Acoustic and Lexical Modeling Techniques for Accented Speech Recognition

Udhyakumar Nallasamy

PHD THESIS PROPOSAL

August 2012

Ph.D. Thesis Proposal

Draft Version: September 10, 2012

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

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

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 wide range of environments including different channels, noise conditions and speakers with varying accents. This thesis focuses on the impact of speakers' accents on the ASR models and techniques to make them robust to such variations. 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. Current approaches to handle accent variations mainly involve adaptation of acoustic models or the pronunciation dictionary.

This thesis examines novel adaptation algorithms capable of modeling changes in phonological realizations, that uniquely characterize accent variations. Techniques that can exploit the contemporary availability of extensive, albeit unlabeled data resources are also investigated. We design experiments under various scenarios where accent adaptation is critical for speech recognition.

In target accent adaptation setup, a source ASR trained on resource-rich accent(s) is adapted to a target accent with limited adaptation data. We propose semi-continuous decision tree adaptation and multi-gram pronunciation models to efficiently model the pronunciation changes between source and target accents. Active and semi-supervised learning are studied to extend the improvements obtained from supervised adaptation. We introduce relevance criteria based data selection to sample additional accent-specific data from large, unlabeled speech corpora with multiple accents.

Finally, we generalize the target accent adaptation techniques to handle multiple accents in the training set. We formulate an accent adaptive training framework using factorized models with shared canonical parameters and accent-specific modules. Our proposed algorithms will be evaluated on Arabic and English accents and compared against existing adaptation techniques.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech recognition research has seen 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 [Matsoukas et al., 2006, Soltau et al., 2009, Gauvain et al., 2005, Hsiao et al., 2009]. This has led to increased adoption of speech recognition technology in desktop, mobile and web platforms for applications such as dictation, voice search [Bacchiani et al., 2008], 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 [Soltau et al., 2011, Nallasamy et al., 2012a]. The Word-Error-Rate (WER) has been to shown to nearly double for mismatched train/test accent pairs in a number of languages such as English [Humphries and Woodland, 1997, Nallasamy et al., 2012a], Arabic [Soltau et al., 2011, Nallasamy et al., 2012a], Mandarin Chinese [Huang et al., 2001a] or Dutch/Flemish accents [Compernolle, 2001]. Moreover, the accent-independent ASR trained on pooled, multiple accents achieves 20% higher WER than accent-specific models [Soltau et al., 2011, Biadsy et al., 2012, Compernolle, 2001].

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, ethnic-

ity 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 [Wells, 1982, 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

- Phonemic inventory - Different accents can have different set of phonemes
- Phonetic realization - Allophones of the same phoneme can be realized differently
- Phonotactic distribution - The distribution of phonemes can be different
- Lexical distribution - Different words can take different phonemes

All of us have an accent and we express both unique and common speech patterns with members of similar accents. 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 [Bael and King, 2003, Tomokiyo, 2000] or automatically learnt using data-driven algorithms [Livescu and Glass, Humphries and Woodland, 1997, Nallasamy et al., 2011]. [Tomokiyo, 2000] used both knowledge-based and data-driven methods to generate pronunciation variants for improving native American ASR on Japanese English. In [Humphries, 1997], 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 [Riley et al., 1999].

The phonetic variations between accents can also be addressed by acoustic adaptation techniques like MLLR/MAP [Leggetter and Woodland, 1995, Gauvain and Lee, 1994] estimation. They are generally model accent variations by linear transforms or Bayesian statistics [Vergyri et al., 2010, Digalakis et al., 1997, Smit and Kurimo, 2011, Tomokiyo, 2000]. However, both MLLR and MAP adaptation are generic adaptation techniques that are not designed to account for the contextual phonological variations presented by the accent. [Clarke and Jurafsky, 2006] 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. Polyphone decision tree in ASR that is used to cluster context-dependent phones based on phonetic question 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. [Wang and Schultz, 2003] 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 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. [Novotney et al., 2011] 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. [Soltau et al., 2011] 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, the 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, [Caballero et al., 2009, Kamper et al., 2012] have used tagged decision trees to train accent-adaptive models. In a similar problem of speaker and language adaptive training in speech synthesis, [Zen et al., 2012] used acoustic factorization to simultaneously train speaker and language adaptive models.

1.3 Expected contributions of the Thesis

- **Target accent adaptation**. We introduce semi-continuous, polyphone decision trees to adapt source accent ASR to target accent using relatively limited adaptation data. We also explore lexical modeling using multigram model based pronunciation adaptation for automatically deriving accent-specific pronunciations using adaptation data. We evaluate these techniques on Arabic and English accents and compare their performance against existing adaptation techniques. *The acoustic adaptation part of the work has been completed and published in [Nallasamy et al., 2012a]. The lexical/pronunciation modeling is part of ongoing work.*

- **Dataselection for Accent adaptation**. We explore active and semi-supervised learning algorithms for the goal of target accent adaptation. We introduce relevance based biased sampling to augment traditional data selection to choose an appropriate subset from a large speech collection with multiple accents. The additional data is used in active or semi-supervised learning to retrain the ASR for additional improvements on the target accent. *Active learning part of the work has been completed and submitted to SLT 2012. The semi-supervised learning has been published in [Nallasamy et al., 2012b].*

- **Accent Robust and Accent Adaptive training**. We introduce an evaluation framework to test various front-ends based on their robustness to accent variations. We analyze the performance of MFCC and Bottle-neck features on a multi-accent Arabic dataset and show that this framework can aid in choosing accent robust features. We propose accent adaptive training using factorized

decision trees and accent-specific dictoinaries to better handle multiple accents in the training data. We evaluate these models on a multi-accented English and Arabic datasets. *Preliminary experiment using the proposed accent robustness criterion has been completed and published in [Nallasamy et al., 2011]. The framework will be extended to English accents with detailed analysis. Accent adaptive training is part of the planned work.*

The tasks listed in each chapter and their status are summarized in Table 5.1.

1.4 Proposal Organization

In chapter 2, we discuss target accent adaptation using semi-continuous polyphone decision tree adaptation. We propose a pronunciation adaptation approach using multigram model to adapt the pronunciation dictionary to target accent. In chapter 3, we explore active and semi-supervised learning in the context of accent adaptation to make use of large amount of easily available unlabeled speech corpus for improved performance on the target test set. We introduce a relevance criterion inaddition to uncertainty or confidence based scores for data selection. In chapter 4, we explore accent robust and accent adaptive techniques to efficiently handle training and test data with multiple accents. We provide a summary of the tasks, their status and the timeline for the remaining work in chapter 5.

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, atleast 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 Previous 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. [Bael and King, 2003] used these rules on the British English BEEP dictionary to create accent-specific ASRs and showed improved performance on cross-accent scenarios. [Tomokiyo, 2000] used both rule-based and data-driven rules to recognize Japanese-accented English. [Humphries and Woodland, 1997, Goronzy et al., 2004, Nallasamy et al., 2011] 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, [Vergyri et al., 2010] used MAP adaptation and compared the performance on multi-accent and cross-accent scenarios. [Livescu, 1999] employed different methods including model interpolation to improve the performance of native American English recognizer on non-native accents. [Smit and Kurimo, 2011] created a stack of transformations to factorize speaker and accent adaptive training and reported improvements on the EMMIE English accent setup. Finally, [Humphries, 1997] 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. [Schultz and Waibel, 2000, Stüker, 2008] 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 [Schultz and Waibel, 2000] and non-native speech [Wang and Schultz, 2003]. 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 3000 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 [Reidhammer et al., 2012]. These models are being widely adopted in ASR having performed better than its counterparts, in both low-resource [Reidhammer et al., 2012] and large-scale systems [Soltau et al., 2009].

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 ≈8000 states, 2.6 times the total number of states in the source ASR (3000)! 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 Experiment Setup - Speech Corpus, Language Model and Lexicon

We evaluate the adaptation techniques on 3 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 [Metze et al., 2010]. 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. 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

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

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 a 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 3K 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 GALE standard test set Dev07. The performance of the baseline SI and SA on source and target accents are shown in Table 3.4. 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 [Soltau et al., 2011]. 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 isnt 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 3K 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 3 different techniques: MAP adaptation, fully-continuous PDTS as formulated in [Schultz and Waibel, 2000] 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 [Huang et al., 2001b]. MLLR is also reported to have issues with accent adaptation [Clarke and

Jurafsky, 2006]. 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 3k states - the adapted fully-continuous PDTS had 256 additional states, while semi-continuous adapted PDTS (SPDTS) ended up with 15K final states (3K codebooks). In a similar fashion, SRC English ASR had 3k states - Adapted English PDTS had 138 additional states while the SPDTS managed 8K final states (3k codebooks). Inspite 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.

