THE APPLICATION OF SPEECH I/O TECHNOLOGY TO INTERACTIVE TELECOMMUNICATION SERVICES

Roland Seidl

Integrated Communication Services Section
Telecom Research Laboratories

ABSTRACT - This paper discusses the evolution and current status of speech I/O technology and its application to telecommunication services. Because speech recognition is the limiting factor in the design of these services, the impact of its constraints on the user/service interface is discussed in the context of some generic applications. The key factors to a potentially successful service are outlined.

INTRODUCTION

It has been said many times that speech is the most natural mode of communications for humans. The convergence of telecommunications and computing technologies has now made a large number of interactive communication services possible. However, there is a gap between the types of services that speech I/O technology can currently deliver compared with the expectations of service providers (in particular those who require commercial success). To a large extent this perceptual gap has come about due to a number of factors, the most significant of which is the constrained nature of the service/user interface, which belies the naturalness of speech communication mentioned above.

Is current speech I/O technology merely technology looking for an application? To a large extent the answer is yes, at least from a service provider's perspective. To understand how this situation has arisen it is instructive to consider the evolution of telecommunication services which employ speech technology.

Evolution Of Voice Interactive Telecommunication Services

The earliest services were the dial-up information services, of which the speaking clock is the oldest and most well known example. With the advent and improvements in digital speech coding technology and techniques and also digital storage technology the range of such services has greatly expanded.

The push button telephone which provided communication via tones, allowed the first interactive voice response services. Users were prompted to select or enter information via the telephone buttons. To allow universal access to these services (since there isn't 100% penetration of touch-tone phones) a natural progression was to allow speech input by means of speaker independent speech recognition. Hence the digits (to replace the push button input) was the first application vocabulary developed. Unfortunately isolated (or connected) digit speaker independent recognition technology is not as reliable as push button input and is rather more expensive, and therefore is not seen as a practical alternative with currently available technology.

This technology development path has also led us to interactive telecommunication services which are menu driven. While menu systems appear to be the "natural" interface for computer systems, especially where a selection process is incorporated, for the end user - employing voice input rather than a keyboard or push button input - a natural language input is most appropriate. Hence there are conflicting requirements (and a consequent need for compromise) for the user/service interface. How these conflicts are resolved are determined by the capabilities of the available technology.

KEY FACTORS IN DESIGNING POTENTIALLY SUCCESSFUL SERVICES

The design of potentially successful speech applications requires consideration of the following:

- Technology capabilities and limitations (current and anticipated technology)
- Application objectives (what is the service required to achieve?)

- How will the service achieve its objectives (what are the technology requirements and the relevant user/service interface)

The phrase "potentially successful" is used since success is generally measured in commercial or strategic terms. The decision concerning which services or applications should be implemented are generally not made by technologists, rather technologists provide information regarding points one and three above. To complete the circle, service planners should provide feedback to the technologists regarding the types of services with potential for commercial success.

In many cases the circle is incomplete because of the perceived inadequacies of speech technology (in particular, recognition) by service planners who are therefore unwilling to consider these types of services. The result is that rather than application considerations driving the development of technology the reverse is true. The question being asked most often is "Given current speech technology (e.g. connected digit recognition) what services can be implemented?" rather than "The following services are of commercial significance, what technologies are required?"

Let us now consider the above factors is more detail.

TECHNOLOGY STATUS

In many cases alternative choices are available, and the correct choice will be guided by the "look" and "feel" of the final application design. The importance of the user/service interface cannot be underestimated. This interface will determine how well the application is perceived by the users and will largely determine its acceptability.

Speech Output Systems

Speech output systems can be used to provide a wide variety of information, by means of voice output, ranging from specific information from a database (in response to a user request), to guidance prompts which evoke specific responses from the user of a voice interactive application. Essentially both of these are instances of information retrieval, where the voice prompts are the information in the latter case. The differentiating parameter between the two cases is the size of the required vocabulary. In the first case the information is potentially unlimited, and hence so is the required vocabulary, whereas in the second case the vocabulary requirements are limited and constrained by the specific application. This leads to a basic dichotomy of techniques for providing speech output, which will be referred to as "compiled" synthesis and "rule based" synthesis.

The term *compiled* synthesis arises from the fact that lexical elements, i.e. words, phrases, or even sentences can be pre-recorded in some fashion and later joined together by an appropriate technique to compile the required output messages. In this case a variety of digital speech coding techniques may be applied for the storage (and subsequent replay) of vocabulary elements. Coding techniques vary in perceived quality but it is generally true that the higher the bit-rate of the encoded speech the better its perceived quality. The higher bit-rate coding algorithms are the least complex, with complexity tending to increase quite significantly with decreasing bit-rate.

