Adaptation techniques to improve ASR performance on accented speakers

Udhyakumar Nallasamy

CMU-LTI-16-012

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:
Florian Metze Chair,
Tanja Schultz Co-Chair,
Alan W. Black,
Monika Woszczyna, M*Modal Inc.

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Language and Information Technologies

Copyright © 2016 Udhyakumar Nallasamy

Abstract

Speech interfaces are becoming pervasive among the common public with the prevalence of smart phones and cloud-based computing. This pushes Automatic Speech Recognition (ASR) systems to handle a wide range of factors including different channels, noise conditions and speakers with varying accents. State-of-the-art large vocabulary ASRs perform poorly when presented with accented speech, that is either unseen or under-represented in the training data. This thesis focuses on problems faced by accented speakers with various ASR configurations and proposes several adaptation techniques to address them.

The influence of accent is examined in three different ASR setups including accent dependent, accent independent and speaker dependent models. In the case of accent dependent models, a source ASR trained on resource-rich accent(s) is adapted to a target accent using a limited amount of training data. Semi-continuous decision tree based adaptation is proposed to efficiently model contextual phonetic changes between source and target accents and its performance is compared against traditional techniques. Active and semi-supervised learning techniques that can exploit the contemporary availability of extensive, albeit unlabeled data resources are also investigated.

In accent independent models, a novel robustness criterion is introduced to evaluate the impact of accent in various ASR front-ends including MFCC and Bottle-neck features. Accent questions are introduced in addition to phonetic ones, to measure the ratio of accent models in the ASR contextual decision tree. Accent aware training is also proposed in the context of deep bottle-neck front-end to derive canonical features robust to accent variations.

Finally, problems faced by accented speakers in speaker-dependent ASR models is addressed. Several neighbour selection and adaptation algorithms are proposed to improve the performance of accented speakers using only a few minutes of data from the target speaker. Extensive analysis is performed to measure the influence of accent in the neighbours selected for adaptation. Neighbour selection using textual features and language model adaptation using neighbours data is also investigated.

Contents

Part I

Introduction

Chapter 1

Introduction

Speech recognition research has made great strides in the recent years and current state-of-the-art ASRs scale to large systems with millions of parameters trained on thousands of hours of audio data. For many tasks such as broadcast news transcription, the Word-Error Rate (WER) has been reduced to less than 10% for a handful of languages [MGA+06, SSK+09, GAAD+05, HFT+09]. This has led to increased adoption of speech recognition technology in desktop, mobile and web platforms for applications such as dictation, voice search [BBS+08], natural language queries, etc. However, these systems suffer high vulnerability towards variations due to accents that are unseen or under-represented in the training data [SMB11, NMS12b]. The WER has been to shown to nearly double for mismatched train/test accent pairs in a number of languages such as English [HW97, NMS12b], Arabic [SMB11, NMS12b], Mandarin Chinese [HCC01] or Dutch/Flemish accents [Com01]. Moreover, accent-independent ASRs trained on multiple, pooled accents achieve 20% higher WER than accent-specific models [SMB11, BMJ12, Com01].

1.1 Accent variations

Human speech in any language, exhibits a class of well-formed, stylized speaking patterns that are common across members that belong to the same clique. These groups can be characterized by geographical confines, socio-economic class, ethnicity or for second-language speakers, by the speakers' native language. These spoken language patterns can vary in their vocabulary, syntax, semantics, morphology and pronunciation. These set of variations are termed as 'Dialects' of a language. Accent

is a subset of dialect variations that is concerned mainly with the pronunciation, although pronunciation can influence other choices such as vocabulary and word-frequency [Wel82, MHu]. Although non-native pronunciations are influenced by the speakers' native language, we do not focus on explicitly modeling L2 variations in this thesis. Pronunciation variations between different accents can be further characterized by

- Phoneme inventory - Different accents can rely on different set of phonemes
- Phonetic realization - The same phoneme can be realized differently in each accent
- Phonotactic constraints - The distribution of phonemes can be different
- Lexical distribution - The choice of words can vary between the accents

Speakers of a specific accent express both unique and common speech patterns with members of other accents of the same language. These accent variations can be represented by contextual phonological rules of the form

$$\mathcal{L} - m + \mathcal{R} \rightarrow s \tag{1.1}$$

where $\mathcal{L}$ represents the left-context, $\mathcal{R}$ the right-context, m the phone to be transformed and s the realized phone. Such rules result in changes to canonical pronunciation including addition, deletion and substitutions of sounds units. [Uni] used such rules in a hierarchical way to convert an accent-independent pronunciation lexicon to a variety of English accents spanning across US, UK, Australia and New Zealand.

1.2 Related work

The two main approaches for accent adaptation include lexical modeling and acoustic adaptation. Lexical modeling accounts for the pronunciation changes between accents by adding accent-specific pronunciation variants to the ASR dictionary. It is accomplished by either rules created by linguistic experts [BK03, Tom00] or automatically learned using data-driven algorithms [LG, HW97, NGM+11]. [Tom00] used both knowledge-based and data-driven methods to generate pronunciation

variants for improving native American ASR on Japanese English. In [Hum97], the transformation rules from source accent (British English) to target accent (American English) pronunciations are automatically learnt using a phone decoder and decision trees. It has also been shown that adding pronunciation variants to the dictionary has a point of diminishing returns, as over-generated pronunciations can lead to ambiguity in the decoder and degrade its performance [RBF+99].

The phonetic variations between accents can also be addressed by acoustic adaptation techniques like MLLR/MAP [LW95, GL94a] estimation. They are generally model accent variations by linear transforms or Bayesian statistics [VLG10, DDNK97, SK11, Tom00]. However, both MLLR and MAP adaptation are generic adaptation techniques that are not designed to account for the contextual phonological variations presented by an accent. [CJ06] showed that MLLR has some limitations in modeling accented speech, particularly if the target accent has some new phones which are not present in the source. The polyphone decision tree in ASR, which is used to cluster context-dependent phones based on phonetic questions is also a candidate for accent adaptation. It decides which contexts are important to be modeled and which ones are merged, thus directly influencing the pronunciation. [WS03] used Polyphone Decision Tree Specialization (PDTS) to model the pronunciation changes between native and non-native accents. One of the limitations of PDTS is that it creates too few contextual states at the leaf of the original decision tree with the available adaptation data, thus having less influence in overall adaptation.

All these supervised adaptation techniques require manually labeled (accent labels) target accent data for adaptation. The adaptation can benefit from additional data, however it is costly to collect and transcribe sufficient amount of speech for various accents. Active and semi-supervised training for the goal of accent adaptation has received less attention in the speech community. [NSK11] uses self-training to adapt Modern Standard Arabic (MSA) ASR to Levantine with limited success. Self-training assumes the unlabeled data is homogeneous, which is not the case for multi-accented datasets. [SMB11] used an accent classifier to select appropriate data for MSA to Levantine adaptation on GALE BC corpus. It requires sufficiently long utterances ($\approx 20s$) for both accents to reliably train a discriminative phonotactic classifier to choose the data.

Finally, real-world datasets have multiple accents and the ASR models should be able to handle such accents without compromising on the performance. The main approaches used in these conditions are multi-style training, which simply pools all the available data to train accent-independent model. Borrowing from Multilingual speech recognition, [CMN09, KMN12] have used tagged decision trees to train

accent-adaptive models. In a similar problem of speaker and language adaptive training in speech synthesis, [ZBB+12] used acoustic factorization to simultaneously train speaker and language adaptive models.

1.3 Thesis contributions

- **Accent dependent modeling**. Semi-continuous, polyphone decision trees are introduced to adapt a source accent ASR to a target accent using relatively limited adaptation data. The performance is evaluated on Arabic and English accents produces a relative improvement of 4-13.6% WER compared to existing adaptation techniques [NMS12b]. The adaptation technique is extended with active and semi-supervised learning algorithms using unlabelled data. Relevance based biased sampling is proposed to augment traditional data selection to choose an appropriate subset from a large speech collection with multiple accents. The selected data is used to retrain the ASR for additional improvements on the target accent. These techniques provide an additional improvement of 8.5%-20.6% relative WER over supervised adaptation in Arabic and English respectively [NMS12c, NMS12a].

- **Accent independent modeling**. An evaluation framework is proposed to test various front-ends based on their robustness to accent variations. The performance of MFCC and Bottle-neck features are analyzed on a multi-accent Arabic and English datasets and showed that this framework can aid in choosing accent robust features. Accent aware training is introduced to efficiently using accent labels in the training data to derive bottle-neck features and compared against accent agnostic training [NMS11, NGM+11].

- **Speaker dependent modeling**. The problems faced by accent speakers in the context of speaker dependent ASR are analyzed and several neighbour selection and adaptation algorithms are proposed. Both maximum likelihood and discriminative versions are investigated. It is shown that using 5 mins of target speaker audio can be used to select neighbours to obtain an improvement of 10.1% relative WER over the MAP adapted model. Finally, neighour selection using text based features and language model adaptation using neighbours data have been investigated. The text based features are shown to augment acoustic based neighbour selection for additional improvements [NFW+13].

1.4 Thesis organization

In section II Accent dependent modeling, target accent adaptation using semi-continuous polyphone decision tree adaptation is discussed. Several existing adaptation techniques have been reviewed and a semi-continuous decision tree adaptation is proposed. Additional gains from unlabelled data is also investigated using active and semi-supervised learning. The section III on accent robust modeling deals with evaluation of accent robustness in the ASR among MFCC and MLP Bottle-neck front-ends. Accent robustness measure is introduced using accent-dependent questions in the ASR decision tree. Finally in section IV, the influence of accent in the performance of speaker-dependent models is discussed in detail.

Part II

Accent dependent modeling

Chapter 2

Target accent adaptation

In this chapter, we investigate techniques that can adapt an ASR model trained on one accent (source) to a different accent (target) with limited amount of adaptation data. With the wide-spread adoption of speech interfaces in mobile and web applications, modern day ASRs are expected to handle speech input from a range of speakers with different accents. The trivial solution is to build a balanced training database with representative accents in the target community. It is quite expensive to collect and annotate a variety of accents for any language, even for the few major ones. While a one-size-fits-all ASR that can recognize seen/unseen accents equally well may be the holy-grail, the practical solution is to develop accent-specific systems, at least for a handful of major accents in the desired language. Since, it is difficult to collect large amount of accented data to train an accent-dependent ASR, the source models are adapted using a relatively small amount of target data. The initial ASR is trained on available training data and adapted to required target accents using the target adaptation data. It is imperative that the adaptation technique should be flexible to efficiently use the small amount of target data to improve the performance on the target accent. The target accent can either be a new unseen accent or it can be a regional accent, under-represented in the training data. In both cases, the source ASR models are adapted to match the target adaptation data better.

2.1 Related work

Two main approaches to target accent adaptation include lexical modeling and acoustic model adaptation. In lexical modeling, the ASR pronunciation dictionary is modified to reflect the changes in the target accent. Both rule-based and data-driven techniques have been used to generate additional pronunciation variants to better match the decoder dictionary to the target accent.

The Unisyn project [Uni] uses a hierarchy of knowledge-based phonological rules to specialize an accent-independent English dictionary to a variety of accents spanning different geographical locations including, US, UK, Australia and New Zealand. [BK03] used these rules on the British English BEEP dictionary to create accent-specific ASRs and showed improved performance on cross-accent scenarios. [Tom00] used both rule-based and data-driven rules to recognize Japanese-accented English. [HW97, GRK04, NGM+11] also used data-driven rules to model different accents in cross-accent adaptation. The main component of these data-driven methods is a phone-loop recognizer which decodes the target adaptation data to recover the ground truth pronunciations. These pronunciations are then aligned with an existing pronunciation dictionary and phonological rules are derived. During decoding, the learnt rules are applied to the existing dictionary to create accent-dependent pronunciation variants.

In the case of acoustic model adaptation, [VLG10] used MAP adaptation and compared the performance on multi-accent and cross-accent scenarios. [Liv99] employed different methods including model interpolation to improve the performance of a native American English recognizer on non-native accents. [SK11] created a stack of transformations to factorize speaker and accent adaptive training and reported improvements on the EMMIE English accent setup. Finally, [Hum97] compared both the lexical and acoustic model adaptation techniques and showed they can obtain complementary gains on two accented datasets. The polyphone decision tree (PDT), in addition to the GMMs can also be a candidate for accent adaptation. [SW00, Stü08] adapted the PDT on the target language/accent and showed improved performance over MAP adaptation.

2.2 PDT adaptation

A polyphone decision tree is used to cluster context-dependent states to enable robust parameter estimation based on the available training data. Phonetic binary

questions such as voiced yes/no, unvoiced yes/no, vowel yes/no, consonant yes/no, etc. are used in a greedy, entropy-minimization algorithm to build the PDT based on the occupational statistics of all the contexts in the training data. These statistics are accumulated by forced-aligning the training data with context-independent (CI) models. The leaves of the PDT serve as final observation density functions in the HMM models. The PDT has great influence in the overall observation modeling as it determines how different contexts are clustered. Since the acoustic variations of different accents in a language are usually characterized by contextual phonological rules, it makes PDT an attractive candidate for accent adaptation.

PDT adaptation has been shown to improve the ASR adaptation for new languages [SW00] and non-native speech [WS03]. It involves extending the PDT trained on the source data with relatively small amount of adaptation data. The extension is achieved by force-aligning the adaptation data with the existing PDT and its context-dependent (CD) models. The occupational statistics are obtained in the same way as before based on the contexts in the adaptation dataset. The PDT training is restarted using these statistics, from the leaves of the original tree. The parameters of the resulting states are initialized from their parent nodes and updated on the adaptation set using a MAP training. The major limitation of this framework is that, each of the newly created states has a set of state-specific parameters (means, variance and mixture-weights) that need to be estimated from the relatively small adaptation dataset. This limits the number of new contexts created to avoid over-fitting.

For example, let us assume we have 3 hours of adaptation data and our source accent model has 3,000 states with 32 Gaussians per state. We enforce a minimum count of 250 frames (with 10ms frame-shift) per Gaussian. The approximate number of additional states that can be created from the adaptation dataset is 135 or only 4.5% of the total states in the source model. Such small number of states have quite less influence on the overall acoustic model. One solution is to significantly reduce the number of Gaussians in the new states, but this will lead to under-specified density functions. In the next section, we review the semi-continuous models with factored parameters to address this issue.

2.3 Semi-continuous PDT adaptation

We propose a semi-continuous PDT adaptation to address the problem of data-sparsity and robust estimation for PDT adaptation. A semi-continuous model extends

a traditional fully-continuous system to incorporate additional states with GMM mixture weights which are tied to the original codebooks. This factorization allows more granulated modeling while estimating less parameters per state, thus efficiently utilizing the limited adaptation data. We briefly review the semi-continuous models and present the use of it in accent adaptation.

In a traditional semi-continuous system, the PDT leaves have a common pool of shared Gaussians (codebooks) trained with data from all the context-dependent states. Each leaf has a unique set of mixture weights (distribution) over these codebooks trained with data specific to the state. The fully-continuous models on the other hand, have state-dependent codebooks (Gaussians) and distributions (mixture weights) for all the leaves in the PDT. Although traditional semi-continuous models are competitive in low-resource scenarios, they lose to fully-continuous models with increasing data. The multi-codebook variant of semi-continuous models can be thought of as an intermediary between semi-continuous and fully-continuous models. They follow a two-step decision tree construction process: in the first level, the scenario is the same as for fully continuous models, with clustered leaves of PDT having individual codebooks and associated mixture-weights. The PDT is then further extended with additional splitting into the second level, where all the states that branched out from the same first level node, share the same codebooks, but have individual mixture-weights. For more details on the difference between fully-continuous, traditional and multi-codebook semi-continuous models, refer to [RBGP12]. These models are being widely adopted in ASR having performed better than its counterparts, in both low-resource [RBGP12] and large-scale systems [SSK+09].

One of the interesting features of multi-codebook semi-continuous models is that the state-specific mixture weights are only a fraction of size of the shared Gaussian parameters, i.e means and variances even in the diagonal case. This allows us to have more states in the second-level tree with robustly estimated parameters, thus more suitable for PDT adaptation on a small dataset of target accent. The codebooks can also be reliably estimated by pooling data from all the shared states. The accent adaptation using this setup is carried out as follows:

- We start with a fully-continuous system and its associated PDT trained on the source accent.
- The CD models are used to accumulate occupation statistics for contexts present in the adaptation data.

- The second-level PDT is trained using these statistics, creating new states with shared codebooks and individual mixture-weights.
- The mixture-weights of the second-level leaves or adapted CD models are then initialized with parameters from their root nodes (fully-continuous leaves).
- Both the codebooks and mixture-weights are re-estimated on the adaptation dataset using MAP training.

Recalling the example from previous section, if we decide to train semi-continuous PDT on a 3 hour adaptation set and a minimum of 124 frames per state (31 free mixture-weight parameters per state), we will end up with ≈8,000 states, 2.6 times the total number of states in the source ASR (3,000)! The MAP update equations for the adapted parameters are shown below.

Table 2.1: *Multi-codebook semi-continuous model estimates.*

Estimate	Equation
Likelihood	$p(o_t\|j) = \sum_{m=1}^{N_k(j)} c_{jm}\mathcal{N}(o_t\ \mu_{k(j),m}, \Sigma\mu_{k(j),m})$
Mixture-weight	$c_{jm}^{MAP} = \frac{\gamma_{jm}+\tau M\hat{c}_{jm}}{\sum_{m=1}^{M}\gamma_{jm}+\tau}$
Mean	$\mu_{km}^{MAP} = \frac{\theta_{km}(\mathcal{O})+\tau\hat{\mu}_{km}}{\gamma_{km}+\tau}$
Variance	$\sigma_{km}^{MAP^2} = \frac{\theta_{km}(\mathcal{O}^2)+\tau(\hat{\mu}_{km}^2+\hat{\sigma}_{km}^2)}{\gamma_{km}+\tau} - \mu_{km}^{MAP^2}$

$\gamma, \theta(\mathcal{O})$ and $\theta(\mathcal{O}^2)$ refer to zeroth, first and second-order statistics respectively. The subscripts j refers to states, k to codebooks and m to Gaussian-level statistics. $k(j)$ refers to state-to-codebook index. τ is the MAP smoothing factor.

2.4 CMU setup - speech corpus, language model and lexicon

We evaluate the adaptation techniques on three different setups on Arabic and English datasets. The training data for Arabic experiments come from Broadcast News (BN) and Broadcast Conversations (BC) from LDC GALE corpus. The BN part consists of read speech from news anchors from various Arabic news channels and the BC corpus consists of conversational speech. Both parts mainly includes

Modern Standard Arabic (MSA) but also various other dialects. LDC provided dialect judgements (Mostly Levantine, No Levantine & None) produced by transcribers on a small subset of the GALE BC dataset automatically chosen by IBM's Levantine dialect ID system. We use 3 hours of 'No Levantine' and 'Mostly Levantine' segments as source and target test sets and allocate the remaining 30 hours of 'Mostly Levantine' segments as adaptation set. The 'No Levantine' test set can have MSA or any other dialect apart from Levantine. The Arabic Language Model (LM) is trained from various text and transcription resources made available as part of GALE. It is a 4-gram model with 692M n-grams, interpolated from 11 different LMs trained on individual datasets [MHJ+10]. The total vocabulary is 737K words. The pronunciation dictionary is a simple grapheme-based dictionary without any short vowels (unvowelized). The Arabic phoneset consists of 36 phones and 3 special phones for silence, noise and other non-speech events. The LM perplexity, OOV rate and number of hours for different datasets are shown in Table 2.2.

We use the Wall Street Journal (WSJ) corpus for our experiments on accented English. The source accent is assumed to be US English and the baseline models are trained on 66 hours of WSJ1 (SI-200) part of the corpus. We assign UK English as our target accent and extract 3 hours from the British version of the WSJ corpus (WSJCAM0) corpus as our adaptation set. We use the most challenging configuration in the WSJ test setup with 20K non-verbalized, open vocabulary task and default bigram LM with 1.4M n-grams. WSJ Nov 93 Eval set is chosen as source accent test set and WSJCAM0 SI_ET_1 as target accent test set. Both WSJ and WSJCAM0 were recorded with the same set of prompts, so there is no vocabulary mismatch between the source and target test sets. We use US English CMU dictionary (v0.7a) without stress markers for all our English ASR experiments. The dictionary contains 39 phones and a noise marker.

2.5 Baseline systems

For Arabic, we trained an unvowelized or graphemic system without explicit models for the short vowels, which are not written. The acoustic models use a standard MFCC front-end with mean and variance normalization. To incorporate dynamic features, we concatenate 15 adjacent MFCC frames (± 7) and project the 195 dimensional features into a 42-dimensional space using a Linear Discriminant Analysis (LDA) transform. After LDA, we apply a globally pooled ML-trained STC transform. The speaker-independent (SI), CD models are trained using an

Table 2.2: *Database Statistics.*

Dataset	Accent	#Hours	Ppl	%OOV
Arabic				
Train-BN-SRC	Mostly MSA	1092.13	-	-
Train-BC-SRC	Mostly MSA	202.4	-	-
Adapt-TGT	Levantine	29.7	-	-
Test-SRC	Non-Levantine	3.02	1011.57	4.5
Test-TGT	Levantine	3.08	1872.77	4.9
English				
Train-SRC	US	66.3	-	-
Adapt-TGT	UK	3.0	-	-
Test-SRC	US	1.1	221.55	2.8
Test-TGT	UK	2.5	180.09	1.3

entropy-based polyphone decision tree clustering process with context questions of maximum width ± 2, resulting in quinphones. The speaker adaptive (SA) system makes use of VTLN and SA training using feature-space MLLR (fMLLR). During decoding, speaker labels are obtained after a clustering step. The SI hypothesis is then used to calculate the VTLN, fMLLR and MLLR parameters for SA decoding. The resulting BN system consists of 6K states 844K Gaussians and and the BC system has 3,000 states and 141K Gaussians. We perform our initial experiments with the smaller BC system and evaluate the adaptation techniques finally on the bigger BN system.

The BC SA system produced a WER of 17.8% on the GALE standard test set Dev07. The performance of the baseline SI and SA on source and target accents are shown in Table 6.1. We note that the big difference in WER between these test sets and the Dev07 is due to relatively clean Broadcast News (BN) segments in Dev07, while our new test sets are based on BC segments. Similar WERs are reported by others on this task [SMB11]. The absolute difference of 7.8-9.0% WER between the two test sets shows the mismatch of baseline acoustic models to the target accent. For further analysis, we also include the WER of a system trained just on the adaptation set. The higher error rate of this TGT ASR indicates that 30 hours is not sufficient to build a Levantine ASR that can outperform the baseline for this task. As expected, the degradation in WER is not uniform across the test sets. The TGT ASR performed 11.1% absolute worse on unmatched source accent while only 0.4% absolute worse on matched target accent compared to the baseline.

The English ASR essentially follows the same framework as Arabic ASR with minor changes. It uses 11 adjacent MFCC frames (± 5) for training LDA and triphone models (± 1 contexts) instead of quinphones. The decoding does not employ any speaker clustering, but uses the speaker labels given in the test sets. The final SRC English ASR has 3,000 states and 90K Gaussians. The performance of TGT ASR trained on the adaptation set is worth noting. Although it is trained on only 3 hours, it has a WER 6.4% absolute better than the baseline source ASR, unlike its Arabic counterpart. This result also shows the difference in performance of ASR in decoding an accent, which is under-represented in the training data (Arabic setup) compared to the one in which the target accent is completely unseen during training (English setup). The large gain of 6.7% absolute for English SA system compared to SI system on the unseen target accent, unlike the Arabic setup, also validates this hypothesis.

Table 2.3: *Baseline Performance.*

System	Training Set	Test WER (%)	
		SRC	TGT
Arabic			
SRC ML SI	Train-SRC	51.2	59.0
SRC ML SA	Train-SRC	47.1	56.7
TGT ML SA	Adapt-TGT	58.2	57.1
English			
SRC ML SI	Train-SRC	13.4	30.5
SRC ML SA	Train-SRC	13.0	23.8
TGT ML SA	Adapt-TGT	33.5	17.4

2.6 Accent adaptation experiments

We chose to evaluate accent adaptation with three different techniques: MAP adaptation, fully-continuous PDTS as formulated in [SW00] and semi-continuous PDTS or SPDTS. MLLR is also a possible candidate, but its improvement saturates after 600 utterances (≈ 1 hour), when combined with MAP [HAH01]. MLLR is also reported to have issues with accent adaptation [CJ06]. The MAP smoothing factor τ is set to 10 in all cases. We did not observe additional improvements by fine-tuning this parameter. The SRC Arabic ASR had 3,000 states - the adapted

fully-continuous PDTS had 256 additional states, while semi-continuous adapted PDTS (SPDTS) ended up with 15K final states (3,000 codebooks). In a similar fashion, SRC English ASR had 3k states - Adapted English PDTS had 138 additional states while the SPDTS managed 8,000 final states (3,000 codebooks). In spite of the difference in the number of states, PDTS and SPDTS have approximately the same number of parameters in both setups. We evaluate the techniques under two different criterion: Cross-entropy of the adaptation data according to the model and WER on the target accent test set

The per-frame cross-entropy of the adaptation data $\mathcal{D}$ according to the model θ is given by

$$H_\theta(\mathcal{D}) = -\frac{1}{T}\sum_{u=1}^{U}\sum_{t=1}^{u_T}\log\ p(u_t|\theta)$$

where U is the number of utterances, u_T is the number of frames in utterance u and $T = \Sigma_u u_T$ refers to total number of frames in the training data. The cross-entropy is equivalent to average negative log-likelihood of the adaptation data. The lower the cross-entropy, the better the model fits the data. Figure 2.1 shows that the adaptation data has the lowest cross-entropy on SPDTS adapted models compared to MAP and PDTS.

The adapted models are used to decode both source and target accent test sets and the WER of all the adaptation techniques are shown in Table 2.4.

Table 2.4: *WER of MAP, PDTS and SPDTS on Accent adaptation.*

System	Test WER (%)	
	SRC	TGT
Arabic		
MAP SA	47.6	51.2
PDTS SA	47.9	50.1
SPDTS SA	48.1	47.6
English		
MAP SA	14.7	16.8
PDTS SA	15.1	15.6
SPDTS SA	16.7	14.5

MAP adaptation achieves a relative improvement of 9.7% for Levantine Arabic

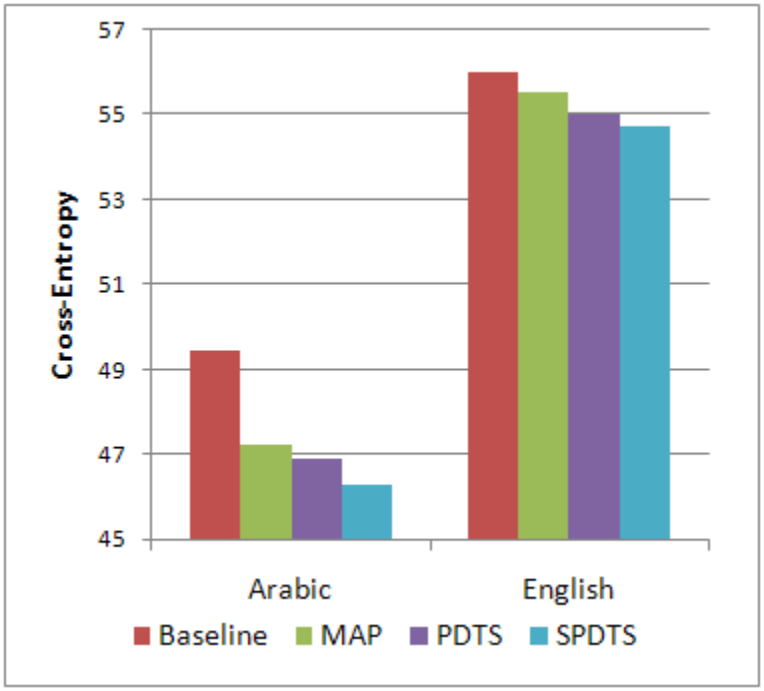


Figure 2.1: *Cross-entropy of adaptation data for various models*

and 29.4% for UK English. As expected, PDTS performs better than MAP in both cases, but the relative gap narrowed down for Arabic. SPDTS achieves additional improvement of 7% relative for Levantine Arabic and 13.6% relative for UK English over MAP adaptation.

