

**Discriminative and Bayesian techniques for
hidden Markov model speech recognition
systems**

by

Darryl William Purnell

Submitted in partial fulfillment of the requirements for the degree of

PHILOSOPHIAE DOCTOR (ENGINEERING)

in the Faculty of Engineering

UNIVERSITY OF PRETORIA

February 2001

Summary

The collection of large speech databases is not a trivial task (if done properly). It is not always possible to collect, segment and annotate large databases for every task or language. It is also often the case that there are imbalances in the databases, as a result of little data being available for a specific subset of individuals. An example of one such imbalance is the fact that there are often more male speakers than female speakers (or *vice-versa*). If there are, for example, far fewer female speakers than male speakers, then the recognizers will tend to work poorly for female speakers (as compared to performance for male speakers).

This thesis focuses on using Bayesian and discriminative training algorithms to improve continuous speech recognition systems in scenarios where there is a limited amount of training data available. The research reported in this thesis can be divided into three categories:

- Overspecialization is characterized by good recognition performance for the data used during training, but poor recognition performance for independent testing data. This is a problem when too little data is available for training purposes. Methods of reducing overspecialization in the minimum classification error algorithm are therefore investigated.
- Development of new Bayesian and discriminative adaptation/training techniques that can be used in situations where there is a small amount of data available.

One example here is the situation where an imbalance in terms of numbers of male and female speakers exists and these techniques can be used to improve recognition performance for female speakers, while not decreasing recognition performance for the male speakers.

- Bayesian learning, where Bayesian training is used to improve recognition performance in situations where one can only use the limited training data available. These methods are extremely computationally expensive, but are justified by the improved recognition rates for certain tasks. This is, to the author's knowledge, the first time that Bayesian learning using Markov chain Monte Carlo methods have been used in hidden Markov model speech recognition.

The algorithms proposed and reviewed are tested using three different datasets (TIMIT, TIDIGITS and SUNSpeech), with the tasks being connected digit recognition and continuous speech recognition. Results indicate that the proposed algorithms improve recognition performance significantly for situations where little training data is available.

Keywords: speech recognition, hidden Markov model training, minimum classification error, Bayesian adaptation, Bayesian learning, maximum a posteriori parameter estimation, sparse data

Samevatting

Die versameling van groot spraakdatabasisse is nie 'n maklike taak nie. Dit is nie altyd moontlik om groot databasisse te versamel, in segmente te verdeel en te annoteer vir elke taak of taal nie. Dit is ook dikwels die geval dat daar 'n wanbalans bestaan in 'n databasis, as gevolg van die onverkrygbaarheid van data vir 'n spesifieke subgroep van individue. Een voorbeeld van so 'n wanbalans is die feit dat daar dikwels meer manssprekers as vrouesprekers is (of andersom). In so 'n geval, sal die herkenner gewoonlik nie goed werk vir vrouens nie, maar sal relatief goed werk vir mans.

Hierdie tesis fokus op die gebruik van Bayes en diskriminerende afrigtingstegnieke om kontinuspraakherkenningstelsels te verbeter in scenarios waar min afrigtdata beskikbaar is. Die navorsing waaroor hier gerapporteer word, kan in drie dele verdeel word:

- Oor-spesialisasie word gekarakteriseer deur goeie herkenning vir die afrigtdata, maar slegte herkenning vir onafhanklike toetsdata. Hierdie probleem ontstaan wanneer te min data beskikbaar is vir afrigtingsdoeleindes. Metodes om oor-spesialisasie in minimum klassifikasiefout afrigting te verminder word dus hier ondersoek.
- Nuwe Bayes en diskriminerende afrigtings- en aanpassingstegnieke om herkenning te verbeter in situasies waar min data beskikbaar is. Een voorbeeld is die situasie waar 'n wanbalans in terme van die aantal vroue- en manssprekers bestaan. Hierdie tegnieke kan gebruik word om die herkenningstempo te verbeter vir die

vrouesprekers.

- Bayes afrigting word gebruik om herkenning te verbeter in situasies waar min data beskikbaar is, en geen taakspesifieke data nie. Hierde metode is uiters berekeningintensief, maar ek glo dat die verbetering in herkenningstempo dit regverdig. Hierdie is, na ons medewete, die eerste keer dat Markov ketting Monte Carlo gebaseerde Bayesian afrigting in verskuilde Markov model spraakherkenningstelsels gebruik word.

