Incorporating Knowledge Sources into Statistical Speech Recognition

Lecture Notes in Electrical Engineering

Incorporating Knowledge Sources into Statistical Speech Recognition
Sakti, Sakriani, Markov, Konstantin, Nakamura, Satoshi, and Minker, Wolfgang
978-0-387-85829-6

Intelligent Technical Systems
Martínez Madrid, Natividad; Seepold, Ralf E.D. (Eds.)
978-1-4020-9822-2

Languages for Embedded Systems and their Applications
Radetzki, Martin (Ed.)
978-1-4020-9713-3

Multisensor Fusion and Integration for Intelligent Systems
Lee, Sukhan; Ko, Hanseok; Hahn, Hernsoo (Eds.)
978-3-540-89858-0

Designing Reliable and Efficient Networks on Chips
Murali, Srinivasan
978-1-4020-9756-0

Trends in Communication Technologies and Engineering Science
Ao, Sio-Iong; Huang, Xu; Wai, Ping-kong Alexander (Eds.)
978-1-4020-9492-7

Functional Design Errors in Digital Circuits: Diagnosis Correction and Repair
Chang, Kai-hui, Markov, Igor, Bertacco, Valeria
978-1-4020-9364-7

Traffic and QoS Management in Wireless Multimedia Networks: COST 290 Final Report
Koucheryavy, Y., Giambene, G., Staehle, D., Barcelo-Arroyo, F., Braun, T., Siris, V. (Eds.)
978-0-387-85572-1

Proceedings of the 3rd European Conference on Computer Network Defense
Siris, V.; Ioannidis, S.; Anagnostakis, K.; Trimintzios, P. (Eds.)
978-0-387-85554-7

Intelligentized Methodology for Arc Welding Dynamical Processes: Visual Information Acquiring, Knowledge Modeling and Intelligent Control
Chen, Shan-Ben, Wu, Jing
978-3-540-85641-2

Proceedings of the European Computing Conference: Volume 2
Mastorakis, Nikos, Mladenov, Valeri, Kontargyri, Vassiliki T. (Eds.)
978-0-387-84818-1

Proceedings of the European Computing Conference: Volume 1
Mastorakis, Nikos, Mladenov, Valeri, Kontargyri, Vassiliki T. (Eds.)
978-0-387-84813-6

Electronics System Design Techniques for Safety Critical Applications
Sterpone, Luca
978-1-4020-8978-7

Data Mining and Applications in Genomics
Ao, Sio-Iong
978-1-4020-8974-9

Continued after index

Sakriani Sakti • Konstantin Markov •
Satoshi Nakamura • Wolfgang Minker

Incorporating Knowledge Sources into Statistical Speech Recognition

Springer

Sakriani Sakti
NICT/ATR Spoken Language
Communication Research Laboratories
Keihanna Science City
Kyoto, Japan

Konstantin Markov
NICT/ATR Spoken Language
Communication Research Laboratories
Keihanna Science City
Kyoto, Japan

Satoshi Nakamura
NICT / ATR Spoken Language
Communication Research Laboratories
Keihanna Science City
Kyoto, Japan

Wolfgang Minker
University of Ulm
Ulm, Germany

ISBN 978-0-387-85829-6 e-ISBN 978-0-387-85830-2
DOI: 10.1007/978-0-387-85830-2

Library of Congress Control Number: 2008942803

© Springer Science+Business Media, LLC 2009
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.
While the advice and information in this book are believed to be true and accurate at the date of going to press, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper.

springer.com

This book is dedicated
to our parents and families
for their support and endless love

Preface

State-of-the-art automatic speech recognition (ASR) systems use statistical data-driven methods based on hidden Markov models (HMMs). Although such approaches have proved to be efficient choices, ASR systems often perform much worse than human listeners, especially in the presence of unexpected acoustic variability. To improve performance, we usually rely on collecting more data to train more detailed models. However, such resources are rarely available, since the presence of variabilities in speech arise from many different factors, and thus a huge amount of training data is required to cover all possible variabilities. In other words, it is not enough to handle these variabilities by relying solely on statistical models. The systems need additional knowledge on speech that could help to handle these sources of variability. Otherwise, only a limited level of success could be achieved.

Many researchers are aware of this problem, and thus various attempts to integrate more explicitly knowledge-based and statistical approaches have been made. However, incorporating various additional knowledge sources often leads to a complicated model, where achieving optimal performance is not feasible due to insufficient resources or data sparseness. As a result, input space resolution may be lost due to non-robust estimates and the increased number of unseen patterns. Moreover, decoding with large models may also become cumbersome and sometimes even impossible.

This book addresses the problem of developing efficient ASR systems that can maintain a balance between utilizing wide-ranging knowledge of speech variability while keeping the training/recognition effort feasible, of course while also improving speech recognition performance. In this book, an efficient general framework to incorporate additional knowledge sources into state-of-the-art statistical ASR systems is provided. It can be applied to many existing ASR problems with their respective model-based likelihood functions in flexible ways.

