A PERSPECTIVE ON TELECOMMUNICATION SERVICES WITH RELEVANCE TO SPEECH PROCESSING

R. A. Seidl

Telecom Australia Research Laboratories

ABSTRACT - Speech processing covers many areas and may be applied not only to facilitate communication between individuals via some communications infrastructure but also to provide "user friendly" access to, and control of, communication services. This paper concentrates on the role of speech processing in the end-to-end communication between individuals. The notion of transparency is introduced to provide a basis for the classification of telecommunication services, and the role of speech processing is discussed against this perspective.

INTRODUCTION

Speech signal processing covers many areas including speech coding and transmission, speech enhancement and noise reduction, speech analysis and reconstruction and speech recognition and synthesis. These techniques will find application in both telecommunication networks and customer premises equipment and telecommunication services in general. The current trend in telecommunication services evolution is one of integration (in a variety of senses) and increasing freedom in the means whereby such services can be implemented. The consideration of telecommunication services based on information types (such as voice, text, data, image, etc.) or on technology groupings (such as mobile, cellular, satellite, wideband, etc.) will become less important as the trend towards service integration becomes established.

The application of speech processing technology and techniques to future telecommunication services requires not only an appreciation of the trends in speech processing activities but also a recognition of the trends in services. This paper provides a classification framework from which viewpoint communication services and related speech processing techniques can be considered. This framework is based on the notion of transparency introduced by Kato and Takemura (1983), where services are viewed from the dimensions of time, space and information.

Speech processing techniques may be applied not only to facilitate communications between individuals via some communications infrastructure but also to provide "user friendly" access to and control of communication services via voice prompts from the service and spoken commands from the service user. This paper concentrates on the end-to-end communications between individuals. The use of speech I/O in telecommunications services and in particular voice response is the subject of a companion paper in these proceedings (Seidl, 1986).

TRANSPARENCY

The notion of transparency refers to the "property of transmitting information without distortion so that the idea behind the originator's communication can be recognized as it is intended to be" (Kato and Takemura, 1983). The capability to meet transparency needs in the three

dimensions of time, space and information provides a basis for the classification of services.

Time transparency enables callers to place messages into a communications network when the called party is busy or unavailable. As well, messages may be received at a time convenient to the intended recipient. Space transparency enables callers to communicate with called parties regardless of location. Further, communication need not only be between individuals but may be among groups of individuals at multiple locations. Finally, information transparency allows a caller to send information to any party regardless of the type of presentation facilities (terminal devices) the recipient may have, or the form that information may have, be it text, voice or image. Figure 1 presents an overview of these three dimensions of transparency.

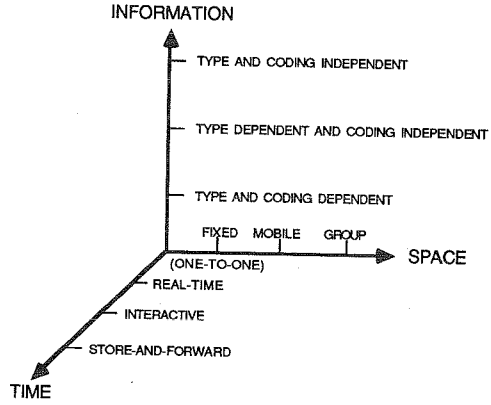

Figure 1. Three dimensions of telecommunications services transparency

Communication services are, in general, used for the sake of conveying information of various types between individuals or between an information source and some individual. Whatever form the communication takes, the unit of information flow will be referred to as a "message".

Time Transparency

The time dimension is segmented according to the amount of delay incurred between the sending of a message and its receipt, or between the sending of a message requesting information (or some action to be performed) and the receipt of that information (or the performance of that action). The time axis has been segmented into three regions referred to as real-time, interactive and store-and-forward.

A real-time service transfers messages with delays not exceeding around one half second. Note that this is not a rigid specification since the allowable limits for a real-time service depend largely upon the specific service being used and user preferences for acceptable delays. For interactive services, which are typified by enquiry/response services and

require a certain amount of time for processing or retrieving information, delays up to tens of seconds may occur. Finally, for store-and-forward services a deliberate delay is introduced between the submission of a message and its subsequent delivery to the intended recipient. This delay may be under the control of either the message originator or the message recipient and some intermediate message storage is required.

Space Transparency

Traditionally, communication services have been largely considered on the basis of communication between two individuals at fixed locations. The space transparency axis extends beyond this simple form of communication to include mobile communications, where individuals may inter-communicate regardless of their location, and group communications, where simultaneous communications between all members of a group is possible.

