interactions during actual operation of a service. Usually, this factor will not be prohibitive in laboratory experiments and field tests. This is the more important if the behaviour of the users in a real paid-for service may differ from the behaviour of test-subjects. This is even true if the test-subjects participated in a near-realistic field test.
Network integration problems often seem to be manageable during laboratory tests with small numbers of customers. However, the PC networks used in these experiments may not scale up to high volume deployment. Unstable platforms are probably the single best way to guarantee that customers will dislike a service and stop using it.

3. OPERATIONAL SERVICES

In this section we describe a number of operational services, in order to better understand why they are successful (or not). It is not our intention to give a comprehensive and up-to-date overview of all operational services and large scale field tests. Rather, we focus on a number of generic services, and attempt to relate their success to the criteria and issues outlined above.
We will only deal with telecommunication services, because we have no 'hard' data on the use and usability of desk top services and C&C applications in warehouses and the like. We want to make one exception, and briefly mention an automatic interpreting system that has been tested in a number of places [12]. A stripped-down version of Dragon Dictate is used as input for a system that plays spoken translations of a large number of commands and questions that are relevant for the treatment of refugees. This application is a good example of how far one can go with spoken commands, provided that intelligent users take care of all necessary exceptions detection and handling. By tightly structuring the setting and the communication protocol it is possible to avoid the need for complicated human-machine dialogs.

3.1 Voice Dialing

Voice dialing is a natural candidate for a killer application. Everybody understands what it means, and the idea that one can call a person just by saying her name has an obvious appeal. Moreover, the use of cell phones while driving is often mentioned as a very important concern in traffic safety. Especially the manual operations necessary to set up a call, and the visual distraction that it causes, is often cited as a major cause of car accidents.
Yet, so far voice dialing has never been successful, neither in the fixed nor in the cellular networks. In the remainder of this section we summarize what we see as the major causes for the surprising lack of success for voice dialing.
First, in the present implementations voice dialing is never completely hands free: in the fixed network one must lift the receiver and perhaps even dial an access number; in the cellular networks one must obtain access to the service by pressing a key or by dialing a short number. Moreover, off-hook access to the service in the fixed network appears to cause problems with dial-up modem connections, and fax transmission. In cellular networks storage of the personal vocabulary in a computer connected to a switch of the service provider causes enormous problems with roaming. Only in very exceptional cases will it be possible to access one's voice dialing application across network boundaries with a short number. These problems are difficult to solve, and difficult for the customer to accept.
If it is necessary to press a key to activate the service, the safety enhancement of voice dialing while driving is minor. Moreover, practically all handsets offer short code dialing, by which up to ten numbers can be dialed with a single key, and they come with displays that show the name connected to the short number. Selecting the proper key and checking the displayed name may not distract the driver any more than the need to monitor the auditory feedback of the voice dialing application or attempts to correct erroneously recognized commands. Thus, Voice Dialing must compete against alternative modes that may be at least as effective and efficient under many conditions.
In the situation where the functionality of voice dialing is most needed, viz. when driving, the acoustic background conditions are often worst. Substantial amounts of research effort are presently being devoted to ASR that is robust against noise. For the time being, however, recognition performance is undoubtedly one of the major impediments to the success of voice dialing.
Interviews with users of voice dialing (and those who tried it but stopped using the service) show that building and maintaining

the phone book is considered as very difficult. Even if the service as such is used several times a day, the phone book maintenance function is only rarely needed. Therefore, customers do not get the opportunity to get acquainted with this function. It has been proposed to integrate the voice dialing phone book with a phone book maintained on a PC or through the Internet, but none of these proposals has been successfully deployed on a large scale. Actually, this integration may only make for a partial solution of the problem. As long as sufficient recognition accuracy requires that the names are acoustically very different, one is faced with the problem of 'inventing' new names for old acquaintances. Although customers tend to think that they need 100+ entries in their personal phone book, most people call only a small number of persons or businesses frequently enough to easily remember their nicknames in the voice dialing service. Therefore, users run into problems when using Voice Dialing to place calls to numbers they only use occasionally, because they do not remember the exact nickname they have trained the system with. This will cause recognition errors, that are naturally blamed on the recognizer.

