

Signals and Communication Technology

For further volumes:
<http://www.springer.com/series/4748>

Qi (Peter) Li

Speaker Authentication

Dr. Qi (Peter) Li
Li Creative Technologies (LcT), Inc.
Vreeland Road 30 A, Suite 130
Florham Park, NJ 07932
USA
e-mail: li@licreativetech.com

ISSN 1860-4862

ISBN 978-3-642-23730-0

e-ISBN 978-3-642-23731-7

DOI 10.1007/978-3-642-23731-7

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011939406

© Springer-Verlag Berlin Heidelberg 2012

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: eStudio Calamar, Berlin/Figueras

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my parents YuanLin Shen and YanBin Li

Preface

My research on speaker authentication started in 1995 when I was an intern at Bell Laboratories, Murray Hill, New Jersey, USA, while working on my Ph.D. dissertation. Later, I was hired by Bell Labs as a Member of Technical Staff, which gave me the opportunity to continue my research on speaker authentication with my Bell Labs colleagues. In 2002, I established Li Creative Technologies, Inc. (LcT), located in Florham Park, New Jersey. At LcT, I am continuing my research in speaker authentication with my LcT colleagues. Recently, when I looked at my publications during the last fifteen years, I found that my research has covered all the major research topics in speaker authentication: from front-end to back-end; from endpoint detection to decoding; from feature extraction to discriminative training; from speaker recognition to verbal information verification. This has motivated me to put my research results together into a book in order to share my experience with my colleagues in the field.

This book is organized by research topic. Each chapter focuses on a major topic and can be read independently. Each chapter contains advanced algorithms along with real speech examples and evaluation results to validate the usefulness of the selected topics. Special attention has been given to the topics related to improving overall system robustness and performance, such as robust endpoint detection, fast discriminative training theory and algorithms, detection-based decoding, and sequential authentication. I have also given attention to those novel approaches that may lead to new research directions, such as a recently developed auditory transform (AT) to replace the fast Fourier transform (FFT) and auditory-based feature extraction algorithms.

For real applications, a good speaker authentication system must first have an acceptable authentication accuracy and then be robust to background noise, channel distortion, and speaker variability. A number of speaker authentication systems can be designed based on the methods and techniques presented in this book. A particular system can be designed to meet required specifications by selecting an authentication method or combining several authentication and decision methods introduced in the book.

Speaker authentication is a subject that relies on the research efforts of many different fields, including, but not limited to, physics, acoustics, psychology, physiology, hearing, auditory nerve, brain, auditory perception, parametric and nonparametric statistics, signal processing, pattern recognition, acoustic phonetics, linguistics, natural language processing, linear and nonlinear programming, optimization, communications, etc. This book only covers a subset of these topics. Due to my limited time and experience, this book only focuses on the topics in my published research. I encourage people with the above backgrounds to consider contributing their knowledge to speech recognition and speaker authentication research. I also encourage colleagues in the field of speech recognition and speaker authentication to extend their knowledge to the above fields in order to achieve breakthrough research results.

This book does not include those fundamental topics which have been very well introduced in other textbooks. This author assumes the reader has a basic understanding of linear systems, signal processing, statistics, and pattern recognition.

This book can also be used as a reference book for government and company officers and researchers working in information technology, homeland security, law enforcement, and information security, as well as for researchers and developers in the areas of speaker recognition, speech recognition, pattern recognition, and audio and signal processing. It can also be used as a reference or textbook for senior undergraduate and graduate students in electrical engineering, computer science, biomedical engineering, and information management.

Acknowledgments

The author would like to thank the many people who helped the author in his career and in the fields of speaker and speech recognition. I am particularly indebted to Dr. Donald W. Tufts at the Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, for his role in guiding and training me in pattern recognition and speech signal processing.

Special thanks are due to Dr. S. Parthasathy and Dr. Aaron Rosenberg who served as mentors when I first joined Bell Laboratories. They led me into the field of speaker verification research. I am particularly grateful to Dr. Biing-Hwang (Fred) Juang for his guidance in verbal information verification research. The work extended speaker recognition to speaker authentication, which has broader applications.