Techniques for the concatenation of vocabulary (message) elements are dependent upon the specific speech coding algorithm being used. The quality of spoken messages can be seriously degraded if the concatenation process is not performed carefully. In particular, the naturalness of the spoken message is susceptible to contextual effects (e.g. the pronunciation of some words vary in different contexts) and these effects must be catered for in the recording of the original vocabulary.

Rule-based text-to-speech synthesis must be used in some cases to avoid excessively large data storage requirements. Note, however, that text-to-speech synthesis can also be used in the limited vocabulary case but might not be adopted for reasons of substantially poorer speech output quality and higher cost. Text-to-speech synthesis does not approach the quality of digitally encoded speech but is more flexible in the extent of the messages which can be produced and the speed at which they can be updated. Their quality can only be described as "synthetic" and must be considered from the viewpoints of intelligibility and naturalness which are inherent in the rule bases of these synthesis systems.

Parameters which affect the choice of speech output systems include:

- Required speech output quality and naturalness
- Vocabulary (message) storage requirements
- Response time requirements (i.e. the period between request for information and its reception, or between a user action and subsequent voice prompt).
- Vocabulary (message) preparation and maintenance requirements.

As well as the technical issues discussed above, human factors issues must also be considered. These latter issues affect the acceptance by users of the technology. Issues to be considered include:

- *Message formulation and duration*: The way in which messages are composed impacts upon their comprehension and intended effect. The length of a message must not exceed the user's short term memory capacity. Rules for good composition are not necessarily the same as for printed messages.
- *Response time*: The delay experienced in system response should be within user expectations.
- *Speech output quality*: This must be acceptable to the users and suited to the application
- *Operational characteristic*: The manner in which applications relying on speech output operate in practical situations can have a significant influence upon the manner in which it is perceived. For example, prompts should vary with the user's experience with a particular application and users should be able to over-ride prompts when desired.

Speech Input Systems

Current speech input (recognition) systems vary widely in capabilities and performance. The parameters which describe such systems include:

- *Speaker capabilities* (speaker dependent or independent)
- *Vocabulary size* (small, medium, or large)
- *Input mode* (isolated words or connected words)
- *Input bandwidth* (high quality microphone or telephone system)

All speech recognition systems must be trained. Once a system has been trained with a specific vocabulary the use of such a system is constrained to applications which can make use of that vocabulary. The training data (speech samples) is crucial to the performance of the recognition system. In particular, for speaker independent systems, speech samples from an appropriately large number of speakers representing the target user population (using the appropriate speech medium, i.e. telephone or high-quality microphone) must be used. For speaker dependent systems, training is usually built into the application installation procedure, and may take several hours to complete (depending upon the size of the vocabulary). It is important to recognize that the training phase of any speech recognition system is generally expensive and time consuming, and hence the application design (in particular vocabulary requirements) needs to be "correct" beforehand.

As the speech input constraints are relaxed, i.e going from speaker dependent to speaker independent recognition, from isolated word to connected word, from small vocabulary to large vocabulary or from wide bandwidth to telephone bandwidth, the speech recognition becomes more difficult. Relaxing more than one of the above constraints further compounds that difficulty, and hence the complexity of the speech recognition system. Speech input systems are error prone. The more difficult the recognition task, the greater the likelihood of error. It is therefore important, when designing applications which use speech recognition, to provide appropriate verification checks (e.g. "The number you require is 123456. Is this correct?") and means whereby the application can recover in cases of incorrect recognition or failure to recognize the speech input. Too many recognition errors, which may result in a user being asked for the same information a number of times can have an adverse effect upon the acceptability of speech recognition systems. In such a case, a fallback strategy should be incorporated, e.g. reversion to a human operator.

An appropriate strategy in the design of applications using speech recognition is to simplify the application as much as possible. One method of achieving this is to use "context sensitive" vocabularies. With such an approach, it is assumed that at any point within the application only a limited number of responses are valid and consequently the recognizer need only discriminate words from a subset of the total application vocabulary. Another approach is to make use of artificial intelligence techniques. Once again, the application is aware of valid response which can be derived from the limitations imposed by the specific application. The knowledge of the "domain of discourse" and the syntax and semantics of the language can be used to improve the overall performance of speech recognition systems.

The ultimate aim of speech input systems is to provide unconstrained natural language speech input to applications. While the technology is still far short of this goal, techniques such as key-word spotting, allied with knowledge based systems, go some way to providing a semblance of unconstrained input. Key-word spotting is the ability to recognize words from the prescribed application vocabulary within a stream of continuous speech without understanding (i.e. rejecting) the remaining words in a spoken input. Rapid developments in such technologies incorporating other natural language understanding techniques and higher level knowledge (e.g. syntactic and semantic) hold the promise for unrestrained input for future applications.