Finally, we tried MAP, PDTS and SPDTS techniques on our 1,100 hour large-scale Arabic BN GALE evaluation ML system. We used a 2-pass unvowelized system trained on the GALE BN corpus for this experiment. It has the same dictionary, phoneset and front-end as the 200 hour BC system and it has 6,000 states and 850K Gaussians. The results are shown below

We get 5.1% relative improvement for SPDTS over MAP in adapting a large-scale ASR system trained on mostly BN MSA speech to BC Levantine Arabic. It is also interesting to note the limitation of PDTS for large systems as discussed in Section 2.2. This experiment shows that Semi-continuous PDT Adaptation can scale well to a large-scale, large vocabulary ASR trained on 1000s of hours of speech data.

Table 2.5: *Per-frame cross-entropy on the adaptation set.*

System	Cross-entropy
Arabic	
Baseline SA	49.43
MAP SA	47.21
PDTS SA	46.89
SPDTS SA	46.28
English	
Baseline SA	55.99
MAP SA	55.53
PDTS SA	55.01
SPDTS SA	54.75

Table 2.6: *Accent adaptation on GALE 1100 hour ML system.*

System	Test WER (%)	
	SRC	TGT
Arabic		
Baseline ML SA	43.0	50.6
MAP ML SA	44.5	49.1
PDTS ML SA	44.9	48.8
SPDTS ML SA	48.9	46.6

2.7 M*Modal setup - speech corpus, language model and lexicon

2.7.1 Database

M*Modal dataset consists of anonymized medical reports in the internal medicine domain, dictated by doctors across various US hospitals. The dictations are fast-paced speech over a 8KHz telephony channel lasting approximately 5 mins each. The dataset has speakers with wide variety of accents, recorded over different devices from cellphones to landline telephones and with varying background noise levels. The training dataset contains 1878 training speakers with a maximum of 1 hour per speaker. The total size of the corpus is 1,450 hours. Native US English is

the majority accent in this dataset, while South-Asian speakers form the next major group, which is our target accent. These speakers are from various countries in the subcontinent including India, Pakistan, Bangladesh and Sri Lanka. We use 168 South Asian speakers as our adaptation set. For the test set, we use the 15 medical reports from these same speakers with approximately an hour of speech per speaker. For comparison we also include a second test set with 15 native US English speakers. Table 2.7 shows the different datasets and their statistics.

Table 2.7: *M*Modal datasets and their statistics.*

Dataset	Accent	Speakers	#Hours	words
Train-SRC	Multiple	1878	1450	1.2M
Adapt-TGT	SouthAsian	168	132	126K
Test-TGT	SouthAsian	10	10.72	86K
Test-SRC	US English	15	6.23	32K

2.7.2 Baseline

The SI system is a fully-continuous, ML trained, GMM-HMM based ASR using 3,000 context-dependent states and 86K gaussians. The system uses MFCC features, appended by first and second derivates and transformed to a 32 dimensional space using a global HLDA matrix trained using the ML criterion. Additional improvements can be obtained by training canonical models using Speaker Adaptive Training (SAT) and Constrained Maximum Likelihood Linear Regression (CMLLR) matrices. However, this is a one-pass system aimed at interactive dictation so we did not include SAT in our baseline. The decoder uses a 4-gram language model with a vocabulary size of 53K words. The language model has a OOV of 0.8% on the test set.

Table 6.1 shows the WER of the SI system on the South Asian and Native US English test sets.

Table 2.8: *Baseline WERs.*

System	Test set	WER
SI	South Asian	45.73
SI	US English	29.89

Table 2.8 shows that SI WER on South Asian accented speakers is significantly worse (53% relative) than native US English speakers, before any adaptation. We conducted MAP and SPDTS adaptation and the results are shown in Table 2.9. The final SPDTS system had 8,000 semi-continuous states, while PDTS ended up with 3200 fully-continuous states.

Table 2.9: *Accent adaptation on M*Modal setup.*

System	Test WER (%)	
	SRC	TGT
Arabic		
Baseline ML SI	29.89	45.73
MAP Adapt	31.17	36.89
PDTS Adapt	32.0	36.37
SPDTS Adapt	33.20	35.09

From the table, SPDTS system obtains 23.3% relative improvement over SI baseline, 4.9% relative improvement over MAP and 3.5% relative improvement over PDTS adaptation. Table 2.10 shows the results for discriminative baseline.

Table 2.10: *Accent adaptation on M*Modal setup.*

System	Test WER (%)	
	SRC	TGT
Arabic		
Baseline DT SI	22.79	34.55
sMBR MAP Adapt	25.41	32.67
PDTS Adapt	26.10	32.32
SPDTS Adapt	27.62	31.30

For discriminative adaptation, we implemented sMBR MAP, using 4 iterations of extended Baum-Welch (EBW) updates. SPDTS is performed on top of the DT adapt model. The mixture weights are calculated with ML training as in the previous experiments. The final SPDTS system produced 4.2% relative improvement over sMBR MAP and 3.2% relative improvement over PDTS. This shows that SPDTS adaptation retains improvement over baseline and MAP adaptation involving discriminative setup.

2.8 Summary

We have introduced semi-continuous based decision tree adaptation for supervised accent adaptation. We showed that the SPDTS model achieves better likelihood on the adaptation data than other techniques. The technique obtains 7-13.6% relative improvement over MAP adaptation for medium-scale and 4.2% - 5.1% relative for large scale systems. SPDTS is evaluated under discriminative SI and adaptation to show that the technique retains most of the improvement under discriminative objective functions.

Chapter 3

Extensions to unlabelled data

Supervised adaptation using MAP/SPDTS requires transcribed accented target data for adapting the source model to the target accent. As discussed in the previous chapter, it is prohibitively costly to obtain large accented speech datasets, due to the effort involved in collecting and transcribing speech, even for a few of the major accents. On the other hand, for tasks like Broadcast News (BN) or Voice search, it is easy to obtain large amounts of speech data with representative accents. However, it is still difficult to reliably identify the accent of the speakers in such a large collection. To make use of these data sets, active and semi-supervised accent adaptation are explored in this chapter, in the context of building accent-dependent models.

3.1 Active learning

Active learning is a machine learning technique commonly used in fields where the cost of labeling the data is quite high [Set09]. It involves selecting a small subset from vast amount of unlabeled data for human annotation. To reduce the cost and ensure minimum human effort, the goal of data selection is to choose an appropriate subset of the data, that when transcribed and used to retrain the model, provides the maximum improvement in the accuracy. Active learning has been applied in natural language processing [TO09], spoken language understanding [THTS05], speech recognition [RHT05, YGWW10, YVDA10, ISJ+12], etc.

Many of the approaches in active learning, rely on some form of uncertainty based measure for data selection. The assumption is that adding the most uncertain

utterances provide the maximum information for re-training the model in the next round. Confidence scores are typically used for active learning in speech recognition [HTRG02] to predict uncertainty. Lattice [YVDA10] and N-best [ISJ+12] based techniques have been proposed to avoid outliers with 1-best hypothesis. Representative criterion in addition to uncertainty have also been shown to improve data selection in some cases [HJZ10, ISJ+12].

In the case of accent adaptation, active learning is used to extend the improvements obtained by supervised adaptation by using additional data from a large speech corpus with multiple accents, but without transcriptions or accent labels. The goal of active learning here, is to choose a relevant subset from this large dataset that matches the target accent. The subset is then manually transcribed and used to retrain the target adapted ASR, to provide additional improvements on the target accent.

3.1.1 Active learning for accent adaptation

Most of the active learning algorithms strive to find the smallest subset from the untranscribed data set that when labeled and used to re-train the ASR will have the same effect of using the entire dataset for re-training, thereby reducing the cost. However, in the case of accent adaptation using a dataset with multiple accents, our goal is not to identify the representative subset but to choose relevant utterances that best match the target test set. Data selection only based on informativeness or uncertainty criterion, can lead to selecting utterances from the mis-matched accent. Such a subset, when used to retrain the ASR, can hurt the performance on the target accent. Hence the key in this case, is to choose both informative and relevant utterances for further retraining to ensure improvements on the target accent.

A relevance criterion is introduced in addition to uncertainty based informative measure for data selection to match the target accent. The experiment starts with the ASR trained on a source accent. A relatively small, manually labeled adaptation data is then used to adapt the recognizer to the target accent. The adapted model is then employed to choose utterances from a large, untranscribed mixed dataset for human transcription, to further improve the performance on the target accent. To this end, a cross-entropy measure is calculated based on adapted and unadapted model likelihoods, to assess the relevance of an utterance. This measure is combined with uncertainty based sampling to choose an appropriate subset for manual labeling. The technique is evaluated on Arabic and English accents and shown to achieve 50-87.5% data reduction for the same accuracy of the recognizer using purely

uncertainty based data selection. With active learning on the additional unlabeled data, the accuracy of the supervised models is improved by 7.7-20.7% relative.

3.1.2 Uncertainty based informativeness criterion

In speech recognition, uncertainty is quantified by the ASR confidence score. It is calculated from the word-level posteriors obtained by consensus network decoding [MBS00]. Confidence scores calculated on 1-best hypothesis are sensitive to outliers and noisy utterances. [YVDA10] proposed a lattice-entropy based measure and selecting utterances based on global entropy reduction. [ISJ+12] observed that lattice-entropy is correlated with the utterance length and showed N-best entropy to be an empirically better criterion. In this work, an entropy-based measure is alse used as an informative criterion for data selection. The average entropy of the alignments is calculated in the confusion network as a measure of uncertainty of the utterance with respect to the ASR. It is given by

$$Informative\ score\ u_i = \frac{\sum_{\mathcal{A}\in u} E_{\mathcal{A}} D_{\mathcal{A}}}{\sum_{\mathcal{A}\in u} D_{\mathcal{A}}} \tag{3.1}$$

where $E_{\mathcal{A}}$ is the entropy of an alignment A in the confusion network and $D_{\mathcal{A}}$ is the duration of the link with best posterior in the alignment. $E_{\mathcal{A}}$ is calculated over all the links in the alignment.

$$E_{\mathcal{A}} = -\sum_{\mathcal{W}\in A} P_{\mathcal{W}} log\ P_{\mathcal{W}} \tag{3.2}$$

3.1.3 Cross-entropy based relevance criterion

In this section, a cross-entropy based relevance criteria is derived for choosing utterances from the mixed set, for human annotation. The source-target mismatch is formulated as a sample selection bias problem [CMRR08, BI10, BBS09] under two different setups. In the multi-accented case, the source data consists mixed set of accents and the goal is to adapt the model trained on the source data to the specified target accent. The source model can be assumed as a background model that has seen the target accent during training, albeit it is under-represented along with other accents in the source data. In the second case, the source and target data belong to two mis-matched accents. The source model is adapted to a completely different target accent, unseen during training. The biased sampling criterion for

both the multi-accented and mis-matched accent cases is derived separately in the following sections.

Multi-accented case

In this setup, the source data contains a mixed set of accents. The target data, a subset of the source represents utterances that belong to a specific target accent. An utterance u in the data set is represented by a sequence of observation vectors and its corresponding label sequence. Let X denote the space of observation sequences and Y the space of label sequences. Let S denote the distribution over utterances $U \in X \times Y$ from which source data points (utterances) are drawn. Let T denote the target set distribution over $X \times Y$ with utterances $\hat{U} \subseteq U$. Now, utterances in T are drawn by biased sampling from S denoted by the random variable $\sigma \in \{0, 1\}$ or the *bias*. When $\sigma = 1$, the randomly sampled $u \in U$ is included in the target dataset and when $\sigma = 0$ it is ignored. Our goal is to estimate the bias $Pr[\sigma = 1|u]$ given an utterance u, which is a measure for how likely is the utterance to be part of the target data. The probability of an utterance u under T can be expressed in terms of S as

$$Pr_T[u] = Pr_S[u|\sigma = 1] \tag{3.3}$$

By Bayes rule,

$$Pr_S[u] = \frac{Pr_S[u|\sigma = 1]Pr[\sigma = 1]}{Pr[\sigma = 1|u]} = \frac{Pr[\sigma = 1]}{Pr[\sigma = 1|u]} Pr_T[u] \tag{3.4}$$

The bias for an utterance u is represented by $Pr[\sigma = 1|u]$

$$Pr[\sigma = 1|u] = \frac{Pr_T[u]}{Pr_S[u]} Pr[\sigma = 1] \tag{3.5}$$

The posterior $Pr[\sigma = 1|u]$ represents the probability that a randomly selected utterance $u \in U$ from the mixed set belongs to the target accent. It can be used as a relevance score for identifying relevant target accent utterances in the mixed set. Since we are only comparing scores between utterances for data selection, $Pr[\sigma = 1]$ can be ignored in the above equation as it is independent of u. Further, we can approximate $Pr_S[u]$ and $Pr_T[u]$, by unadapted and adapted model likelihoods. Substituting and changing to log domain,

$$Relevance\ Score\ u_r \approx \log\ Pr[u|\lambda_T] - \log\ Pr[u|\lambda_S] \tag{3.6}$$

The utterances in the mixed set can have different durations, so we normalize the log-likelihoods to remove any correlation of the score with the duration. The length normalized log-likelihood is also the cross-entropy of the utterance given the model [ML10, NMS12c] with sign reversed. The score that represents the relevance of the utterance to target dataset is given by

$$Relevance\ Score\ u_r = (-H_{\lambda_T}[u]) - (-H_{\lambda_S}[u]) \tag{3.7}$$

where

$$H_\lambda(u) = -\frac{1}{T_u}\sum_{t=1}^{T_u} \log\ p(u_t|\lambda) \tag{3.8}$$

is the average negative log-likelihood or the cross-entropy of u according to λ and T_u is the number of frames in utterance u.

Mis-matched accents case

In this case, source and target correspond to two different accents. let A denote the distribution over observation and label sequences $U \in X \times Y$. Let S and T be the source and target distributions over $X \times Y$ and subsets of A, $U_S, U_T \subseteq U$. The source and target utterances are drawn by biased sampling from A governed by the random variable $\sigma \in \{0, 1\}$. When the bias $\sigma = 1$, the sampled utterance u is included in the target dataset and $\sigma = 0$ it is included in the source dataset. The distributions S and T can be expressed in terms of A as

$$Pr_T[u] = Pr_A[u|\sigma = 1]; Pr_S[u] = Pr_A[u|\sigma = 0] \tag{3.9}$$

By Bayes rule,

$$Pr_A[u] = \frac{Pr[\sigma = 1]}{Pr[\sigma = 1|u]} Pr_T[u] = \frac{Pr[\sigma = 0]}{Pr[\sigma = 0|u]} Pr_S[u] \tag{3.10}$$

Equating LHS and RHS

$$\begin{aligned} \frac{Pr_S[u]}{Pr_T[u]} &= \frac{Pr[\sigma = 1]}{Pr[\sigma = 0]} \frac{Pr[\sigma = 0|u]}{Pr[\sigma = 1|u]} \\ &= \frac{Pr[\sigma = 1]}{Pr[\sigma = 0]} \left[\frac{1}{Pr[\sigma = 1|u]} - 1\right] \end{aligned} \tag{3.11}$$

As in the previous case, we can ignore the constant terms that don't depend on u as we are only comparing the scores between utterances. The relevance score, which is an approximation of $Pr[\sigma = 1|u]$ is given by

$$Relevance\ score\ u_r \approx \frac{Pr_T[u]}{Pr_T[u] + Pr_S[u]} \tag{3.12}$$

Changing to log-domain,

$$\begin{aligned} Relevance\ score\ u_r &\approx \log\ Pr_T[u] \\ &\quad - \log\ (Pr_T[u] + Pr_S[u]) \\ &= \log\ Pr_T[u] \\ &\quad - \log\ \left(Pr_T[u]\left[1 + \frac{Pr_S[u]}{Pr_T[u]}\right]\right) \\ &= -\log\ \left(1 + \frac{Pr_S[u]}{Pr_T[u]}\right) \end{aligned} \tag{3.13}$$

log is a monotonous function, hence $\log(1 + x) > \log(x)$ and since we are only comparing scores between utterances, we can replace $\log(1 + x)$ with $\log(x)$. The relevance score is then the same as the multi-accented case

$$\begin{aligned} Relevance\ Score\ u_r &\approx \log\ Pr_T[u] - \log\ Pr_S[u] \\ &\approx \log\ Pr[u|\lambda_T] - \log\ Pr[u|\lambda_S] \end{aligned}$$

Normalizing the score to remove any correlation with utterance length,

$$Relevance\ Score\ u_r = (-H_{\lambda_T}[u]) - (-H_{\lambda_S}[u]) \tag{3.14}$$

3.1.4 Score combination

Our final data selection algorithm uses a combination of relevance and uncertainty scores for active learning. The difference in cross-entropy is used a measure of relevance of an utterance. The average entropy based on the confusion network is used as a measure of uncertainty or informativeness. Both the scores are in log-scale and we use a simple weighted combination to combine both the scores [ISJ+12]. The final score in given by

$$Final\ score\ u_F = u_r * \theta + u_i \tag{3.15}$$

The mixing weight, θ is tuned on the development set. The final algorithm for active learning that uses both the relevance and informativeness scores is given below.

Algorithm 1 Active learning using relevance and informativeness scores

Input: $\mathcal{X}_T :=$ Labeled Target Adaptation set ; $\mathcal{X}_M :=$ Unlabeled Mixed set ; $\lambda_S :=$ Initial Model ; $\theta :=$ Mixing weight $minScore :=$ Selection Threshold

Output: $\lambda_T :=$ Target Model

$\lambda_T := Adapt(\lambda_S, \mathcal{X}_T)$
for all x in $\mathcal{X}_M$ **do**
  $Loglike_S := -CrossEntropy(\lambda_S, x)$
  $Loglike_T := -CrossEntropy(\lambda_T, x)$
  $Len := Length(x)$
  $RelevanceScore := (Loglike_T - Loglike_S)/Len$
  $InformativeScore := -AvgCNEntropy(\lambda_T, x)$
  $FinalScore := RelevanceScore * \theta + InformativeScore$
  if $(FinalScore > minScore)$ **then**
    $\mathcal{L}_x := QueryLabel(x)$
    $\mathcal{X}_T := \mathcal{X}_T \cup (x, \mathcal{L}_x)$
    $\mathcal{X}_M := \mathcal{X}_M \setminus x$
  end if
end for
$\lambda_T := Adapt(\lambda_S, \mathcal{X}_T)$
return λ_T

3.1.5 Experiment setup

Datasets

Active learning experiments are conducted on both multi-accented and mis-matched accent cases. Multi-accented setup is based on GALE Arabic database discussed in the previous chapter. 1100 hours of Broadcast News (BN) is used as the source training data. It contains mostly Modern Standard Arabic (MSA) but also varying amounts of other dialects. Levantine is assigned as the target accent and randomly selected 10 hours from 30 hour LDC Levantine annotations and created our adaptation dataset. The remaining 20 hours of Levantine speech is mixed with 200 hours of BC data to create the Mixed dataset. This serves as our unlabeled dataset for active learning.

For mis-matched accent case, English WallStreet Journal (WSJ1) is chosen as the source data, as in the previous chapter. British English is used as the target accent and the British version of WSJ corpus (WSJCAM0) for adaptation. 3 hours from

WSJCAM0 are randomly sampled for the adaptation set. The remaining 12 hours of British English speech is mixed with 15 hours of American English from WSJ0 corpus to create our mixed dataset. The test sets, LM and dictionary are similar to our earlier setup. Table 3.1 provides a summary of the datasets used.

Table 3.1: *Database Statistics.*

Dataset	Accent	#Hours	Ppl	%OOV
Arabic				
Training	Mostly MSA	1092.13	-	-
Adaptation	Levantine	10.2	-	-
Mixed	Mixed	221.9	-	-
Test-SRC	Non-Levantine	3.02	1011.57	4.5
Test-TGT	Levantine	3.08	1872.77	4.9
English				
Training	US	66.3	-	-
Adaptation	UK	3.0	-	-
Mixed	Mixed	27.0	-	-
Test-SRC	US	1.1	221.55	2.8
Test-TGT	UK	2.5	180.09	1.3

Baseline systems

HMM-based, speaker-independent ASR systems are built on the training data. They are Maximum Likelihood (ML) trained, context-dependent, fully-continuous systems with global LDA and Semi-Tied Covariance (STC) transform. More details on the front-end, training and decoding framework are explained in [MHJ+10, NMS12b]. We initially adapt our baselines systems on the relatively small, manually labeled, target adaptation dataset. We used semi-continuous polyphone decision tree adaptation (SPDTS) [NMS12b] for the supervised adaptation. The WER of the baselines and supervised adaptation systems are given in Table 3.2.

3.1.6 Implementation details

We use the supervised adapted systems to select utterances from the mixed set for the goal of target accent adaptation. Our mixed sets were created by combining two

Table 3.2: *Baseline and Supervised adaptation WERs.*

System	# Hours	Test WER (%)	
		SRC	TGT
Arabic			
Baseline	1100	46.3	53.7
Supervised Adapt	+10	51.4	52.1
English			
Baseline	66	13.4	30.5
Supervised Adapt	+3	21.0	17.9

datasets, American and British English or BC and Levantine Arabic. We evaluate 3 different data selection algorithms for our experiments: Random sampling, Uncertainty or informative sampling and relevance augmented uncertainty sampling. In each case, we select fixed amounts of audio data alloted to each bin and mix it with the adaptation data. We then re-adapt the source ASR on the newly created dataset. For this second adaptation, we reuse the adapted polyphone decision tree from the supervised case, but we re-estimate the models on the new dataset using Maximum A Posteriori (MAP) adaptation.

In random sampling, we pick at random the required number of utterances from the mixed dataset. The performance of the re-trained ASR directly depends on the composition of source and target utterances in the selected subset. Thus, ASR re-trained on randomly sampled subsets will exhibit high variance in its performance. To avoid varying results, we can run random sampling multiple times and report the average performance. The other solution is to enforce that the randomly selected subset retains the same composition of source and target utterances in the mixed set. We use the latter approach for the results reported here.

For uncertainty based sampling, we used average entropy calculated over the confusion networks (CN) as explained in section 3.1.2. We decode the entire mixed set and choose utterances that have the highest average CN entropy. In the case of relevance augmented uncertainty sampling, we use a weighted combination of relevance and uncertainty or informativeness scores for each utterance. The relevance score is derived from adapted and unadapted model cross-entropies with respect to the utterance. We calculate cross-entropy or average log-likelihood scores using the lattices produced during decoding. The uncertainty score is calculated using average CN entropy as before. We tuned the mixing weights on the English development set and we use the same weight (0.1) for all the experiments. We

selected 5, 10, 15, 20 hour bins for English and 5, 10, 20, 40, 80 bins for Arabic. We choose utterances for each bin and combine it with the initial adaptation set, re-adapt the ASR and evaluate it on the target test set.

Table 3.3 shows WER of the oracle and select-all benchmarks for the two datasets. The oracle involves selecting all the target (relevant) data for human transcription, that we combined with source data to create the mixed dataset. The selected data is added to the initial adaptation set and used to re-adapt the source ASR. We note that in the case of Arabic, the source portion (BC) of the mixed dataset can have additional Levantine utterances, so oracle WER is not the lower bound for Arabic. Select-all involves selecting the whole mixed dataset for manual labeling. From Table 3.3, we can realize the importance of the relevance measure for active learning. In the case of Arabic, one-tenth of relevant data produces better performance on the target test set than the whole mixed dataset. The case is similar for English, where half of the relevant utterances help ASR achieve better performance than presenting all the available data for labeling.

Table 3.3: *Oracle and Select-all WERs.*

System	# Hours	Target WER
Arabic		
Oracle	$10 + 20$	48.7
Select-all	$10 + 221.9$	50.8
English		
Oracle	$3 + 12$	14.2
Select-all	$3 + 27$	14.9

3.1.7 Active learning results

The results for active learning for Arabic is shown in Figure 3.1. It is clear from the plot that the weighted combination of relevance and informative scores perform significantly better than uncertainty based score and random sampling techniques. We observe a 1.7% absolute WER reduction at the peak (40hours) for the weighted score when compared to the CN entropy based data selection technique. Also, with only 5 hours, the weighted score reaches WER of 49.5% while the CN-entropy based technique required 40 hours of data to reach a similar WER of 49.8%. Thus the combined score requires 87.5% less data to reach the same accuracy of CN-entropy based sampling. It is also interesting to note that our algorithm has identified

additional Levantine data than the oracle from the generic BC portion of the mixed set which resulted in further WER reductions.

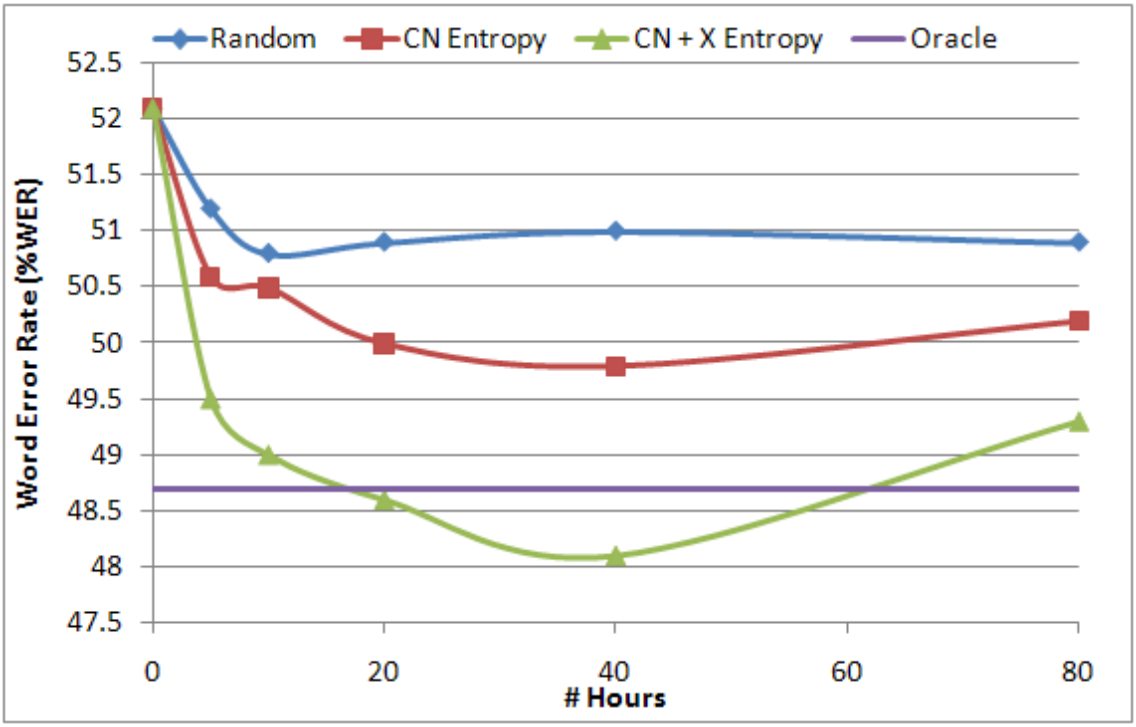


Figure 3.1: *Active learning results for Arabic*

Figure 3.2 shows the equivalent plots for English. The combined score outperforms other techniques in terms of the WER and reaches the performance of the oracle benchmark. It obtains similar performance with 10 hours of data (14.5%) compared to CN-entropy based technique at 20 hours (14.8%), thus achieving a 50% reduction in labeling costs.