3.2 Financial Services

Information services, especially financial information (and perhaps transaction) services, have been known to be good candidates for speech driven implementations. Stock quotation information is a good example of a service where a single command is sufficient, but where the number of items from which to select is much too long to be able to offer access through a menu. Demonstrators of stock quotation services have been available for many years [9]. The service offered by Charles Schwab has been in operation for over two years, and it is considered as very successful, both from a commercial and an operational point of view. Speech recognition accuracy is claimed to be in the high nineties. Although we do not have access to actual data on the use and the users of that service, we surmise that the high recognition rates can, at least to some extent, be explained by the fact that a large proportion of the customers use the service frequently. Thus, they develop habits that fit with the requirements of the recognition system. Moreover, callers who start using the service and keep having recognition problems are bound to abandon it. This kind of self-selection process cannot but bias the recognition rates to the high end of the scale. Of course, there is nothing against this bias, as long as developers of other, new services are aware of this fact, and not assume that the same accuracy will carry over without the initial customer self-selection process.

Since the user population is not really known, it is not possible to say whether the callers to the speech driven Schwab service would have alternative ways of obtaining the information they are looking for, and if these alternatives exist, whether they are more expensive, faster, etc.

In the Netherlands a speech driven stock quotation service, limited to the 50-odd most frequently traded stocks on the Amsterdam Exchange, has been in successful operation for almost ten years. Access is not restricted to subscribers or customers of the bank which operates the service, but callers pay the normal long distance tariff. This service is called because users have no simple and fast alternatives for accessing stock prices, unless they are in a place where videotext service is available. Today, the information is also available through the Internet, but despite the rapid growth of Internet terminals one can never be sure to find an Internet terminal at any time in every place.

3.3 Information and reservation services

Train Timetable and Air Traffic information and reservation tasks have been very popular among the developers of dialog systems. This is probably because these services make for a good compromise between linguistic complexity on the one hand and conceptual simplicity on the other. In many cases these systems have been used to experiment with 'natural language dialogs', a term that is used by different people for quite different things.

In Europe a number of Train Timetable Information services are operational. The first one to be on line was the service of the Swiss Railways. The Dutch, Italian and Swedish Railways operate similar services. Except for the Italian system, that has been designed and built by CSELT, all others are based on the technology developed by Philips Speech Processing. Although superficially similar to the financial information services described above, there is one essential difference: on the average the callers who need timetable information call at most three or four times per year. Since callers will be connected to a live operator if one is available, even frequent users of the service will have difficulty in learning how to deal with the automatic systems. Moreover, it appears that a stable 20% of the callers who get connected to the automatic service hang up almost immediately, even if they have been given the explicit choice between waiting in queue for an operator or using the automatic service. The percentage of immediate hang-ups does not diminish as more customers have accessed it once or twice. If these aborted calls are included in an evaluation of the performance of these systems, dialog or service success rates cannot exceed a ceiling of 80%.

Despite the fact that many laboratory timetable information systems used natural language or mixed initiative dialogs, two of the systems (the one in Italy and the one in Sweden) that are in operation in Europe are based on a system driven dialog strategy. In the EC-funded project ARISE is was found that most customers are able to adhere to the structure imposed by a system driven question-answering protocol. At the same time it appeared that very few callers take advantage of the fact that they can give all relevant information in a single turn in the mixed initiative systems. It is not really possible to determine why callers are reluctant to use mixed initiative, but it is certain that there are at least two factors that play a role: if the system asks directed questions, many callers spontaneously think that a computer is not able to do more than to answer that question [10]; at the same time it appears that recognition performance suffers from the relatively open language model and the large vocabulary that are needed to process mixed initiative input [5]. In this context it is interesting that the system in Sweden, that went into operation only recently, uses the system driven interaction style.

