The University Of Sheffield

Methods for Addressing Data Diversity in Automatic Speech Recognition

Mortaza Doulaty Bashkand

Machine Intelligence for Natural Interfaces (MINI) Lab,
Speech and Hearing (SPandH) Research Group,
Department of Computer Science,
University of Sheffield

This dissertation is submitted on January 2017 for the degree of
Doctor of Philosophy

"*Everything not saved will be lost.*"
–Nintendo "Quite Screen" message

Abstract

The performance of speech recognition systems is known to degrade in mismatched conditions, where the acoustic environment and the speaker population significantly differ between the training and target test data. Performance degradation due to the mismatch is widely reported in the literature, particularly for diverse datasets.

This thesis approaches the mismatch problem in diverse datasets with various strategies including data refinement, variability modelling and speech recognition model adaptation. These strategies are realised in six novel contributions.

The first contribution is a data subset selection technique using likelihood ratio derived from a target test set quantifying mismatch. The second contribution is a multi-style training method using data augmentation. The existing training data is augmented using a distribution of variabilities learnt from a target dataset, resulting in a matched set.

The third contribution is a new approach for genre identification in diverse media data with the aim of reducing the mismatch in an adaptation framework.

The fourth contribution is a novel method which performs an unsupervised domain discovery using latent Dirichlet allocation. Since the latent domains have a high correlation with some subjective meta-data tags, such as genre labels of media data, features derived from the latent domains are successfully applied to the genre and broadcast show identification tasks.

The fifth contribution extends the latent modelling technique for acoustic model adaptation, where latent-domain specific models are adapted from a base model. As the sixth contribution, an alternative adaptation approach is proposed where subspace adaptation of deep neural network acoustic models is performed using the proposed latent-domain aware training procedure.

All of the proposed techniques for mismatch reduction are verified using diverse datasets. Using data selection, data augmentation and latent-domain model adaptation methods the mismatch between training and testing conditions of diverse ASR systems are reduced, resulting in more robust speech recognition systems.

Declaration

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specified in the text. This dissertation is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other university. This dissertation does not exceed the prescribed limit of 80 000 words.

Mortaza Doulaty Bashkand
January 2017

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Thomas Hain. Without his continuous support and endless guidance this thesis would not have been possible.

I would also like to thank Oscar Saz and Raymond W. M. Ng for their kind support and the useful discussions we had throughout my PhD. The current and past members of the MINI group were all very helpful during my studies and I thank them all.

I wish to thank Richard Rose and Olivier Siohan of Google Inc. New York for having me as an intern in summer 2015.

My PhD was supported by the Engineering and Physical Sciences Research Council (EPSRC) programme grant EP/I031022/1 Natural Speech Technology (NST). I am grateful to the NST and EPSRC for the studentship they provided to fund my PhD research and to the Department of Computer Science of the University of Sheffield for funding the overseas element of the tuition fees.

I had a wonderful time during lunch breaks everyday chatting, watching videos and discussing the latest world and technology news with Rosanna Milner and Salil Deena.

Last but not least, I would like to thank my mother and deceased father for their unconditional support and devotion to their son. My deepest appreciation goes to my partner, Fariba Yousefi. Her wholehearted support made my PhD life a lot easier.

Contents

List of Acronyms

AM Acoustic Model

ASR Automatic Speech Recognition

BBC British Broadcasting Corporation

BN Bottleneck

BP Back Propagation

CAT Cluster Adaptive Training

CD Context Dependent

CE Cross-Entropy

CHN Cepstral Histogram Normalisation

CI Context Independent

CMLLR Constrained Maximum Likelihood Linear Regression

CMN Cepstral Mean Normalisation

CMVN Cepstral Mean and Variance Normalisation

CRF Conditional Random Field

CTC Connectionst Temporal Classification

CTS Conversational Telephone Speech

CVN Cepstral Variance Normalisation

DAT Device Aware Training

DBN Deep Belief Network

DNN Deep Neural Network

EM Expectation Maximisation

fDLR Feature Discriminative Linear Regression

fMLLR Feature-space Maximum Likelihood Linear Regression

GD Gender Dependent

GMM Gaussian Mixture Model

HMM Hidden Markov Model

idf Inverse Document Frequency

IR Information Retrieval

iVector Identity Vector

KLD Kullback-Leibler Divergence

LDA Latent Dirichlet Allocation

LHN Linear Hidden Network

LIN Linear Input Network

LM Language Model

LON Linear Output Network

LSI Latent Semantic Indexing

MAP Maximum A Posteriori

MCMC Markov Chain Monte Carlo

MFCC Mel-Frequency Cepstral Coefficients

MGB Multi-Genre Broadcast

ML Maximum Likelihood

MLLR Maximum Likelihood Linear Regression

MMI Maximum Mutual Information

MPE Minimum Phone Error

MTR Multistyle Training

NAT Noise Aware Training

oDLR Output-feature Discriminative Linear Regression

PCA Principle Component Analysis

PDF Probability Density Function

PLP Perceptual Linear Prediction

pSLI Probabilistic Latent Semantic Indexing

RNN Recurrent Neural Network

ROVER Recognizer Output Voting Error Reduction

SA Speaker Adaptive

SAT Speaker Adaptive Training

SD Speaker Dependent

SGD Stochastic Gradient Descent

SGMM Subspace Gaussian Mixture Model

SI Speaker Independent

SVD Singular Value Decomposition

SVM Support Vector Machine

tf Term Frequence

tf-idf Term Frequency - Inverser Document Frequency

VQ Vector Quantisation

VTLN Vocal Tract Length Normalisation

WER Word Error Rate

WMER Word Matching Error Rate

List of Figures

List of Tables

CHAPTER 1

Introduction

Automatic speech recognition (ASR) is the task of transcribing spoken language into text. It has a very wide range of applications, including but not limited to: voice dictation, voice command and control, home automation, personal assistants, automatic translation, language learning, hands-free computing, automatic subtitling, interactive voice responders and medical reporting. As this technology improves and produces fewer errors, its application domain extends.

ASR can be considered as a mapping function that maps a variable length acoustic signal into a variable length sequence of words:

$$f(\mathcal{O}) = \mathcal{W} \tag{1.1}$$

where $\mathcal{O}$ is an acoustic signal and $\mathcal{W}$ is the sequence of words spoken in the acoustic signal. Statistical approaches can be used for solving this problem and the mapping function can be defined in a probabilistic way:

$$\hat{\mathcal{W}} = \arg\max_{\mathcal{W} \in \mathscr{L}} P(\mathcal{W}|\mathcal{O}) \tag{1.2}$$

this changes the ASR problem to a search problem: finding the most likely sequence of words from all of the possible word sequences of language $\mathscr{L}$ given the acoustic signal. Applying Bayes' theorem to equation 1.2 yields:

$$\hat{\mathcal{W}} = \arg\max_{\mathcal{W} \in \mathscr{L}} \frac{P(\mathcal{O}|\mathcal{W})P(\mathcal{W})}{P(\mathcal{O})} = \arg\max_{\mathcal{W} \in \mathscr{L}} P(\mathcal{O}|\mathcal{W})P(\mathcal{W}) \tag{1.3}$$

where $P(\mathcal{O}|\mathcal{W})$ is the observation likelihood computed by an acoustic model (AM) and $P(\mathcal{W})$ is the prior probability of the word sequence computed by a language model (LM). Since the probability of the observation itself, $P(\mathcal{O})$, is independent from the most likely word sequence, it can be omitted from the search. The reason for

using the Bayes' theorem is that the probabilities on the left-hand side of equation 1.3 is not directly computable.

Finding the most likely sequence of words, $\hat{\mathcal{W}}$, is called decoding. The Viterbi algorithm (Viterbi, 1967) is usually used for decoding and during the search, scores from the AM and LM are combined together. For practical reasons and to speed up the search process, parts of the search space are usually pruned.

To assess performance of ASR systems, word error rate (WER) is the commonly applied metric. It is based on the minimum edit distance between the output of the ASR system and the reference text. The error is computed as the ratio between the total count of insertions, deletions and substitutions required to convert the hypothesised text to the reference text vs. total number of words in the reference text.

1.1 Acoustic and language modelling

1.1.1 Acoustic modelling

The observation sequence is usually sampled into frames. These frames are then transformed into some form of spectral representation such as Mel-frequency cepstral coefficients (MFCC) features. The AM is then used to compute the likelihood of the feature vectors given some linguistic units. There are several approaches for acoustic modelling and two of the most popular approaches will be introduced briefly in this section.

Using a lexicon each word can be represented as a sequence of sub-word units, such as phones. Usually these sub-word units are modelled with five-state hidden Markov models (HMM) where the first and last states are non-emitting states which are used for concatenating these units. Considering the coarticulation effect, where each phone is pronounced differently depending on the neighbouring phones, in modern ASR systems context-dependent (CD) phones are modelled instead of context-independent (CI) phones. Since there are exponentially more CD phones compared to CI phones, and not all phone combinations are seen in the training data or sometimes not even possible at all, the HMM states of the CD phones are tied together for parameter sharing.

Gaussian mixture models (GMM), deep neural networks (DNN), support vector machines (SVM) or conditional random fields (CRF) can be used to model the probability density function (PDF) of emitting states of the HMMs (Jurafsky and Martin, 2000). Prior to 2012, GMMs were very popular for modelling the PDFs in acoustic modelling. With the raise of deep learning in 2012, several studies showed

that DNNs can be used to further improve acoustic modelling (Hinton et al., 2012; Yu and Deng, 2015).

With GMMs, each HMM state is usually modelled with an 8–64 component mixture model. Parameters of the model (including GMM weights, means and co-variances and HMM state transition probabilities) are learnt using the Baum-Welch algorithm. It uses the expectation maximisation (EM) algorithm to find the maximum likelihood (ML) estimate of the model parameters given the observation vectors.

To train the acoustic models the transcripts of the speech segments are commonly provided at the word level, however the modelling is performed on sub-word units such as phones. Usually a uniform distribution of the phones in the utterance is assumed and the initial models are trained with this initial alignment. Then, these models are used to acquire better state-level alignments and re-train more accurate models.

Initial proposals to use neural networks in HMM-based speech recognisers date back to the early 90's (Renals et al., 1994). However, the success of those early attempts were not comparable to the state-of-the-art GMM-HMM systems. Around 2012, neural networks became popular again and several studies promoted the use of deep neural networks in speech recognition with some promising results. This was mostly because of having more computation power and more data available. It was shown that the use of DNNs could reduce the WER around 15–25% relative compared to the conventional GMM-HMM systems (Hinton et al., 2012; Yu and Deng, 2015).

There are two popular approaches to integrate DNNs with HMM-based speech recognition systems: bottleneck and hybrid setups (Grézl et al., 2007; Renals et al., 1994). In both setups, a DNN is trained for classifying the frames into phone classes or tied HMM states of the CD phones. The state level alignment for training these DNNs is usually acquired by an initial GMM-HMM system.

In the bottleneck setup, the DNN acts as a feature extractor for the GMM-HMM system. Usually a bottleneck layer, which has a smaller number of neurons compared to other layers, is used in the neural network and outputs of the neurons from the bottleneck layer (either before or after the activation function) are used as a new representation of the input frames. These features are then used in a conventional GMM-HMM system as input features (either in solo mode or by augmenting the existing MFCC features).

With hybrid systems, GMMs are replaced by DNNs. In this setup, emissions of the HMM states are modelled by DNNs. Since HMM-based speech recognition systems require the likelihood computation and DNNs output posterior probabilities,

these scores are converted to likelihood scores using Bayes' theorem.

An alternative approach for solving the speech recognition problem is the so-called end-to-end systems. Unlike phonetic-based systems where different components such as the AM, LM and lexicon are trained separately, the end-to-end techniques try to learn all of the components jointly. One of the first attempts was the connectionist temporal classification (CTC) approach proposed by Graves et al. (2006) which used recurrent neural networks (RNN) and CTC objective function to jointly learn the lexicon and AM without any explicit frame-level alignment. Other approaches also tried to map the acoustic signal directly to characters or even words (Bahdanau et al., 2016; Chan et al., 2016).

The scope of this thesis will be limited to HMM-based speech recognition systems and the ASR related contributions will be evaluated using DNN-based acoustic models.

1.1.2 Language modelling

In equation 1.3, $P(\mathcal{W})$ is the prior probability of the word sequence which is modelled by a language model. N-grams are a simple form of count-based LMs and can be used to assign a probability to a word sequence or find the conditional probability of the next word given a history of $n-1$ words. They are widely used in ASR, hand writing recognition and machine translation.

With advances in deep learning, neural network based LMs are outperforming n-grams in many tasks and will most likely replace them (Mikolov et al., 2010). More specifically, neural networks with recurrent units have better modelling capabilities for language modelling compared to feed-forward networks and most of the state-of-the-art LMs are based on RNNs (Mikolov et al., 2011). Since the focus of this thesis will be on acoustic modelling, LMs will not be studied in depth.

1.2 Motivation

Training acoustic models is usually considered as a supervised learning task which requires labelled training data. Speech data has various characteristics such as type of speech (fluent, natural and conversational), acoustic environment (noisy vs. clean), accent of the speakers, etc. These characteristics vaguely define the conventional domains in ASR (Deng and Li, 2013). However, the concept of a domain is complex and not bound to specific criteria. Training AMs from utterances that match the target speaker population, speaking style or acoustic environment is generally considered to be the easiest way to optimise ASR performance. Furthermore, speech recognition performance is known to degrade when the acoustic environment

and the speaker population in the target utterances are significantly different from the conditions represented in the training data. However, the matched ASR systems usually require in-domain data, e.g. data which has the same underlying distribution as the target domain data. Mismatch happens when the underlying distributions of the training and test data are not the same and depending on level of mismatch, performance of the ASR systems can degrade significantly.

There are several approaches to address the mismatch problem in different levels of the ASR training process. One approach is to create a matched training set to a target test set and train ASR systems with the matched training set. For creating a matched set, often similarity measures are used and data selection is performed based on maximising the similarity measure. The training set can be selected from a fixed set of utterances, where the mismatch minimisation problem turns into a data subset selection problem. The objective in the data subset selection problem is to select a subset of utterances from a pool of available utterances that matches a target set. If the pool of utterances can be extended by generating new samples or augmenting existing samples, then this problem turns into a data generation/augmentation problem. With this approach the training set can be extended by various data generation and augmentation techniques to create a matched training set for a target test set.

With data selection/augmentation techniques, the task is to select/augment the existing data for training the models from scratch. In the case of having some already trained and possibly mismatched models, an alternative approach is to update the model parameters to better match the target test set. This mismatch minimisation technique is typically called *model adaptation*.

Reducing the mismatch typically improves the performance of ASR systems and the main motivation of this thesis is to study how different techniques can be used to reduce the mismatch between training and test conditions. For this purpose new techniques for data selection and augmentation are proposed. Furthermore, new representations of acoustic variability present in speech data are proposed which uses latent modelling techniques. These latent representations are then used for mismatch reduction of the ASR models.

In summary, this thesis studies various different techniques for minimising the mismatch between training and testing conditions in diverse datasets with the ultimate aim of improving performance of the ASR systems.

1.3 Contributions

The contributions of this thesis are listed below.

1. **data selection based on similarity to a target set**: developing a new data selection algorithm based on similarity to a target set for mismatch minimisation (chapter 3)

2. **data augmentation based on the identified levels of variations**: developing a new algorithm for learning the distributions of variations present in a target set and augmenting the training data with the learnt distribution for mismatch minimisation (chapter 3)

3. **genre identification using background tracking features**: identifying genres of media data using local background tracking features for improving the in-domain ASR systems (chapter 4)

4. **genre and broadcast-show identification using latent Dirichlet allocation**: identifying genres and broadcast-shows of media data using latent Dirichlet allocation-based features and investigating the required sources of information for reaching high levels of accuracy (chapter 4)

5. **adaptation of acoustic models to latent domains**: identifying latent domains in diverse datasets and adapting acoustic models to latent domains (chapter 5)

6. **latent domain aware training of DNNs**: organising broadcast media using latent modelling and adapting DNNs to the latent domains (chapter 5)

1.3.1 Data selection based on similarity to a target set

In this work the mismatch minimisation problem was studied as a data subset selection problem. The motivation of this study was to reduce the mismatch between training and test data by data selection techniques. Given a target test set and a pool of diverse training utterances, the task was to select a subset of training data such that the performance of the ASR system trained with this subset should be comparable to the model that is trained with all of the available data. The likelihood ratio was used to decide whether data resembels a target set. This approach was evaluated on a diverse dataset, covering speech from radio and TV broadcasts, telephone conversations, meetings, lectures and read speech. Experiments demonstrated that the proposed technique both finds the relevant data and limits the effects of negative transfer (negative transfer happens when the extra data affects the performance negatively). Results on a 6-hour test set showed relative WER improvements of up to 4% with the proposed data selection technique over using all of the available training data.

Relevant publication: Mortaza Doulaty, Oscar Saz, Thomas Hain, "Data-selective transfer learning for multi-domain speech recognition," in *Proceedings of Interspeech*, Dresden, Germany, 2015.

1.3.2 Data augmentation based on the identified levels of variations

The motivation of this work was to study how data augmentation techniques can be used for mismatch reduction. An alternative approach to address the mismatch problem is to augment the training data by perturbing the utterances in an existing uncorrupted and potentially mismatched training speech corpus to better match target test set utterances. An approach was proposed that, given a small set of utterances from a target test set, automatically identified an empirical distribution of perturbation levels that could be applied to utterances in an existing training set. Distributions were estimated for perturbation types that included acoustic background environments, reverberant room configurations, and speaker related variations such as frequency and temporal warping. The end goal was for the resulting perturbed training set to match the variabilities in the target domain and thereby optimise ASR performance. An experimental study was also performed to evaluate the impact of this approach on ASR performance when the target utterances were taken from a simulated far-field acoustic environment. Using the proposed approach, 10% relative improvement of the WER over the uniform perturbation baseline was achieved.

This work was performed during an internship of the author at Google Inc., New York. The original idea of this internship project was proposed by Richard Rose and Olivier Siohan and all of the follow-up research, implementation and experimental work was performed by Mortaza Doulaty Bashkand with collaboration of his co-authors.

Relevant publication: Mortaza Doulaty, Richard Rose, Olivier Siohan, "Automatic optimization of data perturbation distributions for multi-style training in speech recognition," in *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*, San Diego, California, USA, 2016.

1.3.3 Genre identification using background tracking features

Tagging diverse media data with labels such as genre has many applications in multimedia information retrieval systems. Since shows within the same genre share similar acoustic conditions, grouping media data based on such labels can be used

for data selection and model adaptation of ASR systems as well. This served as a motivation to study the genre identification task in more depth. In this work using a set of local features describing the most likely background environment for each frame, higher level concepts such as genres were identified. These local features were based on the output of an alignment that fits multiple asynchronous parallel background-based linear transformations to the input audio signal. These features can be used to keep track of changes in background conditions, such as presence of music, laughter, applause and etc. The proposed approach was tested on a set of 332 shows from the British Broadcasting Corporation (BBC). Using different classifiers such as HMMs and SVMs, an accuracy of 83% was achieved on this dataset.

Note that at the time of publishing this work, there were no external baselines available for comparison. Relevant baselines are provided in the corresponding section. Access to this data is available with a license agreement with the BBC.

The original asynchronous factorisation work was performed by Oscar Saz and the use of features derived from background indexes for the genre classification task was a joint work between Oscar Saz and Mortaza Doulaty Bashkand.

Relevant publication: Oscar Saz, Mortaza Doulaty, Thomas Hain, "Background-tracking acoustic features for genre identification of broadcast shows," in *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*, Lake Tahoe, Nevada, USA, 2014.

1.3.4 Genre and show identification using latent Dirichlet allocation

Since media data has a complex structure, acoustic latent Dirichlet allocation was proposed for modelling the media data. It was assumed that there was a set of latent factors that contributed to the generation of the media data and each show can be described as a mixture of those latent factors. Experiments were conducted on more than 1200 hours of TV broadcasts from the BBC, where the task was to categorise the broadcasts into 8 genres or 133 show identities. Furthermore, extra sources of information such as show transcripts and meta-data were studied for improving the classification performance. On a 200-hour test set, accuracies of 98.6% and 85.7% were achieved for genre and show identification respectively, using a combination of acoustic and textual features with meta-data.

Relevant publication: Mortaza Doulaty, Oscar Saz, Raymond W. M. Ng, Thomas Hain, "Automatic genre and show Identification of broadcast media," in *Proceedings of Interspeech*, San Francisco, California, USA, 2016.

1.3.5 Adaptation of acoustic models to latent domains

Posterior Dirichlet parameters from acoustic latent Dirichlet allocation (LDA) models have discriminatory information and were successfully used for genre and show identification tasks. The motivation of this study was to explore how this information can be used for acoustic model adaptation. In this work using a diverse dataset, a novel method to perform unsupervised discovery of latent domains using acoustic LDA was proposed. A set of hidden domains was assumed to exist in the data, whereby each audio segment can be considered to be a weighted mixture of the latent domain properties. The classification of audio segments into latent domains allowed the creation of latent domain specific acoustic models. Experiments were conducted on a dataset of diverse speech data covering speech from radio and TV broadcasts, telephone conversations, meetings, lectures and read speech, with a joint training set of 60 hours and a test set of 6 hours. Maximum A Posteriori (MAP) adaptation to latent domains was shown to yield relative WER improvements of up to 10%, compared to the models adapted with human-labelled prior domain knowledge.

Relevant publication: Mortaza Doulaty, Oscar Saz, Thomas Hain, "Unsupervised domain discovery using latent Dirichlet allocation for acoustic modelling in speech recognition," in *Proceedings of Interspeech*, Dresden, Germany, 2015.

1.3.6 Latent domain aware training of deep neural networks

It was shown that more latent domains were beneficial for the genre and show identification tasks. However, with the previous MAP adaptation approach the full potential of the acoustic LDA models could not be exploited, mostly because of data sparsity issues. This served as a motivation to study alternative approaches to incorporate acoustic LDA information for acoustic model adaptation. This work was focused on transcription of multi-genre broadcast media, which is often only categorised broadly in terms of high level genres such as sports, news, documentary, etc. However, in terms of acoustic modelling these categories are coarse. Instead, it is expected that a mixture of latent domains can better represent the complex and diverse behaviours within a TV show, and therefore lead to better and more robust performance. Using LDA modelling, these latent domains were identified and used to adapt DNNs using the one-hot vector representation of the LDA domains. Experiments were conducted on a set of BBC TV broadcasts, with more than 2,000 shows for training and 47 shows for testing. It was shown that latent domain aware training of the DNNs reduced the WER by up to 13% relative compared to the baseline hybrid DNN models.

This technique was also used in parts of the Sheffield multi-genre broadcast

(MGB) 15 system. The relevant LDA-DNN experiments were all conducted by Mortaza Doulaty Bashkand. Other models that were used for comparison and the overall Sheffield system were a joint work between the co-authors of the Sheffield MGB 15 system.

Relevant publication 1: Mortaza Doulaty, Oscar Saz, Raymond W. M. Ng, Thomas Hain, "Latent Dirichlet allocation based organisation of broadcast media archives for deep neural network adaptation," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, Arizona, USA, 2015.

Relevant publication 2 (for the Sheffield MGB 15 system)**:** Oscar Saz, Mortaza Doulaty, Salil Deena, Rosanna Milner, Raymond W. M. Ng, Madina Hasan, Yulan Liu, Thomas Hain, "The 2015 Sheffield system for transcription of multigenre broadcast media," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, Arizona, USA, 2015.

1.4 Organisation

The remainder of this thesis is organised as follows: in chapter 2 a unified view of the domain mismatch problem will be defined and AM adaptation techniques will be reviewed. Chapter 3 will study data selection and augmentation techniques in the context of domain mismatch reduction. Chapter 4 will introduce two new techniques for identifying genre and show entities in the diverse datasets using local expert features. Chapter 5 is devoted to the study of incorporating latent domain representations of the speech data in the framework of acoustic model adaptation for mismatch reduction. Finally, chapter 6 provides a summary of this thesis and outlines the possible directions for future work.

To demonstrate how the chapters are related to each other, figure 1.1 presents dependencies between them. Chapter 2 and chapter 4 can be read directly, but reading chapter 3 requires reading chapter 2 first. Furthermore, chapter 2 and chapter 4 are the prerequisites for reading chapter 5.

1.5 Published work

This section lists the peer-reviewed and published papers during the PhD studies. The first six publications are already introduced in section 1.3 and contain the main contributions of this thesis. The remainder of the publications contain auxiliary

work related to this thesis.

1. Mortaza Doulaty, Oscar Saz, Thomas Hain, "Data-selective transfer learning for multi-domain speech recognition," in *Proceedings of Interspeech*, Dresden, Germany, 2015.

2. Mortaza Doulaty, Oscar Saz, Thomas Hain, "Unsupervised domain discovery using latent Dirichlet allocation for acoustic modelling in speech recognition," in *Proceedings of Interspeech*, Dresden, Germany, 2015.

3. Mortaza Doulaty, Oscar Saz, Raymond W. M. Ng, Thomas Hain, "Latent Dirichlet allocation based organisation of broadcast media archives for deep neural network adaptation," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, Arizona, USA, 2015.

4. Mortaza Doulaty, Oscar Saz, Raymond W. M. Ng, Thomas Hain, "Automatic genre and show Identification of broadcast media," in *Proceedings of Interspeech*, San Francisco, California, USA, 2016.

5. Mortaza Doulaty, Richard Rose, Olivier Siohan, "Automatic optimization of data perturbation distributions for multi-style training in speech recognition," in *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*, San Diego, California, USA, 2016.

6. Oscar Saz, Mortaza Doulaty, Thomas Hain, "Background-tracking acoustic features for genre identification of broadcast shows," in *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*, Lake Tahoe, Nevada, USA, 2014.

7. Oscar Saz, Mortaza Doulaty, Salil Deena, Rosanna Milner, Raymond W. M. Ng, Madina Hasan, Yulan Liu, Thomas Hain, "The 2015 Sheffield system for transcription of multi-genre broadcast media," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, Arizona, USA, 2015.

8. Rosanna Milner, Oscar Saz, Salil Deena, Mortaza Doulaty, Raymond WM Ng, Thomas Hain, "The 2015 Sheffield system for longitudinal diarisation of broadcast media," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Scottsdale, Arizona, USA, 2015.

9. Raymond W. M. Ng, Mortaza Doulaty, Rama Doddipatla, Wilker Aziz, Kashif Shah, Oscar Saz, Madina Hasan, Ghada AlHarbi, Lucia Specia, Thomas Hain,

"The USFD spoken language translation system for IWSLT 2014," in *Proceedings of International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, Nevada, USA, 2014.

10. Salil Deena, Madina Hasan, Mortaza Doulaty, Oscar Saz, Thomas Hain, "Combining feature and model-based adaptation of RNNLMs for multi-genre broadcast speech recognition," in *Proceedings of Interspeech*, San Francisco, California, USA, 2016.

11. Thomas Hain, Jeremy Christian, Oscar Saz, Salil Deena, Madina Hasan, Raymond WM Ng, Rosanna Milner, Mortaza Doulaty, Yulan Liu, "webASR 2 - improved cloud based speech technology," in *Proceedings of Interspeech*, San Francisco, California, USA, 2016.

12. Raymond W. M. Ng, Mauro Nicolao, Oscar Saz, Madina Hasan, Bhusan Chettri, Mortaza Doulaty, Tan Lee, Thomas Hain, "The Sheffield language recognition system in NIST LRE 2015," in *Proceedings of the Speaker and Language Recognition Workshop Odyssey*, Bilbao, Spain, 2016.

13. Erfan Loweimi, Mortaza Doulaty, Jon Barker, Thomas Hain, "Long-term statistical feature extraction from speech signal and its application in emotion recognition", in *Proceedings of International Conference on Statistical Language and Speech Processing (SLSP)*, Budapest, Hungary, 2015.

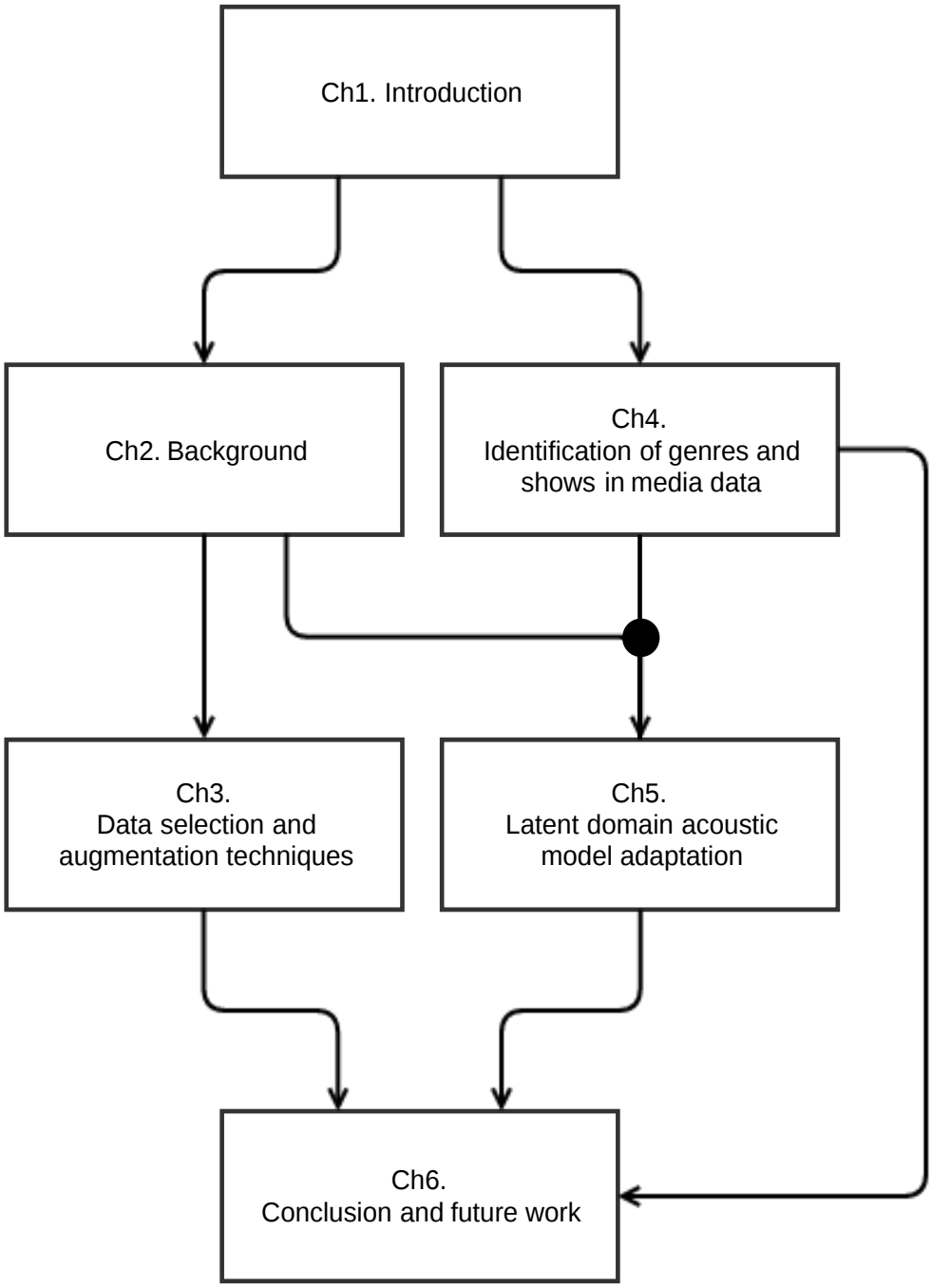


Figure 1.1: Dependencies of the chapters

CHAPTER 2

Background

2.1 Introduction

Often the term *domain* is used to vaguely define collections of speech data that share the same acoustic attributes and variabilities, such as type of speech (read vs. spontaneous), communication channel, background conditions and number of speakers. Conventional ASR domains often include broadcast news, meetings, telephony speech, audio books, lectures and talks (Benesty et al., 2007; Huang et al., 2001; Jurafsky and Martin, 2000). However, the concept of a domain is complex and not bound to specific criteria. In this section a new definition of a domain from a statistical point of view is provided based on the notations introduced in (Pan and Yang, 2010).

A domain is defined as a pair which consists of a feature space and a marginal probability distribution of data in that space:

$$\mathcal{D} = \{\mathcal{X}, P(X)\} \tag{2.1}$$

where $\mathcal{X}$ is a feature space, $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ is a dataset and $P(X)$ is the marginal probability distribution of the data in the feature space.With this notation two domains are different when either their feature spaces are different or they have different marginal probability distributions or both.

For the ASR task, $\mathcal{X}$ is the space of all arbitrary length segments of i.e. 39-dimensional MFCC feature vectors, X is a training dataset and $x_i \in X$ is a particular speech segment. The conventional domains in ASR such as meetings, read speech or talks can be considered to share the same feature space, but have different marginal probability distributions.

A task is defined as:

$$\mathcal{T} = \{\mathcal{Y}, f()\} \tag{2.2}$$

where $\mathcal{Y}$ is a label space and $f()$ is a prediction function which maps some input sequence to some output sequence:

$$f : \mathcal{X} \rightarrow \mathcal{Y}. \tag{2.3}$$

Two tasks are considered different when their label spaces are different or they have different prediction functions or both.

In supervised learning, the training data consists of (x_i, y_i) pairs such that $f(x_i) = y_i$ and $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, $X_{trn} = \{x_1, \ldots, x_n\}$ and $Y_{trn} = \{y_1, \ldots, y_n\}$. In a probabilistic learning framework, $f()$ can be viewed as $P(y|x)$, the posterior probability of the output, y, given the input, x. This function is usually not observed directly and learned from the training data.

In the speech recognition example, $\mathcal{Y}$ is the set of all possible sequences of words in English (defined as $\mathscr{L}$ in chapter 1) and $f()$ is a mapping function which maps an audio segment to a sequence of words. Using the same audio signal for speech recognition and emotion identification (where the task is to identify the emotion of the speaker) can be considered as two different tasks, since the label space as well as the prediction functions are different, but both tasks share the same input to their prediction functions.

In many machine learning problems, the source and target domains (underlying distributions of the training and test data) are assumed to be the same: $\mathcal{D}_{trn} = \mathcal{D}_{tst}$. Furthermore the tasks are identical as well: $\mathcal{T}_{trn} = \mathcal{T}_{tst}$. But in realistic scenarios the domains are usually different and this causes mismatch between the training and test domains. The next section is devoted to the domain mismatch problem.

2.2 Domain mismatch

One of the key assumptions in many statistical approaches for machine learning problems is that the training and test data are drawn from the same underlying distribution (Hermansky et al., 2015; Pan and Yang, 2010). However, in practice this assumption is not always true and the mismatch in training and test data degrades the performance. Actually in practical scenarios even if the training and test data are drawn from the same underlying distributions, after the deployment of the models and over time the new test data will be inevitably different from the original training and test data and this will cause a mismatch in the model (Yu and Deng, 2015).

Domain mismatch happens when $\mathcal{D}_{trn} \neq \mathcal{D}_{tst}$ where $\mathcal{D}_{trn}$, $\mathcal{D}_{tst}$ are the training and test domains respectively and it implies $\mathcal{X}_{trn} \neq \mathcal{X}_{tst}$ and/or $P_{trn} \neq P_{tst}$ (where P_{trn}, P_{tst} are the marginal probability distributions of the training and test sets). Domain adaptation aims at reducing the mismatch and is studied under different names in different fields. In econometrics it is called sample bias selection (Zadrozny, 2004), in statistical learning it is called covariate shift (Shimodaira, 2000), in machine learning it is called domain adaptation or transductive transfer learning (Arnold et al., 2007; Daume III and Marcu, 2006; Pan and Yang, 2010) and in the speech recognition community it is also called domain adaptation.

The performance of automatic speech recognition systems when applied to a particular domain depends on the degree to which the acoustic models provide an accurate representation of that domain. Training acoustic models from utterances that match the target speaker population, speaking style, acoustic environment, etc. (the factors that characterise the marginal probability distribution of the data in the feature space) is generally considered to be the easiest way to optimise the ASR performance. However, there are many scenarios where speech corpora of sufficient size that characterise the sources of variability existing in a particular target domain are not available. For example, it has been shown that ASR performance in many applications benefits from using many thousands of hours of speech utterances collected from a similar domain (Jaitly et al., 2012). Having enough matched high quality training data is rarely a practical option and training ASR systems with mismatched data results in poor performance. Adaptation techniques try to address these issues. Even if matched data exists, after deployment of the ASR systems new data will not be as good a match as it used to be before. This is mostly due to new environments, unseen speakers or even changes to voice of the current speakers over time (Yu and Deng, 2015). This further motivates the studies conducted in this thesis for mismatch compensation.

