



# **Spoken Language Processing**

**Edited by  
Joseph Mariani**

**ISTE**

 **WILEY**

TN912.34  
S762

# Spoken Language Processing

Edited by  
Joseph Mariani



ISTE



 WILEY

First published in France in 2002 by Hermes Science/Lavoisier entitled *Traitement automatique du langage parlé 1 et 2* © LAVOISIER, 2002

First published in Great Britain and the United States in 2009 by ISTE Ltd and John Wiley & Sons, Inc.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

ISTE Ltd  
27-37 St George's Road  
London SW19 4EU  
UK

[www.iste.co.uk](http://www.iste.co.uk)

John Wiley & Sons, Inc.  
111 River Street  
Hoboken, NJ 07030  
USA

[www.wiley.com](http://www.wiley.com)

© ISTE Ltd, 2009

The rights of Joseph Mariani to be identified as the author of this work have been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

---

Library of Congress Cataloging-in-Publication Data

Traitement automatique du langage parlé 1 et 2. English  
Spoken language processing / edited by Joseph Mariani.  
p. cm.

Includes bibliographical references and index.

ISBN 978-1-84821-031-8

1. Automatic speech recognition. 2. Speech processing systems. I. Mariani, Joseph. II. Title.  
TK7895.S65T7213 2008  
006.4'54--dc22

2008036758

---

British Library Cataloguing-in-Publication Data

A CIP record for this book is available from the British Library

ISBN: 978-1-84821-031-8

---

Printed and bound in Great Britain by CPI Antony Rowe, Chippenham, Wiltshire.



Cert no. SGS-COC-2953  
[www.fsc.org](http://www.fsc.org)  
© 1996 Forest Stewardship Council

## Spoken Language Processing

## 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



# Table of Contents

<b>Preface</b> . . . . .	xiii
<b>Chapter 1. Speech Analysis</b> . . . . .	1
Christophe D’ALESSANDRO	
1.1. Introduction. . . . .	1
1.1.1. Source-filter model . . . . .	1
1.1.2. Speech sounds. . . . .	2
1.1.3. Sources . . . . .	6
1.1.4. Vocal tract . . . . .	12
1.1.5. Lip-radiation. . . . .	18
1.2. Linear prediction. . . . .	18
1.2.1. Source-filter model and linear prediction . . . . .	18
1.2.2. Autocorrelation method: algorithm . . . . .	21
1.2.3. Lattice filter . . . . .	28
1.2.4. Models of the excitation. . . . .	31
1.3. Short-term Fourier transform . . . . .	35
1.3.1. Spectrogram . . . . .	35
1.3.2. Interpretation in terms of filter bank. . . . .	36
1.3.3. Block-wise interpretation . . . . .	37
1.3.4. Modification and reconstruction . . . . .	38
1.4. A few other representations . . . . .	39
1.4.1. Bilinear time-frequency representations . . . . .	39
1.4.2. Wavelets . . . . .	41
1.4.3. Cepstrum. . . . .	43
1.4.4. Sinusoidal and harmonic representations . . . . .	46
1.5. Conclusion . . . . .	49
1.6. References . . . . .	50

