ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Enhanced Polyphone Decision Tree Adaptation for Accented Speech Recognition

Udhyakumar Nallasamy[1], *Florian Metze*[1] *and Tanja Schultz*[1,2]

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Cognitive Systems Labs, Karlsruhe Institute of Technology, Karlsruhe, Germany
{unallasa, fmetze, tanja}@cs.cmu.edu

Abstract

State-of-the-art Automatic Speech Recognition (ASR) systems struggle to handle accented speech, particularly if the target accent is under-represented in the training data. The acoustic variations presented by an unfamiliar accent render the ASR polyphone decision tree (PDT) and its associated Gaussian mixture models (GMM) misfit to the test data. In this paper, we improve on the previous work of adapting the polyphone decision tree, using a semi-continuous model based approach to address the problem of data sparsity. We extend the existing PDT to introduce additional states with shared parameters, corresponding to the new contextual variations identified in the adaptation data, while still robustly estimating the state-specific parameters on a relatively small dataset. We conduct ASR experiments on Arabic and English accents and show that our technique performs better than Maximum A-Posteriori (MAP) adaptation and a previous implementation of polyphone decision tree specialization (PDTS). Compared to MAP adapted system, we obtain 7% relative improvement in Word Error Rate (WER) for Arabic and 13.7% relative improvement for English accent adaptation.
Index Terms: automatic speech recognition, accent adaptation

10.21437/Interspeech.2012-516

1. Introduction

With the pervasive adoption of speech interfaces in mobile and web applications, modern day ASRs are expected to handle speech input from a wide range of speakers with different accents. While a one-size-fits-all ASR may be the holy-grail, there is still practical benefit in adapting the existing ASR to a new regional accent with limited data. We note that formally the term 'Accent' refers to only pronunciation changes, while the term 'Dialect' stands for the ensemble of variations in vocabulary, syntax, pronunciation including prosody. Since we focus only on acoustic variations in this work, we use the term 'accent' throughout this paper.

ASR accent adaptation is usually carried out by adapting the following components: pronunciation dictionary by including accent-specific variants [1], polyphone decision tree [2] or Gaussian parameters in the acoustic model i.e means, variances and mixture weights using Maximum Likelihood Linear Regression (MLLR) or MAP estimation [3]. In this paper, we focus on the PDT to adapt an existing ASR to match the target accent that is either under-represented or unseen in the training data. PDTs are used to cluster context-dependent states in an ASR and hence play a crucial role in modeling contextual acoustic variants, which are usually the main source of accent differences. PDT adaptation is also complimentary to pronunciation and Gaussian adaptation and can be easily combined with those techniques for additional improvements.

Previous work on PDT adaptation has focused on extending the tree [2], pruning it [4] or both [5], based on the adaptation data. The newly created states (or contexts) estimated from the adaptation set better match the target data, resulting in improved performance. A drawback of these approaches is that the GMM parameters of these extended states need to be estimated from the relatively small adaptation set, which significantly limits the number of new contexts created. Given that any modification to the decision tree happens at the leaves in these techniques, few additional states have only a limited influence in the overall likelihood. To address this issue, we explore the idea of using semi-continuous models with tied parameters for PDT adaptation. These models have a very low ratio of state-specific to shared parameters, thus enabling far higher number of new states to be created and their parameters be reliably estimated from the small adaptation set.

We evaluate our technique, Semi-continuous PDT specialization (SPDTS) on Arabic and English accents and compare it with MAP adaptation and, PDTS without shared parameters. We show that SPDTS achieves better likelihoods on the adaptation data than other techniques. It also performs favorably in terms of WER on an unseen target accent test set.

2. Accent Adaptation

In this section, we discuss the role of PDT in the overall Hidden Markov Model (HMM) observation modeling and motivate the need for PDT adaptation to handle accent variations. We review the previous work in PDT adaptation and their limitations. We then describe the multi-codebook semi-continuous models and explore the use of these models and their associated two-level decision tree for PDT-based accent adaptation on a relatively small target dataset.

2.1. Polyphone Decision Tree

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/unvoiced, vowel/consonant, 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 [2] and non-native speech [6]. 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 overfitting.

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.

2.2. Semi-continuous Modeling

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 [7]. These models are being widely adopted in ASR having performed better than its counterparts, in both low-resource [7] and large-scale systems [8].

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 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.

3. Experiment Setup - Speech Corpus, Language Model and Lexicon

The training data for Arabic experiments come from Broadcast Conversations (BC) part of LDC GALE corpus. The BC corpus consists of conversational speech, which 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 [9]. 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. The higher perplexity of target accent dataset shows that there is a mismatch between the source and target accents at the text level, in addition to the acoustic variations. However, previous work [10] showed very small gains from LM adaptation compared to AM adaptation.

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 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. A UK English dictionary may be more appropriate to decode the target accent, but our experiments with BEEP UK dictionary didn't achieve any gains over CMU dict, when used during decoding or both training and decoding.

Table 2: *Database Statistics.*

Dataset	Accent	#Hours	Ppl	%OOV
Arabic				
Train-SRC	Mixed	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

4. 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 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 system consists of 3K states and 141K Gaussians.

