

1

Introduction

1.1 Introduction

One of the most fascinating characteristics of humans is their capability to communicate ideas by means of speech. This capability is undoubtedly one of the facts that has allowed the development of our society. Man has been always attracted by the possibility of creating machines capable of producing and recognizing speech, imitating ourselves. An automatic speech recognition (ASR) system can be defined as a mechanism capable of decoding the signal produced in the vocal and nasal tracts of a human speaker into the sequence of linguistic units contained in the message that the speaker wants to communicate.

The final goal of ASR is the man-machine communication. This natural way of interaction has found many applications because of the fast development of different hardware and software technologies. The most relevant are access to information systems, aid to handicapped, automatic translation or oral system control.

The present book is focused on a wide class of applications that involve access through speech to remote information systems or services. Among other applications, we could mention entertainment information, flight reservation, or help in street navigation. This type of applications has been clearly boosted by the fast development of the digital networks (cellular and Internet) during the last 10 years. These systems involve a client-server architecture in which the server contains the information, and the client picks the speech, which is transmitted to the server in a suitable form. This speech-based interface offers a natural interaction and can be particularly useful in the case of very small user interfaces. Also, it can be part of a multimodal and multidevice service.

The place at which the speech is processed (at the client, at the server or at both) and the way this processing is performed define the system architecture. The most direct solution is known as embedded speech recognition. In this case, all speech processing (including recognition) is carried out at the local device. The request is sent to the remote information server, which gives back the corresponding response. The obvious problem of this approach is that the task of installing, maintaining and upgrading the speech recognition system falls on the user. Moreover, if the recognition system has a certain degree of complexity (large vocabulary, flexibility, etc.), it can be an arduous task to embed it in a device like a PDA or a cellular phone. However, work in this direction is

being developed in order to provide efficient and robust embedded ASR (Deligne *et al.*, 2002; Haeb-Umbach, 1997; Varga *et al.*, 2002), since this is an attractive solution for some applications, such as voice dialing or control operation in portable devices or hands-free operation in cars.

A very flexible and powerful alternative is what we will call *remote speech recognition* (RSR). In this case, a local device (telephone, mobile phone, PDA, a speech analysis and encoding program running in a computer, etc.) sends the speech signal or parameters through a transmission channel to a remote server that includes the speech recognition engine. This approach has several advantages, such as the use of a simpler client (which involves lower costs for the user), language portability, centralized server upgrading and maintenance, or the use of more sophisticated recognition techniques, which can make it more attractive. Figure 1.1 shows a general diagram of a remote information system using the RSR approach. After recognition, the server provides and transmits the required information to the client. The information returned could be text, data or even a vocal response either generated by a text-to-speech system or previously recorded. The server may include other functions such as a call control block or a VoiceXML browser, which manages the dialog between the user and the service. Figure 1.2 shows a block diagram of a possible system architecture including these features. The (information) contents may be even out of the voice platform, and accessed by HTTP. An example of how RSR can be integrated in a multimodal service can be found in D. Pearce *et al.* (2005).

There have been several initiatives to promote and study RSR. First, we can mention the COST action 249, entitled “continuous speech recognition over the telephone” developed during the period 1994–2000 (Martens, 2000). The goal of this project was to gather the achievements of different research groups for the definition of a voice-activated information service. The scientific objectives covered different issues of speech recognition (specially those related to multilinguality) and dialog systems. This action had its extension in the COST project 278, called *spoken language interaction in telecommunication*. The most productive activity has been developed by the European Telecommunications Standards Institute (ETSI)- Speech, Transmission Planning and Quality of Service