Information Transparency

This dimension is by far the most difficult to categorise. One method of classification may be ascribed according to the level of dependence of the service on the information type used to convey the "messages" and the manner in which messages are encoded. (A specific example of what is meant by information type and encoding is, for example, a facsimile service where the "message" is of type "image" and the image information is encoded by a standardised algorithm.) In general, for current telecommunication services, specific information type and encoding relate to specific terminal devices (such as voice terminals, facsimile machines, VDUs, etc.).

In the simplest case telecommunication services are information type and coding dependent, that is, voice terminals may only communicate with voice terminals, facsimile with facsimile, and so on, provided that the information is encoded in such a way that it can be decoded by the communicating terminals.

The second category on the information transparency dimension is one which enables communication between terminal devices which are capable of displaying or outputting the same, or perceptually similar, type of information (e.g. text, voice, image) but where the information may have been encoded in a different format by the originating communication terminal than that acceptable to the receiving terminal device. The coding independence is achieved by the use of appropriate code (and protocol) conversion facilities wihin the communications network.

The third category in this dimension is complete information type and encoding independence. To enable this type of communication requires conversion not only of encoded information between communicating devices, but also the conversion between information types (e.g. text to speech).

SPEECH PROCESSING AND TRANSPARENCY

This section describes the areas of speech processing which enable the communication of voice information in the three dimensions of transparency outlined above. It should be noted that the communication may not be entirely voice, but that the communication may require conversion from speech to some other type or vice versa to enable delivery of the message. It is also possible that the voice information could be accompanied by other information types (e.g. text) as part of a "compound document".

Time transparency

Real-time commmunications has provided great impetus to work in the area of speech encoding at reduced bit-rates in order to conserve communication channel bandwidth. The parameters of concern are the quality of the received decoded speech after the encoded speech has been passed through a communications channel, and the robustness of the coding algorithms in the case of noisy channels (especially mobile radio and satellite). Two international standards for encoding telephone bandwidth speech currently exist, 64 kbit/s PCM (CCITT, 1984a) and 32 kbit/s ADPCM (CCITT, 1984b). Standardisation activity is proceeding for 16 kbit/s encoded speech and for wideband (7kHz) encoded speech at 64 kbits/s.

The emergence of new communication systems, in particular networks employing packet switching raises new problems for the transmission of encoded speech in real-time. The speech packetisation process introduces a fixed delay between the time the speech samples are encoded and transmitted, and the transmission process itself introduces a variable delay, which depends on the level of traffic on the network. Delay and delay variability have a significant effect upon the acceptability of the received speech.

Speech enhancement techniques such as echo cancellation (especially for long distance communication circuits or where other speech processing introduces a significant delay), and noise reduction (e.g. removal of background noise for loud-speaking telephony) are other areas which contribute to the improvement in speech communication quality. The employment of digital speech interpolation (which essentially removes the "silences" in a conversation and makes the communication channel available for other "active" conversations) can provide more efficient utilisation of communication channels. However, the interpolation process can remove parts of the conversation and can therefore adversely affect the conversation in progress. The speech activity/"silence" detection process therefore is a critical aspect of digital speech interpolation. Note also that this process introduces a transmission delay and its attendant drawbacks.

Interactive services are typified by enquiry/response services. Speech processing techniques which provide spoken prompts to, and/or speech input from, service users, are relevant to this class of service. Of particular importance are text-to-speech synthesis for voice response systems requiring potentially infinite vocabularies and speech coding processes, and their attendant editing and concatenation processes which enable the recording, editing, and playback of speech messages, for limited vocabulary applications. This class of service is discussed in more detail in the companion paper in this proceedings (Seidl, 1986). A key parameter in this class of service is the delay between an action on the users part (e.g. pressing a DTMF button on a telephone keypad, or entering a spoken command) and the response from the service.

Store-and-forward (or alternatively, messaging) services are an emerging class of telecommunication services. A voice store-and-forward service as well as containing a deliberate delay between the recording of a message and its subsequent delivery (hence store-and-forward) must invoke a variety of speech processing functions. In particular the speech message recording and playback are in real-time, and as well there is an interactive component in the interface between the user and the service. The quality of the recorded message and subsequent playback is important, and therefore the amount of storage to provide such services becomes an important factor.