Most topics in this book were prepared based on previously published peer-reviewed journal and conference papers where I served as the first author. I would like to thank all the coauthors of those publications, namely Dr. Donald Tufts, Dr. Peter Swaszek, Dr. S. Parthasarathy, Dr. Aaron Rosenberg, Dr. Biing-Hwang Juang, Dr. Frank Soong, Dr. Chin-Hui Lee, Qiru Zhou, Jinsong Zheng, Dr. Augustine Tsai, and Yan Huang. Also, I would like to

thank the many anonymous reviewers and editors for their helpful comments and suggestions.

The author also would like to thank Dr. Bishnu Atal, Dr. Joe Olive, Dr. Wu Chou, Dr. Oliver Siohan, Dr. Mohan Sondhi, Dr. Oded Ghitza, Dr. Jingdong Chen, Dr. Rafid Sukkar, Dr. Larry O’Gorman, Dr. Richard Rose, and Dr. David Roe, all former Bell Laboratories colleagues, for their useful discussions and their kind help and support on my research there. Also, I would like to thank Dr. Ivan Selesnick for our recent collaborations.

Within Li Creative Technologies, the author would like to thank Yan Huang and Yan Yin for our recent collaborations in speaker identification research. I also would like to thank my colleagues Dr. Manli Zhu, Dr. Bozhao Tan, Uday Jain, and Joshua Hajicek for useful discussions on biometrics, acoustic, speech, and hearing systems.

From 2008 to 2010, the author’s research on speaker identification was supported by the U.S. AFRL under the contract number FA8750-08-C-0028. I would like to thank program managers Michelle Grieco, John Parker, and Dr. Stanly Wennedt for their help and support. Some of the research results have been included in Chapter 7 and Chapter 8 of this book. Other results will be published later.

I would like to thank my colleague Craig B. Adams and my daughter Joy Y Li for their work in editing this book. I would like to thank Uday Jain, Dr. Manli Zhu and Dr. Bozhao Tan for their proofreading. Also, this book could not have been finished without the support of my wife Vivian for the many weekends which I spent working on it.

The author also would like to thank the IEEE Intellectual Property Rights Office for permissions to use the IEEE copyright materials which I previously published in IEEE publications in the book.

Finally, I would like to thank Dr. Christoph Baumann, Engineering Editor at Springer, for his kind invitation to prepare and publish this book.

Florham Park, NJ

Qi (Peter) Li
July 2011

Contents

- 1 Introduction** 1
 - 1.1 Authentication 1
 - 1.2 Biometric-Based Authentication 3
 - 1.3 Information-Based Authentication 5
 - 1.4 Speaker Authentication 6
 - 1.4.1 Speaker Recognition 7
 - 1.4.2 Verbal Information Verification 9
 - 1.5 Historical Perspective and Further Reading 11
 - 1.6 Book Organization 13
 - References 18

- 2 Multivariate Statistical Analysis and One-Pass Vector Quantization** 23
 - 2.1 Multivariate Gaussian Distribution 23
 - 2.2 Principal Component Analysis 25
 - 2.3 Vector Quantization 27
 - 2.4 One-Pass VQ 28
 - 2.4.1 The One-Pass VQ Algorithm 28
 - 2.4.2 Steps of the One-Pass VQ Algorithm 32
 - 2.4.3 Complexity Analysis 33
 - 2.4.4 Codebook Design Examples 34
 - 2.4.5 Robustness Analysis 38
 - 2.5 Segmental K -Means 39
 - 2.6 Conclusions 40
 - References 40

- 3 Principal Feature Networks for Pattern Recognition** 43
 - 3.1 Overview of the Design Concept 43
 - 3.2 Implementations of Principal Feature Networks 46
 - 3.3 Hidden Node Design 48
 - 3.3.1 Gaussian Discriminant Node 49

