Learning pronunciation variation

A data-driven approach to rule-based lexicon adaptation for automatic speech recognition

by

Ingunn Amdal

A dissertation submitted in partial fulfilment
of the requirements for the degree of

DOKTOR INGENIØR

Department of Telecommunications
Norwegian University of Science and Technology
N-7491 Trondheim
Norway

2002

Norwegian University of Science and Technology
Department of Telecommunications
N-7491 Trondheim, Norway

Dr. ing.-avhandling 2002:100
Rapport nr.: 420206

ISBN 82-471-5502-8
ISSN 0809-103X

Ever'body says words different … Arkansas folks says 'em different, and Oklahomy folks says 'em different. And we seen a lady from Massachusetts, an' she said 'em differentest of all. Couldn' hardly make out what she was sayin'.

John Steinbeck
The Grapes of Wrath (1939)

Abstract

In this dissertation a complete data-driven approach to rule-based lexicon adaptation is presented, where the effect of the acoustic models is incorporated in the rule pruning metric.

Robust speech recognition is an important research topic, which can contribute to make systems based on automatic speech technology more user-friendly. To achieve a robust system the variation seen for different speaking styles must be handled. In this dissertation we have therefore investigated how to model pronunciation variation for different speaking styles.

The method presented in this dissertation consists of data-driven solutions to all the steps in rule-based pronunciation modelling:

- First an alternative transcription is generated from phone recognition of each utterance, using the acoustic models in order to observe the variation without the restriction of the recognizer's lexicon. We use the same acoustic models as we later will use in the recognition phase for a consistent rule derivation and assessment.

- Alignment of the transcriptions is performed by the traditional dynamic programming approach or by a time synchronous approach. For the dynamic programming a data-driven method to deriving phone-to-phone substitution costs based on the statistical co-occurrence of phones, association strength, is introduced.

- Rules for pronunciation variation are derived from this alignment. The rules are pruned using a new metric based on the acoustic log likelihood. Well trained acoustic models are capable of modelling much of the variation seen, using the acoustic log likelihood to assess the pronunciation rules prevents the lexical modelling from adding variation already accounted for.

- The pruned rules are then used to generate pronunciation variants and the lexicon is modified.

- Adding variants to the lexicon not only corrects errors, but may also introduce new errors. Controlling the added confusability is therefore important. A framework for confusability measures based on decision theory is introduced.

The experiments start with a general investigation of standard automatic speech recognition techniques for different speaking styles. The speaking styles investigated are read speech and spontaneous dictation from native speakers and read non-native speech. A general purpose pronunciation lexicon containing variants and marked canonical pronunciations is used to compare acoustic modelling and lexical modelling of pronunciation variation. Performance of acoustic models of different levels of complexity is also compared. The results show that the lexical modelling using the general purpose variants gave small improvements, but the errors differed compared with using only one canonical pronunciation per word. Modelling the variation using the acoustic models (using context dependency and/or speaker dependent adaptation) gave a significant improvement, but the resulting performance for non-native and spontaneous speech was still far from read speech. Data-driven pronunciation variation modelling is therefore investigated for these two speaking styles.

For the non-native task data-driven pronunciation modelling by learning pronunciation rules gave a significant performance gain. Acoustic log likelihood rule pruning performed better than rule probability pruning. The largest improvement was seen when incorporating the variation for all the speakers in one lexicon, making it possible to use the same lexicon and acoustic models for all speakers.

For spontaneous dictation the pronunciation variation experiments did not improve the performance. The answer to how to better model the variation for spontaneous speech seems to lie neither in the acoustical nor the lexical modelling. One of the main differences between read and spontaneous speech is the grammar used as well as disfluencies like restarts and long pauses. The language model may therefore be the best choice for more research to achieve better performance for this speaking style.

Finally, alignment methods are compared. The association strength scheme for deriving substitution costs is shown to give better performing rules than other dynamic programming methods. The usual dynamic programming approach has difficulties: 1) for transcription errors and disfluencies (usual in spontaneous speech) and 2) when aligning different words (for confusability measures). We also show examples of alignments where a time synchronous alignment may be beneficial to ensure that we compare the same acoustic segments.

Preface

This dissertation is submitted in partial fulfilment of the requirements for the doctoral degree of *doktor ingeniør* at the Norwegian University of Science and Technology (NTNU). The advisors have been Professor Torbjørn Svendsen and Associate Professor Magne Hallstein Johnsen, both at the Department of Telecommunications, NTNU.

The main work has been conducted in the period from April 1998 to October 2001. In addition to the research activity, the work included compulsory courses corresponding to one year full-time studies, as well as half a year of teaching assistant duties. Physically I have spent approximately half the time at the Signal Processing Group in the Department of Telecommunications, NTNU, Trondheim, and half the time at Telenor Research and Development, Kjeller. In the period from January 2000 to July 2000 I stayed at Bell Labs, Lucent Technologies, Murray Hill, New Jersey, USA and worked under the supervision of Filipp Korkmazskiy, Arun C. Surendran and Chin-Hui Lee. Further, finishing touches of the work were given while working as a research scientist at Telenor Research and Development, Fornebu, as well as at a second stay at Bell Labs from February 2002 to May 2002 under the supervision of Eric Fosler-Lussier.

The work has been funded by the Norwegian Research Council (NFR) as a part of the project SPODIS (SPOken DIalogue Systems for telephone services) with additional funding from Telenor. In addition, for both my stays in USA I received partial funding from Bell Labs.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor Professor Torbjørn Svendsen for his support, encouragement, and guidance throughout this work. My sincere thanks also go to Associate Professor Magne Hallstein Johnsen who has always been available for questions and discussions and for being invaluable help in the finishing phase. In the starting phase of my work Trym Holter was of great help and contributed significantly to this work with his ideas on data-driven pronunciation modelling.

During the course of this work I have had the privilege of working with many helpful colleagues at NTNU, Bell Labs, and Telenor. There is a number of people who have given me a lot of help and encouragement along the way, and provided an enjoyable working atmosphere. I would like to thank each of them and I especially appreciate the efforts by Bojana Gajić, Tor André Myrvoll, Ole Morten Strand, Hallstein Lervik, Kirsten Ekseth, and Arne Kjell Foldvik at NTNU. At Telenor Research and Development, I would especially like to thank Knut Kvale, Britt Kjus, Lars Hafskjær, and Endre Skolt.

I am grateful to Chin-Hui Lee at Bell Labs, Lucent Technologies, New Jersey, who gave me the opportunity to stay with his group for six months from January 2000 to July 2000. My supervisors Filipp Korkmazskiy and Arun C. Surendran contributed significantly to the work in this dissertation. I also want to thank Olivier Siohan for his helpful support. In my second three month stay at Bell Labs from February 2002 to May 2002, Eric Fosler-Lussier provided an excellent discussion partner with his comprehensive knowledge on pronunciation modelling.

Finally, I thank my family for they love and support. My parents have always given their encouragement and support in a non-intrusive way I appreciate deeply. Especially, I am grateful to my husband Øyvind Eilertsen for his love, support and unwavering confidence in me. His contribution in practical matters like proof-reading is greatly appreciated, but most important is his contributions in all the matters that are beyond a dissertation.

