

Lecture Notes in Artificial Intelligence 2560

Subseries of Lecture Notes in Computer Science
Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science
Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer

Berlin

Heidelberg

New York

Barcelona

Hong Kong

London

Milan

Paris

Tokyo

Silke Goronzy

Robust Adaptation to Non-Native Accents in Automatic Speech Recognition



Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Author

Silke Goronzy
Sony International (Europe) GmbH, SCLE, MMI Lab
Heinrich-Hertz-Straße 1, 70327 Stuttgart, Germany
E-mail: goronzy@sony.de

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek.
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data is available in the Internet at <<http://dnb.ddb.de>>.

CR Subject Classification (1998): I.2.7, I.2, J.5, H.5.2, F.4.2

ISSN 0302-9743

ISBN 3-540-00325-8 Springer-Verlag Berlin Heidelberg New York

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, re-use of illustrations, recitation, broadcasting, reproduction on microfilms 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-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag Berlin Heidelberg New York,
a member of BertelsmannSpringer Science+Business Media GmbH

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2002
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Boller Mediendesign
Printed on acid-free paper SPIN: 10871801 06/3142 5 4 3 2 1 0

Preface

Speech recognition technology is being increasingly employed in human-machine interfaces. Two of the key problems affecting such technology, however, are its robustness across different speakers and robustness to non-native accents, both of which still create considerable difficulties for current systems.

In this book methods to overcome these problems are described. A speaker adaptation algorithm that is based on Maximum Likelihood Linear Regression (MLLR) and that is capable of adapting the acoustic models to the current speaker with just a few words of speaker specific data is developed and combined with confidence measures that focus on phone durations as well as on acoustic features to yield a semi-supervised adaptation approach. Furthermore, a specific pronunciation modelling technique that allows the automatic derivation of non-native pronunciations without using non-native data is described and combined with the confidence measures and speaker adaptation techniques to produce a robust adaptation to non-native accents in an automatic speech recognition system.

The aim of this book is to present the state of the art in speaker adaptation, confidence measures and pronunciation modelling, as well as to show how these techniques have been improved and integrated to yield a system that is robust to varying speakers and non-native accents.

The faculty of electrical engineering of the Technical University Carolo-Wilhelmina of Braunschweig has accepted this book as a dissertation and at this point I would like to take the opportunity to thank all those who supported me during this research work. First of all I would like to thank Prof. Dr.-Ing. Erwin Paulus for his valuable suggestions and his support. Likewise I would like to thank Prof. Dr.-Ing. Günther Ruske from the Technical University of Munich for giving the second opinion.

The research presented in this book was conducted while I was working at Sony in Stuttgart and I would like to express my gratitude that I got the permission to conduct and publish this research work.

I am also indebted to my colleagues, who supported me a lot by being very cooperative and helpful. I would like especially to thank Dr. Krzysztof Marasek and Andreas Haag for their co-operation in the field of confidence measures, Dr. Stefan Rapp for the valuable discussions and Dipl. Ling. Manya

Sahakyan for conducting the accompanying experiments in the area of pronunciation modelling.

Furthermore I would like to thank Dr. Elmar Noeth, Dr. Richard Stirling-Gallacher, Dr. Franck Giron and particularly Dr. Roland Kuhn for their many very helpful comments, which played an important part in improving this book.

Special thanks go to Dr. Ralf Kompe, whose support was an important contribution to this book and who, in spite of his many other duties, always found time to support me.

Most especially, I would like to thank my dear parents, Waldemar and Renate Goronzy, without whose affectionate upbringing my entire education and this book would not have been possible and who provided me with a set of values I could not have gained through any other academic education process.

Fellbach, September 2001

Silke Goronzy

Foreword

The present state of the art in automatic speech recognition - under certain conditions - permits man-machine-communication by means of spoken language. Provided that speech recognition is tuned to the common native language (target language) of the users, speaker-independent recognition of words from a small vocabulary is feasible, especially if the words are spoken in isolation, and for larger vocabularies at least speaker-dependent recognition performance is satisfactory. The most elaborate up-to-date speech recognition systems manage large vocabularies even in speaker-independent connected speech recognition. However, perfect zero error rates cannot be achieved, and therefore such systems can be used only in applications that to some degree are fault tolerant, e.g. within dialog systems that offer appropriate feedback to the user and allow him to correct recognition errors in a convenient way.