3.1.8 Analysis

In this section we analyze the influence of relevance score in choosing the utterances that match the target data in both the setups. We plot the histogram of both CN-entropy and weighted scores for each task. Figure 3.3 shows the normalized histograms for the American and British English utterances in the mixed set. We note that the bins for these graphs are in the ascending order of their scores. Data

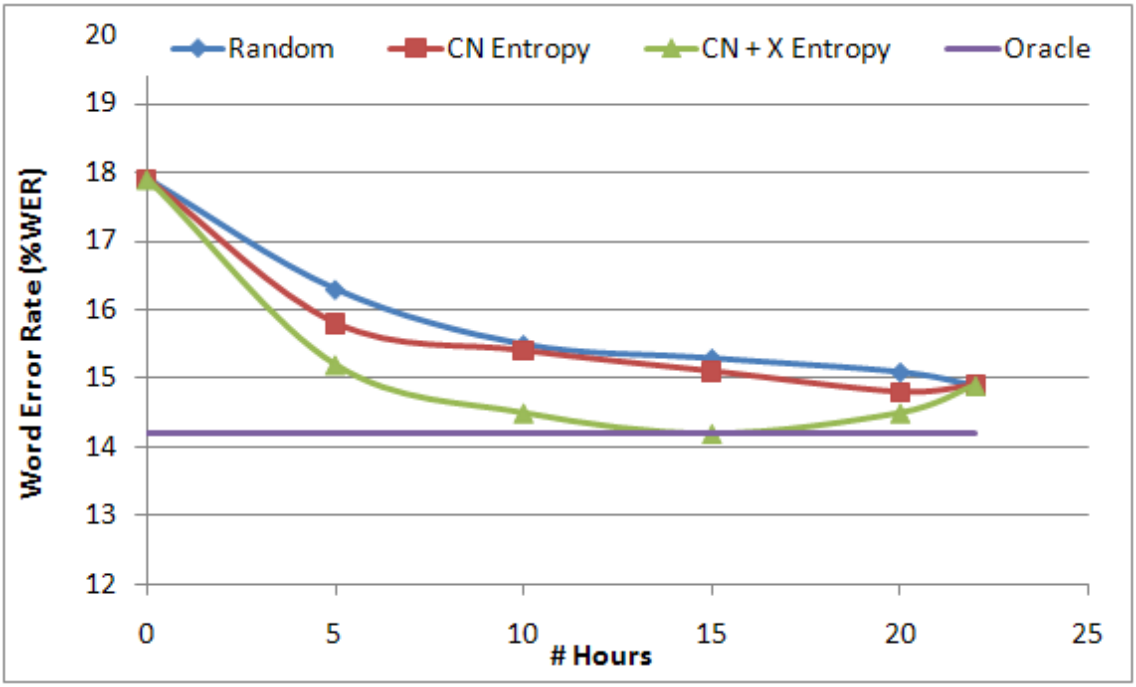
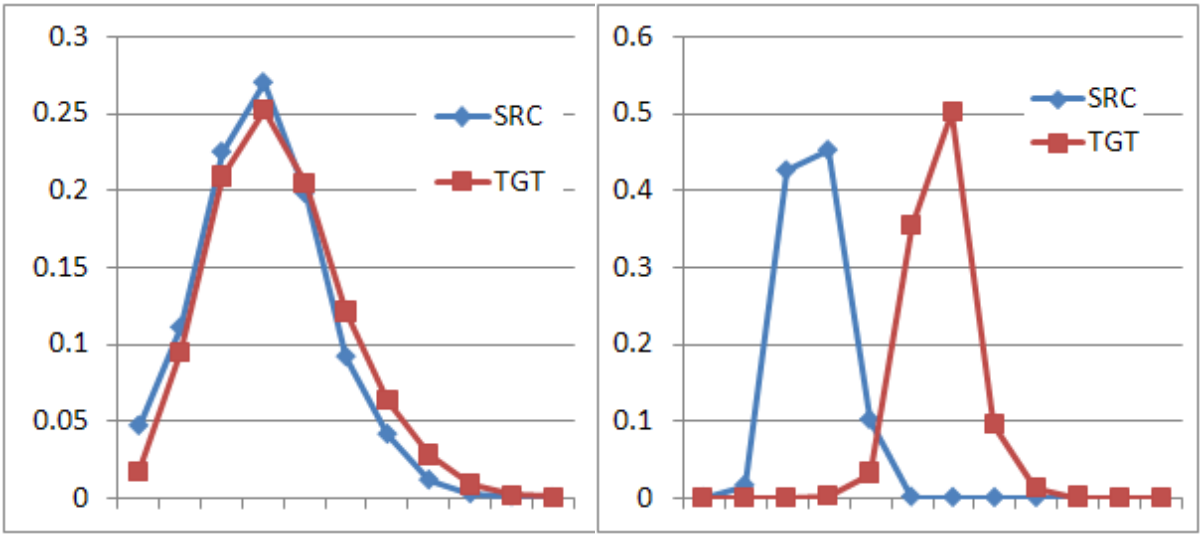


Figure 3.2: *Active learning results for English*

selection starts with the high-scoring utterances, hence the utterances from the right side of the plot are chosen first during active learning. Figure 3.3(a) shows the entropy scores for source (American English) and target (British English) are quite similar and the algorithm will find it harder to differentiate between relevant and irrelevant utterances based solely on uncertainty score. Figure 3.3(b) shows the influence of adding relevance scores to uncertainty scores. In this case, the target utterances have higher scores than source utterances and the algorithm chooses relevant ones for re-training the ASR.

Figure 3.4 shows similar plots for Arabic. The distinction between CN-entropy and the weighted score in source/target discrimination is less clear here compared to English plots. However, we can still see that target utterances achieve better scores with weighted combination than the CN-entropy score. We observed many of the utterances from 'LBC_NAHAR' shows, part of the BC portion of the mixed set, ranked higher in the weighted score. The plot of LBC scores in the histogram shows these utterances from the BC portion have high scores in the weighted case. They are recording of the 'Naharkum Saiid' (news) programmes from Lebanese Broadcasting Corporation originating from the Levantine region and likely to have



(a) Entropy (b) Weighted Score

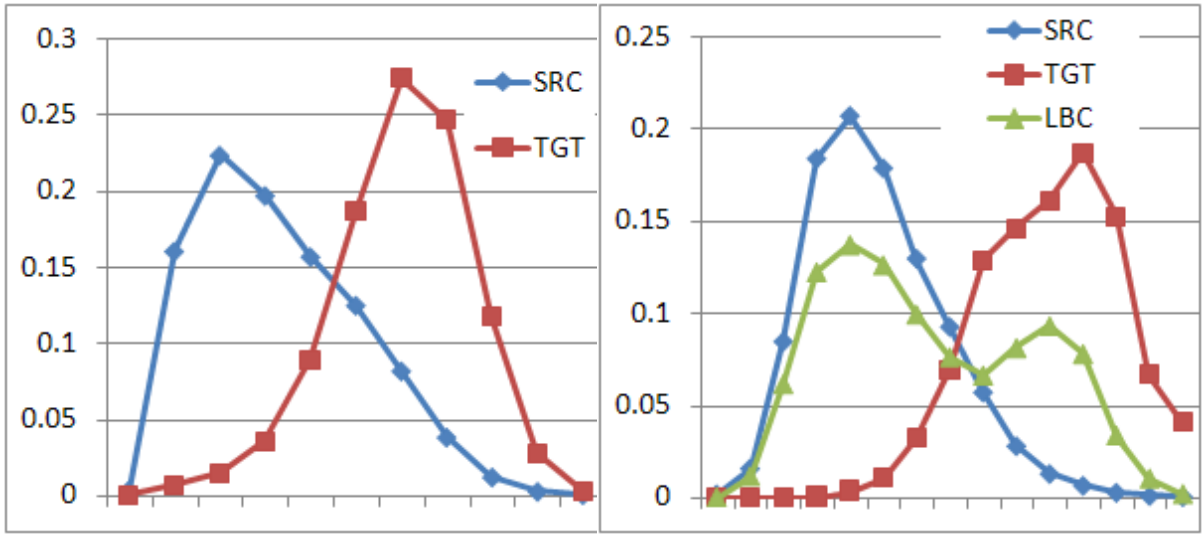
Figure 3.3: Histogram of source and target scores for English.

Levantine speech. This observation shows that the relevance score identifies additional Levantine speech from the BC utterances.

3.2 Semi-supervised learning

Semi-supervised learning has become attractive in ASR given the high cost of transcribing audio data. Unlike active learning, where one chooses a subset of the untranscribed data for manual transcription, semi-supervised learning uses the existing ASR to choose and transcribe the required data for further training.

Self-training is a commonly used technique for semi-supervised learning in speech recognition [YGWW10, WN05, KW99, Ram05, MS08], whereby the initial ASR trained using carefully transcribed speech is used to decode the untranscribed data. The most confident hypotheses are chosen to re-train the ASR. Self-training has been successfully employed under matched training conditions where the labeled training set used to train the seed ASR and the unlabeled dataset have similar acoustic characteristics. It has also enjoyed some success in cross-domain adaptation where the source seed ASR is adapted using untranscribed data from a different target language, dialect or channel [LGN09, NSK11]. In the latter task the



(a) Entropy (b) Weighted Score

Figure 3.4: Histogram of source and target scores for Arabic.

target data, while different from the initial source training dataset, is still assumed to be homogeneous. Our work differs from these setups as the unannotated data in our experiments is not homogeneous. It can have multiple accents, with or without transcriptions. The goal is to select the relevant subset to match the target accent. Hence, choosing hypotheses solely based on confidence scores is not ideal for accent adaptation in this case.

In this section we discuss cross-entropy based data selection to identify speakers that match our target accent, before filtering the utterances by confidence scores. The seed ASR is initially adapted on the target accent using limited, manually labeled adaptation data. We then make use of the adapted and unadapted models to select speakers based on their change in average likelihoods or cross-entropy under adaptation. We couple the speaker selection with confidence based utterance-level selection to choose an appropriate subset from the unlabeled data to further improve the performance on the target accent. We evaluate our technique with Arabic and English accents and show that we achieve between 2.0% and 15.9% relative improvement over supervised adaptation using cross-entropy based data selection. Self-training using only confidence scores fails to achieve any improvement over the initial supervised adaptation in both tasks.

Semi-supervised learning for ASR adaptation involves three steps - training/adapting

initial ASR on limited target data with manual labels, decoding the unlabeled data with the initial adapted model and selecting a suitable subset to re-train the seed ASR, thereby improving its performance on the target test set. The criteria to select an utterance for further re-training, can be based on the following:

- Confidence - How confident is the system about the newly generated hypothesis for the utterance?
- Relevance - How relevant is the utterance for additional improvement in the target test set?

3.2.1 Self-training

Self-training employs confidence scores to select the data for re-training. Confidence scores in ASR are computed using word-level posteriors obtained from consensus network decoding [MBS00]. The selection can be done at utterance, speaker or session level. The average confident score for the appropriate level is calculated as

$$CS_{\mathcal{S}} = \frac{\sum_{\mathcal{W}\epsilon\mathcal{S}} C_{\mathcal{W}} T_{\mathcal{W}}}{\sum_{\mathcal{W}\epsilon\mathcal{S}} T_{\mathcal{W}}} \tag{3.16}$$

where $\mathcal{S}$ can be utterance or speaker or session, $CS_{\mathcal{S}}$ is average confidence score for $\mathcal{S}$ and $C_{\mathcal{W}}$, $T_{\mathcal{W}}$ are the word-level score and duration respectively for the 1-best hypothesis. To avoid outliers with 1-best hypothesis, lattice-level scores have also been proposed for semi-supervised training [YVDA10, FSGL11]. One of the issues with self-training is that it assumes all the data to be relevant and homogeneous. So, data selection is based only on ASR confidence and the relevance criteria is ignored. In our experiments, the unlabeled data has speakers with different accents and data selection based entirely on confidence scores fails to find suitable data for further improvement with re-training.

3.2.2 Cross-entropy based data selection

In this section, we formulate cross-entropy based speaker selection to inform relevance in addition to confidence based utterance selection for semi-supervised accent adaptation. Let us assume that the initial model λ_S is trained on multiple accents from unbalanced training set. It is then adapted on a limited, manually labeled

target accent data set to produce the adapted model λ_T. We have available a large mixed dataset without any accent labels. The goal is to select the target speakers from this mixed dataset and re-train the initial ASR for improved performance on the target test set. We formulate the problem of identifying target data in a mixed dataset similar to sample selection bias correction [BI10, CMRR08, BBS09]. We follow the same derivation as the active learning, but we calculate the relevance at the speaker-level, as we work with speaker-adapted systems in the following experiments.

The final score for target data selection for both the multi-accented and mismatched accents case is given by

$$Selection\ Score = (-H_{\lambda_T}[s]) - (-H_{\lambda_S}[s]) \tag{3.17}$$

where

$$H_\lambda(s) = -\frac{1}{T_s}\sum_{u=1}^{U_s}\sum_{t=1}^{u_T}\log\ p(u_t|\lambda) \tag{3.18}$$

is the average negative log-likelihood or the cross-entropy of s according to λ, U_s is the number of utterances for s, u_T is the number of frames in utterance u and $T_s = \Sigma_u u_T$ refers to total number of frames for s.

We can now sort the speakers in the mixed dataset using this selection score and choose the top scoring subset based on a threshold. The algorithm 2 shows the pseudo code for cross-entropy based semi-supervised learning for target accent adaptation.

3.2.3 Implementation details

We start with a GMM-HMM model trained on the source data. We adapt this model to the target accent using a small amount of manually transcribed target data. We use enhanced polyphone decision tree adaptation based on semi-continuous models (SPDTS) [NMS12b] for supervised adaptation. It involves using the fully continuous source model to collect occurance statistics for each state in the target data. These statistics are used to grow a semi-continuous, second-level decision tree on the adaptation dataset to better match the new contexts with the target accent. We then use Maximum A Posteriori (MAP) adaptation [GL94a] to refine the Gaussians (codebooks) and associated mixture weights (distributions) on the adaptation data. SPDTS gives additional improvements over the traditional MAP adaptation.

Algorithm 2 Cross-entropy based semi-supervised learning

Input: $\mathcal{X}_T$:= Target Adaptation set ; $\mathcal{X}_M$:= Mixed set ; λ_S := Initial Model ; $minScore$:= Selection Threshold
Output: λ_T := Target Model

$\lambda_T := Adapt(\lambda_S, \mathcal{X}_T)$
for all x in $\mathcal{X}_M$ **do**
    $Loglike_S := Score(\lambda_S, x)$
    $Loglike_T := Score(\lambda_T, x)$
    $Len := Length(x)$
    $Score := (Loglike_T - Loglike_S)/Len$
    if $(Score > minScore)$ **then**
        $\mathcal{X}_T := \mathcal{X}_T \cup x$
        $\mathcal{X}_M := \mathcal{X}_M \setminus x$
    end if
end for
$\lambda_T := Adapt(\lambda_S, \mathcal{X}_T)$
return λ_T

We use the target accent adapted ASR as the baseline and select suitable data from the mixed set for further improvements on the target test set. Data selection can be performed at multiple level segments: utterance, speaker or session. In our experiments, we rely on both speaker-level and utterance-level scores for both self-training and cross-entropy based data selection. All our baselines are speaker adapted systems, so we need a reasonable amount of speaker-specific data (minimum 15s) for robust Constrained Maximum Likelihood Linear Regression (CMLLR) based speaker-adaptive training [PY12a]. Utterance-level selection alone does not ensure this constraint. Secondly, the accent information (relevance) and hypothesis accuracy (confidence) can be asserted reliably at the speaker and utterance levels respectively. For self-training, we sort the speakers based on speaker-level, log-likelihood scores normalized by number of frames. For each best-scoring speaker in the list, we enforce the additional limitation that the selected speaker should have at least 15s of utterances that passed the minimum confidence threshold. This allows us to choose speakers with enough utterances for reliable CMLLR based speaker-adaptive (SA) training. For cross-entropy based data selection, we replace the speaker-level confidence score with the difference of length normalized log-likelihoods as specified in Equation 3.17.

We experiment with two different setups. In the first task, the mixed set has

transcriptions available, but doesn't have accent labels. The goal is to choose a relevant subset of audio and its transcription for re-training the initial model. We evaluate both self-training and cross-entropy based data selection for choosing useful data from the mixed set. Given that we have transcriptions available, we omit confidence-based filtering at the utterance level during data selection for this task. In self-training, we use the adapted model to Viterbi align the transcription with the audio for the utterances of each speaker in the mixed set. The confidence score in Equation 3.16 is replaced with the speaker-level, length normalized alignment score for this task. We then select different amounts of data by varying the threshold and re-train the seed ASR to test for improvements. In cross-entropy based data selection, the normalized log-likelihoods corresponding to the adapted and unadapted models are used to select the relevant speakers. Given the transcriptions for each utterance of speaker s, Equation 3.18 becomes

$$H_\lambda(s) = -\frac{1}{T_s}\sum_{u=1}^{U_s}\sum_{t=1}^{u_T}\log\ p(u_t|\lambda, W_r) \tag{3.19}$$

where W_r is the transcription of the audio.

For the second task, the mixed set does not have either transcriptions or accent labels available. Self-training in this case, relies on confidence scores obtained by consensus network decoding [MBS00]. The speaker-level scores are used to choose the most confident speakers and for each speaker, utterances that have an average confidence score greater than 0.85 are selected. 0.85 threshold was chosen as it gave us a good trade-off between WER and amount of available data for selection. Additionally, we enforce the 15s minimum constraint for all selected speakers as explained above. In the case of cross-entropy based selection, we replace the speaker-level confidence score with difference in cross-entropy between adapted and unadapted models. The cross-entropy of a speaker with a model is calculated based on the lattice instead of 1-best hypothesis to avoid any outliers. The lattice-based cross-entropy can be calculated as

$$H_\lambda(s) = -\frac{1}{T_s}\sum_{u=1}^{U_s}\sum_{t=1}^{u_T}\log\ p(u_t|\lambda, W) \tag{3.20}$$

where W is the set of paths in the lattice of the decoded hypothesis and

$$p(u|\lambda, W) = \sum_{w=1}^{W} p(u|\lambda, w)p(w) \tag{3.21}$$

where $p(w)$ is LM prior probability of path w. We choose best scoring speakers on the cross-entropy based selection score and for each speaker, we select utterances same as self-training with minimum confidence score of 0.85. Speakers are constrained to have minimum of 15s duration as above. We re-train the seed ASR using the additional data and report improvements on the test set.

3.2.4 Experiment setup

We used the same setup as active learning for semi-supervised learning experiments. However, unlike the SI baseline in active learning experiments, we used a speaker-adaptive setup with CMLLR-SAT training and MLLR based model adaptation during decoding. For semi-supervised learning, we start off with supervised adaptation of baseline systems on the target accent using limited, manually labeled *Adaptation set*. These adapted systems are used as seed models to select an appropriate subset from the *Mixed set* to further improve the performance on the target accent. Table 3.4 shows the WER of the baseline and adapted systems.

Table 3.4: *Baseline and Supervised adaptation WERs.*

System	# Hours	Test WER (%)	
		SRC	TGT
Arabic			
Baseline	1100	43.0	50.6
Supervised Adapt	+10	44.0	47.8
English			
Baseline	66	12.9	23.6
Supervised Adapt	+3	13.7	14.5

Semi-supervised learning experiments

In this section we study semi-supervised learning on the *Mixed set* in two different setups. In the first, we assume that the *Mixed set* is transcribed, but with no accent labels. We compare self-training and cross-entropy data selection based on Viterbi alignment scores to select appropriate speakers for improving the initial system. In the second setup, we assign the *Mixed set* to have neither transcriptions nor accent labels. In this experiment, we decode the utterances using initial ASR(s) to

obtain the likely hypotheses. We then use lattice likelihoods and confidence scores to choose the appropriate subset for accent adaptation.

Task 1 - Mixed set with transcriptions, no accent labels

For English, we choose 5, 10, 12, 15, 20 hours of audio from the mixed set to re-train the initial ASR in the case of self-training and cross-entropy based selection. We selected 10, 20, 30, 40 and 50 hours of audio data for Arabic from the mixed set. Figure 1 shows the WER of English and Arabic semi-supervised data selection with self-training and cross-entropy difference. The bin 0 corresponds to the supervised adaptation on manually labeled adaptation data. The graphs contain two baselines in addition to self-training and cross-entropy plots. Select-ALL refers to the scenario where all of the available data in the mixed set (27 hours for English and 222 hours for Arabic) are selected for re-training. This corresponds to the lower bound for semi-supervised learning. ORACLE refers to selection of all of the target data in the mixed set. This includes 12 hours of British accent in the case of English and 20 hours of Levantine for Arabic. We note that, ORACLE is only included for comparison and doesn't correspond to the upper bound for our task. A robust data selection would exclude utterances with noise, wrong transcriptions, etc. which will improve the accuracy of the re-trained model. In the case of Arabic, 20 hours of Levantine only correspond to data annotated by LDC. The remaining BC data can have more Levantine speech, which will also help improve on the ORACLE.

In both Arabic and English, self-training does not produce any improvements from semi-supervised learning over the supervised adaptation baseline. In Table.3.4, the WER on the target test set is higher than the source test set, even for the adapted systems. Hence, log-likelihood or confidence based data selection based on the adapted model cannot differentiate between relevant data (target accent) and irrelevant data (source accent). The initial speakers selected for self-training belong exclusively to the source accent which is the reason for the poor performance of re-trained models. This experiment clearly shows that data selection based only on confidence scores fails when the source ASR is adapted on a limited target data and the unlabeled data is not homogeneous. Cross-entropy based selection on the other hand, relies on change in log-likelihood before and after adaptation to identify the relevant speakers from the mixed set. It obtains an improvement of 2.3% absolute (or 15.9% relative @12 hours) for English and 1.8% absolute (or 3.8% relative @20 hours) for Arabic over the supervised baseline.

It is also interesting to note that in the case of English 90% of the selected

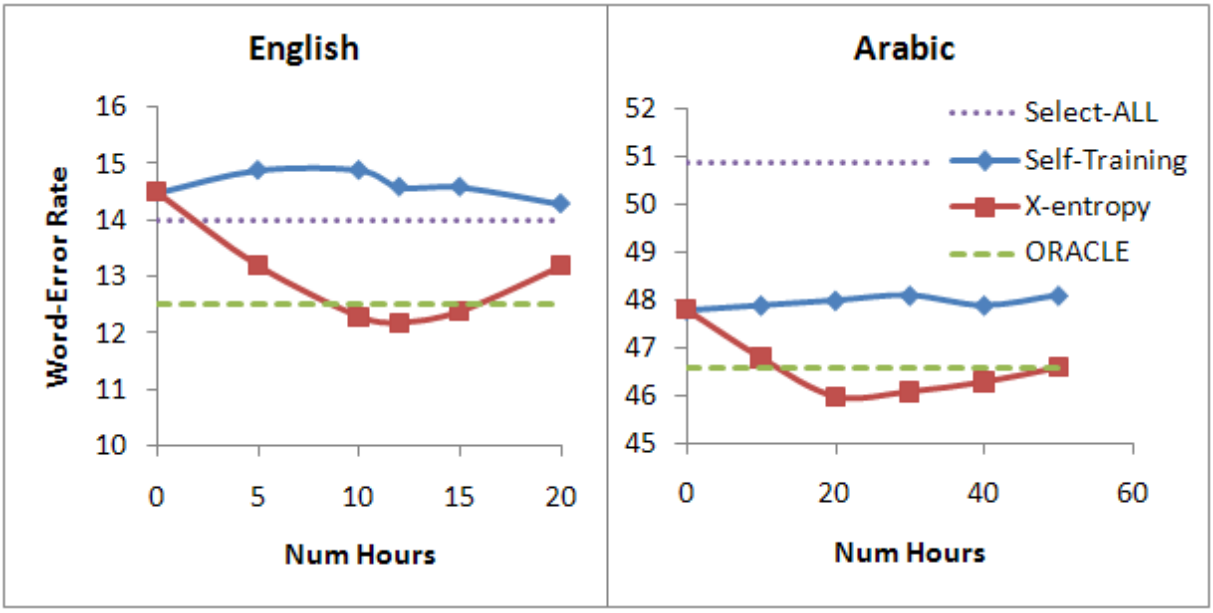


Figure 3.5: *Semi-supervised data selection with transcriptions*

speakers at 12 hours were WSJCAM0 (British English) speakers, while only 40% of the Arabic speakers at 20 hours were from the LDC annotated Levantine set. It is also shown that some of the remaining speakers from the target accent left out for data selection, had worse scores due to transcription errors, etc. This is probably the reason for slight improvement of the best semi-supervised system over the ORACLE (or fully-supervised) adaptation.

Task 2 - Mixed set without transcriptions and no accent labels

The same framework and bins are used as in the previous experiment. For self-training, speaker and utterance selection rely on confidence scores as in Eq. 3.16. For cross-entropy based data selection, speaker level selection is based on the difference in lattice likelihoods as in Eq 3.20. Figure 2 shows the WER of semi-supervised data selection with self-training and cross-entropy difference for English and Arabic datasets. The Select-ALL and ORACLE numbers correspond to 1-best hypothesis from the adapted target ASR.