3.3.2	Fisher’s Node Design	51
3.4	Principal Component Hidden Node Design	52
3.4.1	Principal Component Discriminant Analysis	53
3.5	Relation between PC Node and the Optimal Gaussian Classifier	53
3.6	Maximum Signal-to-Noise-Ratio (SNR) Hidden Node Design ..	55
3.7	Determining the Thresholds from Design Specifications	56
3.8	Simplification of the Hidden Nodes	56
3.9	Application 1 – Data Recognition	56
3.10	Application 2 – Multispectral Pattern Recognition	58
3.11	Conclusions	59
	References	59
4	Non-Stationary Pattern Recognition	61
4.1	Introduction	61
4.2	Gaussian Mixture Models (GMM) for Stationary Process	62
4.2.1	An Illustrative Example	63
4.3	Hidden Markov Model (HMM) for Non-Stationary Process	66
4.4	Speech Segmentation	68
4.5	Bayesian Decision Theory	68
4.6	Statistical Verification	70
4.7	Conclusions	71
	References	72
5	Robust Endpoint Detection	75
5.1	Introduction	76
5.2	A Filter for Endpoint Detection	78
5.3	Real-Time Endpoint Detection and Energy Normalization	81
5.3.1	A Filter for Both Beginning- and Ending-Edge Detection	82
5.3.2	Decision Diagram	82
5.3.3	Real-Time Energy Normalization	83
5.3.4	Database Evaluation	85
5.4	Conclusions	88
	References	89
6	Detection-Based Decoder	93
6.1	Introduction	94
6.2	Change-Point Detection	96
6.3	HMM State Change-Point Detection	97
6.4	HMM Search-Space Reduction	100
6.4.1	Concept of Search-Space Reduction	100
6.4.2	Algorithm Summary and Complexity Analysis	102
6.5	Experiments	104
6.5.1	An Example of State Change-Point Detection	104
6.5.2	Application to Speaker Verification	104
6.6	Conclusions	108

References	109
7 Auditory-Based Time Frequency Transform	111
7.1 Introduction	112
7.1.1 Observing Problems with the Fourier Transform	112
7.1.2 Brief Introduction of the Ear	114
7.1.3 Time-Frequency Analyses	117
7.2 Definition of the Auditory-Based Transform	118
7.3 The Inverse Auditory Transform	120
7.4 The Discrete-Time and Fast Transform	123
7.5 Experiments and Discussions	124
7.5.1 Verifying the Inverse Auditory Transform	124
7.5.2 Applications	126
7.6 Comparisons to Other Transforms	127
7.7 Conclusions	131
References	131
8 Auditory-Based Feature Extraction and Robust Speaker Identification	135
8.1 Introduction	135
8.2 Auditory-Based Feature Extraction Algorithm	138
8.2.1 Forward Auditory Transform and Cochlea Filter Bank	138
8.2.2 Cochlear filter cepstral coefficients (CFCC)	140
8.2.3 Analysis and Comparison	141
8.3 Speaker Identification and Experimental Evaluation	142
8.3.1 Experimental Datasets	142
8.3.2 The Baseline Speaker Identification System	143
8.3.3 Experiments	145
8.3.4 Further Comparison with PLP and RASTA-PLP	146
8.4 Conclusions	147
References	149
9 Fixed-Phrase Speaker Verification	151
9.1 Introduction	151
9.2 A Fixed-Phrase System	152
9.3 An Evaluation Database and Model Parameters	154
9.4 Adaptation and Reference Results	155
9.5 Conclusions	155
References	156
10 Robust Speaker Verification with Stochastic Matching	157
10.1 Introduction	157
10.2 A Fast Stochastic Matching Algorithm	158
10.3 Fast Estimation for a General Linear Transform	160
10.4 Speaker Verification with Stochastic Matching	161