MAP adaptation achieves a relative improvement of 9.7% for Levantine Arabic 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 1100 hour large-scale 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 6000 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

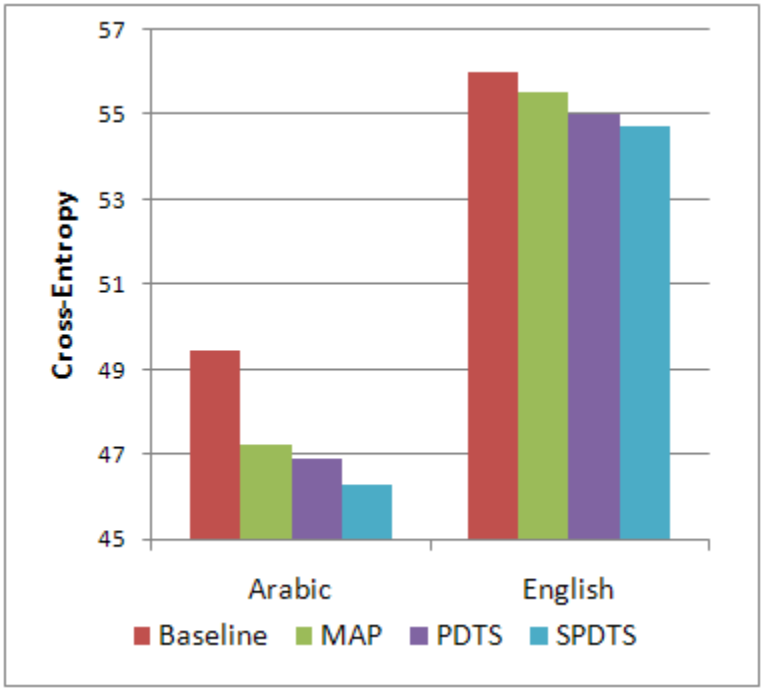


Figure 2.1: *Cross-entropy of adaptation data for various models*

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.

We observe that the adapted models perform better on the target accent, while their performance on the source accent gets worse. We propose to perform a detailed error analysis between the baseline and adapted models for MAP, PDTS and SPDTS techniques to determine the influence of the adapted decision tree on the target accent performance. We aim to verify the hypothesis that the second-level decision tree captures phonological variations, specific to the target accent.

2.7 Proposed Work: Pronunciation adaptation for Accented speech

PDT Adaptation improves the performance of the source model on the target model by modeling the contextual variations between the source and target accents. How-

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

ever, acoustic adaptation by itself cannot account for pronunciation variations introduced by phonetic insertions and deletions. [Jurafsky et al., 2001] investigated triphone based acoustic modeling and concluded that while context-dependent models can account for phone substitutions and prosodic changes, they cannot handle syllable deletions. [Clarke and Jurafsky, 2006] showed that phonetic insertions introduced majority of the errors in MLLR based accent adaptation. To handle all the variations between source and target speech, acoustic adaptation methods should be complemented by a higher-level lexical changes in the pronunciation dictionary.

Pronunciation modeling has a rich history in ASR literature [Strik and Cucchiarini, 1999, Riley et al., 1999] in general and accent adaptation in particular [Humphries, 1997, Livescu and Glass, Tomokiyo, 2000]. It usually involves applying transformation rules to the canonical pronunciations to better match the target accent. These transformation rules are formulated either by expert knowledge or derived automatically from the adaptation data. One of the common techniques is to use a phone-loop recognizer to obtain a ground-truth phonetic transcription. This phone sequence is then compared with the canonical pronunciation obtained from the dictionary to derive the edit rules. The phone-loop recognizer is error-prone as it does not use any lexical information while decoding. Hence, the rule learning should account for noisy observations in the ground-truth phonetic sequence.

[Humphries, 1997] employed decision trees to automatically learn the transformation rules using the aligned phonetic sequences. The same linguistic questions used by the ASR polyphone decision tree was used during training. Once the decision trees are trained, they are employed to edit the canonical dictionary to

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

create new variants. The resulting dictionary is used in decoding of the test data. Experiments on a British English dictionary modified for American English showed consistent improvements on the target test set. One of the limitations of this decision tree based models is that the final phonetic sequence is obtained by a series of locally optimal decisions that may not be globally optimal. It is also cumbersome to generate n-best pronunciations using decision tree based model, for any further re-scoring. Each transformation rule originates a possible pronunciation variant, so the number of possible variants soon become quite large.

A similar problem of automatic grapheme-to-phoneme (G2P) conversion replaced decision tree based models using probabilistic multigram or graphone models with significant improvements [Bisani and Ney, 2008]. [Li et al., 2007] extended the idea for pronunciation adaptation using acoustic information on a name recognition task. They interpolated the graphone LM trained on limited target data with

a background LM trained on generic pronunciations to accomplish pronunciation adaptation. [Li et al., 2011] used a similar LM interpolation technique to perform G2P on dialectal Arabic words with limited adaptation data and large amount of Modern Standard Arabic pronunciations. The background model trained on large canonical dictionary is combined with the small amount of pronunciations obtained from the adaptation data. Both interpolation based smoothing and data combination are compared as methods to address the data sparsity in training a multigram graphone LM on the adaptation data which is used to transform grapheme into phones in the target domain. The authors found that just combining the source and target data produced better improvements than interpolation. This is mainly due to poor alignments obtained solely from the adaptation data before interpolation. The alignment requires large amount of data to reliably match the grapheme and phone sequences.

We propose to improve on graphone LM adaptation by using a multigram LM trained on canonical and ground-truth phone sequences. Such a model will transform the canonical pronunciations in the dictionary to accent-specific variants. The following section will introduce our proposed model for pronunciation adaptation.

2.8 Multigram pronunciation adaptation model

One of the drawbacks of graphone model adaptation using LM interpolation [Li et al., 2007, 2011], is that it requires the target graphone LM which is trained on limited adaptation data. Training a graphone LM requires reliable alignment of phone and grapheme sequences, which is not accurate with limited examples. As a result, pronunciation adaptation using LM interpolation is as good or worse than training a graphone model by pooling both the background and target pronunciations [Li et al., 2007]. To avoid the harder problem of learning graphone alignments using a small amount of adaptation data, we propose to learn a mapping between canonical and accent adapted phone sequences. These sequences are in the common (phonetic) space and the alignment can be performed with simple dynamic string matching. This allows us to train the phone-to-phone transformation LM that can adapt the canonical pronunciations to the target accent. We first use the available canonical dictionary to train a graphone model. The resulting LM is converted into a finite state transducer, which produces a phonetic sequence given a grapheme sequence. We then use a phone-loop ASR to obtain the ground-truth phonetic sequence. We employ acoustically adapted models during decoding. We then use the canonical

and the ground-truth phone sequences to train a second-level, phone-to-phone multigram LM. The resulting FST is then composed with the original canonical G2P transducer to obtain accent-specific pronunciations. The training process is illustrated in the following diagram.

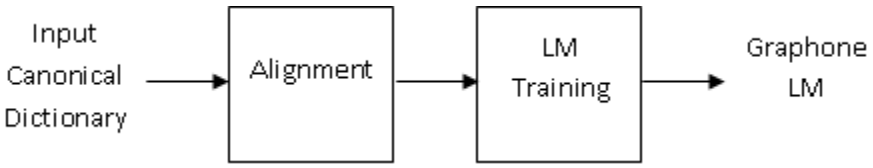


Figure 2.2: *Step 1: Graphone LM Training*

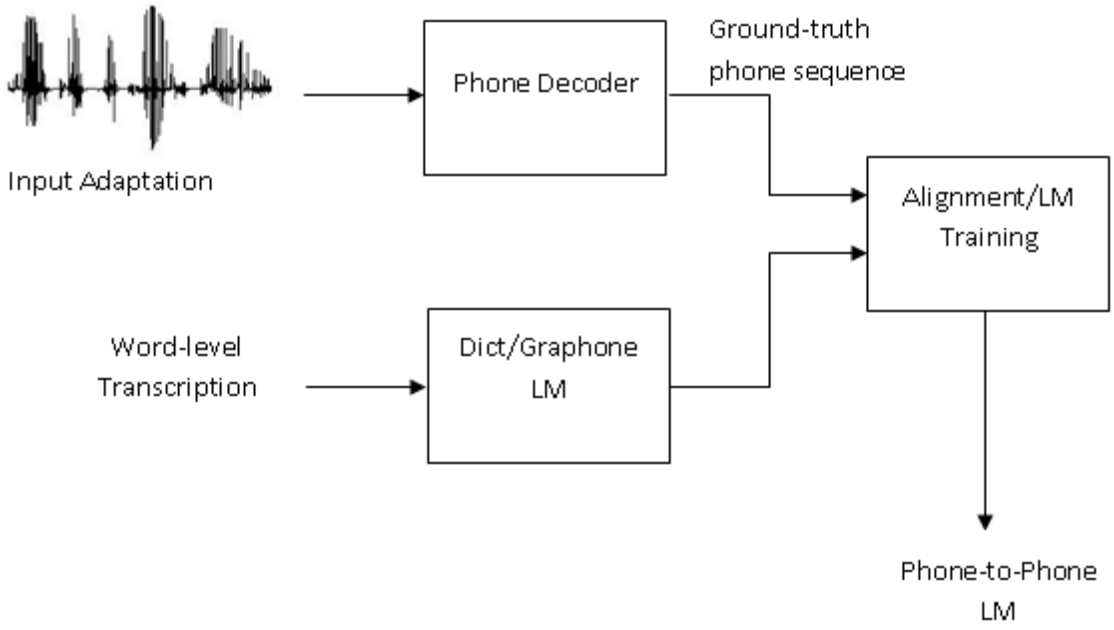


Figure 2.3: *Step 2: Phone-to-Phone LM Training*

In the first step, the canonical dictionary is used to train a graphone based G2P model. We use a many-to-one alignment with LM training on 1-best match between grapheme and phone sequence. During adaptation, we obtain the phone sequence for each utterance using the canonical dictionary. The pronunciations for words not present in the dictionary are derived from the graphone LM. The speech data

for the utterance is used to produce a ground-truth phonetic sequence using a phone-loop decoder. Adapted context-independent models are used to generate the phone-sequence. Confidence scores can be used to prune away any outliers in the hypothesis of the decoder. We align the canonical and ground-truth phone sequences using string matching and train a phone-to-phone multigram LM.