As expected, the results are similar to the previous experiment as self-training fails to obtain any additional improvements with the mixed data. 2% absolute

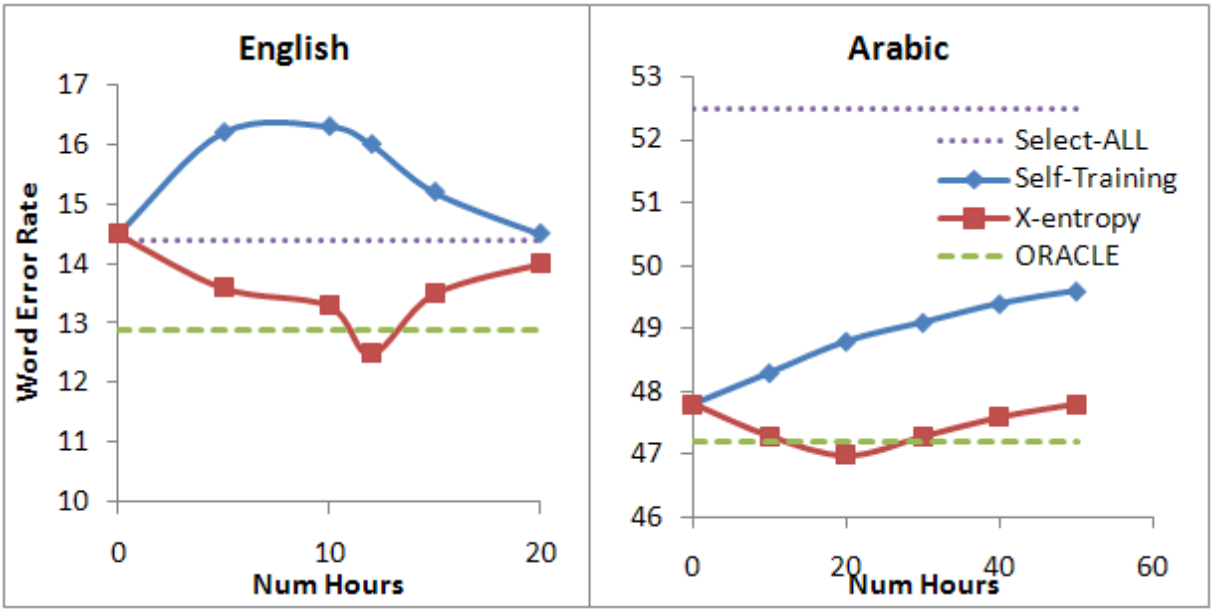


Figure 3.6: *Semi-supervised data selection without transcriptions*

(or 13.8% relative @12 hours) improvement is obtained over supervised baseline for English and 0.8% absolute (or 2.0% relative @12 hours) for Arabic. The total improvement is lower for Arabic compared to English (2.0-3.8% relative vs. 13.8-15.9% relative). However, it is comparable to the gain obtained using a dialect classifier on a similar setup [SMB11].

3.3 Summary

In this chapter, the use of additional untranscribed data for the goal of accent adaptation is investigated. A relevance criterion based biased sampling is proposed, in addition to the informativeness criterion for data selection. The combined criterion was evaluated under active and semi-supervised learning scenarios. It performed better than random and informative sampling techniques in identifying the relevant data for additional improvements on the target test set.

Part III

Accent robust modeling

Chapter 4

Robustness analysis

In this chapter, we deal with training ASR models on datasets with multiple accents. Given that real-world datasets often have speakers with varying accents, it is necessary for ASR to cope with such diversity in the training data. It can be achieved in two different ways. In accent normalization, we seek models that are robust to acoustic variations presented by different accents. As we discussed earlier, these variations can include pronunciation changes, prosody and stress. In accent adaptive training, we use a factorized model with accent-specific parameters and accent-independent, canonical models. The goal is that the accent-specific parameters will learn the intricate variations specific to a particular accent, while the canonical models will learn the shared patterns between different accents. We explore both the approaches in this chapter. Finally, we show that making the models aware of the accent during training or accent-aware training, allows to reduce the influence of accent in the final model.

4.1 Related work

Accent normalization has very little prior work in ASR, however robust ASR models to compensate for other variations such as noise, channel, gender, etc. have been investigated in the past. The normalization can be performed at the feature-level or model-level. At the feature-level, front-ends such as PLP [HJ91] and RASTA [HM94] have been proposed earlier. Probabilistic front-ends based on Multi-Layer Perceptron (MLP) have also been tested for their noise robustness [IMS+04]. A review of feature-based and model based techniques for noise robustness in speech

recognition is presented in [Den11, Gal11a]. The idea behind the design of noise-robust features is that these front-ends are independent of the noise conditions, while still maintaining the discrimination in the phonetic space. Thus, when trained on datasets with multiple noise conditions, the ensuing models are unaffected by these variations. In a similar manner, we seek to evaluate different front-ends based on their robustness to different accents.

Accent adaptive training has mainly involved techniques borrowed from multilingual speech recognition. They include simple data pooling based multi-style training, using accent-tags in the phonetic decision tree for data sharing [Che01, CMN09, KMN12] and using individual distributions while sharing the codebooks [KMN12]. [SK11] introduced stacked transforms, a two-level MLLR transforms to integrate accent and speaker adaptation, similar to factorized CMLLR proposed in [SA11]. As in normalization, accent adaptive training has also commonalities with speaker [Gal11b] and noise [KG09] adaptive training.

4.2 Accent normalization or robustness

We focus on seeking robust features that will ensure accent-independent acoustic models when trained on datasets with multiple accents. We formulate a framework which can be used to evaluate different front-ends on their ability to normalize the accent variations. We use ASR phonetic decision trees as a diagnostic tool to analyze the influence of accent in the ASR models. We introduce questions pertaining to accent in addition to context in the building of the decision tree. We then build the tree to cluster the contexts and calculate the number of leaves that belong to branches with accent questions. The ratio of such 'accent' models to the total model size is used as a measure for accent normalization. The higher the ratio, the more models are affected by the accent, hence less normalization and vice versa.

4.2.1 Decision tree based accent analysis

Phonetic decision trees have been traditionally used in ASR to cluster context-dependent acoustic models based on the available training data. The number of leaves in a phonetic decision tree refers to the size of the acoustic model. In our training process, the decision tree building is initialized by cloning the CI models to each available context in the training data. Two iterations of Viterbi training are performed to update the distributions while the codebooks remain tied to their

respective CI models. Several phonetic classes of the underlying phones such as voiced/unvoiced, vowels/consonants, rounded/unrounded, etc are presented as questions, to the decision tree algorithm. The algorithm then greedily chooses the best question at each step that maximizes the information gain in a top-down clustering of CD distributions. The clustering is stopped once the desired model size is reached or when the number of training samples in the leaves has reached the minimum threshold.

In this framework, we combine questions about the identity of accents with contextual questions and let the entropy-based search algorithm to choose the best question at each stage. The resulting decision tree will have a combination of accent and contextual phonetic questions. An example is shown in Figure 4.1. As shown in the figure, the begin state of phoneme /f/ is clustered into 4 contexts. f-b(1) and f-b(2) are considered accent-dependent contexts, as they are derived by choosing a accent question (Is current phone belong to IRAQI accent?). f-b(3) and f-b(4) are accent-independent contexts, because their derivation does not involve a accent question in the decision tree. The earlier the question is asked, the greater its influence on the ensuing models. In the above tree, a robust front-end should push the accent questions as low as possible in the tree, so only a few models are influenced by them. Hence, the ratio of accent leaves to total model size is used as an estimate to evaluate MFCC and MLP front-ends. We build a decision tree using the combined set of questions. For each leaf node, we traverse the tree back to the root node. If we encounter a accent question in a node, then that leaf is assigned as a accent-dependent model. The ratio of accent-dependent to total leaves is then calculated. The experiment is repeated by varying the model size.

4.3 Accent aware training

In this section, we try to leverage accent labels on each speaker available during training of ASR models. Accent labels can be obtained by eliciting from speakers during enrollment or determined automatically using auxiliary information such as geographic region or even manually labelling a small number of speakers. We introduce the accent label as an additional feature in the training of bottle-neck neural network. The accent is encoded as a 1-hot vector to the input of the neural network and weights of these augmented features are trained along with other feature weights during back-propagation. We compare the performance of the accent-aware BN features to the one that is agnostic to the accent of the speaker.

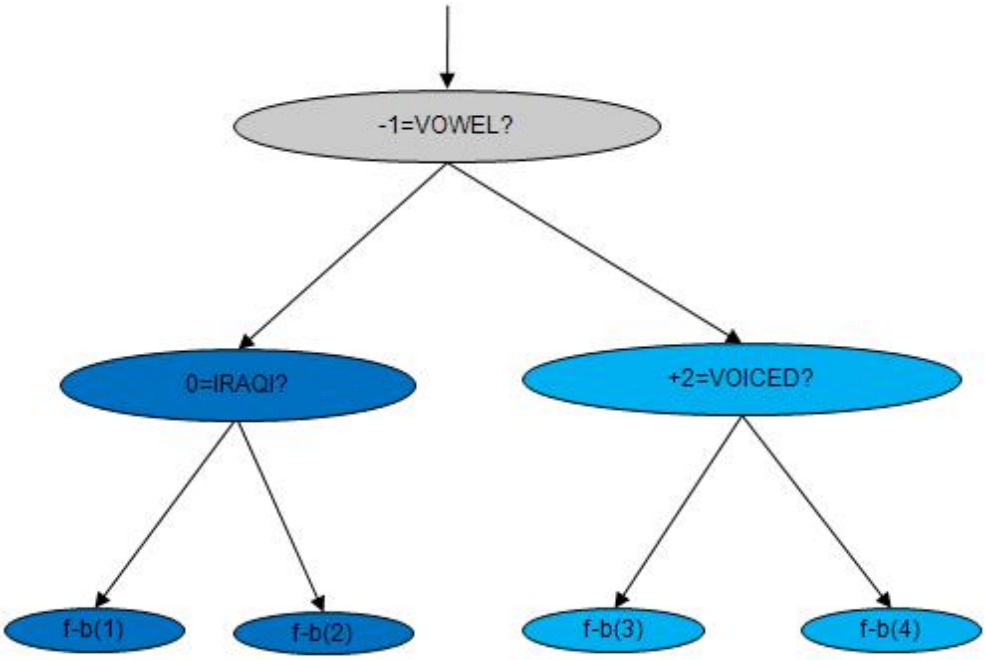


Figure 4.1: *Decision tree for begin state of /f/*

We also compare the influence of the accent in accent-aware Vs accent-agnostic BN front-ends using the decision tree based analysis described in the section above.

4.4 Experiments

Our experiments are carried out on the Pan-Arabic dataset provided by AFRL and medical transcription by M*Modal. We analyze the influence of accents in two different front-ends, MFCC and MLP. The following sections describe the experiments carried out in each setup.

4.4.1 RADC setup

Dataset

RADC database consists of Arabic speech collected from regional Arabic speakers, corresponding transcriptions and lexicons for 5 different accents - United Arab Emirates (UAE), Egyptian, Syrian, Palestinian and Iraqi. It is a balanced data set with approximately 50 recording sessions for each accent, with each session comprising two speakers. The amount of data broken down according to accent is shown in Table 4.1 below.

Table 4.1: *PanArabic Dataset*

Dataset	Num. Hours
UAE (AE)	29.61
Egyptian (EG)	28.49
Syrian (SY)	28.51
Palestinian (PS)	29.29
Iraqi (IQ)	24.92
Total	140.82

Each speaker is recorded on separate channels, including long silences between speaker-turns. Hence the actual conversational speech in the dataset amounts to around 60 hours. The transcriptions of the speech are fully diacritized and included both UTF8 and Buckwalter representations. The first 5 sessions in each accent are held out and used as test data, while the remaining form the training set. The

database also contains accent-specific pronunciation dictionaries. All the accents have a common phone set, except for one minor variation. UAE, Egyptian and Iraqi have the voiced postalveolar affricate, /dZ/ phone. Palestinian and Syrian have the voiced post-alveolar fricative, the /Z/ phone instead. These two phones are merged into one, while designing the ASR phone set. The final phone set contains 41 phones, including, 6 vowels, 33 consonants in SAMPA representation plus a noise and a silence phone.

Baseline

The baseline ASR is trained on speech data pooled from all five accents. The individual, accent-specific dictionaries are merged to form a single ASR dictionary which contains pronunciation variants derived from each accent. The total vocabulary size is 75046 words with an average of 1.6 pronunciations per word. The language model is a 3-gram model trained on the training transcriptions and Arabic background text, mainly consisting of broadcast news and conversations. The OOV rate of the LM on the test data is 1.8%. The perplexity of LM on the test set is 112.3.

We trained two sets of acoustic models based on MFCC and MLP features. For MFCC features, we extract the power spectrum using an FFT with a 10 ms frame-shift and a 16 ms Hamming window from the 16 kHz audio signal. We compute 13 MFCC features per frame and perform cepstral mean subtraction and variance normalization on a per-speaker basis. To incorporate dynamic features, we concatenate 15 adjacent MFCC frames (7) and project the 195 dimensional features into a 42-dimensional space using a Linear Discriminant Analysis (LDA) transform. After LDA, we apply a globally pooled ML-trained semi-tied covariance matrix. For the development of our context dependent (CD) acoustic models, we applied an entropy-based, poly-phone decision tree clustering process using context questions of maximum width 2, resulting in quinphones. The system uses 2000 states with a total of 62K Gaussians with diagonal covariance matrices assigned using merge and split training. The total number of parameters in the acoustic model amounted to 7.8M.

In addition to MFCC system, we trained another set of acoustic models using MLP Bottle-neck features [GF08, FWK08]. A multi-layer perceptron is trained using ICSI's QuickNet MLP package [Qui]. We stack 7 MFCC frames, which serve as input to the MLP. The context-independent (CI) state labels are used as targets. The MLP has a 4-layer architecture - input (195), 2 intermediate (1000, 42) and output (125) layers, with a total of 243,292 parameters. The training data for the MLP

is derived from the ASR training set, 90% of the training speaker list is used for training MLP while the remainder 10% of the speakers is used as a development set. For each training iteration MLP's accuracy on the development set is calculated. The training is stopped when the accuracy saturates on the development set. In our case, MLP training took 5 epochs and reached a frame-level accuracy of 63.86% on the training data and 63.56% on the development data. The activations in the third layer, also called the bottle-neck layer [GKKC07] are used as inputs to build GMM-based HMM acoustic models. Apart from MLP parameters, the MFCC and MLP acoustic models used same number of parameters. The baseline Word Error Rate (WER) for the MFCC and MLP system is given in Table 4.2 below. The WER of MLP ASR system is 0.6% (absolute) lower than the MFCC system. The speaker adapted system produces a WER of 26.8%

Table 4.2: *Baseline Performance.*

Accent	Baseline ASR	
	MFCC	MLP
AE	28.7	28.2
EG	30.0	29.5
SY	27.9	27.2
PS	29.4	28.6
IQ	27.7	27.0
Average	28.7	28.1

Accent robustness

In the first experiment, we examine the influence of accent in MFCC front-end. Table 4.3 summarizes the accent analysis for different model sizes.

We observe that speaker adaptation, including vocal tract length normalization (VTLN) and feature space adaptation (FSA) training, only marginally reduce the influence of accent ($\approx 0.5\%$ absolute) in the acoustic models. In the resulting decision trees, we observe that the /Z/ appears very early in the split. This is the phone we merged from /dZ/ and /Z/ that belongs to two different accent classes. accent questions in the decision tree allowed the phone to split into its accent counterparts. The distribution of different accents for each model size is shown in Figure 4.2.

Table 4.3: *Ratio of accent nodes in MFCC decision tree.*

Model Size	Accent Nodes	Non-Accent Nodes	Ratio
	MFCC		
1000	13	987	1.3%
2000	82	1918	4.1%
3000	224	2776	7.5%
4000	483	3517	12.1%
	MFCC (VTLN + FSA)		
1000	9	991	0.9%
2000	72	1928	3.6%
3000	226	2774	7.5%
4000	465	3535	11.6%

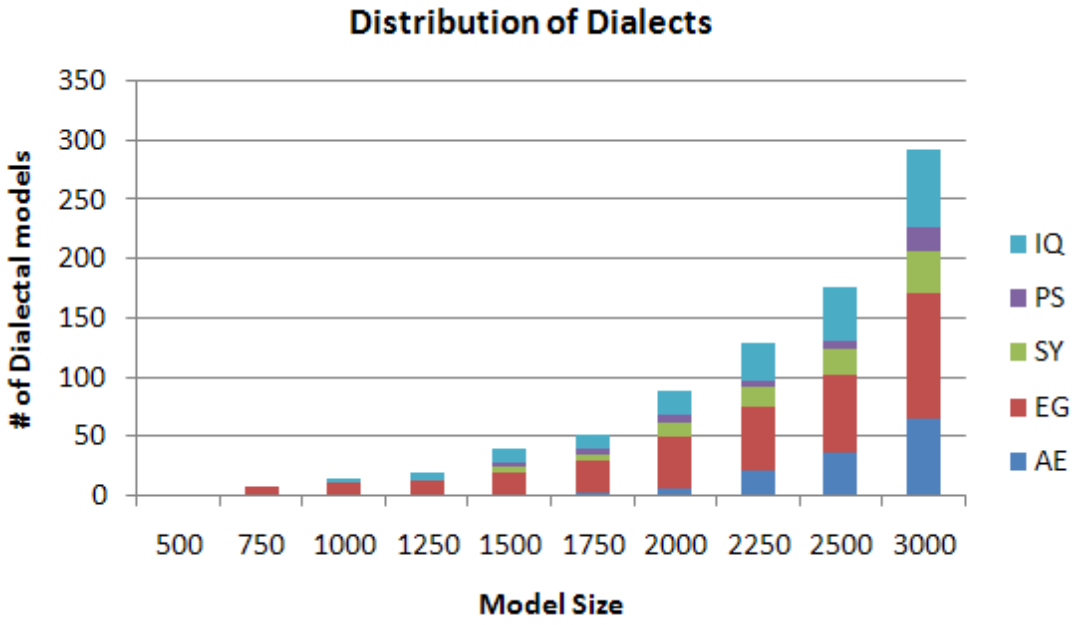


Figure 4.2: *Accent Distribution in MFCC models*

We noticed that most accent models belong to Egyptian across different model sizes. This behavior is consistent with the results found in the literature, where Egyptian is found to be most distinguishable from other accents [BHC10]. We also observed that vowels are more influenced by accent than consonants. Table 4.4 shows the ratio of accent models to all clustered models for vowels and consonants. Except for the case of model size 1000, vowels have more accent models and hence more accent influence, than consonants. This result is in line with the fact that the majority of differences between Arabic accents are characterized by vowels. These observations indicate that decision trees can be used as an effective analytic tool to evaluate the effect of different accents in acoustic models.

Table 4.4: *Ratio of accent models for vowels and consonants.*

Model Size	Accent models	Ratio of Accent Models	
		Vowels	Consonants
1000	13	1.1%	1.4%
2000	82	6.2%	2.9%
3000	224	10.8%	5.4%
4000	483	17.1%	8.8%

MFCC vs. MLP accent analysis

In this section, we examine the influence of accent in MLP and MFCC front-ends. The number of accent models for MLP and MFCC systems is shown in Figure 4.3. From the graph, it can be seen that speaker adaptation marginally reduces the influence of accent in the final models, in both MFCC and MLP. Comparing, the two front-ends, MFCC has less accent models than MLP for all cases.

To confirm the hypothesis that MLP features are more sensitive to accent, we created a more rigorous setup. The pilot experiment used a combined dictionary obtained by composing individual, accent-specific dictionaries. The use of different "accent" pronunciation variants can render the models to be insensitive to accent variations. Hence, in our next experiment, we constrained the dictionary to have only one pronunciation for each word. The training data is force-aligned with the combined dictionary and the most frequent pronunciation variant is selected for each word, which is the only variant used in the experiment. Also, in the previous experiment only singleton accent questions (eg. Is current phone IRAQI?) were used. We experimented with combinations of accent questions in the following

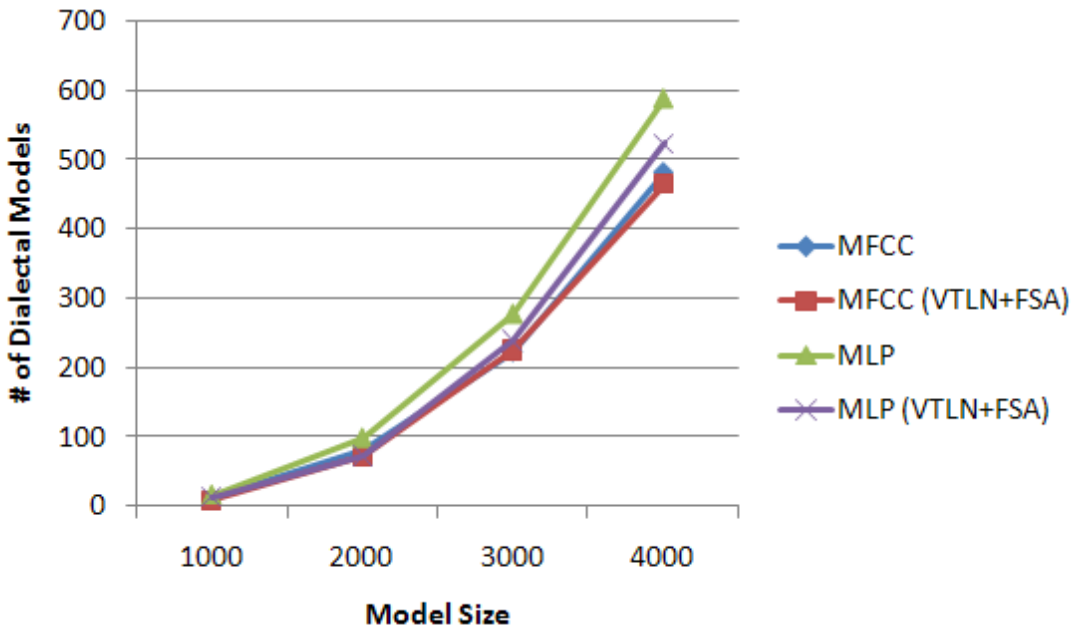


Figure 4.3: *MLP vs. MFCC models*

setup (eg. Is current phone IRAQI OR EGYPTIAN?). This would allow more accent questions to be available for clustering. Figure 4.4 shows the results of the new setup. It can be observed that more MLP models are influenced with accent than in the case of MFCC. These results show that MLP features are more sensitive to linguistic variations, i.e. accent. We also note that similar framework has been used for gender analysis and we find that both MLP and FSA based speaker adaptation greatly reduce the influence of gender in the clustered models.

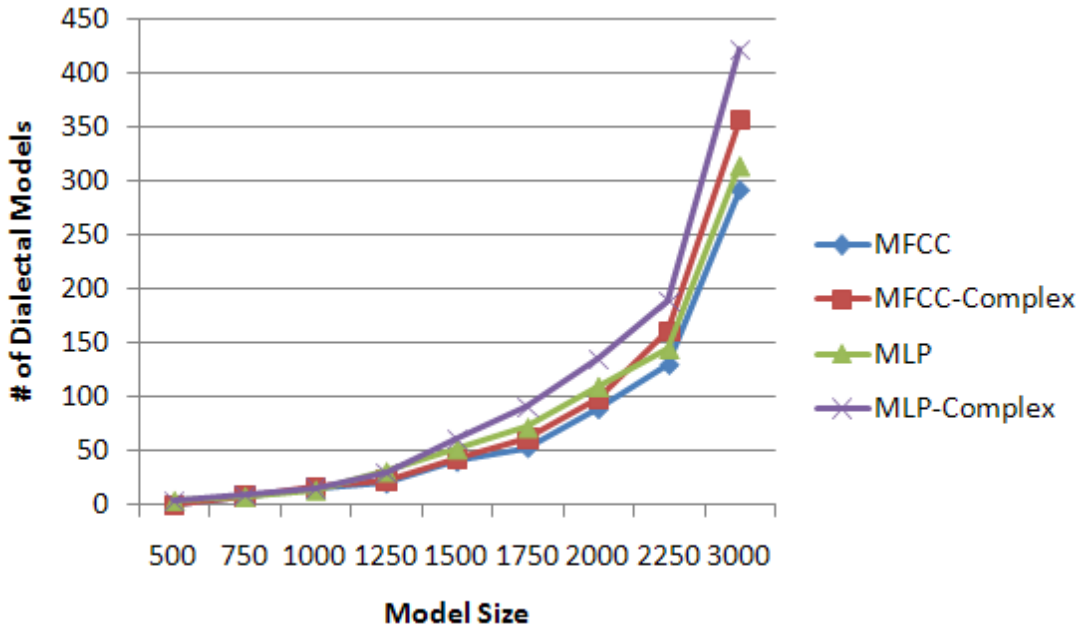


Figure 4.4: *Single Pronunciation models*

To analyze the accent sensitive behavior of MLP, we calculated the frame-level accuracy of vowels and consonants in the MLP outputs on the development set. The average accuracy for vowels and consonants is shown in Table 4.5.

It is clear from Table 4.5 that MLP frame level accuracy is higher for vowels than consonants. We already observed that accented models are dominated by vowels, which indicates that most accent variations occur in vowels in Arabic. Hence, we hypothesize that the low MLP frame accuracy for vowels, rendered MLP to be more sensitive to accent variations.

Table 4.5: *MLP frame accuracy for Vowels and Consonants.*

Phone Class	MLP Accuracy
Vowels	26.41
Consonants	40.80
Noise/Silence	85.78

4.4.2 M*Modal setup

The dataset and baseline for the M*Modal medical transcription setup has been explained in chapter 3. In this section, we evaluate the accent robustness of MFCC and Neural-network bottle-neck features on accented English.

Dataset and baselines

The 1876 training speakers in the M*Modal data set were manually classified into one of 8 different accents. Table 4.6 shows the breakdown of the training speakers interms of their accents.

Table 4.6: *Accent distribution in M*Modal training set.*

Accent	# Speakers	Ratio (%)
US	1397	74.4
South Asian	168	9.3
East Asian	106	5.6
Hispanic	93	4.9
Middle Eastern	52	2.7
African	24	1.2
Eastern European	17	0.9
Western European	11	0.5
Others	10	0.6

The MLP system included a 4-layer bottle-neck MLP trained using frame-level cross-entropy with CI targets. Table 4.7 gives the WER of MFCC and MLP based systems. We use two test sets as before, South Asian and US Native English.

Table 4.7: *WER of MFCC and MLP models* .

System	WER (%)	
	Native	South Asian
MFCC	45.73	19.89
MLP	42.07	28.75

Accent analysis

Accent questions are introduced in the decision tree build process as in Arabic system. The accent models for different model sizes are calculated and shown in Figure 4.5. The figure shows very minimal difference between MFCC and MLP based systems.