<b>Chapter 2. Principles of Speech Coding</b> . . . . .	<b>55</b>
Gang FENG and Laurent GIRIN	
2.1. Introduction. . . . .	55
2.1.1. Main characteristics of a speech coder . . . . .	57
2.1.2. Key components of a speech coder . . . . .	59
2.2. Telephone-bandwidth speech coders . . . . .	63
2.2.1. From predictive coding to CELP. . . . .	65
2.2.2. Improved CELP coders . . . . .	69
2.2.3. Other coders for telephone speech . . . . .	77
2.3. Wideband speech coding . . . . .	79
2.3.1. Transform coding. . . . .	81
2.3.2. Predictive transform coding. . . . .	85
2.4. Audiovisual speech coding. . . . .	86
2.4.1. A transmission channel for audiovisual speech . . . . .	86
2.4.2. Joint coding of audio and video parameters . . . . .	88
2.4.3. Prospects . . . . .	93
2.5. References . . . . .	93
<b>Chapter 3. Speech Synthesis.</b> . . . .	<b>99</b>
Olivier BOËFFARD and Christophe D'ALESSANDRO	
3.1. Introduction. . . . .	99
3.2. Key goal: speaking for communicating . . . . .	100
3.2.1. What acoustic content? . . . . .	101
3.2.2. What melody? . . . . .	102
3.2.3. Beyond the strict minimum . . . . .	103
3.3 Synoptic presentation of the elementary modules in speech synthesis systems . . . . .	104
3.3.1. Linguistic processing. . . . .	105
3.3.2. Acoustic processing . . . . .	105
3.3.3. Training models automatically . . . . .	106
3.3.4. Operational constraints . . . . .	107
3.4. Description of linguistic processing . . . . .	107
3.4.1. Text pre-processing. . . . .	107
3.4.2. Grapheme-to-phoneme conversion . . . . .	108
3.4.3. Syntactic-prosodic analysis . . . . .	110
3.4.4. Prosodic analysis . . . . .	112
3.5. Acoustic processing methodology . . . . .	114
3.5.1. Rule-based synthesis . . . . .	114
3.5.2. Unit-based concatenative synthesis . . . . .	115
3.6. Speech signal modeling. . . . .	117
3.6.1. The source-filter assumption . . . . .	118
3.6.2. Articulatory model . . . . .	119
3.6.3. Formant-based modeling . . . . .	119

3.6.4. Auto-regressive modeling . . . . .	120
3.6.5. Harmonic plus noise model . . . . .	120
3.7. Control of prosodic parameters: the PSOLA technique . . . . .	122
3.7.1. Methodology background . . . . .	124
3.7.2. The ancestors of the method . . . . .	125
3.7.3. Descendants of the method . . . . .	128
3.7.4. Evaluation . . . . .	131
3.8. Towards variable-size acoustic units . . . . .	131
3.8.1. Constitution of the acoustic database . . . . .	134
3.8.2. Selection of sequences of units . . . . .	138
3.9. Applications and standardization . . . . .	142
3.10. Evaluation of speech synthesis. . . . .	144
3.10.1. Introduction . . . . .	144
3.10.2. Global evaluation . . . . .	146
3.10.3. Analytical evaluation . . . . .	151
3.10.4. Summary for speech synthesis evaluation. . . . .	153
3.11. Conclusions . . . . .	154
3.12. References. . . . .	154

## **Chapter 4. Facial Animation for Visual Speech . . . . . 169**

Thierry GUIARD-MARIGNY

4.1. Introduction. . . . .	169
4.2. Applications of facial animation for visual speech. . . . .	170
4.2.1. Animation movies . . . . .	170
4.2.2. Telecommunications . . . . .	170
4.2.3. Human-machine interfaces . . . . .	170
4.2.4. A tool for speech research. . . . .	171
4.3. Speech as a bimodal process. . . . .	171
4.3.1. The intelligibility of visible speech . . . . .	172
4.3.2. Visemes for facial animation . . . . .	174
4.3.3. Synchronization issues. . . . .	175
4.3.4. Source consistency . . . . .	176
4.3.5. Key constraints for the synthesis of visual speech. . . . .	177
4.4. Synthesis of visual speech . . . . .	178
4.4.1. The structure of an artificial talking head. . . . .	178
4.4.2. Generating expressions . . . . .	178
4.5. Animation. . . . .	180
4.5.1. Analysis of the image of a face. . . . .	180
4.5.2. The puppeteer . . . . .	181
4.5.3. Automatic analysis of the speech signal . . . . .	181
4.5.4. From the text to the phonetic string . . . . .	181
4.6. Conclusion . . . . .	182
4.7. References . . . . .	182