During dictionary generation, the input vocabulary is appended with canonical pronunciation using the dictionary. Again, the words not in the dictionary are processed with the graphone LM. The canonical phone sequence is processed with the phone-to-phone LM to obtain the accent-specific pronunciations. The newly created dictionary is used to decode the target test data. We plan to evaluate the model on WSJ task, adapting American English CMU dictionary to British English. We will test the effect of pronunciation adaptation on the test data and compare it against a manually created British English BEEP dictionary. We will also experiment the technique on a large scale M*Modal dataset with native American and South-Asian English accents. The large amount of speaker-specific data (≈ 100 hours of speech per speaker) this database will allow us to build accent-specific and even speaker-specific pronunciation dictionaries and evaluate their performance. Experiments on the combination of pronunciation and acoustic adaptation will be carried out and results will be reported.

2.9 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 5.1% relative for large scale systems. We have proposed to conduct an error analysis to determine the influence of the semi-continuous, decision tree in modeling the phonological variations in the target accent. We have proposed a multi-gram based pronunciation model for accent adaptation at the lexical level. We will investigate its performance on the WSJ and M*Modal accent adaptation tasks and report the results. If successfull, we will also explore different ways of combining acoustic and pronunciation adaptation techniques for further improvements.

Chapter 3

Dataselection for Accent Adaptation

Supervised adaptation using MAP/SPDTS requires transcribed target data for adapting the source model to the target accent. As we 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, such datasets seldom have accent markers or transcriptions. To make use of these large speech collections, we explore active and semi-supervised accent adaptation in this chapter.

3.1 Active Learning

Active learning is a commonly used machine learning technique in fields where the cost of labeling the data is quite high Settles [2009]. 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 Tomanek and Olsson [2009], spoken language understanding Tür et al. [2005], speech recognition Riccardi and Hakkani-Tür [2005], Yu et al. [2010b,a], Itoh et al. [2012], etc.

Many of the approaches in active learning, relied 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 Hakkani-Tür et al. [2002] to predict uncertainty. Lattice Yu et al. [2010a] and N-best Itoh et al. [2012] 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 Huang et al. [2010], Itoh et al. [2012].

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. This corpus has neither transcriptions nor accent labels. The goal of active learning here, is to choose 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, which 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.

We introduce a relevance criterion in addition to uncertainty based informative measure for data selection to match the target accent. We start with the ASR trained on a source accent. We use a relatively small, manually labeled adaptation data to adapt the recognizer to the target accent. We employ the adapted model to choose utterances from a large, untranscribed mixed dataset for human transcription, to further improve the performance on the target accent. To this end, we calculate cross-entropy based measure based on adapted and unadapted model likelihoods, to assess the relevance of an utterance. We combine this measure with uncertainty based sampling to choose an appropriate subset for manual labeling. We evaluate our

technique on Arabic and English accents and we 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 Mangu et al. [2000]. Confidence scores calculated on 1-best hypothesis are sensitive to outliers and noisy utterances. Yu et al. [2010a] proposed lattice-entropy based measure and selecting utterances based on global entropy reduction. Itoh et al. [2012] observed that lattice-entropy is correlated with the utterance length and showed N-best entropy to be an empirically better criterion. In this work, we also use a entropy-based measure as informative criterion for data selection. We calculate the average entropy of the alignments in the confusion network as a measure of uncertainty of the utterance with respect to the ASR. It is given by

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

where $E_{\mathcal{A}}$ is the entropy of an alignment A in the confusion network and $T_{\mathcal{A}}$ 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, we derive cross-entropy based relevance criteria for choosing utterances from the mixed set, for human annotation. We formulate the source-target mismatch as a sample selection bias problem Cortes et al. [2008], Blitzer and III [2010], Bickel et al. [2009] 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. We derive the biased sampling criterion for both the multi-accented and mis-matched accent cases separately in the following section.

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 Moore and Lewis [2010], Nallasamy et al. [2012b] 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 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 Itoh et al. [2012]. 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

We conducted active learning experiments 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. We assinged Levantine as our 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, we chose English WallStreet Journal (WSJ1) as our source data, as in the previous chapter. We used British English as our target accent and the British version of WSJ corpus (WSJCAM0) for adaptation. We randomly

sampled 3 hours from WSJCAM0 for our 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

We built HMM-based, speaker-independent ASR systems 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 Metze et al. [2010], Nallasamy et al. [2012a]. We initially adapt our baselines systems on the relatively small, manually labeled, target adaptation dataset. We used semi-continuous polyphone decision tree adaptation (SPDTS) Nallasamy et al. [2012a] for the supervised adaptation. The Word Error Rate (WER) of the baselines and supervised adaptation systems are given in Table 3.2.

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

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 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.

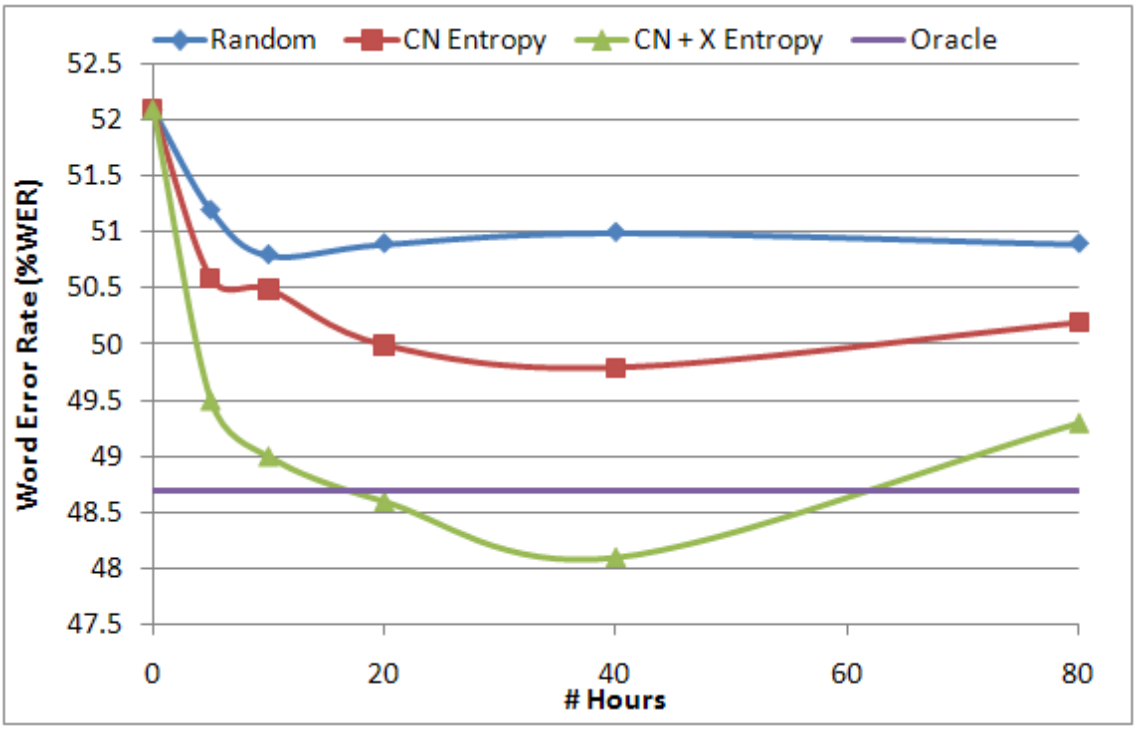


Figure 3.1: *Active learning results for Arabic*

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.