2.3 Relations to transfer learning

Adaptation techniques are a subset of a broader set of techniques in machine learning called transfer learning. Transfer learning aims to improve the performance of a machine learning algorithm using the knowledge acquired in a different domain or task (Pan and Yang, 2010). E.g. given a source and a target domain and their corresponding learning tasks: $\mathcal{D}_S$, $\mathcal{D}_T$, $\mathcal{T}_S$, $\mathcal{T}_T$, transfer learning aims to improve the performance of the objective predictive function in the target task $f_T()$ using $\mathcal{D}_S$, $\mathcal{T}_S$ where the source and target domains are different: $\mathcal{D}_S \neq \mathcal{D}_T$ or the source and target tasks are different: $\mathcal{T}_S \neq \mathcal{T}_T$.

The ideas behind transfer learning have very close resemblance to many natural and real-world problems. For humans, knowing how to drive a car is beneficial in learning how to drive a tractor. Humans are very good at transfer learning and most of the time they use this skill unconsciously and without any extra effort by leveraging some similar knowledge and skills they learned in the past (Pan and Yang, 2010).

2.3.1 Positive and negative transfer

The ultimate aim of transfer learning is to improve performance. When knowledge is transferred successfully and the performance is improved, it is called a positive transfer. However, in some cases it might happen that the transferred knowledge not only did not help to improve the performance, but also damaged it. When the transferred knowledge is harmful, the transfer is called a negative transfer. Measuring positive and negative transfer effects is usually trivial during the model training phase, where labelled data is available for the evaluation. However, after deployment of the model, it is not always easy to measure these effects on the new and unlabelled data.

In the literature the effects of negative transfer are not well studied (Pan and Yang, 2010). This served as a motivation to study the effects of negative transfer in the context of data selection for ASR in this thesis. The details of this study will be presented in chapter 3.

2.3.2 Transductive transfer learning

As discussed earlier, domain adaptation can be considered as a subset of transfer learning techniques. In the machine learning literature, a special term is used for this: transductive transfer learning (Arnold et al., 2007; Daume III and Marcu, 2006; Pan and Yang, 2010). A similar definition to transfer learning can be provided for transductive transfer learning: given a source and a target domain and their corresponding learning tasks: $\mathcal{D}_S$, $\mathcal{D}_T$, $\mathcal{T}_S$, $\mathcal{T}_T$, the aim is to improve the performance of the objective predictive function in the target task $f_T()$ using $\mathcal{D}_S$, $\mathcal{T}_S$ where the source domains are different but tasks are the same: $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S = \mathcal{T}_T$.

This is equivalent to domain adaptation in speech recognition. An overview of adaptation techniques for ASR is provided in the next section.

2.4 Adaptation for mismatch compensation

As discussed in the previous sections, the mismatch between training and test data causes a degradation in performance of ASR systems. To compensate for the mismatch between training and test conditions, adaptation and normalisation techniques can be used. Adaptation techniques are typically divided into model-based and feature-based techniques (Huang et al., 2001). The former, updates the model parameters to better fit the data and the later transforms the features to better fit the model.

The mismatch can be caused from various sources of variability, such as speaker or environment variabilities. In the ASR literature, mostly speaker variabilities are studied in the context of speaker adaptation where the aim is to compensate for the speaker variations. However, some speaker adaptation techniques can be used to compensate for other sources of variability, such as environment, device or the more generic notion of domain (Yu and Deng, 2015).

The conventional GMM-HMM based speaker dependent (SD) systems have a lower WER by a factor of two to three compared to speaker independent (SI) systems which are trained with similar amounts of data (Woodland, 2001). This demonstrates the impact of mismatch between the training and test conditions on the WER. On the other hand, it is not always easy to train SD systems, since they require a reasonable amount of transcribed data from the same speaker and the process of acquiring data and transcribing is time consuming and needs manual work in most cases (Cox, 1995; Woodland, 2001). So this makes speaker adaptive (SA) systems more interesting, as they fill the gap between the SD and SI systems. SI system are typically used to create a SA system.

The notion of dependency to speaker can be generalised to other intrinsic or extrinsic variabilities. For example if an ASR system for the South African English accent is to be trained, usually the best choice of training data would be from the same accent. However, if such training data is not available or only small amounts are available, then an accent independent system with the existing data can be trained and later it can be adapted to that specific accent.

Adaptation can be performed in different modes: it can be either supervised or unsupervised. It can also be in batch or incremental mode. In supervised adaptation, the correct transcription (word level) of the adaptation data is known. However, in case of unsupervised adaptation, the transcription for the adaptation data is a hypothesis which is generated by an ASR system. The problem with unsupervised adaptation is the quality of the estimated transcription which can make the system become even worse. Therefore, a confidence measure can be used to determine

the quality of the estimated transcription (Woodland, 2001; Yu, 2006; Zavaliagkos et al., 1998). With correct confidence measures, systems with even 80% WER can be improved (Zavaliagkos et al., 1998).

Adaptation can also be in either batch (static or block) or incremental (dynamic) mode. In batch adaptation the system is presented with the whole adaptation data before the final system is produced, however, in incremental mode the adaptation data is presented gradually and the system is adapted over the time (Yu, 2006). Depending on the application type, one can choose the most appropriate mode of adaptation (Kumar et al., 2013).

As introduced in chapter 2, the PDFs of the HMM-based acoustic models are often modelled by GMMs or DNNs. Various adaptation techniques are proposed for both techniques and a brief summary of them are provided in the next section.

2.5 Overview of acoustic model adaptation techniques

Adaptation techniques can be categorised into these three main schemes (Woodland, 2001; Yu and Deng, 2015):

- **transformation-based adaptation** where model parameters or features are transformed using (linear) transformations learned from the adaptation data
- **model re-training or conservative training** where some of the model parameters (or all of them) are re-estimated from the adaptation data
- **subspace adaptation** where the model parameters are updated based on the representation of the adaptation data in some subspaces

2.5.1 Transformation-based adaptation

2.5.1.1 GMM-based acoustic models

The mismatch between training and test conditions can be minimised using a transform to alter the model parameters. In transformation-based approaches for adaptation of GMM-based speech recognisers, model parameters including means and/or covariances of the Gaussian components can be transformed using a linear transformation to maximise the likelihood of the adaptation data, given the model. Especially when the amount of adaptation data is limited and a fast adaptation is desired, linear transformation-based approaches are used (Gales and Young, 2008). Two types of linear transformations are introduced next.

Maximum likelihood linear regression

The ML criterion is typically used for training initial AMs where the likelihood of the training data given the model parameters and the correct transcription is maximised. ML estimation is defined as:

$$\hat{\lambda}_{ML} = \arg\max_{\lambda} \ \mathrm{p}(\mathcal{O}|\mathcal{W}, \lambda) \tag{2.4}$$

where $\mathcal{O}$ is the training data, $\mathcal{W}$ is the correct transcription and λ is the parameter set.

The ML estimator can be used to learn a linear transform to maximise the likelihood of the adaptation data. Differences in speakers mostly affect the means of the feature vectors (Leggetter and Woodland, 1995) and thus, transforming the means would neutralise that effect. Means are transformed linearly using:

$$\hat{\mu} = \mathbf{W}\mu + \mathbf{b} \tag{2.5}$$

where $\mathbf{W}$ is a weight matrix and $\mathbf{b}$ is a bias vector.

Using the EM algorithm, a closed form solution to estimate $\mathbf{W}$ and $\mathbf{b}$ is derived (Leggetter and Woodland, 1995). One global transformation can be learned and applied to all of the Gaussian components, or different transformations for each subset of the Gaussian components can be learned and applied. Depending on the amount of available adaptation data a number of the transforms can be defined, e.g. for a few seconds of adaptation data, it is better to have a single transformation and as the amount of the adaptation data increases, more transformations can be used. This is achieved by grouping the Gaussian components together and learning a transform for each group. These groups are called *base classes*. Grouping Gaussian components can be performed in either a static or dynamic way (Gales and Young, 2008). A simple grouping can be performed based on the phonetic characteristics such as silence, vowels, stops, glides, nasals, fricatives, etc. (Young et al., 2006; Yu, 2006). One problem with the static methods is that since the number of groups are fixed and predefined, the amount of adaptation data does not change the number of these base classes and thus, all benefits of having more adaptation data is not exploited. However, dynamic methods can deal with adaptation data more efficiently and can have a variable number of groups depending on the availability of adaptation data. A simple dynamic grouping method can be based on the closeness of the Gaussian components in the acoustic space, e.g. centroid splitting algorithm with Euclidean distance (Young et al., 2006; Yu, 2006). This process is also called transformation sharing.

Covariance matrices can also be transformed. There are two alternative ways of transforming covariances: constrained MLLR and unconstrained MLLR (Shinoda, 2011). In the case of unconstrained MLLR, the variance is transformed using:

$$\hat{\Sigma} = LHL^T \tag{2.6}$$

where H is the Choleski factor of Σ, the original covariance matrix.

Mean adaptation is usually more effective than covariance adaptation and since the later is computationally expensive as well, mean adaptation is often preferred. Relative improvement of up to 15% with mean transformation is reported in the literature for telephony speech, meetings and broadcast news (Woodland, 2001), while adapting covariance yields only 2% WER reduction for most of those tasks (Shinoda, 2011).

Constraint maximum likelihood linear regression

In the constraint maximum likelihood linear regression (CMLLR), the covariance matrix is transformed with the same transformation which is used to transform the means:

$$\begin{aligned} \hat{\mu} &= \mathbf{W}\mu + b \\ \hat{\Sigma} &= \mathbf{W}^T\Sigma\mathbf{W}. \end{aligned} \tag{2.7}$$

This can be considered as applying the transforms at the feature level:

$$\hat{o}_t = \mathbf{W}^{-1}o_t + b. \tag{2.8}$$

When calculating the likelihood of the Gaussians, a factor $|\mathbf{W}|$ is required and expressed as:

$$\mathcal{N}(\mathcal{O}, \hat{\mu}, \hat{\Sigma}) = |\mathbf{W}|\,\mathcal{N}(\mathbf{W}^{-1}\mathcal{O} + \mathbf{W}^{-1}b; \mu, \Sigma). \tag{2.9}$$

Since the parameters of the model do not change in CMLLR, it becomes a good choice for situations where the speaker and acoustic environment change rapidly (Gales and Young, 2008).

CMLLR is also called feature-space maximum likelihood linear regression (fMLLR).

2.5.1.2 DNN-based acoustic models

Similar to the transformation-based adaptation of the GMM-based acoustic models, parameters of the DNN-based acoustic models can be updated using (linear) transformations.

In the following sections and also throughout the thesis, the following notation for DNNs will be used. $\mathbf{v}^0$ is the input to the network (equivalent to $\mathbf{o}$ in the previous notation), $\mathbf{z}^i$, $\mathbf{v}^i$ are the output of ith layer before and after the activation function (called excitation and activation vectors respectively), $\mathbf{W}^i$ and $\mathbf{b}^i$ are the weight matrix and bias vector of the layer i and $f()$ is the activation function. With this notation, excitation and activation of the ith layer are defined as:

$$\mathbf{z}^i = \mathbf{W}^i \mathbf{v}^{i-1} + \mathbf{b}^i, \tag{2.10}$$

$$\mathbf{v}^i = f(\mathbf{z}^i). \tag{2.11}$$

Transformation-based adaptation techniques are one of the most common adaptation methods where a linear transformation is applied (by the means of an extra layer) to either input features (Abrash et al., 1995), input to the softmax layer (Li and Sim, 2010) or activation of the hidden layers (Gemello et al., 2007).

Linear input network

When the transformation is applied to the input layer, it is called linear input network (LIN) or feature discriminative linear regression (fDLR) (Seide et al., 2011). It assumes that the SD features can be linearly transformed to an average speaker's features. For each speaker, a weight matrix and a bias vector is learned together with the other parameters using the back propagation (BP) algorithm. In other words, the speaker independent feature of $\mathbf{v}^0$ is transformed linearly into:

$$\mathbf{v}^1_{LIN} = \mathbf{W}_{LIN} \mathbf{v}^0 + \mathbf{b}_{LIN}. \tag{2.12}$$

The architecture of the network is shown in figure 2.1, where a speaker-dependent linear transformation layer is inserted between the input and first hidden layer.

Linear output network

The transformations can also be applied to the output layer, which is then called linear output network (LON) or output-feature discriminative linear regression (oDLR) (Seide et al., 2011; Yu and Deng, 2015). In the literature, this has been applied to either after or before application of the original weight matrix. In case it is applied before the softmax layer weights:

$$\mathbf{z}^L = \mathbf{W}^L \mathbf{v}^{L-1}_{LON} + \mathbf{b}^L, \tag{2.13}$$

$$\mathbf{v}^{L-1}_{LON} = \mathbf{W}_{LON} \mathbf{v}^{L-1} + \mathbf{b}_{LON} \tag{2.14}$$

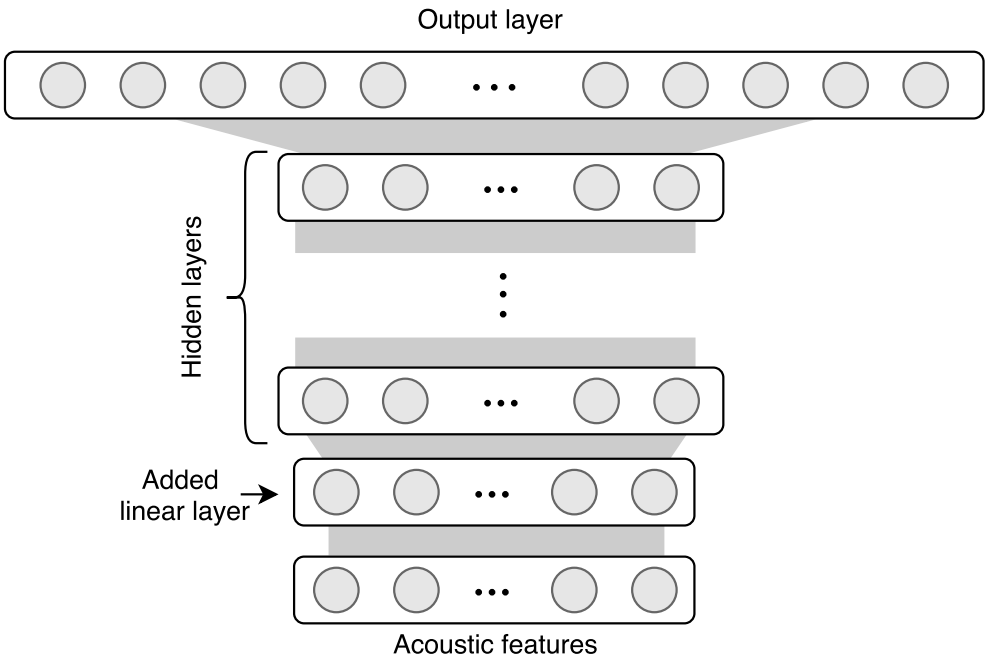


Figure 2.1: Linear input network architecture, adapted from Yu and Deng (2015)

where L is the last layer. And in case it is applied after the softmax layer weights:

$$\mathbf{z}_{LON}^{L} = \mathbf{W}_{LON}\mathbf{v}^{L} + \mathbf{b}_{LON}. \tag{2.15}$$

The architecture of both networks are depicted in figure 2.2 and 2.3. Depending on where the transformation is applied, the number of parameters to be learned can vary a lot, since the output layer is usually larger than the hidden layers (because of the number of tied CD-HMM states).

Linear hidden network

Finally the linear transformations can be applied to the hidden layers, which is called linear hidden network (LHN) (Yu and Deng, 2015). Similar to LON, the transformation can be applied before or after the weigh matrix of the hidden layer, but unlike the LON, number of parameters does not vary much in this case, since the size of hidden layers are usually the same or not vastly different.

There is no clear superiority of these three techniques and their variations to each other and the level of success usually depends on the size of the adaptation data and number of parameters and is very task dependent (Seide et al., 2011; Yu and Deng, 2015).

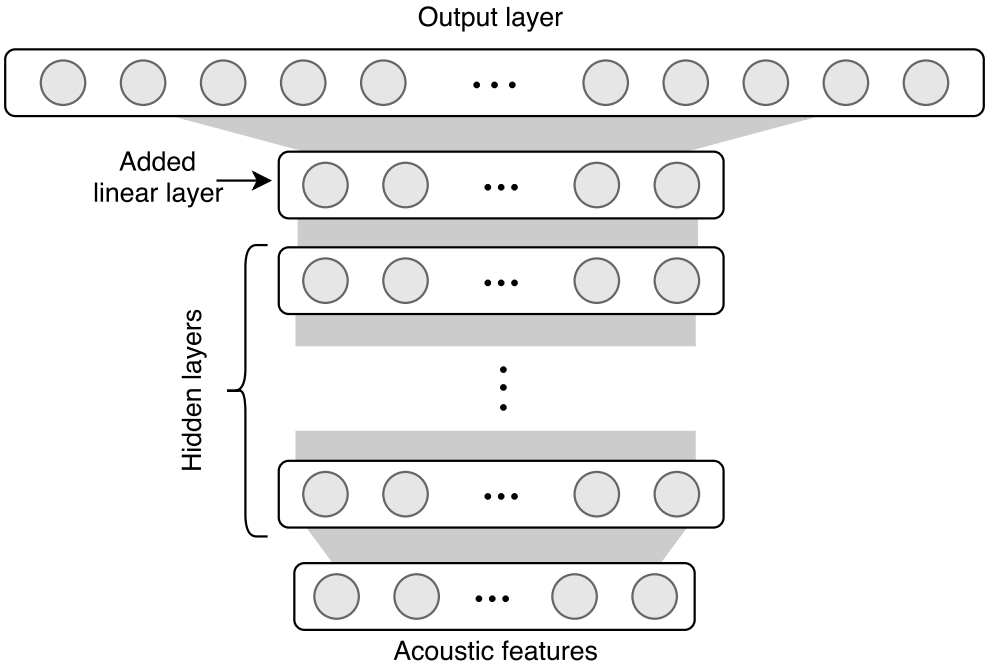


Figure 2.2: Linear output network architecture, before softmax weights, adapted from Yu and Deng (2015)

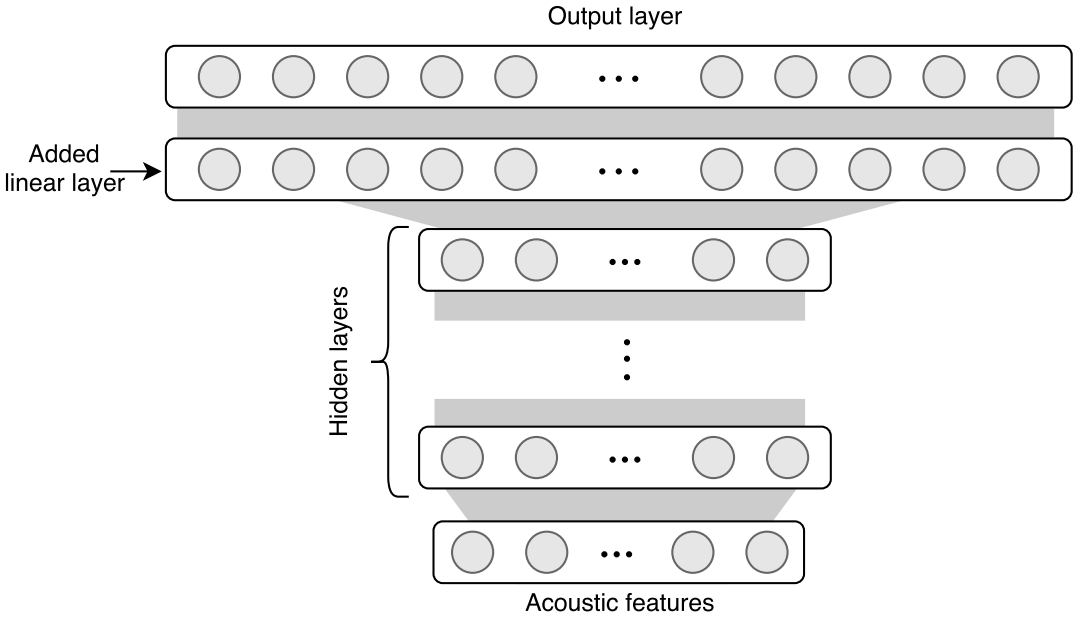


Figure 2.3: Linear output network architecture, after softmax weights, adapted from Yu and Deng (2015)

2.5.2 Model re-training or conservative training

An alternative approach for adaptation can be updating the model parameters with the adaptation data. In this section a brief overview of adaptation techniques that require model re-training will be provided.

2.5.2.1 GMM-based acoustic models

Maximum a posteriori adaptation

Model parameters can be re-estimated from the adaptation data, e.g. using the ML estimation. However, as the amount of adaptation data is limited there is a risk of over-fitting to the adaptation data. To overcome this problem *maximum a posteriori* adaptation can be used. In MAP, rather than maximising the likelihood, the posterior distribution of the HMM parameters is maximised (Yu, 2006):

$$\hat{\lambda}_{MAP} = \arg\max_{\lambda} P(\lambda|\mathcal{O}, \mathcal{W}) = \arg\max_{\lambda} P(\mathcal{O}|\lambda, \mathcal{W})P(\lambda) \tag{2.16}$$

where $\mathcal{O}$ is the training data, $\mathcal{W}$ is the correct transcription, λ is the HMM parameter set and $P(\lambda)$ is the prior distribution of the HMM parameter set. The use of this prior distribution means that when only a limited amount of adaptation data is available, the chances of over-training is less likely (Gauvain and Lee, 1994).

Using MAP criterion, model parameters are estimated using an iterative EM algorithm, similar to the ML training. A complete list of re-estimation equations is given at (Gauvain and Lee, 1994).

The advantages of MAP adaptation is that as the amount of adaptation data increases, the MAP estimate becomes similar to the ML estimate (converges in infinity). On the other hand, the limitation of MAP is that it will only update those Gaussian components that are observed in the adaptation data and others will not be updated. Also since in a large vocabulary speech recognition system there are many Gaussian components, updating all of them will require a considerable amount of adaptation data and a lot of training time. Thus, MAP may not be a good choice of adaptation with small amounts of adaptation data in a reasonable amount of time.

There are other MAP variants, such as the MMI-MAP (Povey et al., 2003), which uses a different prior compared to the ML prior in the conventional MAP adaptation. The MMI-MAP approach is based on a discriminative objective function called the maximum mutual information (MMI) as the prior. Results on several tasks have suggested that MMI-MAP outperforms the (ML)-MAP technique (Povey et al., 2003). However, computing the MMI objective function often requires lattices

which increases the computations required for the adaptation. Rather than the MMI objective function, other discriminative objective functions can be used, such as minimum phone error (MPE) (Povey and Woodland, 2002). Similar results to MMI-MAP are reported by using MPE-MAP (Povey et al., 2003).

2.5.2.2 DNN-based acoustic models

Re-training (Doddipatla et al., 2014; Stadermann and Rigoll, 2005) or conservative training (Yu and Deng, 2015) uses the already trained networks and updates some of the parameters of the network (usually not all of them). With this approach the architecture of the network is not usually changed and in this regard, it is unlike the linear transformation where the structure of the network was changed by adding the extra layers. The reason this technique is also called conservative training is that because of the small amounts of adaptation data, it is not desirable to update all of the model parameters (because of the over-fitting issues) and a conservative approach is more desired where only a subset of the model parameters are updated.

Using the adaptation data and the back-propagation algorithm, the parameters of some layers in the network are updated (often the last layers). For example Doddipatla et al. (2014) proposed a speaker dependent bottleneck layer where the parameters of the bottleneck layer were updated using the speaker specific adaptation data. On a meeting task, WER improvements of update to 4.2% were reported using their proposed conservative training approach.

2.5.3 Subspace adaptation

Models which are discussed so far considered all of the training data as a single block, without any information about the segments. Subspace adaptation techniques make different decisions based on the segments and they will be introduced in this section.

2.5.3.1 GMM-based acoustic models

In subspace adaptation (or speaker clustering or speaker space family), speakers are clustered into different groups and for each group a model is trained. The simplest form of this approach is the gender dependent (GD) systems in which two different models for male and female speakers are trained. Other similar systems also try to cluster speakers into more groups based on other similarities e.g. a distance measure between speakers or other sources of variability. Gender dependent systems usually show improvements (Shinoda, 2011), but there are certain restrictions with these clustering approaches if the number of clusters increases. The first issue is the hard decision in assigning a speaker to a cluster. This hard clustering process

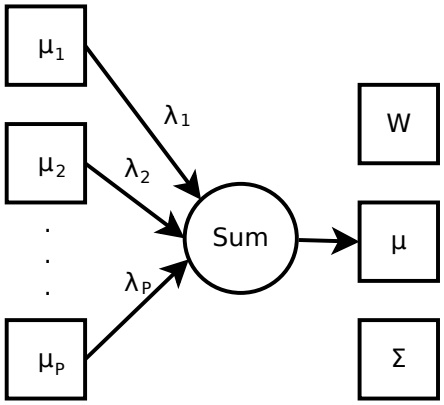


Figure 2.4: Cluster adaptive training (adapted from (Gales, 2000))

may not result in good clusters at all times. Furthermore, the other issue is the fragmentation of training data and less training data means less accurate models and worse recognition rates. Rather than grouping by speakers, this technique can also be extended to different environments, etc. To deal with the problem of hard assignments and data fragmentation, soft assignments can be used which are discussed next.

Cluster adaptive training

To overcome the problem of making a hard decision about the speaker cluster, an alternative approach can be representing a cluster as the weighted sum of means of all other clusters. In cluster adaptive training (CAT), each speaker is represented as a weighted sum of individual speaker cluster models. In CAT, variance and mixture weights are shared between clusters and the mean of each cluster is the linear interpolation of all cluster means (Gales, 2000). Assuming that there are P clusters with M Gaussian components each, for a particular speaker s, the mean of a particular Gaussian component m is defined as:

$$\hat{\mu}^{(sm)} = \mathbf{M}^{(m)} \lambda^{(sm)} \tag{2.17}$$

where $\mathbf{M}^{(m)}$ is the matrix of stacked mean vectors of all P clusters for a particular Gaussian component m:

$$\mathbf{M}^{(m)} = [\ \mu_1^{(m)} \ \ \ldots \ \ \mu_P^{(m)}\ ] \tag{2.18}$$

and $\lambda^{(sm)}$ is the weight vector of speaker s for the Gaussian component m:

$$\lambda^{(sm)} = [\ \lambda_1^{(sm)} \ \ \ldots \ \ \lambda_P^{(sm)}\ ]^T. \tag{2.19}$$

Experiments showed that having different cluster weights for different Gaussian components (e.g. grouped using similarities in acoustic space) yields better performance and of course increases the parameter count (Gales, 2000). So partitioning the Gaussians into R disjoint clusters, the formulas can be rewritten as:

$$\hat{\mu}^{(sm)} = \mathbf{M}^{(m)} \lambda^{(sr_m)} \tag{2.20}$$

$$\lambda^{(sr_m)} = [\ \lambda_1^{(sr_m)}\ \ \ldots\ \ \lambda_P^{(sr_m)}\ ]^T \tag{2.21}$$

where r_m is the cluster weight class of the Gaussian component m.

Once the new means are estimated, the canonical model $\mathcal{M}$ can be represented as:

$$\mathcal{M} = \left\{ \{\mathbf{M}^{(1)}, \ldots, \mathbf{M}^{(M)}\}, \{\Sigma^{(1)}, \ldots, \Sigma^{(M)}\} \right\} \tag{2.22}$$

and the speaker specific cluster weight vectors:

$$\mathbf{\Lambda} = \left\{ \{\lambda^{(11)}, \ldots, \lambda^{(1R)}\}, \ldots, \{\lambda^{(S1)}, \ldots, \lambda^{(SR)}\} \right\} \tag{2.23}$$

where $\Sigma^{(m)}$ is the covariance matrix of the Gaussian component m.

Training is performed using an iterative EM algorithm. Cluster weight vector of speaker s of cluster weight class r is estimated using an ML estimator. A complete list of re-estimation formulas are given in (Gales, 2000).

After computing the cluster weight vectors, the cluster parameters are calculated. In CAT, cluster means are represented in two ways:

- model-based clusters, in which for each cluster, means are explicitly represented.
- transform-based clusters, in which the means of each cluster is a linear transformation of canonical means.

Using CAT in an SI task with very little adaptation data usually reduces the WER. Moreover, when it is used with other adaptation schemes, a 5% relative reduction in the WER compared to a speaker independent system is expected (Gales, 2000).

Eigenvoices

Similar to CAT, the eigenvoice technique (Kuhn et al., 1998) creates canonical speaker models. For adaptation, a weighted sum of those canonical HMMs are created. From a set of T speaker dependent models, T vectors of dimension D

are derived using principal component analysis (PCA) or other similar techniques. These T vectors are called eigenvoices (a similar analogy to eigenfaces used in face recognition (Woodland, 2001)). For a new speaker, the model is constrained to be in a K dimensional space (spanned by the first K eigenvectors, where K is very small compared to the original dimension D) and adaptation estimates the K eigenvoice coefficients for that speaker. Similar to MLLR and CAT, in this approach only means are adapted.

During adaptation, for estimating the eigenvoice coefficients, maximum likelihood eigen decomposition is used (Kuhn et al., 1998). This algorithm is identical to the CAT weight estimation algorithm in the adaptation step.

Usually gains from eigenvoice and CAT are not comparable in performance to techniques such as CMLLR. Furthermore, unlike other adaptation methods, as the amount of adaptation data increases, the performance of the system does not improve accordingly. As a result, these techniques are often used in combination with other adaptation techniques, such as MLLR and MAP (Woodland, 2001).

Subspace Gaussian mixture models

Unlike CAT and eigenvoices where speaker variations are modelled, in Subspace GMMs (SGMM) phone variabilities are modelled in a subspace (Povey et al., 2010, 2011) which is similar to factor analysis in speaker recognition (Povey, 2009). In SGMMs each context-dependent phonetic state is modelled by a GMM whose parameters are associated with a vector-valued quantity, and there is a global shared mapping from these state-vectors to the mixture weights and means. The SGMM model can be described with these equations:

$$\mathrm{p}(\mathbf{o}|j) = \sum_{i=1}^{I} w_{ji} \mathcal{N}(\mathbf{o}, \mu_{ji}, \Sigma_i) \tag{2.24}$$

$$\mu_{ji} = \mathbf{M}_i \mathbf{s}_j \tag{2.25}$$

$$w_{ji} = \frac{\exp(\mathbf{w}_i^T \mathbf{s}_j)}{\sum_{i'=1}^{I} \exp(\mathbf{w}_{i'}^T \mathbf{s}_j)} \tag{2.26}$$

where $\mathbf{o} \in \mathcal{O}$ is a D dimensional feature vector, j is the context dependent speech state, $\mathbf{s}_j$ is the S dimensional state vector ($S \simeq D$), also considered as the subspace dimension. Each state is modelled with an I mixture GMM with all parameters shared between states. $\mathbf{M}_i$ and $\mathbf{s}_j$ are the globally shared parameters. These parameters are learned iteratively using the EM algorithm, similar to the training of conventional GMM-HMM systems.

Other variants of this model are also introduced where there are other sub-

spaces, such as the speaker subspace (Povey et al., 2010). The same notion can be generalised to environment, etc. These models usually yield lower word error rates compared to conventional GMM models and can also be further improved with other adaptation techniques, such as CMLLR. SGMM models were first introduced in 2010 (Povey et al., 2010) and before gaining popularity in the community, quickly became outdated with the rise of deep learning techniques which were outperforming them in comparable setups.

2.5.3.2 DNN-based acoustic models

For subspace methods, a speaker or environment subspace is estimated and then neurons' weights or transformations are computed, based on the subspace representation of the speaker or environment. The PCA-based adaptation approach (Dupont and Cheboub, 2000), identity vector (iVector) based speaker-aware training (Saon et al., 2013) can be considered as subspace methods. Figure 2.5 represents the network architecture where the input features are augmented with the subspace information. Adding subspace information is equivalent to:

$$\begin{aligned} \mathbf{v}^1_{Subspace} &= f\Big(\begin{bmatrix} \mathbf{W}^1_{\mathbf{v}} \mathbf{W}^1_{\mathbf{d}} \end{bmatrix} \begin{bmatrix} \mathbf{v}^0 \\ \mathbf{d} \end{bmatrix} + \mathbf{b}^1_{Subspace} \Big) \\ &= f\Big(\mathbf{W}^1_v \mathbf{v}^0 + \underbrace{\mathbf{W}^1_d \mathbf{d} + \mathbf{b}^1_{Subspace}}_{\text{subspace specific bias}} \Big) \end{aligned} \tag{2.27}$$

where $\mathbf{d}$ is a vector which represents the input in the subspace, such as speaker variability space represented by e.g. iVectors. The notation $\begin{bmatrix} \mathbf{W}^1_{\mathbf{v}} \mathbf{W}^1_{\mathbf{d}} \end{bmatrix}$ represents stacking two matrices (and the same for stacking vectors). This can be considered as a form of bias-adaptation, where a new bias vector is learned for each subspace variation.

This form of adaptation is often considered to be simpler than other approaches, as it does not have an extra adaptation stage. The only modification is augmenting the inputs of the network and the rest of the process is implicit in the training.

2.6 Normalisation for mismatch compensation

As the name suggests, the aim of normalisation techniques is to normalise the features/models, e.g. to a canonical speaker or remove other effects such as those due to the transmission channel, and thus reduce the mismatch. These techniques are mostly applicable to both GMM-based and DNN-based acoustic models.

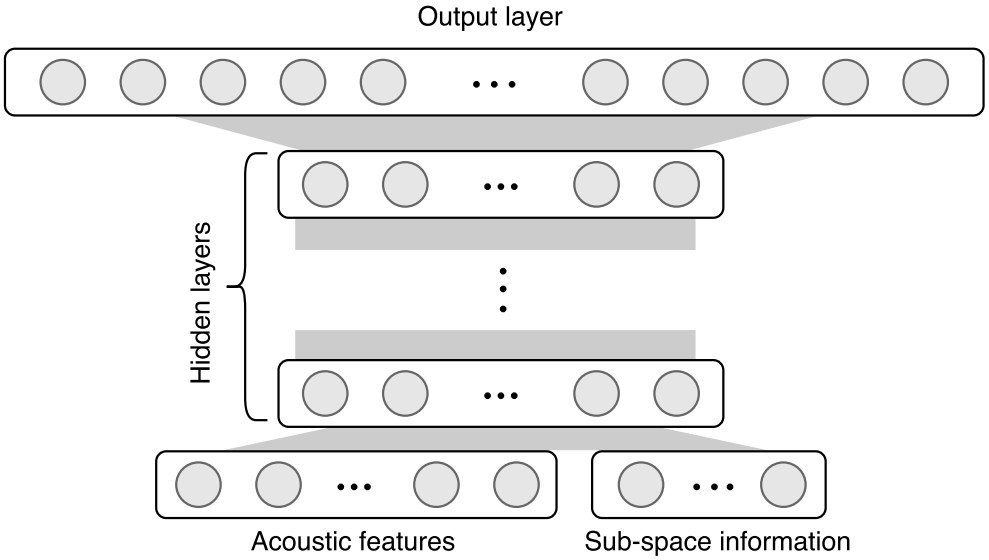


Figure 2.5: Subspace DNN architecture, adapted from Yu and Deng (2015)

2.6.1 Cepstral mean and variance normalisation

Different characteristics of the channel degrade the performance of speech recognition systems. Cepstral mean normalisation (CMN) is a very simple method which tries to minimise the effects of differences in the channels. It requires calculating the cepstral mean across the utterance and subtracting it from each frame. Given a set of cepstral vectors $\mathbf{o}_t$, the mean can be computed as:

$$\mu_{\mathbf{o}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{o}_t. \tag{2.28}$$

Normalising by the mean, gives the new output vector $\hat{\mathbf{o}}_t$:

$$\hat{\mathbf{o}}_t = \mathbf{o}_t - \mu_{\mathbf{o}}. \tag{2.29}$$

This method is not usable for live audio streams. There is an alternative version of CMN, called dynamic CMN which uses a linear combination of an initially estimated mean and as more data comes in, it re-estimates and uses the new mean.

Higher moments can be normalised as well. Cepstral variance normalisation (CVN) is similar to CMN, and they are often used together as cepstral mean and variance normalisation (CMVN):

$$\begin{aligned} \sigma_{\mathbf{o}}^2 &= \frac{1}{T}\sum_{t=1}^{T}(\mathbf{o}_t^2 - \mu_{\mathbf{o}}^2)\ , \\ \hat{\mathbf{o}}_t &= \frac{\mathbf{o}_t - \mu_{\mathbf{o}}}{\sigma_{\mathbf{o}}}\ . \end{aligned} \tag{2.30}$$

2.6.2 Cepstral histogram normalisation

Other normalisation techniques include cepstral histogram normalisation (CHN) which can be considered as an extension to CMVN and is equivalent to normalising each moment of data to match a target distribution (Gales and Young, 2008). In this context, Gaussianisation can be considered as a special case where the desired distribution is a Gaussian distribution and features are altered to match a Gaussian distribution. This is performed by finding a transformation on the input features $\mathcal{O}$ that yields a normal distribution with zero mean and unit variance:

$$\hat{\mathbf{o}} = f(\mathbf{o}) \tag{2.31}$$

$$\hat{\mathbf{o}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{2.32}$$

Performing Gaussianisation on the full feature set is considered to be a complex task (Gales and Young, 2008), but simpler approaches exist where each element of the feature vector is assumed to be independent. It requires generation of a series of random numbers from the normal distribution for each utterance so that the number of random numbers matches the length of the utterance and then sorting these random numbers and assigning them to the first dimension of the feature vector and continuing the same process for other dimensions.