10.5 Database and Experiments	163
10.6 Conclusions	164
References	164
11 Randomly Prompted Speaker Verification	165
11.1 Introduction	165
11.2 Normalized Discriminant Analysis	168
11.3 Applying NDA in the Hybrid Speaker-Verification System	169
11.3.1 Training of the NDA System	169
11.3.2 Training of the HMM System	171
11.3.3 Training of the Data Fusion Layer	173
11.4 Speaker Verification Experiments	173
11.4.1 Experimental Database	173
11.4.2 NDA System Results	174
11.4.3 Hybrid Speaker-Verification System Results	174
11.5 Conclusions	175
References	176
12 Objectives for Discriminative Training	179
12.1 Introduction	179
12.2 Error Rates vs. Posterior Probability	180
12.3 Minimum Classification Error vs. Posterior Probability	181
12.4 Maximum Mutual Information vs. Minimum Classification Error	183
12.5 Generalized Minimum Error Rate vs. Other Objectives	185
12.6 Experimental Comparisons	186
12.7 Discussion	186
12.8 Relations between Objectives and Optimization Algorithms	187
12.9 Conclusions	188
References	189
13 Fast Discriminative Training	191
13.1 Introduction	191
13.2 Objective for Fast Discriminative Training	193
13.3 Derivation of Fast Estimation Formulas	195
13.3.1 Estimation of Covariance Matrices	196
13.3.2 Determination of Weighting Scalar	196
13.3.3 Estimation of Mean Vectors	197
13.3.4 Estimation of Mixture Parameters	198
13.3.5 Discussions	198
13.4 Summary of Practical Training Procedure	200
13.5 Experiments	200
13.5.1 Continuing the Illustrative Example	200
13.5.2 Application to Speaker Identification	204
13.6 Conclusions	204

References 205

14 Verbal Information Verification 207

14.1 Introduction 207

14.2 Single Utterance Verification 209

 14.2.1 Normalized Confidence Measures 211

14.3 Sequential Utterance Verification 212

 14.3.1 Examples in Sequential-Test Design 214

14.4 VIV Experimental Results 215

14.5 Conclusions 219

References 220

15 Speaker Authentication System Design 223

15.1 Introduction 223

15.2 Automatic Enrollment by VIV 224

15.3 Fixed-Phrase Speaker Verification 226

15.4 Experiments 227

 15.4.1 Features and Database 227

 15.4.2 Experimental Results on Using VIV for SV Enrollment 228

15.5 Conclusions 229

References 230

Index 231

List of Tables

1.1	List of Biometric Authentication Error Rates	5
2.1	Quantizer MSE Performance	35
2.2	Comparison of One-Pass and LBG Algorithms	36
2.3	Comparison of Different VQ Design Approaches	37
2.4	Comparison for the Correlated Gaussian Source	37
2.5	Comparison on the Laplace Source	38
3.1	Comparison of Three Algorithms in the Land Cover Recognition	58
5.1	Database Evaluation Results (%)	88
7.1	Correlation Coefficients, σ_{12}^2 , for Different Sizes of Filter Bank in AT/inverse AT	126
8.1	Summary of The training, Development, and Testing Set.	143
8.2	Comparison of MFCC, MGFCC, and CFCC Features Tested on the Development Tet.	144
9.1	Experimental Results in Average Equal-Error Rates of All Tested Speakers	155
10.1	Experimental Results in Average Equal-Error Rates (%)	163
11.1	Segmentation of the Database	174
11.2	Results on Discriminant Analysis	175
11.3	Major Results	175
12.1	Comparisons on Training Algorithms	187
13.1	Three-Class Classification Results of the Illustration Example ..	203
13.2	Comparison on Speaker Identification Error Rates	204

14.1 False Acceptance Rates when Using Two Thresholds and Maintaining False Rejection Rates to Be 0.0%.....	216
14.2 Comparison on Two and Single Threshold Tests.....	216
14.3 Summary of the Experimental Results on Verbal Information Verification.....	219
15.1 Experimental Results without Adaptation in Average Equal-Error Rates.....	229
15.2 Experimental Results with Adaptation in Average Equal-Error Rates.....	229

List of Figures

1.1	Speaker authentication approaches.	6
1.2	A speaker verification system.	7
1.3	An example of verbal information verification by asking sequential questions. Similar sequential tests can also be applied in speaker recognition and other biometric or multi-modality verification.	10
2.1	An example of bivariate Gaussian distribution: $\rho_{11} = 1.23$, $\rho_{12} = \rho_{21} = 0.45$, and $\rho_{22} = 0.89$	24
2.2	The contour of the Gaussian distribution in Fig. 2.1.	25
2.3	An illustration of a constant density ellipse and the principal components for a normal random vector \mathbf{X} . The largest eigenvalue associates with the long axis of the ellipse and the second eigenvalue associates with the short axis. The eigenvectors associate with the axes.	26
2.4	The method to determine a code vector: (a) select the highest density cell; (b) examine a group of cells around the selected one; (c) estimate the center of the data subset; (d) cut a “hole” in the training data set.	29
2.5	The Principal Component (PC) method to determine a centroid. 31	
2.6	Left: Uncorrelated Gaussian source training data. Right: The residual data after four code vectors have been located.	35
2.7	Left: The residual data after all 16 code vectors have been located. Right: The “+” and “o” are the centroids after one and three iterations of the LBG algorithm, respectively.	36
2.8	Left: The Laplace source training data. Right: The residual data, one-pass designed centroids “+”, and one-pass+2LBG centroids “o”.	38