Generally, error rates for speaker-dependent recognition are lower than for speaker-independent recognition, and error rates between these limits are achieved by systems that by default are preset to speaker-independent recognition and can be tailored to a certain speaker by means of a more or less demanding adaptation procedure. For adaptation purposes the prospective user of the system is required to produce a rather large number of certain prescribed utterances. Although some speaker-adaptive systems of this kind are already available, it is still a matter of research, as to how the required number of utterances can be reduced, in order to make the adaptation procedure less demanding for the prospective user. However, at several places more ambitious research work is directed towards speech recognition systems that continuously adapt to the current speaker without requiring a separate adaptation phase at each change between speakers. The work that is presented here is a major step towards such a system and - due to a remarkable new approach - adaptation can be successful even in the case of a non-native speaker with a foreign accent.

In an elaborate speech recognition system there are several knowledge sources among which the pronunciation dictionary is of special interest for measures against non-native accents. Usually, this dictionary denotes the native pronunciations for all items of the vocabulary. In order to make the system ready for non-native accents, also the respective non-native pronunciations must be entered into the pronunciation dictionary. Up to now these addi-

tional entries either had to be specified by an expert for the special pair of target language and foreign language or had to be extracted automatically from a large number of spoken examples. The new approach is based upon the idea of processing the vocabulary and pronunciation dictionary of the target language with special attention to the phoneme inventory of the foreign language. Thus, any desired pair of target language and foreign language can be conveniently managed without the help of special experts and without any spoken examples. This is the most important among the contributions of this work to the fundamentals for the development of future speech recognition systems.

November 2002

E. Paulus

Table of Contents

1	Introduction	1
1.1	Outline of This Book	4
2	ASR: An Overview	7
2.1	General Overview	7
2.2	Automatic Processing of Speech	10
2.3	Evaluation of ASR Systems	11
2.4	Adaptation in ASR Systems	12
3	Pre-processing of the Speech Data	15
3.1	A/D Conversion	15
3.2	Windowing	15
3.3	Filter Bank Analysis	18
4	Stochastic Modelling of Speech	21
4.1	Hidden Markov Models (HMMs)	22
4.2	Solving the Three HMM Problems	25
4.2.1	Recognition	25
4.2.2	Finding the Optimal State Sequence	26
4.2.3	Training	27
5	Knowledge Bases of an ASR System	31
5.1	Acoustic Models	31
5.2	Pronunciation Dictionary	32
5.3	Language Models (LMs)	35
6	Speaker Adaptation	37
6.1	The State of the Art in Speaker Adaptation	38
6.1.1	Feature-Based Approaches	40
6.1.2	Model-Based Approaches	41
6.2	Maximum Likelihood Linear Regression	42
6.2.1	MLLR for Small Amounts of Adaptation Data	44
6.2.2	The Weighted MLLR Approach	46
6.2.3	Implementation Issues	48

6.2.4	Experiments and Results	49
6.3	Summary	54
7	Confidence Measures	57
7.1	The State of the Art in Confidence Measures	58
7.1.1	Statistical Hypothesis Testing	59
7.1.2	Using a Set of Features	60
7.2	Neural Networks	61
7.2.1	Activation Function	62
7.2.2	Output Function	64
7.2.3	Learning in NNs	64
7.3	Evaluating Confidence Measures	66
7.4	CM Features	66
7.4.1	Phone-Duration Based Features	67
7.4.2	Additional Features	69
7.4.3	Combining the NN Classifier with Speaker Adaptation	72
7.5	Experiments and Results	74
7.5.1	Evaluation of the NN Classifier	74
7.5.2	Semi-supervised Adaptation	76
7.6	Summary	78
8	Pronunciation Adaptation	79
8.1	The State of the Art in Pronunciation Modelling	80
8.1.1	Rule-Based Approaches	82
8.1.2	Data-Driven Approaches	83
8.1.3	Combined Approaches	84
8.1.4	Miscellaneous Approaches	86
8.1.5	Re-training the Acoustic Models	86
8.2	Pronunciation Modelling of Accented and Dialect Speech	87
8.3	Recognising Non-native Speech	88
8.4	Generating Non-native Pronunciation Variants	94
8.4.1	Classification Trees	97
8.4.2	Experiments and Results	99
8.5	Summary	103
9	Future Work	105
9.1	Dynamic Selection of Pronunciation Rules	106
10	Summary	109
	Bibliography	113
	Index	125
	Glossary	127

A	Databases and Experimental Settings	131
A.1	The German Database	131
A.1.1	Pre-processing of the Speech Data	132
A.2	Settings for NN Training and Testing	132
A.3	The British English WSJ Database	133
A.4	ISLE Database	133
B	MLLR Results	135
C	Phoneme Inventory	139
C.1	German Symbol Inventory	140
C.2	English Symbol Inventory	142
C.3	Manually Derived Pronunciation Rules for the ISLE Corpus .	142