Keller, E. (1994). Preface. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges* (pp. xi-xiii). Chichester: John Wiley.

Preface

After years of unfulfilled promises, speech has finally arrived on the personal computer. We can finally ask the computer in so many words to do something — and more often than not, it will follow our instructions. And after years of affecting a Mickey Mouse voice, computers can now finally speak with something resembling a human voice.

There were some excellent reasons why it took so long. Speech is an exceedingly complex human activity, and it took a great deal of scientific exploration, experimentation, development, as well as considerable computational power to deal successfully with speech. Moreover, the challenge hasn't been fully met yet. Change the way you pronounce your commands ever so subtly, and the computer won't understand what it understood a minute ago. Or put even good computer speech and human speech side by side, and most listeners will still show a clear preference for human speech.

As a result, the race is on to improve computer speech interfaces, a race that has recently been joined by some of the world's major computer firms. It can be expected that over the next ten years, major efforts will be launched to create competing products, and to improve on existing software. Moreover, the world is a multilingual place, and at least the synthesis side of computer speech will have a distinctly international flavour. The recognition side may initially be handled in language-independent fashion, but in the long run, even in this area, "one size may not fit all". Increasingly, the aim will be to create and understand natural, error-free and truly human-like speech. Efforts will thus have to be made to deal appropriately not only with those aspects of speech that are similar from one language to the next, but also with those that determine their differences.

An additional major challenge will be to produce and recognise so-called "connected speech". Moving from simple, one- or two-word commands chosen from a small vocabulary, to complete sentences, i.e., to connected speech, involving tens of thousands of different words, is not simply a matter of multiplying existing techniques or of increasing computing power. To attain truly satisfactory performance, entirely new types of knowledge about speech and language will likely have to be incorporated into speech algorithms.

So all signs point to increasing world-wide efforts to create speech-based computer interfaces during the next decade and beyond. Attempts will be made to bring to the personal computer logic and programs that have so far required mainframe resources. A whole new generation of computer scientists will thus wonder "what speech is all about", and how to bring it to the computer in new and better ways. This was the main motivation for creating this book. The student new to computer speech has few places to turn to acquire the basic

notions of this field. The (excellent) books that do exist are generally too technical, or tend to require too much prior knowledge and familiarity with terminology from adjoining fields, such as signal processing, linguistics, or phonetic science. This volume was carefully written and edited to be as clear as possible, to be incrementally structured, and to provide explanations of terminology that is not necessarily familiar to computer scientists.

There is another apparent lack in the current literature, somewhat more elusive in nature. Technical volumes on speech synthesis and speech recognition tend to concentrate on the "tried and true" methods, as they have been developed over the last two decades. Few pose the question of the challenges of tomorrow, that is, issues such as how computer speech can be made to sound more natural and more human-like, how to deal with the problem of speaker-to-speaker variation, or how to develop techniques to distinguish input speech from background noise or from other, irrelevant voices in the environment. And yet, those are exactly the issues that will confront young scientists entering the field in a few years' time. Many will be asked not to re-invent the wheel and create yet another speech I/O system, but rather, to develop techniques that will advance the field in general. Consequently, a major section of this book was devoted to the question of how humans seem to solve these problems — explorations that should provide a rich source of ideas of how such problems could be solved by machines.

To sum up, the present volume was designed to accomplish three overall objectives, each of which is associated with a major section of the book. A few introductory chapters about the nature of speech and speech signals are given at the beginning of the book (the *"Background"* section). These are designed to bring the well-motivated computer scientist "up to speed", and to facilitate an understanding of those adjoining areas that are least likely to be taught in computer science courses. These chapters not only present some of the well-known aspects of speech, such as articulation and its basic relationship with the acoustic signal, but they also point out some of the less evident aspects that speech devices must deal with, such as intonation, timing and pausing.

The second objective is to provide a quick overview of existing methods in speech synthesis and speech recognition. These chapters have been collected in the *State of the Art* section of the volume. Because of the great volatility of the field, the emphasis was placed on fundamental principles of operation, rather than on the presentation of commercially available systems. Three chapters deal with synthesis and two with recognition. Further general information on recognition is found in the introduction to Section 3.

The third objective is to present a number of areas that represent challenges for future development in computer speech devices. These articles are presented in the *Challenges* section of the book. How can computer speech be made more human-like? Is a close modelling of human articulatory behaviour the answer? What improvements can arise from a study of the detailed timing structure of speech? And what about the infamous variability of human speech? Have we simply not yet discovered the least variable parameters of speech, or do we need large-scale combinatorial logic to eliminate alternative interpretations? What techniques are required to synthesise or to recognise connected speech? And what about the identification of speech in "difficult"

circumstances? How does the human ear perform this task? Can we learn something from human processing that can be translated into successful algorithms?

Clearly, it is impossible to cover *all* questions of this sort. However, a sufficient sampling of future challenges has been collected here to stimulate the desire to dig deeper, to know more, to experiment, and to implement. The authors of this volume have made a conscientious effort to write clearly, and to cover the major concepts of their fields. It is hoped that readers will find much of interest here to advance their understanding of speech synthesis and speech recognition. Readers new to the field are strongly encouraged to study the introductions to each of the sections. Particularly the introduction to the *Challenges* section is useful for a good understanding of the relevance of these more technical chapters of the volume.

A great many thanks are due to the authors contributing to this volume. It is not an easy task — nor necessarily a very prestigious one — to write readily accessible chapters in a clear, even a didactic style. This is of course especially true of the many authors whose native language is not English. But after an initial discussion, everyone agreed that the need was real, and that the effort was worthwhile. The papers all arrived in time, and publication of the volume became possible with a minimum of delay. Also, the financial aid of the 3e Cycle Suisse-Romand is acknowledged which permitted the organisation of an unforgettable conference in the Swiss Alps, where many of these articles were presented in their initial form.

Lausanne, March 1994 *Eric Keller*