

Preface

It is without doubt that we live in an interconnected world where we are always within reach of smartphone, tablet or telephone system and through which we are always communicating with friends and family, colleagues and workmates or an automated voicemail or interactive dialog system. Otherwise we just relax and switch on the radio, stream some music or watch a movie. These activities are part of our everyday lives. They have been made possible through the advances in speech and audio processing and recognition technologies which only in the last decade have seen an explosion in usage through a bewildering array of devices and their capabilities.

Speech coding refers to the digital representation of the information-bearing analog speech signal, with emphasis on removing the inherent redundancies. Efficient coding of speech waveforms is essential in a variety of transmission and storage applications such as traditional telephony, wireless communications (e.g., mobile phones), internet telephony, voice-over-internet protocol (VoIP) and voice mail. Many of these applications are currently going through an impressive growth phase.

Speech recognition encompasses a range of diverse technologies from engineering, signal processing, mathematical statistical modelling and computer science language processing necessary to achieve the goal of human–computer interaction using our most natural form of communication: speech. Applications of speech recognition have exploded due to the advent of smartphone technology where the use of the traditional keyboard and mouse has given way to touch and speech and in enterprise automated computer voice response services for enquiries and transactions. We are now experiencing an exponential growth in the adoption of speech recognition in smartphone and mobile technology, in information and transaction services and increased R&D effort on efficient low-cost and low-power implementations, robustness in the presence of ambient noise and reliable language understanding and dialog management.

In this book we provide readers with an overview of the basic principles and latest advances across a wide variety of speech and audio areas and technologies across ten chapters. These are organized into three parts from front end signal processing involved with speech coding and transmission and the more sophisticated approaches deployed for speech enhancement to the back end user interface involved with speech recognition to the latest “hot” research areas in emotion recognition and speaker diarization. This book brings together internationally recognized researchers across these diverse fields spanning many countries including the USA, Australia, Singapore and Japan from leading research universities, industry experts from Microsoft and Qualcomm and front line research institutions like Microsoft Research, USA, Institute for Infocomm Research, Singapore and NTT Labs, Japan.

We have divided the book into three parts: “Overview of Speech and Audio Coding”, “Review and Challenges in Speech, Speaker and Emotion Recognition” and “Current Trends in Speech Enhancement”.

Part I comprises four chapters.

The first chapter traces a historical account of speech coding from a front-row seat participant and is titled “From ‘Harmonic Telegraph’ to Cellular Phones”. The second chapter gives an introduction to speech and audio coding, emerging topics and some challenges in speech coding research. In the third chapter, we present scalable and multirate speech coding for Voice-over-Internet Protocols (VoIP) networks. We also discuss packet-loss robust speech coding. The fourth chapter details the recent speech coding standards and technologies. Recent developments in conversational speech coding technologies, important new algorithmic advances, and recent standardization activities in ITU-T, 3GPP, 3GPP2, MPEG and IETF that offer a significantly improved user experience during voice calls on existing and future communication systems are presented. The Enhanced Voice Services (EVS) project in 3GPP that is developing the next generation speech coder in 3GPP is also presented.

Part II includes four chapters which cover the depth and breadth of speech and audio interfacing technologies. The part starts with two overview chapters presenting the latest advances and thoughts in statistical estimation and machine learning approaches to feature modelling for speech recognition, specifically ensemble learning approaches and dynamic and deep neural networks. This is followed by two chapters representing new and emerging research and technology areas which extend speaker recognition to: how speech can be used to detect and recognize the emotional state of a speaker instead, to the deployment in the real world task of speaker diarization in room conversations, that is who spoke when.

Part III presents two different alternative paradigms to the task of speech enhancement. Assuming the availability of multiple microphone arrays the first chapter in this part deals with speech enhancement in the widest sense where speech is degraded by interfering speakers, ambient noise and reverberations and provides a framework which integrates both spatial and spectral features for a blind source separation and speech enhancement solution. The second and final chapter in this part presents a more fundamental approach for signal channel speech enhancement in the presence of ambient additive noise based on the modulation

spectrum approach for differentiating and separating time-frequency speech features from the additive interfering noise features.

The convergence of technologies as exemplified by smartphone devices is a key driver of speech and audio processing. From the initial speech coding and transmission to the enhancement of the speech and the final recognition and modelling of the speech we have in our hands that smart phone device that can capture, transmit, store, enhance and recognize what we want to say. This book provides a unique collection of timely works representing the range of these processing technologies and the underlying research and should provide an invaluable reference and a source of inspiration for both researchers and developers working in this exciting area.

We thank all the chapter contributors which includes Bishnu Atal, Jerry Gibson, Koji Seto, Daniel Snider, Imre Varga, Venkatesh Krishnan, Vivek Rajendran, Stephane Villette, Yunxin Zhao, Jian Xue, Xin Chen, Li Deng, Vidhyasaharan Sethu, Julien Epps, Eliathamby Ambikairajah, Trung Hieu Nguyen, Eng Siong Chng, Haizhou Li, Yasuaki Iwata, Tomohiro Nakatani, Takuya Yoshioka, Masakiyo Fujimoto, Hirofumi Saito, Kuldip Paliwal and Belinda Schwerin for their work.

We thank Springer Publishers for their professionalism and for support in the process of publishing the book. We especially thank Chuck Glasser and Jessica Lauffer.

We hope the material presented here will educate new comers to the field and also help elucidate to practicing engineers and researchers the important principles of speech/audio coding, speech recognition and speech enhancement with applications in many devices and applications such as wireless communications (e.g., mobile phones), voice-over-IP, internet telephony, video comm., text-to-speech, etc. which are ubiquitous today.

Santa Clara, CA, USA
Crawley, WA, Australia
Santa Clara, CA, USA
June 2014

Tokunbo Ogunfunmi
Roberto Togneri
Madihally (Sim) Narasimha

Speech and Audio Processing for Coding, Enhancement
and Recognition

Ogunfunmi, T.; Togneri, R.; Narasimha, M.S. (Eds.)

2015, X, 345 p. 79 illus., 32 illus. in color., Hardcover

ISBN: 978-1-4939-1455-5