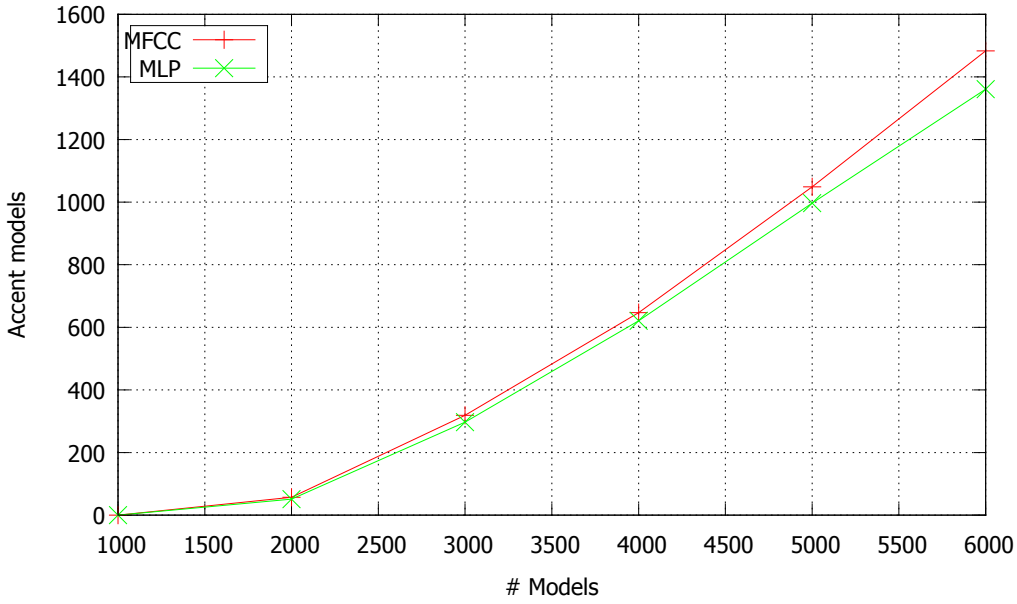


Figure 4.5: *MFCC Vs MLP models*

Given the recent interest in deep architectures for neural networks in ASR [HDY+12], different Neural networks are trained to measure the robustness of features with respect to accents. Table 4.8 shows 3 different NN setups. The first system is trained with CI targets and 4 layer MLP as discussed in the previous experiment. The second one has the same number of layers, but trained with CD

labels. The 3000 context-dependent states are reduced to 1000 CD states by pruning the decision tree. The alignments are still produced by the 3000 state system, while they are mapped to 1000 CD states before NN training. The last system is a ”“deep”” NN with 7 layers and Bottle-neck layer in the 5th layer. All NN models have the same number of 1M parameters. The hidden layers are adjusted accordingly to achieve the desired parameter size. From the table, it is clear that Deep Bottle-Neck Features (DBNF) perform the best with WER close to MFCC sMBR system.

Table 4.8: *WER of MFCC and DBNF models* .

Target	# layers	WER (%)	
		South Asian	Native
CI	4	42.07	28.75
CD	4	41.52	28.21
CD	7	34.75	24.96

The corresponding robustness analysis plot for the 3 systems are shown in 4.6. The graph clearly shows that DBNF can do better normalization of accents than MFCC and other 4 layer NNs.

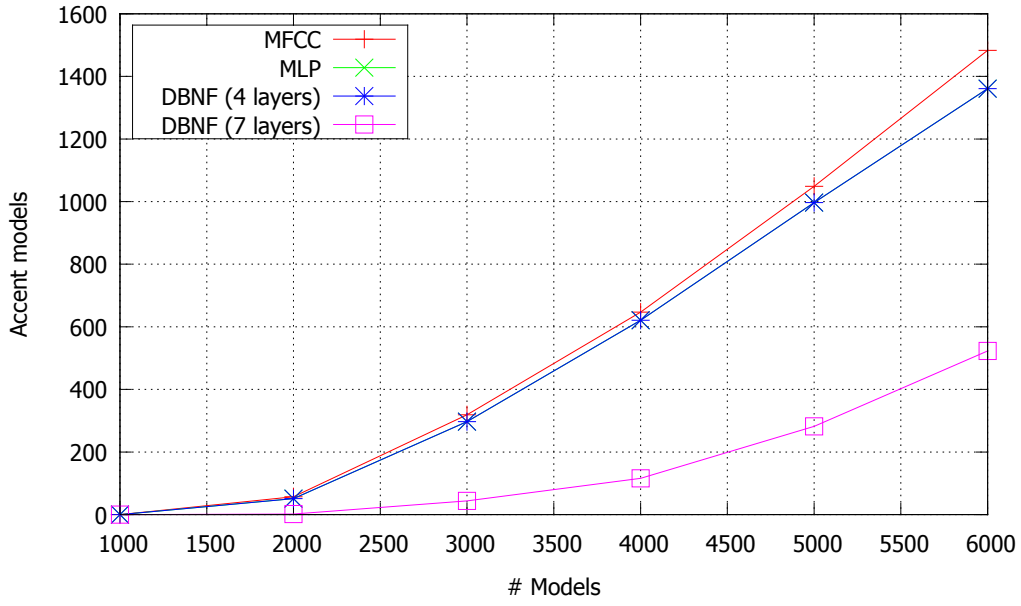


Figure 4.6: *MFCC Vs DBNF models*

The effectiveness of accent normalization at different layers of DNN is explored in Table 4.9. The table shows that the deeper layers provide better accuracy and better accent normalization. Figure 4.7 shows accent robustness at different layers of the DNN.

Table 4.9: *WER of varying bottle-neck layer* .

Bottle-neck Layer #	WER (%)	
	South Asian	Native
3	42.10	28.86
4	37.33	25.87
5	35.74	24.96

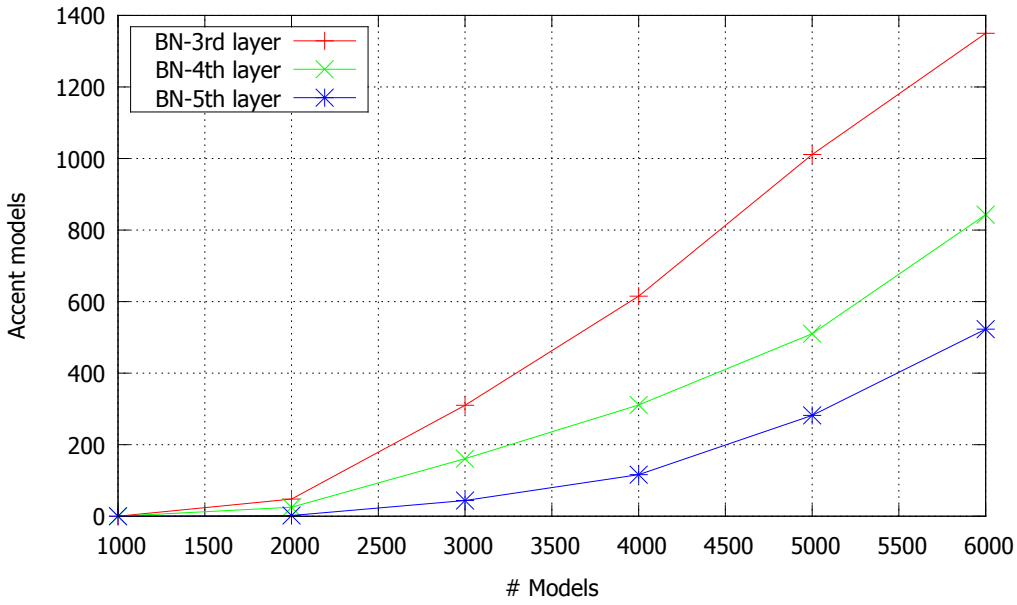


Figure 4.7: *Robustness analysis of Bottle-neck layers*

Accent aware training

In this experiment, we augment the input spectral features to the BN neural network with the accent label of each training speaker. In the M*Modal dataset, each speaker is assigned a label corresponding to one of 9 accent groups. The baseline neural

network has an input dimension of 150 while the accent-aware neural network has 150 + 9 = 159 input features. Their weights are learnt along with other parameters of the BN network during back-propagation. During testing, the accent of the test speaker is added to the input to generate their BN features. The performance of the accent-aware BN features is shown in Table 4.10.

Table 4.10: *WER of accent-ware BN features.*

Target	# Training Hours	WER (%)	
		South Asian	Native
Accent agnostic	150	35.74	25.41
Accent-aware	150	33.38	24.21
Accent agnostic	450	34.75	24.96

The results show that the accent-aware BN front-end has 6.6% lower relative WER than accent-agnostic BN front-end for South Asian speakers and 4.7% lower relative WER for Native US speakers. The table also shows the benefit of labeling a subset of the training data with accent labels. Accent-aware model trained on 150 hours of training data with accent annotations performs better (33.38% Vs 34.75% and 24.21% Vs 24.96%) than accent-agnostic model trained on 450 hours of speech data. We also use decision tree framework to analyze the influence of accent in accent-aware Vs accent agnostic front-ends. Figure 4.8 shows that the accent-aware model is more robust to accent variations than the accent agnostic model.

4.5 Summary

We have presented an evaluation framework to test different front-ends for their ability to normalize accent variations. We analyzed MFCC and MLP front-ends and showed that decision tree based accent robustness is an effective tool to measure tolerance to accent variations. Various architectures of MLP are analyzed and their robustness towards accent variations is evaluated. It is shown that Deep neural networks with bottle-neck layer close to the output layer produces lower WER and better accent normalization criterion. We also show that annotating a subset of speakers with accent labels and using them in an accent-aware training reduces the WER over an accent agnostic ASR model trained on a larger dataset.

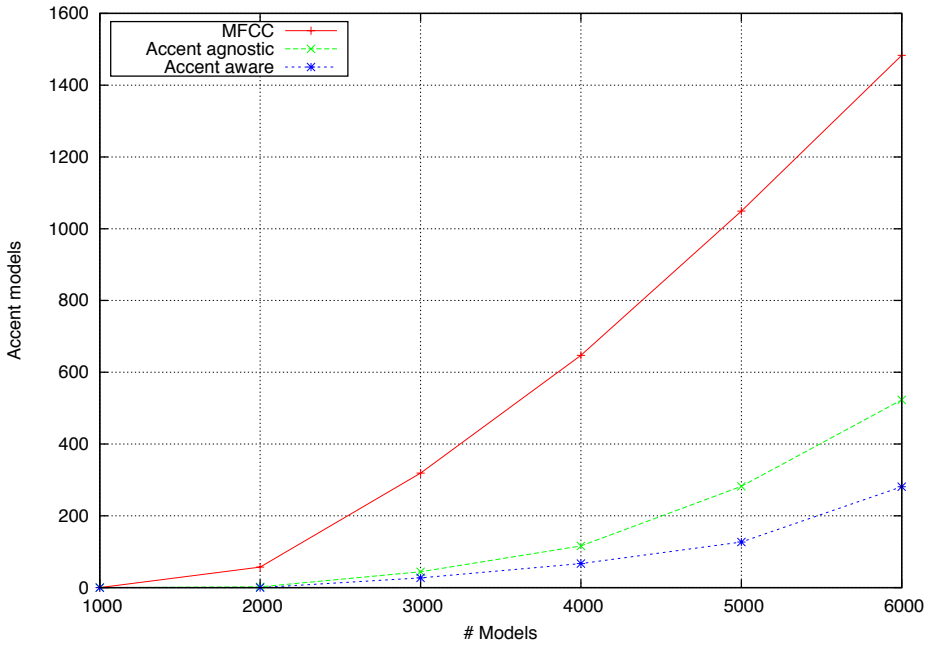


Figure 4.8: *Robustness analysis of accent aware BN front-end*

Part IV

Speaker dependent modeling

Chapter 5

Speaker dependent ASR and accents

Speaker dependent (SD) models trained with sufficient data from a target speaker perform significantly better than speaker independent (SI) models trained by pooling data from multiple speakers. They form an important class of ASR models, widely employed in applications characterized by personal and sustained usage over an extended length of time. Traditionally, this use case was limited to dictation tasks such as medical or law transcription services where a single user can potentially dictate over 100s of hours of reports throughout his/her career on a personal computer. However, with the recent adoption of smartphones and other handhelds, speech interfaces are becoming more pervasive outside of desktop computers. These devices also sport a large resident memory and increased computational power to accommodate SD ASR models customized for the target user. This section of the thesis is devoted to analyzing the influence of accents in SD ASR and addressing some of the issues faced by accented speakers through speaker-based data selection techniques.

5.1 Motivation for SD models

Speaker specific characteristics such as age, gender, vocal-tract length and accent have significant influence on the ASR acoustic models. Speaker-independent systems trained by pooling data from wide range of speakers perform poorly for speakers whose features are under-represented in the training set, e.g. non-natives, female speakers, etc. Speaker-adaptive systems handle these variations by mapping the input signal to a canonical space using various normalization [ZW97] and adap-

tation techniques [Gal98] and training the ASR models in this new feature space. Speaker-dependent systems on the other hand, address this issue by fitting the ASR models on the target speaker's acoustic space using increased training data specific to the speaker. This is achieved by either training SD models solely on the target speaker or adapting an SI model using speaker-specific data. The following graphs show the effectiveness of 1 hour of target speaker data on ASR performance.

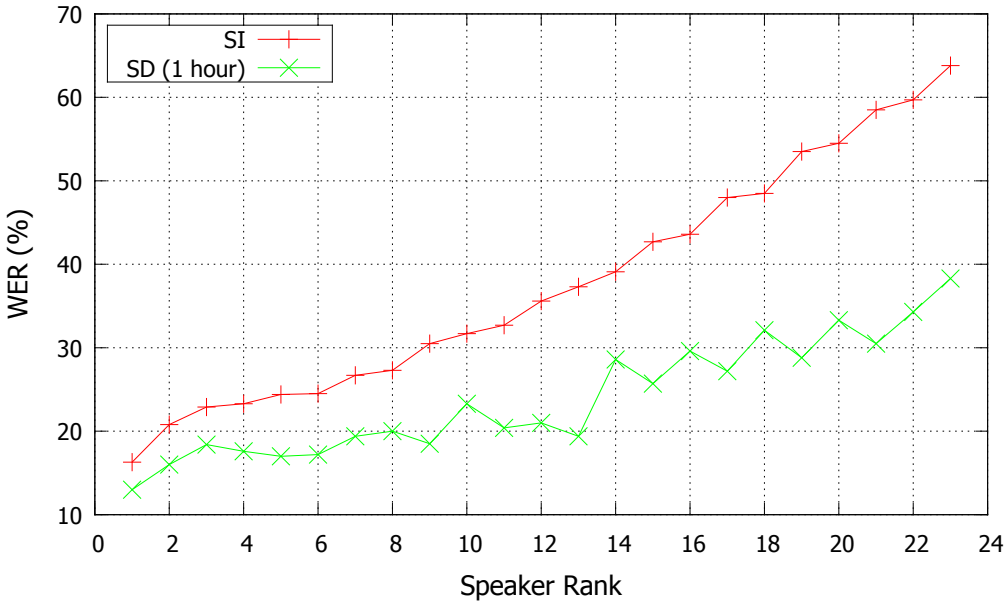


Figure 5.1: WER of SI Vs 1-hour SD models

Figure 5.1 shows test speakers from the two test sets discussed in 2.7.1 arranged in ascending order of their SI WER. The graph also shows the WER of the same speakers after adapting the SI models using 1 hour of speaker-specific data using MAP adaptation. It is clear from the graph that there is significant improvement across all speakers with SD data. More importantly, the speakers at the tail of the WER curve improve a lot (5̃0% relative) compared to the speakers who have a low SI WER. Hence, SD systems trained on data from a specific speaker have the potential to increase the reach of ASR to wider range of speakers, who would otherwise find a generic ASR unusable. This is also an important advantage of SD models over SI counterpart in the context of unsupervised training. Given the availability of large amounts of speech data in many applications, modern day ASR systems employ a combination of first pass hypothesis and confidence based data selection for training. The speakers with high SI WER will have erroneous transcriptions or get filtered

during data selection due to low confidence and hence are unlikely to have lower WER with SI models even if more data becomes available. SD systems on the other hand, can produce significant improvements by adapting the SI model using less speaker-specific data.

SD models, by virtue of being trained on the target speaker are sharper than the SI models. They are more accurate, so they can effectively prune unlikely hypotheses early in the decoding resulting in faster run-time performance. This is crucial for ASR applications that require strict real-time performance on resource-constrained environments such personalized assistants. Figure 5.2 plots WER and real-time performance of SI and SD models for the same speakers in Figure 5.1.

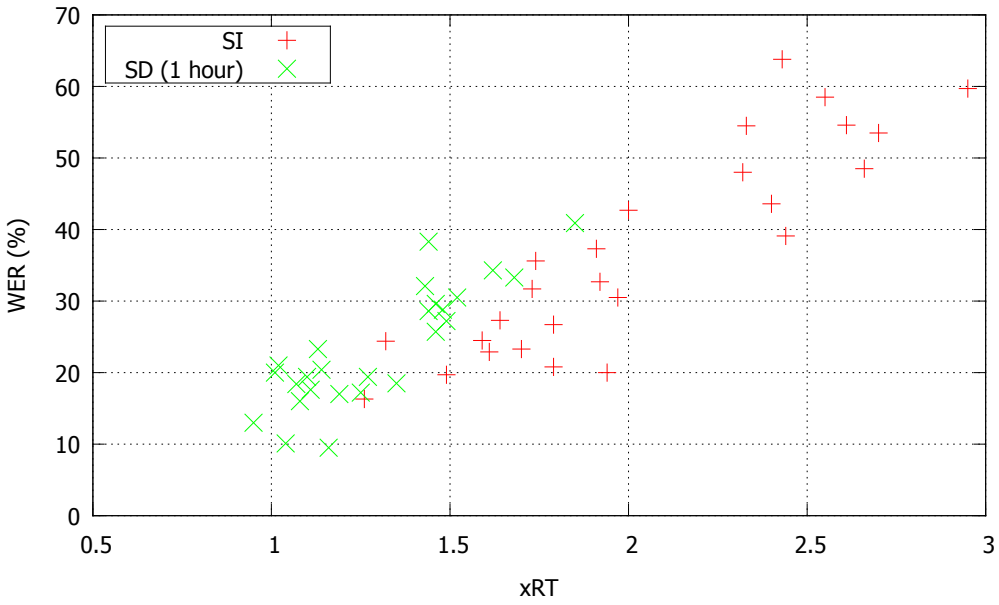


Figure 5.2: WER of SI Vs 1-hour SD models

The Real-Time Factor (RTF) is calculated for the same beam and other pruning settings in the decoder for both the models. The graph shows that SD models are 2-3 times faster than SI models across a wide range of WERs. If there is a real time constraint on the ASR, more aggressive pruning has to be employed for SI models which will result in further degradation of WER. SD models on the other hand, are already close to 1xRT constraints and thus require less aggressive or no pruning. These characteristics ensure even less powerful SD models residing on the client to perform better for the target speaker than powerful SI models hosted on the server.

5.2 Accent issues in SD models

Although it has been established that SD models perform significantly better compared to an SI ASR, they require sufficient data from the target speaker, which is time-consuming to collect. In a real-world task, the user starts off with SI models out of the box and the system adapts to the speaker's data with continued usage. This transition period from SI to a significantly better speaker-specific model with data collected over time, is termed as the adaptation phase. A new speaker has to painfully navigate through this adaptation phase with a low accuracy SI system until he/she has produced sufficient data for adapting the initial model. Customer satisfaction and adoption rates for commercial ASR systems suffer significantly throughout this transition period.

This issue is more serious for accented speakers who encounter significantly higher word error rates with SI system compared to native speakers. Figure 5.3 shows the same xRT Vs WER graph, but with additional accent information. It is clear that accented speakers start off nearly 50% relative worse than native speakers. After an hour of SD data, there is significant improvement in both the WER and RTF for these speakers. However, during this period, they have to rely on inaccurate SI system, which is almost unusable for applications such as dictation.

This section of the thesis aims to address the problem of building reliable speaker-dependent models using limited target speaker's data to reduce the adaptation phase for accented speakers. The performance of the adaptation algorithms will be measured on both accented and native speakers to analyze the influence of accent.

5.3 Related work

Speaker adaptation has a long history in ASR with popular techniques such as Maximum A-Poteriori (MAP) adaptation [GL94b], Maximum Likelihood Linear Regression (MLLR) [LW94] and Constrained-MLLR (CMLLR) [Gal98]. All of these techniques are confined to the available adaptation data, which is only a few minutes in our case. Several techniques have been exploreed in the literature to address the problem of speaker adaptation with limited data. They can be classified into 3 groups: subspace based adaptation, speaker clustering and speaker selection.

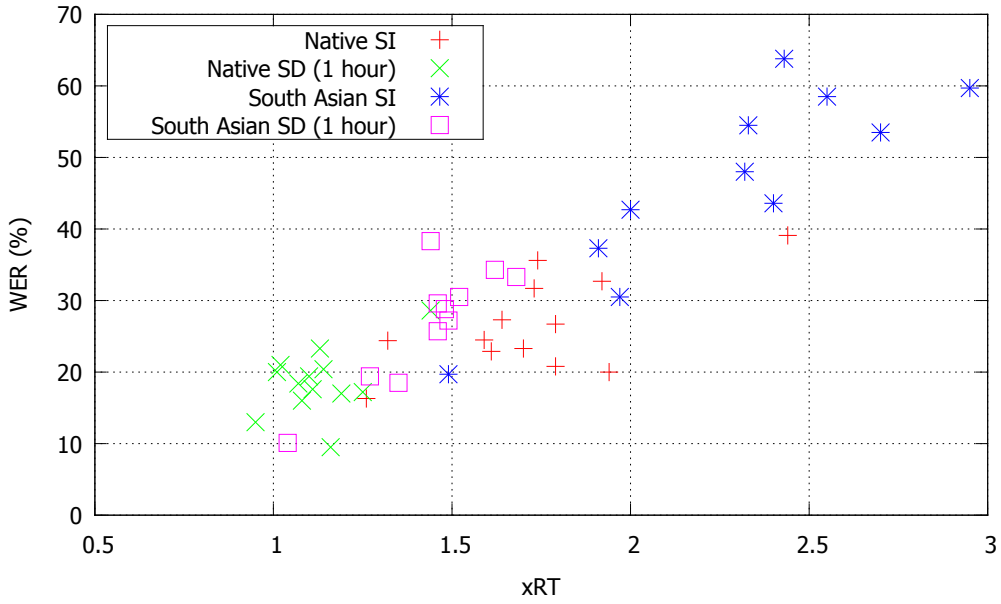


Figure 5.3: WER of SI Vs 1-hour SD models

5.3.1 Subspace based speaker adaptation

Motivated by the success of Eigen faces in face recognition task, [KNJ+98] introduced Eigen voices for speaker adaptation. It involves identifying a subspace of model parameters that best explains the speaker variability in the training data using fewer parameters. Principal Component Analysis (PCA) was performed on speaker-dependent mean supervectors to project them onto a low-dimensional subspace. The first few eigen vectors are chosen as representative dimensions or Eigen voices and speakers are represented as a weighted combination of these basis vectors. During adaptation, the weights for the new speaker in this low-dimensional speaker space can be efficiently estimated on adaptation data using Maximum Likelihood Eigen Decomposition (MLED). [NWJ99] proposed maximum likelihood based instead of PCA to directly estimate the low-dimensional basis vectors referred to as Maximum Likelihood Eigen Space (MLES). It used traditional Baum-Welch like iterative estimation and the technique also supported Bayesian MAP estimation with gender or dialect based priors, known as MAP Eigen Space (MAPES). [MH07] introduced non-linear extensions to Eigen voices using kernel techniques with additional improvements.

An alternative to low-dimensional projection of SD mean vectors is to apply eigen based methods for the estimation of transform parameters. [CLWL00] proposed Eigen-MLLR, which is a low-dimensional estimation of MLLR transform on few seconds of speaker's data. The MLLR matrix space is decomposed into a set of basis eigen matrices using PCA on the training speakers. The weights of these basis matrices are estimated on the adaptation data using maximum likelihood. MAPLR [CLWL00] extended MAP adaptation using eigen-based priors. It used Probabilisitic PCA (PPCA) to derive the eigen voices and during adaptation the transformed model is used as a prior for further MAP estimation. [PY12b] provided a detailed recipe for estimating CMLLR using subspace projection and dynamically varying the basis dimension based on the available adaptation data. Recently, [Bac13] introduced efficient implementation of iVector based speaker adaptation for short utterances using Map-Reduce framework. All of these subspace based techniques provided significant improvements over unadapted baseline, for data between 5 and 15 seconds of speech. The performance saturates to traditional methods after 20 seconds of adaptation data. In the current experiments, the adaptation data is assumed to be few minutes of audio for the target speaker obtained either by initial speaker enrollment or after dictating the first report.

5.3.2 Speaker clustering

In clustering based techniques, the training dataset is grouped into multiple clusters based on some similarity criterion. Model templates are trained for each cluster, reducing the variability in the overall speaker space. During adaptation, the new speaker is assigned to a specific cluster based on the same similarity criterion between the adaptation data and the individual models. The target cluster models are then used to decode the new speaker. [Fur89] introduced hierarchical clustering and adaptation in the context of Vector Quantization (VQ) based ASR. This concept was extended to HMM-GMM based ASR in [SBD95]. A hierarchical cluster tree is built using agglomerative clustering based on relative entropy distance between the training speakers. Instead of training individual models from scratch, only transformation parameters are stored for each cluster. For a new test speaker, top N clusters are chosen based on the relative entropy distance and their parameters are averaged to derive the adapted model. [BVS10] experimented with KL divergence on context-independent gaussian posteriors as a distance measure to build the cluster tree. The hierarchical clustering was initialized with PCA and refined using distributed VQ in a MAP-Reduce framework. Further analysis of the resulting

clusters showed that pitch and loudness had significant influence which showed that the clusters represented different gender and noise conditions.

[Gal00] introduced cluster-adaptive training (CAT), a form of soft-clustering of the training speakers. The mean vectors of the models are represented as a weighted combination of cluster means. The technique is similar to the eigen voices, however both the cluster weights and the resulting clusters themselves are iteratively trained using ML criterion. The eigen voices can also be used to initialize the clusters for CAT training. [Gal01] evaluated various forms of CAT training schemes including model, transform and bias clusters. [YG06] proposed a discriminative version of CAT. The clustering based adaptation techniques are computationally efficient as training the clusters can be done offline and only the cluster assignment is carried out during adaptation. However, it is sub-optimal to precluster the speakers as its difficult to obtain representative clusters for different factors such as age, gender, accent, etc. that influence the acoustic space of each speaker.

5.3.3 Speaker selection

Speaker selection techniques extend the clustering based schemes by considering one cluster for each speaker in the training data. The adaptation phase is then a nearest neighbour selection problem, where the source (training) speakers are ranked based on a similarity score calculated on the adaptation data of the target speaker. It can also be viewed as an instance of exemplar-based technique [SRN+12] at the speaker level. Top N neighbours are selected and their data is used to adapt the SI model to represent the new speaker's acoustic space. Many similarity criteria have been proposed in the literature to compute acoustically similar speakers and adaptation techniques to compute the adapted model. [PBNP98] trained single gaussian source SD models and selected top N neighbours based on their likelihood on the adaptation data. [WC01] used MLLR to approximate source speakers and the likelihood of the adapted models as a similar criterion to the target speaker. [HCC02] experimented with different configuration of source models and similarity scores and concluded that GMM based likelihood score was both accurate and efficient to compute compared to HMM based score and PCA on MLLR matrices. Motivated by the work in speaker recognition using MLLR matrix supervectors, [VRF10] proposed a distance measure adaptation matrices to select similar speakers.

Once the neighbours are selected for a new speaker, several adaptation techniques have been attempted on the neighbours' data. [HP99] experimented with

different ways to compute the SD models, including weighted MLLR means and statistics, MAP adaptation and their combination. [YBM+01] used a simple strategy to store SD accumulators separately which allowed quick re-estimation of the adapted model on the selected speaker. [WC01] used a single-pass re-estimation carried out on the neighbour's data, weighted by the similarity scores. [HCC02] proposed adaptive model combination to compute the adapted model using the parameters of the source models. MLLR and MAPLR have been compared in [VRF10] in supervised and unsupervised setups. They also contrasted the speaker selection approach with eigen voices and showed that its hard to scale the latter to large systems due to significant computational issues in calculating the basis vectors.