Contents

Abstract i

Preface iii

Acknowledgements v

List of abbreviations xi

1 Introduction 1

1.1 Background 1

1.2 Robust automatic speech recognition 3

1.3 Pronunciation modelling 4

1.4 This dissertation 6

1.4.1 Contributions of the dissertation 7

1.4.2 Outline of the dissertation 8

1.4.3 List of publications 9

2 The automatic speech recognition system 11

2.1 Phonetics and phonology 11

2.1.1 Sub-word units for automatic speech recognition 16

2.2 Basics of automatic speech recognition (ASR) 16

2.2.1 Speech pre-processing 18

2.2.2 Acoustic models 20

2.2.3 Lexicon 24

2.2.4 Language model 26

2.3 Discriminative methods in ASR 26

2.4 How to model pronunciation variation in ASR 28

2.5 Statistical considerations 29

3 Speaking styles 31
3.1 Dialects and accented speech (including non-native speech) . . 32
3.2 Spontaneous speech 34
3.3 Language modelling and pronunciation variation 35
3.3.1 Spontaneous speech characteristics in human-human versus human-machine dialogues 36
3.3.2 Language modelling and disfluencies 36

4 Acoustic model adaptation 39
4.1 Linear transformation (MLLR) 40
4.2 MAP adaptation 40
4.3 Other methods 42
4.4 Acoustic models and pronunciation variation 43
4.4.1 Speaker adaptation of acoustic models 43
4.4.2 Accent and dialect adaptation of acoustic models 43
4.4.3 Choice of speech recognition units 44
4.4.4 Dynamic HMMs 45

5 Lexicon adaptation 47
5.1 Knowledge based pronunciation modelling 49
5.2 Direct data-driven pronunciation modelling 51
5.3 Indirect data-driven pronunciation modelling 53
5.4 Step-by-step data-driven rule derivation 55
5.4.1 Reference and alternative transcriptions 55
5.4.2 Alignment 56
5.4.3 Rule derivation 57
5.4.4 Rule assessment and pruning 58
5.4.5 Pronunciation variant generation, assessment and pruning 58
5.4.6 Retranscription 59
5.5 Confusability reduction 59

6 Acoustic log likelihood based pronunciation modelling 61
6.1 Decision theory applied to pronunciation modelling 62
6.1.1 Misclassification measures for one baseform 64
6.1.2 Misclassification measures for sets of baseforms 66
6.1.3 Maximum mutual information 70
6.1.4 Limitations and practical considerations 72
6.2 Direct modelling of pronunciation variants 73
6.3 Indirect modelling using pronunciation rules 74
6.3.1 Reference and alternative transcriptions 74
6.3.2 Alignment 75

6.3.3 Rule derivation 79
6.3.4 Rule assessment and pruning 80
6.3.5 Pronunciation variant generation, assessment and pruning 82
6.3.6 Retranscription 83
6.4 Confusability reduction 83

7 Comparison of ASR performance for different speaking styles 85
7.1 Introduction 86
7.2 Experimental procedure 87
7.2.1 Speaking styles 87
7.2.2 Lexica 88
7.2.3 The HTK reference recognizer 89
7.3 Results 91
7.3.1 Context-independent models 91
7.3.2 Context-dependent models 92
7.3.3 Comparison of context-independent and dependent models 93
7.3.4 Acoustic model speaker adaptation 95
7.3.5 Error analysis 97
7.3.6 Pronunciation probabilities 98
7.3.7 Language model effect 99
7.4 Comparisons with other systems 100
7.5 Discussion 101
7.6 Summary of the main results 102

8 Data-driven pronunciation modelling for non-native speech 105
8.1 Introduction 105
8.2 Experimental procedure 106
8.2.1 The database 106
8.2.2 The BLASR reference recognizer 107
8.2.3 Rule derivation 108
8.3 Results 113
8.3.1 Individual pronunciation rules 113
8.3.2 Individual maximum likelihood baseforms for words . . 114
8.3.3 Common pronunciation rules for all speakers 116
8.3.4 Retranscription 119
8.3.5 Log likelihood based rule pruning 120
8.3.6 Simple log likelihood based confusability reduction . . . 121
8.3.7 Other confusability reduction experiments 123
8.4 Discussion 124
8.5 Summary of the main results 126

9 Data-driven pronunciation modelling for spontaneous speech 127
9.1 Introduction . . . 127
9.2 Experimental procedure . . . 127
9.2.1 The database . . . 127
9.2.2 The HTK reference recognizer . . . 128
9.2.3 Rule derivation . . . 128
9.3 Results . . . 131
9.4 Discussion . . . 134
9.5 Summary of the main results . . . 135

10 Comparison of alignment methods 137
10.1 Introduction . . . 137
10.2 Individual rules for non-native speakers . . . 138
10.2.1 Experimental procedure . . . 138
10.2.2 Results . . . 139
10.3 Alignment of spontaneous dictation . . . 139
10.4 Alignment of error transcriptions . . . 141
10.5 Discussion . . . 142
10.6 Summary of the main results . . . 142

11 Concluding summary 143
11.1 Variation modelling for different speaking styles . . . 144
11.2 Data-driven pronunciation rule derivation and assessment . . . 144
11.3 Alignment . . . 145
11.4 Conclusions . . . 145
11.5 Some directions for further work . . . 146

A Phonetic alphabets 149

B CMU lexicon specification 153

C Confidence intervals 155

D Decision theory 159
D.1 Bayes classifier . . . 159
D.2 Discriminant functions . . . 161
D.3 Optimization . . . 162

E Detailed experimental results for different speaking styles 167
E.1 Choice of retranscription scheme for variant training . . . 167

Bibliography 171

List of abbreviations

ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BLASR	Bell Labs Automatic Speech Recognizer
BPW	Baseforms Per Word
CART	Classification And Regression Trees
CMN	Cepstral Mean Normalization
CMS	Cepstral Mean Subtraction
CMU	Carnegie Mellon University
CRPR	Combined Rule PRobability (individual alignment and joint rule derivation)
GPD	Generalized Probabilistic Descent
HMM	Hidden Markov Model
HTK	Hidden markov model Tool-Kit
IPA	International Phonetic Alphabet
JRPR	Joint Rule PRobability (joint alignment and joint rule derivation)
LDC	the Linguistic Data Consortium
LL	Log Likelihood pruning measure for an acoustic segment
LLH	Log LikeliHood pruning measure for a rule
LPC	Linear Prediction Coefficients
MAP	Maximum A Posteriori
MCE	Minimum Classification Error
MFCC	Mel-Frequency Cepstral Coefficients
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Regression
MMI	Maximal Mutual Information
MRPR	Merged Rule PRobability (merging of individually derived rules)
PER	Phone Error Rate
RPR	Rule PRobability

SAMPA	Speech Assessment Methods Phonetic Alphabet
SI-284	WSJ Speaker Independent 284-speaker set
SI-84	WSJ Speaker Independent 84-speaker set
VTLN	Vocal Tract Length Normalization
WER	Word Error Rate
WSJ	Wall Street Journal

Chapter 1

Introduction

Applications using speech technology have for long been a natural part of many futuristic descriptions in books, films, TV-series, and commercials. The technology has over the last years reached a state where we see speech technology based products also being used in the real world. The acceptance by the public increases as the technology improves, but we are still far from the conversational interfaces encountered in fiction. The speech technology applications in real use are limited in one way or the other to give acceptable performance. Limitations often used are e.g. restricted vocabulary, restricted environments, and systems tailored to one or a few users.

Among the most popular type of applications used by the general public are travel information, call centres, and ordering phones. Many different ways of speaking must be tolerated, and the dialogues are usually kept quite restricted to simplify the task for the system. Another use of speech technology is automatic dictation e.g. for medical personnel, where the system is tailored to one person and thus can manage a larger vocabulary.

1.1 Background

A speech technology based system consists of a speech recognition system for speech input, a speech synthesizer for speech output, and a dialogue manager to handle the dialogue. In this dissertation, only the automatic speech recognition (ASR) system will be considered.

Speech input is preferable when the users need to have their hands and eyes free to do other tasks, as in a car or control room environment, or when the device to be controlled is out of reach. Complex actions like asking for several information items at the same time are easier using speech than a graphical interface. Compared to a graphical interface a speech based system can easily refer to invisible objects. On the other hand there may be ambiguities in referring to the objects. (If a users says “this” the system must interpret what object “this” refers to.) The advantages by using speech rely on a well functioning speech recognition system. With many recognition errors the disadvantage for the user who must perform error-corrections will outweigh the advantages. The advantages of interaction through speech can best be utilized in a multimodal setting. Multimodal user interfaces give the user the opportunity to choose the most suitable mode depending on the situation.