3.1	An illustrative example to demonstrate the concept of the PFN: (a) The original training data of two labeled classes which are not linearly separable. (b) The hyperplanes of the first hidden node (LDA node). (c) The residual data set and the hyperplanes of the second hidden node (SNR node). (d) The input space partitioned by two hidden nodes and four thresholds designed by the principal feature classification (PFC) method.	45
3.2	A parallel implementation of PFC by a Principal Feature Network (PFN).	47
3.3	A sequential implementation of PFC by a Principal Feature Tree (PFT).	47
3.4	(a) Partitioned input space for parallel implementation. (b) Parallel implementation.	48
3.5	(a) Partitioned input space for sequential implementation. (b) Sequential (tree) implementation.	49
3.6	(a) A single Gaussian discriminant node. (b) A Fisher's node. (c) A quadratic node. (d) An approximation of the quadratic node.	50
3.7	When using only Fisher's nodes, three hidden nodes and six thresholds are needed to finish the design.	54
3.8	Application 1: (a) (bottom) The sorted contribution of each threshold in the order of its contribution to the class separated by the threshold. (b) (top) Accumulated network performance in the order of the sorted thresholds.	57
4.1	Class 1: a bivariate Gaussian distribution with $m_1 = [0 \ 5]$, $m_2 = [-3 \ 3]$, and $m_3 = [-5 \ 0]$. $\Sigma_1 = [1.41 \ 0; 0 \ 1.41]$, $\Sigma_2 = [1.22 \ 0.09; 0.09 \ 1.22]$, and $\Sigma_3 = [1.37 \ 0.37; 0.27 \ 1.37]$	64
4.2	Class 2: a bivariate Gaussian distribution with $m_1 = [2 \ 5]$, $m_2 = [-1 \ 3]$, and $m_3 = [0 \ 0]$. $\Sigma_1 = [1.41 \ 0; 0 \ 1.41]$, $\Sigma_2 = [0.77 \ 1.11; 1.11 \ 1.09]$, and $\Sigma_3 = [1.41 \ 0.04; 0.04 \ 1.41]$	64
4.3	Class 3: a bivariate Gaussian distribution with $m_1 = [-3 \ -1]$, $m_2 = [-2 \ -2]$, and $m_3 = [-5 \ -2]$. $\Sigma_1 = [1.41 \ 0; 0 \ 1.41]$, $\Sigma_2 = [0.76 \ 0.11; 0.11 \ 1.09]$, and $\Sigma_3 = [1.41 \ 0.04; 0.04 \ 1.41]$	65
4.4	Contours of the <i>pdf</i> 's of 3-mixture GMM's: the models are used to generate 3 classes of training data.	65
4.5	Contours of the <i>pdf</i> 's of 2-mixture GMM's: the models are trained from ML estimation using 4 iterations.	66
4.6	Enlarged decision boundaries for the ideal 3-mixture models (solid line) and 2-mixture ML models (dashed line).	66
4.7	Left-to-right hidden Markov model.	67
5.1	Shape of the designed optimal filter.	81
5.2	Endpoint detection and energy normalization for real-time ASR.	81