2.6.3 Vocal tract length normalisation

The underlying idea of vocal tract length normalisation (VTLN) is a law of physics: resonances in an acoustic tube (such as the vocal tract) occur at frequencies that are proportional to the inverse of the tube's length. Different speakers have different vocal tract properties and these differences can be normalised by linearly scaling the filter bank centre frequencies in the ASR's front-end feature extraction to approximate a canonical formant frequency scaling. VTLN is effective for telephony speech where speakers are clearly identified, however, for other tasks such as broadcast news, it is not very effective, since the speaker changes must be inferred from the data and usually there is not enough speech for each speaker to have a robust estimate of the parameters (Gales and Young, 2008).

2.6.4 Speaker adaptive training

Rather than features, models can also be normalised. In SI systems the aim is to include as many speakers as possible and try to learn the model parameters from those speakers, hoping that differences between spoken words are captured well. However, differences from all those speakers are also learnt which was not the

original aim. One way to overcome this problem is to normalise the model using speaker adaptive training (SAT).

SAT is a model normalisation technique and tries to have a canonical model which encodes all of the differences between spoken words of different speakers and not the speaker differences. To achieve this, first a set of transforms for each speaker is learned (like MLLR mean transforms), then those transforms are applied to the seed model and a new canonical model is estimated (Anastasakos et al., 1996). The notion of adaptive training can be generalised to any source of variability, such as noise, which is called noise adaptive training (NAT) (Kalinli et al., 2010; Saz et al., 2015) or device, which is called device adaptive training (DAT) (Yu and Deng, 2015).

2.7 Multi-style training for mismatch compensation

Multi-style training (Lippmann et al., 1987) (MTR) aims to improve the accuracy of speech recognition systems by training the model with the data which has different variations (rather than just clean training data), such as noisy environment of different speaking styles. It helps the model to generalise well to those conditions which might not necessarily be present in the clean training data. It involves perturbing the utterances in an existing uncorrupted and potentially mismatched speech corpus to better match a given target domain. MTR techniques were first introduced by Lippmann et al. (1987) for the GMM-HMM based systems. For an isolated word recognition task, clean training data was re-recorded by adding different speaking variations (such as fast, loud, question-pitch) and it helped to improve the accuracy by a factor of two. They became more popular again with the rise of neural networks. Since DNNs are discriminative models, they are more sensitive to mismatches in training and testing conditions, and one way of overcoming this problem is to actually train with mismatched conditions.

2.7.1 Data augmentation

Data augmentation is an extension of the MTR notion and refers to the practice of generating multiple versions of each utterance in a corpus where each version corresponds to a different type or different degree of perturbation (Cui et al., 2014; Karafiát et al., 2015; Ko et al., 2015; Ragni et al., 2014; Ravanelli and Omologo, 2014). This technique is used for robust ASR as well as under-resourced scenarios (where the amount of training data is very limited and usually not enough for training a model with reasonable performance).

Augmenting data can be performed using various perturbation sources. Jaitly and Hinton (2013) used a warping factor for the frequency axis to create multiple versions of the same utterance. This approach is called vocal tract length perturbation and was shown to improve the accuracy. Random warping factors were used in and improvements of up to 1% was achieved on a small task. For large vocabulary tasks, Cui et al. (2014) used a similar approach to augment the training data by a factor of 4 and reported similar improvements.

Rather than frequency, the time axis can be warped to create multiple versions of the same utterance. It is equivalent to tempo modification and is shown to improve the performance (Ko et al., 2015). Note that unlike the pitch modification, with tempo modification since the length of utterance changes, the augmented copies can not be directly used in the DNN training (where frame level alignments are required) and a further alignment step is required. Ko et al. (2015) proposed to use three copies of the data with speed factors of 0.9, 1.0 and 1.1 and reported improvements of up to 4.3% across various large vocabulary speech recognition tasks.

Rather than speaker variability, environment variability can be considered for the data augmentation task. E.g. variations in room impulse responses can be used to create multiple copies of utterances to simulate different rooms (Ravanelli and Omologo, 2014) and this was shown to improve the robustness of the models. Adding various background noises to the clean data and training with the noisified data was also shown to improve the robustness of the speech recognition systems in noisy conditions (Jaitly et al., 2012).

Despite all the successes that have been reported for multi-style training, there are some practical issues in these studies. First, the choice of the type of data perturbation and the parameterisation of the perturbation method is often ad hoc. Second, it is often the case that the impact of a given source of perturbation is significantly different from one domain to another. Finally, determining the impact of a given data augmentation approach requires perturbing the training data, training an acoustic model from the augmented training set, and evaluating the WER using a test set from the target domain. These practical issues mean that MTR is not applicable in all scenarios. In chapter 3 these issues will be studied and discussed in depth.

2.8 Summary

In this chapter a definition for the domain was provided and used to formulate the problem of mismatch in training and test conditions of machine learning problems. Techniques used for compensating the mismatch are studied under various names

in different fields. In the speech recognition community they are called adaptation techniques and mostly speaker adaptation is studied. However, most of the speaker adaptation techniques can be generalised to the other sources of variation, such as background, device and the more generic notion of domain.

The majority of this chapter was an overview of speaker adaptation techniques for both GMM-HMM and DNN-HMM acoustic models and where possible, the generalisation to other sources of variability such as the domain was discussed. Techniques developed for the GMM-HMM models usually cannot be directly used for the adaptation of DNN-HMM models. However, a unified categorisation of techniques for both models was provided in this chapter, which was: transformation-based approaches, retraining and sub-space methods. The relevant techniques for both acoustic model types were studied in this chapter with references for more details. Usually the selection of one method over the other is task dependent, e.g. if the amount of adaptation data is very limited, then not all approaches are applicable. The use cases of these approaches were also provided in the corresponding sections.

The remaining part of this chapter was devoted to normalisation techniques where either features are transformed to better fit the model or the models are transformed to better match the features. Finally another family of mismatch compensation techniques called multi-style training was introduced and the relevant studies were briefly introduced.

All of the different approaches discussed in this chapter have the ultimate goal of mismatch reduction and boosting performance. The remainder of this thesis will be focused on further improving some of the existing techniques by addressing their shortcomings, and also introducing some novel techniques. The next chapter will be about mismatch compensation using data selection techniques, followed by a new approach for modelling the latent domains in speech and its applications in named domain identification and acoustic model adaptation.

CHAPTER 3

Data selection and augmentation techniques

3.1 Introduction

For many machine learning problems and in almost all practical problems the underlying distributions of the training and test data are different, and this causes a mismatch in the training and test conditions which usually degrades the performance (Pan and Yang, 2010). The same problem exists in speech recognition as well, the differences in the underlying distributions from which the training and test data are sampled causes a mismatch in the training and testing conditions and this increases the WER (Yu and Deng, 2015).

Training acoustic models from utterances that match the target speaker population, speaking style, or acoustic environment is generally considered to be the easiest way to optimise ASR performance. However, there are many scenarios where speech corpora of sufficient size, that characterise the sources of variability existing in a particular target domain, are not available. In practical situations even if the training data of sufficient size that matches the target domain is available and used for training the ASR models, after the deployment of the ASR system and over time, the new test data will be different from the initial test data and this will again cause mismatch between the training and test data (Yu and Deng, 2015). This motivates the study conducted in this chapter to explore various techniques that can be used for minimising the mismatch between training and test data.

There are several approaches to address the mismatch problem between training and test data, including adaptation techniques that were introduced in chapter 2 and data selection and augmentation techniques that will be introduced in this chapter. The aim of the data selection/augmentation/generation techniques is to

create perfectly matched training data to a target test set. The matched training corpus is created by either selecting data from an existing pool of data, augmenting some existing data, or generating new data.

To assess the quality of the selected/augmented/generated training data, usually distance measures are defined and used as a proxy value for the WER, such that reducing the distance between the training and test data usually decreases the WER on the target test data. The reason for using proxy values rather than the actual WER is mostly for practical considerations. Computing WER on each subset is not considered to be a practical option because of the time required for training and evaluating the ASR models. Thus, the proxy function should be fast and easy to compute.

If the amount of training data is fixed and known beforehand and the task is to select a subset of that data, then the mismatch minimisation problem turns into a data selection problem. In the data selection problem, given a target test set the aim is to select a subset of the training data that, when a model is trained with the selected training data, will have the lowest WER compared to using any other subset of the available training data.

If the amount of training data is not fixed and the training data can be augmented, e.g. by generating artificial data or perturbing the existing data, and the task is to generate a training set, then the mismatch minimisation problem turns into a data augmentation problem. The aim of the data augmentation techniques is to create a training set that better matches a target test set. The data augmentation problem is also used in low resource scenarios, where the amount of training data is usually not enough to train models with reasonable performance. One approach to solve this problem is to augment the existing training data (Ragni et al., 2014). Data augmentation is also used in MTR scenarios, where the aim is to have diverse conditions (background noise, speaking style, speaker characteristics, etc.) present in the data so that the model generalises better to different conditions in the test set (Lippmann et al., 1987).

The main research question of this chapter is how to create a training set (by either selecting or augmenting data) that best matches a target test set. To address this question, first a unified view of the mismatch minimisation problem is provided based on the notation introduced in chapter 2, and then an overview of data selection and augmentation techniques are provided in section 3.2. Two similarity measures for data selection and data augmentation are provided in section 3.3 and 3.4 respectively, followed by the conclusion of the chapter in section 3.5.

3.2 Data selection for mismatch compensation

The mismatch caused by the variations in acoustic and channel conditions and speaker characteristics in the training and test data degrades the performance of speech recognition systems. This has been shown in several studies (Cox, 1995; Deng et al., 2000; Gales and Young, 2008; Gong, 1995; Hamidi Ghalehjegh, 2016; Seltzer et al., 2013; Yu and Deng, 2015) and will be confirmed in the experimental work described in section 3.3.2. There are several techniques for mismatch compensation, including adaptation techniques that were introduced in chapter 2 and data selection techniques (Kapralova et al., 2014; Lin and Bilmes, 2009; Siohan, 2014; Siohan and Bacchiani, 2013; Wu et al., 2007) that are introduced in this chapter.

In speech recognition and other supervised learning tasks, parameters of the model are usually estimated from a training set, X_{trn}, and its performance is evaluated on an independent test set, X_{tst}. As defined in chapter 2, if X_{trn} is sampled from the distribution P_{trn} and X_{tst} is sampled from P_{tst}, theoretically the mismatch can happen when: $P_{trn} \neq P_{tst}$. A distance measure can be defined over these two marginal distributions as: $\Phi\big(P_{trn}, P_{tst}\big)$ and the aim of data selection techniques is to reduce this distance and thus reduce the mismatch. The assumption is that when the mismatch is reduced, the performance should improve (Pan and Yang, 2010). The data selection problem can be formulated as finding the distribution which is closest to the distribution of the test set:

$$\hat{P} = \arg\min_{P} \Phi(P, P_{tst}) \tag{3.1}$$

and sample the training data from the new distribution $\hat{P}$.

3.2.1 Overview of data selection techniques for ASR

Data selection in the context of automatic speech recognition usually refers to these similar problems: data subset selection for training or adaptation of acoustic models, batch active learning and semi/lightly supervised acoustic model training (Lin and Bilmes, 2009; Nagroski et al., 2003; Wei et al., 2013; Wu et al., 2007).

Data subset selection refers to the problem of selecting a subset of data from a pool of available training data, such that if a model is trained with the selected data it will achieve comparable performance to the model trained with all of the available training data. This problem is also called minimal representative data selection (Lin and Bilmes, 2009; Nagroski et al., 2003). The motivations for the minimal representative data selection problems include reduced training time and fast deployment time (Wu et al., 2007).

Batch active learning (Riccardi and Hakkani-Tür, 2003; Settles, 2010; Tur et al., 2003) addresses the problem of selecting a subset of data from a pool of unlabelled data for labelling subject to some constraints (e.g. a budget in terms of amount of data) (Lin and Bilmes, 2009). This is very similar to the previous problem and the only difference is the initial purpose of the data selection: in the data subset selection the purpose is to train/adapt acoustic models directly with the selected data while in batch active learning, the purpose is to first label the selected data and then train/adapt the acoustic models.

Data selection is also used in semi/lightly supervised training. When the quality of the transcripts are not very reliable, data selections techniques are used to filter the poor quality segments (Lanchantin et al., 2013; Siohan, 2014; Wessel and Ney, 2005). Again, this problem is very similar to the other two problems, however in the ASR literature these problems are studied in different contexts (Itoh et al., 2012; Lin and Bilmes, 2009; Nagroski et al., 2003; Wei et al., 2014a, 2013, 2014b; Wu et al., 2007).

3.2.1.1 Ranking and selecting data

Most of the data selection problems consist of ranking and selecting, where all of the available training data are ranked according to some scores and then the top N samples are selected. Ranking can be performed based on some similarity metrics. The similarity metrics can be purely acoustic or a combination of acoustic and phonetic/linguistic features. These similarity metrics can then be computed pairwise between all of the available data samples or computed individually for each of the samples (as a similarity score to a target set).

For the ranking functions two criteria are usually evaluated: informativeness and representativeness (Itoh et al., 2012). Informativeness measures how beneficial a data point is when added to the training set and representativeness means the frequency of finding a sample in the data samples. Two popular informativeness measures are usually used in the literature: *uncertainty sampling* and *query by committee* (Lin and Bilmes, 2009; Seung et al., 1992).

In the uncertainty sampling techniques, first a model is bootstrapped with the initial labelled data and then is used to assign scores to the unlabelled data. The system then queries the samples which it is uncertain or very certain about (depending on the task).

Typically, confidence based or entropy based scores are used in the uncertainty sampling methods (Lin and Bilmes, 2009). In case of confidence scores, they are used to select data with the most reliable transcriptions, as in semi-supervised training (Kapralova et al., 2014; Wessel and Ney, 2005), or to select data for manual

transcription as in active learning scenarios (Riccardi and Hakkani-Tür, 2003; Tur et al., 2003). Scores derived from the entropy-based approaches can be used as well. The entropy-based methods aim to pick data that, for instance, fits a uniform distribution of some target units (phones, words, etc.), resulting in maximum entropy (Lin and Bilmes, 2009; Wu et al., 2007; Zhang and Rudnicky, 2006) or pick data that have a similar distribution to a target set (Gouvea and Davel, 2011; Siohan, 2014; Siohan and Bacchiani, 2013).

In the query by committee techniques, a set of distinct models are trained and then used for selecting or rejecting data points based on voting and majority agreement. Scores similar to the uncertainty sampling scores can be used here as well.

3.2.1.2 Related work

Nagroski et al. (2003) proposed an uncertainty based sampling technique which uses a combination of various features to compute the representativeness scores of the data points. The features included an aggregated distance between the centroids of the clustered phones in the training and test data, length of each utterance, etc. With combining these features, they assigned a score to each utterance and then selected the top utterances to satisfy a budget (based on amount of data). They reported significant improvements over random selection for a connected digit recognition task. However, the effectiveness of this simple approach on larger datasets was not verified.

Wu et al. (2007) used an entropy based score for the minimal representative data selection problem. Their main objective was to reduce the training time by selecting a subset of available training data which yields a comparable performance to a system which is trained with a much larger dataset. They used the maximum entropy principle, which states that when a distribution is uniform, its entropy is maximised. Data selection was performed in a way that guaranteed the distribution of some base units (such as phones, words or characters) were as close as possible to the uniform distribution. Using a greedy selection technique they added speech segments that increased the entropy by some threshold. It was shown that their maximum entropy-based selection outperformed random selection, and they could select a subset of 150 hours from a pool of 800 hours without a drastic change in the WER of the trained models.

Rank and select algorithms are known to be affected by outliers (Lin and Bilmes, 2009; Siohan, 2014; Siohan and Bacchiani, 2013), as they tend to query the outliers often or the selection leads to having a training set which no longer has the properties of the target set. To avoid this problem, Itoh et al. (2012) used an entropy-based score for data selection that combines informativeness and representativeness. They

used the entropy of the N-best list word hypothesis and combined it with cosine distances of the tf-idf representation of the phone n-grams for the utterances. With this approach they reported a 11% relative WER reduction over the confidence based selection techniques in a 400-hour voice mail corpus.

Rather than trying to have a uniform distribution of some target units, such as phones or characters as proposed by Wu et al. (2007), the distribution of these units can be matched to the distribution of a target set (or a representative sample drawn from a target set). In these techniques usually the aim is to select training data for a target set subject to some budget criterion. First the distributions of these units are estimated from a target set and then data selection is performed trying to match that distribution. Siohan and Bacchiani (2013) used iVector representation of the utterances as their representative scores. iVectors (Dehak et al., 2011) are a low dimensional vector representation of the utterances in the acoustic space. They tried to match the distribution of iVectors in the training set to the distribution of the iVectors in the target test set. As a natural choice to compare two distributions, they used Kullback-Leibler divergence (Gouvea and Davel, 2011). Starting from an initial set, they added new utterances to the training set only if adding them did not increase the divergence value. To avoid quick saturation, they repeated this process with many independent initial sets until the budget criteria was satisfied. Their proposed data selection technique was tested on a voice search task with various budgets, ranging from 25 hours to 125 hours and in all of the cases they could beat the random selection baseline. Unlike other similarity approaches listed in this section, iVector based scores are pure acoustic scores and does not take into account the phonetic or linguistic contents of the utterances.

Similar to the iVector distribution matching (Siohan and Bacchiani, 2013), the distribution of other units can be matched against a target set as well. Siohan (2014) matched the distribution of the CD state symbols in the training set to their distribution in a target set. First the state level alignments of the transcripts are acquired for both the target set and the training data, and then they are used in the data selection criterion. Using a similar procedure to the iVector distribution matching, the distribution of the context dependent state symbols are also matched using KL-divergence to form a training set. Experiments were conducted on a voice search dataset and significant improvements were reported compared to a baseline of random selection of high confidence utterances.

3.2.1.3 Diminishing returns and sub-modular functions

Word error rate reduction curves of large scale ASR systems often show diminishing returns with increasing amounts of training data (Lin and Bilmes, 2009; Moore,

2003). This could be due to redundancy, noisiness or irrelevancy of the additional training data (Wei et al., 2014a). Sub-modular functions are a familti of set function that have the property of diminishing returns (Wei et al., 2014b) and since the performance of the ASR systems when trained with larger datasets exhibit the same effect of diminishing returns, sub-modular functions have been studied for the data selection problem in the ASR literature (Lin and Bilmes, 2009; Wei et al., 2014a, 2013, 2014b). Informally, the value of the submodular functions has the property that the incremental value difference of adding new elements decreases as the size of the input set increases. This is analogous to the diminishing returns in the ASR systems, as the amount of training data increases, the improvement in the WER decreases. Submodular functions have many applications in approximation algorithms, game theory, economics and electrical networks. In machine learning field, these functions have been used in automatic summarisation, multi-document summarisation, feature selection, active learning, sensor placement, image collection summarisation, etc.

A sub-modular function is defined as any set function $f : 2^{\Omega} \rightarrow \mathbb{R}$ that fulfils:

$$f(S) + f(T) \geq f(S \cup T) + f(S \cap T), \forall S, T \subseteq \Omega \tag{3.2}$$

where S and T are two sets. The data selection problem can be formulated as a sub-modular maximisation problem, where the objective is to find a subset S from the complete training set Ω so that any new subset T added to S will not increase the value of the sub-modular function f:

$$\underset{S \subseteq \Omega}{\operatorname{argmax}} \{f(S) | f(S \cup T) < f(S),\ T \subseteq \Omega \setminus S\}. \tag{3.3}$$

Since the problem of sub-modular maximisation is NP-hard (Krause and Golovin, 2014; Wei et al., 2014a), greedy solutions are proposed where the subset S is increased iteratively by the item $s \in \Omega$ that maximises the value of f when added to S:

$$s = \underset{s \in \Omega \setminus S}{\operatorname{argmax}} \{f(S \cup \{s\})\}. \tag{3.4}$$

The set S is obtained when either the optimal answer is found ($f(S) > f(S \cup \{s\})$), or the budget constraint is satisfied: $|S| \leq N$ (budget is defined as the maximum size of the set S).

The function f is normalised if $f(\emptyset) = 0$ and is monotone if $f(S) \leq f(T)$ whenever $S \subseteq T$. If the function f is a normalised monotone sub-modular function, then the greedy algorithm provides a good approximation of the optimal solution.

The sub-optimal solution is no worse than a constant value $(1-1/e)$ from the optimal solution (Krause and Golovin, 2014; Nemhauser et al., 1978).

For the data selection problem, one can define a normalised and monotone sub-modular similarity function and use the greedy algorithm to select data samples and the solution will have theoretical guarantees of being optimal. In the data selection literature several functions are proposed which will be introduced briefly in this section.

The uncapacitated facility location function is defined as (Lin and Bilmes, 2009):

$$f_{fac}(S) = \sum_{i \in V} \max_{j \in S} w_{i,j} \tag{3.5}$$

where w_{ij} is the similarity of utterance i to utterance j. This function measures the similarity of subset S to V (the whole set).

To measure the similarity of two utterances with variable lengths, Fisher kernels (Jaakkola et al., 1999) were proposed by Lin and Bilmes (2009). For any generative model, such as GMMs, the Fisher score can be computed by taking the derivative of the log-likelihood function with respect to any of the model's parameters. Fisher scores have a fixed dimensionality which is the dimensionality of the parameter set. Fisher score for a variable length sequence $\mathbf{o}$ is defined as:

$$U = \frac{\partial}{\partial \theta} \log p(\mathbf{o}|\Theta) \tag{3.6}$$

where Θ is the parameter set of the model and U is the Fisher score which has the same length as the number of parameters in Θ. With Fisher scores, the variable-length observations are mapped to a fixed-length representation and the Fisher kernel can be used to compute the pairwise similarity. Several kernel functions were proposed, such as cosine similarity, radial basis function kernel similarity and ℓ_1 norm similarity.

Lin and Bilmes (2009) proposed to use facility location function with ℓ_1 norm similarity of Fisher scores to select data using the greedy algorithm. They experimented with various budgets for selecting subsets of the training data for model training. When only using 10% of the training data they reported relative phone error rate reduction of around 4% compared to random selection. Their proposed technique also outperformed the confidence based selection by 2% relative.

Wei et al. (2013) proposed to use the graph cut sub-modular function for the data selection problem. The graph cut sub-modular function is defined as:

$$f_{gc}(S) = \sum_{i \in V \setminus S} \sum_{j \in S} w_{ij} \tag{3.7}$$

which measures the similarity of S to the remainder of V (that is not included in S). They also used Fisher kernel similarity with ℓ_1 norms and reported around half percent relative WER reduction compared to the facility location objective function. Other sub-modular functions were also proposed in the literature such as a feature-based sub-modular function (Wei et al., 2014a) or diversity reward function (Wei et al., 2014b) which slightly outperform the other two functions in various different tasks.

The Fisher score based similarity is a pure acoustic similarity measure. Other similarity measures such as string kernels were proposed by Wei et al. (2013) which moves beyond pure acoustic similarity. They used a phone tokeniser to derive a tf-idf representation of the segments and then compute the similarity scores. It should be noted that this measure requires extra models and computations for tokenising the phones, and depending on the availability of a reliable tokeniser, this technique might not be applicable in all data selection problems.

In summary, to use the sub-modular data selection framework, the requirements are to select a sub-modular objective function which is normalised and monotonic and then the greedy algorithm can be used for optimising the objective function. In the next section a similarity measure based on likelihood ratio will be introduced for the data selection problem.

3.3 Likelihood ratio based distance

In this section a new data selection technique based on likelihood ratio similarity to a target test set is proposed and its effectiveness is studied with experimental work on a highly diverse dataset.

As discussed in this chapter, different similarity metrics such as iVectors, Fisher scores, etc. can be used to measure the similarity of the training set utterances to a target test set. Based on the similarity scores, data selection can then be performed by trying to maximise those scores (as defined in equation 3.1). Here a sub-modular function based on the accumulated values of likelihood ratios will be used to select the training data.

Using a generative model such as a GMM with K mixture components, the likelihood of each frame can be computed as:

$$p(\mathbf{o}_t|\Theta) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{o}_t|\mu_k, \Sigma_k), \tag{3.8}$$

$$\mathcal{N}(\mathbf{o}_t|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{o}_t - \mu_k)^{\mathsf{T}}\Sigma_k^{-1}(\mathbf{o}_t - \mu_k)\Big) \tag{3.9}$$

where d is the dimensionality of the frames ($\mathbf{o}_t$), Θ is the parameter set defined as $\Theta = \{\pi_k, \mu_k, \Sigma_k; k = 1, \ldots, K\}$ and π_k, μ_k, Σ_k are the weight, mean and co-variance of the kth component.

$p(\mathcal{O})$ can be used as a measure of similarity for the data selection problem. However, the likelihood value varies hugly depending on the length of $\mathcal{O}$ and its phonetic content. The use of likelihood ratio can alleviate this problem by normalising the likelihood values. The likelihood ratio (LR) of a frame can be computed by dividing the likelihood scores of two models and is defined as:

$$\mathrm{LR}(\mathbf{o}_t) = \frac{p(\mathbf{o}_t|\Theta_1)}{p(\mathbf{o}_t|\Theta_2)} \tag{3.10}$$

where Θ_1 and Θ_2 are two GMM models. The total LR of an utterance is defined as the mean of the frame-based LR values, assuming frame independence:

$$\mathrm{LR}(\mathbf{o}) = \frac{1}{T}\sum_{t=1}^{T}\frac{p(\mathbf{o}_t|\Theta_1)}{p(\mathbf{o}_t|\Theta_2)}. \tag{3.11}$$

The likelihood ratio can be used to decide whether the data bears resemblance to a target set. If this value is high, then there is a high chance that $\mathbf{O}$ is a good match to the target data.

For this purpose Θ_1 is trained with the target set's data and Θ_2 is trained with the pool of training data.

One can define a modular function (Krause and Golovin, 2014) based on the accumulated LRs of all utterances included in a subset of the pooled training data in the following form:

$$f_{\mathrm{LR}}(\mathcal{O}) = \sum_{\mathbf{o}\in\mathcal{O}}\big(\mathrm{LR}(\mathbf{o})\big) \tag{3.12}$$

This function is then maximised by picking the utterances with the highest score to create the training set.

Modular functions are a special case of sub modular functions (Krause and Golovin, 2014) where the greater than or equal sign in equation 3.2 changes to the equal sign. This way, the proposed function f_{LR} is sub modular as well. And since all of the values for LR are non-negative, and therefore any sum of these numbers, as constituted by the function f, the function is necessarily monotonic with expanding sets ($A \subseteq B \subseteq \Omega, f(A) \leq f(B)$). This function is also normalised ($f(\emptyset) = 0$) and as discussed in section 3.2.1.3, the greedy solution can be used to select the training data with a budget in terms of maximum number of hours to be selected. In section 3.3.4, this proposed data selection technique will be evaluated using a diverse dataset. The next section defines the dataset and baseline results.

This measure is a purely acoustic measure and in practical applications where there is a chance of having hugely imbalanced datasets or other extreme cases such as very short utterances, this approach should be used cautiously. Ideally, this approach should be used in combination with other techniques that take into account the phonetic content of the selected utterances. Current experiments presented in section 3.3.4 shows the applicibility of this technique in diverse and multi-domain datasets.

3.3.1 Data selection and transfer learning experiments with a diverse dataset

An experimental study is conducted in this chapter to first study the effects of using mismatched training and test data in the performance of ASR systems, and then to study the effectiveness of the proposed approach in reducing the mismatch and improving the performance. For the experimental work of this chapter, a very diverse dataset was required so that the mismatched conditions can be easily experimented on. For this purpose an artificially diverse data set was created by combining six different datasets. Details of the dataset are provided in section 3.3.1.1. The mismatch in components of this dataset makes it a good choice for the following experimental work, where the effects of using mismatched training data is studied in section 3.3.2. Section 3.3.3 further investigates the positive and negative transfer effects when using cross-domain data and finally in section 3.3.4 a new approach for data selection based on similarity to a target test set is presented using the likelihood ratio function defined in this section.

3.3.1.1 Dataset definition

For the data selection experiments, a highly diverse simulated dataset was created by combining 6 different types of data widely used in ASR experiments:

- Radio (RD): BBC Radio4 broadcasts on February 2009 (Bell et al., 2015b)
- Television (TV): broadcasts from BBC on May 2008 (Bell et al., 2015b)
- Telephone speech (CT): from the Fisher corpus[1] (Cieri et al., 2004)
- Meetings (MT): from AMI (Carletta et al., 2006) and ICSI (Janin et al., 2003) corpora
- Lectures (TK): from TedTalks (Ng et al., 2014)

[1] All of the telephone speech data was up–sampled to 16 kHz to match the sampling rate of the rest of the data.

Table 3.1: Amount of data used from each component dataset for the training set of the diverse dataset and their related statistics (durations are in hh:mm:ss format)

Dataset	Duration	#Segments	#Words	#Unique Words	#Speakers
RD	10:00:05	3,685	116,015	9,827	518
TV	10:00:07	6,774	118,190	10,928	1,745
CT	10:00:01	10,200	114,188	6,029	100
MT	10:00:34	4,088	104,368	5,484	80
TK	10:00:00	5,143	108,927	10,088	100
RS	10:00:04	3,963	84,299	8,902	89
Total	**60:00:52**	**35,279**	**645,987**	**25,374**	**2,632**

- Read speech (RS): from the WSJCAM0 corpus (Robinson et al., 1995)

A subset of 10h from each component dataset was selected to form the training set (60h in total), and 1h from each component dataset was used for the test set (6h in total). The selection of these component datasets aimed to cover the most common and distinctive types of audio recordings used in ASR tasks. Table 3.1 and 3.2 summarises the statistics of the datasets. Each of the component datasets have their own particular attributes; some of them are listed in the statistics table. For example, the MT dataset has only 80 speakers for the 10 hour training set, while TV with similar amount of data has more than 1,700 speakers. Also in terms of the number of unique words, CT and MT have around 6,000 unique words, however, TV and TK have more than 10,000 unique words for the same amount of data (in terms of duration). This shows the diversity of words used in TV and TK compared to CT and MT. Comparing the total number of words, RS has the lowest count which shows that the average speaking rate is lower than the others. In terms of type of speech, all of the datasets can be considered to be spontaneous speech, except the RS which is read speech. However, parts of RD and TV have read speech as well (e.g. news programmes). These differences plus other variabilities, such as speaking style, background conditions, etc. characterise each of these components and shows the diversity of this dataset.

3.3.1.2 Baseline models

Since the dataset consists of various different component datasets and to evaluate the difficulty of each component, baseline models were trained. One set of baseline models were trained for each component separately, and also another baseline model was trained using all of the available pooled data. These models were then evaluated

Table 3.2: Amount of data used from each component dataset for the test set of the diverse dataset and their related statistics (durations are in hh:mm:ss format)

Dataset	Duration	#Segments	#Words	#Unique Words	#Speakers
RD	1:00:00	282	10,872	2,596	68
TV	1:00:01	802	11,379	2,871	90
CT	1:00:01	721	12,727	1,696	71
MT	1:00:02	397	10,026	1,618	53
TK	1:00:04	359	10,321	2,399	19
RS	1:00:01	410	8,743	2,378	20
Total	**6:00:12**	**2,971**	**64,068**	**7,869**	**321**

on the test set. Details for the baseline models are provided in this section.

Two types of systems were used for the experiments: a GMM-HMM system and a bottleneck DNN-GMM-HMM system. For the GMM-HMM system, 13 dimensional PLP (Hermansky, 1990) features plus their first and second derivatives were used (in total 39 dimensional). For the DNN-GMM-HMM system, a 65 dimensional feature vector concatenating the 39 dimensional PLP features and 26 dimensional bottleneck (BN) features were used. The BN features were extracted from a 4 hidden layer feed-forward DNN trained with the 60 hours of the training data. For the DNN, 31 adjacent frames (15 frames to the left and 15 frames to the right) of 23 dimensional Mel-scale log-filter bank energy features were concatenated to form a 713 dimensional super vector; a discrete cosine transform was applied to this super vector to de-correlate and compress it to 368 dimensions and then it was fed into the neural network. The network had 4 hidden layers of size 1,745 followed by a bottleneck layer of size 26 and a softmax output layer of 4,000 context dependent triphone states. The objective function used for training was frame-level cross-entropy (CE) and the optimisation was performed using the stochastic gradient descent (SGD) algorithm. For both types of features, MLE-based GMM-HMM models were trained with 5-state crossword triphones and 16 Gaussian components per state. For the bottleneck system, the frame level alignments were acquired from the initial GMM-HMM system. The language model was based on a 50,000 word vocabulary and was trained by combination of component language models for each of the 6 domains. The interpolation weights were tuned using an independent development set.

Table 3.3: WER (%) of the baseline models on the test set of the diverse dataset, ordered in terms of difficulty

Features	Model	RS	RD	TK	CT	MT	TV	Overall
PLP	ML	17.3	18.4	34.1	46.6	44.0	51.1	**36.0**
	ML in-domain	16.9	19.1	35.1	44.4	44.0	52.9	**36.3**
	MAP	14.6	16.8	31.8	43.5	40.4	49.6	**33.6**
PLP+BN	ML	13.0	13.3	23.5	33.5	32.2	42.0	**26.8**
	ML in-domain	12.6	14.0	25.0	34.3	33.2	44.0	**27.9**
	MAP	12.1	12.8	23.1	32.5	30.6	41.5	**26.2**

3.3.1.3 Baseline results

Table 3.3 presents results using both types of acoustic features with three different types of models: ML, ML in-domain and MAP. ML models were trained with the ML criterion using all of the pooled training data. ML in-domain were the 6 individual models trained with the in-domain 10h data and each model was then used to decode the corresponding test set. Finally, the initial ML model is MAP adapted to each of the 6 domains and the new adapted models were used to decode the corresponding test set.

These results show a large variation in the performance among domains, from 17% and 18% for the read speech and radio broadcasts to 51% for the television broadcasts. The use of PLP+BN features provides a 20–25% relative improvement in performance against the PLP features in all three types of the models; which is consistent across domains and follows the results previously seen in the literature (Hinton et al., 2012; Yu and Deng, 2015). The results using in-domain data models is overall worse than the pooled data models (e.g. 26.8% vs. 27.9% with PLP+BN features) which suggests that more data is helpful for this task. In both types of features the MAP adapted models yielded the best performance which sets MAP as a preferred setup for domain adaptation in the context of GMM-HMM models. Among other adaptation techniques, MLLR adaptation did not consistently improve the performance compared to the MAP adaptation and was not considered for the domain adaptation task with this amount of data.

3.3.2 Effects of using mismatched training data

In this section the effects of using mismatched training data is studied. The motivation for the study in this section is to investigate how using mismatched data for training affects the performance of speech recognition systems. Furthermore, in this

Table 3.4: WER (%) on the test set of the diverse dataset using the domain-specific models

Domain	RD	TV	CT	MT	TK	RS	Overall
RD	**19.1**	55.1	72.1	57.2	50.7	24.9	47.8
TV	26.5	**52.9**	77.3	63.8	52.1	35.2	52.5
CT	82.3	90.1	**44.4**	71.9	67.9	86.6	72.6
MT	44.9	72.3	69.2	**44.0**	51.1	41.1	54.7
TK	39.8	62.8	69.3	56.1	**35.1**	55.4	53.6
RS	29.9	66.2	84.1	67.2	68.9	**16.9**	57.4

section the effects of using similar data (to the test set) from other domains on the performance of ASR systems is studied. For this purpose, with the 10h training data of each domain, an ASR system was trained and used to evaluate the WER of the test sets for each of the components. Rows of table 3.4 show the training domains and columns are the test domains. The lowest WER for each of the domains are from the in-domain models (the diagonal line) and in case of using similar data (e.g. TV and RS for RD), WERs are lower compared to using a completely different data set (e.g. MT for RD). These results suggest that using similar data can be beneficial, but still the performance is not comparable to the case of using in-domain data. This further suggests to evaluate the performance of models trained with in-domain data plus some cross-domain data.