The trend in messaging services is towards the integration of information types into what is referred to as compound documents. Standardisation of compound document architectures is being led by the International Standards Organisation (ISO) TC97/SC18, and a recent review can be found in Horak (1985). Services which support compound document architectures should provide capabilities for the generation, output, and editing of such documents as well as their transmission. If editing is provided for the voice component of such a document the questions of, what level of editing should be provided and how should such editing facilities be implemented, naturally arise. One contention (Poggia, 1985) is that, in the context of compound documents, "voice is a very informal medium and is typically used to convey small amounts of information". If this contention is correct (and since it is based on a limited experimental basis this is questionable) then a limited editing capability only is required (e.g. erase and re-record), as well as limited storage for the speech component of such a compound document. Standards for the transmission of such documents are being developed both by ISO and are incorporated to some extent within the CCITT X.400 series of recommendations on "Message Handling".

Space Transparency

Many aspects in this dimension have been previously covered. In particular, fixed communications covers the range of services and their related speech processing aspects described in the time dimension.

Mobile communications necessarily implies a hostile environment, both from the presence (at times) of significant background noise as well as a noisy communications channel. In this case the robustness of the speech encoding algorithm and its quality in the context of the communications environment are important parameters. Speech enhancement techniques including echo cancellation and noise cancellation can be used to improve the performance of the communication system. Since such communication channels are relatively expensive it is important to produce the highest quality speech at the lowest possible bit-rate.

Group communications (or conferencing) requires a range of speech processing techniques. For example, for audio conferencing where each participant uses a separate voice terminal, a technique for the combination of all the digital speech signals into a composite signal and then redistributing that signal to all participants, is required. Where an audio conference is between two groups of people at separate sites, then special terminals might be used. These terminals could have either separate headphones for individuals or a loudspeaker for the speech output, and separate or a single microphone. Where a terminal has a single microphone in conjunction with a loudspeaker, then acoustic echo must be controlled. For conferencing employing both audio and video, the trend in video coding which has produced low bit-rate coders (384 kbits/s) has given rise to a significant acoustic echo control problem due to the processing delay in the video coding process and the necessity to synchronise the audio and video channels. Wideband encoded speech can improve the perceived speech quality for services using loudspeaker speech output.

Information Transparency

Information and coding dependent services describe the majority of current telecommunication services (e.g. telephones may only communicate with telephones, facsimile machines may only communicate with facsimile machines which have the same information encoding algorithm, etc). Type dependent,

but coding independent communication services are facilitated by either or both protocol and code conversion. Of particular importance is the conversion between various speech encoding algorithms in the digital domain without an intervening analog conversion to limit the accumulation of quantisation distortion.

Services which are both type and encoding independent require not only code and possibly protocol conversions but also content conversions between various information types. In this context conversion between speech and text (speech recognition and speech understanding) as well as from text and image to speech would be required. Obviously some conversions are more practicable than others. What processes would be involved in the conversion of image information to speech are difficult to imagine. To maintain the "idea behind the communication" for some of these conversion processes would require a level beyond simple speech processing (i.e. artificial intelligence and so-called expert systems techniques). An example in this domain is the translation from one spoken language to another.

CONCLUSION

Advances in technology will provide an increasing freedom in the implementation of telecommunication services. The trend in such services is towards an integration, especially in the types of information transported. The aim of this paper has been to provide a framework within which telecommunications services of ever increasing complexity can be classified. The notion of transparency provides a useful perspective on such services and the role of speech processing can be identified in this context. It will become increasingly more important to consider speech, not in isolation, but in conjunction with other types of information (e.g. in conjunction with text or video, etc.).

ACKNOWLEDGEMENT

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

REFERENCES

CCITT (1984a), Recommendation G.711, "Pulse Code Modulation (PCM) of Voice Frequencies", Red Book III.3, 85-93.

CCITT (1984b), Recommendation G.721, "32 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)", Red Book III.3, 125-160.

HORAK, W. (1985), "Office Document Architecture and Office Document Interchange Formats: Current Status of International Standardisation", Computer 18, 50-60.

KATO, T., & TAKEMURA, T. (1983), "A Concept of Future Telecommunication Networks: Transparency and ISDN Implementation", Hitachi Review 32, 141-146.

POGGIA, A., et al. (1985), "CCCUS: A Computer Based Multimedia Information System", Computer 18, 92-103.

SEIDL, R.A. (1986), "Voice Response Techniques for Telecommunications Applications", Proceedings of 1st Aust. Conf. on Speech Science and Technology.