Preface

This book, entitled *Spoken Language Processing*, addresses all the aspects covering the automatic processing of spoken language: how to automate its production and perception, how to synthesize and understand it. It calls for existing know-how in the field of signal processing, pattern recognition, stochastic modeling, computational linguistics, human factors, but also relies on knowledge specific to spoken language.

The automatic processing of spoken language covers activities related to the analysis of speech, including variable rate coding to store or transmit it, to its synthesis, especially from text, to its recognition and understanding, should it be for a transcription, possibly followed by an automatic indexation, or for human-machine dialog or human-human machine-assisted interaction. It also includes speaker and spoken language recognition. These tasks may take place in a noisy environment, which makes the problem even more difficult.

The activities in the field of automatic spoken language processing started after the Second World War with the works on the *Vocoder* and *Voder* at Bell Labs by Dudley and colleagues, and were made possible by the availability of electronic devices. Initial research work on basic recognition systems was carried out with very limited computing resources in the 1950s. The computer facilities that became available to researchers in the 1970s made it possible to achieve initial progress within laboratories, and microprocessors then led to the early commercialization of the first voice recognition and speech synthesis systems at an affordable price. The steady progress in the speed of computers and in the storage capacity accompanied the scientific advances in the field.

Research investigations in the 1970s, including those carried out in the large DARPA "Speech Understanding Systems" (SUS) program in the USA, suffered from a lack of availability of speech data and of means and methods for evaluating the performance of different approaches and systems. The establishment by DARPA, as part of its following program launched in 1984, of a national language resources center, the Linguistic Data Consortium (LDC), and of a system assessment center, within the National Institute of Standards and Technology (NIST, formerly NBS), brought this area of research into maturity. The evaluation campaigns in the area of speech recognition, launched in 1987, made it possible to compare the different approaches that had coexisted up to then, based on "Artificial Intelligence" methods or on stochastic modeling methods using large amounts of data for training, with a clear advantage to the latter. This led progressively to a quasi-generalization of stochastic approaches in most laboratories in the world. The progress made by researchers has constantly accompanied the increasing difficulty of the tasks which were handled, starting from the recognition of sentences read aloud, with a limited vocabulary of 1,000 words, either speaker-dependent or speaker-independent, to the dictation of newspaper articles for vocabularies of 5,000, 20,000 and 64,000 words, and then to the transcription of radio or television broadcast news, with unlimited size vocabularies. These evaluations were opened to the international community in 1992. They first focused on the American English language, but early initiatives were also carried out on the French, German or British English languages in a French or European context. Other campaigns were subsequently held on speaker recognition, language identification or speech synthesis in various contexts, allowing for a better understanding of the pros and cons of an approach, and for measuring the status of technology and the progress achieved or still to be achieved. They led to the conclusion that a sufficient level of maturation has been reached for putting the technology on the market, in the field of voice dictation systems for example. However, it also identified the difficulty of other more challenging problems, such as those related to the recognition of conversational speech, justifying the need to keep on supporting fundamental research in this area.

This book consists of two parts: a first part discusses the analysis and synthesis of speech and a second part speech recognition and understanding. The first part starts with a brief introduction of the principles of speech production, followed by a broad overview of the methods for analyzing speech: linear prediction, short-term Fourier transform, time-representations, wavelets, cepstrum, etc. The main methods for speech coding are then developed for the telephone bandwidth, such as the CELP coder, or, for broadband communication, such as "transform coding" and quantization methods. The audio-visual coding of speech is also introduced. The various operations to be carried out in a text-to-speech synthesis system are then presented regarding the linguistic processes (grapheme-to-phoneme transcription, syntactic and prosodic analysis) and the acoustic processes, using rule-based approaches or approaches based on the concatenation of variable length acoustic units. The different types of speech signal modeling – articulatory, formant-based, auto-regressive, harmonic-noise or PSOLA-like – are then described. The evaluation of speech synthesis systems is a topic of specific attention in this chapter. The extension of speech synthesis to talking faces animation is the subject of the next chapter, with a presentation of the application fields, of the interest of a bimodal approach and of models used to synthesize and animate the face. Finally, computational auditory scene analysis opens prospects in the signal processing of speech, especially in noisy environments.

The second part of the book focuses on speech recognition. The principles of speech recognition are first presented. Hidden Markov models are introduced, as well as their use for the acoustic modeling of speech. The Viterbi algorithm is depicted, before introducing language modeling and the way to estimate probabilities. It is followed by a presentation of recognition systems, based on those principles and on the integration of those methodologies, and of lexical and acoustic-phonetic knowledge. The applicative aspects are highlighted, such as efficiency, portability and confidence measures, before describing three types of recognition systems: for text dictation, for audio documents indexing and for oral dialog. Research in language identification aims at recognizing which language is spoken, using acoustic, phonetic, phonotactic or prosodic information. The characteristics of languages are introduced and the way humans or machines can achieve that task is depicted, with a large presentation of the present performances of such systems. Speaker recognition addresses the recognition and verification of the identity of a person based on his voice. After an introduction on what characterizes a voice, the different types and designs of systems are presented, as well as their theoretical background. The way to evaluate the performances of speaker recognition systems and the applications of this technology are a specific topic of interest. The use of speech or speaker recognition systems in noisy environments raises especially difficult problems to solve, but they must be taken into account in any operational use of such systems. Various methods are available, either by pre-processing the signal, during the parameterization phase, by using specific distances or by adaptation methods. The Lombard effect, which causes a change in the production of the voice signal itself due to the noisy environment surrounding the speaker, benefits from a special attention. Along with recognition based solely on the acoustic signal, bi-modal recognition combines two acquisition channels: auditory and visual. The value added by bimodal processing in a noisy environment is emphasized and architectures for the audiovisual merging of audio and visual speech recognition are presented. Finally, applications of automatic spoken language processing systems, generally for human-machine communication and particularly in telecommunications, are described. Many applications of speech coding, recognition or synthesis exist in many fields, and the market is growing rapidly. However, there are still technological and psychological barriers that require more work on modeling human factors and ergonomics, in order to make those systems widely accepted.

The reader, undergraduate or graduate student, engineer or researcher will find in this book many contributions of leading French experts of international renown who share the same enthusiasm for this exciting field: the processing by machines of a capacity which used to be specific to humans: language.

Finally, as editor, I would like to warmly thank Anna and Frédéric Bimbot for the excellent work they achieved in translating the book *Traitement automatique du langage parlé*, on which this book is based.

Joseph Mariani
November 2008