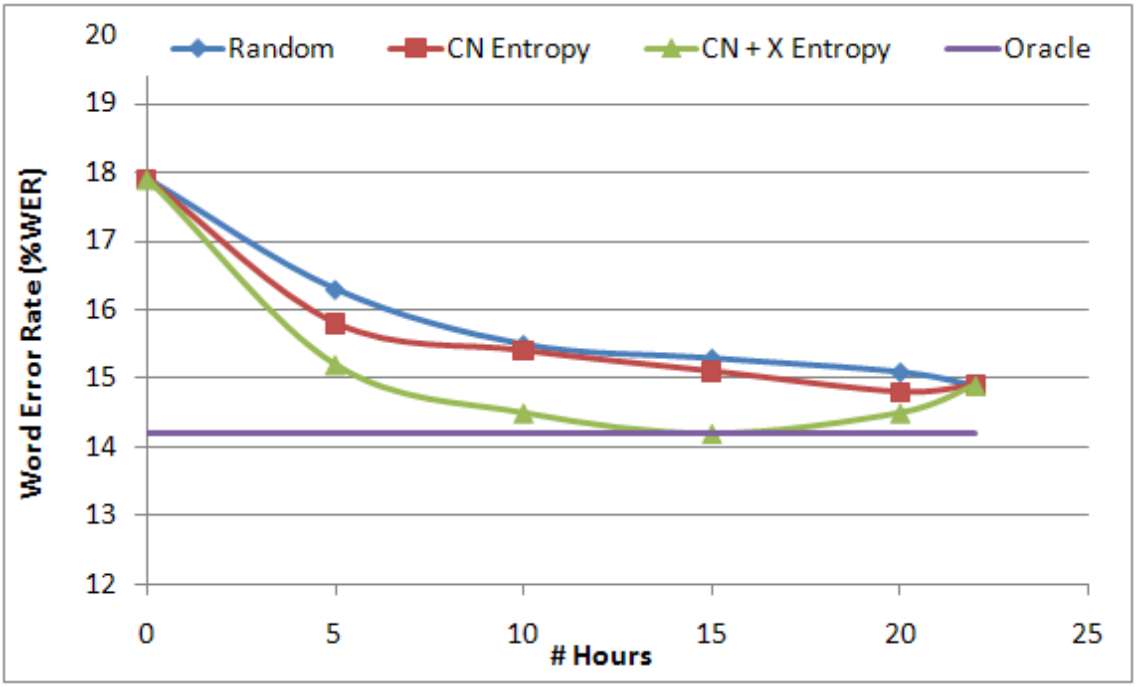
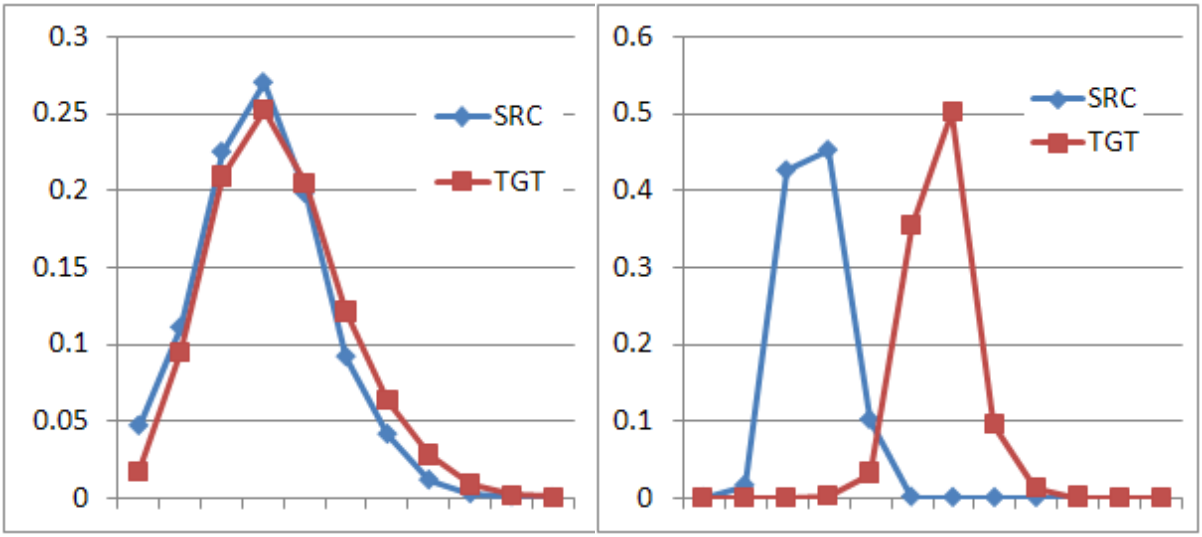


Figure 3.2: *Active learning results for English*

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 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



(a) Entropy (b) Weighted Score

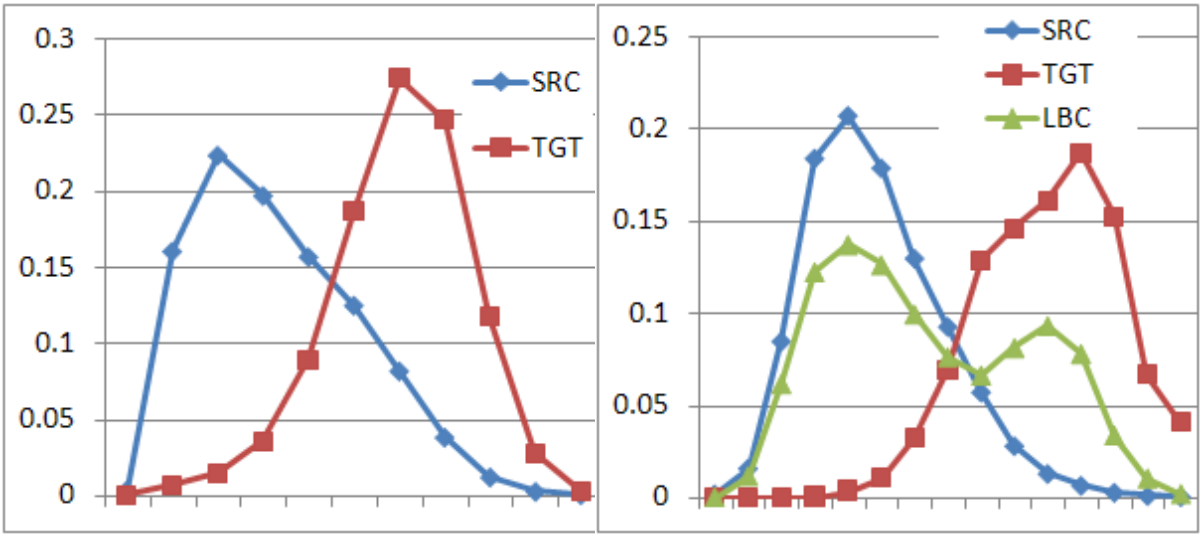
Figure 3.3: Histogram of source and target scores for English.

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

3.2 Semisupervised 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 Yu et al. [2010b], Wessel and Ney [2005], Kemp and Waibel



(a) Entropy (b) Weighted Score

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

[1999], Ramabhadran [2005], Ma and Schwartz [2008], 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 Lööf et al. [2009], Novotney et al. [2011]. In the latter task the 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 Mangu et al. [2000]. 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 Yu et al. [2010a], Fraga-Silva et al. [2011]. 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 Blitzer and III [2010], Cortes et al. [2008], Bickel et al. [2009]. 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

Algorithm 2 Cross-entropy based semi-supervised learning

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

$\lambda_T := Adapt(\lambda_S, \mathcal{X}_T)$
for all x in $\mathcal{X}_M$ **do**
    $Loglike_S := Score(\lambda_S, x)$
    $Loglike_T := Score(\lambda_T, x)$
    $Len := Length(x)$
    $Score := (Loglike_T - Loglike_S)/Len$
    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

use enhanced polyphone decision tree adaptation based on semi-continuous models (SPDTS) Nallasamy et al. [2012a] 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 Gauvain and Lee [1994] to refine the Gaussians (codebooks) and associated mixture weights (distributions) on the adaptation data. SPDTS gives additional improvements over the traditional MAP adaptation.

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 Povey and Yao [2012]. 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 Mangu et al. [2000]. 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. For baseline, 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 Word-Error Rates (WER) of the baseline and adapted systems.

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.