INTERACTIVE VOICE APPLICATIONS - GENERIC CATEGORIES

To consider application objectives, as mentioned previously, requires input of a commercial nature which is beyond the scope of this paper. However, consideration of generic categories of services can provide some insight into technology requirements.

Much of the literature tends to categorize speech I/O applications based on arbitrary commercial sectors, e.g. banking, education, telecommunications, aids for disabled, etc. Alternatively applications may be segmented on a technology basis. e.g. speaker independent or speaker dependent. However, none of these classifications are generic in nature. All the above commercial and technologically based applications can be categorised by being either:

- *Information retrieval applications*
- *Control applications*

In information retrieval applications the goal of the application is to retrieve information relevant to some perceived need (e.g. What's my bank account balance?) or to convey some information (such as "I wish to order X gallons of wine Y") to a specific application. Note the information retrieval in the latter case is by the application from the user. Control applications on the other hand elicit responses from a user (from a list of allowed responses) to perform some task (e.g. mail sorting, telephone dialling, etc.).

For telecommunications, the information retrieval applications fall mostly in the area of what are currently operator assisted services (e.g. directory enquiries, changed number service enquiries, pay by phone, etc.). The control applications will emerge with developments in network technology such as the "intelligent network" to facilitate, for example, the selection of network based services such as call diversion, conference calls, etc.

Essentially, the applications differ only in the detailed specification of the speech I/O systems. These details range from the content of the prompt messages, to the vocabulary required by the speech input system. The design of the user interface is critical to the acceptance of the application. The choice of speech output and input systems should be determined from the application requirements and will impact on the acceptability of the application. The simulation of applications to ensure an appropriate application design is a very good idea.

USER/SERVICE INTERFACE DESIGN ISSUES

The design of user/service interfaces is highly application dependent. As well, limitation in technology, in particular speech recognition, will require compromises to be made. Many of the human factors issues relevant to the design of user/service interfaces have already been raised in the section discussing the status of speech technology and so will not be repeated here. One critical factor in the implementation of interactive speech applications is the design of such interfaces to optimise the level of performance of the speech recognizer. A consequence of poor recognition performance will be reflected in poor service performance (i.e. in the achievement of the service objectives) which results in a loss of user confidence and hence in poor acceptance of such a service.

Many parameters contribute to the performance characteristics of speech recognizers, only some of which the interface designer may control. These parameters include operational factors such as the required service response time, human factors such as the sex of the speaker, language factors such as the active vocabulary, algorithmic factors such as the recognition strategy used, channel and environmental factors such as circuit noise and microphone type and performance factors such as the type of feedback and error correction available.

An Example - Speech Input Payphone

Consider the following elementary application. It is required to implement a payphone which uses speech input, rather than a dial or push-buttons, to determine the number to be dialled. Obviously the objective of this application is to determine from the user the required telephone number and to subsequently dial it. Since the user will be paying for the call connection, assuming a connection can be established, there should be some confirmation from the user that the correct number has been input before the call establishment proceeds.

At first glance this appears to be a trivial example. All that would seem to be needed is a simple dialogue which firstly asks the user to input a telephone number and then, after receiving the speech input, repeats the number (as interpreted by the recognizer) back to the user and finally asks whether or not this is the number to be dialled. Depending upon whether a yes or no reply was received, the number would then be dialled or a new number solicited. The recognizer vocabulary requirements would therefore be the digits, "yes" and "no". Since the payphone would be available to the general public, speaker independent recognition is required and ideally should be capable of recognizing connected digits. This application should be easily implemented by current technology, shouldn't it?

Unfortunately as a former politician was once noted to have said, "Life wasn't meant to be easy!" The speech recognizer needs to be capable of accepting telephone bandwidth speech to allow usage of current telephone handsets. Although manufacturers of speech recognition technology have promised connected digit recognition for some time it is generally not available, and so the first compromise that must be made is the use of isolated digit recognition rather than connected digit. This particular compromise significantly affects the interface design. In particular a method is required to force the user to segregate the input of the individual digits.

Once a particular recognition system has been selected, which meets the vocabulary and telephone bandwidth requirements of this application, it is necessary to assess the performance of the recognizer. One particular, commercially available, system which meets the above requirements and also provides additional telephony interfaces and appropriate support software was tested using a database of 97 *Australian* speakers (48 male, 49 female). The resultant confusion matrix is presented in table I. As well as recognizing words within its own vocabulary, the recognizer also provided an "Unrecognized" response where it could not provide a significant match. Other useful features include the ability to output a "beep" to prompt users for speech input, and other error outputs which indicate that a user had spoken before the beep or that the user was speaking too softly or alternatively the signal to noise ratio was too low.