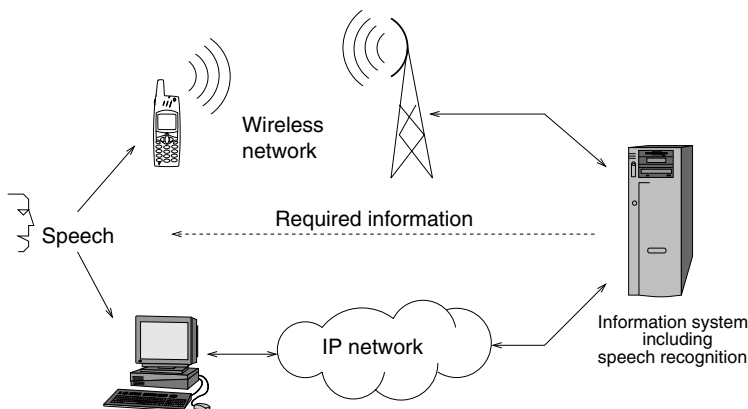


Figure 1.1 General scheme of a speech-driven remote information system

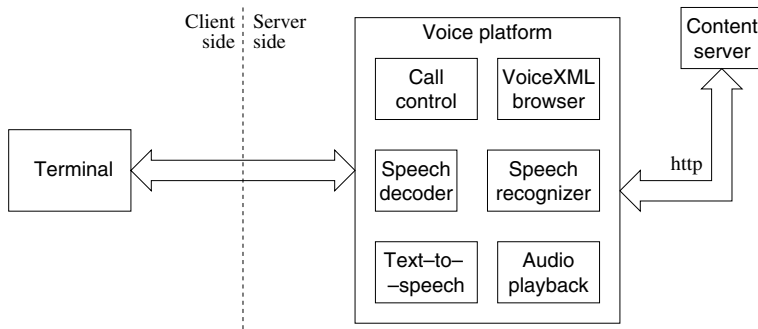


Figure 1.2 Block diagram of a possible RSR-based speech-enabled service architecture

(STQ) Aurora working group. This group is integrated in the STQ committee of the ETSI, and has developed several standards for RSR that are commented on later in this chapter.

The study of RSR systems involves a number of topics. However, the most characteristic issue that must be addressed and that differentiates this type of systems with respect to other ASR systems is robustness against degradation. There are two main sources of degradation that may affect a client–server ASR system. The first one is the noise introduced by the adverse acoustic environment. The acoustic noise is the unavoidable consequence of the noisy environment in which the speaker is usually placed (conference centers, airports, train stations, etc.) and can be treated at both the client and the server. The second one is the distortion introduced by the transmission channel. Typically, this distortion is the consequence of either a degraded wireless link (due to a bad coverage, fading, etc.) or of packet loss (in the case of transmission under the IP protocol), which can be treated by both, a suitable channel encoding at the client side or an error concealment technique applied in the server.

There are several possibilities for the implementation of an RSR system. The basic criterion to classify them is the way in which the speech is coded, transmitted and decoded. The first approaches to RSR were implemented over an analog (or a non-fully digital) transmission infrastructure, such as the public switched telephone network (PSTN). In this case, the speech signal is transmitted through the telephone line and sampled, analyzed and recognized at the server. This approach has the advantage of using a well-deployed network. Also, the whole speech signal is available at the server and can be fully processed there. However, its main drawback is the degradation introduced by the telephone channel, which has a narrow (250–3400 Hz) nonflat frequency response with unknown gain and phase, and this also introduces additive tones and stationary noise, impulse noise, amplitude and phase jitter, and so on. (Moreno and Stern, 1994). Besides, it does not benefit from the potential of a digital environment. The alternative is an RSR system implemented over a digital infrastructure. Researchers and developers have concentrated their attention on this possibility because of the multiple advantages it offers, such as robustness and the possibility of access to a wide range of services. In the following chapters, we will develop different issues regarding the implementation of RSR systems over digital channels, the problems that may arise, and some solutions to them.

1.2 RSR over Digital Channels

There are several possibilities for the implementation of an RSR system over a digital channel. In the first approach, usually known as *network speech recognition* (NSR), the recognition system resides in the network from the client's point of view. In this case, the speech is compressed by a speech codec in order to allow a low bitrate transmission and/or to use an existing speech traffic channel (as in the case of mobile telephony). The recognition is usually performed over the features extracted from the decoded signal, although it is also possible to extract the recognition features directly from the codec parameters. Figure 1.3 shows a scheme of this system architecture. In the case where implementation is over an IP network, a VoIP (Voice over IP) codec can be employed.

However, the approach that has received more attention during the last few years is the one known as *distributed speech recognition* (DSR). In this case, the client includes a local front end that processes the speech signal in order to directly obtain the specific features (usually cepstrum) used by the remote server (back end) to perform recognition, thus avoiding the coding/decoding process required by NSR. The conceptual scheme of DSR is shown in Figure 1.4. DSR has several advantages over NSR:

- It avoids the use of a speech codec, which can reduce the recognition performance because of compression.
- The bitrate involved in DSR is usually smaller than that of NSR.
- In mobile environments, DSR allows the increase of the system robustness. First, the front end located at the client can carry out some type of acoustic noise compensation. Also, the transmission can be carried out over a data channel instead of over a voice channel, so that the system is more robust against channel errors.
- It naturally enables multimodal interfaces by sending the speech features along with other information through a data channel.

On the other hand, the main advantage of NSR is that there is no need for modifying the existing clients in the case of mobile telephony networks.