The goal of all research on ASR is to improve the performance. Improved ASR will help make systems based on speech technology more user-friendly and also increase the number of applications that can be speech-enabled. For languages where the ASR research effort has been large, (e.g. US and UK English, Japanese, German) the current technology has reached such a state that useful applications based on speech recognition are possible, e.g. products for controlled tasks like dictation [130]. In the future more and more advanced speech and language based services are expected [16].

With the promising performance of state-of-the-art ASR systems it is possible to strive for further improvements that will handle more speaker and environmental variation. To make a speech recognition product a success there must be few restrictions on the customer’s behaviour and environment. If ASR systems can cope with conversational dialogues it is possible to use mixed initiative, not only machine-driven dialogues [132]. This development calls for a deeper understanding of the underlying principles of spoken language [42]. It also calls for an understanding of the recognizer to ensure that variation in speech is addressed correctly. Changes in one part of the recognition system will influence other parts of the system.

Current ASR systems have difficulties with spontaneous speech encountered in conversational dialogues, as well as accent and dialect variability. The substantial differences between non-native speech and native speech will also challenge a “native” ASR system. More subtle differences that are easily handled by humans (e.g. Australian versus US English) may still cause problems for ASR. The speaker dependent variations will mainly be caused by inter-speaker variability, but we will also encounter intra-speaker variability, i.e. the same speaker may behave differently depending on the context (e.g. environment and task) and over time when getting used to a system. In this dissertation we concentrate on the variability that is not on the acoustic level (e.g. the variability due to the differences in the vocal apparatus between individuals), but on the variation on the lexical level that will be similar for groups of speakers. We call this variation different *speaking styles*.

Our objective is to investigate both the similarities and differences among the different speaking styles and how to best model the variation seen. Even if there are different types of variation, the methods for treating them may be similar.

1.2 Robust automatic speech recognition

All ASR systems must handle variation. The same word spoken several times by the same speaker will vary both in length and acoustical content. For speaker independent speech recognition the voice quality and characteristics will vary even more. The term *robust* automatic speech recognition is used when we consider variation beyond the inter- and intra-speaker acoustic differences we see even for read speech.

The variation in speech input to a speech technology based service may be divided into three groups:

1. Pronunciation variation
2. Grammar and vocabulary variation
3. Channel and noise variation

Pronunciation, grammar and vocabulary variation will be speaker dependent whereas channel and noise variation will depend on the environment. Trying to make ASR handle both speaker and environment variation is crucial in robust modelling. The research groups interested in one of these two issues are often interested in the other one as well, as both areas must be handled for example in public telephone services based on speech recognition.

Different speakers using different speaking styles will use different pronunciations. Spontaneous speech and different dialects or accents are examples of speaking styles with pronunciations that differ from the canonical ones often found in pronunciation dictionaries. The population of most countries becomes more and more multinational, and non-native users will increase the observed variability in pronunciation even more. There will also be differences between expert and novice users of a speech based service (e.g. fast versus over-articulated speech). User-friendly systems should be able to recognize the pronunciations judged appropriate by the user. One of the eight golden rules in user interface design is to "Support internal locus of control" [102]. The user should be spared surprising system actions when using "non-surprising" speech. This will help the user to keep a consistent mental model of the system, which is of great importance for a well-designed dialogue system. If the recognizer makes errors when the user is using rare words or pronunciations, this will be understandable for the user. To make dialogue systems using speech recognition more user-friendly, robustness to common pronunciation variation is needed.

The task that the speech based application is intended for gives requirements for the recognizer's vocabulary and grammar, and presents another source of variation. The vocabulary and grammar preferred by the user may vary dependent on e.g. non-nativeness, dialect and sociolect, as well as differences between expert and novice users. One cannot assume that the users of a speech technology application will stick to a well-defined grammar (as perhaps for dictation systems). Users may be unwilling to normalize both pronunciation and grammar. They may also be unaware of their own peculiarities. Many perceive their own speaking style as normalized but all these "normalized" variants differ. There was for example an unexpected amount of variation in pronunciation among professional speakers when searching for "model" speakers for Austrian German [83]. Hesitations, restarts, and other disfluencies are also characteristics of spoken language that vary among speakers and must be handled by the language model in a speaker-independent system. Robustness is therefore also needed in modelling grammar variability.

The third main variable that needs to be addressed in speech recognition-based applications is non-speech variation, e.g. noise and channel variation. Both pronunciation and grammar variation are dependent on the user and therefore quite different from this last type of variation that is dependent on the environment. To control how we model the observed variation, we should treat the environment and speaker variation separately. Acoustic model adaptation techniques can model both speaker and environment variations. Explicitly separating these two effects is recognized as an important area for future research [125].

1.3 Pronunciation modelling

It is desirable for speech recognition systems to manage diverse speaking styles, (e.g. spontaneous speech, accents and dialects, and speech from users with different mother tongues), but such variation in user input is difficult for the current state-of-the-art recognizers. One way of improving this is better modelling of pronunciation variation. The pronunciation dictionary is therefore an important part of the ASR. In this dissertation it is called a *lexicon*, a familiar term in the speech community. A lexicon defines the transcription of the words in terms of the acoustic model units of the recognizer. This transcription will not necessarily look like an entry in a pronunciation dictionary made for humans and not machines. This will be treated in more detail in section 2.2.3.

Pronunciation modelling is by no means a new issue in the ASR community, early efforts are reported in e.g. [8] and [95]. Pronunciation variation modelling is still an important issue in ASR research, and overviews are for example given in [113] and [114]. More recently multilingual ASR has become an interest [1], which introduces new challenges for pronunciation modelling.

Pronunciation variation can be captured using linguistic knowledge, i.e. specific knowledge about how people with different accents pronounce words, but this knowledge is not always sufficient for pronunciation modelling. As an example, a transcription of spontaneous US English speech (Switchboard) revealed at least 80 variants of the word "and" [43]. Non-native speech varies even more, and the phonological rules governing the variation will probably be different for speakers with different mother tongues. In such cases a data-driven approach may be more suitable where we try to extract information from a database containing the speech we want to model. The resulting pronunciation rules will depend on the database and thus on the language, as well as the task and speaking style. This may be favourable for a tailored system as we then only model the variation seen for this specific task. Nevertheless, a method that is independent of the specific database and language is preferable, as it can be reused for other tasks without major modifications.

Linguistic knowledge does not give sufficient information to optimize an automatic speech recognizer. The knowledge varies from language to language, but even for the most studied languages pronunciations are constantly changing and the number of different speaking styles with their characteristics makes it infeasible to have a complete picture. The speech recognition models used today are therefore based on statistical representation and analysis, which must be kept in mind when optimizing the system. A handbook in dialogue design for speech technology by Balentine and Morgan [12] says (page 70):

> "Such models contain statistical-processing artifacts that bear no direct relations to human hearing, and consequently speech recognition often makes mistakes humans would not make."

All parts of the recognizer except the lexicon are usually optimized with respect to objective criteria. A data-driven approach will enable us to use the same criteria for the lexicon as for the other parts of the recognizer, allowing a unified optimization of the whole system. We therefore believe a data-driven approach to pronunciation modelling should be preferred. There is no reason to ignore linguistic knowledge, but the effects of the pronunciation rules derived using either method should be verified on representative speech data.

The statistically based acoustic models of current ASR systems are capable of handling much of the variation seen in speech, also pronunciation variation.

More complex acoustic models will for example handle many allophonic variations in a suitable way. Adaptation of the acoustic models is a successful method to further improve individual recognizers.