The systems that are in operation were designed to give factual data to callers who were –at least tacitly-- supposed to know the exact functionality of the system. However, analysis of the dialogs recorded in the framework of the ARISE project has shown that a fair proportion of the callers to information services actually are looking for help in planning a trip. If, for instance, a journey takes much less time than the caller anticipated, she may want to enter a negotiation dialog episode, in order to obtain a more suitable travel advice. Present systems have only limited capacity for navigation and negotiation, and it appears to be very

difficult to convey these limitations to occasional users. This confirms the finding of [4] that naïve callers have difficulty to build an accurate mental picture of 'natural language dialog systems'.

One of the issues that was studied in the ARISE project was operator fallback. If a service can be accessed by the general public, it is almost certain that some proportion of the callers will not manage to accomplish their goals. If one intends to offer a very high quality of service, there is no way around some form of operator fallback. In the Dutch timetable information system callers are handed over to an operator if the dialog manager persistently fails to fill an additional slot in the query form. Also, the caller can press *9* to request forwarding to an operator. There are numerous system integration and Human Factors issues left open with respect to operator fallback. The navigation strategy implemented in the Dutch ARISE system, that is actually a speech copy of the screen display of the operator, is obviously suboptimal. In more general terms one may ask whether timetable information is suitable for a speech-only service. Presently, we are designing experiments to investigate possible ways of combining speech input with text displayed on a screen as output and as basis for navigation and negotiation.

It is difficult to determine the 'success' of the European train timetable services. On the one hand, a persistent 20% premature hang-up rate can be construed as an indication that the service is not very successful. But the Italian Railway company claim that they handle more than twice as many calls to their information service with the same number of operators on duty. Also, the system in the Netherlands is handling a stable number of 4,000+ calls per day. With these numbers one cannot but conclude that the service does satisfy a need.

Evaluating operational services is extremely difficult, and even that may be an understatement. The EURESCOM project MIVA is on the other extreme of the scale: it was designed as a laboratory experiment that allowed close control of a large number of experimental factors [11]. The MIVA application gave information on telephone services, aimed at foreigners who do not speak the language of the country they are in. The same services were offered in parallel in a number of languages. One of the most interesting results of the MIVA project is that it provides data on the cultural differences with respect to the appreciation of speech driven services and the impact of several 'objective' performance measures on that appreciation.

3.4 Personal Call Assistants

Since the first announcement of the Wildfire™ service (and probably even before that moment) the idea of bringing the friendly, intelligent operator who had been replaced by automatic switches back into the network has appealed to many people. It exerts the same kind of attraction as voice dialing. However, in actual reality this kind of service meets with problems which are strongly reminiscent of what we have seen in the case of voice dialing.
Experiments and market studies in the Netherlands have shown that only a small number of professionals from the ICT industry readily understand the functions offered by a comprehensive PCA. Even professionals find it hard to devise an optimal way for integrating the use of their home phone, business phone and mobile phone. If calls addressed to different phone numbers can end up in the PCA mail box, people find it difficult to navigate and hear the messages that are relevant at a given time and in a given situation. If calls to different numbers end up in different mailboxes, they find it difficult to locate mails, especially when callers have decided to use another number than the one that was most 'likely' for the person called. In short: the full fledged PCA is the type of service that is best reserved for the professional market, where customers can afford to receive intensive training. For the SOHO and consumer markets comprehensive PCA services will only become suitable if large proportions of the population have grown accustomed to forwarding calls between wireline and mobile terminals and voice mail and messaging systems.
If customers are uncertain about the functions of a service, and about the way in which these functions must be accessed, the ASR and NLP modules in the interface are easily overtaxed. This is the more so if the ASR anticipates a wide range of possible commands and queries. However, monitoring of the practical use customers make of a service like Wildfire shows that a small number of functions covers almost the complete use [13]. This may well motivate a simplification of the service functionality, with an attendant reduction of the size of the lexicon and the complexity of the grammar, the more so since these simplifications go hand in hand with an increase in recognition performance. This is another example of the maxim 'simple is beautiful'. The CallMinder™ service [14], a speech driven voice mail service, offered by BT, is a point in case. The functionality of the service is limited to just a simple voice mail, but it has over 2 million subscribers. More complex versions of the service require a new generation ASR/NLP and a new generation AI to provide access to domain knowledge and common sense reasoning that people expect from a human assistant. It is probably advisable to avoid appealing but misleading terms like 'assistant' until we have the technology that is needed to implement services which are reminiscent of a real human secretary.