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

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

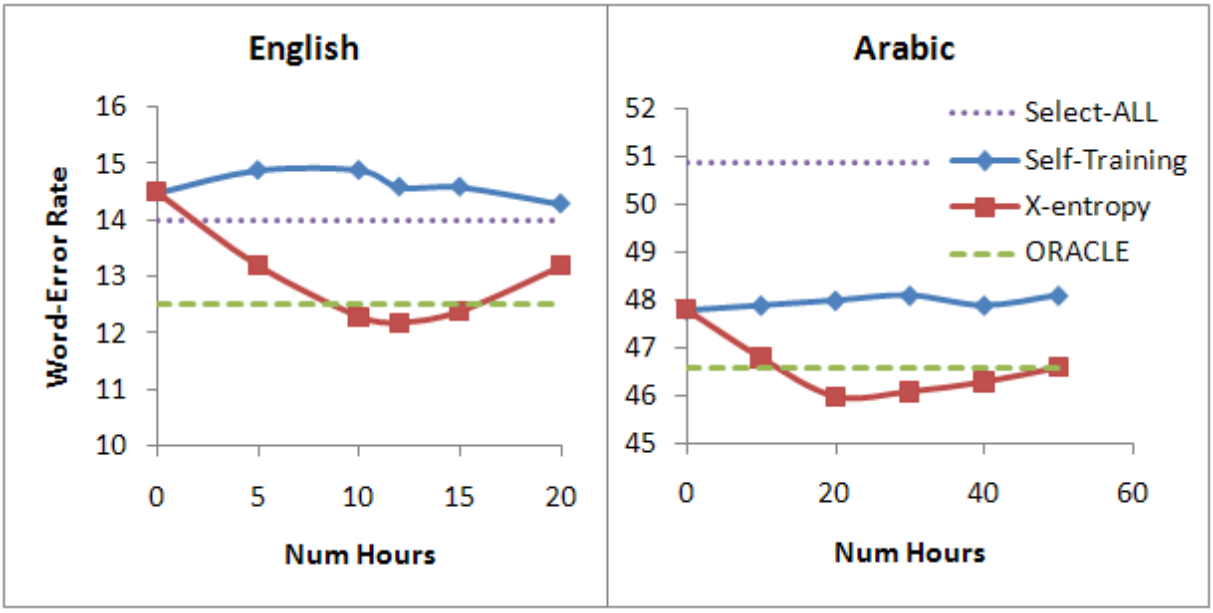


Figure 3.5: *Semi-supervised data selection with transcriptions*

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 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. We also found 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. More analysis is needed to explore the characteristics of the speakers selected for Arabic from the BC portion of the mixed set.

Task 2 - Mixed set without transcriptions and no accent labels

We used the same framework and bins 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.

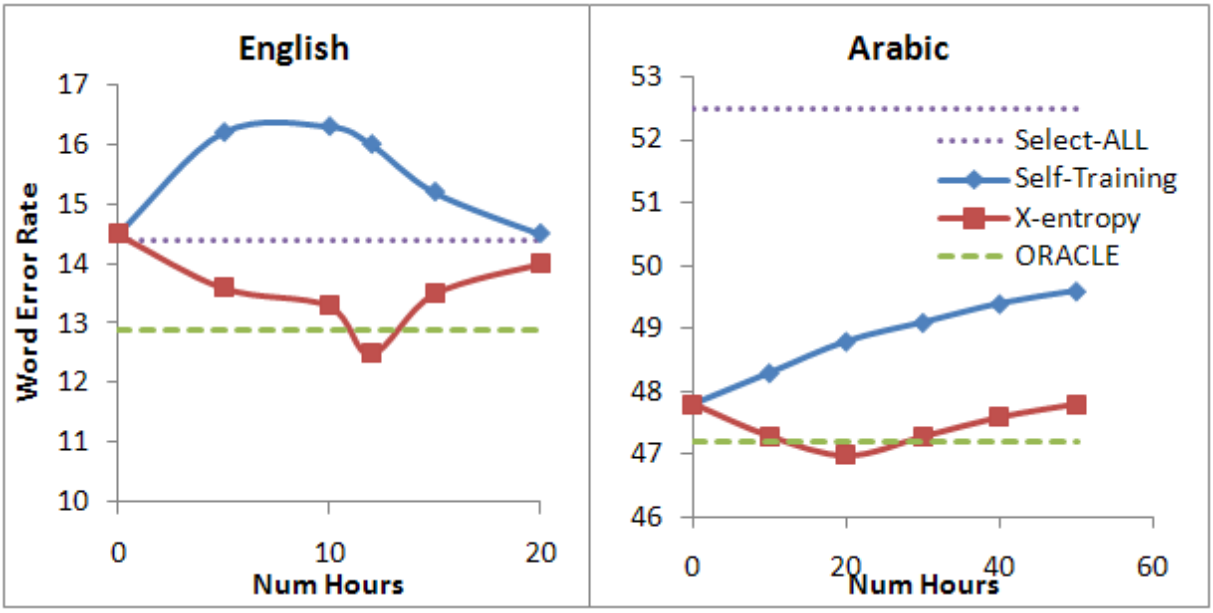


Figure 3.6: *Semi-supervised data selection without transcriptions*

As expected, the results are similar to the previous experiment as self-training fails to obtain any additional improvements with the mixed data. We get 2% absolute (or 13.8% relative @12 hours) improvement 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 Soltau et al. [2011].

3.3 Summary

In this chapter, we investigated the use of additional untranscribed data for the goal of accent adaptation. We proposed a relevance criterion based biased sampling, 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.

Chapter 4

Accent Robust and Accent Adaptive training

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 topics in this chapter.

4.1 Previous 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 [Hermansky and Jr., 1991] and RASTA [Hermansky and Morgan, 1994] have been proposed earlier. Probabilistic front-ends based on Multi-Layer Perceptron (MLP) have also been tested for their noise robustness [Ikbal et al., 2004]. A review of feature-based and

model based techniques for noise robustness in speech recognition is presented in [Deng, 2011, Gales, 2011b]. 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 [Chengalvarayan, 2001, Caballero et al., 2009, Kamper et al., 2012] and using individual distributions while sharing the codebooks [Kamper et al., 2012]. [Smit and Kurimo, 2011] introduced stacked transforms, a two-level MLLR transforms to integrate accent and speaker adaptation, similar to factorized CMLLR proposed in [Seltzer and Acero, 2011]. As in normalization, accent adaptive training has also commonalities with speaker [Gales, 2011a] and noise [Kim and Gales, 2009] 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 is 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 which 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.

4.2.2 Dataset

All our experiments are carried out on the Pan-Arabic dataset provided by AFRL. The 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 of 2 speakers. The amount of data broken down according to accent is shown in Table 4.1 below.

Each speaker is recorded in separate channels, including long silences between

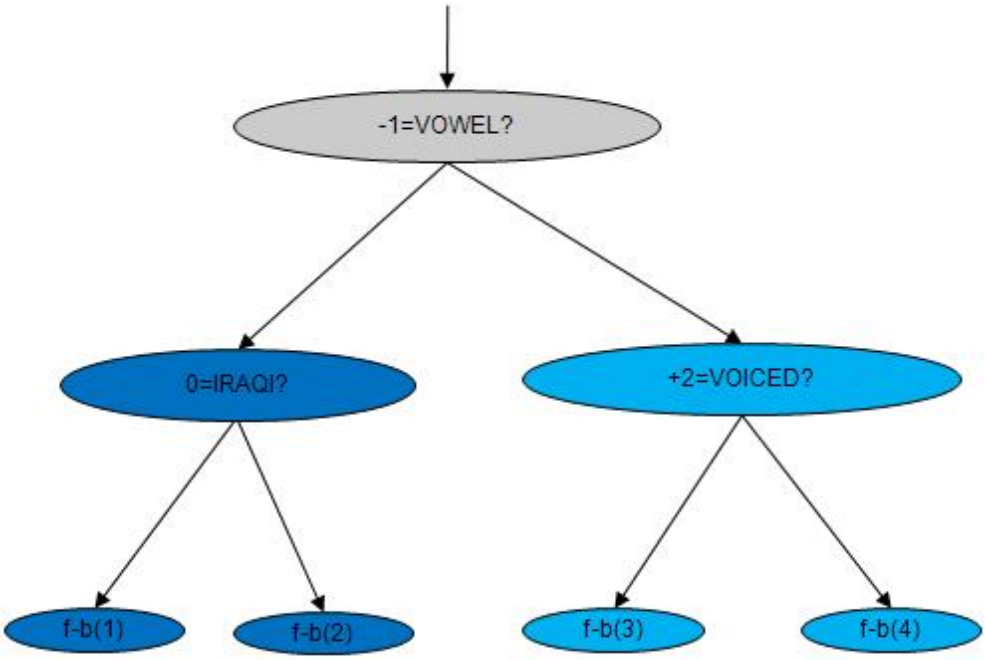


Figure 4.1: *Decision tree for begin state of /f/*

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 postalveolar fricative, the /Z/ phone instead. These 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.

4.2.3 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

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

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 [Grézl and Fousek, 2008, Frankel et al., 2008]. A multilayer 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 [Grézl et al., 2007] 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

4.2.4 Preliminary experiments

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.