Some pronunciation variation can be described as phonological, e.g deletions, insertions, and larger changes (larger both in length and acoustic variation.) This kind of variation may be better handled at the lexical level. Modelling of a group of very different speakers by adapting the acoustic models may result in diffuse models, and in these cases pronunciation modelling by changing the lexicon may give better performance. The two techniques for capturing variation should be combined using the method that gives the best result; acoustic model adaptation for the pronunciation variation at the allophonic level, and lexicon adaptation for the more phonological variation.

Since large vocabulary recognizers always include a language model, the effect of this model should be incorporated in the pronunciation modelling techniques.

One of the main problems in pronunciation modelling is to make sure that we know which variation we are modelling. The effect of the acoustic models, the lexicon, and the language model will interact. The possibility of adding superfluous complexity by modelling the same variation several times, or even worse, adding contradicting changes, must be avoided.

1.4 This dissertation

In this dissertation the focus is on using the lexicon to capture speaker variation, using the same lexicon and the same acoustic models for all speakers. Experiments on individual lexicon adaptation as well as acoustic model adaptation are also presented.

We believe all parts of the system should be optimized in a consistent way. The training of the acoustic and language parts of an ASR system is based on objective criteria, objective criteria should be used for optimizing the lexicon also. This calls for data-driven methods in pronunciation modelling. Knowledge about human perception and production of speech, as well as linguistics and phonetics, is important, but must be formalized for building the ASR system.

Different measures can be used in data-driven pronunciation modelling. We believe the acoustic likelihood should be utilized as a measure in pronunciation modelling. We then take into consideration the variation already modelled by the acoustic models and thereby give a measure consistent with the optimization of the whole ASR system.

One of the major drawbacks with data-driven modelling is that we are restricted to variation present in the lexicon adaptation data. Direct modelling of pronunciation variation by deriving alternative pronunciations for words present in sufficient numbers in the adaptation data, has been shown to give improvement, e.g. in [49]. To model pronunciations for words not present in the lexicon adaptation data ("unseen words"), it is necessary to extend the method to modelling pronunciation rules, and thereby generalize the variation seen in the adaptation data. This gives many new challenges on how to derive the rules and how to generate and assess pronunciation variants from them.

One reason for the modest improvements achieved in pronunciation modelling is the lack of a way to control the confusability between pronunciations. To make lexica tailored to a person or group we cannot rely on just adding extra pronunciations, we must also remove confusable ones. The use of discriminative methods in choosing which pronunciations to add to the lexicon is one way of solving this problem.

Pronunciation modelling consists of several steps, including alignment of the reference and alternative transcriptions. Although this is a small part of pronunciation modelling (and therefore the whole system), we have investigated alignment methods based on objective criteria in order to be consistent in all parts of the pronunciation modelling.

Last, but not least, it is important to gain knowledge on the effect of the different variation modelling techniques for different speaking styles. The impact of various standard ASR techniques on different speaking styles has been therefore investigated.

1.4.1 Contributions of the dissertation

- **Data-driven pronunciation rule assessment using acoustic likelihood:**
  In this dissertation the acoustic likelihood derivation and selection of pronunciation variants presented in [49] is expanded to derivation and selection of pronunciation *rules*. The advantage of consistent optimization when using an acoustic likelihood based metric is combined with the possibility to model pronunciations for unseen words.

- **Data-driven alignment using dynamic programming with phone-to-phone substitution costs derived using the data:**
  Current alignment methods use costs for phone-to-phone substitutions based on phonetic knowledge or phone identity only. In this dissertation we present a data-driven approach to derive the costs. This method is inspired by the grapheme-to-phoneme conversion in [68]. These costs may be more suitable than phonologically based costs when the alternative

transcription to be aligned is automatically derived without phonological constraints. This method also gives the possibility of non-symmetric mappings.

- **Data-driven pronunciation rule derivation using time synchronous alignment:**
  Another alignment alternative using the time information provided by the ASR system is presented in this dissertation. Using this alignment method we assure we compare the transcriptions of the same acoustic segments. This method also provides pronunciation rules without the need for an extra rule derivation step. Corresponding acoustic likelihood scores are also derived as a by-product.

- **Comparisons of standard variation modelling techniques for different speaking styles:**
  State-of-the art variation modelling techniques are evaluated in this dissertation for different speaking styles. The study is focused on the comparison of acoustic and lexical modelling.

- **A framework for decision theory applied to pronunciation modelling:**
  An assessment method for confusability is important. In this dissertation known decision theory methods, e.g. from [61] and [67], are used to derive confusability measures given a lexicon and an appropriate lexicon adaptation set.

1.4.2 Outline of the dissertation

Chapter 2 describes the basics of the automatic speech recognition system at the level necessary to understand the experiments in this dissertation. The description starts with a short explanation of elements from phonetics and phonology that are relevant for ASR systems. Three ways of modelling pronunciation variation in different parts of the recognizer are outlined. A section on statistical considerations is also included in this chapter. The kinds of variation in spoken language that can be addressed by pronunciation modelling, in this dissertation called "speaking styles", are described in chapter 3. Language modelling topics concerning speaking style variation are also described. Modelling pronunciation variation is also closely related to the adaptation of the acoustic models; this subject is described in chapter 4. Lexicon adaptation is the main subject of this dissertation and is described in chapter 5. Chapters 3, 4, and 5 give an overview of previous work in the field. The theory underlying the experiments in this dissertation is given in chapter

6. Experiments are described in the result chapters 7, 8, 9, and 10. These 4 result chapters conclude with discussions and short summaries. Finally, a concluding summary is given in chapter 11.

1.4.3 List of publications

Parts of the results presented in this dissertation are published in the following publications:

- I. Amdal and T. Svendsen, "Evaluation of pronunciation variants in the ASR lexicon for different speaking styles," in *Proc. LREC-2002*, (Las Palmas de Gran Canaria, Spain), pp. 1290–1295, 2002.

- I. Amdal, F. Korkmazskiy, and A. C. Surendran, "Joint pronunciation modelling of non-native speakers using data-driven methods," in *Proc. ICSLP-2000*, (Beijing, China), pp. III:622–625, 2000.

- I. Amdal, F. Korkmazskiy, and A. C. Surendran, "Data-driven pronunciation modelling for non-native speakers using association strength between phones," in *Proc. ISCA ITRW ASR2000*, (Paris, France), pp. 85–90, 2000.

The results on maximum likelihood variants in section 8.3.2 are based on similar experiments published in:

- I. Amdal, T. Holter, and T. Svendsen, "Modellering av uttalevariasjon for automatisk talegjenkjenning" in *Nordlyd (Tromsø University working papers on language & linguistics)*, vol. 28, pp. 74-87, 2000.

- I. Amdal, T. Holter, and T. Svendsen, "Maximum likelihood pronunciation modelling of Norwegian natural numbers for automatic speech recognition," in *Proc. Norwegian Signal Processing Symposium (NORSIG)*, (Asker, Norway), pp. 145–150, 1999.

Further work on confusability metrics is presented in:

- E. Fosler-Lussier, I. Amdal, and H.-K. J. Kuo, "On the road to improved lexical confusability metrics," in *Proc. ISCA ITRW Pronunciation Modeling and Lexicon Adaptation (PMLA)*, (Estes Park (CO), USA), pp. 53–58, 2002.

Earlier work with relevance to pronunciation modelling is presented in:

- K. Kvale and I. Amdal, "Improved automatic recognition of Norwegian natural numbers by incorporating phonetic knowledge," in *Proc. ICASSP-97*, (Munich, Germany), pp. 1763–1766, 1997.
- K. Kvale and I. Amdal, *Automatic recognition of Norwegian natural numbers over telephone lines*, Telenor R&D report 19/97, 1997.

Chapter 11

Concluding summary

In this dissertation pronunciation variation modelling for different speaking styles has been investigated. We have interpreted the term "speaking styles" widely, we have included not only read and spontaneous speech, but also non-native speech. A goal for the research on ASR is to make applications based on speech technology more user-friendly. To achieve this it is important that the system can accept pronunciations that are perceived as normal by the user. Spontaneous and non-native speech that is "normal" in the sense that humans recognize it without problems, causes problems for current ASR systems. We believe that a better modelling of pronunciation variation will give a more robust system for different speaking styles.