3.5 Directory Assistance and Auto-attendants

One application that is in high demand –by the service providers, but not necessarily also by the customers- is automated Directory Assistance. The reason is clear: the service is expensive to offer through operators. However, automation poses extremely high demands on the capabilities of the ASR system, since it must be able to recognize the millions of different proper names that occur in most countries and languages. And perhaps more importantly: the ASR system must be able to understand that the likelihood that it has not correctly understood a name is too high to take the output for granted. Name recognition in automatic attendant applications is equally difficult, be it that the number of different names even in a big company is much smaller than in the public telephone directory.
For a automated DA application to be successful it is necessary that the dialog is tightly structured, so that a single information item is prompted per turn. It is not necessarily clear what is meant by 'single information item'; for instance, a persons first and surname can be considered as one or two items. There are essentially two different strategies that can be used in an automated DA service: the system can prompt for individual items, send the response to an ASR system, and revert to an operator if the recognition result has a confidence score below a preset

threshold. The alternative strategy acquires all items, does ASR and combines the output with a database search to find the most likely unique entry [15]. The latter approach seems to be most advantageous, but its superior performance may come at the cost of quite substantial additional processing. This is the more so because a substantial proportion of the DA queries refer to organizations or companies, for which a unique listing may be identified before all 'standard' information is acquired. We estimate that less than 30% of the DA queries handled by KPN Telecom refer to human individuals. Often, callers do not even know the exact address information. Callers asking for toll-free numbers may not even know the city where the company is located. When operator fallback is necessary after one or more items have been acquired with above-threshold confidence, the question arises whether it is useful to fill in the corresponding fields on the operator's screen. The answer is probably 'yes', but we plan to investigate this issue in experiments with real callers and real operators.

Many person names are ambiguous without additional information, mainly the home address. In some countries, like the Netherlands, the amount of ambiguity is increased because most persons are listed with their last name plus initials. It is questionable whether one should try to engage in a –potentially lengthy- dialog between the system and the caller to disambiguate the names.

Recently, operator based DA has seen increasing competition from services based on the Internet. For the service provider Internet access to DA information is most probably less expensive, the more so if it can be combined with clever advertising. The advent of screen phones with proper typing and browsing facilities is likely to reduce the demand for operator based DA services further. It remains to be seen whether Internet access for DA services can be implemented in the cellular networks. In any case, ASR is aiming at a moving target, as alternative modes for delivering the service become more widely available.

Attempts to use ASR in services like DA and Auto-attendants has shown that the acoustic decoding that we have presently available is far less effective than what humans can accomplish. Even in the case of simultaneous optimization of the recognition results for all items with the help of the underlying database the perplexity is extremely high, leaving the bulk of the work to acoustic decoding.

4. CONCLUSIONS

Automatic Speech Recognition has certainly outgrown its laboratory age. However, it is evident that it is not yet fully mature. Although the number of successful speech driven services is growing, it remains necessary to constantly ask ourselves whether ASR is the right type of user interface mode for a new service. The maxim that speech is man's most natural communication mode may not be generalized to mean that speech is a suitable medium for any service and any kind of information. On the contrary, applications that return many factual or numerical information items most probably are better handled by a system that is able to provide some type of graphical output.

As long as the recognition performance of ASR systems falls short of human performance by almost an order of magnitude, extra caution is needed in trying to deploy the technology. It is unlikely that major performance improvements for telephony services like stock quotations, travel advice or DA information can be obtained from better language models. Instead, research must focus on improved acoustic modeling.