Since there are various types of knowledge sources from different domains, it may be difficult to formulate a probabilistic model without learning the dependencies between the sources. To solve such problems in a unified way, the work reported in this book adopts the Bayesian network (BN) framework. This approach allows the probabilistic relationship between information sources to be learned. Another advantage of the BN framework lies in the fact that it facilitates the decomposition of the joint probability density function (PDF) into a linked set of local conditional PDFs based on the junction tree algorithm. Consequently, a simplified form of the model can be constructed and reliably estimated using a limited amount of training data.

This book focuses on the acoustic modeling problem as arguably the central part of any speech recognition system. The incorporation of various knowledge sources, including background noises, accent, gender and wide phonetic knowledge information, in modeling is also discusses. Such an application often suffers from a sparseness of data and memory constraints. First, the additional sources of knowledge are incorporated at the HMM state distribution. Then, these additional sources of knowledge are incorporated at the HMM phonetic modeling. The presented approaches are experimentally verified in the large-vocabulary continuous-speech recognition (LVCSR) task. The book closes with a summary of the described methods and the results of the evaluations.

Contents

List of Figures

List of Tables

Glossary

AM	Acoustic model
ARPA	Advanced Research Projects Agency
ASR	Automatic speech recognition
A-STAR	Asian speech translation advanced research
ATR	Advanced Telecommunication Research
AUS	Australian
BN	Bayesian network
BRT	British
BTEC	Basic travel expression corpus
BU	Boston University
C1	Center monophone unit
C3	Center triphone context
Csk3	Center skip-triphone context
C5	Center pentaphone context
CCCC	CSR corpus coordinating committee
CNRS-LIMSI	France's National Center for Scientific Research
CPD	Conditional probability distribution
CPT	Conditional probability table
CSR	Continuous speech recognition
C-STAR	Consortium for speech translation advanced research
CU	Cambridge University
DAG	Directed acyclic graph
DARPA	Defense Advanced Research Projects Agency
DBN	Dynamic Bayesian network
DCT	Discrete cosine transform
DEL	Deletions
DI	Deleted interpolation
DSR	Distributed speech recognition
EDB	English database
ELRA	European language resources association

EM	Expectation-maximization
EPPS	European Parliament Plenary Sessions
fLRC-HMM/BN	Full HMM/BN for left, right and center state
fLRCA-HMM/BN	Full HMM/BN for left, right and center state, including accent dependency
fLRCAG-HMM/BN	Full HMM/BN for left, right and center state, including accent and gender dependency
fLRG-HMM/BN	Full HMM/BN for left, right and center state, including gender dependency
FFT	Fast Fourier transform
GDHMM	Gender-dependent Hidden Markov model
GFIKS	Graphical framework to incorporate additional knowledge sources
GIHMM	Gender-independent Hidden Markov model
GMM	Gaussian mixture model
HMM	Hidden Markov model
ICASSP	International conference on acoustics, speech and signal processing
ICSI	International Computer Science Institute
ICSLP	International conference on spoken language processing
IEEE	Institute of Electrical and Electronics Engineers
IEICE	Institute of Electronics, Information and Communication Engineers
Imp	Improvement
INS	Insertions
L3	Left triphone context
L4	Left tetraphone context
LM	Language model
LPC	Linear prediction coefficients
LRC-HMM/BN	HMM/BN for left, right and center state
LR-HMM/BN	HMM/BN for left and right state
Lsk3	Left skip-triphone context
LVCSR	Large-vocabulary continuous-speech recognition
MAD	Machine translation aided dialogue
MAP	Maximum *a posteriori*
MDL	Minimum description length
MFCC	Mel-frequency cepstral coefficients
MIT	Massachusetts Institute of Technology
ML	Maximum likelihood
MLLR	Maximum likelihood linear regression
MSG	Modulation-filtered spectrogram
MT	Machine translation
NIST	National Institute of Standards and Technology
NOVO	Noise voice composition
PDF	Probability density function

PLP	Perceptual linear prediction
PMC	Parallel model combination
R3	Right triphone context
R4	Right tetraphone context
Rel	Relative
Resc	Rescoring
Rsk3	Right skip-triphone context
S2ST	Speech-to-speech translation
SD	Speaker dependent
SI	Speaker independent
SIL	Silence
SLC	Spoken Language Communication
SNR	Signal-to-noise ratio
SSS	Successive state splitting
STQ	Speech processing, transmission and quality
SUB	Substitutions
SWB	Switchboard
TC-STAR	Technology and corpora for speech to speech translation research
TI	Texas Instrument
US	United States
VQ	Vector quantization
WER	Word error rate
WFST	Weighted finite state transducers
WSJ	Wall Street journal