Pronunciation variation can be modelled in different parts of the ASR system: the acoustic models, the lexicon, or the language model. This dissertation is focused on lexical modelling, but the pronunciation modelling technique presented utilizes assessment metrics incorporating both acoustic models and implicitly the language model. A joint optimization can prevent adding variation in one part of the system that is already sufficiently modelled in another part of the system, or even worse; adding contradicting variations. We therefore strive for a unified optimization using objective criteria for all parts of the ASR system, including the lexicon.

A complete data-driven approach to pronunciation rule derivation was presented in this dissertation. Using rules we can generalize from the variation seen in the adaptation data to words not present in these data. The method presented shows how to use the acoustic log likelihood as a metric to assess the rules.

11.1 Variation modelling for different speaking styles

We have shown in chapter 7 that augmenting the recognizer lexicon with pronunciation variants found in a general purpose lexicon gave small performance gains, and most for read native speech. Error analysis showed that the system using a single canonical pronunciation generated different errors than the one using pronunciation variants, although the word error rate was similar. For non-native speech we observed no improvement for context-dependent acoustic models compared to context-independent models. This speaking style had the largest gain using speaker dependent acoustic model adaptation, but the performance was still far from the results for native speech. For spontaneous speech we observed less improvement by speaker adaptation than for read speech.

11.2 Data-driven pronunciation rule derivation and assessment

Baseform variant generation by using data-driven rule derivation can be described in five steps (repeated from section 5.4):

1. Automatically generate alternative transcriptions
2. Align the reference and alternative transcriptions
3. Derive rules from the alignment
4. Assess and prune the rules
5. Generate baseform variants from the rules, assess the variants, prune or assign weights, and modify the lexicon

In this dissertation data-driven approaches were presented for all steps, but the main contributions are on steps 2, 4 and 5. For step 2 we have introduced association strength as a way to derive phone substitution costs from the data. For step 4 we have introduced a metric based on acoustic log likelihood and for step 5 we have presented a framework for a confusability metric based on decision theory.

In chapter 8 the methods presented were evaluated on non-native speech. The results show that the acoustic log likelihood pruning gave improved performance compared with the more traditional rule probability pruning. We observed a better performance when modelling the non-native speakers jointly than individually. This was a surprising result, as the speakers had quite different language backgrounds, but may be because the amount of data used was small. Even if the joint set of non-native speech was more diverse, the larger amount of data was beneficial to get a more reliable rule selection for the data-driven methods investigated. The confusability metric gave inconsistent results for this task, showing the vulnerability for the method when the vocabularies of the adaptation and test sets differ strongly.

For spontaneous dictation the results in chapter 9 showed no benefit by using the same pronunciation modelling techniques as used for non-native speech, neither for rule probability pruning nor acoustic log likelihood pruning. One reason may be that the rather simple rules investigated in this dissertation only modelled variation that already was sufficiently modelled by the acoustic models. The reason may also be that pronunciation modelling is not the main answer to better modelling of this speaking style.

11.3 Alignment

In chapter 10 some shortages of current alignment methods were identified. To compare dynamic programming alignment methods with different substitution cost schemes, we have compared WER after rule based pronunciation variation modelling using the different alignments. The non-native task was chosen for this experiment and the alternative transcription was made using a phone loop. The statistically based association strength was shown to produce equal or better performing rules than uniform or phonetically based substitution costs.

The usual dynamic programming approach has difficulties: 1) for transcription errors and disfluencies (usual in spontaneous speech) and 2) when aligning different words (for confusability measures). We have shown examples of this where a time synchronous alignment may be beneficial to ensure that we compare the same acoustic segments.

11.4 Conclusions

Current ASR systems perform substantially worse for both non-native and spontaneous speech than for read speech. In this dissertation we have therefore investigated a data-driven approach to rule based lexicon adaptation for these two speaking styles.

For a spontaneous dictation task the proposed method did not give any improvement, nor did more traditional rule probability based pronunciation modelling. The recognized strings were similar to the baseline result; the variation modelled by the rules did neither correct, nor introduce errors. Baseform variants from a general purpose lexicon did not give any improvement either, but here the errors differed compared with using one canonical baseform entry.

For the non-native task we observed the same effect as for spontaneous speech when adding general purpose variants: The errors differed, but the performance did not. The proposed rule-based lexicon adaptation gave significant improvements for this task, and we observed larger gain for the new acoustic log likelihood metric compared with a rule probability metric.

The results indicate that we should choose different ways to model pronunciation variation for these two speaking styles. However, the rules investigated in this dissertation are rather simple and this may be one reason why they were better able to model the larger shifts in pronunciation present in non-native speech. Further research combining more sophisticated rule derivation with the proposed acoustic log likelihood pruning should be tried before we reject the hypothesis that pronunciation variation modelling also will give a better performance for spontaneous speech.

One of the main differences between read and spontaneous speech is the grammar used as well as disfluencies like restarts and long pauses. The language model may therefore be the best choice for more research to achieve better performance for this speaking style.

Even if speaker adaptation was shown to give large improvements for non-native speech, the resulting performance was worse than for native speech on the same task. To achieve results more comparable to native speech, a combination of lexical and acoustic adaptation may be beneficial. It is then crucial to use a metric for the pronunciation rule and variant pruning that incorporates the variation accounted for by the acoustic models. The proposed metric in this dissertation is a step towards this goal.

11.5 Some directions for further work

The rules investigated in this dissertation are rather simple. We have only looked at phone identity as context for the transformation. We will then only be able to model the phone sequences seen in the adaptation data. CART based rule modelling is one way of generalizing the context automatically from the data, and a combination of CART based rule derivation and acoustic log likelihood assessment would therefore be interesting for further studies. With a better generalization (or more data) more sophisticated rules could be derived using more information, e.g. stress and syllable information or for non-natives especially, orthographic information. Coarticulaton effects across word borders should also be included in the variation modelling.

The step from rules to variants needs more attention. We have in this dissertation assessed only the most important variants by the restriction of applying only one rule to each baseform. A rule hierarchy based on an acoustic log likelihood metric must be formulated to extend the method presented to using several rules in one baseform.

The unigram language model was implicitly incorporated in the rule pruning metric. This was done by assessing the total effect instead of the relative effect of the rules. We then incorporate the word probability found in the adaptation set. Higher order language models should be incorporated, and more explicitly by using the probabilities from the language model defined for the task. The probabilities given will then be the same as used in the recognition phase. The task language model is usually trained on much larger amounts of text data than in the adaptation data and will give more reliable estimates for the probabilities. One problem with assessing the total effect is that we then needed the extra restriction of using only one rule for each rule condition. A clustering procedure where we ensure that each acoustic segment only contributes once may be beneficial.

Controlling the confusability is important. A joint optimization using acoustic log likelihood in the pronunciation rule pruning will to some extent help to reduce the confusability by preventing addition of superfluous variation. For a proper confusability metric the combined effect of the baseforms should be assessed. We have in this dissertation presented a framework for discriminative pronunciation assessment with confusability measures based on decision theory. Assessing this framework on data with better agreement between the adaptation and test data than the non-native task should be investigated. When a suitable metric is found it can be used in data-mining approaches to find the optimal set of baseforms.

One difference between the non-native and spontaneous tasks we have investigated is that for the non-native speakers we derived pronunciation rules from an adaptation set with the same speakers as in the test set. Deriving speaker dependent lexica is an interesting task for future research where the main challenge is that we normally will have small amounts of data for each speaker.

Dialects and native accented speech are not investigated in this dissertation, but we may assume that these speaking styles will behave more similar to non-native speech than spontaneous speech. An investigation of the proposed method, and refinements of it for dialect pronunciation modelling, should be investigated.