The 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. 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 [10]. 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 doesn't 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 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

5. Accent Adaptation Experiments

We chose to evaluate accent adaptation with 3 different techniques: MAP adaptation, fully-continuous PDTS as formulated in [2] and semi-continuous PDTS or SPDTS. MLLR is also a possible candidate, but its improvement saturates after 600 utterances ($\approx$ 1 hour), when combined with MAP [11]. MLLR is also reported to have issues with accent adaptation [12]. The MAP smoothing factor τ is set to 10 in all cases. We didn't 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 1 shows that the adaptation data has the lowest cross-entropy on SPDTS adapted models compared to MAP and PDTS.

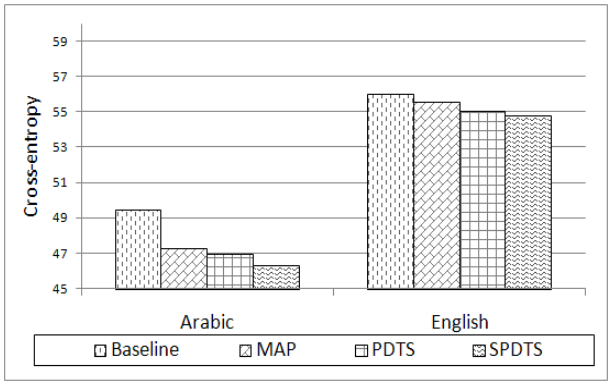

Figure 1: *Cross-entropy of adaptation data for various models*

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 4.

Table 4: *Evaluation of accent adaptation techniques.*

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

MAP adaptation achieves a relative improvement of 9.7% for Levantine Arabic 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 GALE evaluation ML system. We used a 2-pass unvowelized system trained on the GALE BN corpus for this experiment. More details of this system can be seen in [9]. The results are shown below

Table 5: *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

We get 5.1% relative improvement for SPDTS over MAP in adapting a large-scale ASR system trained on mostly BN MSA speech to BC Levantine Arabic. It is also interesting to note the limitation of PDTS for large systems as discussed in Section 2.1. 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.

6. Conclusion and Future work

We presented a semi-continuous approach to PDT specialization (SPDTS) and compared it with MAP and fully-continuous PDTS for the problem of accent adaptation. We showed that SPDTS with the adapted PDT and its associated models, performs better than MAP and PDTS interms of both average likelihood on the adaptation data and WER on the target test set. One of the interesting characteristic of SPDTS two-level tree is that, it can decouple accent independent and accent dependent parameters. This framework allows us to setup accent adaptive training for datasets with multiple accents, similar to SA training. The system will have accent-specific semi-continuous trees and states analogous to SAT transforms, while sharing common canonical codebooks. This idea will be explored in the near future.

7. References

[1] Humphries, J. J., Woodland, P. C., "Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition", Proc. Eurospeech, Rhodes, 1997.

[2] Schultz, T. and Waibel, A., "Polyphone Decision Tree Specialization for Language Adaptation", Proc. ICASSP, Istanbul, 2000.

[3] Smit, P. and Kurimo, M., "Using Stacked Transformations for Recognizing Foreign Accented Speech", Proc. ICASSP, Prague, 2011.

[4] Singh, R., Raj, B. and Stern, R. M., "Domain Adduced State Tying for Cross-Domain Acoustic Modelling", Proc. Eurospeech, Budapest, 1999.

[5] Stuker, S., "Modified Polyphone Decision Tree Specialization for Porting Multilingual Grapheme based ASR Systems to New Languages", Proc. ICASSP, Las Vegas, 2008.

[6] Wang, Z. and Schultz, T., "Non-Native Spontaneous Speech Recognition through Polyphone Decisision Tress Specialization", Proc. Interspeech, Geneva, 2003.

[7] Riedhammer, K., Bocklet, T., Ghoshal, A., and Povey, D., "Revisting Semi-continuous Hidden Markov Models", Proc. ICASSP, Tokyo, 2012.

[8] Soltau, H., et. al, "Advances in Arabic Speech Transcription at IBM Under the DARPA GALE Program", IEEE Trans on Audio, Speech and Language Proc., 17(5), 2009.

[9] Metze, F., Hsiao, R., Jin, Q., Nallasamy, U., and Schultz, T., "The 2010 CMU GALE Speech-to-Text System", Proc. Interspeech, Makuhari, 2010.

[10] Soltau, H., Mangu, L., and Biadsy, F., "From Modern Standard Arabic to Levantine ASR: Leveraging GALE for Dialects", Proc. ASRU, Hawaii, 2011.

[11] Huang, X., Acero, A., Hon, H. W., Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Prentice Hall, 2001.

[12] Clarke, C. and Jurafsky, D., "Limitations of MLLR Adaptation with Spanish-Accented English: An Error Analysis", Proc. Interspeech, Pittsburgh, 2006.

[13] Burget, L., Schwarz , P., et. al, "Multilingual Acoustic Modeling For Speech Recognition Based On Subspace Gaussian Mixture Models", Proc. ICASSP, Dallas, 2010.