CROSS-LANGUAGE ACOUSTIC ADAPTATION FOR AUTOMATIC SPEECH RECOGNITION

by

Christoph Nieuwoudt

Submitted in partial fulfilment of the requirements for the degree

Philosophiae Doctor

in the

Faculty of Engineering

UNIVERSITY OF PRETORIA

April 2000

Summary

Speech recognition systems have been developed for the major languages of the world, yet for the majority of languages there are currently no large vocabulary continuous speech recognition (LVCSR) systems. The development of an LVCSR system for a new language is very costly, mainly because a large speech database has to be compiled to robustly capture the acoustic characteristics of the new language.

This thesis investigates techniques that enable the re-use of acoustic information from a source language, in which a large amount of data is available, in implementing a system for a new target language. The assumption is that too little data is available in the target language to train a robust speech recognition system on that data alone, and that use of acoustic information from a source language can improve the performance of a target language recognition system.

Strategies for cross-language use of acoustic information are proposed, including training on pooled source and target language data, adaptation of source language models using target language data, adapting multilingual models using target language data and transforming source language data to augment target language data for model training. These strategies are allied with Bayesian and transformation-based techniques, usually used for speaker adaptation, as well as with discriminative learning techniques, to present a framework for cross-language re-use of acoustic information. Extensions to current adaptation techniques are proposed to improve the performance of these techniques specifically for cross-language adaptation. A new technique for transformation-based adaptation of variance parameters and a cost-based extension of the minimum classification error (MCE) approach are proposed.

Experiments are performed for a large number of approaches from the proposed framework for cross-language re-use of acoustic information. Relatively large amounts of English speech data are used in conjunction with smaller amounts of Afrikaans speech data to improve the performance of an Afrikaans speech recogniser. Results indicate that a significant

reduction in word error rate (between 26% and 50%, depending on the amount of Afrikaans data available) is possible when English acoustic data is used in addition to Afrikaans speech data from the same database (i.e. both sets of data were recorded under the same conditions and the same labelling process was used). For same-database experiments, best results are achieved for approaches that train models on pooled source and target language data and then perform further adaptation of the models using Bayesian or discriminative techniques on target language data only. Experiments are also performed to evaluate the use of English data from a different database than the Afrikaans data. Peak reductions in word error rate of between 16% and 35% are delivered, depending on the amount of Afrikaans data available. Best results are achieved for an approach that performs a simple transformation of source model parameters using target language data, and then performs Bayesian adaptation of the transformed model on target language data.

Keywords: multilingual speech recognition, cross-language acoustic adaptation, Bayesian adaptation, transformation-based adaptation, minimum classification error adaptation

Opsomming

Spraakherkenningstelsels is reeds ontwikkel vir die groot tale van die wêreld, maar vir die meerderheid van tale bestaan daar tans geen groot-woordeskat kontinuespraakherkenningstelsels nie. Die ontwikkeling van 'n groot-woordeskat kontinuespraakherkenningstelsel vir 'n nuwe taal is baie duur, hoofsaaklik omdat 'n groot databasis opgestel moet word om die akoestiek van 'n nuwe taal op robuuste wyse te vervat.

Die tesis ondersoek tegnieke wat die hergebruik van akoestiese inligting van 'n brontaal, waarvoor 'n groot hoeveelheid data beskikbaar is, toe te laat in die implementering van 'n stelsel vir 'n nuwe teikentaal. Die aanname word gemaak dat te min data beskikbaar is vir die teikentaal om 'n robuuste spraakherkenningstelsel mee af te rig, en dat akoestiese inligting in 'n brontaal gebruik kan word om die herkenning van 'n teikentaalstelsel te verbeter.

Strategieë vir die gebruik van akoestiese inligting oor taalgrense heen word voorgestel en sluit in: afrigting op gepoelde brontaal- en teikentaaldata, aanpassing van brontaalmodelle met teikentaaldata, aanpassing van multitaalmodelle met teikentaaldata en transformasie van brontaaldata om teikentaaldata aan te vul vir afrigting van modelle. Hierdie strategieë word met Bayes en transformasie tegnieke, wat gewoonlik vir sprekeraanpassing gebruik word, en diskriminerende afrigtingstegnieke gebruik om 'n raamwerk vir die gebruik van akoestiese inligting oor taalgrense daar te stel. Uitbreidings van bestaande tegnieke word voorgestel om die herkenning van die tegnieke te verbeter vir kruis-taal aanpassing. 'n Nuwe tegniek vir transformasie van variansieparameters en 'n kostegebaseerde uitbreiding van die minimum klassifikasiefout tegniek word voorgestel.