Bibliography

[1] M. Adda-Decker, “Towards multilingual interoperability in automatic speech recognition,” *Speech Communication*, vol. 35, pp. 5–20, 2001.

[2] J. B. Allan, “How do humans process and recognize speech?,” *IEEE Transactions on speech and audio processing*, vol. 2, pp. 567–577, October 1994.

[3] I. Amdal, T. Holter, and T. Svendsen, “Maximum likelihood pronunciation modelling of Norwegian natural numbers for automatic speech recognition,” in *Proc. Norwegian Signal Processing Symposium (NORSIG)*, (Asker, Norway), pp. 145–150, 1999.

[4] I. Amdal, F. Korkmazskiy, and A. C. Surendran, “Data-driven pronunciation modelling for non-native speakers using association strength between phones,” in *Proc. ISCA ITRW ASR2000*, (Paris, France), pp. 85–90, 2000.

[5] I. Amdal, F. Korkmazskiy, and A. C. Surendran, “Joint pronunciation modelling of non-native speakers using data-driven methods,” in *Proc. ICSLP-2000*, (Beijing, China), pp. III:622–625, 2000.

[6] I. Amdal and T. Svendsen, “Evaluation of pronunciation variants in the ASR lexicon for different speaking styles,” in *Proc. LREC-2002*, (Las Palmas de Gran Canaria, Spain), pp. 1290–1295, 2002.

[7] B. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” *Journal of the Acoustical Society of America*, vol. 55(6), pp. 1304–1312, 1974.

[8] L. R. Bahl, R. Bakis, J. Bellegarda, P. F. Brown, D. Buhrstein, S. K. Das, P. V. de Souza, P. S. Gopalakrishnan, F. Jelinek, D. Kanevsky, R. L. Mercer, A. J. Nadas, D. Nahamoo, and M. A. Picheny, “Large vocabulary natural language continuous speech recognition,” in *Proc. ICASSP-89*, (Glasgow, Scotland), pp. 465–467, 1989.

[9] L. R. Bahl, S. Das, P. V. de Souza, M. Epstein, R. L. Mercer, B. Merialdo, D. Nahamoo, M. A. Picheny, and J. Powell, "Automatic phonetic baseform determination," in *Proc. ICASSP-91*, (Toronto, Canada), pp. 173–176, 1991.

[10] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Decision trees for phonological rules in continuous speech," in *Proc. ICASSP-91*, (Toronto, Canada), pp. 185–188, 1991.

[11] J. K. Baker, "The DRAGON system approach – An overview," *IEEE Transactions on acoustics, speech and signal processing*, vol. ASSP-23, pp. 24–29, February 1975.

[12] B. Balentine and D. P. Morgan, *How to build a speech recognition application.* Enterprise Integration Group, 1999.

[13] W. J. Byrne, M. Finke, S. P. Khudanpur, J. McDouough, H. Nock, M. D. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Pronunciation modelling using a hand-labelled corpus for conversational speech recognition," in *Proc. ICASSP-98*, (Seattle (WA), USA), pp. 313–316, 1998.

[14] W. Chou, C.-H. Lee, and B.-H. Juang, "Minimum error rate training based on N-best string models," in *Proc. ICASSP-93*, (Minneapolis (MN), USA), pp. II:652–655, 1993.

[15] *CMU Pronunciation Dictionary.* [online], 1998. [cited 2002-03-01]. URL: `http://www.speech.cs.cmu.edu/cgi-bin/cmudict/`.

[16] R. V. Cox, C. A. Kamm, L. R. Rabiner, J. Schroeter, and J. G. Wilpon, "Speech and language processing for next-millennium communication services," *Proceedings of the IEEE*, vol. 88(8), pp. 1314–1337, 2000.

[17] N. Cremelie and J.-P. Martens, "Automatic rule-based generation of word pronunciation networks," in *Proc. EUROSPEECH-97*, (Rhodes, Greece), pp. 2459–2462, 1997.

[18] N. Cremelie and J.-P. Martens, "In search of better pronunciation models for speech recognition," *Speech Communication*, vol. 29, pp. 115–136, 1999.

[19] C. Cucchiarini, "Assessing transcription agreement: methodological aspects," *Clinical linguistics & phonetics*, vol. 10 (2), pp. 131–155, 1996.

[20] V. Diakoloukas, V. V. Digalakis, L. G. Neumeyer, and J. Kaja, "Development of dialect-specific speech recognizers using adaptation methods," in *Proc. ICASSP-97*, (Munich, Germany), pp. 1455–1458, 1997.

[21] V. V. Digalakis and L. G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Transactions on speech and audio processing*, vol. 4, pp. 294–300, July 1996.

[22] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis.* Wiley-interscience, 1973.

[23] *The European Language resources Distribution Agency (ELDA).* [online description], 2002. [cited 2002-08-01]. URL: `http://www.elda.fr/`.

[24] R. T. Endresen, *Fonetikk og fonologi: Ei elementær innføring.* Universitetsforlaget, (Oslo, Norway), 1991.

[25] M. Finke and I. Rogina, "Wide context acoustic modeling in read vs. spontaneous speech," in *Proc. ICASSP-97*, (Munich, Germany), pp. 1743–1746, 1997.

[26] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modelling in large vocabulary conversational speech recognition," in *Proc. EUROSPEECH-97*, (Rhodes, Greece), pp. 2379–2382, 1997.

[27] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, (Santa Barbara (CA), USA), pp. 347–354, 1997.

[28] L. D. Fisher and G. van Belle, *Biostatistics*, ch. Hypothesis testing for binomial variables, pp. 182–183. Wiley, 1993.

[29] E. Fosler-Lussier, "Multi-level decision trees for static and dynamic pronunciation models," in *Proc. EUROSPEECH-99*, (Budapest, Hungary), pp. 463–466, 1999.

[30] E. Fosler-Lussier, I. Amdal, and H.-K. J. Kuo, "On the road to improved lexical confusability metrics," in *Proc. ISCA ITRW Pronunciation Modeling and Lexicon Adaptation (PMLA)*, (Estes Park (CO), USA), pp. 53–58, 2002.

[31] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on pronunciations in conversational speech," *Speech Communication*, vol. 29, pp. 137–158, 1999.

[32] E. Fosler-Lussier and G. Williams, "Not just what, but also when: Guided automatic modeling of Broadcast News," in *Proc. DARPA Broadcast News Workshop*, (Herndon (VA), USA), pp. 171–174, 1999.

[33] J. E. Fosler-Lussier, *Dynamic pronunciation models for automatic speech recognition.* PhD thesis, University of California, Berkeley, 1999.

[34] B. Gajić, *Feature extracion for automatic speech recognition in noisy acoustic environments.* PhD thesis, NTNU (Norwegian University of Science and Technology), 2002.

[35] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on speech and audio processing*, vol. 9, pp. 358–366, May 2001.

[36] J.-L. Gauvain and C.-H. Lee, "Bayesian learning for hidden Markov model with Gaussian mixture state observation densities," *Speech Communication*, vol. 11, pp. 205–213, 1992.

[37] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on speech and audio processing*, vol. 2, pp. 291–298, April 1994.

[38] E. P. Giachin, A. E. Rosenberg, and C.-H. Lee, "Word juncture modeling using phonological rules for HMM-based continuous speech recognition," *Computer Speech and Language*, vol. 5, pp. 155–168, 1991.

[39] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP-89*, (Glasgow, Scotland), pp. 532–535, 1989.

[40] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP-92*, (San Francisco (CA), USA), pp. I:517–520, 1992.

[41] S. Goronzy, R. Kompe, and S. Rapp, "Generating non-native pronunciation variants for lexicon adaptation," in *Proc. ISCA ITRW Adaptation methods for speech recognition*, (Sophia-Antipolis, France), pp. 143–146, 2001.