Die voorgestelde algoritmes is getoets met drie verskillende spraakdatabasise: TIMIT, TIDIGITS en SUNSpeech. Die take is kontinussyferherkenning en kontinuspraakherkenning. Resultate toon dat die voorgestelde algoritmes goed werk vir situasies waar min afrigtingsdata beskikbaar is.

Sleutelwoorde: spraakherkenning, verskuilde Markov model afrigting, minimum klasifikasiefout, Bayes aanpassing, Bayes afrigting, maksimum a posteriori parameter skatting, min afrigdata

Acknowledgements

First and foremost, I would like to thank Professor Liesbeth Botha for the advice and guidance she gave me during the last three years.

I would also like to thank Christoph Nieuwoudt and Johann Holm, with whom I had many interesting and thought-provoking discussions.

Finally, I would like to thank my family and friends for their support and encouragement throughout.

Contents

1	Introduction	1
1.1	Adaptation	3
1.1.1	Bayesian adaptation	4
1.1.2	Transformation-based adaptation	6
1.1.3	Hybrid adaptation algorithms	7
1.2	Training	7
1.2.1	Discriminative training	8
1.2.2	Bayesian learning	9
1.3	Problem statement	10
1.4	Organization of this thesis	11
1.5	Contributions of this thesis	12
2	Background	14

2.1	Hidden Markov models	14
2.1.1	Feature extraction	15
2.1.2	Continuous density hidden Markov models	16
2.1.3	Training	18
2.1.4	Duration modeling	22
2.1.5	Search	24
2.2	Overtraining	26
2.2.1	The Bias/Variance Dilemma	26
2.3	Experimental procedure	29
2.4	Speech datasets	31
2.4.1	TIMIT	31
2.4.2	TIDIGITS	35
2.4.3	SUNSPEECH	37
3	Minimum classification error training	40
3.1	Introduction	40
3.2	Minimum classification error training	44
3.2.1	Bayes risk	45
3.2.2	Optimization criterion	47

3.2.3	Optimization methods	48
3.2.4	Parameter transformation	50
3.2.5	Parameter adaptation	51
3.3	Embedded MCE	55
3.4	Discussion and Experiments	57
3.4.1	Batch-mode versus online optimization	59
3.4.2	Smoothness of the loss function	60
3.4.3	Need for a zero-one loss function	62
3.4.4	Overtraining in MCE	64
3.4.5	Summary and discussion of results	73
3.5	Summary	76
4	Bayesian adaptation	77
4.1	Introduction	78
4.1.1	Bayes' theorem	79
4.1.2	Sequential nature of Bayes' theorem	80
4.1.3	Bayesian learning and prediction	81
4.1.4	Maximum <i>a-posteriori</i> probability estimate	82
4.1.5	MAP adaptation in speech recognition	83

5.1.1	Bayes' theorem	155
5.1.2	Bayesian learning and prediction	156
5.1.3	Bias/variance problem	159
5.1.4	Hierarchical models	160
5.2	Monte Carlo methods	160
5.2.1	Gibbs sampling	162
5.2.2	The stochastic dynamics method	163
5.2.3	The hybrid Monte Carlo algorithm	168
5.3	Implementation of Bayesian HMM learning	169
5.3.1	HMM constraints	170
5.3.2	HMM prior and hyperparameters	171
5.3.3	Refreshing hyperparameters	176
5.3.4	Implementation of stochastic dynamics method	177
5.3.5	Recognition	178
5.4	Experiments	181
5.4.1	SUNSpeech	182
5.4.2	TIMIT	189
5.4.3	TIDIGITS	192

5.4.4	Summary of results	194
5.5	Summary	195
6	Conclusion	197
6.1	Overview	197
6.2	Summary by Chapter	199
6.3	Future research	202
	Bibliography	204
A	Probability distributions	217
A.1	The normal distribution	217
A.2	The Wishart distribution	218
A.3	Dirichlet distribution	219
A.4	The gamma distribution	219
A.5	Conjugate families of distributions	220