In this thesis, the existing ASR models are used to directly compute the similarity score between the source and target speakers. Several variations of this technique are explored and empirically evaluated based on their performance on the target speaker test set. Different parameters involved in the selection including the number of neighbours, size of adaptation data, etc are also investigated in detail.

5.4 Neighbour selection and adaptation

This section of the thesis addresses the challenge of building SD models by automatically selecting acoustically similar speakers to the target accented speaker, or *neighbours* from a large and diverse set of existing users with large amounts of training data. A few minutes of the speaker's data is used to select the neighbours, so the adaptation can be performed sooner than waiting for sufficient data from the user. The neighbours' data is utilized to build an initial SD system for the target speaker. Such a neighbours initialized SD system performs significantly better compared to the baseline SI models, thus helping to reduce the adaptation interval for the accented speakers. Several speaker selection and adaptation techniques are investigated in this section. Chapters 6 and 7 deal with using acoustic data to select the target speaker's neighbours. The neighbour selection and adaptation using maximum likelihood criterion are explained in 6. Chapter 7 investigates experiments with discriminative methods including State-level Minimum Bayes Risk (sMBR) training and Deep Bottle-neck Features (DBNF). Chapter 8 explores the use of text data in selecting neighbours for improved performance.

Chapter 6

Maximum likelihood neighbour selection and adaptation

This chapter explores building SD models using only a few minutes of adaptation data using neighbour selection and adaptation under maximum likelihood (ML) criterion. Various parameters involved in selection including the number of neighbours, the amount of adaptation data, etc are empirically evaluated. The influence of accent in automatic neighbour selection is analyzed and the automatic selection is compared to accent adaptation using manual annotations. Finally, unsupervised adaptation is explored in the context of both neighbour selection and adaptation.

6.1 Neighbour selection

The goal of the neighbour selection is to identify a group of speakers who are acoustically close to the target speaker using few mins of adaptation data. Two different speaker selection techniques are studied in this chapter - likelihood based and transformation based. The likelihood based approach aims to find source speakers in the training set who are *close* to the given target speaker. It is performed using the following steps:

- The SI model is adapted to each of the source speakers in the database.
- The resulting source SD models are used to calculate the likelihood of the target speaker's data, which serves as a similarity score between source and

target speakers. Given a source model λ_S for the source speaker S, adaptation utterances U_T and their reference transcriptions W_r, for the target speaker T, the similarity score $SS_T(S)$ is calculated as

$$SS_T(S) = \sum_{u \in U_T} \log P(O_u, W_r | \lambda_S) \tag{6.1}$$

- The training speakers are ranked based on their likelihoods and top N speakers are selected for target speaker adaptation.

The transformation based approach attempts to choose neighbours who can be *transformed* into the target speaker. It follows the likelihood based technique in computing the source SD models. In addition, the SD models are adapted on the target speaker's data before calculating the likelihood score. The similarity score in transformation based approach is given by

$$SS_T(S) = \sum_{u \in U_T} \log P(O_u, W_r | f_T(\lambda_S)) \tag{6.2}$$

where $f_T(\lambda_S)$ is the source model adapted on the target speaker's data. A regression-tree based MLLR [Gal96] is used as the transformation function f_T in the experiments. The source speakers are ranked as before for the selection. The final SD model will be calculated by adapting the source model on the neighbours' data. Hence any mismatch between the source and target speakers, that can be modeled by linear transformations, e.g. channel variations can be ignored during neighbour selection. This is achieved by the additional adaptation step before calculating the similarity score.

6.2 Experiments

The M*Modal data set described in section 2.7.1 is used for all experiments in this chapter. Five mins for each speaker in the development data is used to choose the neighbours and the results are evaluated on the test set. As a first step, the SI model is adapted on 5 mins of the development set using regression-tree based MLLR. The number of transforms for MLLR is automatically selected based on the amount of available adaptation data. In the experiments reported, an average of 10 MLLR transforms are used given 5 mins of adaptation data for each target speaker.

Table 6.1 shows the WER of SI and MLLR adapted systems. The MLLR adapted system produces a relative improvement of 10.4% over the SI model. Additional improvements can be obtained by training canonical models using SAT and CMLLR. However, the CMLLR matrices for the test speakers have to be computed on the adaptation data as this is a one-pass dictation system. Such a SAT setup didn't give any significant improvement on top of regression-tree based MLLR adaptation with 5 mins of speaker-specific data in our previous experiments, so SAT is not included in the baseline.

Table 6.1: *Baseline WERs.*

System	Test set	WER
SI	South Asian	45.73
SI + MLLR	South Asian	40.99

In likelihood based selection, the SI model is adapted to the source speaker using MAP adaptation. The likelihood of the target speaker's data on the adapted model is computed. The source speakers are ranked based on their similarity score. In the transformation based technique, an additional regression-tree based MLLR is computed for the source model on the target data before the likelihood computation. 20 neighbours are selected using each criteria. The neighbours are constrained to have at least 15 minutes of speech to ensure sufficient data for adaptation. Once the neighbours are selected, MAP is used to adapt the SI model on the neighbours' data. The neighbour initialized model is then further adapted using MLLR on the target speaker's data. The final target SD models are used to decode the test set. Table 6.2 shows the WER for the likelihood and transformation based selection. The results show that transformation based neighbour selection outperforms the likelihood based approach. It also has 26.3% relative lower WER than the SI and 17.8% relative lower than MLLR adapted baseline.

The setup of creating SD models for each source speaker might seem computationally demanding. However, the neighbours are chosen from a set of existing speakers with large amounts of data. Hence the SD models need to be created only once for all new speakers in the development set, which can be done offline. Moreover these speakers already have SD systems trained for their own dictation. Only their parameters need to be accessed instead of creating source SD models from scratch during selection. Only after the neighbours are chosen for a new speaker, the data of these neighbours are accessed for adapting the SI model. Table 6.3 shows speaker-wise WER for the SI, SI + MLLR and neighbour MAP + MLLR

Table 6.2: *WER for neighbour selection techniques on South Asian speakers.*

System	Selection		WER (%)
	Source	Target	
SI	-	-	45.73
SI + MLLR	-	-	40.99
Likelihood	MAP	-	36.32
Transformation	MAP	MLLR	33.71

systems. The improvements over SI+MLLR are between 10.9% and 31.2% on this test set. This shows that neighbour selection produces improvements for speakers over a wide range of WERs. The following sections will analyze varying these parameters and their influence on target WER. All the experiments from here on will use transformation based neighbour selection.

Table 6.3: *Adaptation WERs for South-Asian speakers.*

Speaker	Test WER (%)			Impr (%)
	SI	SI + MLLR	Neighbour	
1	19.7	15.9	12.9	18.9
2	30.5	27.6	24.6	10.9
3	42.7	37.4	32.6	12.8
4	37.3	30.2	23.3	22.9
5	48.0	44.9	30.9	31.2
6	58.5	52.5	41.4	21.1
7	63.8	56.1	45.5	18.9
8	59.7	55.7	49.1	11.8
9	53.5	48.6	41.6	14.4
10	43.6	41.0	35.2	14.1
Avg	**45.73**	**40.99**	**33.71**	**17.8**

6.2.1 Varying number of neighbours

In this section, the number of neighbours chosen are varied from 1 to 80 and used to initialize the SD system for the target speaker. Table 6.4 lists the WERs of adapting with varying the neighbours. The adaptation step after neighbour selection involves

MAP adaptation on the neighbour data followed by MLLR adaptation on the target data. The respective WERs are listed below.

Table 6.4: *Analysis of varying neighbours.*

Neighbours	WER (%)	
	Neighbour MAP	+ Target MLLR
1	43.21	41.58
5	38.32	34.57
10	36.81	34.18
20	36.49	33.71
40	40.72	36.48
80	42.21	37.81

From Table 6.4, it is clear that 20 neighbours produces the lowest WER. However, neighbours 5 and 10 are very close to the WER of 20 neighbours.

6.2.2 Varying the amount of adaptation data

In this experiment, the amount of target adaptation data is varied to measure its effect on the neighbour selection. It should be noted that, to select different amounts of target speaker data the utterances for each speaker are added to the development set until the desired time in minutes is reached. However, the audio is not excised in the middle of utterance to meet the time limit, so the exact duration will be slightly higher than the expected length. The target data is varied by 1, 2, 5 and 60 minutes for neighbour selection. In each case, 20 neighbours are chosen and adaptation is carried out using neighbours-MAP followed by target-MLLR. Table 6.5 lists the WERs for neighbour selection carried out for different amounts of development data. To add clarity, the exact amount of target adaptation data (averaged across speakers) chosen for each case is listed in the table.

Neighbour MAP WER in Table 6.5 can be used to compare the speaker selection across different amount of adaptation data. The results show several interesting properties. Focusing on the first three rows, better neighbours are obtained with increasing target data. However, the neighbours chosen with just 2.58 minutes perform quite close to the ones selected with 5.82 minutes of target speaker data. This is attractive as most dictation systems perform an enrollment step which guides the new user to read out a few phonetically balanced sentences. The average

Table 6.5: *Analysis of varying adaptation data.*

Target speaker data	WER (%)	
	Neighbour MAP	+ Target MLLR
1.63	37.21	36.02
2.68	36.93	35.17
5.82	36.46	33.87
60.85	36.48	30.40

amount of data collected during the enrollment step is from 2 to 5 minutes which could be used to select the neighbours to build a better SD model, rather than waiting for the speaker to start using the system.

The last row reports results for neighbours selected with 1 hour of target speaker data. It doesn't perform any better than neighbours selected with 5 minutes. This shows that neighbour selection can be performed with just a few minutes and they don't need to be re-selected them as more SD data becomes available. The Target MLLR results for 60 minutes on the other hand is better than 5 minutes due to additional speaker's data available for adaptation and not because of better neighbours.

6.2.3 Varying target speaker's data

In this section, the behavior of SI and neighbours initialized SD models are examined with increasing adaptation data. For each datapoint, both systems are adapted on the chosen amount of adaptation data and evaluated on the test set. The MLLR is calculated on the target data and the transformed means act as a prior model for the ensuing MAP adaptation. The combined adaptation performed better than using MLLR or MAP alone. Figure 6.1 shows the WER plot for SI and neighbour initialized models. The datapoint at zero SD data, refers to the SI baseline. It is interesting to note that, although the neighbours are chosen with only 5 minutes of target speech, the neighbours initialized system continues to perform better than adapted SI model with increased data.

To understand the impact of the neighbour adaptation technique on native speakers, the same experiment was conducted on the test set of 15 US English speakers. As in the South-Asian case, 5 minutes of each speaker is used to select the neighbours. Figure 6.2 shows the WER plots for SI-Init and Neighbours-Init systems

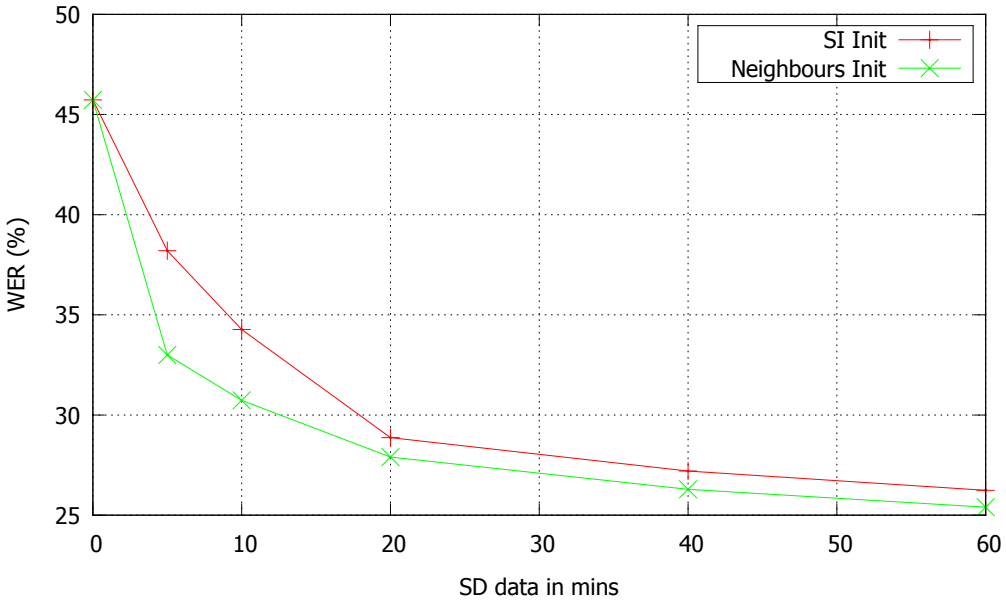


Figure 6.1: WER for SouthAsian speakers

with increasing adaptation data. The same pattern is shows for native speaker as seen with accented speakers. However, the total improvement is less (7% relative at 5 minutes) compared to the South-Asian case (15% relative at 5 minutes), which is expected as the majority of training set data is from native speakers.

6.3 Analysis

This section conducts an analysis of the neighbours selected and the influence of accent and gender in automatic selection is reported.

6.3.1 Influence of gender and accent

The training data is manually annotated with accent and gender labels. For a few speakers without gender labels, it is assigned based on VTLN [ZW97] warp factors. These annotations are used to measure the influence of gender and accent on the neighbours selected for a target speaker. 99% of the neighbours selected match the gender of the target speaker. Hence, it can be concluded that gender has a

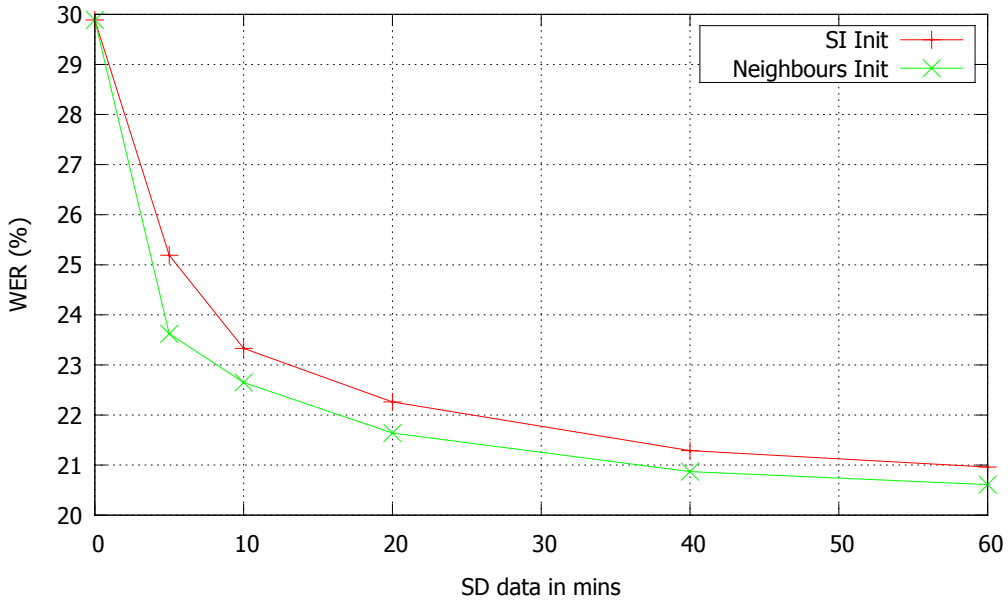


Figure 6.2: WER for Native speakers

decisive impact in the neighbour selection. Figure 6.3 shows the cumulative count of South-Asian and non-South-Asian neighbours in each rank added across all target speakers.

The graph clearly shows that South-Asian speakers are ranked higher than others in the neighbours list. Mann-Whitney U test [MW47], a non-parametric rank test is conducted to verify the influence of accent. 100 ranked neighbours selected for each speaker are grouped into South-Asian and non-South-Asian categories. The test showed significant difference ($p < 0.001$) between the ranks of the two groups, thus confirming accent has significant influence on choosing neighbours.

6.3.2 Automatic selection vs. manual annotations

In this experiment, automatic selection is compared against choosing neighbours based on the manual annotations. There are 168 South-Asian speakers labeled in the training set. Gender and accent labels are used to explicitly choose neighbours that match the target speaker and the WER of the resulting SD models is compared against the automatic selection technique. In all cases, once the neighbours are decided adaptation is performed on neighbours' data using MAP and MLLR on

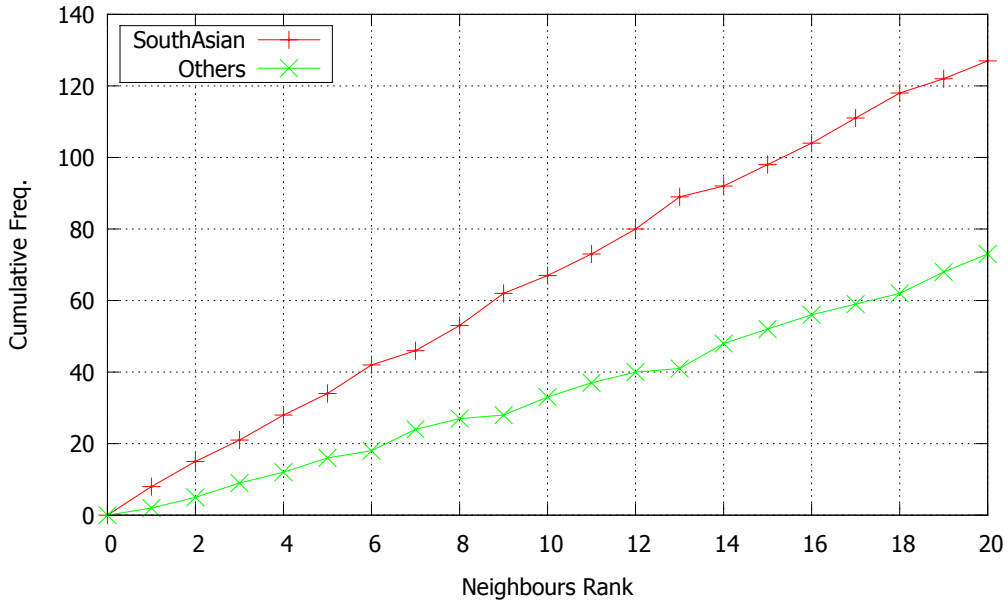


Figure 6.3: WER of SI Vs 1-hour SA models

the target speaker's speech. Table 6.6 shows the WERs of adapted systems on automatically selected neighbours and the ones based on manual labels.

Table 6.6: *Automatic selection Vs. Manual annotations.*

System	Neighbours	Selection	WER (%)
Transform	20	Automatic	33.71
Accent	168	Accent	36.89
Random	20	Accent & Gender	36.35

The first row represents the best automatic selection technique, transformation based 20 neighbours selection using 5 minutes of target speaker's data. The second row shows the WER of SI model adapted on the South-Asian subset. It is 3.2% absolute worse than transformation based automatic selection. In the third row, randomly selected 20 neighbours from a set of matched accent and gender speakers is shown. The results were averaged across 5 trials in this case. Still the adapted system is 2.6% absolute worse than the best system. Both of these results show that, although gender and accent have significant influence on neighbours, the automatic selection is better than using accent and gender labels for choosing neighbours.

In the second set of experiments, automatic selection is combined with manual annotations, by running transformation based neighbour search on the accent subset instead of the whole training set. Table 6.7 lists the WER of automatic selection without and with manual annotations.

Table 6.7: *Automatic selection using manual annotations.*

System	Neighbours	Selection	WER (%)
Transform	20	Automatic	33.71
Accent	20	Automatic + Accent	33.73

The results show no major difference in performance between the two systems. From both the above experiments, it can be concluded that gender and accent have significant influence in automatic neighbours selection. However, the manual annotations of these speaker characteristics don't provide any additional benefits over transformation based approach, whether used by themselves or combined with automatic selection, except for reducing the search space.

6.4 Unsupervised adaptation

The experiments so far assumed availability of manual transcriptions for both the neighbours and target speaker. However in many real-world cases only a subset of available speakers are transcribed. Any additional target speaker data obtained over time is folded into the training set using initial hypothesis obtained from the existing ASR, as transcriptions. This section analyzes unsupervised training in the context of neighbour selection and adaptation. Figure 6.4 shows the difference between supervised and unsupervised adaptation with respect to the SI model. In the unsupervised case, the adaptation makes use of automatic transcriptions obtained using the SI model.

For the experiments in this section, the speakers in the training set and hence the neighbours are assumed to have manual transcriptions. In the first task, the supervised neighbour selection and adaptation is compared against the unsupervised case. Unsupervised neighbour selection involves automatically transcribing the beginning 5 mins of the target speaker and using it for selecting acoustically close neighbours. Once the neighbours are selected, their manual transcriptions are used to adapt the initial SI model and further adaptation is performed on the

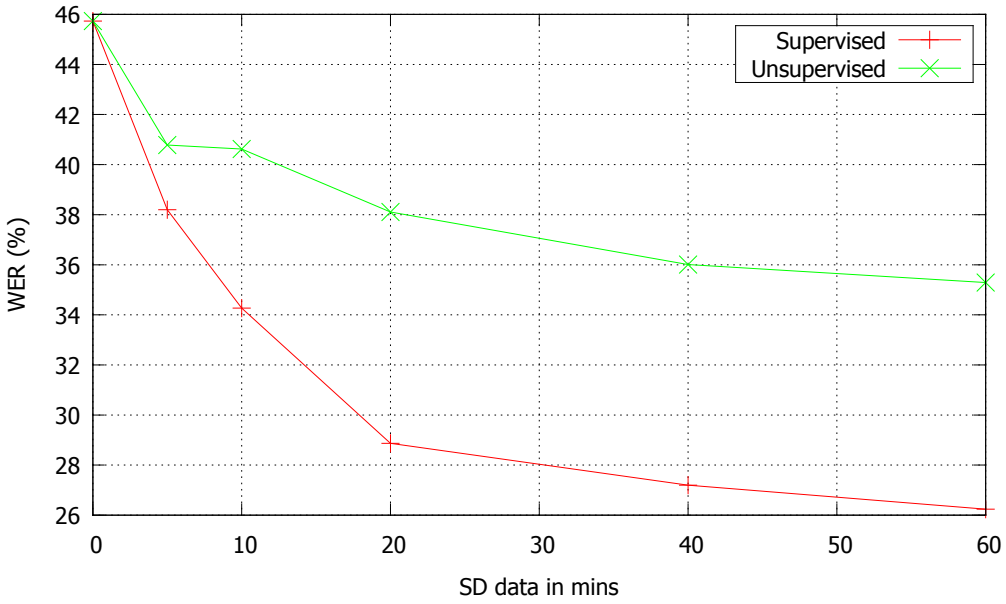


Figure 6.4: WER of supervised Vs unsupervised adaptation

automatic transcriptions from the target speaker. As more data from the target speaker becomes available, it is further automatically transcribed and used for adaptation.

Figure 6.5 shows the WER for South Asian speakers using SI- and neighbour-initalized models with additional unsupervised data. Neighbour selection achieves large improvements in the unsupervised case. With only 5 mins of target speaker data, neighbour initialized models achieve WER of 35.60%, while SI-initialized models require around 1 hour of data to obtain a similar improvement of 35.29%.

In the second experiment, unsupervised neighbour selection is compared againt the supervised case. It is assumed that the beginning 5 mins of target speaker data has manual transcriptions. As explained in section 6.2.2, this is practically feasible with the enrollment step for the new speakers. Figure 6.6 shows the difference between supervised and unsupervised neighbour selection. Table 6.8 compares the quality of neighbours selected using manual Vs automatic transcription of the first 5 mins of target speaker data. The supervised selection produces better neighbours, however the difference is very minimal, only 0.19% absolute.

Figure 6.6 shows the effect of additional target speaker data with supervised and unsupervised neighbour selection. To directly compare both experiments,

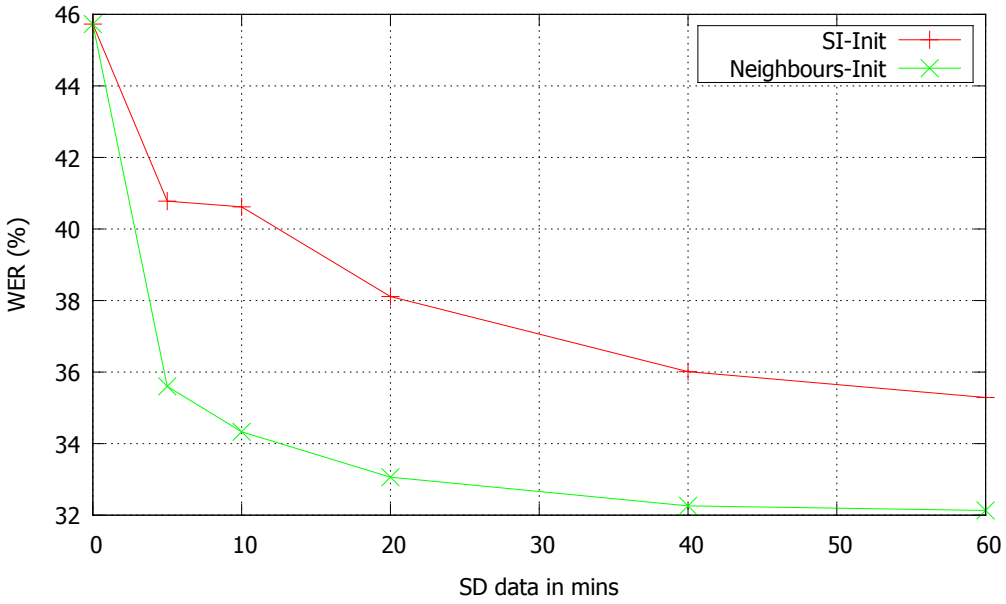


Figure 6.5: WER of unsupervised neighbour selection andadaptation

Table 6.8: *WER for supervised Vs unsupervised neighbour selection.*

Neighbour selection	WER (%)
None (SI)	45.73
Unsupervised	38.12
Supervised	37.93

once the neighbours are selected only automatic transcriptions are used for further adaptation. In this graph, the manual transcriptions for the first 5 mins of the target speaker are only used for selection and not adaptation. As in Table 6.8, although supervised selection produces lower WER, it is only 0.5-1.3% absolute lower than the unsupervised case.