3.3.3 Effects of adding cross-domain data: positive and negative transfers

In this section the effect of adding cross-domain data is studied. The motivation of this experiment is to study how the WER changes when cross-domain data is added to the training set. This involves training 30 models in total. E.g. for radio data, the training data from other 5 domains are combined with the in-domain data one-by-one to train five different models and the performance is evaluated on the RD test set. Figure 3.1 presents the relative WER change over the baseline in-domain results. Test domains are listed vertically and cross-domain training data are listed horizontally. For example the first row shows that adding TV data to the radio data reduces the WER on the target radio data by 7% relative and adding telephony data (CT) increases the WER by 8% relative compared to training with only radio data. These effects are called positive transfer and negative transfer. Positive transfer happens when the newly added data helps improve the performance and negative

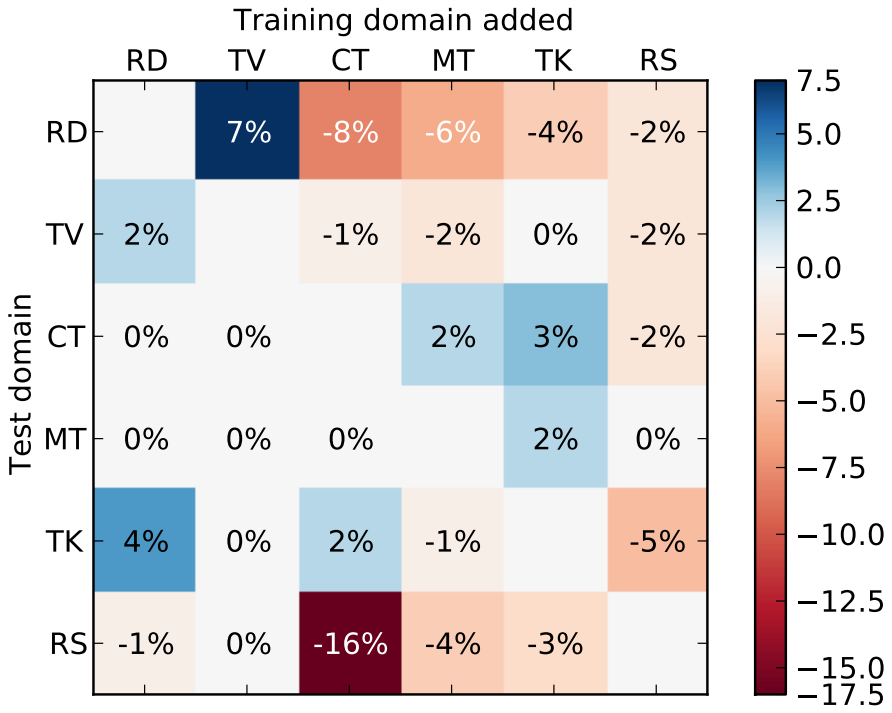


Figure 3.1: Heatmap of relative WER change by adding cross-domain data to in-domain models

transfer happens when the new addition degrades the performance.

From this plot degrees of similarity between data can also be inferred, e.g. TV data and radio data are similar and in cross-domain scenarios they help each other. The same is true for telephony speech and lectures. However this relation is not necessarily symmetric, e.g. adding lectures data helps for the target meeting data, but adding meeting does not help lectures data. It is also worth noting some extreme negative effects, e.g. adding telephony data degrades the performance of read speech data considerably (16%).

These results showed that positive and negative transfer occurred across domains, possibly due to similarities and differences in speech styles, acoustic channels and background conditions. However a rule-based optimisation of the best model for each target domain would require a complex and error-prone process. The next experiments aimed to evaluate how an automatic selection of training could benefit from the positive transfer, while restricting the negative transfer.

3.3.4 Data selection based on likelihood ratio similarity to a target set

As introduced in section 3.3, likelihood ratio can be used as a measure of similarity to a target test set. To evaluate the proposed approach in the multi-domain diverse

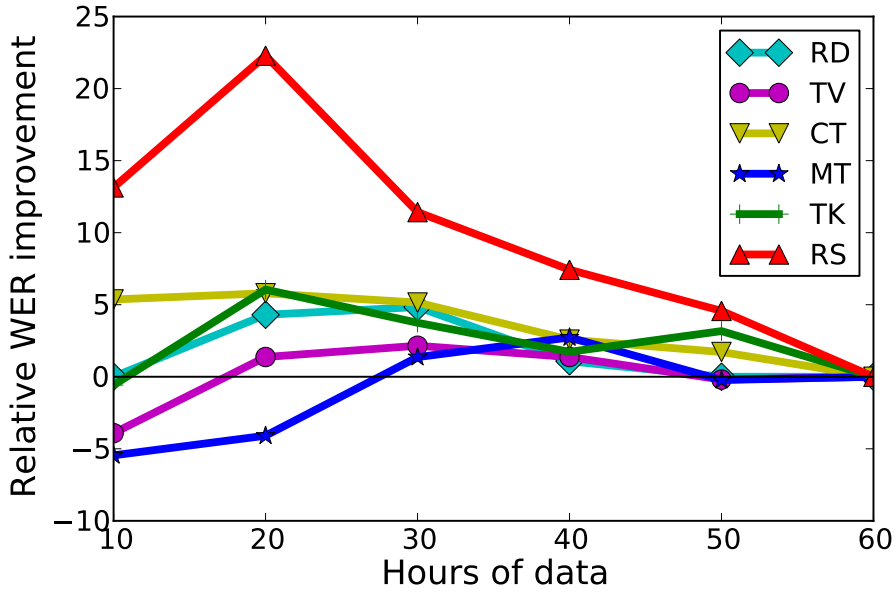


Figure 3.2: Relative WER (%) improvement with budget–based data selection

dataset, two GMMs with 512 mixture components were trained with ML criterion and mix-up procedure, one for the target test data and one for the training data. They were then used to compute the LR for each utterance in the training set in order to select the training data according to the acoustic similarity. These experiments are conducted first using a fixed budget, in terms of the maximum number of hours, which is presented in section 3.3.5. An alternative approach to derive the budget automatically is proposed next in section 3.3.6 followed by a summary of the data selection experiments.

3.3.5 Data selection based on budget

The first evaluation was performed using data selection based on a budget. Five possible budgets of 10, 20, 30, 40 and 50 hours were chosen for each test domain and the respective training data was chosen using the $f_{LR}(S)$ sub-modular function, as in equation 3.12 with the budget constraints from the pooled training data. Figure 3.2 shows relative improvement for each domain and budget against the results with the 60-hour model. The graphs show that all domains improve performance as the budget increases until a certain limit is reached, then negative transfer decreases the performance, converging to the WER achieved with the 60-hour trained model.

In order to observe which types of data were selected for each domain with different budgets, figure 3.3 presents the percentage of training data selected for each test domain with a 10-hour budget. While the majority of the data was chosen from the same domain, some cross-domain data was also selected, indicating positive transfer between domains. This occurred, for instance, with TV and read speech

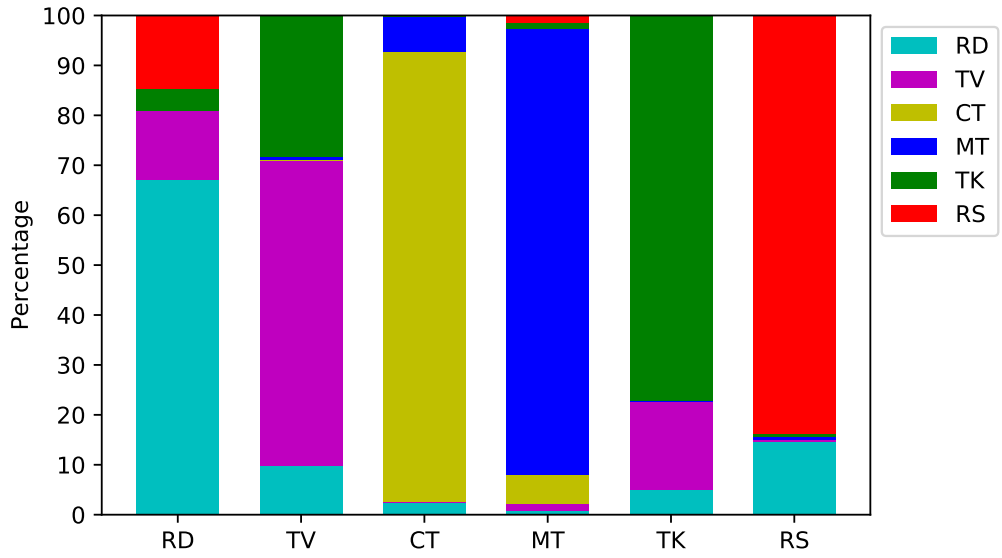


Figure 3.3: Types of data selected for a 10-hour budget using likelihood ratio similarity measure from the diverse dataset

data towards radio data; and lectures data towards TV data. This shows that the similarity measure prefers in-domain data plus some other similar data.

The distribution of the phones in the selected data for each domain was analysed to see if it is biased to any group of the phones or not (e.g. as reported by Siohan (2014)) and it was observed that it was not biased towards any specific phone set or context dependent phone set.

A natural extension to this work is to infer the budget automatically. In the next section a method is proposed to find the right budget for each of the domains.

3.3.6 Automatic decision on budget

An issue that can arise with the evaluated budget-based proposal is the fact that a decision on a budget has to be made, and as the results in figure 3.2 suggest, the optimal budget varies across different domains. A method for deciding a budget for a given target domain was proposed by selecting only utterances whose likelihood-ratio is above a threshold defined as the mean of the highest-weighted component of a GMM fitted to the distribution of likelihood ratios. The use of the component with the highest weight avoids the influence of outliers in the distribution of the LR values and since it is selected by a data driven method, no manual threshold setting is required. Note that the automatic threshold finding process is repeated for each of the target domains so that each target domain has its own threshold value selected by the processed defined above.

The experiments with an automatic budget decision were performed for both

Table 3.5: WER (%) of the baseline models with the diverse dataset

Features	Method	RS	RD	TK	CT	MT	TV	Overall
PLP	Budget-30h	17.7	50.0	44.2	43.4	33.4	15.5	**34.9**
	Auto. decision	17.7	49.7	44.2	43.8	32.9	15.1	**34.7**
PLP+BN	Budget–30h.	13.0	41.5	32.6	32.1	22.5	12.1	**26.3**
	Auto. decision	12.7	41.4	32.5	32.3	22.4	11.8	**26.2**

Table 3.6: Amount of data (hours) selected by the automatic budget decision

Domain	Amount
RD	41.2
TV	35.8
CT	21.9
MT	35.6
TK	31.4
RS	17.1

types of features, PLP and PLP+BN. Table 3.5 presents the results for these experiments and compares them to the outcome of data selection based on a 30-hour budget, which was the best fixed budget from figure 3.2. The results showed that the use of an automatically derived threshold improved the results obtained with a fixed budget for both types of features, indicating that the proposed method could estimate the right amount of data to select for each target domain.

The amount of data selected for each domain is presented in table 3.6. This table shows how read speech and conversational telephone speech are the ones which benefited from a lower amount of training data (20 hours or less), while the rest of the domains preferred more data (from 30 to 40 hours). These values were consistent with the patterns of positive and negative transfer observed in Figure 3.2 and suggest that the automatic budget decision is having the benefits of positive transfer, while avoiding the effects of negative transfer.

3.3.7 Summary

The effect of positive and negative transfer across widely diverse domains in ASR was explored and it was confirmed that the use of more data does not always reduce the WER. The effects of adding cross-domain data was studied and patterns of positive and negative transfer were studied. A sub-modular function based on likelihood

ratio was proposed and used to perform an informed and efficient selection of data for different target test sets. The evaluation of selection techniques based on budget and on automatic budget decision has achieved gains of 4% over a 60-hour MLE model for PLP features and 2% for PLP+BN features.

3.4 Phone posterior probability based distance

The likelihood ratio based data selection technique proposed in section 3.3 was purely acoustic and did not use any phonetic or linguistic information. That technique is suitable for selecting data in an unsupervised manner for bootstrapping new models. However, when initial models exist and the aim is to further improve the performance, Siohan (2014) showed that using a pure acoustic similarity measure for data selection biases the selection towards less informative data samples. For example in very large and possibly redundant datasets, it was observed that very short utterances with the same transcript were being selected with the pure acoustic metrcis Siohan (2014). To alleviate this problem, in this section another similarity measure is proposed which takes into account the phonetic contents of the utterances.

Phone posterior probabilities contain phonetic discriminatory information for phone recognition. They also contain other valuable acoustic information which can be used as a measure of similarity. A preliminary study has been conducted where the same set of utterances were perturbed with different levels of SNR (by additive background noise) and a reference model is used to compute the phone posterior probabilities. Figure 3.4 is an example of how the differences in the background noise affects the phone posteriors and shows a sample utterance that has been perturbed by additive background noise so the resulting signals have signal to noise ratio of 10 dB and 25 dB respectively. The horizontal axis in each posteriorgram corresponds to time in milliseconds and the vertical axis corresponds to the indices of context independent HMM states. Each point in the plots corresponds to CI posteriors computed by averaging DNN activations across context dependent states with the same centre context, resulting in a total of 121 context independent state posteriors. The DNN had four hidden layers with an output layer of size 4k. Further details about this network and the dataset are provided in section 3.4.3.2.

It can be seen in figure 3.4 that, for the mismatch caused by the additive background noise perturbation, there is an obvious impact on phone confusability as the SNR is reduced. While this impact on phone posteriors might not be as visually obvious for all mismatched conditions, this example suggests that it may be reasonable to use distance measures derived from these posteriors as a similarity measure.

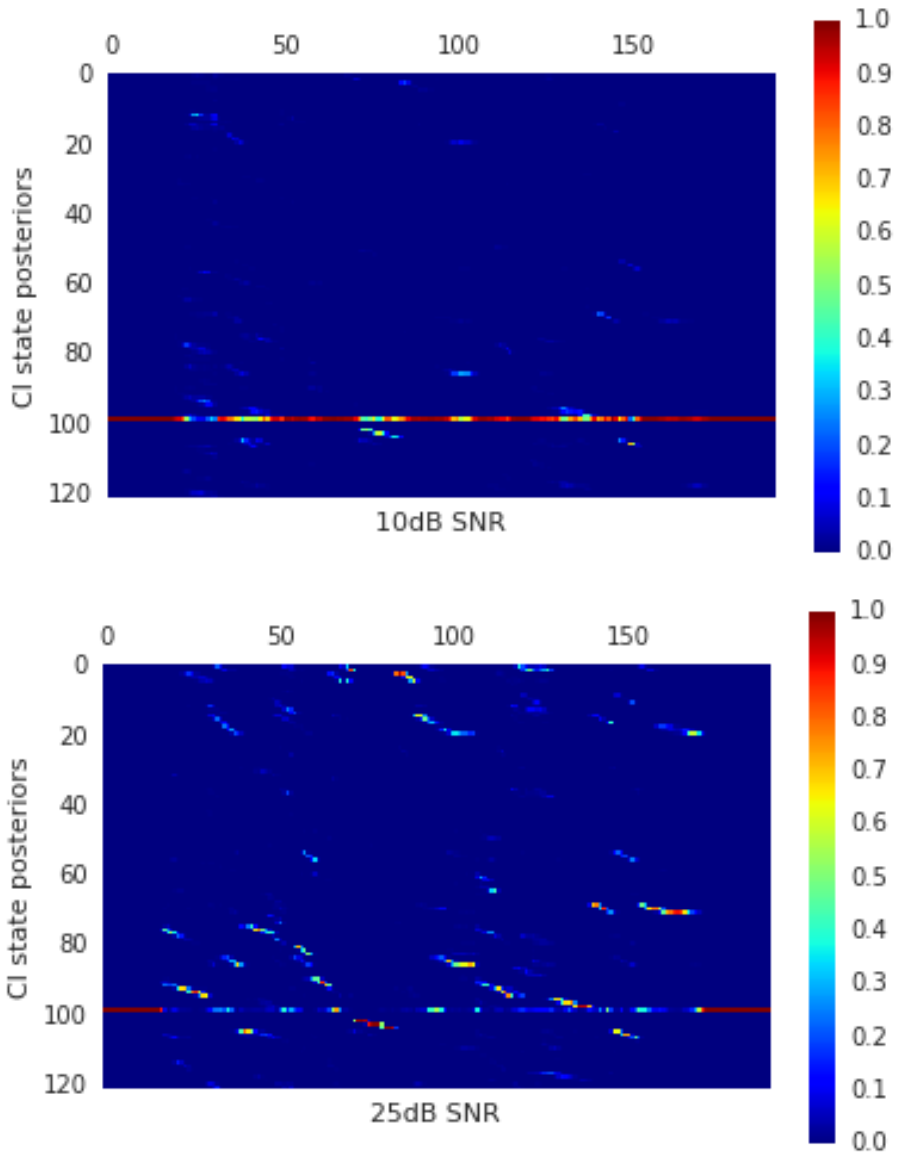


Figure 3.4: Impact of noise on phone posteriors for 10dB (top) and 25dB SNR (bottom) on the same 2 sec. utterance

Based on the arguments about the usefulness of phone posterior based features, several similarity measures were investigated for the function Φ is equation 3.1. A similarity measure based on the cosine distance between DNN phone posterior vectors that are averaged over a set of target domain utterances and training utterances proved to be effective in the initial experiments. The average phone posterior for a set of utterances $\mathcal{O}$ can be computed as:

$$P \cong \frac{1}{|\mathcal{O}|} \sum_{\mathbf{o} \in \mathcal{O}} \frac{1}{T} \sum_{t=1}^{T} r_{\mathbf{o},t} \tag{3.13}$$

where $r_{\mathbf{o},t}$ is the phone posterior probability of segment $\mathbf{o}$ at time t. With these

averaged posterior vectors, the distance function can be re-written as:

$$\Phi\big(P_{trn}, P_{tst}\big) = 1 - \frac{P_{trn}P_{tst}}{||P_{trn}||\ ||P_{tst}||}. \tag{3.14}$$

As already shown in the preliminary study in this section, average phone posteriors reflect the variabilities present in the data and thus might be good at quantifying the variabilities as well. By quantifying the variabilities, similarities and dissimilarity between data points can be identified and it can be used as a metric for the data selection and augmentation problem. To further study the suitability of this metric for measuring similarity, an empirical study was conducted to first verify the effectiveness of this approach in identifying and quantifying the levels of variability and secondly to find the optimal number of utterances needed to obtain a robust estimate of the statistics required for determining the variation levels (e.g. the size of $\mathcal{O}$).

3.4.1 Robust estimate of variability levels

In the experimental work conducted in this section, a source of variability was introduced to the utterances by artificially perturbing them by some SNR levels. A set of utterances from the target test set are perturbed by a fixed SNR level, $\alpha = 10$dB, and the training set utterances are perturbed with a set of perturbation levels corresponding to SNR values ranging from 0dB to 20dB at 2dB intervals. The motivation for this experiment is to show that the distance metric defined by equation 3.14 can capture the various levels of variability and in this particular case, different level of SNR values, and to select the utterances from the pool of training utterances that have the shortest distance to the target test set's utterances to create a training corpus. Note that the ultimate aim of this experiment is not to estimate the SNR levels, as there are dedicated algorithms for acquiring such estimations more accurately, but to provide a generic metric that can be used for quantifying other sources of variability present in the data as well. By quantifying the variabilities present in a target data set, similar utterances can be selected or augmented for creating a matched training corpus. The ultimate aim of creating a training corpus that has similar variabilities to a target test set is to reduce the WER on that target test set. A list of variability sources and levels are provided in section 3.4.3.1.

3.4.1.1 Identifying SNR perturbation levels

The underlying assumption in this experiment is that one can determine whether a given type and a given level of variability is present in a set of target domain utterances by perturbing an uncorrupted set of training utterances and measure

the similarity between the two data sets with the defined measure. This leads to the following procedure which is summarised in figure 3.5. Given an uncorrupted training set, X^{tr}, and utterances X_i^{ta} representing the ith sample of utterances from the target domain, determine the closest matching perturbation level, $\hat{\alpha}^t \in \mathcal{A}_t$, for perturbation type t. To do this, the similarity measure which is defined between the target utterances and the training utterances that have been perturbed by a given perturbation level, α, is used.

This similarity measure is defined over phone posterior probabilities obtained from perturbed training and target utterances. The posteriors are modelled by the outputs of an existing reference DNN, as shown in figure 3.5, whose inputs are features derived from the perturbed training utterances. The architecture of this reference DNN and other details about the training procedure are described later in this chapter. The posterior probability for phone index, k, given training observation vector, $x_{l,f}^{tr}(\alpha)$, from frame f of training utterance l when the utterance is perturbed by perturbation level α is given by $r_{l,f}^{\alpha}(k) = p(k|x_{l,f}^{tr}(\alpha))$. Hence, each observation frame is represented by a K dimensional vector of posterior probabilities, $r_{l,f}^{\alpha}$, where K is the number of phone classes.

Posterior probabilities are computed for both the perturbed training utterances and also for the target domain utterances. Figure 3.5 illustrates how this is done by computing statistics from posteriors derived from the perturbed training utterances and the utterances, X_i^{ta}, sampled from the target domain. These are depicted in the figure as C_{α}^{P} and C_i^T respectively. The similarity measure, $\Phi(C_{\alpha}^{P}, C_i^T)$, is then used to find an optimum perturbation level as:

$$\hat{\alpha}_i = \arg\min_{\alpha} \Phi(C_{\alpha}^{P}, C_i^T). \tag{3.15}$$

The statistics, C_{α}^{P} and C^T, for both sets are accumulated with varying numbers of utterances sampled from both the target and training data. Using the procedure defined in figure 3.5, SNR levels in the data set sampled from the target domain are then estimated. For example to estimate the perturbation level present in C_i^T (a sample from the target data with an unknown perturbation level), its distance is compared with all of the C_{α}^{P}, which are the utterances from the training set with known levels of perturbation (α). Then based on the cosine distances, the perturbation level of the C_{α}^{P} with closest distance is assigned as the perturbation level of C_i^T.

Figure 3.6 shows the classification accuracies where the blue bars are the classification accuracies when the target is the exact 10dB value and the red bars are the classification accuracies when the target is a window of 8dB to 12dB. The plot suggests that 300 utterances are enough to have a robust estimate of the statistics.

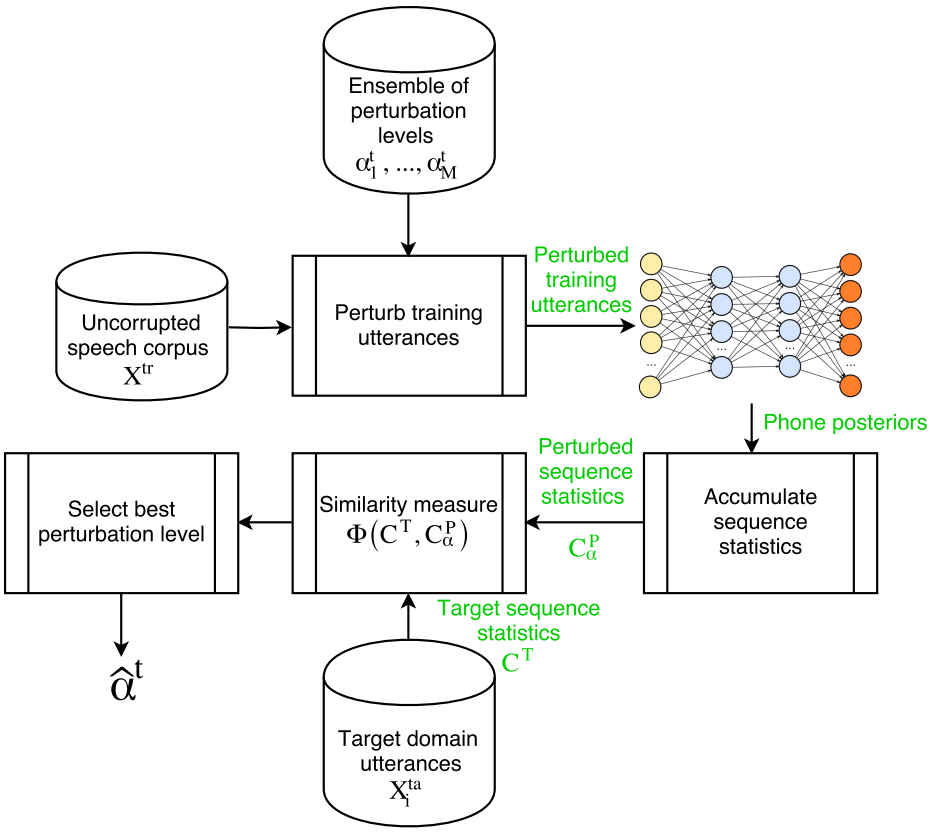


Figure 3.5: Perturbation level determination procedure

As discussed earlier, there are clearly many approaches for SNR estimation; however, similar behaviour was observed for the other perturbation types that will be listed in the remainder of this chapter. Hence, the perturbation level classification accuracy illustrated in figure 3.6 suggests that, with enough data, this can serve as a general approach for estimating perturbation levels for other perturbation sources.

3.4.1.2 Generalisation of the proposed approach to other sources of variability

To further investigate this distance measure and to examine the suitability of it for other intrinsic and extrinsic variabilities, an experimental study is conducted in this section. Extrinsic variability refers to ambient noise which includes a range of noise levels (signal-to-noise-ratios), background noise types, and room characteristics. Intrinsic variability corresponds to speaker and speaking style variation which is modelled in this work by introducing simulated frequency and tempo perturbation to the speech waveform (Ko et al., 2015; Verhelst and Roelands, 1993).

It is also assumed that each variability type, t, is represented by a discrete set of M variability levels, $\mathcal{A}_t : \{\alpha_1^t, \ldots, \alpha_M^t\}$.

The goal of this approach is to select optimum levels that, when applied to the

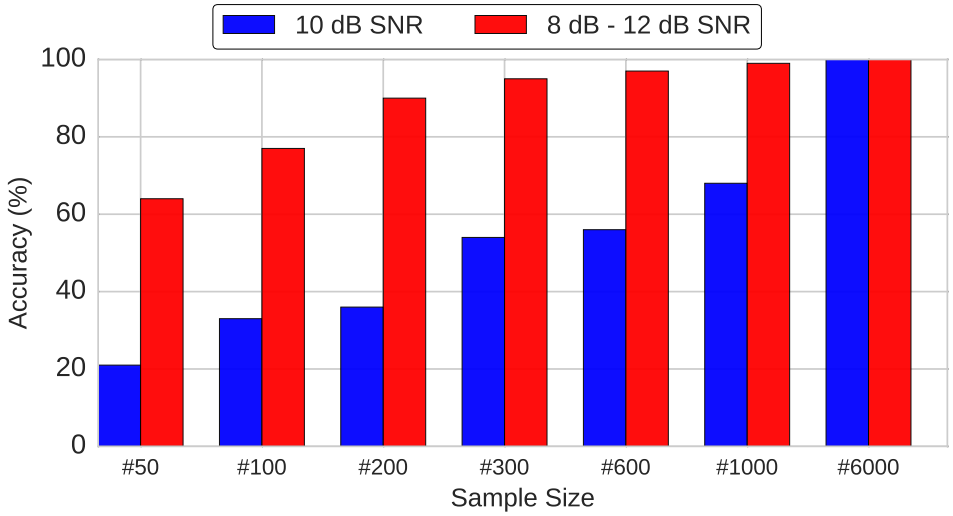


Figure 3.6: Classification accuracy of perturbation level over a range of dataset sample sizes

utterances in a potentially mismatched training corpus, provides a "best match" to the empirical distribution of the target domain utterances. This involves solving two problems. The first is to find the perturbation level, $\hat{\alpha}^t$, that provides the best match to a set of sample utterances, X_i^{ta}. This problem is already discussed in the context of similarity measure.

The second problem is to find a distribution, $p_t()$, of perturbation levels that provides the best match to the utterances from the available N sets of sample utterances. Section 3.4.2 provides a description of the process of identifying a set of distributions to be used for perturbing training utterances for creating an augmented training set.

The larger goal of these experiments is to find a better training set which is more similar to the target test set, e.g. in this case having the distribution of the variations in the training set similar to the distribution of variations in the target test set. This will ultimately reduce the mismatch and improve the WER. For this purpose, an experimental study is presented in section 3.4.3 where simulated target domains are created by introducing multiple levels of intrinsic and extrinsic variability. It will be shown that performing MTR with these estimated distributions results in a WER that approaches the "best case" WER obtained when performing MTR with distributions that are matched to the known target domain perturbation distributions.

3.4.2 Identifying perturbation distributions

This section addresses the problem of obtaining distributions of perturbation levels. Levels will be drawn from these distributions when perturbing training utterances to create a multi-style training corpus. Section 3.4.2.1 describes the approach used for finding a distribution of these levels for a single perturbation type to model a set of utterances taken from a target domain. Section 3.4.2.2 describes how this approach can be extended to identifying distributions of perturbation levels for multiple perturbation types.

3.4.2.1 Empirical distributions for a single perturbation type

The procedure for estimating a distribution, $p_t()$, over perturbation levels, $\mathcal{A}_t$, for a single perturbation type, t, is summarised by algorithm 1. The goal is for this distribution to assign weight to a given perturbation level based on the frequency with which data perturbed with that level is found to most closely match based on the distance measure to a set of utterances selected from the target domain. Given an uncorrupted training set, X^{tr}, and N sets of utterances, $X_1^{ta}, \ldots, X_N^{ta}$, sampled from the target domain, the procedure in algorithm 1 determines a distribution of perturbation levels, $p_t()$, that best matches all N data sets from the target domain. Then, the multi-style training set, X^{MTR}, can be generated by perturbing utterances with levels sampled from $\mathcal{A}_t : \{\alpha_1^t, \ldots, \alpha_M^t\}$ according to perturbation distribution, $\hat{p}_t$.

Algorithm 1 Perturbation distribution estimation procedure

Given: Training data X^{tr}, data sets $X_1^{ta}, \ldots, X_N^{ta}$ sampled from target domain, and perturbation levels $\mathcal{A}_t : \{\alpha_1^t, \ldots, \alpha_M^t\}$ for perturbation type t
Initialize Counts: $f_t(\alpha) \leftarrow 0 \quad \forall \alpha \in \mathcal{A}_t$
for All $X_i^{ta} \in \{X_1^{ta}, \ldots, X_N^{ta}\}$ **do**
  Compute target posteriors and stats (Fig 3.5): C_i^T
end for
for All $\alpha \in \mathcal{A}_t$ **do**
  Perturb training utterances: $X^{tr}(\alpha) = \mathcal{F}_t(X^{tr}, \alpha)$
  Compute training posteriors and stats (Eq. 3.13): C_α^P
  for All $X_i^{ta} \in \{X_1^{ta}, \ldots, X_N^{ta}\}$ **do**
    Compute similarity measure (Eq. 3.14): $\Phi(C_\alpha^P, C_i^T)$
    Perturb. level (Eq. 3.15): $\hat{\alpha}_i = \arg\min_\alpha \Phi(C_\alpha^P, C_i^T)$
    $f_t(\hat{\alpha}_i) = f_t(\hat{\alpha}_i) + 1$
  end for
end for
for $\alpha \in \mathcal{A}_t$ **do**
  $\hat{p}_t(\alpha) = f_t(\alpha)/N$
end for

Estimation of $\hat{p}_t$ can be described as follows. First, as illustrated in Figure 3.5, DNN posteriors are derived from the data sets, X_i^{ta}, and sequence statistics, C_i^T, are estimated from the posteriors. Second, X^{tr} is perturbed with each $\alpha \in \mathcal{A}_t$ to produce M perturbed versions of the training set, $X(\alpha)^{tr}$, $\forall \alpha \in \mathcal{A}_t$. The notation $\mathcal{F}_t(X^{tr}, \alpha)$ in algorithm 1 signifies the process of perturbing the training data set with a perturbation type t. Third, an optimum $\hat{\alpha}_i^t$ is identified for each data set sampled from the target domain. This corresponds to the perturbation level that, when applied to the training data, best matches the ith sample of utterances from the target data set according to the distance measure defined in equation 3.15. The frequency count, $f_t(\hat{\alpha}_i^t)$, associated with $\hat{\alpha}_i^t$ is incremented, and the perturbation distribution is obtained from the normalised counts, $\hat{p}_t(\alpha) = f_t(\hat{\alpha}^t)/N$.

Having estimated the perturbation distribution from multiple subsets of the target domain, this distribution is then used for perturbing the training utterances to create a final multi-style training set which has the minimum mismatch according to the defined measure to the target test set. For each training utterance, a perturbation level is randomly selected from the set $\mathcal{A}_t$ according to distribution $\hat{p}_t$. Section 3.4.3 describes how this multi-style set is used to train a DNN acoustic model and is then evaluated on utterances sampled from the same target domain.

3.4.2.2 Extension to multiple perturbation types

The procedures outlined in sections 3.4.1.1 and 3.4.2.1 address the problem of identifying a distribution of perturbation levels associated with a single perturbation type. The more general case would be to estimate a multi-variate distribution of perturbation levels across a set of P perturbation types. It is possible to combine the perturbation levels from all perturbation types and estimate a single multi-variate distribution. However, in these experiments, multiple univariate perturbation distributions are estimated, one for each perturbation type.

A sequential procedure is used for estimating distributions of perturbation levels for multiple perturbation types. The general outline of this procedure is summarised in figure 3.7. The process begins with sets of perturbation levels for P perturbation types, $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_P$. At each step of the process, an optimum level $\hat{\alpha}^t$ is selected using the procedure described in Section 3.4.1.1. Then, this $\hat{\alpha}^t$ is used to perturb the training utterances for all succeeding steps of the process when selecting perturbation levels for other perturbation types. For example, if perturbation set $\mathcal{A}_1$ corresponds to the set of possible noise levels and set $\mathcal{A}_2$ corresponds to room configurations, the first step of the sequential process would be to estimate the optimum noise level $\hat{\alpha}^1$. Then the training utterances would be corrupted using this noise level before selecting the closest matching room configuration, $\hat{\alpha}^2$, in the second step.

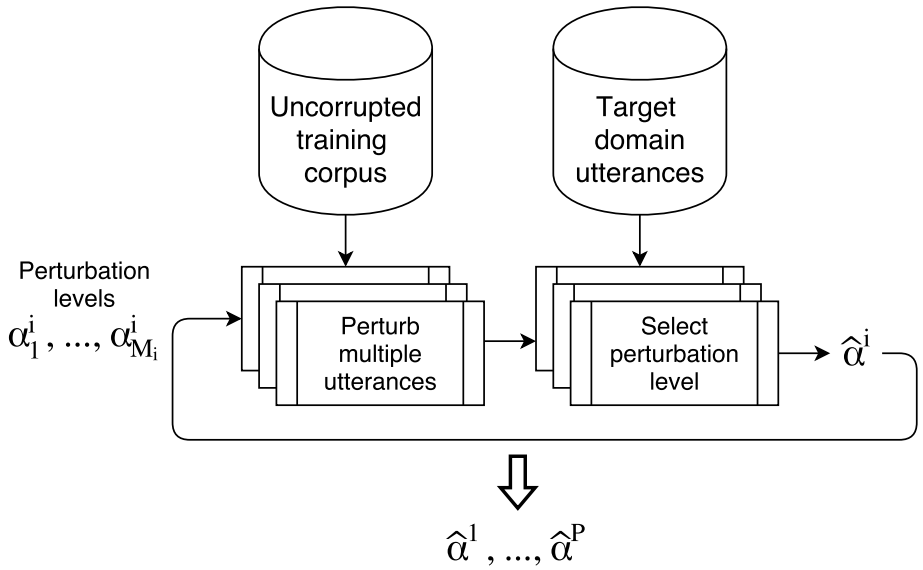


Figure 3.7: Sequential estimation of perturbation levels for multiple perturbation types

This process is repeated until perturbation levels for all P perturbation types have been identified. It was observed that finding the extrinsic variabilities first and then finding the intrinsic variabilities yield better performance and in the following experiments the same order was followed.

3.4.3 Experimental study

This section presents an experimental study to show how the defined similarity measure can be used to improve the WER in a simulated far-field speech recognition task. First, the speech corpus and the multi-style training scenario are described in section 3.4.3.1, and the baseline acoustic models are described in section 3.4.3.2. Then an evaluation of the approach described in section 3.4.2 for estimating perturbation distributions to best match a set of utterances sampled from a target domain is given in section 3.4.4. The goal of these experiments is to determine how these distributions, when applied to perturbing a training set in a multi-style training scenario, can reduce ASR WER on a set of simulated target domain utterances.

3.4.3.1 Simulated datasets and baseline models

MTR experiments were carried out by creating a corpus of utterances corrupted using a set of simulated perturbation types. These perturbations represent a range of room characteristics and acoustic background conditions, along with a range of speaker characteristics introduced by warping the time and frequency scales of the utterances. The simulated distortions were applied to a large set of anonymised American English voice search utterances, used internally at Google. The training

set consists of 200 hours of spontaneous speech consisting of 300,000 utterances. The test set contains 20 hours of spontaneous speech consisting of 30,000 anonymised American English voice search utterances. The utterances in these data sets were chosen to have relatively high SNR in order to approximate as close as possible a scenario where perturbation types are applied to clean utterances.

In these experiments, a set of $P = 4$ perturbation types were used to perturb both the training and target datasets. This implies that the types of perturbations that might be expected in a target domain are assumed to be known in the experimental study. Of course, this is not in general a practical assumption. As a result, it is important to note that the results reported here reflect the ability of this approach to match the given simulated domain, and there is no guarantee that this simulated domain is a completely accurate model of utterances arising from an actual far field acoustic environment or speaker population. However, it also assumed that the absence of a given source of perturbation is automatically determined by allowing the automated procedure to select a "no perturbation" level. For example, selecting a high SNR level implies the absence of additive noise, or selecting frequency or time warping equal to unity implies that speaker variation has minimal effect.

