# VOICE RESPONSE TECHNIQUES FOR TELECOMMUNICATIONS APPLICATIONS

R. A. Seidl

Telecom Australia Research Laboratories

ABSTRACT - A wide range of services can be implemented by making use of voice response techniques. This paper provides an overview of the alternative technologies applicable to such services and the parameters affecting their performance are discussed. Some issues which do not relate to the technology itself, but rather to its use are also presented.

## INTRODUCTION

A wide range of enhanced services can be implemented by making use of voice response techniques, which are finding increasing application in the provision of spoken messages by computer (processor) control. Essentially there are two broad categories of application: firstly, the provision of information (in aural form) from an information database in response to an information seeking request; and secondly, in the provision of a "user friendly" interface between service users and the service itself, by providing appropriate spoken prompts to facilitate its use. There are several alternative techniques which may be applied to these instances, however the particular application has a significant bearing on the technology likely to be the most successful. The choice of an inappropriate technique can lead to user resistance and poor acceptance of the technology.

This paper provides an overview of the techniques applicable to voice response systems (VRS). The parameters which affect the performance, both from a user's viewpoint and from a provider's viewpoint, will be discussed. Some particular points which do not relate to the technology itself, but more specifically to the use of the technology, will also be examined.

## SPEECH OUTPUT TECHNOLOGY ALTERNATIVES

Services which use speech output can be classified according to the amount of information which must be presented in spoken form. For the user guidance or feedback application, the vocabulary is limited and specific to the particular service application, whereas to present information from a database the spoken vocabulary is potentially quite large. Hence a dichotomy of application possibilities for speech output technologies exists, and this dichotomy is determined by the extent of the required vocabulary.

The fundamental techniques which apply to this dichotomy will be termed "compiled" synthesis in the limited vocabulary case, and "rule-based" synthesis in the unlimited vocabulary case. The term "compiled" arises from the fact that lexical elements can be pre-recorded in some fashion and later concatenated to compile the required output message. In this case, a variety of digital speech coding techniques may be applied for the storage of vocabulary elements. The large vocabulary case must employ text-to-speech (TTS) synthesis techniques. Text-to-speech synthesis can also be applied in the limited vocabulary case but might not be adopted for reasons of substantially poorer speech output quality and higher cost.

## THE SPEECH MESSAGE PROCESSING SYSTEM

The basic elements of a speech message processing system are depicted in figure 1. Its primary function is to receive message requests (from an as yet undefined sources) and output the corresponding voice messages. It must also provide facilities to maintain, store and update the corresponding vocabulary elements. The various components and the associated speech processing requirements are determined by the speech output technology adopted in the message output system for the voice response application (i.e. TTS or "compiled" synthesis). The elements can be segregated into the following functions:

VOCABULARY PREPARATION, EDITING and UPDATE
VOCABULARY STORAGE
MESSAGE COMPOSITION
MESSAGE OUTPUT

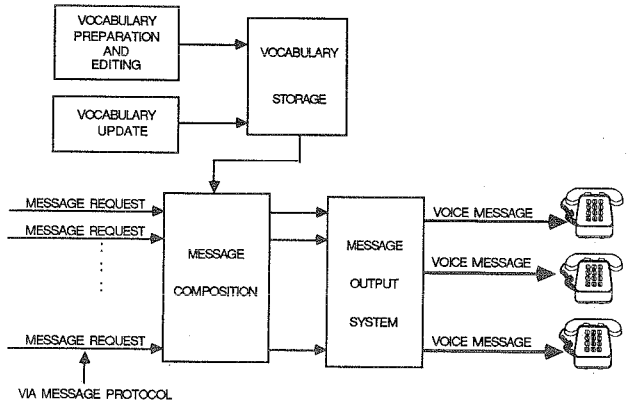


Figure 1. Elements of a speech message processing system.

### Vocabulary Preparation, Editing and Update

For a system employing TTS synthesis output, these functions are relatively simple requiring only a simple text editor. Systems using "compiled" synthesis, where vocabulary elements consist of segments of encoded speech, must employ a variety of techniques which vary in sophistication and complexity depending upon the particular coding algorithm.