5.3	State transition diagram for endpoint decision.	83
5.4	Example: (A) Energy contour of digit “4”. (B) Filter outputs and state transitions.	84
5.5	(A) Energy contours of “4-327-631-Z214” from original utterance (bottom, 20 dB SNR) and after adding car noise (top, 5 dB SNR). (B) Filter outputs for 5 dB (dashed line) and 20 dB (solid line) SNR cases. (C) Detected endpoints and normalized energy for the 20 dB SNR case, and (D) for the 5 dB SNR case.	85
5.6	Comparisons on real-time connected digit recognition with various signal-to-noise ratios (SNR’s). From 5 to 20 dB SNR’s, the introduced real-time algorithm provided word error rate reductions of 90.2%, 93.4%, 57.1%, and 57.1%, respectively.	87
5.7	(A) Energy contour of the 523th utterance in DB5: “1 Z 4 O 5 8 2”. (B) Endpoints and normalized energy from the baseline system. The utterance was recognized as “1 Z 4 O 5 8”. (C) Endpoints and normalized energy from the real-time, endpoint-detection system. The utterance was recognized correctly as “1 Z 4 O 5 8 2”. (D) The filter output.	89
6.1	The scheme of the change-point detection algorithm with $t_\delta = 2$: (a) the endpoint detection for state 1; (b) the endpoint detection for state 2; and (c) the grid points involved in p_1 , p_2 and p_3 computations (dots).	98
6.2	Left-to-right hidden Markov model.	98
6.3	All the grid points construct a full search space Ψ . The grid points involved in the change-point detection are marked as black points. A single path (solid line) is detected from the forward and backward change-point detection.	102
6.4	A “hole” is detected from the forward and backward state change-point detection. A search is needed only among four grid points, (8,3), (8,4), (9,3) and (9,4). The solid line indicates the path with the maximum likelihood score.	102
6.5	A search is needed in the reduced search space Ω which includes all the black points in between the two dashed lines. The points along the dashed lines are involved in change-point detection, but they do not belong to the reduced search space.	103
6.6	A special case is located between (11,4) and (18,6), where the forward boundary is under the backward one. A full search can be done in the subspace $\{(t, s_t) \mid 11 \leq t \leq 18; 4 < s_t < 6\}$	103
6.7	The procedure of sequential state change-point detection from state 1 (top) to state 7 (bottom), where the vertical dashed lines are the detected endpoints of each state.	105
6.8	The procedure of sequential state change-point detection from state 8 (top) to state 13 (bottom).	106

6.9 (a) Comparison of average individual equal-error rates (EER's); (b) Comparison on average speedups. 107

7.1 Male's voice: "2 0 5" recorded simultaneously by close-talking (top) and hands-free microphones in a moving car (bottom). 112

7.2 The speech waveforms in Fig. 7.1 were converted to spectrograms by FFT and displayed in Bark scale from 0 to 16.4 Barks (0 to 3500 KHz). The background noise and the pitch harmonics were generated mainly by FFT. 113

7.3 The spectrum of FFT at the 1.15 second time frame from Fig. 7.2: The solid line represents the speech from a close-talking microphone. The dashed line is from a hands-free microphone mounted on the visor of a moving car. Both speech files were recorded simultaneously. The FFT spectrum shows 30 dB distortion at low frequency bands due to background noise and pitch harmonics as noise. 114

7.4 Illustration of human ear and cochlea. 115

7.5 Illustration of a stretched out cochlea and a traveling wave exciting a portion of the basilar membrane. 115

7.6 Impulse responses of the BM in the AT when $\alpha = 3$ and $\beta = 0.2$. They are very similar to the research results reported in hearing research. 121

7.7 The frequency responses of the cochlear filters when $\alpha = 3$: (A) $\beta = 0.2$; and (B) $\beta = 0.035$ 122

7.8 The traveling wave generated by the auditory transform from the speech data in Fig. 7.1. 123

7.9 A section of the traveling wave generated by the auditory transform. 124

7.10 Spectrograms from the output of the cochlear transform for the speech data in Fig. 7.1 respectively. The spectrogram at top is from the data recorded by the close-talking microphone, while the spectrogram at bottom is from the hands-free microphone. 125

7.11 The spectrum of AT at the 1.15 second time frame from Fig. 7.10: The solid line represents the speech from a close-talking microphone. The dashed line is from a hands-free microphone mounted on the visor of a moving car. Both speech files were recorded simultaneously. 125

7.12 Comparison of speech waveforms: (A) The original waveform of a male voice speaking the words "two, zero, five." (B) The synthesized waveform by inverse AT with the bandwidth of 80 to 5K Hz. When the filter numbers are 8, 16, 32, and 64, the correlation coefficients σ_{12}^2 for the two speech data sets are 0.74, 0.96, 0.99, and 0.99, respectively. 126

