Preface

When I was being interviewed at the handwriting recognition group of IBM T.J. Watson Research Center in December of 1990, one of the interviewers asked me why, being a mechanical engineer, I was applying for a position in that group. Well, he was an electrical engineer and somehow was under the impression that handwriting recognition was an electrical engineering field! My response was that I had done research on Kinematics, Dynamics, Control, Signal Processing, Optimization, Neural Network Learning theory and lossless image compression during the past 7 years while I was in graduate school. I asked him what background he thought would have been more relevant to do research in handwriting recognition.

Anyhow, I joined the on-line handwriting recognition group which worked side-by-side with the speech recognition group. Later, I transferred to the speech recognition group and worked on speaker recognition. Aside from the immediate front-end processing, on-line handwriting recognition, signature verification, speech recognition and speaker recognition have a lot in common. During the 10 years at IBM I also worked on many complementary problems such as phonetics, statistical learning theory, language modeling, information theoretic research, etc. This continued with further work on real-time large-scale optimization, interactive voice response systems, standardization and more detailed speaker recognition research at Recognition Technologies, Inc. to the present date, not to mention the many years of code optimization, integer arithmetic, software architecture and alike within the past 25 years.

The reason for sharing this story with the reader is to point out the extreme multi-disciplinary nature of the topic of speaker recognition. In fact, every one of the fields which I mentioned above, was quite necessary for attaining a deep understanding of the subject. This was the prime motivation which lead me to the writing of this book. As far as I know, this is the first textbook (reference book) on the subject which tries to deal with every aspect of the field, as much as possible. I have personally designed and implemented (coded) two full-featured speaker recognition systems and in the process have had to deal with many different aspects of the subject from theory to

practice.

One problem with which many researchers are faced, when dealing with highly multi-disciplinary subjects such as speaker recognition, is the scattered information in all the relevant fields. Usually, most treatments of the subject try to use hand-waving to get the reader through all the different aspects of the subject. In their treatment, most survey papers, throw a plethora of references at the reader so that he/she would follow up on each of the many leads – which is usually impractical. This causes a half-baked understanding of the subject and its details which will be carried over from master to apprentice, leaving the field crippled at times.

In the above description, while qualifying this book, I used the word textbook, but I also parenthetically referred to it as a reference book. Well, originally when I was asked to write it, we had a textbook approach in mind. However, as I delved deeper into the attempt of presenting all the necessary material, based on the motivation which was stated earlier, the coverage of the different subjects grew quickly from an intended 300 page textbook to nearly 900 pages which probably qualifies as a reference book. In fact, most of the book may be used as reference material for many related subjects.

In my many years of teaching different courses at Columbia University, such as Speech Recognition, Signal Recognition, and Digital Control, I have noticed the following. Since today's technologies are built layer-upon-layer on top of existing basic technologies, the amount of underlying knowledge necessary for understanding the topics at the tips of these theoretical hierarchies has grown exponentially. This makes it quite hard for a researcher in a multi-disciplinary topic to grasp the intricacies of the underlying theory. Often, to deal with the lack of time, necessary for an in-depth understanding of the underlying theory, it is either skipped or left to the pursuance of the students, of their own volition.

In this book, I have tried to cover as much detail as possible and to keep most of the necessary information self-contained and rigorous. Although, you will see many references presented at the end of each chapter and finally as a collection in a full bibliography, the references are only meant for the avid reader to follow up into the nitty-gritty details upon interest. Most of the high-level details are stated in the 26 chapters which make up this book.

To be able to present the details, and yet have a smooth narrative in the main text, a large amount of the detailed material is included in the last 4 chapters of the book, categorized as *Background Material*. These chapters start with the coverage of some necessary *linear algebra* and related mathematical bases followed by a very detailed chapter on *integral transforms*. Since integral transforms are central to the *signal processing* end of the subject, and they heavily rely on an intimate knowledge of *complex variable theory*, Chapter 24 tries to build that foundation for the reader. Moreover, the essence of theoretical subjects such as *neural networks* and *support*

vector machines is the field of *numerical optimization* which has been covered in some detail in Chapter 25. The last chapter covers details on *standards*, related to the speaker recognition field. This is a practical aspect which is usually left out in most textbooks and, in my opinion, should be given much more attention.

The main narrative of the book has *three major parts*:
Part I covers the *introductory and basic theory* of the subject including *anatomy*, *signal representation*, *phonetics*, *signal processing and feature extraction*, *probability theory*, *information theory*, *metrics and distortion measures*, *Bastian learning theory*, *parameter estimation and leaning*, *clustering*, *parameter transformation*, *hidden Markov modeling*, *neural networks*, and *support vector machines*.

The second part, *advanced theory*, covers subjects which deal more directly with speaker recognition. These topics are *speaker modeling*, *speaker recognition implementation*, and *signal enhancement and compensation*.

Part III, *practice*, discusses topics specifically related to the implementation of speaker recognition or related issues. These are *representation of results*, *time-lapse effects*, *adaptation techniques*, and finally, *overall design issues*.

Every effort has been made to deliver the contents of the book in a hierarchical fashion. In other words, think of writing an efficient program in a class-based programming language where the main program is simply a few lines. The main program, in this, case would be the chapters in *Part III (Practice)*. The classes that are instantiated within the main program, mainly come from *Part II (advanced theory)* and they in-turn include more specialized classes from *Part I (basic theory)*. Part II and Part I classes make calls to methods in *Part IV (background material)*.

Yorktown Heights, New York, August 2011 *Homayoon Beigi*

http://www.springer.com/978-0-387-77591-3

Fundamentals of Speaker Recognition

Beigi, H.

2011, LXI, 942 p. 177 illus., Hardcover
ISBN: 978-0-387-77591-3