The implementation of these perturbation types and the size, M, of the perturbation sets are given as follows. The first is the signal-to-noise ratio associated with additive background noise. There are $M = 13$ levels ranging from 0dB to 24dB with approximately 60% of the target utterances corrupted with SNR levels above 15dB. The second perturbation type is the room impulse response produced from room specifications using the image model (Allen and Berkley, 1979). A total of 11 rooms are simulated, with reverberation times uniformly selected from values 0, 0.6, 0.77, 0.84, 0.88. The simulated distances between source and microphone ranged from approximately 0.3 meters to 2 meters.

The third perturbation type was frequency warping to approximate physiological differences within the speaker population. A total of 11 values were used, uniformly sampled over the range from 0.9 to 1.1. Finally, warping of the time axis was used to approximate speaking rate variation. Here, 11 values were used, uniformly sampled over the range from 0.9 to 1.1. For the frequency and time warping perturbations, waveform similarity overlap-add algorithm (WSOLA) was used (Ko et al., 2015; Parviainen, 2015; Verhelst and Roelands, 1993). Note that the time-stretching using WSOLA does not affect the pitch.

3.4.3.2 Baseline acoustic models

The acoustic models used for determining perturbation levels as depicted in Figure 3.5 are hybrid feed-forward DNNs. The input features to the network consist

of 26 stacked frames of 40 dimensional Mel-scale log-filter bank energies. The network has 4 hidden layers with 1280 nodes per layer and a 4000 node output layer where the output nodes correspond to context dependent HMM states. The posterior vectors, $r_{l,f}^{\alpha}$, used in equation 3.13 correspond to $K = 121$ dimensional CI state posteriors obtained by summing over the CD state activations with the same centre phone context. The DNNs used in Figure 3.5 are trained with the CE criterion from the uncorrupted 300k training utterances.

The acoustic models used to evaluate ASR performance for multi-style training have the same form as those described above. After perturbing the training data using one of the MTR scenarios described in Section 3.4.4, the perturbed training utterances are used for training the DNN. In the MTR training scenario, this DNN, after being initially trained from clean data using CE training, is sequence trained with the state level minimum Bayes-risk criterion (Gibson and Hain, 2006) from the perturbed training set. The first two rows of table 3.7 presents the WER for the cases where the DNN is sequence trained from the uncorrupted (clean) data and evaluated on the clean and the noisy data respectively. Here, the noisy evaluation data is created by perturbing the 30,000 utterance test set with perturbation levels sampled from the above perturbation sets. It is clear from the table that the error rate more than doubles when DNN models trained from uncorrupted data are used for recognition on noisy test utterances, which shows the level of mismatch in training and test data.

3.4.4 Optimised perturbation distribution

The performance of the approach for estimating perturbation distributions was evaluated using the following steps. First, the sequential procedure described in section 3.4.2.2 is used to find the perturbation distributions for all four perturbation types in the order of background noise level, room impulse response, frequency warping, and time warping. For each perturbation type, the procedure outlined in algorithm 1 is used to estimate distributions over perturbation levels. Second, these estimated distributions were used to select perturbation levels from the four perturbation types for perturbing the utterances of the training set described in section 3.4.3.1. Finally, this training set was used for sequence training of the DNN model described in section 3.4.3.2, and this model was used for decoding on the simulated target domain test set.

The WER obtained for this model on the noisy test set is shown in the third row of table 3.7. The WER obtained for the estimated perturbation distributions is almost 20% absolute lower than the WER obtained using the DNN trained from the uncorrupted training set. However, the impact of using these estimated perturba-

Table 3.7: WER (%) using MTR training scenarios

Training Set	Test Set	WER%
Clean	Clean	24.7
Clean	Noisy	55.1
Estimated perturbation	Noisy	35.2
Oracle perturbation (best case)	Noisy	33.5
Uniform perturbation (worst case)	Noisy	39.3

tion distributions for perturbing the data set relative to other perhaps more adhoc approaches is not clear from this comparison. To provide a better comparison, two additional MTR scenarios are considered. The first is a "best case" scenario (the oracle experiment) corresponding to perturbing the training set by selecting perturbation levels from perturbation distributions that match the target domain test data. The second, "worst case" scenario, corresponds to using training utterances that are perturbed using uniform random perturbation distributions. In both of these scenarios, the DNN models are sequence trained using the perturbed training sets and decoding is performed on the target domain test data. The WERs for these "best case" and "worst case" MTR scenarios are shown in rows four and five respectively of table 3.7. It is clear that the WER obtained for the estimated perturbation distributions is over 4% absolute lower than the worst case scenario and begins to approach the best case WER.

3.4.5 Summary

The goal of the experiments presented in this section was to capture the distributions of variations present in a target test set in order to create a training corpus which matches those variations, with the ultimate goal of reducing the mismatch in the training and test conditions and to improve the WER. For this purpose, a similarity measure was proposed together with a technique to learn the empirical distributions of variations present in the data. A multi-style training set was generated for a far-field speech simulated target domain by automatic optimisation of perturbation distributions. The training set resulting from performing MTR training using these estimated distributions was evaluated by measuring WER on a simulated far-field test set. It was found that the WER obtained using these distributions approaches that obtained for the best case scenario corresponding to a perturbation distribution that matches the target domain, and is considerably lower than the WER obtained for the worst case where distributions are randomly chosen.

3.5 Conclusion

In this chapter an overview of data selection techniques for ASR was provided. The data selection techniques were studied in the context of reducing the mismatch between the training and test conditions with the ultimate goal of improving the recognition accuracy. Two new approaches for data selection were introduced. A similarity measure based on likelihood ratio was proposed where the training data is selected based on similarity to a target test set and the experimental results were provided using a highly diverse dataset. In this dataset, data from six different domains were pooled together and various mismatched conditions when using out-of-domain data and cross-domain data were studied. It was shown that using the proposed method, the WER can be improved under different mismatched conditions.

The second approach was based on phone posteriors computed by a reference model. First the effectiveness of the proposed metric in quantifying the variations present in the data in the form of signal-to-noise ratios was studied and then it was generalised to learn the distributions of other sources of variability. Then these distributions were used to create a training corpus which matches the distributions of variations present in a target test set. It was shown that using this MTR training corpus reduces the WER significantly when compared to using a uniformly perturbed training corpus.

CHAPTER 4

Identification of genres and shows in media data

4.1 Introduction

The amount of digital media is growing larger and larger every day due to digital televisions, online streaming services and social media. There are over 28,000 TV broadcasting stations in the world, every minute more than 300 hours of video is being uploaded just to YouTube and on a daily basis, users around the globe spend more than 100 million hours watching Facebook videos (Central Intelligence Agency, 2016; Facebook, 2016; YouTube, 2016). This creates a huge demand for effective techniques for automatic processing of these digital media so that their content can be easily searched, retrieved and navigated.

Multimedia data may have some associated meta-data which facilitates the automatic processing for the downstream tasks such as indexing. Meta-data can be either structured or unstructured. Examples of structured meta-data include genre labels, number of speakers, speaker labels, duration, date and time of production, date and time of broadcast, broadcast type, broadcast media, etc. Examples of unstructured meta-data include description and textual summary. Genre labels may include news, sports, comedy, documentary and drama, which are categories that imply more than purely semantic information. For example shows that belong to the same genre may share similar acoustic conditions.

Some of these meta-data are objective, such as duration and some are subjective, such as genre labels. The objective properties are usually observable and measurable, while the subjective properties might not be measurable easily and in some cases impossible to measure. Even for humans assigning subjective tags can be challenging, for example a news programme that discusses oil price rise after death

of a political figure can be considered as belonging to either of these genres: news, finance or politics.

The extra information provided by meta-data is usually used for efficient querying, navigation, browsing and discovery (Chowdhury, 2010). For example classification of multimedia data into genres or other categories makes content discovery easier for the users of information retrieval systems.

The meta-data tags might not always be available for the data, especially for the subjective properties such as the genre labels. Also with huge historical digital archives, there might be some inconsistency in the tags, especially if the tagging was performed manually by several people. Manual labelling of the digital archives is usually not considered as a viable option even for the medium-sized archives, especially with budget and time constraints. Thus, automatic labelling of genres or other similar tags is an important task for multimedia and information retrieval systems and is the main motivation of the study in this chapter. Furthermore, since shows that belong to the same genre usually share similar acoustic conditions, this information can be used in acoustic model adaptation for mismatch reduction as well. This further motivates the study conducted in this chapter. The empirical results to support this argument for improving the WER of ASR systems is provided in the next chapter. In this chapter, the main aim is to automatically tag the media data with genre and show labels.

Research in the media processing field is further motivated by initiatives such as the "MediaEval benchmarking for multimedia evaluation" (Larson et al., 2013), or the "Robust, as accurate as human genre classification for video" challenges within the multimedia grand challenges of the ACM multimedia conference (Challenge, 2010).

Given the applications of genre labelling in multimedia information retrieval systems and their potential applications in acoustic model adaptation in ASR systems, the main research question this chapter is trying to answer is how broadcast media data can be classified into subjective tags such as genre labels using audio. It further investigates which sources of information are required for further improving genre classification accuracy. To answer these questions, two techniques for genre identification are proposed in this chapter. The first approach is based on a set of local features called background tracking features and the second approach is based on a latent modelling technique called latent Dirichlet allocation (LDA). The LDA approach is also used for the show identification task for the first time. An overview of genre identification techniques is provided in the next section.

4.2 Overview of genre identification

Genre identification or genre ID is the task of assigning a genre label from a set of predefined labels to media data. Genre labels can include advice, children's, comedy, competition, documentary, drama, events and new. Research on genre ID tasks typically report accuracies of over 90% (Ekenel and Semela, 2013; Kim et al., 2013; Montagnuolo and Messina, 2007, 2009). Typical datasets that are used by the researchers in this field are the RAI dataset (Montagnuolo and Messina, 2007), Quaero dataset (EU Quaero Programme, 2011) and some custom YouTube videos. The RAI and Quaero datasets are around 70 hours each and most of other datasets have similar or smaller sizes.

Genre labelling is difficult even for humans, mostly because of its subjective nature. Labels of genres differ among datasets and this makes interpretations of results difficult. Also the chosen labels do not always fully reflect all of the possible genres that appear in TV programmes; for instance the RAI dataset has 7 genres labels. These 7 genres are cartoon, commercial, football, music show, news, talk show and weather forecast, which seem to be in some cases very specific, e.g. football which can be considered as a subset of a broader sport genre. For example, it is not very clear that if a tennis programme is added to this dataset, what the genre label would be, sports or tennis.

Visual and acoustic features can be used for the genre identification task. Meta-data can also be used to further improve the accuracy. Multi-modal approaches also try to further improve the accuracy by combining audio-visual features. The focus of this chapter will be mostly on acoustic-based approaches. However, textual and meta-data will also be studied.

Initial approaches for genre identification include the use of generative models such as GMMs with short-term features, such as MFCCs or PLPs. Kim et al. (2013) reported an accuracy of 93.6% on a 11.5h test set with the RAI dataset using GMMs trained with the MFCC features. These features represent short-term characteristics of speech, such as the spectral properties of phonemes and speakers. In smaller and more homogeneous datasets where the same shows and speakers might often reoccur, the classification accuracy with these features is usually much better than the accuracies obtained on larger and more heterogeneous datasets (Saz et al., 2014). Later in this chapter and in section 4.5.4 the performance of GMMs with short-term features on a large dataset is empirically studied and it will be shown that these features are not very suitable for larger and more heterogeneous datasets.

Using video-based features, an accuracy of 99.2% was reported by Ekenel and

Semela (2013) for the RAI dataset (compared to a baseline of 93.6% with GMMs). For other similar datasets such as the Quaero dataset, similar classification accuracies were reported, e.g. 94.5% by the same authors. On a custom YouTube dataset, they also reported an accuracy of 87.3% which was further improved to 89.7% by the use of meta-data.

The probabilistic approach using GMMs can be further extended using latent modelling techniques. Kim et al. (2013) reported an absolute improvement of 0.7% over their 93.6% GMM baseline on the RAI dataset using acoustic topic models. They used vector quantisation (VQ) to represent frames by discrete symbols and trained acoustic latent Dirichlet allocation models (Kim et al., 2009a) followed by support vector machine classifiers. Some parts of the work presented in section 4.4 are built upon this work and a detailed review of the latent modelling techniques is presented in that section.

Neural networks have been also applied to the genre identification task. Montagnuolo and Messina (2007) reported an accuracy of 92% on the RAI dataset using a feed forward neural network with one hidden layer using acoustic features. In an other work, the same authors further improved the genre classification accuracy to 95% using various types of audio-visual features and different neural network architectures (Montagnuolo and Messina, 2009).

Sageder et al. (2016) tried to pool various types of features and then group and select a subset using canonical correlation analysis in order to identify low-correlated and complementary features. These features were then used to train different classifiers such as K-nearest neighbour, random forest and support vector machine. They reported very good classification performance on different datasets including some custom RAI and BBC shows, however the amount of data is tiny (less than 55h in total and in case of BBC, 4.5h with just 3 classes) and thus it is hard to directly compare with other approaches that uses larger and more heterogeneous datasets.

Other approaches try to identify certain audio-visual events and learn the semantics of broadcast shows or YouTube videos (Castan and Akbacak, 2013; Lee and Ellis, 2010). These techniques are considered to be structural analyses methods, since the main aim is to identify some certain events such as applause, music, etc. and identify the genre from a sequence of these events. This might be similar to how humans identify genres using audio. For example hearing lots of laughter in a show might be a good indication of a comedy genre, or hearing a cacophony of cheering and applauses may indicate a sport genre. These methods were mostly experimented on using small and limited datasets, such as a small selection of YouTube videos (less than 50 hours for training and testing) (Castan and Akbacak, 2013) and

their generalisation on larger and more heterogeneous datasets has not been studied.

From the reviewed approaches in this section, structural analysis methods and latent modelling techniques will be further investigated in the remainder of this chapter. This chapter introduces two new contributions of this thesis: a structural analysis method and a latent modelling technique for the genre ID task. Also for the first time, show identification task will be studied in this chapter.

A new structural analysis method is introduced in section 4.3 for tracking the background events such as music, laughter, applause, etc. and using features derived from these local descriptors, genre identification task will be performed. Furthermore, because of the complex nature of media data, modelling the variabilities as some latent variables might be better than explicit modelling. In section 4.4, a new technique for genre and show identification based on latent modelling techniques will be presented.

4.3 Background tacking features for genre identification

In this section a technique for extracting a new type of features, called background tracking features is provided followed by an experimental study of using these features for the genre identification task. The experimental study is based on this assumption: the composition of background conditions present in a show can be used for identifying the genre labels. These local descriptor features are derived from the asynchronous factorisation of speaker and background in the audio signal. A brief overview of the asynchronous factorisation is provided next.

4.3.1 Asynchronous factorisation of background and speaker

As shown in chapter 2, CMLLR transformations can be used for speaker and environment adaptation. These transformations are usually applied to whole utterances, assuming the stationary nature of the variability in the utterances. In case of using CMLLR for speaker adaptation, that assumption is usually valid, however, when such transformations are applied for environment adaptation, such assumption is not very realistic. In real datasets, such as TV recordings, even within a short utterance, the background and environment can change rapidly. This served as the motivation for Saz and Hain (2013), where instead of learning and applying a linear transformation on the utterance level, the transformations are applied on the frame

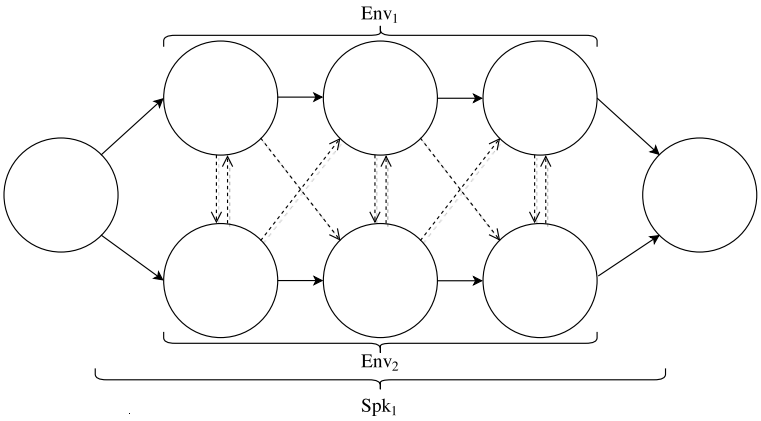


Figure 4.1: Asynchronous HMM topology with two environments, adapted from Saz and Hain (2013)

level (with some constraints). The CMLLR adaptation can be written as:

$$\hat{\mathbf{o}} = \mathbf{W}\mathbf{o} + \mathbf{b} \tag{4.1}$$

where $\mathbf{o}$ is the original utterance, $\mathbf{W}$ is a weight matrix, $\mathbf{b}$ is a bias vector and $\hat{\mathbf{o}}$ is the transformed utterance.

For the same utterance, different transformations can be applied in a cascaded manner, such as one for the environment and one for the speaker (Seltzer and Acero, 2011). The cascaded transformation can be written as:

$$\hat{\mathbf{o}} = \mathbf{W}^{spk}(\mathbf{W}^{env}\mathbf{o} + \mathbf{b}^{env}) + \mathbf{b}^{spk} \tag{4.2}$$

where the subscript *spk* and *env* are the transformations for the speaker and the environment. Saz and Hain (2013) proposed an asynchronous version of the cascaded transformation for each frame:

$$\hat{\mathbf{o}}_t = \mathbf{W}^{spk}(\mathbf{W}_t^{env}\mathbf{o}_t + \mathbf{b}_t^{env}) + \mathbf{b}^{spk} \tag{4.3}$$

where the subscript t denotes time.

To learn and apply these transformations, one can opt for a frame classification approach, where the presence of any background condition for each frame is determined by an external classifier. However, Saz and Hain (2013) proposed a modification to the structure of an HMM, where for each type of the background conditions, the emitting states are duplicated in parallel and extra arcs are added to allow transitions between the parallel states. Figure 4.1 presents the modified topology for two environments, where two transformations are for the environments and one transformation is for the speaker.

Depending on how the transitions between the parallel states (the dashed lines

in figure 4.1) are handled, two different types of models can exist: fully synchronous models have all of the possible transitions, and phone synchronous models where the transitions specified with the dashed line are removed. In the phone synchronous model, the assumption is that during the time that a phone is being uttered, the background conditions do not change. Results in (Saz and Hain, 2013) suggest that having phone synchronous models during adaptation yields better performance and during decoding, a fully synchronous model is preferred.

For more detailed information about the training procedure of these models, refer to Saz and Hain (2013). With these models, the sequence of states can be extracted during a decoding or if the transcript is known, during an alignment to the correct transcript. This way each frame will have a label, which corresponds to the index of the most likely environment for that frame. These indexes are then represented by one-hot-vector encoding and averaged over a sequence of P frames for the audio segment of length T frames. This yields $M = {}^{T}/_{P}$ vectors with the dimensionality of number of environments: $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$

$$\mathbf{y}_p = \frac{1}{P} \sum_{t=P(p-1)+1}^{Pp} \mathbf{x}_t \tag{4.4}$$

where $\mathbf{x}_t$ is the one-hot-vector encoding of the environment index. A graphical description of the process is provided in figure 4.2. 12 frames are presented and the number of environments is 4 and $P = 12$ which yields a single 4 dimensional vector ($M = {}^{12}/_{12} = 1$). The final representation is based on the normalised counts of the environments.

These aggregated vectors can then be used as features in different classifiers. The experimental setup's details are provided in the next section.

4.3.2 Experimental setup

4.3.2.1 Dataset

To experiment with the background tracking features in a genre ID task, data from 332 shows with a total amount of 231 hours which were broadcast by the BBC during the first week of May, 2008 were used. According to the internal genre classification of the BBC, these shows are classified into these eight genres:

- advice: consumer, do-it-yourself and property shows
- children's: cartoons and educational shows
- comedy: situational comedy and light entertainment shows

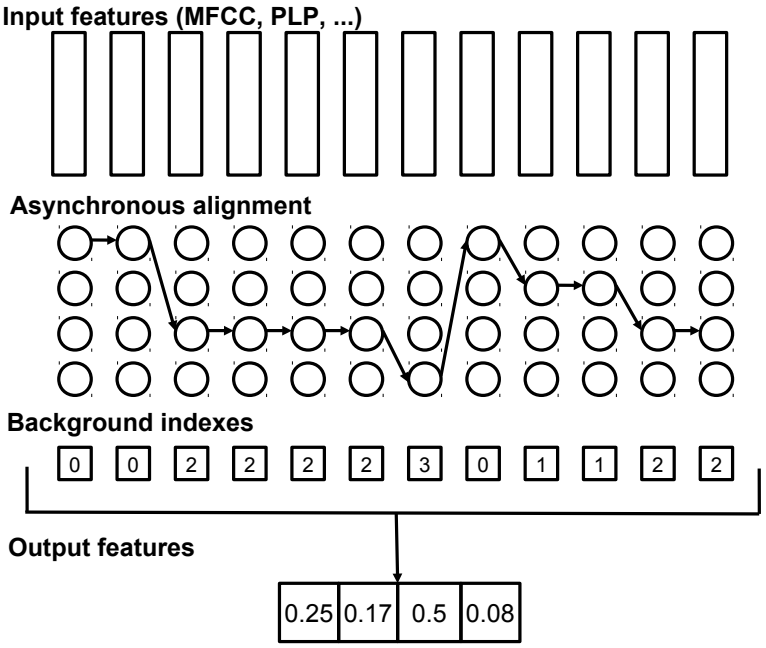


Figure 4.2: Background tracking features extraction process, adapted from Saz and Hain (2013)

- competition: quiz shows and other contest shows
- documentary: including fly-on-the-wall shows
- drama: soap operas and other serialised dramas
- events: live events, sports and concerts
- news: broadcast news and current affair shows

Since the shows cover a whole week, it includes a mixture of the genres and in this sense it is a more realistic scenario compared to the limited RAI dataset. These genres are very heterogeneous as well, for example events genre covers live sports as well as music shows.

The split between the training and test set was performed by selecting 285 shows for the training set and 47 shows for the test set. Amount of data and number of shows per genre for the training and test set is presented in table 4.1. This dataset is called dataset A in the remainder of this chapter.

4.3.2.2 Extracting background tracking features

As discussed earlier, if the correct transcript of the data is available, then the background tracking features can be extracted by aligning the transcripts to the audio signals and keeping track of the best path through the states. However in the case of

Table 4.1: Amount of training and test data (hours) per genre in dataset A

Genres	Training set		Test set	
	#Shows	Duration	#Shows	Duration
Advice	34	24.5	4	3.0
Children's	45	18.5	8	3.0
Comedy	20	9.7	6	3.2
Competition	37	25.9	6	3.3
Documentary	41	29.8	9	6.8
Drama	19	14.4	4	2.7
Events	23	29.8	5	4.3
News	66	50.3	5	2.0
Total	**285**	**203.0**	**47**	**28.3**

dataset A, only subtitles were available. A lightly supervised training procedure as described in (Lanchantin et al., 2013) was used for the training of the GMM-HMM acoustic models which were then used for the forced alignment of the data.

Seven initial CMLLR transformations were trained on a modified version of the WSJCAM0 (Robinson et al., 1995) corpus, as described in Saz and Hain (2013). These seven transformations correspond to these acoustic backgrounds: clean speech, classical music, contemporary music, applause, cocktail party noise, traffic noise and wildlife noise and were retrained asynchronously on the BBC dataset. After this initial stage, the feature vectors were processed using $P = 100$ which yielded 7 dimensional feature vectors.

4.3.2.3 Visualising the background tracking features

Using the procedure described in section 4.3.2.2, the features were extracted and aggregated. Each aggregated feature vector corresponds to one second of the audio segment. Figure 4.3 visualises 60 seconds of these features for four different shows. The 7-dimensional features are represented by bar plots in each column (which corresponds to one second). Visually inspecting these plots and trying to synchronise it with the audio, the changes in the distribution of the feature vectors correspond to the events happening in the background. For example for figure 4.3a, the news programme starts with music, then changes to street noise, then to clean studio speech and finally ends with some street noise. Figure 4.3b, is a cut from a music event show and shows music changes from rock music to solo singing and ends with instrumental rock music. Figure 4.3c presents a historical documentary show that

Table 4.2: Genre classification accuracy (%) with GMM models and short-term PLP features on dataset A

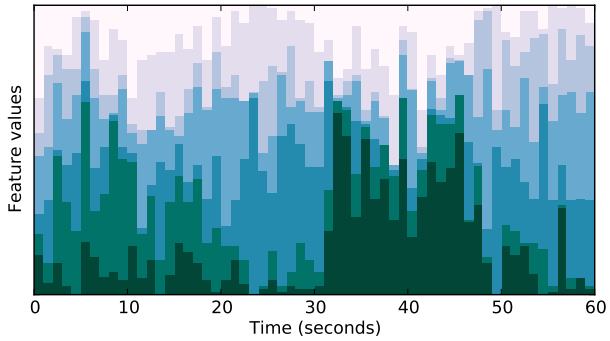
#Components	Accuracy
8	44.7
16	48.9
32	48.9
64	48.9
128	53.2
256	53.2
512	61.7
1024	59.6
2048	61.7

starts with bell sounds and whistles, then continues with some music, followed by some clean speech and ends with some birds song and seaside noises. Figure 4.3d corresponds to a minute cut from a light entertainment show and has portions of speech with long laughter bursts.

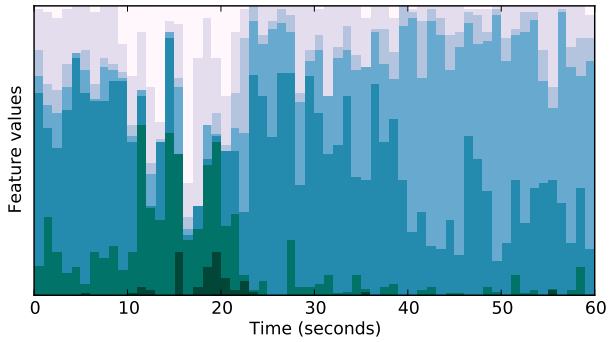
4.3.2.4 Baseline

To evaluate the performance of the proposed approach for the genre identification task, first the baseline experiments are performed. As a baseline classifier, GMMs were trained with the PLP features. The 13 dimensional PLP features were extracted every 10ms and their first and second derivatives were added to form a final 39 dimensional feature vector. GMMs with a varying number of mixture components were trained using the EM algorithm and the mix-up procedure for each of the 8 genres. The label assignment to the new data was based on computing the overall likelihood of the frames with all of the 8 models and picking the GMM with the highest likelihood. This baseline enables the comparison of the dataset and the proposed approach with other related techniques which were introduced in section 4.2. Table 4.2 summarises the classification accuracy with GMM classifiers with varying number of mixture components.

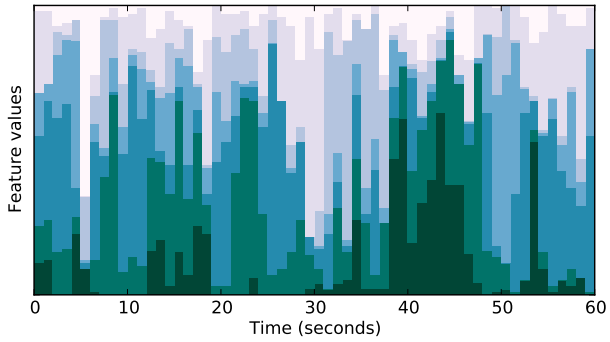
Comparing the results obtained here with the results reported in the literature on other datasets such as the RAI dataset, shows how challenging this BBC dataset is. Best accuracy for this dataset is obtained with a GMM with 512 mixtures, which is 61.7%, while the best accuracy with the GMMs for the RAI dataset was reported as 93.6% (Kim et al., 2013).



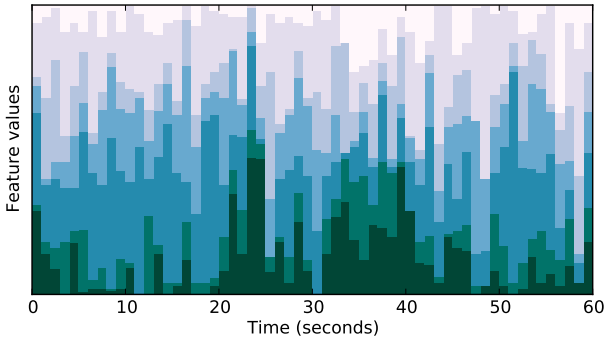
(a) News genre (broadcast news)



(b) Events genre (live music show)



(c) Documentary genre (history show)



(d) Comedy genre (light entertainment)

Figure 4.3: One-minute samples of background tracking features for four different shows, adapted from Saz et al. (2014)

Table 4.3: Genre classification accuracy (%) with GMM models and background tracking features on dataset A

#Components	Type	
	Subtitles	Decoding
8	59.6	59.6
16	66.0	63.8
32	66.0	68.0
64	72.3	68.1
128	70.2	68.1
256	68.1	66.0
512	70.2	70.2
1024	59.6	59.6
2048	53.1	49.0

4.3.3 GMM classification with the background tracking features

The proposed background tracking features can be used as input features to train the GMM models. These 7 dimensional features were augmented by adding first and second derivatives to form a 21 dimensional vector. Also note that since $P = 100$ were used, the length of the new representations are 100 times smaller than the baseline experiments. A total of 73,528,233 frames were available for the training of the GMMs with the PLP features, however, for the training of the new GMMs with the background tracking features only 730,621 frames were used. As discussed, the background tracking features can be derived by alignment (when the ground truth labels are available) or by decoding (when the ground truth labels are not available). In this section both cases were studied. The first experiment was conducted using the provided subtitles with the lightly supervised training procedure (Lanchantin et al., 2013) and in the second experiment outputs of an ASR system were used. Table 4.3 summarises the classification accuracy using GMMs and background tracking features with the subtitles and the output of decoding.

The differences between using subtitles and the output of decoding does not vary much and it shows robustness of this approach to the use of inaccurate transcripts. The ASR system used in the experiments had around 30% WER on this dataset. In the following experiments, the background tracking features are derived from aligning to the subtitles.

4.3.4 HMM classification with the background tracking features

After the successful application of the GMMs with the background tracking features, in this section HMMs are studied with the same input features. HMMs can model the temporal transitions among the hidden states that exist in the data and for this task they should provide extra gains. For the HMM experiments it was found that models with 8 states provide the best results and therefore were used in the experiments. The emissions of the HMMs were modelled with GMMs and all of the parameters were learned using the EM algorithm. Similar to the GMM experiments, 1 HMM for each genre was trained and the assignment of the data to the class was performed by picking the model with the highest likelihood. The classification accuracies using the HMMs are provided in figure 4.4 and compared against the other classifiers. HMMs outperformed GMMs in this task and the best accuracy with the HMMs was 78.7% which was achieved with a HMM with 8 states and 32 mixture components in each state.

4.3.5 SVM classification with background tracking features

GMMs and HMMs can deal with variable length inputs and were applied successfully for the genre ID task. The use of discriminative classifiers, such as support vector machines is studied in this section. Since SVMs can not deal with variable length inputs, the inputs should be mapped to a fixed length. For this purpose a similar approach to what is used in the speaker ID tasks was used. GMM parameters were adapted to the shows using the MAP adaptation approach and the mean vectors of the Gaussian components were stacked to form a super-vector. The super-vectors were used in the SVM classifiers with radial basis function kernels (Cortes and Vapnik, 1995). The parameters of the SVM classifiers were tuned using grid search over a range of values and a held-out cross validation set was used for the evaluation. SVMs are binary classifiers and here one-against-one approach (Knerr et al., 1990) was used for handling multiple classes. This approach requires training $n\frac{(n-1)}{2}$ classifiers where n is the number of classes. These classifiers are trained for each pair of the classes and during the prediction time a majority voting scheme is used for assigning the predicted class label.

The classification accuracy using the SVMs are provided in figure 4.4 and compared against the accuracy of GMM and HMM classifiers. Both HMMs and SVMs outperformed the GMMs and also SVMs outperformed the HMMs, with less complex models (16 mixture components compared to the 256 components). The best accuracy for this task was 80.9% which was achieved by the SVMs.

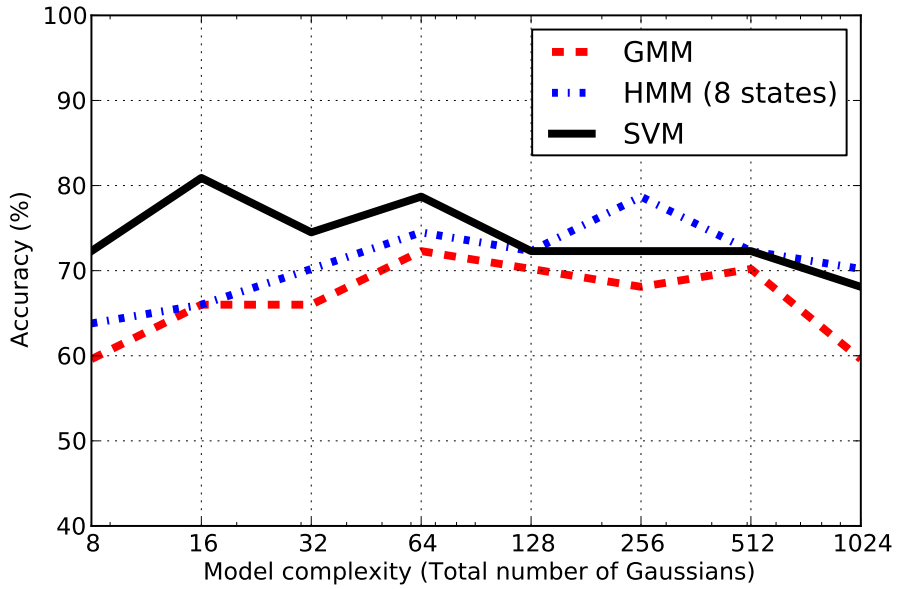


Figure 4.4: Genre classification accuracy (%) using GMMs, HMMs and SVMs on dataset A

4.3.6 System combination

Using the confidence scores from the HMM and SVM models, the outputs of the two systems can be combined. The assumption in most of the system combination tasks is that different models have different types of errors and by combining their output, the overall accuracy can be improved. The confidence scores from the HMM models were based on the likelihood scores of the selected HMMs, while for the SVMs the scores were calculated using a five-fold cross-validation (Silva and Ribeiro, 2009; Wu et al., 2004). When both systems output the same class, then it will be used as the final label, in case they disagree, the output of the system with the higher confidence will be used. The system combination further improved the classification accuracy to 83.0% which is 2.1% absolute improvements over the best single model accuracy (80.9% with the SVMs).

4.3.7 Summary

In this section a new approach for genre identification based on background tracking features was described. These features were extracted from the application of background audio specific transformations and keeping track of which environment is selected for each frame and then using the index of the environment to form a new feature vector. These vectors are then aggregated by averaging over a moving window and used as input features in different classifiers such as GMMs, HMMs and SVMs. The dataset used in the experiments was over 230h of BBC broadcasts in 8

genres. The best single system was the SVM classifiers which yielded an accuracy of 80.9%. Using system combination the accuracy was further improved to 83.0%. The results suggest that these features can be used for media data in more realistic scenarios. One issue with this approach is the extra computation required for decoding or aligning the transcripts. In the next section, an alternative approach is introduced where the transcripts are not required.

4.4 Discovering latent domains in media data

4.4.1 Latent modelling using latent Dirichlet allocation

Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a Bayesian probabilistic generative model for collections of discrete items. It describes how every item within the collection is generated using a finite mixture of latent variables, typically referred as topics in text modelling. Each topic is also described by a finite mixture of an underlying set of topic probabilities. This model was originally proposed in the context of text modelling, where documents within a collection of documents are modelled with a mixture of topics. Furthermore, topics are distributions of words. However this model is a generic model and can be used for other types of data. Section 4.4.1.4 summarises other use cases of LDA model beyond text modelling. This model seems to fit the complex nature of diverse data in terms of modelling each audio segment with a mixture of latent domains that contributed in generating that data. In the context of text modelling, a brief overview of similar techniques is presented next, followed by an introduction to the LDA model and its acoustic variant.

4.4.1.1 Latent semantic indexing

In the context of text modelling and information retrieval (IR), a popular approach to represent a collection of documents with a varying number of words is to use term frequency–inverse document frequency (tf-idf) representation (Salton and McGill, 1986). First a set of words or discrete symbols are defined and then the frequency of each word in each document is computed. Then document frequencies are also computed for each word which represents the count of documents in the whole corpus that contains that word. The term frequencies are combined with the inverse of the document frequency (usually on a log scale) to yield the final representation of each document. These vectors for each document are stacked to form a matrix which represents the whole corpus. The tf-idf metric shows the importance of each term in a document and in the entire corpus and can be thought of as a dimensionality

reduction technique, where documents with variable length of words are mapped to a fixed length representation.