7.13	(A) and (B) are speech waveforms simultaneously recorded in a moving car. The microphones are located on the car visor (A) and speaker's lapel (B), respectively. (C) is after noise reduction using the AT from the waveform in (A), where results are very similar to (B).	127
7.14	Comparison of FT and AT spectrums: (A) The FFT spectrogram of a male voice "2 0 5", warped into the Bark scale from 0 to 6.4 Barks (0 to 3500 KHz). (B) The spectrogram from the cochlear filter output for the same male voice. The AT is harmonic free and has less computational noise.	128
7.15	Comparison of AT (top) and FFT (bottom) spectrums at the 1.15 second time frame for robustness: The solid line represents speech from a close-talking microphone. The dashed line represents speech from a hands-free microphone mounted on the visor of a moving car. Both speech files were recorded simultaneously. The FFT spectrum shows 30 dB distortion at low-frequency bands due to background noise compared to the AT. Compared to the FFT spectrum, the AT spectrum has no pitch harmonics and much less distortion at low frequency bands due to background noise.	129
7.16	The Gammatone filter bank: (A) The frequency responses of the Gammatone filter bank generated by (7.19). (B) The frequency responses of the Gammatone filter bank generated by (7.19) plus a equal loudness function.	130
8.1	Schematic diagram of the auditory-based feature extraction algorithm named cochlear filter cepstral coefficients (CFCC). . .	138
8.2	Comparison of MFCC, MGFCC, and the CFCC features tested on noisy speech with white noise.	145
8.3	Comparison of MFCC, MGFCC, and CFCC features tested on noisy speech with car noise.	146
8.4	Comparison of MFCC, MGFCC, and CFCC features tested on noisy speech with babble noise.	147
8.5	Comparison of PLP, RASTA-PLP, and the CFCC features tested on noisy speech with white noise.	148
8.6	Comparison of PLP, RASTA-PLP, and the CFCC features tested on noisy speech with car noise.	148
8.7	Comparison of PLP, RASTA-PLP, and the CFCC features tested on noisy speech with babble noise.	149
9.1	A fixed-phrase speaker verification system.	153

10.1	A geometric interpretation of the fast stochastic matching. (a) The dashed line is the contour of training data. (b) The solid line is the contour of test data. The crosses are the means of the two data sets. (c) The test data were scaled and rotated toward the training data. (d) The test data were translated to the same location as the training data. Both contours overlap each other.	159
10.2	A phrase-based speaker verification system with stochastic matching.	162
11.1	The structure of a hybrid speaker verification (HSV) system.	167
11.2	The NDA feature extraction.	170
11.3	The Type 2 classifier (NDA system) for one speaker.	171
13.1	Contours of the <i>pdf</i> 's of 3-mixture GMM's: the models are used to generate three classes of training data.	201
13.2	Contours of the <i>pdf</i> 's of 2-mixture GMM's: the models are from ML estimation using four iterations.	201
13.3	Contours of the <i>pdf</i> 's of 2-mixture GMM's: The models are from the fast GMER estimation with two iterations on top of the ML estimation results. The overlaps among the three classes are significantly reduced.	202
13.4	Enlarged decision boundaries for the ideal 3-mixture models (solid line), 2-mixture ML models (dashed line), and 2-mixture GMER models (dash-dotted line): After GMER training, the boundary of ML estimation shifted toward the decision boundary of the ideal models. This illustrates how GMER training improves decision accuracies.	202
13.5	Performance improvement versus iterations using the GMER estimation: The initial performances were from the ML estimation with four iterations.	203
14.1	An example of verbal information verification by asking sequential questions. (Similar sequential tests can also be applied in speaker verification and other biometric or multi-modality verification.)	208
14.2	Utterance verification in VIV.	210
14.3	False acceptance rate as a function of robust interval with SD threshold for a 0% false rejection rate. The horizontal axis indicates the shifts of the values of the robust interval τ	218
14.4	An enlarged graph of the system performances using two and three questions.	219
15.1	A conventional speaker verification system	224

15.2	An example of speaker authentication system design: Combining verbal information verification with speaker verification	225
15.3	A fixed-phrase speaker verification system	226