Handbook of Standards and Resources for Spoken Language Systems

Dafydd Gibbon, Roger Moore, and Richard Winski (editors)
(University of Bielefeld, DRA Speech Research Unit, and Vocalis)

Berlin: Mouton de Gruyter, 1997,
xxx+886 pp and CD-ROM; hardbound,
ISBN 3-11-015366-1, DM 298.00

Reviewed by
Jan P. H. van Santen
Lucent Technologies—Bell Laboratories

1. Introduction

This handbook is different from most books reviewed on these pages in that it is the result of, and a driving force in, a complex process of planned European-style international cooperation. Thus, it should be reviewed not only for content, but also for its role in this process.

Let me start by briefly describing the cooperative process. EAGLES (Expert Advisory Group on Language Engineering Standards) is an initiative of the Commission of the European Union, whose purposes include producing specifications and guidelines, and encouraging cooperation between industry and academia, and between European countries. The initiative comprised five working groups, one of which—the Spoken Language Working Group, WG5—is responsible for the handbook under review.

2. Content

The authors interpret the concept of spoken language systems broadly, and include any system whose key components consist of speech technologies such as automatic speech recognition (ASR), speech synthesis (SS or TTS), speaker verification, and coding.

This book is not meant as a textbook for these technologies, although the authors mention that it may serve as an introductory text in some cases. Instead, its stated goals are that of supplying reference materials for researchers and developers working in these areas, providing information on common practice in Europe for non-Europeans, and delivering guidance on systems specification and assessment for corporate end users.

Besides a massive bibliography and a 170-page section listing spoken language reference materials, the book has three main sections. The first, "Spoken language and corpus design," covers system design (written by Choukri), SL corpus design (den Os), SL corpus collection (Draxler), and SL corpus representation (den Os); the second section, "Spoken language characterization," contains SL lexica (Gibbon), language models (Ney), and physical characterization and description (Knohl and Kraft); and the third section, "Spoken language system assessment," has chapters on methodologies and experimental design (Howell), recognition systems (van Leeuwen and Steeneken), speaker verification systems (Bimbot and Chollet), synthesis systems (van Bezooijen and van Heuven), and interactive systems (Fraser).

These chapters vary widely in depth and level of sophistication. The chapters

by Ney and by Bimbot and Chollet are quite technical—and also quite good—but require the reader to almost be a specialist in their respective areas. On the other hand, the chapter by Howell on statistics and experimental design is at the level of a first statistics course for undergraduate psychology students. Because much assessment is multidimensional, I would have liked to see more material on multivariate statistical inference, multidimensional scaling, and clustering. But the chapter is clearly written and well-organized.

The chapter by Ney on language modeling techniques for ASR raises the issue of why the book does not contain similar chapters on speech synthesis components. For example, a chapter on certain broad classes of text analysis such as morphological decomposition, pronunciation dictionaries, or sense disambiguation would not have been out of place.

An issue that I find insufficiently emphasized in the section on assessment is that of coverage and generalizability across applications and textual genres. In the case of ASR, the standard assessment procedure is to use the bulk of a corpus for training, and the remainder for tests. Since both test and training samples are drawn from the same corpus, and since most corpora have some degree of homogeneity (e.g., in terms of vocabulary, syntactic style, or topic) this means that good performance on test samples may be a poor predictor of performance on samples outside of the corpus. This raises issues about the relevance of such assessment results when a user wants to use a system in an application with textual materials different from those used in the assessment.

In the case of TTS assessment, with some notable exceptions, most evaluation procedures use a fixed, known, and generally small set of test items. Since many components of TTS systems can be easily fine-tuned, there is no reason why developers would not make sure that their system operates properly for some of these textual materials. The reason that TTS systems can be so easily fine-tuned is that several of their components are *list-like*, such as pronunciation dictionaries, concatenative units, and various types of rules (assuming interactions that are not too severe). With such components, one can alter a few items without having much effect on how the system processes other inputs. This contrasts with speech coding systems, where alterations in components generally affect performance on any input, not merely on the 150 most frequent Belgian surnames.