[42] S. Greenberg, “Recognition in a new key – towards a science of spoken language,” in *Proc. ICASSP-98*, (Seattle (WA), USA), pp. 1041–1044, 1998.

[43] S. Greenberg, “Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation,” *Speech Communication*, vol. 29, pp. 159–176, 1999.

[44] S. Greenberg, J. Hollenback, and D. Ellis, “Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus,” in *Proc. ICSLP-96*, (Philadelphia (PA), USA), pp. Addendum 24–27, 1996.

[45] T. Hain and P. C. Woodland, “Dynamic HMM selection for continuous speech recognition,” in *Proc. EUROSPEECH-99*, (Budapest, Hungary), pp. 1327–1330, 1999.

[46] T. Hain and P. C. Woodland, “Modelling sub-phone insertions and deletions in continuous speech recognition,” in *Proc. ICSLP-2000*, (Beijing, China), pp. IV:172–175, 2000.

[47] E. Harborg, *Hidden Markov models applied to automatic speech recognition.* PhD thesis, NTH (Norwegian Institute of Technology), 1990.

[48] P. A. Heeman, D. Cole, and A. Cronk, “The U.S. SpeechDat-Car data collection,” in *Proc. EUROSPEECH-2001*, (Aalborg, Denmark), pp. 2031–2034, 2001.

[49] T. Holter, *Maximum likelihood modelling of pronunciation in automatic speech recognition.* PhD thesis, NTNU (Norwegian University of Science and Technology), 1997.

[50] T. Holter and T. Svendsen, “Incorporating linguistic knowledge and automatic baseform generation in acoustic subword unit based speech recognition,” in *Proc. EUROSPEECH-97*, (Rhodes, Greece), pp. 1159–1162, 1997.

[51] T. Holter and T. Svendsen, “Maximum likelihood modelling of pronunciation variation,” *Speech Communication*, vol. 29, pp. 177–191, 1999.

[52] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: A guide to theory, algorithm, and system development.* Prentice Hall, 2001.

[53] J. J. Humphries and P. C. Woodland, "The use of accent-specific pronunciation dictionaries in acoustic model training," in *Proc. ICASSP-98*, (Seattle (WA), USA), pp. 317–320, 1998.

[54] J. J. Humphries and P. C. Woodland, "Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition," in *Proc. EUROSPEECH-97*, (Rhodes, Greece), pp. 2367–2370, 1997.

[55] J. J. Humphries, P. C. Woodland, and D. Pearce, "Using accent-specific pronunciation modelling for robust speech recognition," in *Proc. ICSLP-96*, (Philadelphia (PA), USA), pp. 2324–2327, 1996.

[56] A. Høyland, *Statistisk metodelære.* Tapir forlag, (Trondheim, Norway), 1986.

[57] *The International Phonetic Alphabet.* [online], 1996. [cited 2002-03-01]. URL: `http://www2.arts.gla.ac.uk/IPA/ipachart.html`.

[58] F. Jelinek, "Workshops on large vocabulary conversational speech recognition at Johns Hopkins," in *Proc. ICSLP-96*, (Philadelphia (PA), USA), pp. Addendum 32–33, 1996.

[59] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, pp. 532–556, 1976.

[60] F. T. Johansen, *Global discriminative modelling for automatic speech recognition.* PhD thesis, NTH (Norwegian Institute of Technology), 1996.

[61] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Transactions on Signal Processing*, vol. 40, pp. 3043–3054, 1992.

[62] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiuyang, and Z. Sen, "What kind of pronunciation variation is hard for triphones to model?," in *Proc. ICASSP-2001*, (Salt Lake City (UT), USA), pp. 577–580, 2001.

[63] D. Jurafsky and J. H. Martin, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Prentice Hall, 2000.

[64] S. Katagiri, B.-H. Juang, and C.-H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," *Proceedings of the IEEE*, vol. 86, pp. 2345–2373, 1998.

[65] J. M. Kessens, H. Strik, and C. Cucchiarini, “A bottom-up method for obtaining information about pronunciation variation,” in *Proc. ICSLP-2000*, (Beijing, China), pp. I:274–277, 2000.

[66] J. M. Kessens, M. Wester, and H. Strik, “Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation,” *Speech Communication*, vol. 29, pp. 193–207, 1999.

[67] F. Korkmazskiy and B.-H. Juang, “Discriminative training of the pronunciation networks,” in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, (Santa Barbara (CA), USA), pp. 137–144, 1997.

[68] F. Korkmazskiy and C.-H. Lee, “Generating alternative pronunciations from a dictionary,” in *Proc. EUROSPEECH-99*, (Budapest, Hungary), pp. 491–494, 1999.

[69] H.-K. J. Kuo, E. Fosler-Lussier, H. Jiang, and C.-H. Lee, “Discriminative training of language models for speech recognition,” in *Proc. ICASSP-2002*, (Orlando (FL), USA), pp. 325–328, 2002.

[70] K. Kvale, *Segmentation and labelling of speech.* PhD thesis, NTH (Norwegian Institute of Technology), 1993.

[71] P. Ladefoged, *A course in phonetics.* Harcourt Brace College Publishers, 1993.

[72] J. Laver, *Principles of phonetics.* Cambridge University Press, 1994.

[73] W. A. Lea, *Trends in speech recognition.* Prentice Hall, 1980.

[74] C.-H. Lee and Q. Huo, “On adaptive decision rules and decision parameter adaptation for automatic speech recognition,” *Proceedings of the IEEE*, vol. 88, pp. 1241–1269, 2000.

[75] C.-H. Lee, C.-H. Lin, and B.-H. Juang, “A study on speaker adaptation of the parameters of continous density hidden Markov models,” *IEEE Transactions on signal processing*, vol. 39, pp. 806–814, April 1991.

[76] C.-H. Lee, F. K. Soong, and K. K. Paliwal, *Automatic speech and speaker recognition: Advanced topics.* Kluwer, 1996.

[77] L. Lee and R. Rose, “A frequency warping approach to speaker normalization,” *IEEE Transactions on speech and audio processing*, vol. 6, pp. 49–60, January 1998.

[78] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP-96*, (Atlanta (GA), USA), pp. 353–356, 1996.

[79] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[80] L. Mayfield Tomokiyo, "Lexical and acoustic modeling of non-native speech in LVCSR," in *Proc. ICSLP-2000*, (Beijing, China), pp. IV:346–349, 2000.

[81] L. Mayfield Tomokiyo, "Linguistic properties of non-native speech," in *Proc. ICASSP-2000*, (Istanbul, Turkey), pp. 1335–1338, 2000.

[82] D. McAllaster, L. Gillick, F. Scattone, and M. Newman, "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," in *Proc. ICSLP-98*, (Sydney, Australia), pp. 1847–1850, 1998.

[83] R. Muhr, R. Höldrich, and E. Wächter-Kollpacher, "The pronouncing dictionary of Austrian German and the other major varieties of German – A phonetic resources database on the pronunciation of German," in *Proc. LREC-2002*, (Las Palmas de Gran Canaria, Spain), pp. 1284–1289, 2002.

[84] C. H. Nakatani and J. Hirschberg, "A corpus-based study of repair cues in spontaneous speech," *Journal of the Acoustical Society of America*, vol. 95(3), pp. 1603–1616, March 1994.

[85] *NIST Spoken Language Technology Evaluations.* [online], 2002. [cited 2002-07-01]. URL: `http://www.nist.gov/speech/tests/`.

[86] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," in *Proc. ICASSP-92*, (San Francisco (CA), USA), pp. I:521–523, 1992.

[87] S. Oviatt, "Predicting spoken disfluencies during human-computer interaction," *Computer Speech and Language*, vol. 9, pp. 19–35, 1995.

[88] H. Printz and P. Olsen, "Theory and practice of acoustic confusability," *Computer Speech and Language*, vol. 16, pp. 131–164, 2002.