<b>Chapter 5. Computational Auditory Scene Analysis</b> . . . . .	189
Alain DE CHEVEIGNÉ	
5.1. Introduction. . . . .	189
5.2. Principles of auditory scene analysis . . . . .	191
5.2.1. Fusion versus segregation: choosing a representation . . . . .	191
5.2.2. Features for simultaneous fusion. . . . .	191
5.2.3. Features for sequential fusion. . . . .	192
5.2.4. Schemes . . . . .	193
5.2.5. Illusion of continuity, phonemic restoration . . . . .	193
5.3. CASA principles. . . . .	193
5.3.1. Design of a representation. . . . .	193
5.4. Critique of the CASA approach. . . . .	200
5.4.1. Limitations of ASA. . . . .	201
5.4.2. The conceptual limits of “separable representation” . . . . .	202
5.4.3. Neither a model, nor a method? . . . . .	203
5.5. Perspectives . . . . .	203
5.5.1. Missing feature theory . . . . .	203
5.5.2. The cancellation principle. . . . .	204
5.5.3. Multimodal integration . . . . .	205
5.5.4. Auditory scene synthesis: transparency measure . . . . .	205
5.6. References . . . . .	206
<b>Chapter 6. Principles of Speech Recognition</b> . . . . .	213
Renato DE MORI and Brigitte BIGI	
6.1. Problem definition and approaches to the solution. . . . .	213
6.2. Hidden Markov models for acoustic modeling . . . . .	216
6.2.1. Definition. . . . .	216
6.2.2. Observation probability and model parameters . . . . .	217
6.2.3. HMM as probabilistic automata . . . . .	218
6.2.4. Forward and backward coefficients . . . . .	219
6.3. Observation probabilities. . . . .	222
6.4. Composition of speech unit models . . . . .	223
6.5. The Viterbi algorithm. . . . .	226
6.6. Language models . . . . .	228
6.6.1. Perplexity as an evaluation measure for language models . . . . .	230
6.6.2. Probability estimation in the language model . . . . .	232
6.6.3. Maximum likelihood estimation . . . . .	234
6.6.4. Bayesian estimation . . . . .	235
6.7. Conclusion . . . . .	236
6.8. References . . . . .	237

<b>Chapter 7. Speech Recognition Systems</b> . . . . .	239
Jean-Luc GAUVAIN and Lori LAMEL	
7.1. Introduction. . . . .	239
7.2. Linguistic model. . . . .	241
7.3. Lexical representation. . . . .	244
7.4. Acoustic modeling. . . . .	247
7.4.1. Feature extraction. . . . .	247
7.4.2. Acoustic-phonetic models. . . . .	249
7.4.3. Adaptation techniques . . . . .	253
7.5. Decoder . . . . .	256
7.6. Applicative aspects . . . . .	257
7.6.1. Efficiency: speed and memory . . . . .	257
7.6.2. Portability: languages and applications . . . . .	259
7.6.3. Confidence measures. . . . .	260
7.6.4. Beyond words . . . . .	261
7.7. Systems . . . . .	261
7.7.1. Text dictation . . . . .	262
7.7.2. Audio document indexing . . . . .	263
7.7.3. Dialog systems . . . . .	265
7.8. Perspectives . . . . .	268
7.9. References . . . . .	270
 <b>Chapter 8. Language Identification</b> . . . . .	279
Martine ADDA-DECKER	
8.1. Introduction. . . . .	279
8.2. Language characteristics . . . . .	281
8.3. Language identification by humans. . . . .	286
8.4. Language identification by machines. . . . .	287
8.4.1. LId tasks . . . . .	288
8.4.2. Performance measures . . . . .	288
8.4.3. Evaluation . . . . .	289
8.5. LId resources . . . . .	290
8.6. LId formulation . . . . .	295
8.7. LId modeling . . . . .	298
8.7.1. Acoustic front-end . . . . .	299
8.7.2. Acoustic language-specific modeling . . . . .	300
8.7.3. Parallel phone recognition. . . . .	302
8.7.4. Phonotactic modeling . . . . .	304
8.7.5. Back-end optimization . . . . .	309
8.8. Discussion . . . . .	309
8.9. References . . . . .	311