We noticed that most accent models belong to Egyptian across different model

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%

sizes. This behavior is consistent with the results found in the literature, where Egyptian is found to be most distinguishable from other accents [Biadsy et al., 2010]. 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%

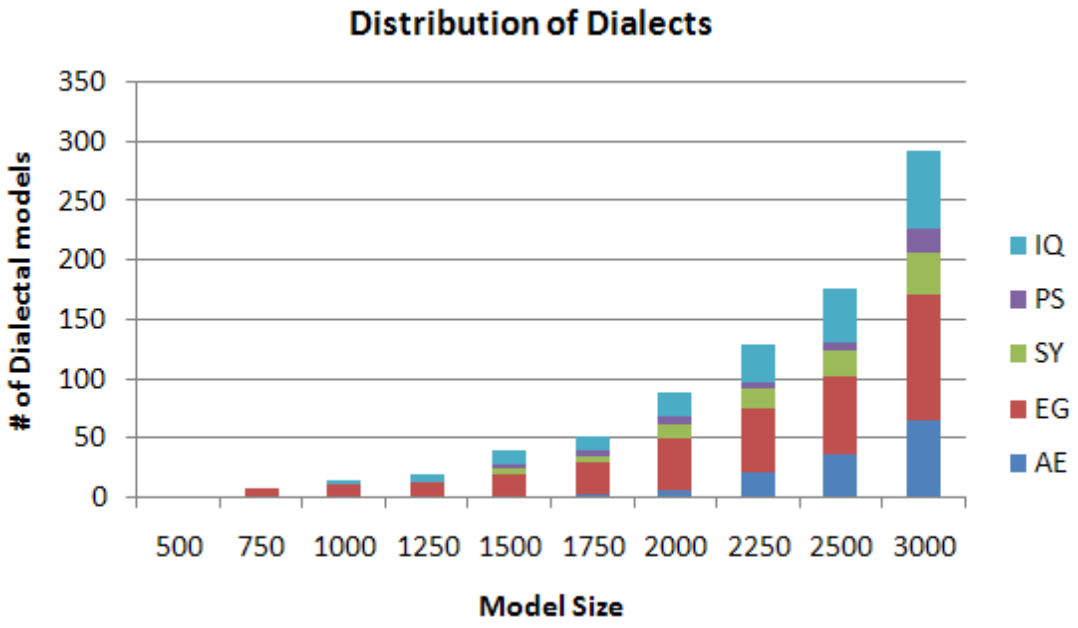


Figure 4.2: *Accent Distribution in MFCC models*

4.2.5 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

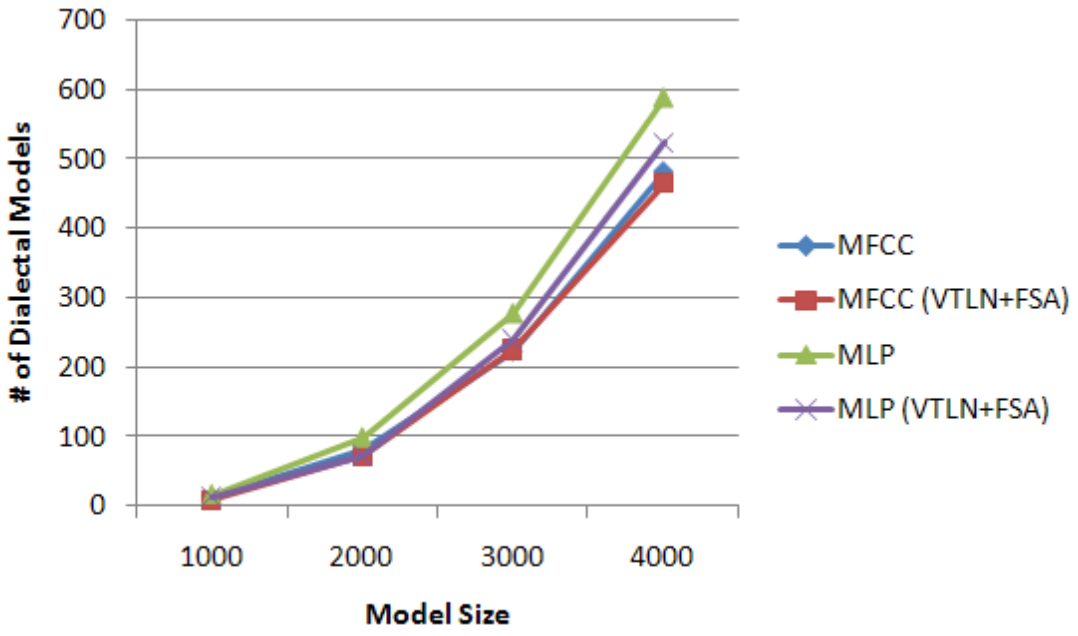


Figure 4.3: *MLP vs. MFCC models*

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.

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.

We have presented an evaluation framework to test different front-ends for their

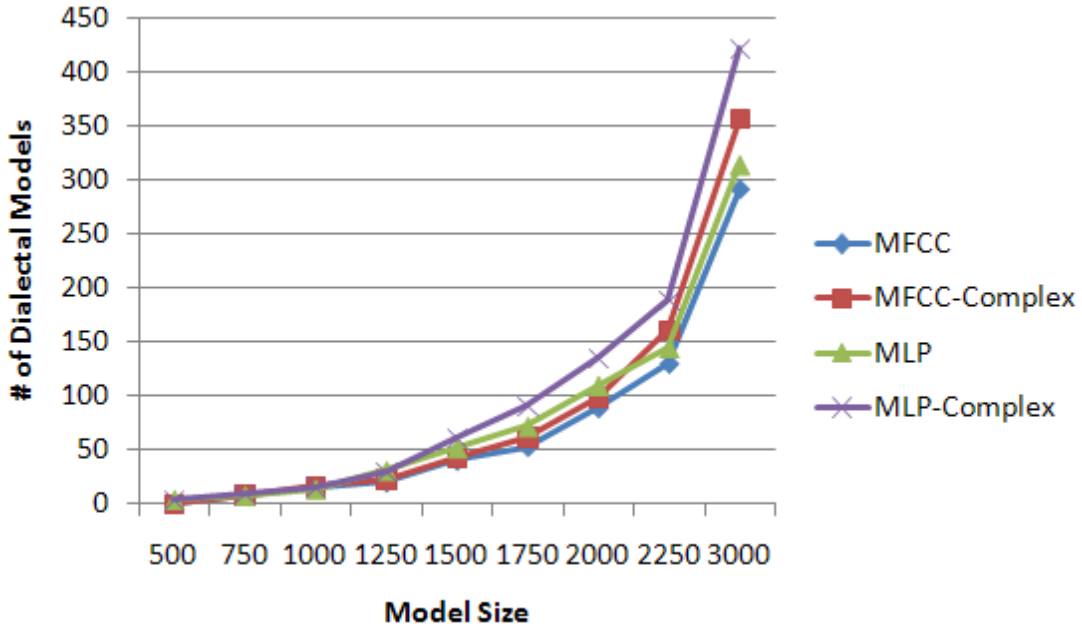


Figure 4.4: *Single Pronunciation models*

ability to normalize accent variations. We analyzed MFCC and MLP front-ends and showed that although MLPs are better at gender normalization than MFCCs, they are slightly more sensitive to accent variations than MFCCs. We investigated the MLP frame accuracies and hypothesized that their sensitivity could stem from lower accuracy on the vowels which are highly influenced by accent variations. The analysis also showed us the characteristics of the accents and how they are related.

We propose to extend this framework to multi-accented English datasets including the ones used in chapters 2 and 3. We will evaluate different front-ends including articulatory features, MLP Bottle-neck features and MFCC to rank them according to the decision tree based accent robustness criterion. We will analyze the robustness criterion of these front-ends to reveal any relation between the characteristics of the English accents used in this experiment. The word error rate of these front-ends on a multi-accented English test set will also be reported.

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

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

4.3 Proposed Work: Accent Adaptive training

In this section, we propose to formulate a training procedure for accent adaptive training. In our framework, we simultaneously train accent-dependent and accent-independent parameters on a dataset with multiple accents. To achieve this, we generalize the techniques proposed for target accent adaptation in chapter 2 for multiple accents. We propose both acoustic and lexical-level adaptive training to efficiently train a accent-adaptive model that can handle different accents in the training set. The following sections details the techniques and experiments for each case.

4.3.1 Accent Adaptive training - Acoustic Level

In this section, we extend the semi-continuous decision tree based adaptation for target accent adaptation to multiple accent scenario. We grow multiple, accent-specific, two-level decision trees and train accent-dependent distributions, while maintaining a common set of shared codebooks. We aim to integrate the accent-adaptive training with speaker-adaptive training, so the factorization schemes benefit from each other. The speaker-specific parameters are characterized by CMLLR transforms, while the accent-specific parameters are chosen to be two-level decision trees. The accent-adaptive model is shown below.

Training Procedure

As shown in Figure 4.5, the acoustic level accent training consists of the following steps:

- We train the CI model, its codebooks and distributions are trained on a common pool of data with all the accents combined.