The dimensionality reduction ability of td-idf representation is limited to the size of the word list and researchers in the IR field have proposed alternative approaches. Deerwester et al. (1990) proposed latent semantic indexing (LSI) which is the singular value decomposition (SVD) of the tf-idf matrix. LSI captures a linear subspace in which the tf-idf features have the highest variance (Blei et al., 2003; Deerwester et al., 1990). It has been shown that LSI can better represent the documents with lower dimensional representations (Deerwester et al., 1990). As an alternative to LSI, its probabilistic version is also proposed as pLSI (Hofmann, 1999) which is a generative model and instead of SVD, the probabilistic approach uses the EM algorithm to fit the model to the training corpus. Substantial performance gains are reported for using the probabilistic approach in various IR tasks such as keyword search (Hofmann, 1999).

Despite the success of the pLSI over LSI, one of its main shortcomings is its incompleteness by not providing a probabilistic model at the document level (Blei et al., 2003; Griffiths and Steyvers, 2004). With the pLSI approach, each document is projected into a lower dimensional space, but there is no generative probabilistic model for this vector. It further suffers from the linear growth of the parameters count with the number of documents which leads to over-fitting issues which are discussed in depth in Blei et al. (2003).

To address the shortcomings of the LSI and pLSI models, latent Dirichlet allocation was introduced by Blei et al. (2003). Most of the comparison studies confirm that LDA usually outperforms LSI and pLSI models in different information retrieval tasks (Blei et al., 2003; Kim et al., 2009a; Lukins et al., 2008; Wei and Croft, 2006). pLSI has also been applied to audio, in the context of audio information retrieval systems (Kim et al., 2009a) and it was again shown that LDA outperforms the pSLI model.

The next section describes the inference procedure with the LDA models and its parameter estimation.

4.4.1.2 Latent Dirichlet allocation inference

As described earlier, LDA is a probabilistic generative model for collections of discrete items. It assumes there exists a latent set of variables that govern the generation of each item within the collection. LDA is mostly studied in the context of text modelling. The generative process of LDA for generating a document in the context of text modelling is described in algorithm 2.

The graphical model representation of the LDA model is shown in figure 4.5 as

Algorithm 2 Generative process of LDA model for a document

Given:

- ξ: the parameter of Poisson distribution
- α: the parameter of Dirichlet distribution
- K: number of topics
- V: vocabulary size
- β: a $K \times V$ matrix for the probability of a word given a topic: $\beta_{ij} = \mathrm{p}(w_n = j | z_n = i)$

Choose number of words: $N \sim \mathrm{Poisson}(\xi)$
Choose topic distribution: $\theta \sim \mathrm{Dir}(\alpha)$
for each $w_n \in \mathbf{w} = \{w_1, \ldots, w_N\}$ **do**
  Choose topic of the word w_n**:** $z_n \sim \mathrm{Multinomial}(\theta)$
  Choose the word: $w_n \sim \mathrm{p}(w_n | z_n, \beta)$
end for

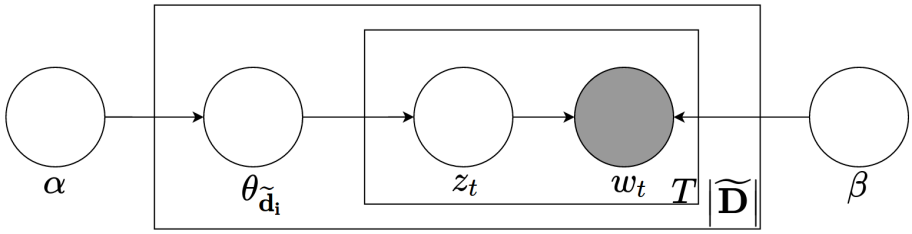


Figure 4.5: Graphical model representation of LDA

a three level hierarchical Bayesian model. The plates represent replicates, where the outer plate is for documents and the inner plate is for the words within the document. Observed variables are shaded and the rest are all latent variable (only the words are observed in the LDA model). α and β are corpus level variables, $\theta_{\tilde{\mathbf{d}}_i}$ is a document level variable and w_n and z_n are word level variables.

The posterior distribution of the latent variables given a document is:

$$\mathrm{p}(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta) = \frac{\mathrm{p}(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)}{\mathrm{p}(\mathbf{w} | \alpha, \beta)}. \tag{4.5}$$

The denominator of equation 4.5 is not computationally feasible because of the intractable integrals for the marginalising over the latent variables. Approximate inference algorithms can be used for inference, such as variational approximation (Blei et al., 2003), Markov chain Monte Carlo (MCMC) (Griffiths and Steyvers, 2004) or gradient descent based optimisation methods (Kim et al., 2012). Based on the results reported by Kim et al. (2009b) and with considerations for the faster convergence, variational approximation was chosen as the approximation algorithm for the LDA models in this thesis.

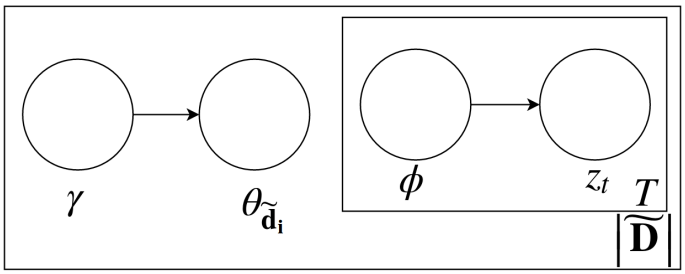


Figure 4.6: Graphical model representation of the simplified distribution for the LDA model

In the variational approximation methods, a simpler distribution is defined and its distance with the real distribution is minimised using Jensen's inequality. The denominator term is equation 4.5 is defined as:

$$\mathrm{p}(\mathbf{w}|\alpha,\beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \Big(\prod_{i=1}^{K} \theta_i^{\alpha_i - 1}\Big)\Big(\prod_{n=1}^{N}\sum_{i=1}^{K}\prod_{j=1}^{V}(\theta_i \beta_{ij})^{w_n^j}\Big) d\theta. \tag{4.6}$$

The summation over the latent topics in equation 4.6 which couples θ and β, causes the intractability of the integral. This problem can be seen in the graphical model representation of the LDA model in figure 4.5 by the edges that connect θ, z_n and w_n. A simplification can be made here by dropping those edges and the w_n node. The graphical model representation of the simplified variational distribution used for approximating the posterior in equation 4.6 is presented in figure 4.6.

The variational distribution is specified as:

$$\mathrm{q}(\theta, \mathbf{z}|\gamma,\phi) = \mathrm{q}(\theta|\gamma)\prod_{n=1}^{N}\mathrm{q}(z_n|\phi_n), \tag{4.7}$$

where γ is the Dirichlet parameter that determines θ and $(\phi_1,\ldots,\phi_N)$ are the multinomial parameters that generate the topics ($\mathbf{z}$).

The optimal values of the variational distribution are found by minimising the KL-divergence of the two distributions specified in equation 4.5 and 4.7 using an iterative procedure:

$$\gamma^*,\phi^* = \arg\min_{\gamma,\phi} \mathrm{KLD}\big(\,\mathrm{q}(\theta,\mathbf{z}|\gamma,\phi)\;||\;\mathrm{p}(\theta,\mathbf{z}|\mathbf{w},\alpha,\beta)\big). \tag{4.8}$$

An iterative process for the minimisation of the KL-divergence is defined as (Blei

et al., 2003):

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{4.9}$$

$$\phi_{ni} \propto \beta_{iw_n} \exp\left(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{k} \gamma_j)\right) \tag{4.10}$$

where Ψ is the derivative of the log gamma function. For the detailed equations, refer to the appendix A.1 and A.3 of the original LDA paper (Blei et al., 2003).

The parameters γ and ϕ are document specific, which means that they are dependent on $\mathbf{w}$. In other words, the variational distribution in equation 4.7 is implicitly dependent on the document as well, and this dependency can be written in a more explicit way as

$$\mathrm{p}(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) \approx \mathrm{q}(\theta, \mathbf{z}|\gamma^*(\mathbf{w}), \phi^*(\mathbf{w})), \tag{4.11}$$

which can be used to approximate the true posterior distribution.

The Dirichlet parameter γ can be used as a representation of the document in the topic space. This representation can be thought of a lower dimensional representation of the documents and can be used in subsequent classifiers as the input features (Blei et al., 2003).

4.4.1.3 Latent Dirichlet allocation parameter estimation

For the parameter estimation, usually the marginal log likelihood of the data is used as the objective function for optimisation:

$$\ell(\alpha, \beta) = \sum_{d=1}^{|D|} \log \mathrm{p}(\mathbf{w}_d|\alpha, \beta). \tag{4.12}$$

As previously discussed, $\mathrm{p}(\mathbf{w}|\alpha, \beta)$ is not computationally tractable and a variational method was described which provided a lower bound for the log likelihood function (Blei et al., 2003). A variational EM procedure was introduced in the original LDA paper to find the model parameters α and β. In the expectation stage for each document the values of γ_d^* and ϕ_d^* are computed. In the maximisation step, using the approximate posterior computed in the expectation step, the lower bound of the log likelihood function is maximised with respect to the model parameters α and β. These two steps are repeated until convergence in the lower bound of the log likelihood function. The full list of equations are provided in appendix A.4 of Blei et al. (2003).

4.4.1.4 Beyond text modelling

Many of the initial use cases of the LDA model were for text modelling tasks. However, LDA is a generic model and can be used for other types of data. In this section a brief overview of using LDA models for other types of data will be provided.

In image processing, LDA models can be used for object categorisation and localisation. Sivic et al. (2005) assumed that each image can be modelled with a finite mixture of object categories that are present in the image. To fit into the text modelling concepts of LDA models, images were analogous to documents, pixel patch codewords were treated as visual words and object categories were used instead of topics. The pixel patches were described by so called SIFT features (Lowe, 1999) and quantised to have a discrete representation. When training the LDA model, the number of topics was specified to match the total number of the unique objects present in the corpus. With these models, topic weight vectors for each image can be inferred and a hard assignment can be performed based on the object category with the highest probability to define the category of the objects in the image. Using the same model that was trained for the object categorisation task, the location of the objects can also be identified. This task is called object localisation. Here for each patch the probability $p(\mathbf{z}|\mathbf{w})$ is computed and patches having high probability (based on a threshold value) for a single object category (topic) are marked as having the object in the patch. Cao and Fei-Fei (2007) tried to improve the quality of the visual words by incorporating some spatial information into the visual word representation and called their model spatial latent topic model. It performed better in object categorisation and localisation task when compared to the LDA models with simpler visual words.

LDA has also been used for audio data in the context of music analysis and audio classification. Hu and Saul (2009) used LDA for harmonic analysis of music. They tried to analyse the underlying harmonic structure of musical pieces for automatic key-finding and modulation detection. In their work that studied western tonal music, the musical keys were assumed to be the latent topics, musical pieces and musical notes were analogous to the documents and words, respectively. With the LDA model, the automatic key-finding task can be performed as well as similarity ranking of classical musical pieces (Hu, 2009).

In the context of audio classification, Kim et al. (2012, 2009a,b, 2010b) used LDA for the classification of unstructured audio clips, where they assumed the existence of latent acoustic topics that govern the generation of the audio clips. Equivalent to words were the quantised MFCC feature vectors, and the audio clips were analogous to the documents in the text modelling world. They also compared the LDA model with other latent modelling techniques, in particular LSI and pLSI and concluded

the superiority of the LDA model. In a subsequent classification task using SVMs and the posterior Dirichlet parameters from the LDA model as the input features to the SVMs, they could improve the classification accuracy by 9% absolute compared to a GMM baseline.

Other similar works from the same author include the supervised variant of the LDA model (Kim et al., 2010a) where the LDA model is modified to include the labels of the data points and can be used to get the posterior class probability for each of the classes directly without having to use an other classifier. With the supervised LDA, they could further improve the audio clip classification accuracy. The same technique can be used for other classification purposes, such as the genre labelling task of broadcast media (Kim et al., 2013). More details of the acoustic LDA work will be provided in the next section.

4.4.2 Acoustic LDA

Since the LDA model deals with a collection of discrete symbols, to accommodate modelling audio within the LDA framework, the continuous acoustic feature vectors need to be represented by some discrete symbols. These discrete symbols are called acoustic words. In most of the related work (Kim, 2010; Kim et al., 2012, 2013, 2010a, 2009a,b, 2010b), vector quantisation techniques such as the Linde-Buzo-Gray algorithm (Gersho and Gray, 1992) were used for this purpose. The only parameter for this technique is the size of codebook (or number of clusters) and the EM algorithm is used to iteratively calculate the data point assignments to clusters and then update the cluster centres. This way each continuous vector is assigned to a discrete symbol and the LDA model can be trained using the procedure outlined in section 4.4.1.2. Alternative LDA models were also proposed, such as the Gaussian-LDA, where instead of the multinomial distribution, a Gaussian distribution was used to avoid the VQ step (Hu et al., 2012). This model was tested on an audio retrieval task. However, the gains from this modification were not consistent (Hu et al., 2012; Kim, 2010) and most of the successful audio based LDA techniques use VQ (Hu and Saul, 2009; Kim, 2010; Kim et al., 2012, 2013, 2010a, 2009a,b, 2010b). In this thesis the discrete LDA model is used, however, instead of VQ an alternative approach which is described below will be used.

As an alternative approach to VQ, in this work a GMM with N mixture components was used to represent each continuous frame with some discrete symbols. This approach was found to outperform the VQ based approach in a genre ID task. In this approach, the index of the Gaussian component with the highest posterior probability is used to represent each frame with a discrete symbol. Assuming $\mathbf{d}_i$ is an speech segment of length T frames: $\mathbf{d}_i = \{\mathbf{u}_1^i, ..., \mathbf{u}_t^i, ..., \mathbf{u}_T^i\}$, the discrete symbol

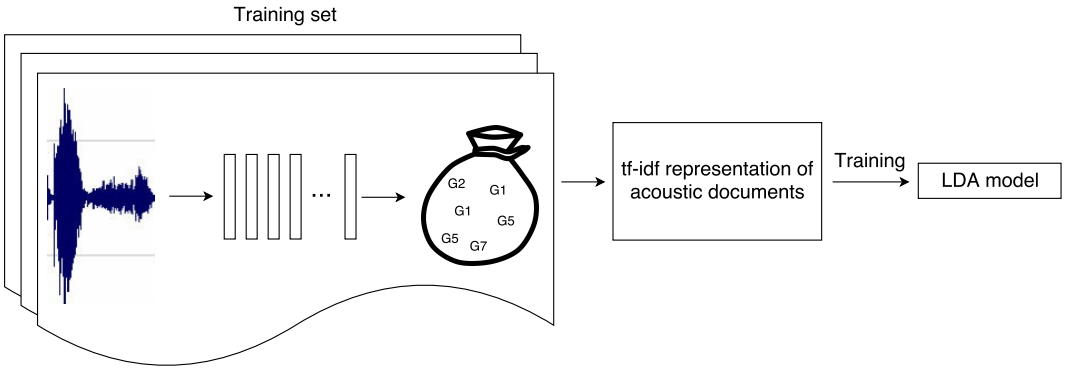


Figure 4.7: Acoustic LDA training procedure

representation of each continuous frame is defined as:

$$v_t^i = \arg\max_n P(G = n|\mathbf{u}_t^i), \;\; 1 \leq n \leq N, \; 1 \leq t \leq T \tag{4.13}$$

where G is a Gaussian component from a mixture of N components and $\mathbf{u}_t^i$ is the continuous feature vector at time t of segment i. With this new representation, $\mathbf{d}_i$ is represented by the new symbol sequence $\mathbf{s}_i = \{v_1^i, ..., v_t^i, ..., v_T^i\}$, here called acoustic document. For each acoustic word (discrete symbol) v_t^i in each acoustic document i, term frequency-inverse document frequency is computed as:

$$\begin{aligned} w_n^i &= \text{tfidf}(v_t^i = n, \mathbf{s}_i, \mathbf{S}) \\ &= \text{tf}(v_t^i = n, \mathbf{s}_i) \;\; \text{idf}(v_t = n, \mathbf{S}) \\ &= \text{tf}(v_t^i = n, \mathbf{s}_i) \;\; \log\big(\frac{|\mathbf{S}|}{\text{df}(v_t = n, \mathbf{S})}\big) \end{aligned} \tag{4.14}$$

where $\mathbf{S}$ is the set of all acoustic documents represented with acoustic words. The final representation of a segment has a fixed dimensionality which is the number of Gaussian components, N.

The assumption in the acoustic LDA with acoustic words is very similar to the bag-of-words assumptions made for text modelling. The order of frames does not matter in each audio clip and this representation can be named as bag-of-acoustic-symbols. LDA training can be performed with the raw counts of symbols in the form of bag-of-words (tf), however, consistent improvements were reported when using the tf-idf vectors (Hong et al., 2011) instead of the raw counts and in this thesis tf-idf representations are used. The acoustic LDA training procedure is depicted in figure 4.7.

The acoustic LDA models can be used to infer the posterior Dirichlet parameter γ of the acoustic documents. Figure 4.8 presents the inference process, where the variable length audio signal is mapped to a fixed length vector of size K, the dimen-

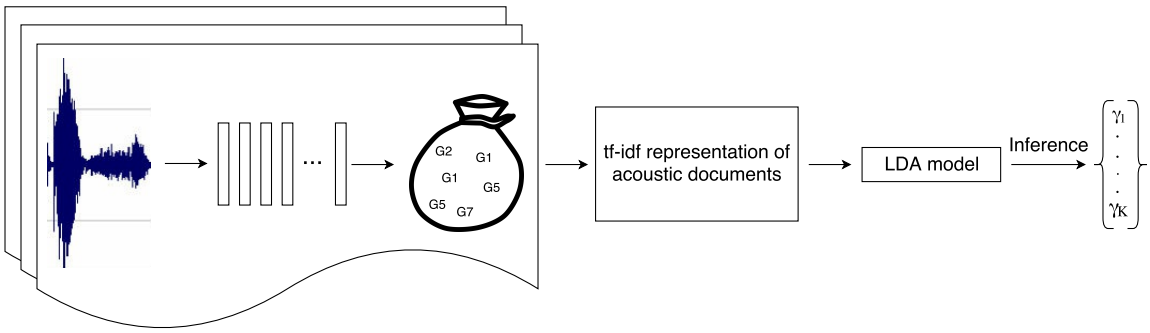


Figure 4.8: Acoustic LDA inference procedure

sion of the latent topic space. With the acoustic LDA, the latent topics are called latent domains in the remainder of this thesis.

4.5 Using latent domains for genre and show identification

LDA models can be used to compute the posterior Dirichlet parameter γ of the acoustic documents. The low-dimensional γ vectors can be considered as a mapping from the original high-dimensional acoustic space to a new low-dimensional domain (topic) space. The new representations are used as input features in the SVM classifiers for the genre and show ID tasks.

First a new baseline for the genre ID task with the dataset A is presented in the next section.

4.5.1 Genre identification with dataset A

To compare the performance of acoustic LDA with the background tracking approach on the genre ID task, the same dataset is used in the initial experiments. Later in this chapter, a much larger dataset is used for both genre ID and show ID experiments.

Using the procedure described in section 4.4.2 and the dataset A defined in section 4.3.2.1, LDA models with varying number of latent domains were trained on whole shows. These models then were used to extract the posterior Dirichlet parameter γ. These γ vectors were used as features in the SVM classifiers. The SVM training procedure was similar to the previous experiments described in section 4.3.5. Table 4.4 presents the classification accuracy for the genre ID task on dataset A. More details about the acoustic LDA training and its differences with the proposed method by Kim et al. (2009a) is provided in section 4.5.5.

The results suggest that the acoustic LDA performed well in this task, with the

Table 4.4: Genre classification accuracy (%) using whole shows on dataset A

#Domains	Accuracy
16	78.7
32	85.1
64	82.9
128	87.2
256	87.2
512	89.4
1024	87.2
2048	89.4

highest accuracy being 89.4%. The best accuracy from a single system using the background tracking features was 80.9%. To further study these models, a larger and more challenging dataset was required. The next section describes a new dataset which is around five times larger than dataset A and in the remainder of this chapter, all of the experiments are performed using with the new dataset. Also, some further experimental details such as the number of mixture components for the GMM is provided in section 5.2.3.

4.5.2 Dataset

Dataset B also consisted of TV broadcasts provided by the BBC (similar to dataset A). Dataset B was identical to the one defined and provided for the 2015 multi-genre broadcast (MGB) challenge (Bell et al., 2015a), but with a different training/testing set definition. The shows were chosen to cover the full range of broadcast show types and categorised in the same 8 genres as present in dataset A: advice, children's, comedy, competition, documentary, drama, events and news. All of the shows were broadcast by the BBC during 6 weeks in April and May 2008. There were more then 2,000 shows in the original MGB challenge data, from which 1,789 shows were selected for the experiments based on having at least 5 episodes so that at least 4 episodes can be used for training and one can be used for testing. 1,501 shows were selected for the training set and 288 shows were selected for test set, with 133 unique shows in total. Appendix A provides the list of shows.

The distribution of the shows (duration and count) across genres for the training set and test set of dataset B is shown in table 4.5. Figure 4.9 shows the distribution of the 133 shows for both training set and test set, where the horizontal axis represents shows and the vertical axis represents the number of episodes in that show. Order

Table 4.5: Amount of training and test data (hours) per genre for dataset B

Genres	Training set		Test set	
	#Shows	Duration	#Shows	Duration
Advice	189	135.3	35	24.4
Children's	301	112.7	60	25.0
Comedy	90	44.1	22	10.8
Competition	224	153.3	45	29.8
Documentary	90	57.4	29	19.3
Drama	102	69.0	21	14.6
Events	98	161.0	21	36.3
News	407	293.0	55	40.2
Total	**1501**	**1025.6**	**288**	**200.4**

of the bars are identical in both plots, e.g. the first bar of both plots represents the same show.

4.5.3 Visualising posterior Dirichlet parameter γ

Given the dataset B described in section 4.5.2 and the training procedure described in section 4.4.2, acoustic LDA models were trained with 64 latent domains. The purpose of this section is to visualise the posterior Dirichlet parameter of the LDA models and verify how it is different across different shows and genres.

LDA models in this experiment were trained on segment level, rather than whole shows as in Kim et al. (2009a). In a later experiment in section 4.5.5 the differences between segment-based and whole-show-based LDA models will be discussed. The segments were the output of an automatic segmentation system and only speech segments were used for training. For the segments of each show, a latent domain assignment based on the highest value of γ components was performed and these assignments were normalised by the length of the segment and accumulated per show and per genre. Figure 4.10 presents the most important 16 latent domains (based on duration) from an acoustic LDA model with 64 latent domains. The top 16 domains are represented first and the rest of the domains are accumulated at the top of the bars. The figure shows the distinct patterns across the genres, for example around 20% of the news genre is described by a latent domain which is plotted in red in the figure, while children's and drama genres have less than 4% of their data described by the same latent domain.

While figure 4.10 shows the differences across genres, it is of high interest to

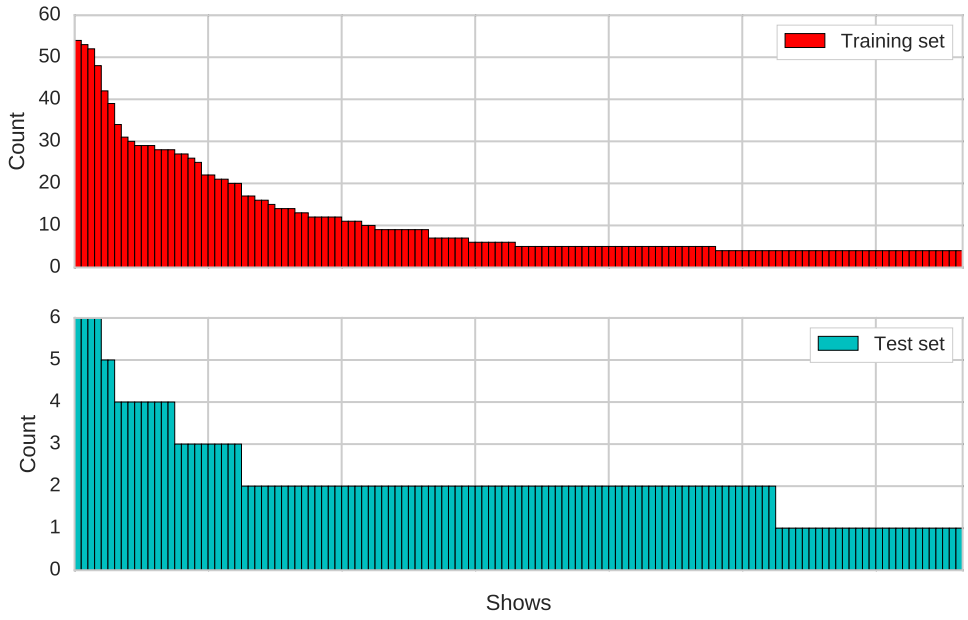


Figure 4.9: Distribution of 133 shows in training and test set of dataset B

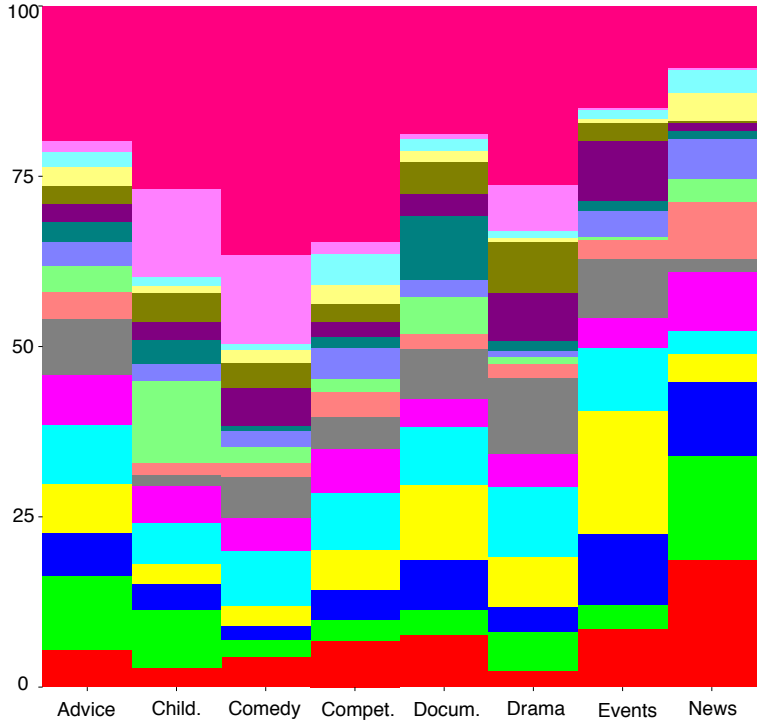


Figure 4.10: Distribution of the most important 16 LDA domains across genres

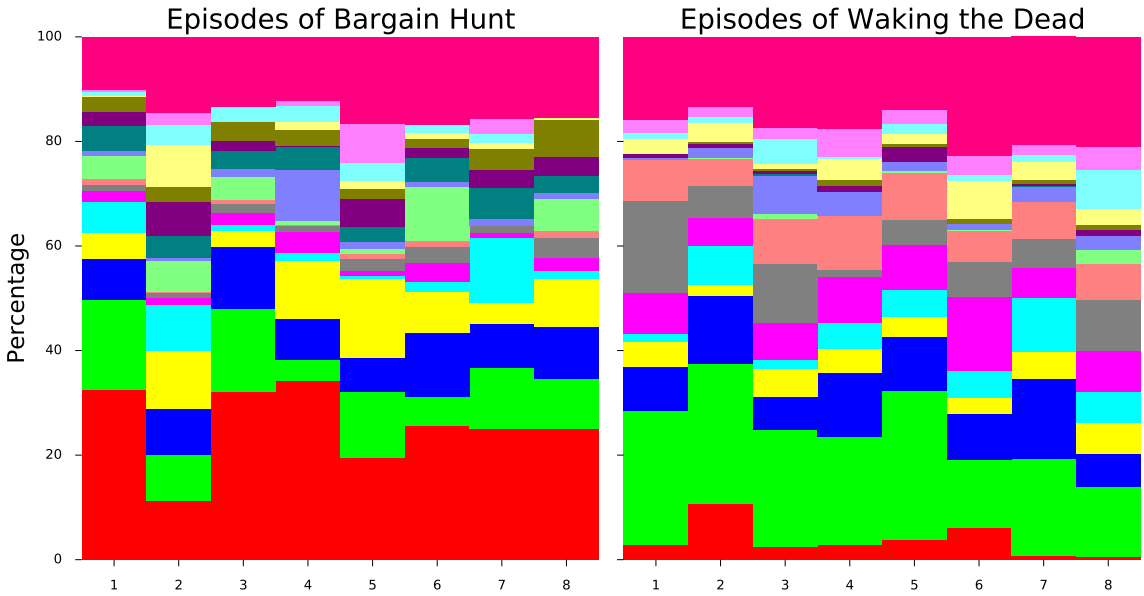


Figure 4.11: Distribution of the most important 16 LDA domains across different episodes of two shows

examine how the distribution of the latent domains are similar or different across the episodes of the same show and also other shows. Figure 4.11 presents the distribution of the most important 16 LDA domains for 8 episodes of "Bargain Hunt" (competition) and 8 episodes of "Waking the Dead" (drama). These 8 episodes for each show are represented by the columns in each of the plots and numbered from 1 to 8 in the horizontal axis.

One can observe that the distribution of the latent domains shows similarity within a genre (e.g., similarities of the red region on the lower left corner or the green area on the lower right corner). However between the two genres clear differences can be observed. One can further observe that more than 50% of each show is typically described by the top 2 or 3 LDA domains, and these differ in case of different genres but agree for the same programme within the genre. This indicates that individual shows are far more consistently described than the accumulated statistic allows to observe.

4.5.4 Baseline

As a baseline, GMM classifiers were used for the genre and show identification tasks. The setup is similar to the previous baseline experiments with dataset A. The genre ID task has 8 target classes and the show ID task has 133 target classes. 13 dimensional PLP features plus their first and second derivatives were used to train the genre-based and show-based GMMs using the EM algorithm and a mix-up procedure to reach 512 mixtures. The optimal number of mixtures for a similar task

Table 4.6: Genre/show classification accuracy (%) with GMMs on dataset B

Model	Genre ID	Show ID
GMM	61.5 (79.2)	70.1

was found to be 512 in the previous experiments. Table 4.6 shows the classification accuracy for both tasks. Since there are fewer target classes, genre ID should be an easier classification task compared to the show ID task. However, GMMs were found to perform better in classifying shows than genres (70.1% compared to 61.5%). One reason for this could be the diversity of data as discussed earlier in this chapter and the fact that PLP features are good for representing speaker specific characteristics (Reynolds, 1994) and for the show ID task the GMMs are learning speakers in re-occurring episodes. However they provide poor generalisation for the genre ID task. If show-to-genre mapping is assumed to be *a priori* knowledge, then the show ID GMMs can be used for the genre ID task. The accuracy for genre ID in such a setting would be 79.2%.

4.5.5 Whole-show and segment-based acoustic LDA experiments

Whole-show or segment-based acoustic LDA models can be trained. In this section both approaches are studied. As the name suggests, the whole-show models require training the LDA models on the whole shows. For the segment-based approach, the shows need to be segmented into speech and non-speech parts and then the LDA models can be trained on these segments. In this work speech segments were used.

4.5.5.1 Experiments

Whole shows were used to train the LDA models with varying number of latent domains. The performance of these models is to be compared with the segment-based LDA models. The segment-based approach requires summing and normalised posterior Dirichlet parameter of the segments that belong to each show and is computed as:

$$\frac{1}{\sum_{j \in segs} |j|} \sum_{i \in segs} |i| \, \gamma(\mathbf{s}_i) \tag{4.15}$$

where $|i|$ represents the length of semgment i and seg is the set of all speech segments that belong to a particular show that these statistics are being aggregated.

Table 4.7: Genre/show classification accuracy (%) using whole show and segment based acoustic LDA models on dataset B

#Domains	Whole show		Segment based	
	Genre ID	Show ID	Genre ID	Show ID
16	73.6	45.1	76.7	46.7
32	71.9	53.8	81.5	57.8
64	78.1	56.6	81.2	63.4
128	77.8	56.9	83.3	66.6
256	76.4	58.0	86.4	67.3
512	80.2	61.8	85.0	66.7
1024	77.1	65.3	85.7	63.8
2048	80.6	65.3	84.7	63.1

With the whole shows the posterior Dirichlet parameter and with segment based approach, the summed and normalised posterior Dirichlet parameter values were used to represent the shows with fixed dimensionality, which is the number of latent domains. These feature vectors were used in SVM classifiers. The SVM training procedure was similar to the previous experiments.

The classification accuracy for the genre ID and show ID tasks are presented in table 4.7. As the performance of segment level models was better than the whole show models, they were used in the rest of the experiments. Segment based models also had higher accuracy with fewer latent domains. E.g. the highest accuracy with the segment based models for genre ID was 86.4% obtained with an LDA model with 256 latent domains. However, the best performance for the whole show models was 80.6%, with 2048 latent domains. A similar pattern was found in the performance of the show ID task as well.

4.5.6 Text-based LDA

Words in transcripts of the shows have valuable information for discriminating genres or shows. In this section, classification of genres and shows is studied based solely on textual features. According to section 35 of the British Broadcasting Act of 1990, public TV broadcasting stations in the UK should provide subtitles of the TV soundtrack, mostly for helping deaf and hard-of-hearing viewers (parliament, 1990). For the MGB data, subtitles provided by the BBC were available but the quality varied considerably by genres (Bell et al., 2015a). For example subtitles of live events and news were mostly re-spoken live ASR output and had higher errors,

however for other genres which did not have the live nature, the quality was usually higher. For a detailed analysis of the subtitles' quality refer to Bell et al. (2015a).

Subtitles were used as-is, without any preprocessing (such as stemming or stop word removal) before training the classifiers for both tasks. Although subtitles can be of varied quality, their correctness is still high. In a second experiment, ASR output is used instead of the subtitles. The ASR systems used here were trained for participation in the MGB challenge. For more details about these ASR systems, refer to chapter 5 and Saz et al. (2015).

The classification task here is similar to a document classification task, where each show's transcript is a document and the classes are either genres or shows. Text based LDA models were trained and the topic posteriors were used as features in the SVM classifiers. A simpler approach could be training SVMs with tf-idf features. However here the LDA model reduces the dimensionality of the tf-idf features to the number of latent topic, which is known to work better than tf-idf only features for the document classification task (Blei et al., 2003). Table 4.8 summarises the results. LDA models trained with the subtitles performed substantially better than models trained on the ASR output. Note that the ASR models used here have around 30% WER on the official development set of the MGB challenge. The performance gap is even wider in case of the show ID task, 22.6% vs. 13.5% absolute difference. This is caused by some specific names that were present in the subtitles, but not in the ASR output and such words have considerable discriminability information.

The overall performance of text based classification with the subtitles is generally better than the audio based classification (96.2% vs. 84.4% for the genre ID task and 81.3% vs. 67.3% for the show ID task) but when considering the ASR output only, the audio based classification is better for the show ID task.

4.5.7 Using meta-data

The data used in the experiments also included some meta-data, such as the BBC broadcast channel number, the date and time of broadcast, and other unstructured information. Using some of the structured meta-data is studied next to learn how the classification accuracy can be further improved. Since these programmes were broadcast during 6 weeks in April and May 2008, using the date was not likely to be helpful which was verified in the experiments as well. Instead, the time of broadcast, splitting 24 hours into 8 chunks, and channel number, in this setup 1–4 corresponding to BBC1, BBC2, BBC3 and BBC4, were appended as one-hot-vectors to the inputs of the SVM classifiers and their effect is studied. Table 4.9 summarises the results of using the meta-data together with the acoustic LDA features. Adding these meta-data improves the accuracy of both tasks. When comparing channel and

Table 4.8: Genre/show classification accuracy (%) using text based LDA models on dataset B

#Topics	Subtitles		ASR output	
	Genre ID	Show ID	Genre ID	Show ID
16	77.4	41.3	70.1	29.2
32	81.3	50.7	71.9	34.0
64	85.4	62.1	81.6	45.8
128	89.2	68.8	87.5	55.2
256	91.0	77.1	88.2	65.6
512	91.0	76.7	87.9	63.9
1024	94.8	81.3	88.5	64.9
2048	96.2	79.9	89.9	64.9
4096	93.1	78.1	89.6	64.2

Table 4.9: Genre/show classification accuracy (%) using meta-data on dataset B

Meta-data	Genre ID	Show ID
Only Channel & Time	46.7	22.0
Baseline (acoustic 256)	86.4	67.3
+ Channel	89.6	72.8
+ Time	89.9	77.7
+ Channel & Time	92.3	82.6

time, in both tasks appending time helps more and the difference is larger in case of the show ID task (72.8% vs. 77.7%). Combining channel information and time of broadcast also helps further improve the classification accuracy in both tasks and overall with meta-data there is 5.9% and 15.3% absolute improvement in accuracies of genre ID and show ID tasks. The first row in table 4.9 shows the accuracy when only meta-data is used (without any acoustic or textual features) which shows the amount of information provided solely by the meta-data.