Problems with recognition performance are aggravated by the failure of customers to understand the exact functionality of a service and the limitations in the grammar and vocabulary of the ASR system. This is part of the explanation why simple interfaces, essentially based on unambiguous commands, work best, despite the limitations they impose on the freedom of the customers to choose their own formulations. In that sense we are far removed for 'natural language' systems. In order for these systems to become a reality we must not only improve pure ASR and NLP performance; in addition, substantial progress in AI engineering is needed.

REFERENCES

[1] Peterson, G. and Barney, H.L. Control Methods used in a study of the vowels. Journal Acoustical Society of America, Vol. 24, 1952, pp 175-184.

[2] Bernsen, N. O., and Dybkjaer, L. A theory of speech in multimodal systems. *Proc. ESCA Workshop on Interactive Dialogue in Multi-modal Systems*, Kloster Irsee, 1999, pp. 105-108.

[3] Sadek, D. Design Considerations on Dialogue Systems: From Theory to Technology – The Case of ARTIMIS. *Proc. ESCA Workshop on Interactive Dialogue in Multi-modal Systems*, Kloster Irsee, 1999, pp. 173-187.

[4] Thomson, D.L. and Wisowaty, J. User Confusion in Natural Language Services. *Proc. ESCA Workshop on Interactive Dialogue in Multi-modal Systems*, Kloster Irsee, 1999, pp. 189-196.

[5] Sturm, J., den Os, E., and Boves, L. Issues in Spoken Dialogue Systems: Experiences with the Dutch ARISE System. *Proc. ESCA Workshop on Interactive Dialogue in Multi-modal Systems*, Kloster Irsee, 1999, pp. 1-4.

[6] Gray, W.D. and Salzman, M.C. Damaged Merchandise? A Review of Experiments that Compare Usability Evaluation Methods. *Human-Computer Interaction*, Vol. 13, 1998, pp. 203-261.

[7] den Os, E.A. and Bloothooft, G. Evaluating various spoken dialogue systems with a single questionnaire: Analysis of the ELSNET Olympics. *Proc. First Intern. Conf. On Language Resources and Evaluation*, Granada, 28030 May 1998, Pp. 51-54

[8] den Os, E.A., Jongebloed, H., Stijsiger, A. and Boves, L. Speaker Verification as a User-Frendly Access for the Visually Impaired. *Proc. Eurospeech-99*, pp. 1263-1266.

[9] Lennig, M., Sharp, D., Kenny, P, Gupta, V. and Precoda, K. Flexible vocabulary recognition of speech. *Proc. ICSLP-92*, pp. 93-96.

[10] den Os, E.A., Boves, L. Lamel, L. and Baggia, P. Overview of the ARISE Project. *Proc. Eurospeech-99*, pp. 1527-1530.

[11] Johnston, D. An Overview of the EURESCOM MIVA Project. *Proc. ESCA-NATO Workshop* Multi-lingual Interoperability in Speech Technology, Leusden, Sept. 13-14, 1999, pp. 31- 37.

[12] Hunt, M., Bamberg, P., Tucker, J. and Anderson, S. A Military Operational Automatic Interpreting System. *Proc. ESCA-NATO Workshop* Multi-lingual Inter-operability in Speech Technology, Leusden, Sept. 13-14, 1999, pp. 75-78.

[13] Jeanreneaud, P., Cockroft, G. and VanderHeijden, A. A multimodal, multilingual telephone application: the Wildfire electronic assistant. *Proc. Eurospeech-99*, pp. 1259-1262.

[14] Beacham, K. and Barrington, S. CallMinder™: the development of BT's new telephone answering service. *British Telecommunications Engineering,* Vol. 15, pt. 2, July 1996.

[15] Gupta, V., Robillard, S and Pelletier, C. Automation of Locality Recognition in ADAS Plus. *Proc. IVTTA-98*, pp. 1-4.