[89] *CALLHOME American English Lexicon (PRONLEX).* [online description], 1995. [cited 2002-03-01]. URL:
`http://morph.ldc.upenn.edu/Catalog/LDC97L20.html`.

[90] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition.* Prentice Hall, 1993.

[91] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals.* Prentice Hall, 1978.

[92] W. Reichl and W. Chou, "Decision tree state tying based on segmental clustering for acoustic modeling," in *Proc. ICASSP-98*, (Seattle (WA), USA), pp. 801–804, 1998.

[93] B. Resch, "Data driven pronunciation modeling for large vocabulary spontaneous speech recognition," Master's thesis, Graz University of Technology, 2002.

[94] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Communication*, vol. 29, pp. 209–224, 1999.

[95] M. D. Riley and A. Ljolje, *Automatic speech and speaker recognition: Advanced topics*, ch. Automatic generation of detailed pronunciation lexicons, pp. 285–301. Kluwer, 1996.

[96] B. D. Ripley, *Pattern recognition and neural networks.* Cambridge University Press, 1996.

[97] *SAMPA computer readable phonetic alphabet.* [online], 2000. [cited 2002-03-01]. URL: `http://www.phon.ucl.ac.uk/home/sampa/home.htm`.

[98] H. Schramm and X. L. Aubert, "Efficient integration of multiple pronunciations in a large vocabulary decoder," in *Proc. ICASSP-2000*, (Istanbul, Turkey), pp. 1659–1662, 2000.

[99] H. Schramm and P. Beyerlein, "Towards discriminative lexicon optimization," in *Proc. EUROSPEECH-2001*, (Aalborg, Denmark), pp. 1457–1460, 2001.

[100] T. Schultz and I. Rogina, "Acoustic and language modeling of human and nonhuman noises for human-to-human spontaneous speech recognition," in *Proc. ICASSP-95*, (Detroit (MI), USA), pp. 293–296, 1995.

[101] K. Shinoda and C.-H. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proc. IEEE Workshop on Automatic*

Speech Recognition and Understanding, (Santa Barbara (CA), USA), pp. 381–388, 1997.

[102] B. Shneiderman, *Designing the user interface: Strategies for effective Human-Computer Interaction.* Addison-Wesley, 1998.

[103] E. Shriberg, "Disfluencies in Switchboard," in *Proc. ICSLP-96*, (Philadelphia (PA), USA), pp. Addendum 11–12, 1996.

[104] E. Shriberg and A. Stolcke, "Word predictability after hesitations: A corpus-based study," in *Proc. ICSLP-96*, (Philadelphia (PA), USA), pp. 1868–1871, 1996.

[105] R. Singh, B. Raj, and R. M. Stern, "Automatic generation of phone sets and lexical transcriptions," in *Proc. ICASSP-2000*, (Istanbul, Turkey), pp. 1691–1694, 2000.

[106] R. Singh, B. Raj, and R. M. Stern, "Structured redefinition of sound units by merging and splitting for improved speech recognition," in *Proc. ICSLP-2000*, (Beijing, China), pp. III:151–154, 2000.

[107] O. Siohan, T. A. Myrvoll, and C.-H. Lee, "Structural maximum *a posteriori* linear regression for fast HMM adaptation," *Computer Speech and Language*, vol. 16, pp. 5–24, 2002.

[108] M. Siu and M. Ostendorf, "Modeling disfluencies in conversational speech," in *Proc. ICSLP-96*, (Philadelphia (PA), USA), pp. 386–389, 1996.

[109] T. Sloboda and A. Waibel, "Dictionary learning for spontaneous speech recognition," in *Proc. ICSLP-96*, (Philadelphia (PA), USA), pp. 2328–2331, 1996.

[110] R. W. Sproat and J. P. Olive, "Text-to-speech synthesis," *AT&T Technical Journal*, vol. 74, pp. 35–44, 1995.

[111] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *Proc. ICSLP-96*, (Philadelphia (PA), USA), pp. 1005–1008, 1996.

[112] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *Proc. ICASSP-96*, (Atlanta (GA), USA), pp. 405–408, 1996.

[113] H. Strik, "Pronunciation adaptation at the lexical level," in *Proc. ISCA ITRW Adaptation methods for speech recognition*, (Sophia-Antipolis, France), pp. 123–130, 2001.

[114] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: overview and comparison of methods," in *Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, (Rolduc, the Netherlands), pp. 137–144, 1998.

[115] H. Strik, C. Cucchiarini, and J. M. Kessens, "Comparing the performance of two CSRs: How to determine the significance level of the difference," in *Proc. EUROSPEECH-2001*, (Aalborg, Denmark), pp. 2091–2094, 2001.

[116] T. Svendsen, K. K. Paliwal, E. Harborg, and P. O. Husøy, "An improved sub-word based speech recognizer," in *Proc. ICASSP-89*, (Glasgow, Scotland), pp. 108–111, 1989.

[117] G. Tajchman, E. Fosler, and D. Jurafsky, "Building multiple pronunciation models for novel words using exploratory computational phonology," in *Proc. EUROSPEECH-95*, (Madrid, Spain), pp. 2247–2250, 1995.

[118] D. Torre, L. Villarrubia, J. M. Elvira, and L. Hernandez-Gomez, "Automatic alternative transcription generation and vocabulary selection for flexible word recognizers," in *Proc. ICASSP-97*, (Munich, Germany), pp. 1463–1466, 1997.

[119] D. Van Compernolle, "Recognizing speech of goats, wolves, sheep and … non-natives," *Speech Communication*, vol. 35, pp. 71–79, 2001.

[120] N. D. Warakagoda, *Nonlinear dynamical systems for automatic speech recognition.* PhD thesis, NTNU (Norwegian University of Science and Technology), 2001.

[121] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass, "Effect on speaking style on LVCSR performance," in *Proc. ICSLP-96*, (Philadelphia (PA), USA), pp. Addendum 16–19, 1996.

[122] M. Wester and E. Fosler-Lussier, "A comparison of data-derived and knowledge-based modeling of pronunciation variation," in *Proc. ICSLP-2000*, (Beijing, China), pp. I:270–273, 2000.

[123] M. Wester, J. M. Kessens, and H. Strik, "Pronunciation variation in ASR: Which variation to model?," in *Proc. ICSLP-2000*, (Beijing, China), pp. IV:488–491, 2000.

[124] M. Wolff, M. Eichner, and R. Hoffmann, "Automatic learning and optimization of pronunciation dictionaries," in *Proc. ISCA ITRW Adaptation methods for speech recognition*, (Sophia-Antipolis, France), pp. 159–162, 2001.

[125] P. C. Woodland, "Speaker adaptation for continuous density HMMs: A review," in *Proc. ISCA ITRW Adaptation methods for speech recognition*, (Sophia-Antipolis, France), pp. 11–19, 2001.

[126] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using HTK," in *Proc. ICASSP-94*, (Adelaide, Australia), pp. II:125–128, 1994.

[127] *Wall Street Journal speech database (WSJ).* [online description], 1993. [cited 2002-03-01]. URL:
`http://morph.ldc.upenn.edu/Catalog/LDC94S13A.html`.

[128] S.-L. Wu, M. L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proc. ICASSP-97*, (Munich, Germany), pp. 987–990, 1997.

[129] Q. Yang and J.-P. Martens, "Data-driven lexical modeling of pronunciation variations for ASR," in *Proc. ICSLP-2000*, (Beijing, China), pp. I:417–420, 2000.

[130] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Processing Magazine*, pp. 45–57, September 1996.

[131] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *HTK Version 3.0.* [online description], 2000. [cited 2002-03-01]. URL: `http://htk.eng.cam.ac.uk/`.

[132] V. Zue, "Conversational interfaces: Advances and challenges," in *Proc. EUROSPEECH-97*, (Rhodes, Greece), pp. KN9–KN18, 1997.