<b>Chapter 9. Automatic Speaker Recognition . . . . .</b>	<b>321</b>
Frédéric BIMBOT.	
9.1. Introduction. . . . .	321
9.1.1. Voice variability and characterization. . . . .	321
9.1.2. Speaker recognition . . . . .	323
9.2. Typology and operation of speaker recognition systems . . . . .	324
9.2.1. Speaker recognition tasks . . . . .	324
9.2.2. Operation. . . . .	325
9.2.3. Text-dependence . . . . .	326
9.2.4. Types of errors . . . . .	327
9.2.5. Influencing factors . . . . .	328
9.3. Fundamentals. . . . .	329
9.3.1. General structure of speaker recognition systems . . . . .	329
9.3.2. Acoustic analysis . . . . .	330
9.3.3. Probabilistic modeling. . . . .	331
9.3.4. Identification and verification scores . . . . .	335
9.3.5. Score compensation and decision . . . . .	337
9.3.6. From theory to practice . . . . .	342
9.4. Performance evaluation. . . . .	343
9.4.1. Error rate . . . . .	343
9.4.2. DET curve and EER . . . . .	344
9.4.3. Cost function, weighted error rate and HTER . . . . .	346
9.4.4. Distribution of errors . . . . .	346
9.4.5. Orders of magnitude . . . . .	347
9.5. Applications . . . . .	348
9.5.1. Physical access control. . . . .	348
9.5.2. Securing remote transactions . . . . .	349
9.5.3. Audio information indexing. . . . .	350
9.5.4. Education and entertainment . . . . .	350
9.5.5. Forensic applications. . . . .	351
9.5.6. Perspectives . . . . .	352
9.6. Conclusions. . . . .	352
9.7. Further reading. . . . .	353
<b>Chapter 10. Robust Recognition Methods . . . . .</b>	<b>355</b>
Jean-Paul HATON	
10.1. Introduction . . . . .	355
10.2. Signal pre-processing methods. . . . .	357
10.2.1. Spectral subtraction . . . . .	357
10.2.2. Adaptive noise cancellation . . . . .	358
10.2.3. Space transformation . . . . .	359
10.2.4. Channel equalization . . . . .	359
10.2.5. Stochastic models . . . . .	360
10.3. Robust parameters and distance measures . . . . .	360

10.3.1. Spectral representations . . . . .	361
10.3.2. Auditory models. . . . .	364
10.3.3 Distance measure . . . . .	365
10.4. Adaptation methods . . . . .	366
10.4.1 Model composition . . . . .	366
10.4.2. Statistical adaptation . . . . .	367
10.5. Compensation of the Lombard effect. . . . .	368
10.6. Missing data scheme. . . . .	369
10.7. Conclusion . . . . .	369
10.8. References. . . . .	370

## **Chapter 11. Multimodal Speech: Two or Three senses are Better than One . . . . .**

377

Jean-Luc SCHWARTZ, Pierre ESCUDIER and Pascal TEISSIER

11.1. Introduction . . . . .	377
11.2. Speech is a multimodal process . . . . .	379
11.2.1. Seeing without hearing. . . . .	379
11.2.2. Seeing for hearing better in noise. . . . .	380
11.2.3. Seeing for better hearing... even in the absence of noise. . . . .	382
11.2.4. Bimodal integration imposes itself to perception . . . . .	383
11.2.5. Lip reading as taking part to the ontogenesis of speech. . . . .	385
11.2.6. ...and to its phylogenesis ? . . . . .	386
11.3. Architectures for audio-visual fusion in speech perception . . . . .	388
11.3.1. Three paths for sensory interactions in cognitive psychology . . . . .	389
11.3.2. Three paths for sensor fusion in information processing . . . . .	390
11.3.3. The four basic architectures for audiovisual fusion . . . . .	391
11.3.4. Three questions for a taxonomy . . . . .	392
11.3.5. Control of the fusion process . . . . .	394
11.4. Audio-visual speech recognition systems . . . . .	396
11.4.1. Architectural alternatives . . . . .	397
11.4.2. Taking into account contextual information . . . . .	401
11.4.3. Pre-processing . . . . .	403
11.5. Conclusions . . . . .	405
11.6. References. . . . .	406