Eksperimente word uitgevoer vir 'n groot aantal benaderings uit die voorgestelde raamwerk vir kruis-taal hergebruik van akoestiese inligting. Relatief groot hoeveelhede Engelse spraakdata word gebruik tesame met kleiner hoeveelhede Afrikaanse spraakdata om die werkverrigting van 'n Afrikaanse herkenningstelsel te verbeter. Die resultate dui aan dat 'n beduidende vermindering in woordfouttempo (tussen 26% en 50%, afhangende van die hoe-

veelheid Afrikaanse data wat beskikbaar is) moontlik is wanneer Engelse data tesame met Afrikaanse data van dieselfde databasis gebruik word (dit wil sê beide datastelle is onder dieselfde toestande opgeneem en dieselfde etiketteringsproses is gebruik). Vir dieselfde-databasis eksperimente word die beste resultate bereik vir benaderings wat modelle afrig op gepoelde brontaal- en teikentaaldata, en wat dan verdere afrigting van modelle volgens Bayes of diskriminasiegebaseerde tegnieke uitvoer met slegs teikentaaldata. Eksperimente word ook uitgevoer om die gebruik van Engelse spraakdata van 'n verskillende databasis as die Afrikaanse data te evalueer. Piek verminderings in fouttempo tussen 16% en 35% word gelewer, afhangende van die hoeveelheid Afrikaanse data wat beskikbaar is. Beste resultate word bereik vir 'n benadering wat 'n eenvoudige transformasie van bronmodelparameters uitvoer met gebruik van teikentaaldata, en dan Bayes aanpassing van die getransformeerde model uitvoer met teikentaaldata.

Sleutelwoorde: multitaalspraakherkenning, kruis-taal akoestiese aanpassing, Bayes aanpassing, parameter transformasie, minimum klassifikasiefout aanpassing

Contents

List of Abbreviations

ANN	Artificial neural network
CBLF	Cost-based loss function
CBMM	Cost-based misclassification measure
CDHMM	Continuous density hidden Markov model
CDR	Connected digit recognition
CMS	Cepstral mean subtraction
CRBMM	Cost and reward-based misclassification measure
DCT	Discrete cosine transform
DTW	Dynamic time warping
EM	Expectation maximisation
ESHMM	Expanded state hidden Markov model
FFT	Fast Fourier transform
GPD	Generalised probabilistic descent
HMM	Hidden Markov model
LDA	Linear discriminant analysis
LVCSR	Large vocabulary continuous speech recognition
MAP	Maximum *a posteriori*
MAPLR	Maximum *a posteriori* linear regression
MCE	Minimum classification error
MFCC	Mel-scaled cepstral coefficient
ML	Maximum likelihood
MLLR	Maximum likelihood linear regression
MLP	Multi-layer perceptron
MMI	Maximum mutual information
MSE	Minimum square error
PCA	Principal component analysis
SA	Speaker adaptive
SD	Speaker dependent
SI	Speaker independent
VTLN	Vocal tract length normalisation
WER	Word error rate

List of Symbols

$\mathbf{A}$	a state transition probability matrix
D	the feature dimension
K	the number of mixtures in a state
M	the number of classes/HMMs
N	the number of states in an HMM
$\mathbf{R}$	a Gaussian precision matrix
$\mathbf{S}$	a sample variance
T	the number of time frames in an observation sequence
$\mathbf{W}$	a transformation matrix
$\mathbf{X}$	an observation sequence
a	an HMM transition probability
c	a mixture weight
$\mathbf{q}$	a state sequence
$\mathbf{m}$	the mean vector of a Gaussian prior distribution
r	a Gaussian precision value
v	a parameter of the mixture weight Dirichlet prior distribution
$\mathbf{v}$	the target variance parameter
w	the relative variance of the mean in the prior
$\mathbf{x}$	a feature vector
$\mathbf{\Lambda}$	the parameters of a set of HMMs
$\mathbf{\Sigma}$	a Gaussian covariance matrix
$\mathbf{\Upsilon}$	the precision of the covariance Wishart prior distribution
α	a parameter of a gamma distribution
β	a parameter of a gamma distribution
γ	the state/mixture occupancy variable
	the slope of the sigmoid in the MCE loss function
ϵ	the MCE update parameter
ζ	the cost associated with a phoneme misclassification
η	a parameter of the transition probability Dirichlet prior distribution
	the scaling factor in the MCE misclassification measure
θ	the offset in the MCE loss function
κ	the reward associated with a phoneme classification decision
$\boldsymbol{\lambda}$	the parameters of an HMM
$\boldsymbol{\mu}$	a Gaussian mean vector
ξ	the transition count variable
$\boldsymbol{\sigma}$	a Gaussian variance vector (for diagonal covariance)
$\boldsymbol{v}$	the extended mean vector
τ	the variance of the mean in the prior (univariate)
ϖ	a learning rate parameter