4.5.8 System combination

With the two systems based on acoustic and textual features, a combination of both systems can be used, assuming that they will make different classification errors and their outputs are complimentary. To combine the scores of the systems, logistic regression was used to find a linear combination of the individual system scores to

Table 4.10: Genre/show classification accuracy (%) with system fusion on dataset B

Method	Genre ID	Show ID
Baseline (acoustic 256)	86.4	67.3
Baseline (text 2048)	96.2	79.9
Acoustic & Text	97.2	85.0
Acoustic + Meta-data & Text	98.6	85.7

maximise the probability of correct classification (Brummer, 2010). System fusion using logistic regression outperforms the simpler confidence based approach that was used with the background tracking features in section 4.3.6. Since dataset B is much larger than the previous dataset, parameters of the fusion system can be learned on large independent sets without over or under-fitting issues.

Table 4.10 shows the classification accuracy with system fusion. The combination of acoustic and text based systems improved the classification accuracy for both tasks, 97.2% and 85.0% accuracy for genre ID and show ID respectively, which shows the complementarity of the individual systems. Moreover, including meta-data further improved the accuracy to 98.6% and 85.7% which is near perfect for the genre ID task.

4.5.9 Summary

In this section methods for the genre classification of broadcast media based on audio were proposed using the acoustic LDA model. Furthermore, the use of other sources of information such as subtitles and meta-data to obtain high levels of accuracy was explored. Also for the first time, a show classification task on very large datasets was studied.

The experiments were conducted on two datasets, one with 230 hours of data and the other with more than 1,200 hours of data. Both datasets included TV shows from the BBC which was broadcast in 2008. The bigger dataset was a part of the MGB 2015 challenge (Bell et al., 2015a). The smaller dataset was used to have a fair comparison of the two proposed techniques on the same dataset (background tracking vs. acoustic LDA). Since accuracy of the genre ID task using dataset A was 89%, a larger dataset was used for the remainder of the experiments. For the genre ID task there were 8 classes and for the show ID task there were 133 classes. Acoustic and textual LDA models were trained with the audio and subtitles to infer the posterior Dirichlet parameters which were then used in SVM classifiers to classify the genres and shows. On a 200h test set, a combination of both acoustic and text

based classifiers had accuracy of 97.2% and 85.0% for the genre ID and show ID tasks respectively. Use of meta-data such as time of broadcast further improved the accuracies to 98.6% and 85.7%. It should be noted that the amount of data used in these experiments was larger by at least an order of magnitude than the amount of data used in other genre ID papers and for the show ID task, this work was one of the first attempts.

4.6 Conclusion

In this chapter, genre and show identification tasks were studied. Automatic labelling of genres and shows in the media data has many applications in information retrieval and media archive systems. Furthermore, such labels can be used to improve the accuracy of ASR systems as will be shown in the next chapter.

For the genre ID task, two approaches were proposed, one with background tracking features and the other with acoustic LDA. Background tracking features were obtained from the CMLLR transformations that were trained on a specific type of background conditions. These transformations were then applied based on maximising the overall likelihood of the data, yielding the index of the environment for each frame. These features were aggregated on frame level to have a wide context and were used as features with different classifiers, such as GMMs, HMMs and SVMs. The dataset used for the evaluation consisted of over 230 hours of broadcast media from the BBC. The results showed that with the SVM classifiers, an accuracy of 80.9% can be achieved. System combination also improved the accuracy by 2.1% absolute. The second approach was based on a latent modelling technique called latent Dirichlet allocation. With the LDA models, broadcast media is assumed to be generated by a mixture of latent domains and these latent domains seem to fit the complex structure of the broadcast media well. Using a larger dataset of around 1,200 hours, accuracies of 97.2% and 85.0% were achieved for the genre ID and show ID tasks respectively. The results were further improved using meta-data.

The acoustic LDA approach outperformed the background tracking approach on the same dataset. Also unlike the background tracking approach where the transcript of the audio was required (either in form of ground truth or the output of an initial decoding), training of the acoustic LDA models was unsupervised. The posterior Dirichlet parameters computed by the LDA model can be also used for domain discovery, e.g. grouping similar data together or to identify new domains. In the next chapter the use of these features in acoustic model adaptation will be studied.

CHAPTER 5

Latent domain acoustic model adaptation

5.1 Introduction

Acoustic LDA models which were introduced in chapter 4 can be used for discovering latent structures in the speech data. With the posterior Dirichlet parameter computed with the LDA models, it was shown that they could be used for discriminating genres and shows. After successful application of the acoustic LDA models for the genre ID and show ID tasks in chapter 4, the motivation for the experimental work conducted in this chapter is to find new approaches for incorporating the information provided by the acoustic LDA models for adaptation of the acoustic models with the ultimate aim of mismatch and thus WER reduction.

It was already shown in several studies that adapting acoustic models to speakers and environments can improve accuracy of the ASR systems considerably, especially in mismatched conditions (Bell et al., 2015a; Doddipatla et al., 2014; Dupont and Cheboub, 2000; Gemello et al., 2007; Kuhn et al., 1998; Leggetter and Woodland, 1995; Li and Sim, 2010; Saon et al., 2013; Shinoda, 2011; Woodland, 2001; Yu and Deng, 2015). This was confirmed again by initial experiments conducted in chapter 3 with a diverse dataset. The research question for this chapter's work is: how to incorporate acoustic LDA information for domain adaptation of acoustic models to improve the accuracy of speech recognition. Acoustic LDA models can be used to discover latent domains and acoustic models can be adapted to these latent domains.

From the adaptation techniques presented in chapter 2, two techniques were considered in this chapter for incorporating the acoustic LDA information for domain adaptation: MAP adaptation for GMM-HMM (and DNN-GMM-HMM) systems and

subspace adaptation for hybrid DNN-HMM systems. The rationale for this decision was based on the model architecture and the amount of available data for adaptation.

With the LDA-MAP approach, first the acoustic LDA models are used to discover the latent domains present in the data, and then the segments are assigned to each of the latent domains based on the maximum posterior inferred by the LDA models. Then the base acoustic model is MAP adapted to each of the latent domains. Details of this technique plus the experimental results are provided in section 5.2.

The second approach can be considered to be a sub-space approach, where each speech segment is mapped to the latent domain space using the posterior Dirichlet vector. These vectors are then augmented the inputs to the neural network based acoustic models for latent domain bias adaptation. This approach is similar to the iVector (Saon et al., 2013) adaptation of DNNs. This approach will be studied in section 5.4.

5.2 LDA-MAP experiments with the diverse dataset

With the LDA models, posterior Dirichlet parameters can be inferred for each speech segment: $\gamma(\mathbf{d}_i)$. This K-dimensional posterior vector can be used to assign a domain to each segment based on the highest posterior value:

$$\text{Domain}(\mathbf{d}_i) = \arg\max_j \gamma_j(\mathbf{d}_i),\ j \in \{1..K\} \tag{5.1}$$

where $\gamma_j(\mathbf{d}_i)$ denotes the jth component of the posterior vector for the ith segment.

With the assignment of the speech segments to the latent domains based on the maximum latent Dirichlet posterior values, the base acoustic model which was trained with all of the available training data can be MAP adapted to each of the latent domains. This technique will yield K new models, one for each of the latent domains. These new models should better represent the latent domains compared to the base generic and un-adapted model. During the test time, first the domain assignment of each segment is determined based on the maximum latent Dirichlet posterior values and then the corresponding model can be used for decoding. This idea will be experimented with two datasets, first with the artificially diverse data set which was described in chapter 3 and then on the more realistic MGB dataset which was introduced in chapter 4.

5.2.1 Dataset

The dataset used in the experiments was the same as the one defined in chapter 3. It consisted of 66h of data from these six diverse ASR domains (10h for training and

Table 5.1: WER (%) of the baseline models on diverse dataset

Features	Model	RS	RD	TK	CT	MT	TV	Overall
PLP	ML	17.3	18.4	34.1	46.6	44.0	51.1	**36.0**
	MAP	14.6	16.8	31.8	43.5	40.4	49.6	**33.6**
PLP+BN	ML	13.0	13.3	23.5	33.5	32.2	42.0	**26.8**
	MAP	12.1	12.8	23.1	32.5	30.6	41.5	**26.2**

1h for testing from each domain):

- Radio (RD): BBC Radio4 broadcasts on February 2009 (Bell et al., 2015b)
- Television (TV): broadcasts from BBC on May 2008 (Bell et al., 2015b)
- Telephone speech (CT): from the Fisher corpus (Cieri et al., 2004)
- Meetings (MT): from AMI (Carletta et al., 2006) and ICSI (Janin et al., 2003) corpora
- Lectures (TK): from TedTalks (Ng et al., 2014)
- Read speech (RS): from the WSJCAM0 corpus (Robinson et al., 1995)

For a more detailed description and statistics of the dataset, refer to table 3.1 and 3.2 of chapter 3.

5.2.2 Baseline

To evaluate performance of the proposed approach, a set of baseline experiments were conducted. The baseline AM models included the ML models and MAP adapted models to the named domains. The ML models were trained with the whole 60-hour training set using the ML criterion. This base model was then MAP adapted to each of the 6 named domains to yield 6 new models. Each of these models were then used to decode the corresponding data during the test time. Table 5.1 summarises the results for two sets of features: PLP and PLP+BN. Note that these results were presented in chapter 3 and full details of the experiments were also provided in that chapter.

For both types of features, the MAP adapted models performed better than the ML models and this suggests that for the LDA models, MAP adaptation may further improve this baseline

5.2.3 Training LDA models

To train the acoustic LDA models, a procedure similar to the defined procedure in chapter 4 was used. There were two hyper parameters to select prior to the training. First, the number of the latent domains K needed to be decided prior to the training. Also, since the audio frames needs to be represented by some discrete symbols, the size of the codebook V also needed to be defined. Representation of the continuous frames with discrete symbols was performed using the same approach that was described in chapter 4. For this purpose, a set of experiments were conducted with different codebook sizes and number of latent domains. Codebooks of size 128 up to 8,192 were used and given a codebook, different LDA models with a varying number of domains from 4 to 64 were trained using the training data described in section 5.2.1.

For the genre ID and show ID tasks, the turnaround time for training and evaluating models with different hyper parameters was reasonably short. However, for the ASR experiments tuning the hyper parameters required much more time and involved training different LDA models followed by training, adapting and evaluating ASR models. To address this problem, a proxy value was required to evaluate the performance with different hyper parameter configurations. An initial way of evaluating how the different latent domains behaved was by measuring the distribution of the data, according to manual labels, which was included in each latent domain. Figure 5.1 presents this distribution for a codebook of size 2,048 and 8 latent domains. From this figure, it is possible to see how telephone speech was separated into two different latent domains (D1 and D3), while meeting speech was mostly assigned to a unique latent domain (D7). Other manually labelled domains, such as radio and television broadcasts were scattered across latent domains (D2, D4 or D8), indicating the presence of previously unseen domains within these types of data.

Following this, the KL divergence (Kullback and Leibler, 1951) was used as an appropriate metric to measure the consistency of the hidden topics discovered by the LDA model. This measured how the distributions of data in latent domains, as in figure 5.1, in different sets, for instance training and test data, were different to each other.

$$\mathrm{KLD}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)} \tag{5.2}$$

where P and Q are the distributions for the training and test data. To compute the divergence, the distributions were smoothed by discounting 3% of the total mass and linearly distributing it across zero counts.

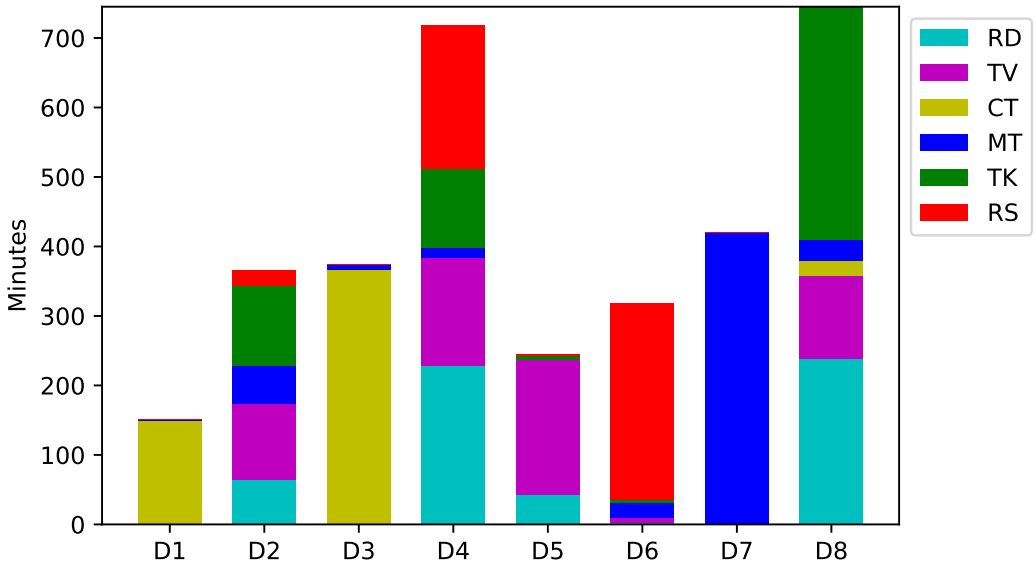


Figure 5.1: Amount of data for each discovered domain ($K = 8$) from the labelled domains using a codebook of size 2,048

Figure 5.2 shows the divergence values of different configurations. Low values of divergence indicated a more consistent set of hidden domains found by LDA modelling and, thus, were preferred over configurations with higher values. In terms of codebook size, codebooks of 2,048 and 8,192 symbols resulted in lower divergence. For the number of domains, increasing to more than 12 resulted in an increase in divergence. For the rest of the experiments, a codebook of size 2,048 was used because of the lower divergence compared to smaller codebooks and better computation time compared to the larger codebooks.

5.2.4 MAP adaptation to the latent domains with the diverse dataset

The experiments were conducted with domains of size 4, 6, 8, 10 and 12 and a codebook of size 2,048. Each MAP adapted domain specific model was used to decode the corresponding speech segments in the test set that were assigned to that domain. Figure 5.3 shows the overall WER on the test set with different number of latent domains using both types of features, PLP and PLP+BN. The lowest WER values, 30.4% for PLP features and 25.4% for PLP+BN, were achieved with 8 domains for both types of features, which was 16% and 5% relative improvement over their respective ML baselines. Comparing with MAP adaptation to human-labelled domains the relative WER reduction was 10% and 3%. The improvements in WER vanished for more than 8 hidden domains, indicating that using a larger number of domains was not beneficial for this task. One explanation for this could

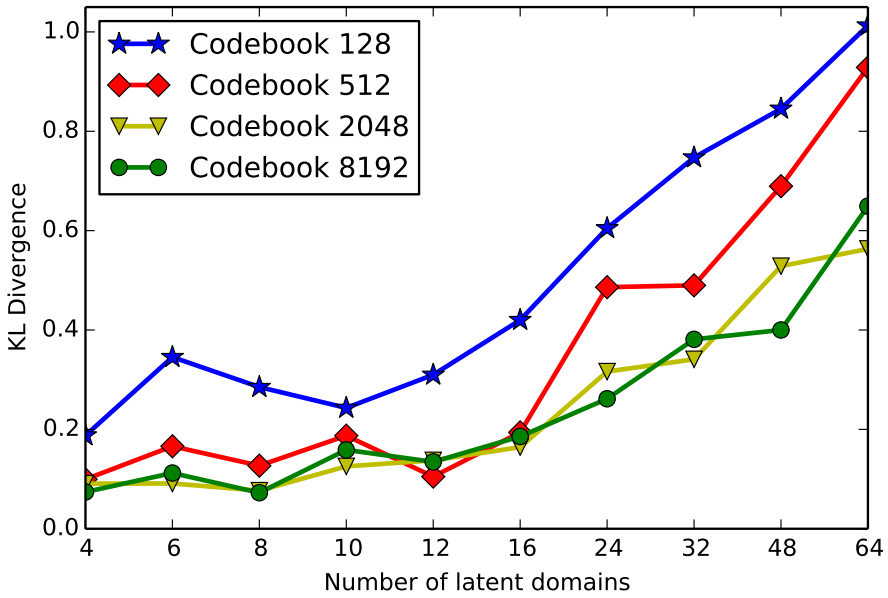


Figure 5.2: KL divergence of the training and test set latent domains

Table 5.2: WER (%) of LDA-MAP models ($K = 8$)

Features	Model	RS	RD	TK	CT	MT	TV	Overall
PLP	MAP	14.6	16.8	31.8	43.5	40.4	49.6	**33.6**
	LDA MAP	12.5	15.3	29.1	38.2	38.5	44.7	**30.4**
PLP+BN	MAP	12.1	12.8	23.1	32.5	30.6	41.5	**26.2**
	LDA-MAP	11.9	12.8	22.3	31.1	31.0	41.0	**25.4**

be the data sparsity issues; as the number of latent domains increases, the amount of available adaptation data decreases.

Table 5.2 presents the breakout of the results using 8 hidden domains across the manually labelled domains. Improvements occur across all of these domains, indicating that the LDA model can benefit all types of speech in this setup. The domains that achieved the highest gains from using LDA-MAP adaptation (with PLP features) were read speech, telephone speech and TV broadcasts, with relative WER reductions of 14%, 12%, 10% respectively compared to the MAP adaptation to the manually labelled domains. The lowest gain, 4% relative, occurred on meeting speech. Similarly, with PLP+BN features telephone speech, lectures and read speech benefited the most, with relative WER reduction of 5%, 4% and 2% respectively.

Finally, Table 5.3 shows the WER across the hidden domains for both types of features with the LDA-MAP models. An interesting observation was that domains such as TV with high WER and domains such as read speech with low WER were split across different hidden domains with WERs closer to the average WER.

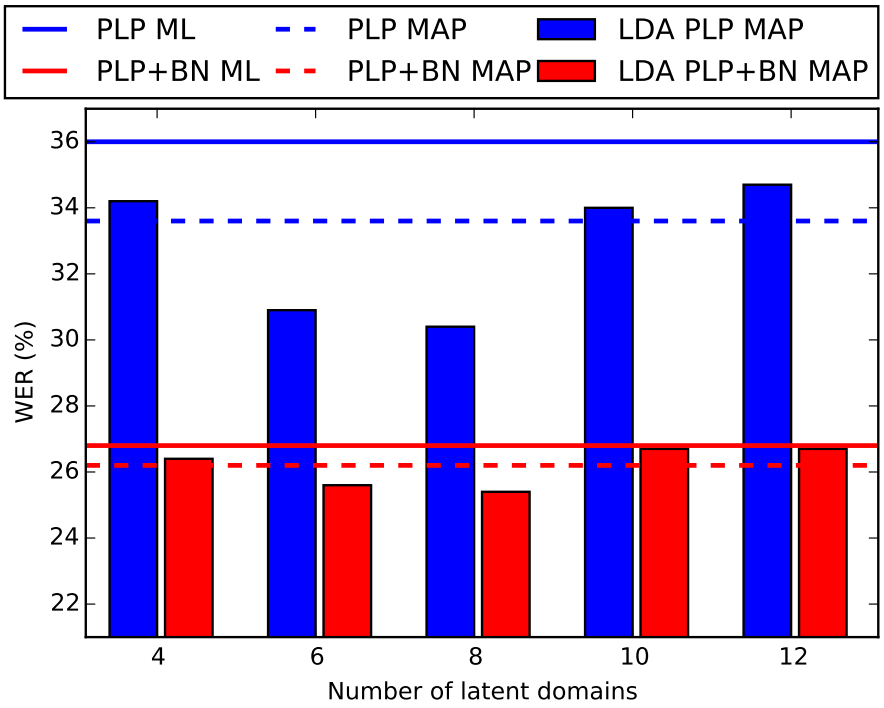


Figure 5.3: WER (%) of LDA-MAP adapted models with different number of latent domains

Table 5.3: WER (%) of LDA-MAP models ($K = 8$) across hidden domains

Features	D1	D2	D3	D4	D5	D6	D7	D8	Overall
PLP	37.3	34.9	39.7	39.2	24.6	17.1	38.7	22.9	**30.4**
PLP+BN	33.9	29.2	30.4	32.8	19.7	12.6	30.9	19.2	**25.4**

5.3 LDA-MAP experiments with the MGB dataset

In this section, the proposed LDA-MAP adaptation technique is be experimented on more realistic datasets, specifically the broadcast media data of the MGB challenge. The dataset was already introduced in section 4.5.2 of chapter 4. In this chapter a subset of the official training set was used for the training and an official development set (`dev.short`) was used as the test set.

For the training data high quality transcription was not provided with the original dataset. Instead only the subtitle text broadcast with each show plus an aligned version of the subtitles was available where the time stamps of the subtitles had been corrected in a lightly supervised manner (Bell et al., 2015a; Long et al., 2013). After this process, the new transcripts for the training shows had two potential problems: first, the subtitle text might not always match the actual spoken words and second, the time boundaries given might have errors arising from the lightly supervised alignment. To alleviate these two problems, only segments with word matching error rate (WMER) of lower than 40% were used, which yielded around 500h of data. The WMER was a by-product of the semi-supervised alignment process that measures how similar the text in the subtitle matched the output of a lightly supervised ASR system for that segment.

For the language model subtitles from shows broadcast from 1979 to March 2008, with a total of 650 million words were used to train statistical language models (Stolcke, 2002). For decoding a 50k lexicon with a highly pruned 3-gram language model was used to generate lattices and then those lattices were re-scored using a 4-gram language model. Both of the language models were trained on the 650M words of the subtitles data.

5.3.1 Baseline

As a baseline for the comparison, PLP+BN features were used in a DNN-GMM-HMM system. The models were similar to the previously trained models for the diverse dataset. Two sets of adaptation results are provided in table 5.5, one with the MAP adaptation to the genres and the other with the CMLLR adaptation to the genres. The results suggest that the MAP adaptation is better than the CMLLR adaptation (CMLLR adaptation was unsupervised). Also the first line in table 5.5 corresponds to the un-adapted model. It should also be noted that the gain from MAP adaptation yields improvements of only 1% relative, which shows the challenging nature of this data and the need for other adaptation techniques.

Table 5.4: Amount of training and test data (hours) per genre for the MGB dataset

Genres	Training set		Test set	
	#Shows	Duration	#Shows	Duration
Advice	264	193.1	4	3.0
Children's	415	168.6	8	3.0
Comedy	148	74.0	6	3.2
Competition	270	186.3	6	3.3
Documentary	285	214.2	9	6.8
Drama	145	107.9	4	2.7
Events	179	282.0	5	4.3
News	487	354.4h	5	2.0
Total	**2,193**	**1580.5**	**47**	**28.3**

Table 5.5: WER (%) of baseline BN models for the MGB dataset by genre

Adaptation	Advc.	Chld.	Cmdy.	Compt.	Doc.	Drm.	Even.	News	Overall
N/A	27.7	31.0	49.4	28.5	30.4	51.7	37.4	17.6	**33.3**
MAP	27.1	28.7	49.6	28.7	30.4	50.9	36.7	17.2	**32.9**
CMLLR	27.6	30.7	49.4	28.1	30.9	50.8	36.1	17.9	**33.2**

Table 5.6: WER (%) of LDA-MAP BN models for the MGB dataset per genre

Adaptation	Advc.	Chld.	Cmdy.	Compt.	Doc.	Drm.	Even.	News	Overall
LDA-MAP	27.5	29.6	49.5	28.1	30.8	50.8	35.1	17.0	**32.8**

5.3.2 LDA-MAP

Similar to the experiments conducted in section 5.2.4, LDA-MAP experiments were conducted using the new MGB dataset. Acoustic LDA models with 8 latent topics were trained and used to MAP-adapt the seed model (which was trained with all of the training data). During test time, the same LDA models were used to assign a latent domain to each of the segments based on the maximum domain posterior values and use the corresponding model for decoding. The results are provided in table 5.6. LDA-MAP improves the baseline with 1.5% relative.

This approach for adaptation suffers from the same problem as before: data splitting. It was shown in chapter 4 that LDA models with more latent domains performed better for the genre and show classification tasks, however, with the LDA-MAP approach their full potential cannot be exploited. Increasing the number of latent domains further splits the data and reduces the amount of available data for each latent domain. This served as the motivation to explore other adaptation techniques to incorporate the acoustic LDA information. The next section describes a subspace adaptation technique with the LDA models.

5.4 Subspace adaptation of deep neural network acoustic models to latent domains

The previous approach for adaptation using the acoustic LDA models was not able to fully exploit the amount of information provided by the LDA models with more latent domains. It was shown in chapter 4 that LDA models with more latent domains were performing better in the genre ID and show ID tasks. However, with the LDA-MAP approach, increasing the number of latent domains was further splitting the amount of adaptation data per latent domain and WER was eventually increasing with more latent domains. In this section a new approach for adapting neural network based acoustic models is described. As discussed in chapter 4, acoustic LDA can be considered as a mapping function from the high-dimensional acoustic space with variable length vectors to a fixed and lower-dimensional domain space. The dimensionality of the latent Dirichlet posteriors is equivalent to the size of the latent domains (which is set prior to training). This information can be provided to

Table 5.7: WER (%) of baseline hybrid models for the MGB dataset

Adaptation	Advc.	Chld.	Cmdy.	Compt.	Doc.	Drm.	Even.	News	Total
N/A	27.6	29.1	47.8	28.2	31.3	52.0	38.1	17.9	**33.3**
SAT	26.2	27.5	46.1	25.9	29.8	49.3	35.8	15.9	**31.4**

a neural network acoustic model to perform bias adaptation, similar to the iVector adaptation of DNNs (Saon et al., 2013), which was introduced in chapter 2.

In the previous experiments in this chapter, DNNs were used in a bottleneck setup where they were used to extract the BN features. These BN features were then concatenated with the PLP features and then were used in the GMM-HMM models. However, in the remainder of this chapter, the DNNs are used in a hybrid setup. The performance of the baseline bottleneck and hybrid setups are comparable, as will be shown in the baseline experiments in table 5.7. The reason for using the hybrid setup in this section was mostly due to the system building efforts to participate in the MGB challenge and the fact that hybrid systems often slightly outperform the BN systems (Yu and Deng, 2015).

Similar to the experiments conducted in section 5.3, the baseline PLP GMM-HMM systems were used to get the frame level alignment for the training data to train the DNNs. 13 dimensional PLP feature with four neighbouring frames on each side were spliced together to form a 117-dimensional feature vector and then projected down to a 40 dimensional feature vector using linear discriminant analysis. The input to the DNN was 440 dimensional PLP features that were ± 5 frames to the left/right of the current 40 dimensional frame. The network had 6 hidden layers of size 2048 and an output layer of size 6478 (corresponding to the CD HMM states). The network was initialised using deep belief network (Hinton et al., 2006) pre-training and then trained to optimise per frame cross entropy objective function with stochastic gradient descent. A speaker adapted DNN was also trained as a second baseline system using SAT-style training (Anastasakos et al., 1996).

SAT-style training of the DNNs requires learning speaker-specific CMLLR transformations. These linear transformations are then applied to the inputs of the DNN (before feeding to the network). With these transformations, the features are mapped to an average speaker's feature space.

Table 5.7 presents the word error rate of the test set with baseline models. Using the speaker-adapted DNN, the WER is reduced by 6% relative compared to the unadapted DNN (31.4% vs. 33.3%). The performance of the LDA adapted DNNs will be compared against the speaker adapter models in the next section.

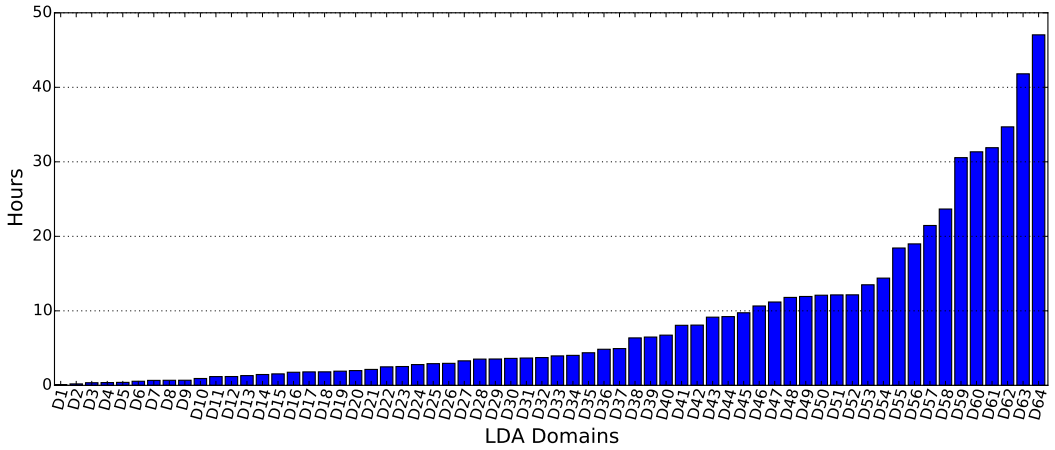


Figure 5.4: Amount of data across LDA domains

5.4.1 LDA-DNN Experiments

Using the acoustic LDA training procedure outlined in chapter 4, models with 64 and 128 latent domains were trained on the speech segments. This leveraged the considerations about the homogeneity and sparsity of the discovered domains discussed in the previous section.

Apart from selecting an appropriate size of domain, cross-agreement data filtering was performed to ensure high domain homogeneity for each acoustic document. A domain-tuple with 8192 items was established. These items come from the Cartesian product of the 64×128 domain mappings from the two corresponding LDA models. It is assumed that the two LDA models share a significant portion of the domains. If there is a high heterogeneity within an acoustic document, maximum-a-posteriori domain assignment from either or both LDA models will not be accurate, and they would appear in the rare classes in the 8192 domain-tuple items. Histogram pruning based on normalised pairs counts was performed to remove those rare items. The pruning cut-off was determined to result in a target training set size of around 500h, which was comparable to the data amount in the previous baseline experiments. Figure 5.4 shows the amount of data (in hours) for each of the 64 LDA domains.

The baseline DNN systems had an input layer of size 440. That input was expanded by augmenting the LDA inferred domain with one-hot encoding. Figure 5.5 represents the network architecture after augmenting with the LDA domain code. It was already shown in chapter 4 that representing the domain posteriors with one-hot encoding was better than raw posterior vectors for the genre and show classification tasks, and this was verified for the ASR experiments as well. The new input had the size of 504 (440+64). The new LDA-DNN was trained similarly to the baseline DNNs. This method is called latent-domain-aware training (LDaT).

As already discussed in chapter 2, augmenting the input features with the latent

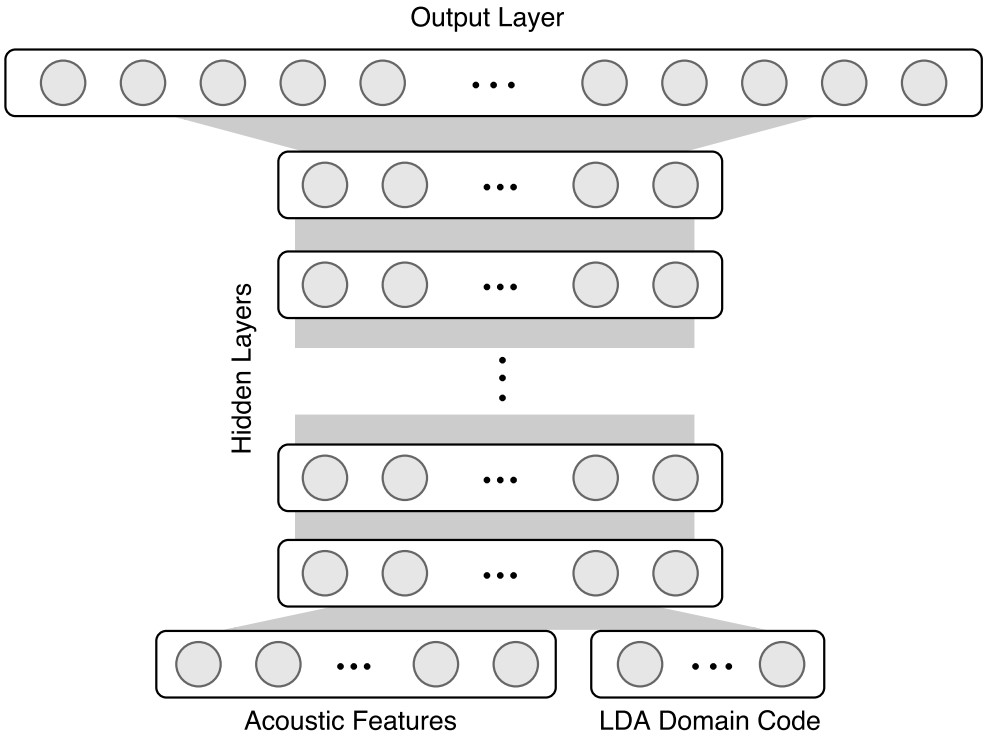


Figure 5.5: DNN architecture with LDaT, adapted from Yu and Deng (2015)

domain codes is equivalent to having a latent domain specific bias. The activation of the first layer in the LDaT architecture can be written as:

$$\begin{aligned} \mathbf{v}^1_{LDaT} &= f\Big(\begin{bmatrix} \mathbf{W}^1_{\mathbf{v}} \mathbf{W}^1_{\mathbf{d}} \end{bmatrix} \begin{bmatrix} \mathbf{v}^0 \\ \mathbf{d} \end{bmatrix} + \mathbf{b}^1_{LDaT} \Big) \\ &= f\Big(\mathbf{W}^1_v \mathbf{v}^0 + \underbrace{\mathbf{W}^1_d \mathbf{d} + \mathbf{b}^1_{LDaT}}_{\text{domain specific bias}} \Big). \end{aligned} \tag{5.3}$$

Table 5.8 presents the WER of baseline and adapted models for all of the eight genres. LDaT training reduces the WER from 33.3% to 30.6%, which is even better than speaker adapted DNN (31.4%). Combining speaker adaptation and domain adaptation (SAT+LDaT, linear input transformation for the speaker and bias adaptation for the latent domain) yields 28.9%, which is 13% relative WER reduction compared to the baseline DNN model and 8% relative improvement over the speaker adapted DNN. This also suggests that LDA inferred domains were not speaker clusters, since combining the two adaptations still improved the performance.

Because of the diverse nature of the data used, WER differs a lot across genres. Comedy and drama had the highest errors (43.8% and 45.0% respectively with LDaT+SAT models) showing the difficult nature of these genres. On the other hand, news had the lowest WER (14.3%). The WER diversity across the genres was consistent between all of the models presented in table 5.8.

Table 5.8: WER (%) of LDaT(+SAT) hybrid models for the MGB dataset

Adaptation	Advc.	Chld.	Cmdy.	Compt.	Doc.	Drm.	Even.	News	Total
SAT	26.2	27.5	46.1	25.9	29.8	49.3	35.8	15.9	**31.4**
LDaT	25.8	27.8	45.1	25.7	28.9	47.7	33.5	15.7	**30.6**
LDaT+SAT	24.2	26.5	43.8	23.6	27.3	45.0	31.6	14.3	**28.9**

5.4.2 Summary

In this section, a new method called LDaT was proposed for the first time to adapt the neural network acoustic models with the domain posteriors from the LDA model. The method employs acoustic latent Dirichlet allocation to identify acoustically distinctive data clusters. These so-called LDA domains were then encoded using one-hot encoding, and used to augment the standard input features for DNNs in training and testing. The results were presented on a diverse set of BBC TV broadcasts, with 500h of training and 28h of test data. WER reduction of 13% relative was achieved using the proposed adaptation method, compared to the baseline hybrid DNNs.

5.5 The Sheffield MGB 2015 system

In this section, a brief overview of the ASR systems developed for participating in the 2015 Multi-Genre broadcast challenge is presented. The proposed LDaT was used in parts of the Sheffield MGB system.

The MGB challenge had four tasks from which two tasks were directly related to ASR. In challenges and evaluations the ultimate aim is to have the lowest possible WER on the test sets. For this purpose, several ASR systems were trained and they were combined together in different stages, such as cross-adaptation, where one model is adapted using the hypothesis text generated by another model, or combining outputs of the individual ASR systems using voting schemes such as ROVER (Fiscus, 1997).

One of the ASR components of the Sheffield system was the proposed LDA-DNN model which was trained with 512 hours of data (TRN1). This training set was selected based on having WMER of less than 40%. Another training set was also used in the Sheffield system, which was around 700 hours of data that was selected based on having high confidence scores (TRN2). The confidence scores were computed from an initial DNN which was trained with the previous dataset. The amount of data for both datasets are presented in table 5.9.

Table 5.9: Amount of training data (hours) for the Sheffield MGB system

Genre	TRN1	TRN2
Advice	72.2	107.8
Children's	54.2	68.9
Comedy	17.3	26.2
Competition	68.5	99.0
Documentary	92.6	113.5
Drama	24.1	36.3
Events	34.2	44.1
News	153.4	203.0
Total	**512.5**	**698.8**

Table 5.10: WER (%) on the MGB dataset using the two training sets

System	Training Data	WER
Hybrid	TRN1	30.6
	TRN2	29.0
Bottleneck	TRN1	34.4
	TRN2	33.3

Bottleneck and hybrid systems were trained using both training sets. The networks were first trained with the CE criterion and then sequence trained using the boosted MMI objective function (Povey et al., 2008). Table 5.9 presents the WER with the two datasets.