6.5 Summary and discussion

This chapter presented an adaptation technique to build SD models with a few minutes of target speaker's data. An improvement of 23% relative over SI is obtained

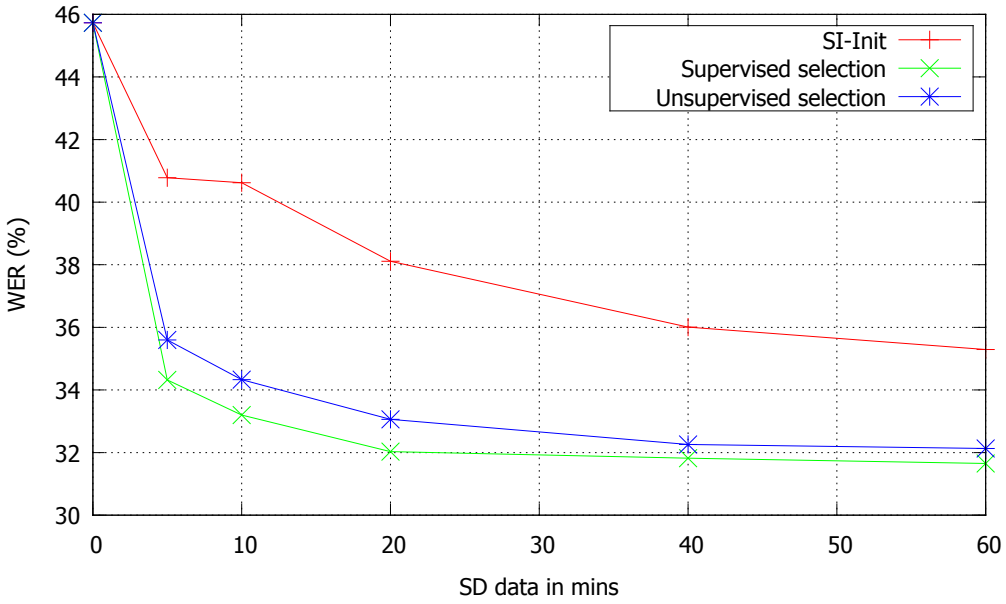


Figure 6.6: WER of supervised Vs unsupervised neighbour selection

with just 5 mins of the adaptation data. In the unsupervised case, 5 mins of target speaker data produced the same WER with neighbour selection as the SI-initialized models with 1 hour of adaptation data. The selected neighbours are analyzed to show that accent and gender play a crucial role in their selection. The automatic selection is compared against choosing neighbours based on manual annotations and concluded that the automatic approach performed better.

Chapter 7

Discriminative neighbour selection and adaptation

This chapter investigates neighbour selection and adaptation in the context of discriminative objective functions. Maximum Likelihood (ML) training aims to maximize the likelihood of the model generating the data, given the reference transcriptions. It guarantees optimal parameters assuming model correctness and infinite data. Real speech is not produced by a Markovian process and there is only a limited training data in practice, thus rendering ML objective function sub-optimal. Discriminative training (DT) has been proposed to compensate for these limitations of ML training. It makes use of information in the reference transcription and competing hypothese to estimate the model parameters. The competitors in the hypothesis space are usually compactly represented by a lattice. Different objective functions for DT have been proposed in the literature including Maximum Mutual Information (MMI) [VOWY97], Minimum Phone Error (MPE) [PW02], State-level Minimum Bayes Risk (sMBR) [GH06], etc. In the experiments described here, sMBR is used as a discriminative objective function. Discriminative training can also be achieved by using bottle-neck features obtained by training a Neural network using frame-level cross-entropy objective function.

In the experiments described in the previous chapter, various steps involved ML objective functions. They are listed as follows along with the specific estimation technique.

- Estimating SI model parameters (ML training).
- Neighbour selection (Likelihood/Transformation based selection).

- Neighbour adaptation (MAP adaptation).
- Target speaker adaptation (MAP adaptation).

All of these steps can use discriminative objective functions to potentially improve the performance of the final ASR. The first section analyzes neighbour selection and adaptation using a discriminatively trained SI model. The following section describes discriminative adaptation as an alternative to ML-based MAP adaptation on neighbours' data. Finally, neighbour selection involving a discriminative score is investigated. Combination of these techniques is also presented.

7.1 Discriminative training of SI model

In this section, the ML-trained SI model is updated with few iterations of discriminative training. The objective function used is state-level MBR which is given by

$$F(\lambda) = \sum_{u \in U} \sum_{W \in W'} P(W|O_u, \lambda) A(W_r, W) \tag{7.1}$$

where λ represents model parameters and U the list of training utterances. The reference transcription for a particular utterance is given by W_r and W' represents its hypothesis space. The loss between the reference and a hypothesis transcription is given by $A(W_r, W)$, which is this case is a hamming distance between frame-level state labels. The optimization is achieved using Extended Baum-Welch (EBW) style iterative updates with I-smoothing in each iteration to back-off to previous iteration. The DT SI model is trained using 8 iterations of sMBR training weight of 350.

Table 7.1: *Discriminative SI model.*

System	WER (%)
ML	45.73
DT (sMBR)	34.55
ML Neighbour Adapt	34.49

From Table 7.1, DT model obtains a WER of 34.55% which is 24.4% relatively lower than ML SI model. The new SI model is used in transformation based speaker

selection approach to choose 20 neighbours. Once the neighbours are chosen, ML MAP adaptation is employed to adapt the SI model on the neighbours' data. Unlike the ML SI case, maximum likelihood neighbour adaptation fails to produce any significant improvement over the DT baseline. Further adaptation with target speaker's data does not provide any improvements as well, as shown in figure 7.1.

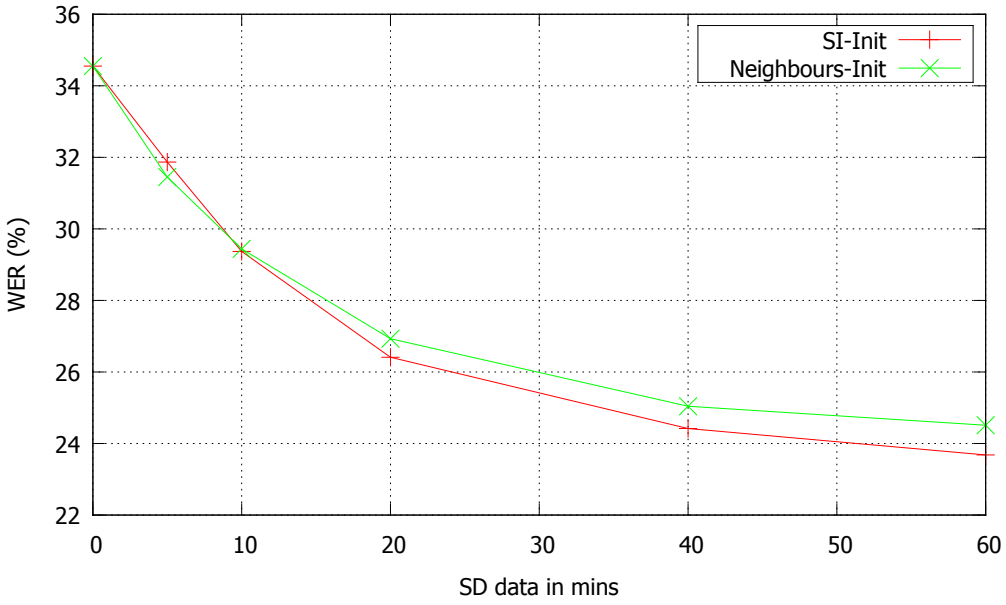


Figure 7.1: WER for Discriminative selection and adaptation

7.2 Discriminative neighbour adaptation

In this section, ML adaptation on the neighbours' data is replaced using discriminative adaptation. Discriminative versions of MAP adaptation including MMI-MAP and MPE-MAP were introduced in [PGKW03]. The technique modifies the standard EBW update equations to include two back-off schemes. The first one is the standard I-smoothing, which borrows back-off statistics from the previous iteration. The second scheme involves the traditional MAP adaptation with back-off statistics borrowed from the ML model parameters. The implementation discussed here involved 4 iterations of sMBR based MAP adaptation on neighbours' data. Table 7.2 shows the WER of different neighbour adaptation techniques. The discriminative adaptation

involving sMBR-MAP on neighbours' data produced 8.5% relative improvement over baseline SI model.

Table 7.2: *Discriminative neighbour adaptation.*

SI training	Neighbour adaptation	WER (%)
sMBR	-	34.55
sMBR	ML MAP	34.49
sMBR	sMBR MAP	31.60

Further addition of target speaker adaptation continued to produce lower WER for Neighbour-initialized model over the SI-initialized one, as shown in figure 7.2. The improvements however is lower compared to ML case. Discriminative adaptation can also be applied to the target speaker data, however, with a maximum of 1 hour of data sMBR-MAP did not produce significant improvement over the ML-MAP hence the ML counterpart retained for this case.

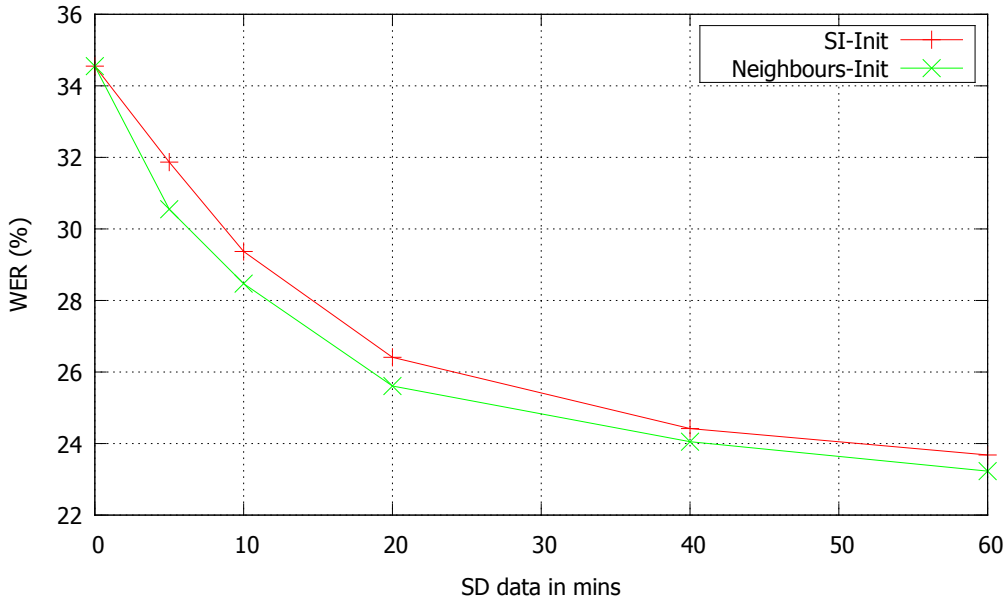


Figure 7.2: WER for Discriminative selection and adaptation

Table 7.3 shows the WER for different selection and adaptation setups. It can be seen that there is almost no improvement by adapting the SI model on the neighbours data using ML-MAP criterion. Discriminative adaptation [PGKW03]

of the SI model on the neighbours data yielded a relative improvement of 8.5% compared to the unadapted model.

7.3 Discriminative neighbour selection

Discriminative criterion can also be used for neighbour selection in addition to adaptation. In this approach, the adaptation data of 5 mins is decoded using the SI model to create lattices for the target speaker. Source speakers are chosen whose models maximize the average sMBR accuracy on the target speaker lattices. The sMBR accuracy is calculated as

$$sMBR\ Accuracy_T(S) = \sum_{u \in U_T} \sum_{W \in W'} \gamma(O_u, W | f_T(\lambda_S)) A(W_r, W) \tag{7.2}$$

where W_r is the reference transcription, W' are the competitor paths in the denominator lattice, $\gamma(O_u, W/\lambda_S)$ is the posterior of a lattice path according to the (adapted) source model $f_T(\lambda_S)$. $A(W_r, W)$ is the raw accuracy between the reference and competitor state sequences. Analogous to likelihood and transformation based neighbour selection, discriminative method leads to choosing neighbours who *make less error* on the target speaker's data.

Table 7.3: *Discriminative selection and adaptation.*

Selection	Adaptation	WER (%)
-	-	34.55 (SI)
ML transformation	ML MAP	34.49
ML transformation	sMBR MAP	31.60
sMBR Acc.	ML MAP	31.97
sMBR Acc.	sMBR MAP	31.00

From Table 7.3, it is interesting to note that neighbours selected using sMBR accuracy produce 7.5% relative improvement over SI model, using ML MAP adaptation. Comparing rows 2 and 3, it is noted that discriminative selection can lead to neighbours who produce less WER on target speaker data than ML based selection. While the gains from discriminative selection and adaptation are not additive, the combined technique still produces the best result of 10.3% relative

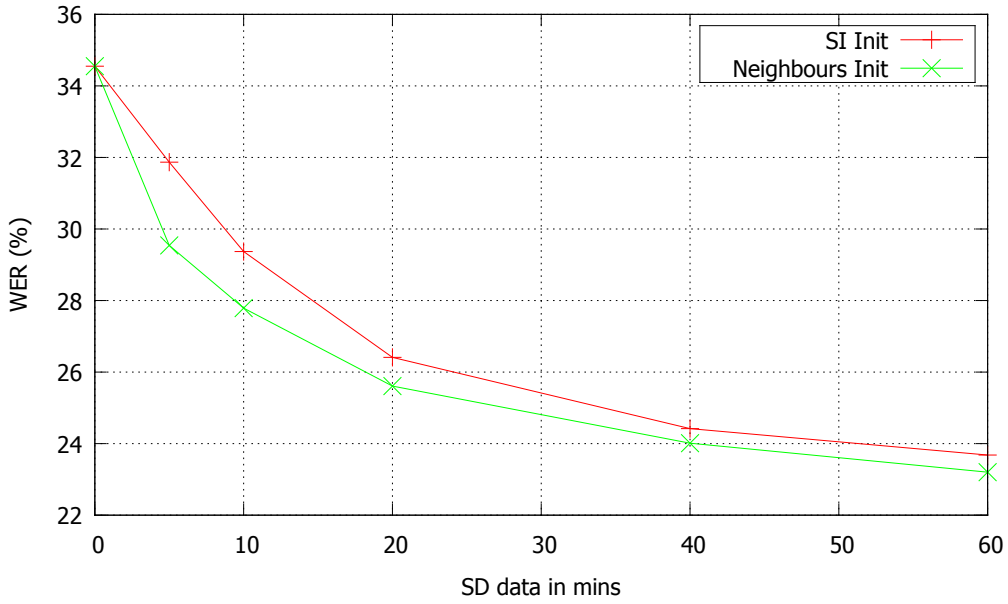


Figure 7.3: WER for Discriminative selection and adaptation

improvement over the unadapted system. Figure 7.3 shows the WER of SI and neighbours-initialized systems with increasing target data.

7.4 Deep bottle-neck features

In this section, neighbour selection and adaptation is studied in the context of a neural network based ASR system. The GMMs in the system are trained with features from the bottle-neck layer of a multi-layered perceptron (MLP). The input to the neural network are the same features the GMMs are trained with in the MFCC based system, i.e MFCC followed by LDA and STC. These features are stacked with a window size of 5 and projected down to 150 dimensions using a second LDA matrix. The second matrix helps with whitening the input features to the MLP in addition to dimensionality reduction. This setup corresponds to the Type IV features investigated in [RPVC13].

The MLP has 4 hidden layers with dimensionality 465 and a bottle-neck layer of size 32. The architecture of the MLP is 150x465x465x465x32x465x1000 and it has a total of 1M parameters. The output layer has 1000 nodes representing

context-dependent states in the MFCC based system. The contextual decision tree with 3000 states is pruned down to 1000 leaves, which are used as targets for the MLP. However, the labels are obtained by force-aligning the utterances with the 3000 state system similar to [SHRY13]. It has been shown that bottle-neck features performed better with a lower number of targets than the original MFCC system [CCR+13]. Once the MLP is trained, the activations from the bottle-neck features are extracted, stacked with a window size of 9 and projected down to 32 dimensions with a final LDA transform.

MLP is trained using backpropagation using the 'newbob' learning rate schedule. The training starts with a learning rate of 0.008 until the accuracy on the held-out data reduces and then it is divided in half for remaining epochs. The training stops if there is no improvement of accuracy on the held-out set. The available training data is subsampled to reduce the training time for the MLP. All the training speakers are retained, but they are restricted to a maximum of 15 mins to ensure there is ample diversity in the MLP training data. 450 hours from 1691 speakers is used as a training set and 15 hour of data from 187 speakers is used as held-out set. The training continued for 9 epochs with a batch size of 256. The final frame accuracy of the MLP on the training set is 46.3% and 43.0% on the held-out set.

7.5 DBNF neighbour selection and adaptation

The neighbour selection for DBNF system follows the transformation based technique using ML criterion. The source speakers are ranked with the similarity score using 5 mins of adaptation data. Top 20 neighbours are selected for each target speaker. Once the neighbours are chosen, MAP adaptation is performed on the neighbours data. Finally MLLR followed MAP adaptation is carried out on the target speaker's data. Figure 7.4 shows the WER plots of neighbour-initialized and SI adapted systems for increasing target speaker's data.

It is interesting to note that although DBNF system used ML based neighbour selection and adaptation, its performance is similar to MFCC system with discriminative adaptation. The additional MLP layers act as discriminatively trained models for the ASR system.

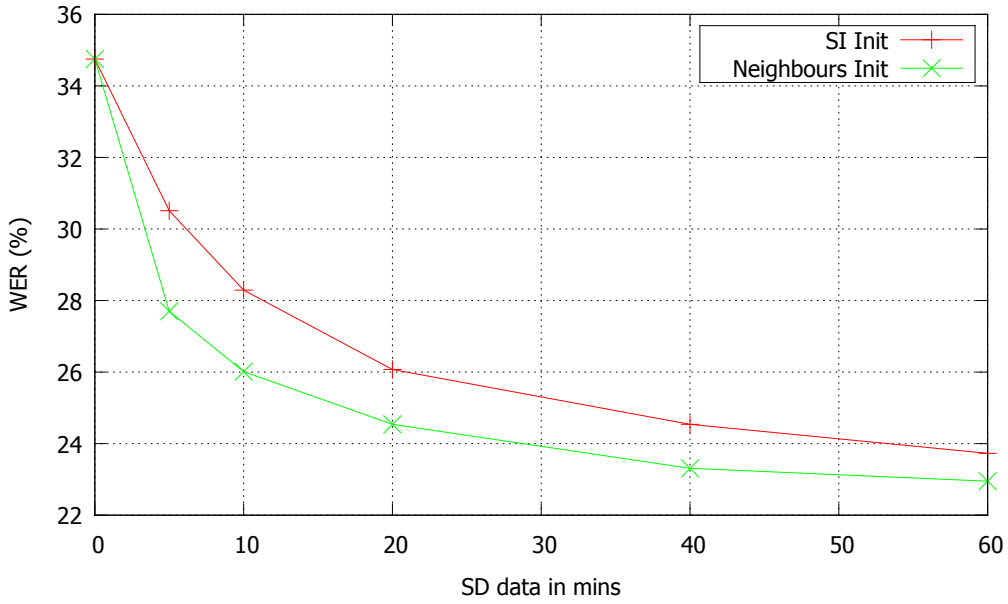


Figure 7.4: WER for DBNF neighbour selection and adaptation

7.6 Summary and discussion

The neighbour selection and adaptation techniques are studied for discriminative models in this chapter. A discriminative version of neighbour selection is introduced and compared with the DT adaptation. Their combination showed additional improvements. Finally, neighbour selection using DBNF features is studied and compared against MFCC based systems.

Chapter 8

Text based neighbour selection

This chapter explores text-based neighbour selection techniques for improved speech recognition. In the first set of experiments, text-based selection is used to augment acoustic data selection. When not enough acoustic data is available, it is shown that text based selection improves the reliability of neighbours selected for adaptation. In the second set of experiments, text based speaker selection is used to select neighbours for language model adaptation. LM neighbours are analogous to acoustic neighbours for a speaker, selected to reduce the perplexity of the target speaker on the test set. Analysis on the neighbours selected based on domain and accent is also presented.

8.1 Textual features for neighbour selection

The previous chapters presented various neighbour selection and adaptation techniques using few mins of audio data from the target speaker. It is shown that as the audio data available for the target speaker decreases, the quality of neighbour selected for adaptation of target SD models also decreases. In this section, text based features are investigated to augment the acoustic data selection. There are many applications where text data for a speaker is available in abundance, while it is difficult to collect the audio data. For example, in the case of medical transcription if the doctor is starting with ASR for the first time, it is easy to access a list of archived reports while the corresponding audio data may not be available. There are also auxilliary text sources available in many applications to help improve the performance of SD models. For example, in the case of desktop dictation, the users'

emails may serve as representative text for the domain.

In this section, text-based features are used to select neighbours for adaptation of the SI model. In the acoustic neighbour selection, source speaker models are used to calculate likelihoods on the target speaker data. The speakers with the highest likelihood are chosen as neighbours for adaptation. It was observed that the majority of selected neighbours belong to the same accent. However when the amount of adaptation data is varied, there was a deterioration in the quality of neighbours selected. Table 6.2.2 showed WER of neighbour adapted system based on the amount of target data available for selection. The goal of the following experiment is to compensate this lack of acoustic data using text-based features. It explores if additional text data can be used to augment acoustic neighbour selection.

A classifier is trained on the source speakers and their top 20 acoustic neighbours using bag-of-word features. During testing, available text data for the test speaker is used to pre-select the neighbours for adaptation. The second round of selection is carried out using the small amount of enrollment audio data. Since the second round of selection is carried out on pre-selected source speakers, it produces better neighbours compared to choosing neighbours considering the whole training data.

8.1.1 Experiment setup

The same M*Modal medical transcription dataset is used here. For each of the 1876 training speakers, neighbours are calculated by scanning through the entire training dataset. Choosing 20 neighbours was found to give the best acoustic adaptation performance in the previous chapter. Hence the top 20 neighbours are used as positive examples for the source speaker and 20 speakers with the lowest likelihood score as negative examples. A vocabulary of 20k words based on frequency is used to calculate the bag of word features. The steps for preparing a training dataset for text-based neighbour predictor is listed as follows.

- For each source speaker, calculate 20 speakers that produce the highest likelihood and 20 which produce the lowest likelihood.
- For each training example, 20k text features are extracted from the source speaker's reports based on unigram counts.
- It is appended with 20k text features extracted from the neighbour's text data.

- The output class is based on whether the neighbour belong to the highest or lowest likelihood group.

Once the training examples are collected for all source speakers, a binary classifier is trained based solely on text features. In this experiment, 90% of the source speakers (1690) formed the training set and remaining 10% is used as a development set. Linear SVM is used to train the text based neighbour predictor with the help of LIBLINEAR [Lib] tool. The regularization constant is tuned using the development set. The classifier had an accuracy of 68% but the classification performance is not the main focus of this experiment. We are mainly interested in the quality of neighbours selected for each test speaker. It is assumed we have 15 reports for each speaker during training and test. In testing, the text data available for the test speaker and the source speaker is used to form the feature vector. The output determines if the source speaker is a good neighbour for the test speaker. Once the neighbours are chosen, acoustic adaptation of the SI model is carried out using the audio data from the neighbours.

8.1.2 Experiments

Table 8.1 shows the WER 20 neighbours selected randomly (5 trials) and using a text-based neighbour predictor. We can see that text-based neighbours provided 5% relative improvement over the baseline. One of the reasons the random selection yield poor results is that the training dataset has less than 10% of South-Asian speakers. Hence most of the random neighbours belong to the different accent. In the text-based selection, there are 40% South-Asian speakers on average, in the 20 neighbours selected for adaptation. It is still less than nearly 70% of neighbours with matched accent using acoustic data. Hence, there is only a moderate improvement with text-based features. However, the text-based predictor can be used as a pre-selection criteria to augment acoustic selection in the case where there is only a small amount of available target speech data.

Table 8.1: *WER of neighbour selection using text-based selection.*

Neighbour selection	WER (%)
- (SI)	45.73
Random	50.12
Text-based	43.20

The text-based predictor can be used as a first step to select matching accent neighbours for cases where enough speech data is not available for the target speaker. These neighbours can be rescored using the available audio data. In this experiment, the amount of target speech data available is varied from 1 to 5 mins. An initial set of 100 neighbours is chosen using the text-based selection and they are rescored to choose 20 neighbours using acoustic neighbour selection as the second step. Table 8.2 shows the WER of acoustic and combined neighbour selection techniques. It can be seen that while reducing the amount of acoustic data from 5 mins to 1 min caused 0.75% absolute reduction in WER, the text based pre-selection reduced it to only 0.4%.

Table 8.2: *WER of neighbour selection using audio and text-based selection.*

Target speaker data	WER (%)	
	Acoustic data	+ Text data
1.63	37.21	36.61
2.68	36.93	36.42
5.82	36.46	36.21

8.2 Language model adaptation

In the previous sections, we focussed our attention on adapting the acoustic model to better model the target speaker. In this section, the neighbour selection and adaptation is carried out for the goal of language model adaptation. The language model can be influenced by a variety of factors including dialect/accent of the doctor, vocabulary choices based on geographical origin, hospital they work for, speciality of dictated content, etc. We use a different dataset with richer annotations to allow us analyze these different factors influencing the language model. The new M*Modal dataset consists of 1350 speakers from 27 different specialties including Internal medicine, Radiology, Pediatrics, etc. Each specialty contained 50 doctors in the training set. Each doctor in the dataset is also annotated with the hospital they are from. The dataset contained 300K reports in total with 72M tokens and a vocabulary of 58K words.

The LM neighbours are chosen as follows

- SI LM is trained using data from all source speakers.

- For each source speaker, SD LM is trained using text data from the source speaker.
- The 2 LMs are interpolated on the target speaker data, aka. transformation based selection.
- The likelihood is then calculated on the target text data.

Once the neighbours are ranked based on the likelihood, the top N neighours are chosen for adaptation. A neighbour LM is trained on their pooled data and interpolated with the SI LM using the target speaker's text data. We used the SI and South Asian adapted acoustic models to label speaker whether they are South-Asian or not based on the respective acoustic likelihood.

A test set of 15 speakers of South Asian accent with 50 reports each is used to analyze the neighbour selection for LM adaptation. An additional 5 reports for each speaker is used for neighbour selection and adaptation. They are randomly selected from different specialties to ensure there is no significant overlap. The following sections analyze the influence of various dialectal factors such as accent, hospital origin and specialty in neighbour selection.

8.3 Analysis

In the first experiment, the influence of different factors in neighbour selection is analyzed. Speciality of the doctor had the most significant influence on the LM neighours compared to the accent and the hospital labels (Mann-Whitney U test, $p < 0.001$). Figure 8.1 shows the neighbours at different ranks separated by those from same domain as the target speaker and others. As the rank increases, the ratio of same domain neighbours increases. There is a cross-over between the two curves at the end, as all the neighbours from the same domain have been selected and the remaining are out of domain ones. The graph clearly shows that the specialty of the target speaker has significant influence in neighbours selected for LM adaptation.