On balance, however, there is a wealth of material here that until now was not available in this convenient form. The editors should also be lauded for the book's overall organization; for one, I found the *recommendation* chapters in each section particularly helpful.

3. Is Speech Technology Ready for a Handbook?

This review would not be complete without discussing the state of speech technology itself. After all, the existence of a handbook may be taken to imply a certain level of maturity, and it is important to ask whether this level indeed has been reached and, if not, what harm there is in pretending it has been reached.

The candid truth about speech technology is that, whereas predictions of future computer hardware performance have been generally proven to be accurate or even pessimistic, predictions of progress in speech technology have been consistently too optimistic. For at least two decades, there has been a sliding five-year window within which speech technology would produce systems that understand us and talk with voices hard to distinguish from human speech. With few exceptions, current speech technologies work well only under conditions that are highly restrictive in terms of

domain, user, or both. For example, ASR systems that are currently sufficiently reliable to be used as more than a toy all lie on a curve between, on the one hand, systems for speaker-independent word-spotting with extremely small vocabularies, and, on the other hand, systems for large-vocabulary speaker-dependent dictation requiring vast amounts of training by the user.

Most of the progress made is due to vast improvements in computer hardware allowing better parameter estimation and more expansive searches, to availability of large speech and text corpora, and to algorithmic and statistical innovations. Remarkably few contributions have come from "science," by which I refer to the study of human speech and language production and understanding in traditional academic fields such as linguistics, psycholinguistics, neuroscience, physiology, and acoustics. A fundamental question is to what degree current disappointing performance levels are due to this state of affairs, and how to remedy this. Although I personally would much like to believe that these traditional academic fields have critically important contributions to make, they may need to be reorganized to become more multidisciplinary and more receptive to concepts and ideas coming from technology.

The benefit of a handbook is that it is an important tool for speech and language researchers in any field, and also helps to bring together researchers from a variety of disciplines. Thus, it could help in making the field more multidisciplinary, and thereby produce a better balance between science and technology in this area.

There is a subtle danger, however, in pretending that the field is more mature than it really is—a pretense not unique to this book; one can sense it at any speech technology conference. The danger is that it sends two messages: a message to corporate and governmental funding organizations that their investments have a high probability of quickly producing concrete results; and a message to students entering the field that a substantial fraction of scientific and technological issues have been resolved. If it is not the case that speech technology has passed the threshold of maturity, then funding organizations misallocate scarce resources and students do not ask sufficiently provocative questions.

4. Conclusion

There is no doubt that the processes that have culminated in the production of this book have been beneficial for the European speech technology community. The book is likely to have a major influence on standards committees and other such organizations. This influence will probably be very beneficial, because the book was written by authors who are experts in their respective fields and do not represent narrow commercial or local interests. In addition, it represents a broader consensus than is usually present in these committees and organizations.

The book brings together in one volume information that up to now was essentially inaccessible because it was so widely dispersed. Some chapters are well-written and lucid and would be useful teaching tools. But there are some omissions from the book, and the level of difficulty varies widely between chapters.

In other words, the book seems to have largely reached its objectives. But, as I have argued above, a book of this high degree of visibility will also have unintended consequences. I have pointed out a somewhat subtle consequence—which is that it strengthens the portrayal of speech technology as scientifically more advanced than it in fact is—and that this may hamper scientific progress. On balance, however, the use of the book as a resource for working scientists is likely to easily outweigh this potential consequence.

Jan P. H. van Santen is co-editor of *Progress in Speech Synthesis* (Springer-Verlag, 1997). Van Santen's address is: Language Modeling Research Department, Bell Labs—Lucent Technologies, Room 2D–431, 600 Mountain Avenue, Murray Hill, NJ 07974; e-mail: jphvs@research.bell-labs.com; http://www.bell-labs.com/project/tts/jphvs.html