In the same way as speech codecs are standardized for mobile telephony or VoIP, it is advised that a standardized feature extractor and encoder be used in DSR clients. The implementation of RSR systems over heterogeneous networks can also be eased by using DSR standards. Figure 1.5 shows two possible scenarios that mix mobile and IP

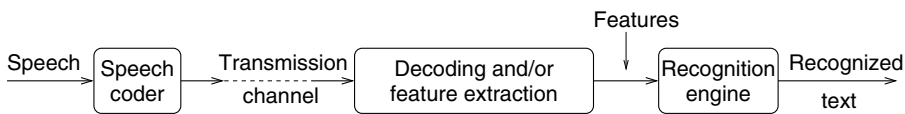


Figure 1.3 Scheme of a network speech recognition (NSR) system

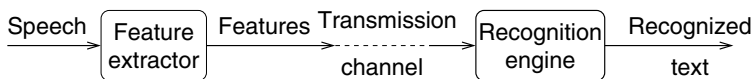


Figure 1.4 Scheme of a distributed speech recognition (DSR) system

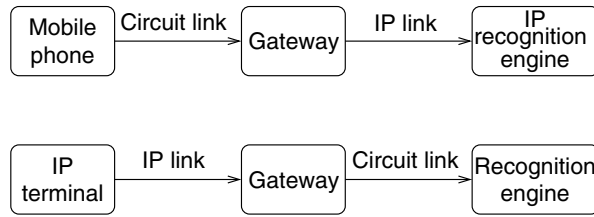


Figure 1.5 Two possible scenarios for a DSR system implemented over a heterogeneous network

networks. In both cases, the gateway must transcode the speech features, which can be straightforward if the mobile and IP payload are the same. On the contrary, NSR over heterogeneous networks may require the use of several codecs (tandeming), which can result in further speech degradation. The ETSI-STQ Aurora working group has been developing several DSR standards during the last few years: a basic standard for DSR (ETSI, 2003a) (front end FE), a DSR standard for working in noisy conditions (ETSI, 2003b) (advanced front end AFE), and two extensions of these two standards to enable tonal language recognition and speech reconstruction (ETSI, 2001, 2003c) (XFE and XAFE). In June 2004, 3GPP (3rd generation partnership project) approved XAFE as the recommended standard for speech-enabled services (SES).

1.3 Organization of the Book

The next chapter deals with the basic concepts required for the development and understanding of a state-of-the-art ASR system. Thus, the fundamentals of cepstral analysis and hidden Markov models (HMM) are introduced. Vector quantization (VQ) is also introduced in this chapter, since VQ is a useful tool in some HMM-based systems. A discussion is provided on how these tools can be used to build an ASR system. The chapter concludes with two specific topics, model adaptation and uncertainty treatment, that we consider specially useful for the development of RSR systems.

As mentioned earlier, an RSR system differs from a classical ASR system in that RSR systems are implemented over digital networks. The introduction of these networks (in particular, mobile and IP networks) in Chapter 3 has a double intention. The first is to provide the reader with those aspects that can be relevant for the implementation of RSR systems. On the other hand, as the subtitle of this book points out, the main topic of this book is robustness in RSR. This is the reason we are also interested in analyzing the degradation that may be introduced when RSR data is transmitted over these networks, which mainly results in data errors and losses. The chapter concludes with an in-depth study of the other source of degradation that may affect an RSR system: environmental noise. This includes the analysis and modeling of additive noise and linear channel distortion.

Chapter 4 deals with the two main architectures for RSR, NSR and DSR and their corresponding techniques for speech compression. First, some fundamentals of speech coding useful for both NSR and DSR are introduced. Then, the two variants of NSR (NSR from decoded speech and NSR from codec parameters) are studied, with special emphasis on degradation and robustness issues. The DSR approach is widely developed

next, since it has attracted the attention of a number of researchers during the last few years. They are classified and studied according to the compression scheme employed, although special attention is paid again to degradation and robustness.

The two subsequent chapters are devoted to the robustness techniques against both transmission channel and environmental degradation. Chapter 5 is devoted to the first type of degradation. The different techniques are classified into two groups: sender-driven (or channel coding) and receiver-based (or error concealment). The first group includes error detection and correction, interleaving and media-specific FEC. The second group includes classical techniques such as interpolation and estimation and also RSR-specific concealment techniques implemented in the speech recognizer. Chapter 6 is concerned with those techniques useful for the development of robust front ends, such as those included in the ETSI standards. In essence, this chapter deals with robust techniques against the environmental degradation, voice activity detection and feature normalization techniques. Robust back-end techniques are out of the scope of this book since they are not specific to RSR systems but apply to ASR systems in general.

The ETSI DSR standards are treated in Chapter 7. The goal of this chapter is to offer a comprehensive approach to these standards, which may facilitate their reading. The four standards are jointly developed so that it is easy to identify both common and differentiating elements. The study is carried out under the scope of the previous chapters, so the different elements contained in the standards can be easily identified and understood.

The book also includes three appendices on alternative representations of the linear prediction (LPC) coefficients, basic digital modulation concepts and a brief review of channel coding techniques. These appendices gather some procedures and concepts that may be useful for the understanding of the previous chapters, although we consider that they are neither the goal of this book nor essential. In particular, the last two appendices are specially developed for an appropriate comprehension of the communication concepts related with RSR and utilized in Chapters 3, 4, 5 and 7.