All "compiled" synthesis voice response systems (i.e. those employing encoded speech) must have a facility for encoding the speech, or for producing parameters derived from the speech for the lower bit-rate algorithms. The processes for deriving the parameters can be extremely complex and computationally intensive. Once the speech, or its derived parameters have been stored, there must exist processes for editing. This includes the ability to mark the beginning and end of vocabulary elements and in some cases (such as with LPC encoded speech) to hand-tailor the parameters to produce a better synthesis. The segmentation process is extremely critical because if it is not correctly done, then the quality of

the resultant "compiled" speech output can be degraded. The editing processes are generally performed interactively using a trained operator since automated processes are susceptible to errors due to the vagaries of human speech. To circumvent these problems the original speech recording must be performed under strict supervision and in a controlled environment.

Updating a vocabulary set also raises some problems which are not technical in nature. If encoded speech forms the basis of the VRS then it is necessary that the original speaker is available for subsequent vocabulary recordings, and that the speaker be in the same or similar state of health. Any variation to this results in noticeable differences in the characteristics of the output messages.

Vocabulary Storage

The vocabulary storage requirements are determined by the particular application of the voice response system (large or small vocabulary) and in the case of "compiled" speech (small vocabulary) the type of coding technique adopted. For high bit-rate encoded speech, requiring large amounts of storage (even for moderate vocabularies), bulk magnetic media will be required. In this case the access time for large storage discs will determine the effectiveness of the VRS as gauged by the response time between the initial message request and the initiation of speech output. For reliable system operation and for speed of access to vocabulary items the vocabulary storage and the vocabulary items must be replicated.

Message Composition

The message composition function is responsible for accepting requests for spoken messages and extracting the appropriate data from the vocabulary store. Essentially it must be capable of handling a multiplicity of incoming requests and maintain directories of the vocabulary storage as well as handling the physical data retrieval.

Message Output

The message output element of the VRS must be able to handle a multiplicity of simultaneous outgoing messages. The message output system contains speech coding or synthesis devices which can translate the data streams delivered by the message composition module into speech.

PLACING VOICE RESPONSE INTO A TELECOMMUNICATION SERVICES ENVIRONMENT

To utilise the voice response facilities provided by the speech message processing system described above requires a communications processor to provide access to the system as well as an applications processor to control that access for specific applications. A voice response service in a telecommunications environment is depicted in figure 2. The functions of the communications and applications processors are described below.

The Communications Processor

This processor provides a variety of communications interfaces between the voice response service and its users. It's functions encompass:

- Telephone line control
- DTMF tone decoding (alternatively, a Speech Recognition System)
- Communications protocols for remote Hosts or other telecommunications service nodes

DTMF tone decoding and/or speech recognition are provided to enable telephone users to reply to voice prompts which have been sent from the speech message processing system. The provision of specific communications protocols allows the same user generated information to be transferred to the communications processor from telephone users who may be connected to a remote host or some other telecommunications service node which requires voice response. These inputs are all decoded by the communications processor and their information content passed to the applications processor for interpretation. A real-time communications protocol is required to pass voice message requests from the applications processor (via the communications processor) to the speech message processing system.

The communications processor can also be used by the application processor to download application dependent information from a remote host (e.g. such details as customer account numbers and corresponding validation information (PINs) for home banking, order entry, etc., and time critical information such as inventory levels to allow real-time confirmation of orders without interrogation of the remote host).

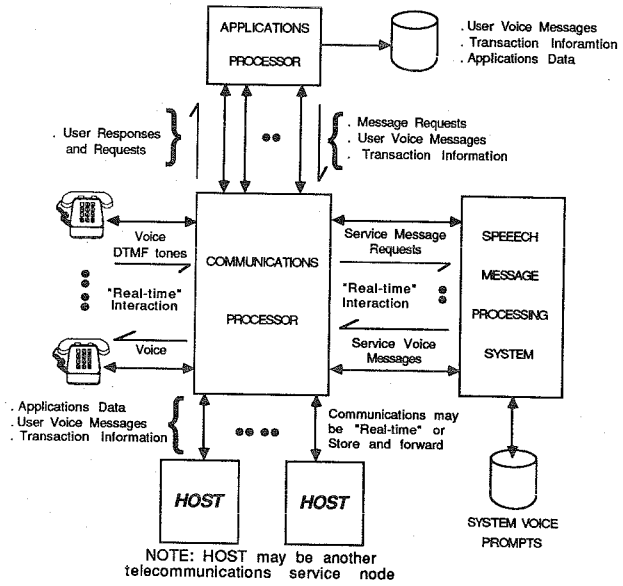


Figure 2. Voice response services generic functions.