These documents are highly specific reports of patients, so it is obvious to see speciality playing a larger role in the language model compared to accent labels automatically derived from acoustics. Comparing the experiments in section 8.1 and 8.2, we can confirm that accent manifests itself in textual content in addition to the speech data. We can automatically identify this influence using a classifier trained solely on the text, albeit with accents labels obtained using acoustic likelihoods. It

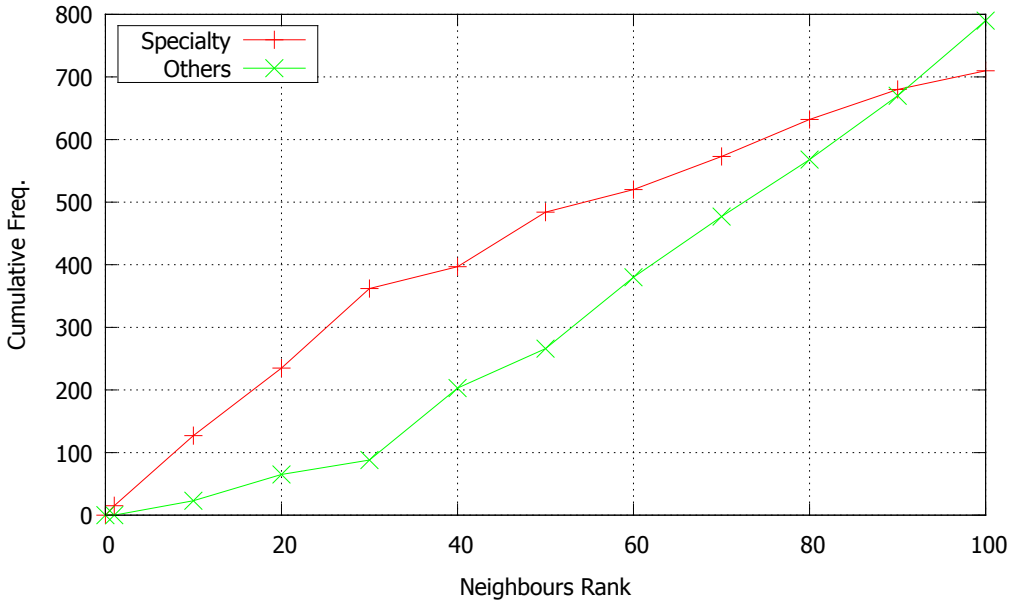


Figure 8.1: Rank of neighbours from same and different specialties

is helpful for choosing neighbours for acoustic model adaptation, particularly when only a small amount of target speech data is available. The neighbour selection and adaptation can be extended to LM adaptation where the technique automatically picks neighbours related to the speciality of the target speaker which has the highest influence on the content of the speech data.

8.4 ASR experiments

In this section, LM neighbour selection and adaptation is carried out for the test speakers in original M*Modal test set. We start with determining the optimal number of neighbours for LM adaptation. Table 8.3 shows that perplexity of the target LM adapted using varying number of neighbours. The best result can be obtained using 30 neighbours.

In the next experiment, 30 neighbours are chosen for each test speaker and their SD LMs are used in ASR decoding. Table 8.5 shows the WER of SI and neighbour adapted LMs on the South Asian test set. It is to be noted that the SI LM trained on the new M*Modal dataset is used to obtain the baseline WER instead of internal

Table 8.3: *Perplexity for varying neighbours in LM adaptation.*

Rank	Perplexity
1	97.0
10	95.1
20	92.7
30	88.1
40	89.5
80	91.3
100	93.7

medicine LM as in the previous experiments. Hence, the baseline WER is worse (47.7%) than reported before (45.73%). From the table, there is 11.5% relative reduction is perplexity which resulted in 6.7% relative reduction in WER.

Table 8.4: *WER for neighbour selection based LM adaptation.*

LM	Perplexity	WER (%)
SI	83.9	47.8
Neighbour adapt	74.2	44.6

Table 8.5 shows speaker WER for unadapted and neighbour adapted LMs. It can be seen that the improvement is consistent across all speakers.

8.5 Summary and discussion

This chapter explored the use of text data for neighbour selection and adaptation. Text-based features can be shown to augment acoustic neighbour selection, when there is limited amount of speech data from the target speaker. The neighbour selection in the context of LM adaptation is discussed. The specialty of the target speaker is shown to influence LM neighbour selection through empirical analysis. The neighbour adapted LM produced 11.5% reduction in perplexity and 6.7% reduction in WER across 10 speakers.

Table 8.5: *WER for LM adaptation using neighbour selection.*

Speaker	WER (%)	
	SI-LM	Neighbour Adaptation
1	20.4	19.3
2	31.2	30.1
3	44.0	41.2
4	38.6	36.7
5	49.5	46.9
6	61.0	58.1
7	65.9	62.5
8	61.2	58.2
9	55.7	52.8
10	44.3	42.3
Avg	**47.8**	**44.6**

Part V

Conclusion

Chapter 9

Summary and Future Directions

9.1 Thesis contributions

This thesis quantified the problems faced by accented speakers in 3 different ASR scenarios and proposed novel research to address them. Techniques are introduced to handle accented speakers in accent dependent systems, accent independent systems and speaker dependent systems. The following list summarizes the contributions in each ASR setup.

- Accent dependent systems: A decision tree based adaptation technique is proposed to specifically handle accent variations from source accent to the target, given the limited amount of adaptation data. Semi-continuous polyphone decision tree specialization (SPDTS) is shown to efficiently use small amount of adaptation data to adapt the baseline model to the target accent. The technique is evaluated on different datasets ranging from medium-scale to large-scale tasks including WSJ, GALE and M*Modal medical dictation in two languages. Compared to traditional MAP adaptation, SPDTS obtained a relative WER reduction of 4.2-13.6% on ML and discriminative baseline ASR models.

  Bias sampling is proposed in the context of active and semi-supervised learning to leverage unlabeled target accent data in a large corpus. Active learning technique resulted in a reduction of labelling cost by 50% and semi-supervised learning provided additional improvement of 2-15.9% relative WER over supervised baseline.

- Accent independent systems: Decision tree based accent robustness analysis is introduced to measure the accent normalization characteristics of different front-ends. Traditional 4-layer bottle-neck MLP is shown to normalize for gender, but are still sensitive to accent variations. Deeper architectures and deeper bottle-neck layers are shown to have better normalization characteristics when compared to MFCC or MLP.

  Accent-aware training is proposed to augment accent labels to spectral features during training of BN front-ends. It is shown that labelling a subset of speakers with accents (150 hours) and using them as additional features in accent-aware training results in a WER of 33.78% which is better than 34.75% WER obtained by training an accent agnostic model on a much larger dataset (450 hours).

- Speaker dependent systems: The problems faced by accented speakers in the context of speaker dependent models are highlighted. Neighbour selection and adaptation techniques are proposed to create target speaker-dependent ASR model using few mins of adaptation data. These technique produce an improvement of 31.2% relative over SI+MLLR baseline. The neighbours are further analyzed to show that accent plays a crucial role in the neighbours chosen automatically for the target speaker. The neighbour selection is extensively analyzed in different experiments including varying the number of neighbours, amount of adaptation data and supervised Vs unsupervised adaptation.

  A discriminative version of the neighbour selection using sMBR is formulated and shown to provide gains of around 10.1% relative over ML neighbour selection. Text-based neighbour selection is explored in the case of lack of sufficient or no audio data from the target speaker and shown to provide consistent gains of around 1-2% over audio-only neighbour selection. The technique is ported to LM adaptation and the neighbour selection is analyzed along different factors such as accent, speciality of the dictation content and place of work. LM adaptation resulted in 11.5% reduction in perplexity and 6.7% relative reduction in WER compared to the unadapted LM.

9.2 Future Directions

This work can have several future directions. Some of them are listed below

- Dialect modeling using RNNLM: The current experiments mainly use Ngram

models for language model adaptation. It will be interesting to think about language model adaptation and interpolation in the context of RNN architecture.

- The thesis has explored on acoustic and language model adaptation in the context of accented speakers. Other models in the ASR such as vocabulary can be adapted based on the dialect of the user.

- The accent dependent chapter introduced decision tree based accent adaptation to model contextual rules in the target accent. This technique is complementary to pronunciation modeling whose aim is to change the surface forms in the dictionary using re-write rules. It will be interesting to compare the interplay between these two modules and investigate tighter integration of both techniques.

Bibliography

[Bac13] Michiel Bacchiani. Rapid adaptation for mobile speech applications. In *ICASSP*, 2013. 74

[BBS+08] Michiel Bacchiani, Françoise Beaufays, Johan Schalkwyk, Mike Schuster, and Brian Strope. Deploying goog-411: Early lessons in data, measurement, and testing. In *ICASSP*, pages 5260–5263, 2008. 3

[BBS09] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10, 2009. 27, 40

[BHC10] Fadi Biadsy, Julia Hirschberg, and Michael Collins. Dialect recognition using a phone-gmm-supervector-based svm kernel. In *INTERSPEECH*, pages 753–756, 2010. 57

[BI10] John Blitzer and Hal Daumé III. ICML tutorial on domain adaptation. http://adaptationtutorial.blitzer.com, June 2010. 27, 40

[BK03] Christophe Van Bael and Simon King. An accent-independent lexicon for automatic speech recognition. In *ICPhS*, pages 1165–1168, 2003. 4, 12

[BMJ12] Fadi Biadsy, Pedro Moreno, and Martin Jansche. Google's cross-dialect arabic voice search. In *ICASSP*, 2012. 3

[BVS10] Françoise Beaufays, Vincent Vanhoucke, and Brian Strope. Unsupervised discovery and training of maximally dissimilar cluster models. In *INTERSPEECH*, pages 66–69, 2010. 74

[CCR+13] Jia Cui, Xiaodong Cui, Bhuvana Ramabhadran, Janice Kim, Brian Kingsbury, Jonathan Mamou, Lidia Mangu, Michael Picheny, Tara N.

Sainath, and Abhinav Sethy. Developing speech recognition systems for corpus indexing under the iarpa babel program. In *ICASSP*, 2013. 97

[Che01] Rathinavelu Chengalvarayan. Accent-independent universal hmm-based speech recognizer for american, australian and british english. In *INTERSPEECH*, pages 2733–2736, 2001. 50

[CJ06] Constance Clarke and Daniel Jurafsky. Limitations of mllr adaptation with spanish-accented english: an error analysis. In *INTERSPEECH*, 2006. 5, 18

[CLWL00] Kuan-Ting Chen, Wen-Wei Liau, Hsin-Min Wang, and Lin-Shan Lee. Fast speaker adaptation using eigenspace-based maximum likelihood linear regression. In *INTERSPEECH*, pages 742–745, 2000. 74

[CMN09] Mónica Caballero, Asunción Moreno, and Albino Nogueiras. Multi-dialectal spanish acoustic modeling for speech recognition. *Speech Communication*, 51(3):217–229, 2009. 5, 50

[CMRR08] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *ALT*, pages 38–53, 2008. 27, 40

[Com01] Dirk Van Compernolle. Recognizing speech of goats, wolves, sheep and ... non-natives. *Speech Communication*, 35(1-2):71–79, 2001. 3

[DDNK97] V. Digalakis, V. Digalakis, L. Neumeyer, and J. Kaja. Development of dialect-specific speech recognizers using adaptation methods. In *ICASSP*, pages 1455–1458, 1997. 5

[Den11] Li Deng. Front-end, back-end, and hybrid techniques for noise-robust speech recognition. In *Robust Speech Recognition of Uncertain or Missing Data*, pages 67–99. 2011. 50

[FSGL11] Thiago Fraga-Silva, Jean-Luc Gauvain, and Lori Lamel. Lattice-based unsupervised acoustic model training. In *ICASSP*, pages 4656–4659, 2011. 39

[Fur89] Sadaoki Furui. Unsupervised speaker adaptation method based on hierarchical spectral clustering. In *ICASSP*, 1989. 74

[FWK08] Joe Frankel, Dong Wang, and Simon King. Growing bottleneck features for tandem asr. In *INTERSPEECH*, page 1549, 2008. 54

[GAAD+05] Jean-Luc Gauvain, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Véronique Gendner, Lori Lamel, and Holger Schwenk. Where are we in transcribing french broadcast news? In *INTERSPEECH*, pages 1665–1668, 2005. 3

[Gal96] M. J. F. Gales. The generation and use of regression class trees for MLLR adaptation. Technical report, Cambridge University, 1996. 78

[Gal98] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech & Language*, 12(2):75–98, 1998. 70, 72

[Gal00] Mark J. F. Gales. Cluster adaptive training of hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 8(4):417–428, 2000. 75

[Gal01] M. J. F. Gales. Multiple-cluster adaptive training schemes. In *ICASSP*, 2001. 75

[Gal11a] M. J. F. Gales. Model-based approaches to handling uncertainty. In *Robust Speech Recognition of Uncertain or Missing Data*, pages 101–125. 2011. 50

[Gal11b] M. J. F. Gales. Model-based approaches to handling uncertainty. In *Robust Speech Recognition of Uncertain or Missing Data*, pages 101–125. 2011. 50

[GF08] Frantisek Grézl and Petr Fousek. Optimizing bottle-neck features for lvcsr. In *ICASSP*, pages 4729–4732, 2008. 54

[GH06] Matthew Gibson and Thomas Hain. Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition. In *INTERSPEECH*, 2006. 91

[GKKC07] Frantisek Grézl, Martin Karafiát, Stanislav Kontár, and Jan Cernocký. Probabilistic and bottle-neck features for lvcsr of meetings. In *ICASSP*, volume 4, pages IV–757 –IV–760, april 2007. 55

[GL94a] Jean-Luc Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994. 5, 40

[GL94b] Jean-Luc Gauvain and Chin-Hui Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994. 72

[GRK04] Silke Goronzy, Stefan Rapp, and Ralf Kompe. Generating non-native pronunciation variants for lexicon adaptation. *Speech Communication*, 42(1):109–123, 2004. 12

[HAH01] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. In *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 2001. 18

[HCC01] Chao Huang, Eric Chang, and Tao Chen. Accent issues in large vocabulary continuous speech recognition. Technical Report MSR-TR-2001-69, Microsoft Research, 2001. 3

[HCC02] Chao Huang, Tao Chen, and Eric Chang. Speaker selection training for large vocabulary continuous speech recognition. In *ICASSP*, pages 609–612, 2002. 75, 76

[HDY+12] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *Signal Processing Magazine*, 2012. 61

[HFT+09] Roger Hsiao, Mark Fuhs, Yik-Cheung Tam, Qin Jin, Ian Lane, and Tanja Schultz. The cmu-interact mandarin transcription system for gale. In *GALE Book*, 2009. 3

[HJ91] Hynek Hermansky and Louis Anthony Cox Jr. Perceptual linear predictive (plp) analysis-resynthesis technique. In *EUROSPEECH*, 1991. 49

[HJZ10] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. In *NIPS*, pages 892–900, 2010. 26

[HM94] Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994. 49

[HP99] Jing Huang and Mukund Padmanabhan. A study of adaptation techniques on a voicemail transcription task. In *EUROSPEECH*, 1999. 75

[HTRG02] Dilek Z. Hakkani-Tür, Giuseppe Riccardi, and Allen L. Gorin. Active learning for automatic speech recognition. In *ICASSP*, pages 3904–3907, 2002. 26

[Hum97] J.J. Humphries. Accent modelling and adaptation in automatic speech recognition. http://svr-www.eng.cam.ac.uk/~jjh11/publications/PhD_thesis.ps.gz, 1997. 5, 12

[HW97] J. J. Humphries and Philip C. Woodland. Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition. In *EUROSPEECH*, 1997. 3, 4, 12

[IMS+04] Shajith Ikbal, Hemant Misra, Sunil Sivadas, Hynek Hermansky, and Hervé Bourlard. Entropy based combination of tandem representations for noise robust asr. In *INTERSPEECH*, 2004. 49

[ISJ+12] N. Itoh, T.N. Sainath, D.N. Jiang, J. Zhou, and B. Ramabhadran. N-best entropy based data selection for acoustic modeling. In *ICASSP*, pages 4133–4136, 2012. 25, 26, 27, 30

[KG09] D. K. Kim and M. J. F. Gales. Adaptive training with noisy constrained maximum likelihood linear regression for noise robust speech recognition. In *INTERSPEECH*, pages 2383–2386, 2009. 50

[KMN12] Herman Kamper, Félicien Jeje Muamba Mukanya, and Thomas Niesler. Multi-accent acoustic modelling of south african english. *Speech Communication*, 54(6):801–813, 2012. 5, 50

[KNJ+98] Roland Kuhn, Patrick Nguyen, Jean-Claude Junqua, Lloyd Goldwasser, Nancy Niedzielski, Steven Fincke, Ken Field, and Matteo Contolini. Eigenvoices for speaker adaptation. In *ICSLP*, 1998. 73

[KW99] Thomas Kemp and Alex Waibel. Unsupervised training of a speech recognizer: recent experiments. In *EUROSPEECH*, 1999. 37

[LG] Karen Livescu and James Glass. Lexical modeling of non-native speech for automatic speech recognition. In *ICASSP*, pages 1683 – 1686. 4

[LGN09] Jonas Lööf, Christian Gollan, and Hermann Ney. Cross-language bootstrapping for unsupervised acoustic model training: rapid development of a polish speech recognition system. In *INTERSPEECH*, pages 88–91, 2009. 37

[Lib] Phoneme. http://www.csie.ntu.edu.tw/~cjlin/liblinear/. 101

[Liv99] K. Livescu. Analysis and modeling of non-native speech for automatic speech recognition. http://www.sls.lcs.mit.edu/sls/publications/1999/msthesis-livescu.pdf, 1999. 12

[LW94] C. J. Leggetter and Philip C. Woodland. Speaker adaptation of continuous density HMMs using multivariate linear regression. In *ICSLP*, 1994. 72

[LW95] C.J. Leggetter and P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185, 1995. 5

[MBS00] Lidia Mangu, Eric Brill, and Andreas Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech & Language*, 14(4), 2000. 27, 39, 42

[MGA+06] Spyridon Matsoukas, Jean-Luc Gauvain, Gilles Adda, Thomas Colthurst, Chia-Lin Kao, Owen Kimball, Lori Lamel, Fabrice Lefevre, Jeff Z. Ma, John Makhoul, Long Nguyen, Rohit Prasad, Richard M. Schwartz, Holger Schwenk, and Bing Xiang. Advances in transcription of broadcast news and conversational telephone speech within the combined ears bbn/limsi system. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1541–1556, 2006. 3

[MH07] Brian Kan-Wing Mak and Roger Wend-Huu Hsiao. Kernel eigenspace-based mllr adaptation. *IEEE Transactions on Audio, Speech & Language Processing*, 15(3):784–795, 2007. 73

[MHJ+10] Florian Metze, Roger Hsiao, Qin Jin, Udhyakumar Nallasamy, and Tanja Schultz. The 2010 cmu gale speech-to-text system. In *INTERSPEECH*, pages 1501–1504, 2010. 16, 32

[MHu] Accents research. http://www.phon.ucl.ac.uk/home/mark/accent. 4

[ML10] Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *ACL (Short Papers)*, pages 220–224, 2010. 29

[MS08] Jeff Z. Ma and Richard M. Schwartz. Unsupervised versus supervised training of acoustic models. In *INTERSPEECH*, pages 2374–2377, 2008. 37

[MW47] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60, 1947. 84

[NFW+13] Udhyakumar Nallasamy, Mark Fuhs, Monika Woszczyna, Florian Metze, and Tanja Schultz. Neighbour selection and adaptation for rapid speaker-dependent asr. In *ASRU*, 2013. 6

[NGM+11] Udhyakumar Nallasamy, Michael Garbus, Florian Metze, Qin Jin, Thomas Schaaf, and Tanja Schultz. Analysis of dialectal influence in pan-arabic asr. In *INTERSPEECH*, pages 1721–1724, 2011. 4, 6, 12

[NMS11] Udhyakumar Nallasamy, Florian Metze, and Thomas Schaaf. Normalization of gender, dialect and speaking style using probabilistic front-ends. In *DAGA*, 2011. 6

[NMS12a] Udhyakumar Nallasamy, Florian Metze, and Tanja Schultz. Active learning for accent adaptation in automatic speech recognition. In *SLT*, 2012. 6

[NMS12b] Udhyakumar Nallasamy, Florian Metze, and Tanja Schultz. Enhanced polyphone decision tree adaptation for accented speech recognition. In *Interspeech*, 2012. 3, 6, 32, 40

[NMS12c] Udhyakumar Nallasamy, Florian Metze, and Tanja Schultz. Semi-supervised learning for speech recognition in the context of accent adaptation. In *MLSLP Symposium*, 2012. 6, 29

[NSK11] Scott Novotney, Richard M. Schwartz, and Sanjeev Khudanpur. Unsupervised arabic dialect adaptation with self-training. In *INTERSPEECH*, pages 541–544, 2011. 5, 37

[NWJ99] Patrick Nguyen, Christian Wellekens, and Jean-Claude Junqua. Maximum likelihood eigenspace and mllr for speech recognition in noisy environments. In *EUROSPEECH*, 1999. 73

[PBNP98] Mukund Padmanabhan, Lalit R. Bahl, David Nahamoo, and Michael A. Picheny. Speaker clustering and transformation for speaker adaptation in speech recognition systems. *IEEE Transactions on Speech and Audio Processing*, 6(1):71–77, 1998. 75

[PGKW03] Daniel Povey, M. J. F. Gales, Do Yeong Kim, and Philip C. Woodland. MMI-MAP and MPE-MAP for acoustic model adaptation. In *INTERSPEECH*, 2003. 93, 94

[PW02] Daniel Povey and Philip C. Woodland. Minimum phone error and i-smoothing for improved discriminative training. In *ICASSP*, pages 105–108, 2002. 91

[PY12a] Daniel Povey and Kaisheng Yao. A basis representation of constrained mllr transforms for robust adaptation. *Computer Speech & Language*, 26(1):35–51, 2012. 41

[PY12b] Daniel Povey and Kaisheng Yao. A basis representation of constrained mllr transforms for robust adaptation. *Computer Speech & Language*, 26(1):35–51, 2012. 74

[Qui] Quicknet toolkit. http://www1.icsi.berkeley.edu/Speech/qn.html. 54

[Ram05] Bhuvana Ramabhadran. Exploiting large quantities of spontaneous speech for unsupervised training of acoustic models. In *INTERSPEECH*, pages 1617–1620, 2005. 37

[RBF+99] Michael Riley, William Byrne, Michael Finke, Sanjeev Khudanpur, Andrej Ljolje, John W. McDonough, Harriet J. Nock, Murat Saraclar, Charles Wooters, and George Zavaliagkos. Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 29(2-4):209–224, 1999. 5

[RBGP12] K. Reidhammer, T. Bocklet, A. Ghoshal, and D. Povey. Revisiting semi-continuous hidden markov models. In *ICASSP*, 2012. 14

[RHT05] Giuseppe Riccardi and Dilek Hakkani-Tür. Active learning: theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(4):504–511, 2005. 25

[RPVC13] Shakti Rath, Daniel Povey, Karel Veselý, and Jan Cernocký. Improved feature processing for deep neural networks. In *ICASSP*, 2013. 96

[SA11] Michael L. Seltzer and Alex Acero. Factored adaptation for separable compensation of speaker and environmental variability. In *ASRU*, pages 146–151, 2011. 50

[SBD95] Ananth Sankar, Françoise Beaufays, and Vassilios Digalakis. Training data clustering for improved speech recognition. In *EUROSPEECH*, 1995. 74

[Set09] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 25

[SHRY13] Andrew Senior, Georg Heigold, Marc'aurelio Ranzato, and Ke Yang. An empirical study of learning rates in deep neural networks for speech recognition. In *ICASSP*, 2013. 97

[SK11] Peter Smit and Mikko Kurimo. Using stacked transformations for recognizing foreign accented speech. In *ICASSP*, pages 5008–5011, 2011. 5, 12, 50

[SMB11] Hagen Soltau, Lidia Mangu, and Fadi Biadsy. From modern standard arabic to levantine asr: Leveraging gale for dialects. In *ASRU*, pages 266–271, 2011. 3, 5, 17, 46

[SRN+12] Tara N. Sainath, Bhuvana Ramabhadran, David Nahamoo, Dimitri Kanevsky, Dirk Van Compernolle, Kris Demuynck, Jort F. Gemmeke, Jerome R. Bellegarda, and Shiva Sundaram. Exemplar-based processing for speech recognition: An overview. *IEEE Signal Process. Mag.*, 29(6):98–113, 2012. 75

[SSK+09] Hagen Soltau, George Saon, Brian Kingsbury, Hong-Kwang Jeff Kuo, Lidia Mangu, Daniel Povey, and Ahmad Emami. Advances in arabic speech transcription at ibm under the darpa gale program. *IEEE Transactions on Audio, Speech & Language Processing*, 17(5):884–894, 2009. 3, 14

[Stü08] Sebastian Stüker. Modified polyphone decision tree specialization for porting multilingual grapheme based asr systems to new languages. In *ICASSP*, pages 4249–4252, 2008. 12

[SW00] T. Schultz and A. Waibel. Polyphone decision tree specialization for language adaptation. In *ICASSP*, 2000. 12, 13, 18

[THTS05] Gökhan Tür, Dilek Z. Hakkani-Tür, and Robert E. Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171–186, 2005. 25

[TO09] Katrin Tomanek and Fredrik Olsson. A web survey on the use of active learning to support annotation of text data. In *Workshop on Active Learning for NLP*, HLT '09, pages 45–48, Stroudsburg, PA, USA, 2009. 25

[Tom00] Laura Mayfield Tomokiyo. Lexical and acoustic modeling of non-native speech in lvscr. In *INTERSPEECH*, pages 346–349, 2000. 4, 5, 12

[Uni] Unisyn lexicon. http://www.cstr.ed.ac.uk/projects/unisyn. 4, 12

[VLG10] Dimitra Vergyri, Lori Lamel, and Jean-Luc Gauvain. Automatic speech recognition of multiple accented english data. In *INTERSPEECH*, pages 1652–1655, 2010. 5, 12

[VOWY97] V. Valtchev, J. J. Odell, Philip C. Woodland, and Steve J. Young. Mmie training of large vocabulary recognition systems. *Speech Communication*, 22(4):303–314, 1997. 91

[VRF10] Ravichander Vipperla, Steve Renals, and Joe Frankel. Augmentation of adaptation data. In *INTERSPEECH*, pages 530–533, 2010. 75, 76

[WC01] Jian Wu and Eric Chang. Cohorts based custom models for rapid speaker and dialect adaptation. In *INTERSPEECH*, pages 1261–1264, 2001. 75, 76

[Wel82] J.C. Wells. *Accents of English:*. Accents of English. Cambridge University Press, 1982. 4

[WN05] Frank Wessel and Hermann Ney. Unsupervised training of acoustic models for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(1):23–31, 2005. 37

[WS03] Zhirong Wang and Tanja Schultz. Non-native spontaneous speech recognition through polyphone decision tree specialization. In *INTERSPEECH*, 2003. 5, 13

[YBM+01] Shinichi Yoshizawa, Akira Baba, Kanako Matsunami, Yuichiro Mera, Miichi Yamada, Akinobu Lee, and Kiyohiro Shikano. Evaluation on unsupervised speaker adaptation based on sufficient hmm statictics of selected speakers. In *INTERSPEECH*, pages 1219–1222, 2001. 76

[YG06] Kai Yu and M. J. F. Gales. Discriminative cluster adaptive training. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1694–1703, 2006. 75

[YGWW10] Kai Yu, Mark J. F. Gales, Lan Wang, and Philip C. Woodland. Unsupervised training and directed manual transcription for LVCSR. *Speech Communication*, 52(7-8):652–663, 2010. 25, 37

[YVDA10] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero. Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion. *Computer Speech & Language*, 24(3), 2010. 25, 26, 27, 39

[ZBB+12] Heiga Zen, Norbert Braunschweiler, Sabine Buchholz, Mark J. F. Gales, Kate Knill, Sacha Krstulovic, and Javier Latorre. Statistical parametric speech synthesis based on speaker and language factorization. *IEEE Transactions on Audio, Speech & Language Processing*, 20(6):1713–1724, 2012. 6

[ZW97] Puming Zhan and Alex Waibel. Vocal tract length normalization for large vocabulary continuous speech recognition. Technical report, Carnegie Mellon University, 1997. 69, 83