## **Chapter 12. Speech and Human-Computer Communication . . . . .**

417

Wolfgang MINKER &amp; Françoise NÉEL

12.1. Introduction . . . . .	417
12.2. Context. . . . .	418
12.2.1. The development of micro-electronics. . . . .	419
12.2.2. The expansion of information and communication technologies and increasing interconnection of computer systems . . . . .	420

- 12.2.3. The coordination of research efforts and the improvement of automatic speech processing systems . . . . . 421
- 12.3. Specificities of speech. . . . . 424
  - 12.3.1. Advantages of speech as a communication mode . . . . . 424
  - 12.3.2. Limitations of speech as a communication mode . . . . . 425
  - 12.3.3. Multidimensional analysis of commercial speech recognition products . . . . . 427
- 12.4. Application domains with voice-only interaction. . . . . 430
  - 12.4.1. Inspection, control and data acquisition . . . . . 431
  - 12.4.2. Home automation: electronic home assistant. . . . . 432
  - 12.4.3. Office automation: dictation and speech-to-text systems . . . . . 432
  - 12.4.4. Training. . . . . 435
  - 12.4.5. Automatic translation . . . . . 438
- 12.5. Application domains with multimodal interaction . . . . . 439
  - 12.5.1. Interactive terminals . . . . . 440
  - 12.5.2. Computer-aided graphic design. . . . . 441
  - 12.5.3. On-board applications . . . . . 442
  - 12.5.4. Human-human communication facilitation . . . . . 444
  - 12.5.5. Automatic indexing of audio-visual documents . . . . . 446
- 12.6. Conclusions . . . . . 446
- 12.7. References. . . . . 447

**Chapter 13. Voice Services in the Telecom Sector . . . . . 455**  
Laurent COURTOIS, Patrick BRISARD and Christian GAGNOULET

- 13.1. Introduction . . . . . 455
- 13.2. Automatic speech processing and telecommunications . . . . . 456
- 13.3. Speech coding in the telecommunication sector . . . . . 456
- 13.4. Voice command in telecom services . . . . . 457
  - 13.4.1. Advantages and limitations of voice command . . . . . 457
  - 13.4.2. Major trends . . . . . 459
  - 13.4.3. Major voice command services . . . . . 460
  - 13.4.4. Call center automation (operator assistance) . . . . . 460
  - 13.4.5. Personal voice phonebook . . . . . 462
  - 13.4.6. Voice personal telephone assistants . . . . . 463
  - 13.4.7. Other services based on voice command . . . . . 463
- 13.5. Speaker verification in telecom services . . . . . 464
- 13.6. Text-to-speech synthesis in telecommunication systems . . . . . 464
- 13.7. Conclusions . . . . . 465
- 13.8. References. . . . . 466

**List of Authors . . . . . 467**

**Index . . . . . 471**



# Chapter 1

## Speech Analysis

### 1.1. Introduction

#### 1.1.1. *Source-filter model*

Speech, the acoustic manifestation of language, is probably the main means of communication between human beings. The invention of telecommunications and the development of digital information processing have therefore entailed vast amounts of research aimed at understanding the mechanisms of speech communication.

Speech can be approached from different angles. In this chapter, we will consider speech as a signal, a one-dimensional function, which depends on the time variable (as in [BOI 87, OPP 89, PAR 86, RAB 75, RAB 77]). The acoustic speech signal is obtained at a given point in space by a sensor (microphone) and converted into electrical values. These values are denoted  $s(t)$  and they represent a real-valued function of real variable  $t$ , analogous to the variation of the acoustic pressure. Even if the acoustic form of the speech signal is the most widespread (it is the only signal transmitted over the telephone), other types of analysis also exist, based on alternative physiological signals (for instance, the electroglottographic signal, the palatographic signal, the airflow), or related to other modalities (for example, the image of the face or the gestures of the articulators). The field of speech analysis covers the set of methods aiming at the extraction of information on and from this signal, in various applications, such as:

---

Chapter written by Christophe D'ALESSANDRO.