The poor recognition performance is due to the fact that this particular recognition system was trained for American speakers, and this is exemplified by the very bad recognition results for "Oh" which is a typical Australian diphthong. The use of "Oh" in this case, should certainly be discouraged. Given the relatively bad performance of this system an alternative is being sought.

Assuming an alternative could not be found we are faced with the problems of providing appropriate user feedback to attempt to optimise the recognition performance. To segregate the user input, one could use the "beep" as a prompt mechanism. However, given such a high error rate it is highly likely that there would be at least one error in a telephone digit string. A better feedback mechanism therefore would be to repeat the digit that was recognized and allow the user to correct his input (if it were misrecognized) on a per digit basis. Thus a control word such as "Cancel" would need to be added to indicate a correction is required.

		Recognizer Output															Percentage		
		1	2	3	4	5	6	7	8	9	0	O	H	C	T	U	Corr	Incor	Unrec
	1	74	0	6	0	3	0	0	0	9	0	0	0	0	0	5	76.29	18.56	5.15
	2	0	87	0	0	0	0	0	0	7	0	0	0	0	0	3	89.69	7.22	3.09
	3	0	1	91	0	0	0	1	0	2	0	1	0	0	0	1	93.81	5.15	1.03
	4	0	3	0	64	2	0	1	0	0	2	5	0	0	0	20	65.98	13.40	20.62
	5	2	0	2	4	69	0	0	0	12	0	0	0	0	0	8	71.13	20.62	8.25
I	6	0	0	0	0	0	78	0	1	0	0	0	0	1	0	17	80.41	2.06	17.53
N	7	0	0	0	0	0	10	67	0	1	2	0	0	3	0	14	69.07	16.49	14.43
P	8	0	0	6	0	1	17	2	42	12	0	1	0	2	0	14	43.30	42.27	14.43
U	9	1	0	7	3	2	1	0	1	77	0	0	0	0	0	5	79.38	15.46	5.15
T	0	0	1	0	0	0	0	3	0	10	78	0	0	0	0	5	80.41	14.43	5.15
	Oh	0	6	0	0	18	0	3	1	25	0	31	0	0	0	13	31.96	54.64	13.40
	Help	0	0	0	0	0	1	0	0	0	1	0	87	3	0	5	89.69	5.15	5.15
	Cancel	0	0	0	0	0	0	0	0	1	0	0	0	86	0	10	88.66	1.03	10.31
	sTop	0	0	0	0	0	0	2	0	0	1	0	4	3	68	19	70.10	10.31	19.59
																OVERALL	73.56	16.20	10.24

Table I. Speech Recognizer Confusion Matrix

Another question that arises is how will the system determine that the telephone input is completed. It is not appropriate to merely count the number of digits that have been input, since telephone numbers are not all of the same length. Two alternatives are possible. The first is to use a timeout mechanism where the system assumes that input is complete if no speech is input for a predetermined time interval. The second alternative is to use another control word such as "End" or "Stop". The first alternative can lead to errors if the time interval is set too short due to a premature termination of input, or to an unacceptable delay if set too long. On the other hand, the use of control words including "Cancel" and "Stop" require a modest amount of user training. This could be incorporated into an opening greeting message, a "how to use card" or even provided via an on-line help ("Help" - yet another control word).

Finally, although not required by this example, there would need to be a recovery mechanism if the recognizer cannot perform correctly. (Obviously if this were the case for this example a user would be inclined to hang up in disgust and go to a standard payphone.) Generally the recovery mechanism of last resort is reversion to a human operator.

CONCLUSION

This paper has raised the issue that with current speech technology there is a mismatch between technological capabilities and the expectations of service providers. More significantly applications are being driven by technology rather than vice versa. As well, there is a mismatch between the required and natural modes of communications between services and their users, imposed by the constraints of the technology. In this environment, the design of the user/service interface is critical to the acceptibility of such services. By means of a simple example it was pointed out that, in general, such interfaces are not as simple as first conceived.

Future developments in technology and natural language understanding techniques which incorporate higher levels of knowledge (e.g. syntactic and semantic information relevant to a given application) will allow easier design of user interfaces by reduced restrictions on speech input.

ACKNOWLEDGEMENT

The permission of the Executive General Manager of Telecom Australia Research Laboratories to publish this paper is acknowledged.