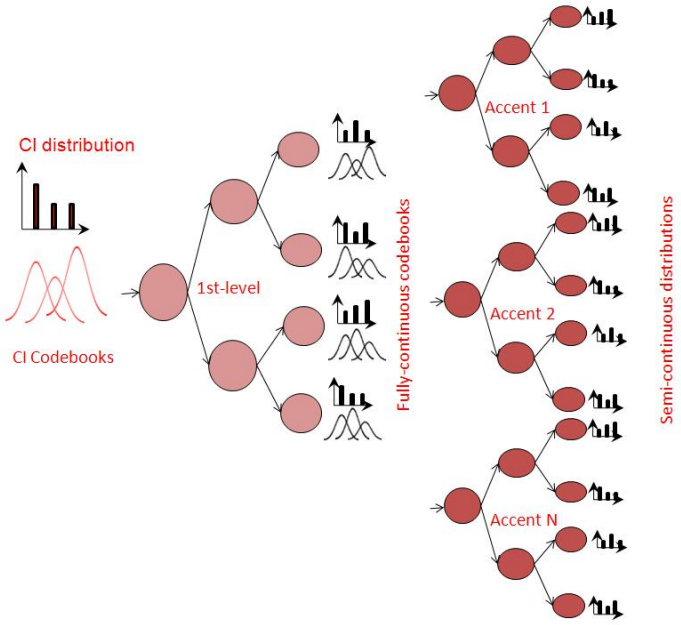


Figure 4.5: *Accent Adaptive Model*

- We accumulate statistics for each context to grow a first-level, fully continuous model.
- This model is then further split, using multiple, two-level decision trees along the distributions for each accent.
- During training procedure, for each speaker, we load the appropriate accent tree based on his/her accent and accumulate the codebook-level and distribution-level statistics.
- In the update phase, codebook-level statistics are merged across all accents to estimate accent-independent codebooks. The distributions for the individual accents are independently updated.

One of the critical aspects of this model is the parameter ratio between shared and accent-specific parameters. Ideally, we would like to delegate accent-independent contextual modeling to the first-level decision tree, while accent-dependent contexts are assined to the multiple second-level decision trees. We propose to identify the accent-dependent contexts by introducing accent-questions as in the accent

normalization analysis. We stop growing the first-level tree once an accent question is encountered. These leaves are then further split using the second-level distribution tree with accent-specific parameters. We will evaluate this model on a multi-accented dataset for Arabic and English accents. Our baseline will be a multi-style trained system which is trained on data pooled from all the accents. We will perform a detailed error analysis to show that the accent-dependent, second-level trees help achieve accent-adaptive training, while sharing accent-independent parameters.

4.3.2 Accent Adaptive training - Lexical Level

Accent adaptive training can also be carried out at the lexical level, by employing accent-specific pronunciations for each accent group during training. We extend our lexical adaptation technique proposed in chapter 2 to transform a canonical dictionary into accent-specific dictionaries, which are then employed in re-training the acoustic models. The transformation rules are derived by comparing the ground-truth sequence obtained by a phone decoder to the canonical phone sequence. A phone-to-phone LM is trained individually for each accent. These FSTs can be applied to the original dictionary to get matching pronunciation variants for each accent. These pronunciation variants are chosen based on the accent of a particular speaker which training or decoding. Accent adaptive training at the lexical level will involve the following steps

- We train the accent-independent model using data from all accents combined.
- We use this model to generate ground truth phone sequence using a phone loop decoding for all the utterances in the training set. These sequences are then paired up with the pronunciations from the canonical dictionary.
- Accent-specific multi-gram phone-to-phone transformation LMs are trained using the paired phone sequences from the respective accents.
- These transformations are applied to the canonical pronunciation dictionary to derive accent-specific dictionaries.
- During training, the appropriate accent-specific pronunciation dictionary is chosen based on the accent of the speaker.

The final model will be evaluated on dataset with multiple accents. During decoding, the same dictionary selection procedure is repeated. The results are compared against the model trained with the canonical dictionary. A detailed error analysis to compare the difference in performance between accent-specific and canonical pronunciations will be performed and the results will be reported. The final system will be combined with acoustic-level adaptive training for any additional gains.

4.4 Summary

We proposed an evaluation framework to test different front-ends on their accent normalization ability. We analyzed MFCC and MLP front-ends and concluded what MLP is more sensitive to accent variations than MFCC. We investigated the results and identified the poor accuracy of the MLP on vowels could lead to its sensitivity to different accents. We have proposed to extend this framework for English accents and analyze the robustness of various front-ends. We also proposed accent adaptive training by using factorized models to model the shared and unique patterns between different accents. We plan to perform experiments with accent adaptive training at acoustic and lexical level and compare its performance against a simple multi-style baseline.

Chapter 5

Tasks and Timeline

In this section, we outline the problems addressed in this proposal, along with the results obtained, work to be done and datasets used. This proposal starts by addressing the simple adaptation problem in accent modeling - Adapting a source accent to a specific target accent using relatively small amount of transcribed adaptation data (Chapter 2). We propose acoustic and lexical/pronunciation adaptation techniques to address this problem. In acoustic adaptation, we introduced semi-continuous, two-level decision tree based accent adaptation and showed that it out-performed conventional adaptation techniques [Nallasamy et al., 2012a]. We will investigate pronunciation adaptation using phone-to-phone multigrams to model lexical level accent variations. We will analyze the results to examine the type of accent variations modeled by each technique and explore a suitable combination to benefit from complimentary nature of acoustic and lexical accent adaptation.

The gains from the supervised adaptation are improved further by making use of large amount of untranscribed data with multiple accents, for target accent adaptation (Chapter 3). Data selection using a relevance criterion is carried out under active and semi-supervised learning to select appropriate subset for retraining the target ASR. This criterion allowed us to identify useful unlabeled target accent data that further improved a ASR adapted on a limited amount of transcribed data, in a supervised manner [Nallasamy et al., 2012b].

Finally, we deal with a practical scenario of handling different accents in the training dataset (Chapter 4). We introduce decision tree based accent normalization criterion by adding accent-level questions in the ASR decision tree. The ratio of accent-dependent to total leaves conveys the normalization ability of a front-end. We found that the normalization or robustness behavior of a front-end is influenced

by language-specific characteristics, such as in the case of MLP features whose low vowel accuracy led to its increased sensitivity to accent variations in Arabic [Nallasamy et al., 2011]. The accent tags on the decision trees can also reveal the relationship between different accents which allows us to model them efficiently. We propose to extend this framework to analyze various front-ends on their robustness to English accents. As in the case of Arabic, our analysis will also reveal interesting relationships between the different accents in the database.

We have proposed accent adaptive training to simultaneously train accent-dependent and canonical parameters from a multi-accented dataset. We will explore both acoustic and lexical level approaches to address this problem. In both cases, we generalize the techniques experimented in target accent adaptation to design algorithms that can handle phonological variations between multiple accents in the database.

Table 5.1 shows a list of tasks that are part of this thesis and their status. Table 5.2 shows the timeline for completion of the thesis.

5.1 Remaining Work

We list the remaining work in each chapter below.

- **Target accent adaptation**.
    1. Pronunciation modeling for accent adaptation
    2. Combination of pronunciation adaptation with acoustic adaptation.
    3. Performance analysis of both approaches and diagnostic experiments to understand their modeling of accent variations.
- **Accent Robust and Adaptive training**.
    1. Accent robust analysis of different front-ends on accented English.
    2. Accent adaptive training model using multiple accent-factorized two-level decision trees.
    3. Extension of pronunciation modeling to multiple accents.

Table 5.1: *Tasks and their status* .

Chapter	Task	Datasets used	Status	Results obtained
Target Accent Adaptation	Acoustic adaptation based on semi-continuous decision trees [Nallasamy et al., 2012a]	WSJ English and GALE Arabic	Experiments Completed. More analysis proposed	7-13.6% relative improvement over MAP adaptation
Target Accent Adaptation	Lexical adaptation based on multi-gram LM	WSJ English and M*Modal English	Proposed	-
Data selection	Active learning [Submitted to SLT 2012]	WSJ English and GALE Arabic	Completed	7.7-20.7% rel improvement over supervised baseline
Data selection	Semi-supervised learning [Nallasamy et al., 2012b]	WSJ English and GALE Arabic	Completed	2.0-15.9% rel improvement over supervised baseline
Accent Robust and Adaptive training	Accent normalization [Nallasamy et al., 2011]	RADC Pan-Arabic dataset and WSJ English/GALE Arabic	Preliminary experiments Completed, Experiments on English accent and more analysis proposed	Evaluated and analyzed MLP and MFCC front-ends on Arabic accents.
Accent Robust and Adaptive training	Accent Adaptive training - Acoustic Level	WSJ English and GALE Arabic	Proposed	-
Accent Robust and Adaptive training	Accent Adaptive training - Lexical Level	WSJ English and M*Modal English	Proposed	-

Table 5.2: *Timeline for the thesis* .