Communications protocols to enable the inter-communication of remote hosts and other telecommunications services with the voice respons service must meet the response requirements of the service users (i.e. the delay between an action performed by the user, which anticipates some response from the service, and the output of the spoken message must be short enough to be perceived as being "acceptable").

The Applications Processor

The functions performed by the applications processor include the interpretation of user input as presented via the communications processor and determining subsequent actions to fulfill the requirements of the voice response service application. These actions include: the determination of the next voice response to be sent to the telephone user (dialog control); the storage of information input by users (transaction data collection), either for validation purposes or subsequent transmission to some remote host (e.g. as in an order entry application); and the transmission of voice message requests to the speech message processing system.

To perform the above functions the application processor must maintain application specific information including menu structures for the interactive dialog, valid response information, and time critical information as previously mentioned.

NON-TECHNICAL ISSUES AFFECTING VOICE RESPONSE SYTEMS

The major parameters for the VRS technology alternatives have been indicated above. In particular the vocabulary requirements are largely dictated by the application within which the VRS will be used, and this in turn influences the type of speech synthesis technology adopted ("compiled" or text-to-speech). In the case of limited vocabulary applications, where "compiled" synthesis is appropriate, a choice of speech coding techniques will be possible and particular attention must be paid to the required speech quality output of the VRS.

As well as the technical compromises, imposed by such constraints as storage capacity, which affect the VRS performance, other factors which affect the performance of the VRS in conjunction with telecommunication services and service users must be considered. The VRS must be able to interact with a variety of external telecommunication services which require the voice response capability to present information or prompts in spoken form by means of an appropriate protocol which must be determined. This protocol would forward message output requests to the VRS in an appropriate format, to which the message processing system would respond by outputting the corresponding spoken message. This spoken message needs to be delivered within, what the service user considers to be, a satisfactory (response) time. Satisfactory response times are subjectively interpreted and are influenced by, among other factors, the user's perception of the service's operation and the perceived complexity of the task being undertaken. For example, if the task is considered simple then a short response time is necessary, whereas for complex tasks (as estimated by the service user) a longer response time will be deemed satisfactory. The need to effect satisfactory response times will influence the choice of VRS technology.

The manner in which messages are formulated will also affect the acceptance and effectiveness of VRS technology. For messages that request a response from the user, then the high information content words should not be at the immediate beginning of the message. All too often listeners are not paying close attention at the start of a message. For messages which direct users to perform some action to achieve an end result it is necessary to formulate such messages so that the action phrase precedes the consequence. Services which employ a VRS to provide a "voice menu" to interactively guide users through service procedures must, in addition to message formulation, consider additional constraints and operational aspects. Such

messages will need to be kept short because of the limited retention capability of the human short term memory. For experienced users, a mechanism whereby the user can circumvent the voice prompt must be included so that it is not necessary to hear the entire prompt for every service access. (An alternative shortened prompt message set should also be considered for such users). For "voice menus" which have a number of menu levels, it has been suggested that a different voice be used at each level to enable the user to keep track of the position in the menu hierarchy. Too many voices (and as a corollary too many menu levels) will still cause confusion.

In summary, not only technical issues must be resolved but also that a variety of operational (including "human factors") considerations must be taken into account in the use of VRS techniques for telecommunication service applications. The resolution of all these issues is greatly influenced by the type of application in which the voice response system is to be implemented.

## CONCLUSIONS

The factors which will influence the choice of speech response technologies are summarised below.

Vocabulary requirements lead to a basic dichotomy of VRS techniques:

For limited vocabulary systems where "compiled" speech synthesis is applicable and bit-rate (for the stored speech) is the prime parameter, and there are tradeoffs between cost and quality, and

Unlimited vocabulary systems which must use text-to-speech synthesis, where speech quality is "synthetic" and hence degree of naturalness and level of intelligibility are of prime concern.

Parameters which influence the choice of appropriate technology include:

- Speech output quality
- Storage requirements for the VRS vocabulary
- VRS response time
- Vocabulary preparation and maintenance
- Vocabulary size

Factors related to the implementation of VRS systems and which may influence acceptance of the technology include:

- Message formulation
- Response time
- Speech output quality
- Operational characteristics (e.g. prompt over-ride for experienced users in voice menu type applications)

The final choice for all the above parameters will be a compromise and will largely be influenced by the type of application that is to be provided.

## ACKNOWLEDGEMENT

The permission of the Director Research, Telecom Australia, to present this paper is hereby acknowledged.