The results suggest that models trained with TRN2, which was based on confidence score selection, yield lower WER in both systems.

After experimenting with different data selection methods, adaptation techniques were studied next. The CE hybrid DNNs that were trained with TRN1 were re-trained using the proposed LDaT procedure. In the interest of time, these models were only CE trained (sequence training was not performed). For the bottleneck systems, NAT was used. In the NAT setup, asynchronous CMLLR transformations were used which were initially trained on 8 different background conditions (as described in chapter 4) and were re-trained using TRN1. Again for the faster experiments turnaround time, the smaller training set was used.

Table 5.11 summarises the results for both systems. The overall best WER was achieved using the proposed LDaT method (28.9%), which was better than NAT

Table 5.11: WER (%) on the MGB dataset using domain and noise adaptation with hybrid and bottleneck systems

Genre	Hybrid		Bottleneck	
	Baseline	Adapted	Baseline	Adapted
Advice	26.9	24.2	25.2	24.6
Children's	26.8	26.5	30.8	29.2
Comedy	45.9	43.8	44.7	43.3
Competition	25.5	23.6	27.3	26.7
Documentary	28.5	27.3	28.9	27.9
Drama	49.1	45.0	42.1	40.8
Events	33.0	31.6	34.9	33.8
News	16.1	14.3	16.6	15.8
Total	**30.7**	**28.9**	**31.0**	**30.0**

training (30.0%). Note that the baseline in this experiment for the hybrid system was different from the baseline system used in the experiments presented in table 5.7.

The final Sheffield MGB system had more components, such as RNN LMs and DNN segmentation. Since all of the components were not directly related to the topic of this thesis, they are not introduced here. For a detailed description, refer to Saz et al. (2015).

The results presented in table 5.12 show the performance of different ASR components and how the system combination improved the WER by 9% relative compared to the average WER of the four individual systems.

5.6 Conclusion

In this chapter two new techniques for adaptation of acoustic models to the latent domains discovered by the acoustic latent Dirichlet allocation models were proposed. The first technique was based on MAP adaptation where the speech segments were assigned to different clusters based on maximum Dirichlet posterior values inferred by the acoustic LDA models, and then the base acoustic models were MAP adapted to the new clusters. Using the proposed method and an artificially diverse dataset where data from six conventional ASR domains were mixed together, 7% relative improvement compared to the baseline un-adapted models was achieved.

One of the shortcomings of the LDA-MAP approach was the data splitting problem, where the full potential of the LDA models with a higher number of latent domains could not be exploited. With the increasing number of latent domains,

Table 5.12: WER (%) of the different components of the Sheffield MGB 15 system on the MGB dataset

Genre	Systems				ROVER
	ASR-P1	ASR P2-1	ASR P2-2	ASR P2-3	
Advice	23.1	22.8	23.0	23.7	21.6
Children's	36.5	31.0	31.2	32.0	27.7
Comedy	45.4	42.9	42.8	45.3	40.9
Competition	25.1	24.1	24.2	25.1	22.7
Documentary	30.0	28.4	28.5	29.3	26.6
Drama	40.8	38.6	39.0	40.5	37.1
Events	36.4	33.6	33.5	34.3	31.3
News	14.1	14.2	13.8	15.0	13.2
Total	**31.2**	**29.4**	**29.4**	**30.5**	**27.5**

the data was also further splitting and as the amount of data per LDA domain was decreasing, MAP adaptation was loosing its effectiveness.

To overcome the data splitting problem and fully exploit the power of acoustic LDA models, an alternative adaptation approach was proposed, where the inputs of the DNN were augmented by the LDA domain posteriors to perform domain specific bias adaptation. With this approach a new bias for each of the latent LDA domains was learned by the model. Experiments were conducted on the multi genre broadcast challenge's dataset with around 500 hours of training data and 28 hours of test data. Significant word error rate reductions compared to the un-adapted and speaker-adapted models were achieved.

CHAPTER 6

Conclusion and future work

A summary of the contributions of this thesis and possible directions for future work are outlined in this chapter. Section 6.1 presents the main contributions of this thesis that were introduced in chapters 3, 4 and 5. Possible directions for future work are suggested in section 6.2.

6.1 Thesis summary

The main objective of this thesis was to study different approaches for reducing the mismatch between training and testing conditions, and improve performance of ASR systems. The strategies for mismatch compensation were realised in six novel contributions.

Adaptation methods are a family of mismatch reduction techniques that were introduced in chapter 2. Creating a matched training set is an alternative approach to address the mismatch problem. Data selection and augmentation techniques were introduced in chapter 3 and two contributions for selecting and augmenting training data were proposed there. Chapter 4 was about labelling complex media data with subjective labels such as genre tags. Media data grouped together with such tags usually have similar acoustic conditions and thus that information can be exploited for mismatch reduction. Two contributions were introduced which both used local expert features. In the first work, background conditions were explicitly detected and used to infer the genre tags. In the second work, influencing factors in creation of complex structure of media data were modelled by some latent variables and local features derived from the latent Dirichlet allocation modelling were used for genre labelling. Also for the first time, the show identification task was studied as well as using other sources of information for improving the classification performance. Chapter 5 investigated two new approaches for adaptation of acoustic models to the

latent domains discovered by the latent modelling technique (proposed in chapter 4). The first technique was about adapting AMs to the latent domains discovered from the data. A second technique was also proposed which used the latent domain information for an implicit bias adaptation of deep neural network AMs.

The remainder of this section provides an overview of the contributions of this thesis.

6.1.1 Chapter 3: Data selection and augmentation techniques

Given a target set, the objective of data subset selection and augmentation techniques are to select a matched subset of the training set or create an augmented training set by perturbing the existing training utterances. Two contributions were introduced in that chapter.

The first contribution was a data subset selection technique. The proposed technique used likelihood ratio based similarity scores to select data from a pool of utterances that were similar to a target test set. The ratios were computed based on the likelihood of a GMM model trained with the target data and another GMM trained with the pooled training data. A modular function based on accumulated likelihood ratios was then maximised by picking the utterances with the highest score to create the training set. The experimental work was performed on an artificially diverse dataset where data from six different domains were mixed together and the task was to find the best subset of training data for each of the six target domains. Using an automatic budget decision, 4% relative gain was achieved over a baseline model which was trained with all of the data.

The second contribution of this chapter was a data augmentation technique. The proposed technique consisted of learning the distributions of some perturbation levels from a target test set and then creating an augmented training set with the learned distribution. The training set utterances were selected from a voice search dataset and the augmented training set was evaluated on a mismatch simulated far-field test set. Using the proposed approach to create an augmented training set, the WER of the models trained with the augmented dataset was significantly lower than the baseline models that were trained with the mismatch data.

6.1.2 Chapter 4: Identification of genres and shows in media data

Media data has a complex and diverse nature and is thus a good choice for the experimental studies conducted in this thesis. The main objective of this chapter was

to study how subjective meta-data such as genre labels can be automatically identified. Genre labels usually imply similar acoustic conditions and this information was shown to be useful in improving ASR performance by reducing the mismatch in training and testing conditions. In this chapter two other contributions of the thesis were introduced. The first contribution was based on a set of local descriptive features that identify the background conditions of frames from a set of predefined conditions such as laughter, applause, music, street noise, etc. The background conditions were extracted from the output of an alignment process that fits multiple linear transformations asynchronously to the input audio data. These local features were then combined together and used in HMM and SVM classifiers for the genre identification task. Experiments were conducted on 332 BBC shows with 8 genres and an accuracy of 83% was achieved by an ensemble system.

Instead of explicitly identifying the background conditions for each frame, in the second contribution the variabilities present in the complex media data were modelled by latent variables. It was assumed that there exists a set of latent factors that govern the generation of the data and each data point can be represented by a mixture of those factors. Latent Dirichlet allocation was used to model speech segments and the latent factors were considered as latent domains. Using the latent domain posteriors as new representations for the speech segments, genre identification task can be performed. For the first time identification of show entities was also studied. To reach high levels of accuracy for both tasks, the use of other sources of information such as textual features and meta-data was studied as well and it was shown that with an ensemble system that uses acoustic, text and meta-data, 98.6% and 85.7% can be achieved for the genre ID and show ID tasks respectively. More than 1,400 hours of TV broadcasts from the BBC were used for the experiments and the genre ID task had 8 target classes and the show ID task had 133 target classes.

6.1.3 Chapter 5: Latent domain acoustic model adaptation

In this chapter two contributions were introduced that studied how latent domains can be used for adaptation of acoustic models. The first contribution was a novel technique based on acoustic LDA to discover the latent domains in highly-diverse speech data. The data set consisted of data from TV and radio shows, meetings, lectures, talks and telephony speech with a 60-hour training set and 6-hour test set. It was assumed that there exists a set of latent domains and each audio segment is a mixture of different properties of those latent domains with different weights. LDA models were used to discover the latent domains and then these domains were used to perform MAP domain adaptation. Results showed relative improvement of up to 16% over the baseline models and up to 10% over the models MAP adapted

to the manually labelled domains.

The second contribution of this chapter was to extend the use of latent domains for adapting DNN AMs. Latent-domain-aware training was proposed where the inputs to the DNN were augmented with a one-hot vector encoding of the latent domains to perform an implicit bias adaptation. The results were presented using a diverse set of BBC TV broadcasts, with 500h of training and 28h of test data. WER reduction of 13% relative was achieved using the proposed adaptation method, compared to the baseline hybrid DNNs.

6.2 Future directions

There are several future directions that can be persued based on the studies conducted in this thesis. A brief summary of these future research directions are provided in this section.

6.2.1 LDA based data selection

Data selection techniques were introduced in chapter 3 and a new data selection technique based on likelihood ratio similarity was proposed. Acoustic LDA modelling was also proposed in chapter 4 where the LDA domain posteriors were used to cluster acoustically distinctive data points in the acoustic space. This metric can be used as a measure of similarity for the data selection problem as well. The initial idea is to map all of the target test set and training utterances to the latent-domain space and based on proximity of the training data points in that space to the target utterances, select a subset of data for training acoustic models. Since this selection yields a matched dataset, performance of the acoustic model trained with this subset should be better especially when a using diverse datasets.

6.2.2 Improving acoustic embedding with LDA posteriors

Latent domain posteriors were introduced in chapter 4 and were successfully used in the genre and show identification task. They were also used in chapter 5 for acoustic model adaptation. These posterior vectors are a form of acoustic embeddings. They can be considered as an acoustic variation of Word2Vec (Mikolov et al., 2013), which is a fixed dimensional vector representation of words in a continuous space where this mapping has some semantic information. Other similar embedding techniques for audio have been proposed as well, such as Audio Word2Vec (Chung et al., 2016). These mappings have many applications, for example they can be used in spoken term detection systems. As a future work, acoustic LDA based embeddings can

be studied for representation of the spoken words in a continuous space and their applications in spoken term detection tasks.

6.2.3 Using background-tracking feature for acoustic LDA training

The current form of acoustic LDA training relies on using short-term spectral features to derive some discrete representation of the frames. Different features, such as the background tracking features that were introduced in chapter 4, can be used for training acoustic LDA models. Since the background tracking features represent the background condition and it was already shown that higher level features derived from these local features can be used for tasks such as genre identification, it is of high interest to train the LDA models with these features. Latent modelling of these features might allow to discover new latent structures that are hidden in the background conditions.

6.2.4 Deep neural network acoustic model adaptation with embeddings

Chapter 5 presented a subspace adaptation approach for deep neural network based acoustic models using acoustic LDA features. The proposed technique augmented the DNN inputs for an implicit bias adaptation. Latent Dirichlet posteriors were converted to one-hot vector encodings of the latent domains. An interesting study could be to explore how improving the representations can further reduce the WER. For example by improving the current representations to include more semantic information, WER could be further reduced. Furthermore, studying other adaptation techniques of DNN based acoustic models with the embeddings can be another research direction.

6.2.5 Alternative adaptation approaches for the latent domains

In chapter 5, MAP adaptation of acoustic models to the latent domains discovered by the LDA model was presented. One issue with that approach was the data sparsity, where the number of the latent domains could not be increased, as increasing them further split the data. With decreasing amounts of data, MAP adaptation was not very helpful. However, as was shown in chapter 4, more latent domains were beneficial for both genre ID and show ID tasks. Alternative adaptation techniques that include parameter sharing between the models for the latent domains might

alleviate the data sparsity issues. For example in the context of DNN acoustic models, parts of the initial DNN can be further trained (fine-tuned) with the latent domain's data, or even some latent domain specific layers can be added to the network. With these changes, the number of parameters to be re-estimated from the adaptation data decreases and thus more latent domains can be used without having to worry about the amount of data.

Bibliography

Abrash, V., Franco, H., Sankar, A., and Cohen, M. (1995). Connectionist speaker normalization and adaptation. In *Proceedings of EuroSpeech*, Madrid, Spain.

Allen, J. B. and Berkley, D. A. (1979). Image method for efficiently simulation room-small acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950.

Anastasakos, T., McDonough, J., Schwartz, R., and Makhoul, J. (1996). A compact model for speaker-adaptive training. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, Pennsylvania, USA.

Arnold, A., Nallapati, R., and Cohen, W. W. (2007). A comparative study of methods for transductive transfer learning. In *Proceedings of International Conference on Data Mining Workshops (ICDMW)*, Omaha, Nebraska, USA.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China.

Bell, P., Gales, M. J. F., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., McParland, A., Renals, S., Saz, O., Webster, M., and Woodland, P. (2015a). The MGB Challenge: Evaluating multi-genre broadcast media recognition. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA.

Bell, P., Gales, M. J. F., Hain, T., Kilgour, J., Lanchantin, P., Liu, X., McParland, A., Renals, S., Saz, O., Wester, M., and Woodland, P. C. (2015b). The MGB challenge: Evaluating multi-genre broadcast media recognition. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA.

Benesty, J., Sondhi, M. M., and Huang, Y. (2007). *Springer handbook of speech processing*. Springer Science & Business Media.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Brummer, N. (2010). FoCal toolkit for evaluation, fusion and calibration of statistical pattern recognisers.

Cao, L. and Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, Vasilis Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2006). The AMI meeting corpus: A pre-announcement. In *Proceedings of International Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Bethesda, Maryland, USA.

Castan, D. and Akbacak, M. (2013). Indexing multimedia documents with acoustic concept recognition lattices. In *Proceedings of Interspeech*, Lyon, France.

Central Intelligence Agency (2016). The World Fact Book: `https://www.cia.gov/library/publications/the-world-factbook/fields/2015.html`.

Challenge, A. M. G. (2009–2010). Multimedia grand challenge.

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China.

Chowdhury, G. (2010). *Introduction to Modern Information Retrieval, Third Edition*. Facet Publishing, 3rd edition.

Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y., and Lee, L.-S. (2016). Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. In *Proceedings of Interspeech*, San Francisco, California, USA.

Cieri, C., Miller, D., and Walker, K. (2004). The Fisher corpus: A resource for the next generations of speech-to-text. In *Proceedings of Language Resources Evaluation Conference (LREC)*, Lisbon, Portugal.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cox, S. (1995). Predictive speaker adaptation in speech recognition. *Computer Speech & Language*, 9(1):1–17.

Cui, X., Goel, V., and Kingsbury, B. (2014). Data augmentation for deep neural network acoustic modeling. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy.

Daume III, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

Deng, L., Acero, A., Plumpe, M., and Huang, X. (2000). Large-vocabulary speech recognition under adverse acoustic environments. In *Proceedings of Interspeech*, Beijing, China.

Deng, L. and Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):1060–1089.

Doddipatla, R., Hasan, M., and Hain, T. (2014). Speaker dependent bottleneck layer training for speaker adaptation in automatic speech recognition. In *Proceedings of Interspeech*, Singapore.

Dupont, S. and Cheboub, L. (2000). Fast speaker adaptation of artificial neural networks for automatic speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Istanbul, Turkey.

Ekenel, H. K. and Semela, T. (2013). Multimodal genre classification of TV programs and YouTube videos. *Multimedia tools and applications*, 63(2):547–567.

EU Quaero Programme (2011). Quaero programme website: `http://www.quaero.org`.

Facebook (2016). Facebook Q3 Earning Report: `https://investor.fb.com/investor-news/press-release-details/2016/Facebook-Reports-Third-Quarter-2016-Results/default.aspx`.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Santa Barbara, California, USA.

Gales, M. J. F. (2000). Cluster adaptive training of hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 8(4):417–428.

Gales, M. J. F. and Young, S. (2008). The application of hidden markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304.

Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

Gemello, R., Mana, F., Scanzio, S., Laface, P., and De Mori, R. (2007). Linear hidden transformations for adaptation of hybrid ANN/HMM models. *Speech Communication*, 49(10):827–835.

Gersho, A. and Gray, R. M. (1992). *Vector quantization and signal compression*. Springer Science & Business Media, Berlin, Germany.

Gibson, M. and Hain, T. (2006). Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition. In *Proceedings of Interspeech*, Pittsburgh, Pennsylvania, USA.

Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech Communication*, 16(3):261–291.

Gouvea, E. and Davel, M. H. (2011). Kullback-Leibler divergence-based ASR training data selection. In *Proceedings of Interspeech*, Florence, Italy.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*, Pittsburgh, Pennsylvania, USA.

Grézl, F., Karafiát, M., Kontár, S., and Cernocky, J. (2007). Probabilistic and bottleneck features for LVCSR of meetings. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, Hawaii, USA.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228–5235.

Hamidi Ghalehjegh, S. (2016). *New Paradigms for Modeling Acoustic Variation in Speech Processing*. PhD thesis, McGill University.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752.

Hermansky, H., Burget, L., Cohen, J., Dupoux, E., Feldman, N., Godfrey, J., Khudanpur, S., Maciejewski, M., Mallidi, S. H., Menon, A., et al. (2015). Towards machines that know when they do not know: Summary of work done at 2014 frederick jelinek memorial workshop. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., r. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York City, New York, USA.

Hong, L., Dan, O., and Davison, B. D. (2011). Predicting popular messages in twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web*, Hyderabad, India.

Hu, D. and Saul, L. K. (2009). A probabilistic topic model for unsupervised learning of musical key-profiles. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan.

Hu, D. J. (2009). Latent dirichlet allocation for text, images, and music. *University of California, San Diego*, 26:2013.

Hu, P., Liu, W., Jiang, W., and Yang, Z. (2012). Latent topic model based on gaussian-lda for audio retrieval. In *Proceedings of Chinese Conference Pattern Recognition (CCPR)*, Beijing, China.

Huang, X., Acero, A., Hon, H.-W., and Foreword By-Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

Itoh, N., Sainath, T. N., Jiang, D. N., Zhou, J., and Ramabhadran, B. (2012). N-best entropy based data selection for acoustic modeling. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan.

Jaakkola, T. S., Haussler, D., et al. (1999). Exploiting generative models in discriminative classifiers. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, Denver, Colorado, United States.

Jaitly, N. and Hinton, G. E. (2013). Vocal tract length perturbation (vtlp) improves speech recognition. In *Proceedings ICML Workshop on Deep Learning for Audio, Speech and Language*.

Jaitly, N., Nguyen, P., Senior, A., and Vanhoucke, V. (2012). Application of pretrained deep neural networks to large vocabulary speech recognition. In *Proceedings of Interspeech*, Portland, Oregon, USA.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The ICSI meeting corpus. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong.

Jurafsky, D. and Martin, J. (2000). *Speech & language processing*. Pearson Education India.

Kalinli, O., Seltzer, M. L., Droppo, J., and Acero, A. (2010). Noise adaptive training for robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):1889–1901.

Kapralova, O., Alex, J., Weinstein, E., Moreno, P., and Siohan, O. (2014). A big data approach to acoustic model training corpus selection. In *Proceedings of Interspeech*, Singapore.

Karafiát, M., Grézl, F., Burget, L., Szőke, I., and Černocký, J. (2015). Three ways to adapt a CTS recognizer to unseen reverberated speech in BUT system for the ASpIRE challenge. In *Proceedings of Interspeech*, Dresden, Germany.

Kim, S. (2010). *Contextual modeling of audio signals toward information retrieval*. PhD thesis, University of Southern California.

Kim, S., Georgiou, P., and Narayanan, S. (2012). Latent acoustic topic models for unstructured audio classification. *APSIPA Transactions on Signal and Information Processing*, 1:e6.

Kim, S., Georgiou, P., and Narayanan, S. (2013). On-line genre classification of tv programs using audio content. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, British Columbia, Canada.

Kim, S., Georgiou, P. G., Narayanan, S., and Sundaram, S. (2010a). Supervised acoustic topic model for unstructured audio information retrieval. In *Proceedings of Annual Summit and Conference of Asia Pacific Signal and Information Processing Association (ASC APSIPA)*, Biopolis, Singapore.

Kim, S., Narayanan, S., and Sundaram, S. (2009a). Acoustic topic model for audio information retrieval. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, USA.

Kim, S., Sundaram, S., Georgiou, P., and Narayanan, S. (2009b). Audio scene understanding using topic models. In *Proceeding of Neural Information Processing System (NIPS) Workshop*, Vancouver, British Columbia, Canada.

Kim, S., Sundaram, S., Georgiou, P., and Narayanan, S. (2010b). Acoustic stopwords for unstructured audio information retrieval. In *Proceedings of IEEE European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark.

Knerr, S., Personnaz, L., and Dreyfus, G. (1990). Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing*, pages 41–50. Springer.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proceedings of Interspeech*, Dresden, Germany.

Krause, A. and Golovin, D. (2014). Submodular function maximization. *Tractability: Practical Approaches to Hard Problems*.

Kuhn, R., Nguyen, P., Junqua, J.-C., Goldwasser, L., Niedzielski, N., Fincke, S., Field, K. L., and Contolini, M. (1998). Eigenvoices for speaker adaptation. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of Mathematical Statistics*, 2(1):79–86.

Kumar, A., Metze, F., Wang, W., and Kam, M. (2013). Formalizing expert knowledge for developing accurate speech recognizers. In *Proceedings of Interspeech*, Lyon, France.

Lanchantin, P., Bell, P. J., Gales, M. J., Hain, T., Liu, X., Long, Y., Quinnell, J., Renals, S., Saz, O., Seigel, M. S., et al. (2013). Automatic transcription of multi-genre media archives. In *Proceedings of the Workshop on Corpora in the Digital Humanities (CEUR)*.

Larson, M., Anguera, X., Reuter, T., Jones, G., Ionescu, B., Schedl, M., Piatrik, T., Hauff, C., and Soleymani, M. (2013). Indexing multimedia documents with acoustic concept recognition lattices. In *Proceedings of MediaEval Multimedia Benchmark Workshop*, Barcelona, Spain.

Lee, K. and Ellis, D. P. (2010). Audio-based semantic concept classification for consumer video. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1406–1416.

Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185.

Li, B. and Sim, K. C. (2010). Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems. In *Proceedings of Interspeech*, Makuhari, Japan.

Lin, H. and Bilmes, J. (2009). How to select a good training-data subset for transcription: Submodular active selection for sequences. In *Proceedings of Interspeech*, Brighton, UK.

Lippmann, R., Martin, E., and Paul, D. (1987). Multi-style training for robust isolated-word speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA.

Long, Y., Gales, M. J. F., Lanchantin, P., Liu, X., Seigel, M. S., and Woodland, P. C. (2013). Improving lightly supervised training for broadcast transcriptions. In *Proceedings of Interspeech*, Lyon, France.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Kerkyra, Greece.

Lukins, S. K., Kraft, N. A., and Etzkorn, L. H. (2008). Source code retrieval for bug localization using latent dirichlet allocation. In *Proceeding of Working Conference on Reverse Engineering*, Antwerp, Belgium.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of Interspeech*, Makuhari, Japan.

Mikolov, T., Kombrink, S., Deoras, A., Burget, L., and Cernocky, J. (2011). Rnnlm-recurrent neural network language modeling toolkit. In *Proceedings of IEEE Workshop of Automatic Speech Recognition Understand (ASRU)*, Honolulu, Hawaii, USA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, Nevada, USA.

Montagnuolo, M. and Messina, A. (2007). TV genre classification using multimodal information and multilayer perceptrons. In *AI*IA 2007: Artificial Intelligence and Human-Oriented Computing*, pages 730–741. Springer.

Montagnuolo, M. and Messina, A. (2009). Parallel neural networks for multimodal video genre classification. *Multimedia Tools and Applications*, 41(1):125–159.

Moore, R. K. (2003). A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proceedings of EuroSpeech*, Geneva, Switzerland.

Nagroski, A., Boves, L., and Steeneken, H. (2003). In search of optimal data selection for training of automatic speech recognition systems. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas, Virgin Islands, USA.

Nemhauser, G., Wolsey, L., and Fisher, M. (1978). An analysis of approximations for maximising submodular set functions i. *Mathematical Programming*, 14(1):265–294.

Ng, R. W. N., Doulaty, M., Doddipatla, R., Saz, O., Hasan, M., Hain, T., Aziz, W., Shaf, K., and Specia, L. (2014). The USFD spoken language translation system for IWSLT 2014. In *Proceedings of IEEE International Workshop on Spoken Language Translation (IWSLT)*, South Lake Tahoe, Nevada, USA.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

parliament, B. (1990). Broadcasting Act 1990: http://www.legislation.gov.uk/id?title=Broadcasting+Act+1990.

Parviainen, O. (2015). SoundTouch audio processing library. http://www.surina.net/soundtouch.

Povey, D. (2009). A tutorial-style introduction to subspace Gaussian mixture models for speech recognition. Technical report, Microsoft.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., Goel, N. K., Karafiát, M., Rastrow, A., Rose, R. C., et al. (2010). Subspace Gaussian mixture models for speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Dallas, Texas, USA.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Rastrow, A., et al. (2011). The subspace gaussian mixture model: A structured model for speech recognition. *Computer Speech & Language*, 25(2):404–439.

Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., and Visweswariah, K. (2008). Boosted MMI for model and feature-space discriminative training. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA.

Povey, D. and Woodland, P. C. (2002). Minimum phone error and i-smoothing for improved discriminative training. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, Florida, USA.

Povey, D., Woodland, P. C., and Gales, M. J. F. (2003). Discriminative map for acoustic model adaptation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong.

Ragni, A., Knill, K. M., Rath, S. P., and Gales, M. J. F. (2014). Data augmentation for low resource languages. In *Proceedings of Interspeech*, Singapore.

Ravanelli, M. and Omologo, M. (2014). On the selection of the impulse responses for distant-speech recognition based on contaminated speech training. In *Proceedings of Interspeech*, Singapore.

Renals, S., Morgan, N., Bourlard, H., Cohen, M., and Franco, H. (1994). Connectionist probability estimators in hmm speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(1):161–174.

Reynolds, D. A. (1994). Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):639–643.

Riccardi, G. and Hakkani-Tür, D. (2003). Active and unsupervised learning for automatic speech recognition. In *Proceedings of Interspeech*, Geneva, Switzerland.

Robinson, T., Fransen, J., Pye, D., Foote, J., and Renals, S. (1995). WSJCAM0: A british english speech corpus for large vocabulary continuous speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit, Michigan, USA.

Sageder, G., Zaharieva, M., and Breiteneder, C. (2016). Group feature selection for audio-based video genre classification. In *MultiMedia Modeling*, pages 29–41. Springer.

Salton, G. and McGill, M. J. (1986). *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Saon, G., Soltau, H., Nahamoo, D., and Picheny, M. (2013). Speaker adaptation of neural network acoustic models using i-Vectors. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic.

Saz, O., Doulaty, M., Deena, S., Milner, R., Ng, R. W., Hasan, M., Liu, Y., and Hain, T. (2015). The 2015 sheffield system for transcription of multi-genre broadcast media. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, Arizona, USA.

Saz, O., Doulaty, M., and Hain, T. (2014). Background-tracking acoustic features for genre identification of broadcast shows. In *Proceedings of IEEE workshop on Spoken Language Technology (SLT)*, South Lake Tahoe, Nevada, USA.

Saz, O. and Hain, T. (2013). Asynchronous factorisation of speaker and background with feature transforms in speech recognition. In *Proceedings of Interspeech - International Speech Communication Association*, Lyon, France.

Seide, F., Li, G., Chen, X., and Yu, D. (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Waikoloa, Hawaii, USA.

Seltzer, M. L. and Acero, A. (2011). Separating speaker and environmental variability using factored transforms. In *Proceedings of Interspeech*, Florence Italy.

Seltzer, M. L., Yu, D., and Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, British Columbia, Canada.

Settles, B. (2010). Active learning literature survey. Technical report, University of Wisconsin, Madison, Wisconsin, USA.

Seung, H. S., Opper, M., and Sompolinsky, H. (1992). Query by committee. In *Proceedings of Conference on Learning Theory Workshop (COLT)*, Pittsburgh, PA, USA.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Shinoda, K. (2011). Speaker adaptation techniques for automatic speech recognition. In *Proceedings of Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Xi'an, China.

Silva, C. and Ribeiro, B. (2009). *Inductive inference for large scale text classification: kernel approaches and techniques*, volume 255. Springer.

Siohan, O. (2014). Training data selection based on context-dependent state matching. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy.

Siohan, O. and Bacchiani, M. (2013). ivector-based acoustic data selection. In *Proceedings of Interspeech*, Lyon, France.

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T. (2005). Discovering objects and their location in images. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Beijing, China.

Stadermann, J. and Rigoll, G. (2005). Two-stage speaker adaptation of hybrid tied-posterior acoustic models. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, Pennsylvania, USA.

Stolcke, A. (2002). SRILM-an extensible language modeling toolkit. In *Proceedings of Interspeech*, Denver, Colorado, USA.

Tur, G., Schapire, R., and Hakkani-Tür, D. (2003). Active learning for spoken language understanding. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong.

Verhelst, W. and Roelands, M. (1993). An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Minneapolis, Minnesota, USA.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Wei, K., Liu, Y., Kirchhoff, K., Bartels, C., and Bilmes, J. (2014a). Submodular subset selection for large-scale speech training data. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy.

Wei, K., Liu, Y., Kirchhoff, K., and Bilmes, J. (2013). Using document summarisation techniques for speech data subset selection. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, Atlanta, Georgia, USA.

Wei, K., Liu, Y., Kirchhoff, K., and Bilmes, J. (2014b). Unsupervised submodular subset selection for speech data. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy.

Wei, X. and Croft, W. B. (2006). Lda-based document models for ad-hoc retrieval. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA.

Wessel, F. and Ney, H. (2005). Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(1):23–31.

Woodland, P. C. (2001). Speaker adaptation for continuous density HMMs: A review. In *Proceedings of International Speech Communication Association (ISCA) Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France.

Wu, T.-F., Lin, C.-J., and Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005.

Wu, Y., Zhang, R., and Rudnicky, A. (2007). Data selection for speech recognition. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Kyoto, Japan.

Young, S., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006). *The HTK book version 3.4*. Cambridge University Engineering Department.

YouTube (2016). YouTube Statistics: `https://www.youtube.com/yt/press/en-GB/statistics.html`.

Yu, D. and Deng, L. (2015). *Automatic Speech Recognition: A Deep Learning Approach*. Springer.

Yu, K. (2006). *Adaptive training for large vocabulary continuous speech recognition*. PhD thesis, University of Cambridge.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of International Conference on Machine Learning (ICML)*, Banff, Alberta, Canada.

Zavaliagkos, G., Siu, M.-H., Colthurst, T., and Jayadev, B. (1998). Using untranscribed training data to improve performance. In *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia.

Zhang, R. and Rudnicky, A. (2006). A new data selection approach for semi-supervised acoustic modeling. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France.

APPENDIX A

List of shows used in chapter 4

List of the shows from the MGB dataset that were used in the experiments conducted in chapter 4 are provided in this appendix.

Table A.1: List of the BBC shows used in the genre ID and show ID experiments

Show name	Genre	Channel
A Perfect Spy	Drama	BBC4
A Question of Sport	Competition	BBC1
Arthur	Children's	BBC2
Bargain Hunt	Competition	BBC1
BBC London News	News	BBC1
BBC News	News	BBC1
BBC News at One	News	BBC1
BBC News at Six	News	BBC1
BBC News at Ten	News	BBC1
BBC Ten O'Clock News	News	BBC1
BBC Young Musician of The Year 2008	Events	BBC4
Beat The Boss	Competition	BBC1
Bill Oddie S Wild Side	Documentary	BBC1
Blue Peter	Children's	BBC1
Boogie Beebies	Children's	BBC2
Breakfast	News	BBC1
Cash In The Attic	Advice	BBC1
Casualty	Drama	BBC1
Chinese School	Documentary	BBC4
Chucklevision	Comedy	BBC2
Coast	Documentary	BBC2

Comedy Map of Britain	Documentary	BBC2
Countryfile	Advice	BBC1
Dad's Army	Comedy	BBC2
Dan Cruickshank's Adventures	Documentary	BBC2
Doctors	Drama	BBC1
Doctor Who	Drama	BBC3
Doctor Who Confidential	Documentary	BBC3
Doctor Who The Daleks	Documentary	BBC4
Dog Borstal	Advice	BBC3
Eastenders	Drama	BBC3
Eggheads	Competition	BBC2
Escape To The Country	Advice	BBC2
Fimbles	Children's	BBC2
Flog It	Advice	BBC2
Gardeners World	Advice	BBC2
Gavin and Stacey	Comedy	BBC3
Gcse Bitesize	Children's	BBC2
Glamour Girls	Documentary	BBC3
Golf Us Masters	Events	BBC2
Graham Norton Uncut	Comedy	BBC2
Grange Hill	Children's	BBC1
Great British Menu	Competition	BBC2
Have I Got a Bit More News For You	Comedy	BBC2
Hedz	Comedy	BBC2
Hider In The House	Competition	BBC2
Holby Blue	Drama	BBC1
Holby City	Drama	BBC1
Homes Under The Hammer	Advice	BBC1
I'd Do Anything	Competition	BBC1
I'd Do Anything Results	Competition	BBC1
Ideal	Comedy	BBC2
In Search of Medieval Britain	Documentary	BBC4
Inside Sport	News	BBC2
In The Night Garden	Children's	BBC2
Jackanory Junior	Children's	BBC2
Johnny S New Kingdom	Documentary	BBC1
Key Stage Three Bitesize	Children's	BBC2
Last Man Standing	Competition	BBC2

Later Live With Jools Holland	Events	BBC2
Later With Jools Holland	Events	BBC2
Life In Cold Blood	Documentary	BBC1
Little Britain	Comedy	BBC3
Love Soup	Drama	BBC1
Mama Mirabelle S Home Movies	Children's	BBC2
Match of The Day	Events	BBC1
Meet The Immigrants	Documentary	BBC1
Missing Live	Advice	BBC1
Natural World	Documentary	BBC2
Newsnight	News	BBC2
Newsnight Review	News	BBC2
Newsround	Children's	BBC1
Open Gardens	Advice	BBC2
Panorama	News	BBC1
Points of View	Advice	BBC1
Premiership Rugby	Events	BBC2
Proms On Four	Events	BBC4
Pulling	Comedy	BBC3
Raven The Secret Temple	Competition	BBC2
Ready Steady Cook	Competition	BBC2
Roar	Children's	BBC1
Schools Hands Up	Children's	BBC2
Schools Look and Read	Children's	BBC2
Schools Primary Geography	Children's	BBC2
Schools Primary History	Children's	BBC2
Schools Science Clips	Children's	BBC2
Schools Something Special	Children's	BBC2
Schools The Way Things Work	Children's	BBC2
Schools Watch	Children's	BBC2
Seaside Rescue	Documentary	BBC1
See Hear	Advice	BBC1
Small Talk Diaries	Children's	BBC1
Snooker Extra	Events	BBC2
Snooker World Championship	Events	BBC1
Snooker World Championship Highlights	Events	BBC2
Something Special	Children's	BBC2
Songs of Praise	Documentary	BBC1

Sound	Advice	BBC2
Space Pirates	Children's	BBC1
Sportsround	Children's	BBC2
Stake Out	Children's	BBC1
Street Doctor	Documentary	BBC1
Stupid	Comedy	BBC2
The Apprentice	Competition	BBC1
The Apprentice You Re Fired	Competition	BBC2
The Book Quiz	Competition	BBC4
The Daily Politics	News	BBC2
The Kids Are All Right	Competition	BBC1
The One Show	News	BBC1
The Politics Show	News	BBC1
The Slammer	Competition	BBC1
The Surgery	Advice	BBC2
The Twenties In Colour The Wonderful	Documentary	BBC1
The Wall	Comedy	BBC3
The Weakest Link	Competition	BBC1
Through The Keyhole	Competition	BBC2
Tikkabilla	Children's	BBC2
To Buy Or Not To Buy	Advice	BBC1
Tommy Zoom	Children's	BBC2
Top Gear	Advice	BBC3
Tracy Beaker	Children's	BBC2
Traffic Cops	Documentary	BBC1
Transatlantic Sessions	Events	BBC4
Trapped	Competition	BBC2
Two Pints of Lager And	Comedy	BBC3
University Challenge The Professionals	Competition	BBC2
Waking The Dead	Drama	BBC1
Watchdog	Advice	BBC1
Weatherview	News	BBC1
Working Lunch	News	BBC2
World News Today	News	BBC4
World Swimming Championships	Events	BBC2
Young Dracula	Drama	BBC1