Task	Time
Pronunciation Modeling	Oct 1, 2012 - Dec 31, 2012
Combination and Diagnostic experiments	Jan 1, 2013 - Mar 1, 2013
Accent Robustness for English accents	Mar 1, 2013 - Apr 1, 2013
Accent Adaptive training - Acoustic level	Apr 1, 2013 - June 1, 2013
Accent Adaptive training - Lexical level	June 1, 2013 - Aug 1, 2013
Thesis writing and wrap up	Aug 1, 2013 - Oct 1, 2013
Defense	Nov 1, 2013

Bibliography

Accents research. http://www.phon.ucl.ac.uk/home/mark/accent. 2

Quicknet toolkit. http://www1.icsi.berkeley.edu/Speech/qn.html. 51

Unisyn lexicon. http://www.cstr.ed.ac.uk/projects/unisyn. 2, 8

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. 1

Christophe Van Bael and Simon King. An accent-independent lexicon for automatic speech recognition. In *ICPhS*, pages 1165–1168, 2003. 3, 8

Fadi Biadsy, Julia Hirschberg, and Michael Collins. Dialect recognition using a phone-gmm-supervector-based svm kernel. In *INTERSPEECH*, pages 753–756, 2010. 53

Fadi Biadsy, Pedro Moreno, and Martin Jansche. Google's cross-dialect arabic voice search. In *ICASSP*, 2012. 1

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10, 2009. 25, 38

Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451, 2008. 18

John Blitzer and Hal Daumé III. ICML tutorial on domain adaptation. http://adaptationtutorial.blitzer.com, June 2010. 25, 38

Mónica Caballero, Asunción Moreno, and Albino Nogueiras. Multidialectal spanish acoustic modeling for speech recognition. *Speech Communication*, 51(3):217–229, 2009. 4, 48

Rathinavelu Chengalvarayan. Accent-independent universal hmm-based speech recognizer for american, australian and british english. In *INTERSPEECH*, pages 2733–2736, 2001. 48

Constance Clarke and Daniel Jurafsky. Limitations of mllr adaptation with spanish-accented english: an error analysis. In *INTERSPEECH*, 2006. 3, 14, 17

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

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *ALT*, pages 38–53, 2008. 25, 38

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. 48

V. Digalakis, V. Digalakis, L. Neumeyer, and J. Kaja. Development of dialect-specific speech recognizers using adaptation methods. In *ICASSP*, pages 1455–1458, 1997. 3

Thiago Fraga-Silva, Jean-Luc Gauvain, and Lori Lamel. Lattice-based unsupervised acoustic model training. In *ICASSP*, pages 4656–4659, 2011. 37

Joe Frankel, Dong Wang, and Simon King. Growing bottleneck features for tandem asr. In *INTERSPEECH*, page 1549, 2008. 51

M. J. F. Gales. Model-based approaches to handling uncertainty. In *Robust Speech Recognition of Uncertain or Missing Data*, pages 101–125. 2011a. 48

M. J. F. Gales. Model-based approaches to handling uncertainty. In *Robust Speech Recognition of Uncertain or Missing Data*, pages 101–125. 2011b. 48

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. 3, 39

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. 1

Silke Goronzy, Stefan Rapp, and Ralf Kompe. Generating non-native pronunciation variants for lexicon adaptation. *Speech Communication*, 42(1):109–123, 2004. 8

Frantisek Grézl and Petr Fousek. Optimizing bottle-neck features for lvcsr. In *ICASSP*, pages 4729–4732, 2008. 51

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. doi: 10.1109/ICASSP.2007.367023. 52

Dilek Z. Hakkani-Tür, Giuseppe Riccardi, and Allen L. Gorin. Active learning for automatic speech recognition. In *ICASSP*, pages 3904–3907, 2002. 24

Hynek Hermansky and Louis Anthony Cox Jr. Perceptual linear predictive (plp) analysis-resynthesis technique. In *EUROSPEECH*, 1991. 47

Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994. 47

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. 1

Chao Huang, Eric Chang, and Tao Chen. Accent issues in large vocabulary continuous speech recognition. Technical Report MSR-TR-2001-69, Microsoft Research, 2001a. 1

Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. In *NIPS*, pages 892–900, 2010. 24

Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. In *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, 2001b. 14

J. J. Humphries and Philip C. Woodland. Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition. In *EUROSPEECH*, 1997. 1, 3, 8

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. 3, 8, 17

Shajith Ikbal, Hemant Misra, Sunil Sivadas, Hynek Hermansky, and Hervé Bourlard. Entropy based combination of tandem representations for noise robust asr. In *INTERSPEECH*, 2004. 47

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. 23, 24, 25, 28

Dan Jurafsky, Wayne Ward, Zhang Jianping, Keith Herold, Yu Xiuyang, and Zhang Sen. What kind of pronunciation variation is hard for triphones to model? In *ICASSP*, pages 577–580, 2001. 17

Herman Kamper, Félicien Jeje Muamba Mukanya, and Thomas Niesler. Multi-accent acoustic modelling of south african english. *Speech Communication*, 54(6): 801–813, 2012. 4, 48

Thomas Kemp and Alex Waibel. Unsupervised training of a speech recognizer: recent experiments. In *EUROSPEECH*, 1999. 35

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. 48

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. 3

T. Li, Philip C. Woodland, Frank Diehl, and Mark J. F. Gales. Graphone model interpolation and arabic pronunciation generation. In *INTERSPEECH*, pages 2309–2312, 2011. 19

Xiao Li, Asela Gunawardana, and Alex Acero. Adapting grapheme-to-phoneme conversion for name recognition. In *ASRU*, pages 130–135, 2007. 18, 19

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. 8

Karen Livescu and James Glass. Lexical modeling of non-native speech for automatic speech recognition. In *ICASSP*, pages 1683 – 1686. 3, 17

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. 36

Jeff Z. Ma and Richard M. Schwartz. Unsupervised versus supervised training of acoustic models. In *INTERSPEECH*, pages 2374–2377, 2008. 36

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. 25, 37, 40

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. 1

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. 12, 30

Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *ACL (Short Papers)*, pages 220–224, 2010. 27

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. 3, 5, 8, 62, 63

Udhyakumar Nallasamy, Florian Metze, and Tanja Schultz. Enhanced polyphone decision tree adaptation for accented speech recognition. In *Interspeech*, 2012a. 1, 4, 30, 39, 61, 63

Udhyakumar Nallasamy, Florian Metze, and Tanja Schultz. Enhanced polyphone decision tree adaptation for accented speech recognition. In *MLSLP Symposium*, 2012b. 4, 27, 61, 63

Scott Novotney, Richard M. Schwartz, and Sanjeev Khudanpur. Unsupervised arabic dialect adaptation with self-training. In *INTERSPEECH*, pages 541–544, 2011. 3, 36

Daniel Povey and Kaisheng Yao. A basis representation of constrained mllr transforms for robust adaptation. *Computer Speech & Language*, 26(1):35–51, 2012. 39

Bhuvana Ramabhadran. Exploiting large quantities of spontaneous speech for unsupervised training of acoustic models. In *INTERSPEECH*, pages 1617–1620, 2005. 36

K. Reidhammer, T. Bocklet, A. Ghoshal, and D. Povey. Revisiting semi-continuous hidden markov models. In *ICASSP*, 2012. 10

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. 23

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. 3, 17

T. Schultz and A. Waibel. Polyphone decision tree specialization for language adaptation. In *ICASSP*, 2000. 8, 9, 14

Michael L. Seltzer and Alex Acero. Factored adaptation for separable compensation of speaker and environmental variability. In *ASRU*, pages 146–151, 2011. 48

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 23

Peter Smit and Mikko Kurimo. Using stacked transformations for recognizing foreign accented speech. In *ICASSP*, pages 5008–5011, 2011. 3, 8, 48

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. 1, 10

Hagen Soltau, Lidia Mangu, and Fadi Biadsy. From modern standard arabic to levantine asr: Leveraging gale for dialects. In *ASRU*, pages 266–271, 2011. 1, 3, 13, 44

Helmer Strik and Catia Cucchiarini. Modeling pronunciation variation for asr: A survey of the literature. *Speech Communication*, 29(2-4):225–246, 1999. 17

Sebastian Stüker. Modified polyphone decision tree specialization for porting multilingual grapheme based asr systems to new languages. In *ICASSP*, pages 4249–4252, 2008. 8

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. URL `http://dl.acm.org/citation.cfm?id=1564131.1564140`. 23

Laura Mayfield Tomokiyo. Lexical and acoustic modeling of non-native speech in lvscr. In *INTERSPEECH*, pages 346–349, 2000. 3, 8, 17

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. 23

Dimitra Vergyri, Lori Lamel, and Jean-Luc Gauvain. Automatic speech recognition of multiple accented english data. In *INTERSPEECH*, pages 1652–1655, 2010. 3, 8

Zhirong Wang and Tanja Schultz. Non-native spontaneous speech recognition through polyphone decision tree specialization. In *INTERSPEECH*, 2003. 3, 9

J.C. Wells. *Accents of English:*. Accents of English. Cambridge University Press, 1982. ISBN 9780521285407. URL `http://books.google.com/books?id=a3-ElL71fikC`. 2

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. 35

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), 2010a. 23, 24, 25, 37

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, 2010b. 23, 35

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. 4